The regulatory potential of splicing 3′ untranslated regions in cancer and stem cell differentiation

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield

School of Biosciences, Faculty of Science

Jack Joseph Riley

October 2024

Acknowledgements

I would like to thank my supervisors Dr. Ian Sudbery and Prof. Ivana Barbaric for their exceptional guidance and support throughout my PhD, I couldn't have asked for better mentors. I am deeply grateful to Ian for all the time and effort put into my bioinformatics training, and for answering any silly questions along the way. I would also like to extend my thanks to Prof. Stuart Wilson for welcoming me into his lab and providing guidance on the molecular biology side of the project. I am thankful to all my colleagues, past and present, in the Sudbery and Wilson labs: Josh, Ivo, Ang, Carmen, Cristina, Charlotte, Mikayla, Chiarra, Ashleigh, Vicky and Pete for their support and training. I am also thankful to my colleagues in the Barbaric lab and wider CSCB: Chris, Dylan, Theo, Gabi, Owen and Paul for their help working with hESCs, especially in the early days where things weren't going so well. It has been an amazing experience working between multiple groups, and I have learned so much from everyone.

I am immensely grateful to my friends and family for always supporting me. A special thanks to my partner Gosia for her unwavering support, especially when things got stressful. Finally, I would like to thank my parents for their encouragement and support over the years; without them, this work would not have been possible.

Abstract

3′ untranslated regions (3′UTRs) mediate various RNA-RNA and RNA-protein interactions and have important functions in the regulation of RNA stability, localization and translation efficiency. Therefore, 3′UTR splicing represents a potential mechanism to regulate these functions in a context-specific manner. However, such events are generally considered to act as signals to elicit transcript degradation via nonsense-mediated decay (NMD) and are often seen as transcriptional noise. By analysing large RNAseq datasets, we found that 3′UTR-spliced transcripts are widespread through the human pluripotent stem cell (hPSC) transcriptome and are often highly expressed. We also found that the usage of 3′UTR-spliced transcripts changed during stem cell differentiation. Additionally, 3′UTR introns are enriched for binding of RNA binding proteins and microRNAs, and 3′UTR-spliced transcripts often appear to evade NMD.

By establishing a hPSC-based cardiomyocyte differentiation model and conducting a transcriptome-wide RNA stability assay, we were able to interrogate the impact of 3′UTR splicing on RNA stability at the individual-event level, and how this changes during differentiation. First, taking a gene-level approach, we observed a transient global destabilization of RNA early in differentiation, followed by global stabilization upon terminal differentiation. Next, by comparing the half-lives of intron-spliced and intron-retaining isoforms, we found that splicing within the protein-coding region usually results in RNA stabilization; however, 3′UTR splicing results in an equal split between stabilization and destabilization. Finally, we found that the effect of 3′UTR splicing on RNA stability changes during differentiation, identified instances of differentiation-stage-specific regulation, and estimated the contribution of stability and splicing rate changes to changes in exon usage observed across differentiation.

In summary, 3′UTR splicing represents a relatively unexplored mechanism of post-transcriptional regulation. Such events are widespread through the hPSC transcriptome and contribute to RNA stability changes during differentiation.

Contents

List of Figures

Abbreviations

3′SS	3′ splice site
3UI	3′UTR intron
3′UTR	3′ untranslated region
4sU	4-thiouridine
5′SS	5′ splice site
ANOVA	analysis of variance
APA	alternative polyadenylation
ARE	AU-rich element
AS	alternative splicing
AS-NMD	alternative splicing coupled nonsense-mediated decay
ASO	antisense oligonucleotide
CDS	coding sequence
DEU	differential exon usage
DGE	differential gene expression
DTE	differential transcript expression
DTU	differential transcript usage
ECDF	empirical cumulative distribution function
EJC	exon junction complex
GO	gene ontology
GSEA	gene set enrichment analysis
hESC	human embryonic stem cell
hiPSC	human induced pluripotent stem cell
HipSci	Human Induced Pluripotent Stem Cell Initiative
hPSC	human pluripotent stem cell
IAA	iodoacetamide
KEGG	Kyoto encyclopedia of genes and genomes
m6A	N6-methyladenosine
miRNA	microRNA
MRE	microRNA response element

mRNA	messenger RNA
mRNP	messenger ribonucleoprotein
NMD	nonsense-mediated decay
PAS	polyadenylation site
PCR	polymerase chain reaction
poly(A)	polyadenosine
PSI	percent spliced in
PSO	percent spliced out
PTC	premature termination codon
qPCR	quantitative polymerase chain reaction
RBP	RNA binding protein
RNAi	RNA interference
SD	standard deviation
SEM	standard error of the mean
siRNA	small interfering RNA
SNP	single nucleotide polymorphism
TCGA	The Cancer Genome Atlas
TPM	transcripts per million
WCSS	within cluster sum of squares

1. Introduction

1.1 Importance of studying gene expression in development

The human body is made up of trillions of cells, comprising hundreds of distinct cell types, each with specialized functions resulting from the unique RNA and protein expression profiles of each cell type. Despite this, all cells share the same genetic material, and their lineage can ultimately be traced back to a single cell, the zygote. Here we introduce the process of early human embryonic development to highlight the importance of understanding the regulation of gene expression. This is particularly apparent where cells are restricted in their differentiation capacity - their potency - as development proceeds.

1.1.1 Early development and formation of the germ layers

Following fertilization, the zygote undergoes a series of mitotic divisions, producing multiple blastomeres (Figure 1.1A). By the 16-cell (morula) stage, blastomeres are present as a compact mass. This is considered the last stage at which the cells are "totipotent", i.e. able to differentiate into any embryonic or extraembryonic tissue. The morula then undergoes differentiation and blastulation (Figure 1.1A). This results in the formation of a blastocyst, a hollow ball of cells (trophoblasts; later contributing to the placenta) with a fluid filled cavity, and the inner cell mass (ICM; Figure 1.1A). Cells of the inner cell mass are "pluripotent", and give rise to the entire embryo. Following implantation, the blastocyst undergoes gastrulation, giving rise to the three primary germ layers: the ectoderm, mesoderm and endoderm. Whilst the 3 germ layers contribute to all the distinct cell types within the human body, each layer alone is restricted to specific cell types, as highlighted in Figure 1.1B. Therefore, these cells represent the first "multipotent" cells in the embryo.

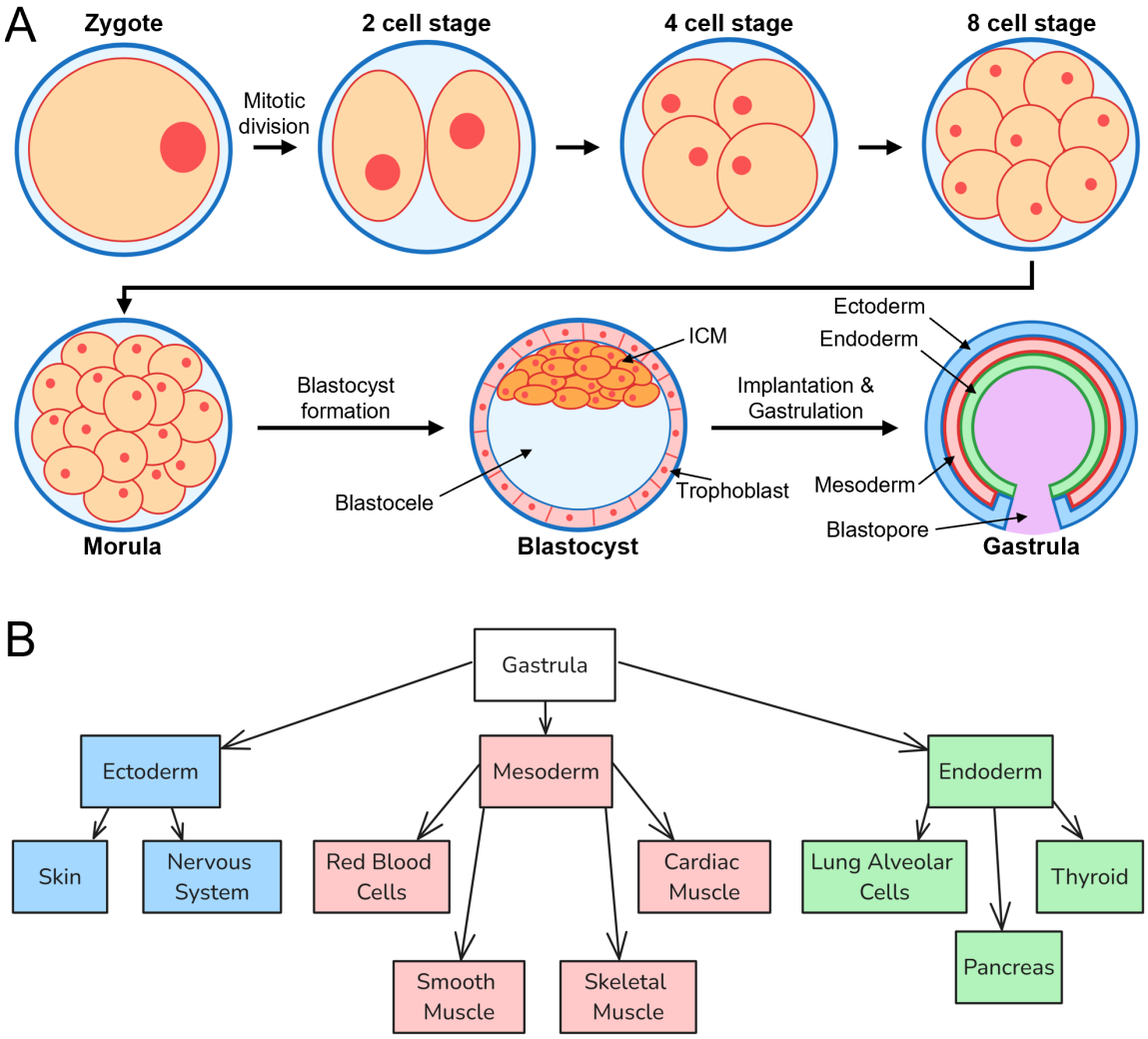

Figure 1.1: Early embryogenesis and germ layers. A) stages of early embryonic development up to the formation of the embryonic germ layers (ectoderm, endoderm, mesoderm) in the gastrula. B) Contribution of each embryonic germ layer to different tissues of the body. Several (non-exhaustive) examples are provided for each lineage.

Understanding how gene expression changes, and the processes underpinning its regulation, are crucial for our understanding of human development, as well as aging, disease states, and cancer. Central to this is regulation of mRNA at both the transcriptional and post-transcriptional levels. Whilst this thesis focuses particularly on regulation of mRNA at the post-transcriptional level, here we briefly introduce the key mechanisms underpinning the processing of mRNA into its mature form, and the

features of mature mRNA.

1.1.2 From DNA to mRNA

Transcription of protein coding genes is performed by RNA polymerase (Pol) II, resulting in the production of precursor messenger RNA (pre-mRNA), which is then subjected to a number of modifications necessary for its maturation into mRNA. Pre-mRNA processing occurs co-transcriptionally, including the addition of the 5′ cap, intron splicing, cleavage and polyadenylation. This leads to the production of mature mRNA (Figure 1.2). Mature mRNA contains an N7-methylguanosine (m7G) cap at the 5′ end, are polyadenylated at their 3′ termini, and their coding region (CDS) is flanked by untranslated regions (UTRs) at the 5′ (5′UTR) and 3′ (3′UTR) end. A more in depth overview of the processes of polyadenylation and splicing are discussed in Sections 1.3 and 1.4 respectively, and in both instances we explore how these processes can be regulated depending on cellular context. In reality, mature mRNA does not exist in the naked form depicted in Figure 1.2; instead, it is bound by various interacting proteins. Together, the mRNA and its associated proteins are termed the messenger ribonucleoprotein (mRNP). Various proteins and protein complexes are added to the mRNP during pre-mRNA processing. For example, the cap binding complex (CBC) binds to the m7G cap, whilst exon junction complexes (EJCs) are deposited following intron splicing (see Section 1.4 for further detail).

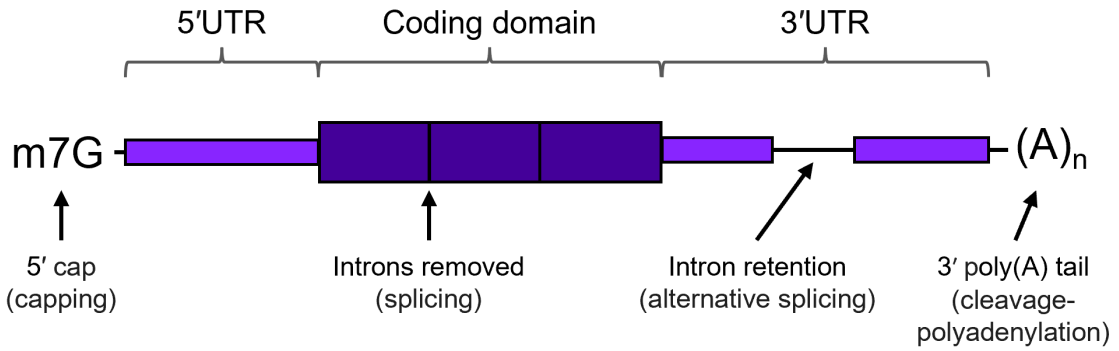

Figure 1.2: Pre-mRNA processing steps and mature mRNA structure. Pre-mRNA processing steps are shown in brackets underneath each of the features they contribute towards. Coding region is shaded dark purple, UTRs are depicted by light purple.

It is important to appreciate how interconnected pre-mRNA processing steps are. For example, 5′ capping is intimately linked with transcription initiation, where capping enzymes associate directly with the C-terminal domain of RNA Pol II following phosphorylation of Ser5 (Cho et al., 1997; Ho and Shuman, 1999). Splicing also impacts Pol II pause release and productive elongation, where inhibiting splicing with pladienolide B leads to increased Pol II pausing in the promoter proximal region (Caizzi et al., 2021). Additionally, the speed of transcription impacts splicing fidelity in yeast, where fast transcribing mutants displayed lower splicing efficiencies compared to wild-type and slow mutants (Aslanzadeh et al., 2018).

1.1.3 From mRNA to proteins

Following its maturation, mRNA is exported from the nucleus by the TREX complex. This process is also intimately linked to pre-mRNA processing, particularly 5′ capping and splicing, where the recruitment of TREX component ALYREF has been shown to be dependent on the CBC and EJC component eIF4A3 (Gromadzka et al., 2016; Viphakone et al., 2019). However, nuclear export of RNA can also occur independently of splicing, given that 8.9% of protein coding genes only have a single exon, and are therefore

unspliced (Jorquera et al., 2016), yet are still exported and translated. Translation of RNA subsequently occurs in the cytoplasm, although the site of translation can also be specialized for function, for example localized translation in neuronal dendrites is important for synaptic plasticity (Holt et al., 2019). Translation is also coupled with RNA degradation. This is the case for truncated mRNAs which contain premature termination codons (PTCs), either as a result of mutation or processing errors. Degradation of such transcripts is facilitated via nonsense-mediated RNA decay (NMD), which will be discussed in detail in Section 1.5.

The composition of the mature mRNA has a substantial impact on how well it is translated, turned over, and localized. The CDS is well studied in this regard, given that it encodes the amino acid chain of the proteins produced from the mRNA; therefore, its codon optimality can impact translation efficiency and stability (Bae and Coller, 2022). However, the contributions of UTRs is generally less well studied. The UTR sequences do not directly contribute to the protein coding sequence; however, they have important regulatory roles in how well RNA is translated, for example the presence of upstream open reading frames (uORFs) in 5'UTRs (Calvo et al., 2009), and the binding of trans factors to the 3'UTR (discussed in Section 1.2). 3'UTRs also have important roles in the regulation of RNA stability and transcript localization (discussed in Sections 1.2.1 and 1.2.2), which subsequently impacts how well, and the sites at which, their proteins are expressed.

1.2 Functions of the 3' untranslated region (3'UTR)

3'UTRs play important roles in regulation at the RNA level through mediating numerous RNA-RNA and RNA-protein interactions within the mRNP. Cis elements within the 3'UTR act as binding sites for trans regulators (Figure 1.3) such as microRNAs (miRNAs) and RNA binding proteins (RBPs), which have downstream effects on RNA stability (Section 1.2.1), localization (Section 1.2.2), and translation efficiency (Section 1.2.3).

Despite these important regulatory functions, 3′UTRs are less well studied than the corresponding CDS (Mayr, 2019). Interestingly, the length of 3′UTRs correlates exponentially with the number of cell types an organism has (Chen et al., 2012), despite a similar number of protein coding genes between organisms studied (Hillier et al., 2005; Ezkurdia et al., 2014; Mayr, 2016). Additionally, the length of 3′UTRs of individual genes is regulated during development through the use of alternative polyadenylation sites (Ji et al., 2009; Miura et al., 2013), as will be discussed in Section 1.3. Together these findings point towards 3′UTRs as important regulators of biological complexity, and a platform for post-transcriptional regulation of gene expression beyond the amino acid sequence. These regulatory avenues will be discussed herein.

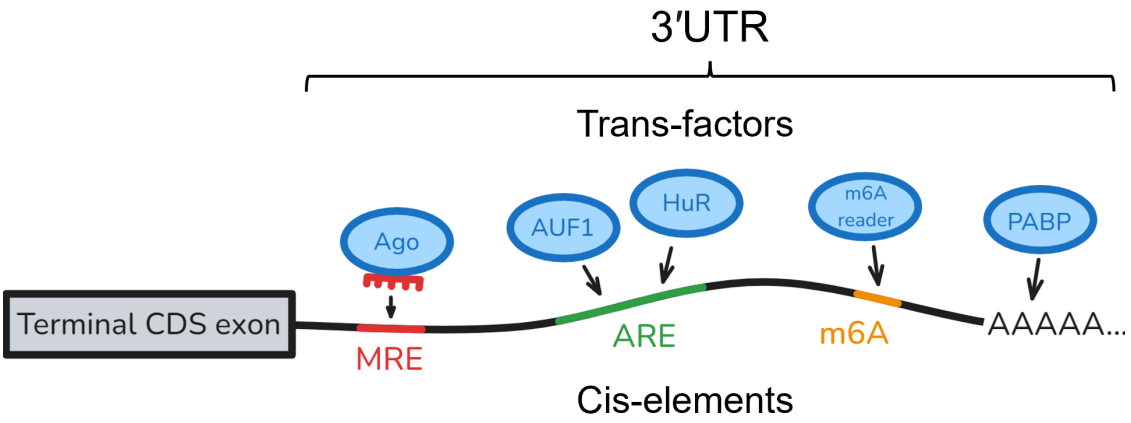

Figure 1.3: 3′UTRs act as a binding platform for trans factors. The 3′UTR contains many cis-elements (below line), several (non-exhaustive) examples are shown: microRNA response elements (MREs; red), AU-rich elements (AREs; green), N^6-methyladenosine (m6A; orange). These cis-element act as binding sites for various trans factors. Several (non-exhaustive) examples are shown (above line).

1.2.1 Regulation of transcript stability

Regulating RNA stability is crucial in the life of an mRNA and the nature of how it is expressed. Where RNA is less stable, changes in the transcriptional output of a gene are more rapidly reflected at the protein level. For example, where transcription is shut off for a given gene, the pool of corresponding RNA will also be rapidly degraded, meaning

proteins cannot be produced. On the other hand, where RNA is very stable, the corresponding RNA pool can act as a buffer between transcriptional shutoff and changes at the protein level. This is also reflected upon upregulation, where very stable RNA will reach their new, higher levels, more slowly (Alon, 2006). Such regulation likely differs on a gene-by-gene basis depending on their function. Additionally, transcript stability may change in a context-specific manner due to modification of the mRNP composition. In this regard, the 3'UTR plays an important role as a binding platform for such trans factors.

3'UTRs regulate RNA stability through the binding of RNA binding proteins (RBPs). Many RBPs bind directly to sequence elements present in the 3'UTR. However, Argonaute proteins act differently, they are first loaded with a miRNA, forming the RNA-induced silencing complex (RISC), and subsequently guided to their targets by miRNA-mRNA interactions. The binding of RBPs to the 3'UTR can have both stabilizing or destabilizing effects on the RNA, depending on the RBP that binds, and also on the cellular context. Additionally, given that multiple RBPs and miRNAs can bind the same mRNA, cooperation or competition between trans factors for occupancy within the mRNP increases the complexity of regulation. Although the full extent of mRNA-RBP interactions is still under investigation (Gerstberger et al., 2014; Van Nostrand et al., 2020), recent studies have attempted to systematically identify RBP interaction modules and attribute these to post-transcriptional regulatory functions (Khoroshkin et al., 2024). In this section we explore several well-established interactions between RBPs and 3'UTRs that impact RNA stability.

1.2.1.1 AU-rich element (ARE) binding proteins

AU-rich elements (AREs) are cis elements within 3'UTRs that bind ARE-binding proteins (Figure 1.3). Following the discovery that 3'UTRs contain cis elements that regulate RNA stability and expression (Miller et al., 1984; Treisman, 1985), Shaw and Kamen (1986) identified a highly conserved AU-rich sequence within the 3'UTR of the

granulocyte-monocyte colony stimulating factor (GM-CSF) gene, which they hypothesised contributed to its instability upon stimulation. By inserting AU-rich sequence elements into the rabbit β-globin 3'UTR, which is very stable endogenously, they observed a substantial destabilization.

One of the first identified AREs was ARE binding factor 1 (AUF1) (Brewer, 1991), which is now more commonly referred to as heterogeneous nuclear ribonucleoprotein D (HNRNPD). Using an in vitro cell-free RNA decay model, Zhang et al. (1993) showed that AUF1 binds to AREs within the 3'UTR of its targets, including *c-fos* and *c-myc*, and its binding was abrogated in ARE mutants. The group later found that the affinity between AUF1 and the AREs within its targets correlated with its RNA destabilizing function (DeMaria and Brewer, 1996). As well as binding AU-rich elements, AUF1 also binds U- and GU-rich elements (Yoon et al., 2014). Interestingly, AUF1 has been shown to bind its own 3'UTR (Wilson et al., 1999), potentially regulating its own stability and expression. This will be discussed further in Section 1.6.1.

Whilst AU-rich elements are often considered a marker of transcript destabilization, they can also facilitate RNA stabilization, depending on the ARE-binding protein they attract. In this regard, overexpression of the ARE-binding protein human antigen R (HuR), also known as ELAV-like RNA binding protein 1 (ELAVL1), led to the stabilization of ARE-containing mRNA in parallel studies (Fan and Steitz, 1998; Peng et al., 1998). Meanwhile, knockdown of HuR had a negative impact on the expression and mRNA stability of GFP-ARE constructs (Raineri et al., 2004). HuR competes with other ARE-binding proteins for occupancy of AREs; however, it also has been shown to overlap with the binding of Argonaute protein binding at or near miRNA sites (Mukherjee et al., 2011; Lebedeva et al., 2011), which may either repress (Bhattacharyya et al., 2006) or promote (Kim et al., 2009) miRNA-mediated destabilization.

1.2.1.2 miRNA-guided binding of Argonaute proteins

miRNAs are short non-coding regulatory RNAs that bind to 3'UTR cis elements known as miRNA response elements (MREs; Figure 1.3), which are complementary to the seed sequence (nucleotides 2-7) of the miRNA (Lewis et al., 2005). Binding of miRNAs to mRNA plays an important role in regulating gene expression. Specifically in the context of RNA stability, miRNA binding is generally associated with transcript destabilization. It is estimated that there are approximately 2300 mature miRNAs in humans (Alles et al., 2019), and selective pressure has maintained these miRNA-mRNA interaction in >60% of human genes (Friedman et al., 2009). miRNAs are ultimately derived from longer precursors known as primary-miRNAs (pri-miRNAs), which are characterized by a hairpin structure. Drosha and its binding partner DGCR8 bind the pri-mRNA hairpin in the nucleus and cleave it, producing a precursor-miRNA (pre-miRNA), which is subsequently exported to the cytoplasm. pre-miRNA is then bound by Dicer and cleaved, to produce a double-stranded miRNA duplex, which associates with Argonaute family proteins. Finally one of the two miRNA strands within the duplex is discarded, and the remaining miRNA guides Argonaute towards its mRNA targets. Destabilization is conferred through recruitment of additional protein complexes such as CCR4-NOT which leads to deadenylation (Fabian et al., 2011). Whilst all 4 Argonaute proteins are capable of this process, Ago2 is the most characterized given that it is the only member with intrinsic endonucleolytic activity, allowing it to cleave its targets directly (Meister et al., 2004). This also enables Ago2 to cleave pre-miRNA independently of Dicer (Cheloufi et al., 2010; Cifuentes et al., 2010), and to process exogenous RNA duplexes, constituting the basis of RNAi.

Much like mRNAs, the expression of miRNAs is regulated during development. The miR-302 cluster represents some of the most abundant miRNAs in undifferentiated hESCs (Bar et al., 2008), and their expression is capable of reprogramming somatic cells (Anokye-Danso et al., 2011). miR-302 cluster miRNAs promote hESC self-renewal and

prevents apoptosis in culture (Zhang et al., 2015), and their expression is dependent on pluripotency transcription factors Oct4 and Sox2 (Card et al., 2008), which were shown to bind directly to miRNA promoters in mouse ESCs (Marson et al., 2008). High levels of miR-302 cluster expression is restricted to the hESC state, and their expression is dramatically reduced upon differentiation (Kwon and Song, 2016). In addition to the miR-302 cluster, many miRNAs are expressed in a cell-type specific manner, where global miRNA expression profiling across differentiation via miRNA-seq has revealed differentiation stage-restricted expression of numerous miRNAs (Fedorova et al., 2023).

As well as the expression of each miRNA, the MRE position within the context of the whole 3'UTR may impact its targeting efficacy, with sites at the beginning of the 3'UTR being more efficacious (Grimson et al., 2007; Wu et al., 2017). Additionally, MREs found in AU-rich regions (Grimson et al., 2007) and those that have more accessible secondary structures (Long et al., 2007) are also more efficacious. Such variables have been incorporated into the TargetScan algorithm (Garcia et al., 2011).

1.2.1.3 N^6-Methyladenosine (m6A) writers, readers, and erasers

N^6-Methyladenosine (m6A) is the most common and most studied base modification found in mRNA (Roundtree et al., 2017). It is estimated that between 0.15-0.6% of adenosine residues found in mammalian mRNA are methylated at the N^6 amino group (He and He, 2021). m6A is found at DRACH motifs; however, the presence of a DRACH motif alone is not sufficient for m6A modification, given that only an estimated 5% of motifs are methylated (He and He, 2021). m6A is deposited by RNA binding proteins known as "m6A writers" and it can be removed through the action of demethylases known as "m6A erasers" (Wang et al., 2022). The reciprocal function of writers and erasers allows m6A modification to be dynamic, and regulated depending on cellular context (Yang et al., 2020; Wang et al., 2022). m6A is recognised by a third class of m6A associated proteins called "m6A readers". m6A readers act as the effector proteins, conveying

downstream effects on translation (Meyer et al., 2015), splicing (Zhao et al., 2014; Pendleton et al., 2017), and RNA stability. Whilst there are examples of RNA stabilization due to m6A modification (Huang et al., 2018; Wu et al., 2019), the majority of examples indicate an overall destabilizing effect (Wang et al., 2014; Du et al., 2016; Bertero et al., 2018; Collignon et al., 2023).

The distribution of m6A in mRNA is uneven, and there is an enrichment of m6A in long internal exons, near stop codons, and in 3'UTRs (Dominissini et al., 2012; Meyer et al., 2012; Linder et al., 2015). This highlights the importance of the 3'UTR as a site to harbour m6A bases, which will subsequently attract m6A readers, and impact RNA stability. Recently, 2 parallel studies showed that the presence of an exon junction complex (EJC) in the mRNP as a result of splicing (see Section 1.4) prevents m6A modification within its locality, creating a so-called "m6A exclusion zone" (Yang et al., 2022; Uzonyi et al., 2023). This indicates that m6A modification is influenced by the local mRNP composition, and could be regulated by splicing the 3'UTR.

1.2.2 Regulation of transcript localization

3'UTRs can also harbour cis elements that bind trans effector proteins with functions related to localization. A classic example of the regulation of transcript localization by 3'UTR cis elements came from the study of *bicoid* mRNA localization in Drosophilla embryogenesis by Macdonald and Struhl (1988). *bicoid* mRNA is transcribed maternally during oogenesis; however, is not translated until the eggs are laid (Driever and Nüsslein-Volhard, 1988). Upon translation, the Bicoid protein forms an anterior to posterior morphogen gradient in the developing Drosophilla embryo (Driever and Nüsslein-Volhard, 1988). It was originally hypothesised that *bicoid* mRNA was deposited only at the anterior pole, and subsequent diffusion of the Bicoid protein contributed to gradient of expression observed (Driever and Nüsslein-Volhard, 1988). However, later studies showed that *bicoid* mRNA also forms a gradient of expression which is highly similar to the protein expression profile (Spirov et al., 2009). This suggests that regulation

of localization at the mRNA level is key. Macdonald and Struhl (1988) revealed a 625 nucleotide region responsible for the anterior localization of *bicoid* mRNA, which was later shown to be dependent on Staufen (St Johnston et al., 1991; Ferrandon et al., 1994). More recently it was shown that the ESCRT-II complex protein Vps36 directly binds stem loop V of the *bicoid* 3'UTR, and mutation of Vps22 interfered with the anterior localization of both Staufen protein and *bicoid* mRNA (Irion and St Johnston, 2007).

The localization of *mbpa* mRNA to myelin sheaths during myelination has been attributed to 3'UTR cis-elements (Ainger et al., 1997; Yergert et al., 2021). Additionally, a 54 nucleotide "zipcode" sequence within the 3'UTR of β-*actin* mRNA has been shown to be necessary for its localization to growth cones (Kislauskis et al., 1994; Zhang et al., 2001). 3'UTR mediated localization in this manner was later shown to be due to the binding of zipcode binding protein 1 (ZBP1) (Oleynikov and Singer, 2003; Farina et al., 2003).

1.2.3 Regulation of translation efficiency

Whilst we previously introduced the structure of the mRNP in a simplified manner in Figure 1.2 as a linearised molecule, in reality the mRNP takes a more complex secondary structure. The regulation of translation efficiency by the 3'UTR highlights the importance of this. There is a great body of evidence pointing towards a "closed loop" conformation (Figure 1.4), where the 5' and 3' ends of mRNA are brought together through interactions between 5' cap- and poly(A) tail-associated proteins (Vicens et al., 2018). Interactions between eukaryotic initiation factor (eIF)4E, eIF4G and poly(A)-binding protein (PABP) have been shown to be central to this process (Wells et al., 1998; Borman et al., 2000). Depletion of PABP resulted in a reduction in translation efficiency of reporter constructs, which could be rescued by reintroducing wild-type PABP, but not mutant PABP which could not interact with eIF4G (Kahvejian et al., 2005). This highlights the importance of the 5'-3' interaction on translation efficiency, and demonstrates that proteins that are generally associated with the 3' end of mRNA can have profound effects on translation.

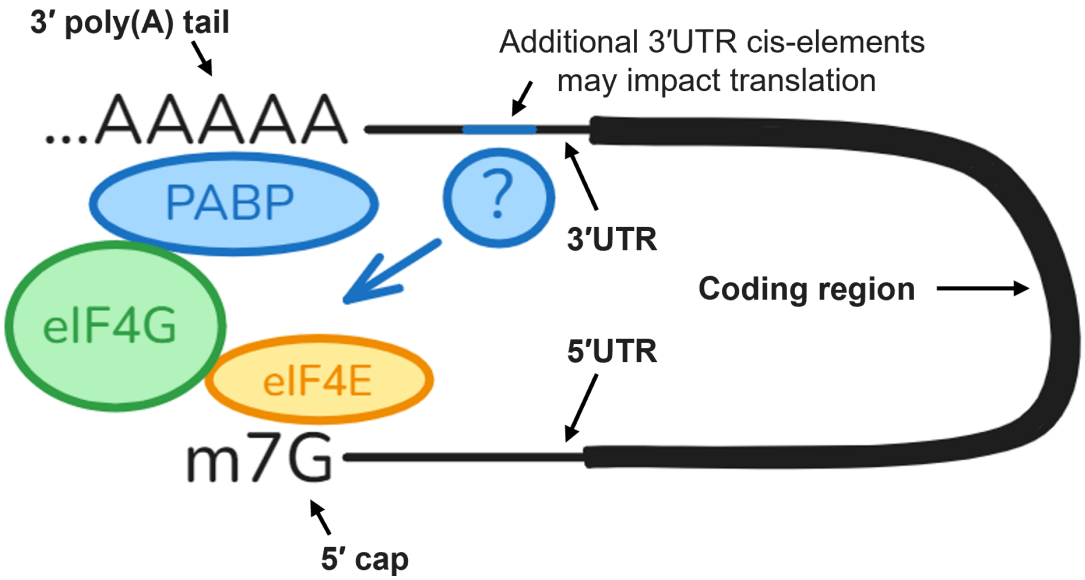

Figure 1.4: Closed loop mRNP conformation. The mRNP forms a closed loop structure through interactions between 3′ and 5′ associated factors. Interactions between eIF4E, eIF4G and PABP are central to this.

Turning back to some previous examples, whilst we previously focused on miRNAs in the context of transcript destabilization, miRNAs can also represses translation (Eichhorn et al., 2014). Whilst miRNA-mediated repression of translation is incompletely understood, it has been strongly linked to cap-dependent translation initiation (Mathonnet et al., 2007; Naeli et al., 2023), whereby the recruitment of CCR4-NOT impacts translation independently of its function as a deadenylase (Cooke et al., 2010; Bawankar et al., 2013). The binding of additional RBPs to the 3′UTR has also been shown to impact translation. For example, SRSF3 inhibits translation of p21 mRNA through binding its 3′UTR, potentially through interaction with eIF4A1 (Kim et al., 2022). Additionally, the binding of HuR to target 3′UTRs can both promote or repress translation, depending on the target (Popovitchenko et al., 2016; Zhang et al., 2020). This demonstrates that the local mRNP composition within the "closed loop" conformation impacts the translation of the mRNA. Therefore, the composition of cis-elements within the 3′UTR could affect this. Regulation in this manner is made more complex through

differential expression of trans factors such as miRNAs and RBPs in different cell types, in combination with the dynamic regulation of 3'UTR composition via alternative polyadenylation (Section 1.3) and alternative splicing (Section 1.4).

1.3 Alternative polyadenylation regulates 3'UTR length

Cleavage and polyadenylation (CPA) defines the 3' end of mRNA during pre-mRNA processing. CPA is carried out by 4 core protein complexes: cleavage and polyadenylation specificity factor (CPSF), cleavage stimulating factor (CSTF), and cleavage factors I and II (CFI and CFII). The site of CPA is determined by the 3'UTR sequence composition, where each of the above complexes associates with different sequence elements (Figure 1.5A). The canonical poly(A) signal (PAS) is defined by the AAUAAA hexamer (Proudfoot and Brownlee, 1976) which is directly bound by CPSF (Keller et al., 1991). 3' cleavage subsequently occurs 10-30nt (median = 21nt) downstream of the PAS (Tian et al., 2005; Gruber et al., 2016). The 3' cleavage site is also flanked downstream by G/U-rich elements which are bound by CSTF (Proudfoot, 1991; Takagaki and Manley, 1997), whilst CFI associates with UGUA motifs upstream of the PAS (Yang et al., 2010). Following cleavage, the free 3' termini is polyadenylated through the action of poly(A) polymerase (PAP)(Figure 1.5A).

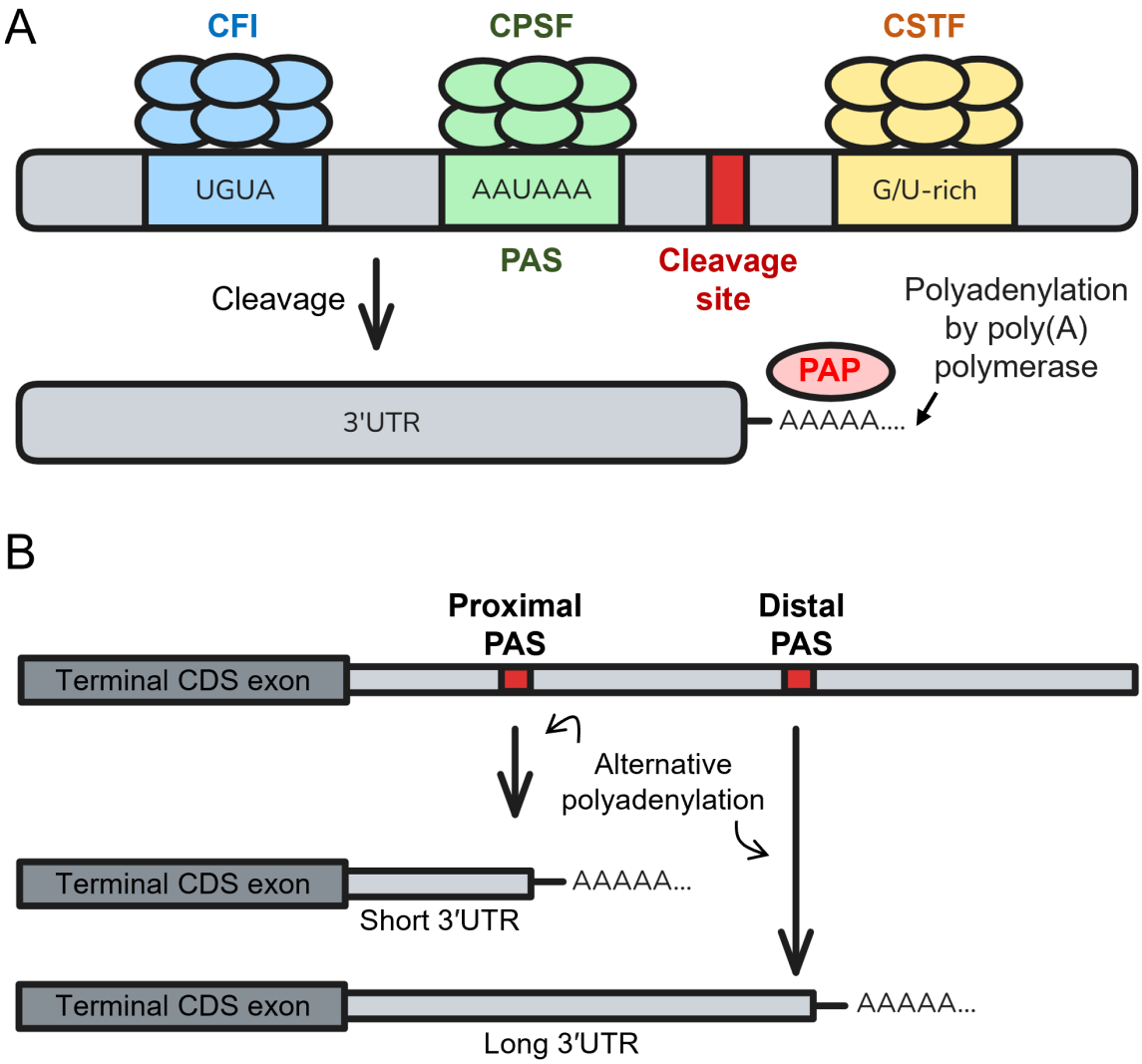

Figure 1.5: Cleavage and Polyadenylation (CPA) and Alternative Polyadenylation (APA). A) Protein complexes (above blocks) and cis elements (on/below blocks) that define 3′ end formation. CFI = cleavage factor I; CPSF = cleavage and polyadenylation specificity factor; CSTF = cleavage stimulating factor; PAP = poly(A) polymerase; PAS = polyadenylation site. B) Where multiple PAS are present, selection of either the proximal or distal PAS can change the length of the 3′UTR, this process is known as alternative polyadenylation.

Additional cis elements within the vicinity of the PAS may recruit trans factors that strengthen or weaken the PAS (Giammartino et al., 2011). This is important given that the majority of human genes contain multiple polyadenylation sites (Tian et al., 2005).

Subsequently, alternative polyadenylation (APA), through differential usage of proximal or distal sites (Figure 1.5B), regulates 3'UTR length in a context-specific manner in development and disease. As such, the post-transcriptional regulatory functions of 3'UTRs are also regulated through modifying the composition of cis-elements in the 3'UTR that recruit trans effector proteins.

1.3.1 APA in development and stem cell differentiation

By measuring the relative usage of distal polyA sites (RUD) during mouse embryonic and postnatal development, Ji et al. (2009) found that 3'UTRs became progressively longer, at the whole organism level, as development proceeds. They also observed progressive lengthening of 3'UTRs in the brain throughout both embryonic development and early post-natal development. Interestingly, in the testes, they observed 3'UTR lengthening only during embryonic development, followed by a switch to proximal polyA site usage between postnatal weeks 5-12. Consistent with these findings, in zebrafish development Li et al. (2012) observed a significant shortening of 3'UTRs between the zygotic and blastula stages, followed by 3'UTR lengthening thereafter. This suggests that 3'UTR shortening may be important at stages where rapid cell division is required. 3'UTR shortening is observed upon mouse T cell activation (Sandberg et al., 2008) and differentiation of hematopoietic stem cells into multipotent progenitor cells (Sommerkamp et al., 2020). One possible explanation for increased proliferation following 3'UTR shortening is the loss of miRNA binding sites within the 3'UTRs of pro-proliferative genes (Sandberg et al., 2008), which could increase their expression at the transcript level, or release their transcripts from miRNA-mediated translation inhibition. Conversely, 3'UTR shortening may increase the efficacy of MREs found near the middle of 3'UTR and 5' to alternative PAS sites. It has been shown that highly conserved MREs are enriched immediately 5' to PAS sites in some pro-differentiation genes (Hoffman et al., 2016), suggesting that increased usage of the proximal PAS sites in these genes might increase the efficacy of the neighbouring 5' MREs, which could

destabilize or translationally repress these transcripts (Nam et al., 2014; Hoffman et al., 2016). In the context of cellular differentiation, transient 3'UTR shortening at the early stages of differentiation may help promote rapid expansion of the cell population in "transit amplifying" cells, which only exist briefly (Mueller et al., 2013).

1.3.2 APA in senescence and cancer

3'UTR lengthening via APA is also associated with cellular senescence. Chen et al. (2018) showed that senescent mouse embryonic fibroblast cells have longer 3'UTRs in genes from pathways related to cellular senescence, compared with cells from earlier passages. Additionally, APA-mediated lengthening of the RRAS 3'UTR in senescent human cells leads to a decrease in its overall protein expression owing to less stable RNA compared to the RRAS short 3'UTR isoform (Chen et al., 2018). Similarly, APA-mediated lengthening of the MDM2 3'UTR coincides with a decrease in gene expression, and a subsequent increase in p53 expression during cellular senescence in rat cells (Wang et al., 2020). Whilst the CDK16 3'UTR is lengthened due to APA in cellular senescence, it is shortened in lung cancers (adenocarcinoma and squamous cell carcinoma) which stabilizes its transcripts through evading miR-485-5p binding, and promotes cell growth (Jia et al., 2022). Global 3'UTR shortening also promotes acute myeloid leukemia by preventing the differentiation of immature cells (Davis et al., 2022). Such global 3'UTR shortening is widespread across multiple tumour types (Xia et al., 2014), often leading to the evasion of miRNA binding, stabilization, and upregulation of oncogenes (Mayr and Bartel, 2009).

1.3.3 Functional consequences of regulating 3'UTR length by APA

By regulating 3'UTR length through APA, the composition of 3'UTR cis-elements, and subsequently the binding of trans factors within the mRNP can be modified in a cell context-specific manner. This in turn will impact how the transcript is regulated at the RNA level, potentially leading to changes in RNA stability, localization and translation

efficiency. Interestingly, 3'UTR-mediated protein-protein interactions between transcripts and the proteins they encode can impart additional functionality to nascent proteins, meaning proteins derived from transcripts with long or short 3'UTRs can have different functions despite sharing the same amino acid sequence (Figure 1.6)(Mayr, 2019).

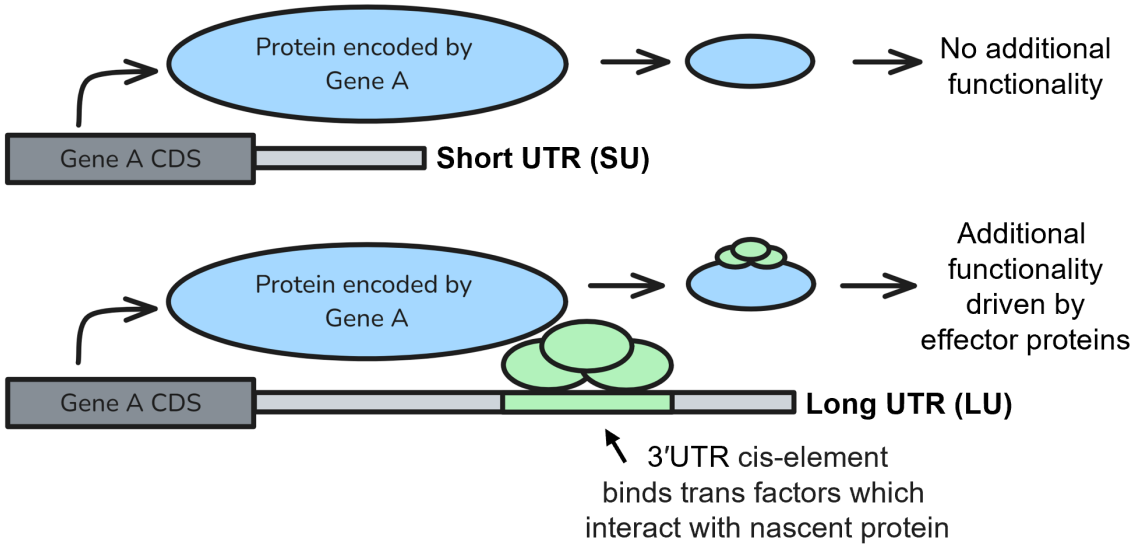

Figure 1.6: 3'UTR-mediated protein-protein interactions. The 3'UTR can harbour cis-elements that bind trans factors that subsequently interact with nascent proteins derived from the transcript. Therefore, where 3'UTR length is changed through APA, the composition of protein complexes produced can be modified. This can impart additional functionality to the nascent protein depending on the function of interacting proteins (e.g. changes in protein stability and localization).

CD47 APA is an example of this, leading to the production of either CD47-LU (encoded by mRNA with long 3'UTRs) or CD47-SU (encoded by mRNA with short 3'UTRs) proteins, which share the same amino acid sequence, but vary in function (Berkovits and Mayr, 2015). CD47-LU is expressed at the cell surface and contributes to cell migration, whilst CD47-SU localizes predominantly with the endoplasmic reticulum. This difference is mediated by HuR, which binds the long CD47 3'UTR but not the short CD47 3'UTR. Knockdown of HuR did not impact total mRNA or protein abundance, but led to a reduction in cell surface CD47 expression. Similar effects on cell surface expression were observed upon knockdown of either SET or RAC1, suggesting that these proteins form a

complex that only interacts with CD47-LU and not CD47-SU.

BIRC3 APA is as a second example, where BIRC3-LU or BIRC3-SU proteins are produced depending on poly(A) site usage, these again share the same amino acid sequence but BIRC3-LU has additional functionality (Lee and Mayr, 2019). This functionality is imparted through the protein complexes that BIRC3-LU forms that BIRC3-SU does not. During BIRC3-LU translation, IQGAP1 and RALA (amongst other factors) are recruited specifically to the long 3'UTR containing transcripts. Subsequently, the BIRC3-LU:IQGAP1:RALA complex facilitates CXCR4 cell surface expression, which is involved in B-cell migration. Such functionality is proposed to promote survival of malignant B-cells in chronic lymphocytic leukaemia, where a significant increase in distal poly(A) site usage is observed compared to non-malignant B-cells.

1.4 RNA splicing

Introns are removed from pre-mRNA co-transcriptionally through the process of splicing. This process is carried out by the spliceosome, a multi-subunit complex consisting of 5 core small nuclear RNAs (snRNAs) and hundreds of associated proteins. The majority of introns are defined by GT and AG dinucleotides at their 5' and 3' termini, respectively. These introns are spliced by the major spliceosome, consisting: U1, U2, U4, U5, U6 snRNA, and associate proteins (Akinyi and Frilander, 2021). However, approximately 0.5% of introns utilise AT-AC terminal dinucleotides and are spliced by the minor spliceosome (Turunen et al., 2013). For the purpose of introduction, here we focus on the major spliceosome.

The spliceosome assembles on the intron and its assembly is dependent on several intron-defining features outlined in Figure 1.7A. Most notable are the 5' splice site (5'SS; GT dinucleotide) at the 5' terminus, which is recognised by U1 snRNP, and the branch point (BP) adenosine, which is recognised by U2 snRNP. Introns are also defined by the 3' splice site (3'SS; AG dinucleotide) at the 3' terminus, and the polypyrimidine tract

(poly(Y)), which sits between the BP and 3'SS. Assembly of U1 and U2 on the intron forms the pre-spliceosome. U4/U6.U5 tri-snRNP is subsequently recruited to the pre-spliceosome, leading to conformational changes producing a catalytically active spliceosome (Nguyen et al., 2015; Wilkinson et al., 2020). The spliceosome subsequently catalyses two reactions: branching and exon ligation (Wilkinson et al., 2020). During branching, the BP adenosine is brought together with the 5'SS, cleaving the exon-intron boundary, and producing a 3' lariat intron (Figure 1.7B). Next, during exon ligation, the 5' exon is brought together with the 3'SS, leading to excision of the lariat intron and ligation of the 5' and 3' exons (Figure 1.7C). Following splicing, the exon junction complex (EJC) is deposited 20-24 nucleotides upstream of the splice junction by the spliceosome. Such precise deposition is mediated by Cwc22, which acts as a "molecular ruler" (Wilkinson et al., 2020). The presence of the EJC effectively acts as a marker to indicate that splicing has been conducted, and its presence is associated with RNA export (as discussed in Section 1.1.3).

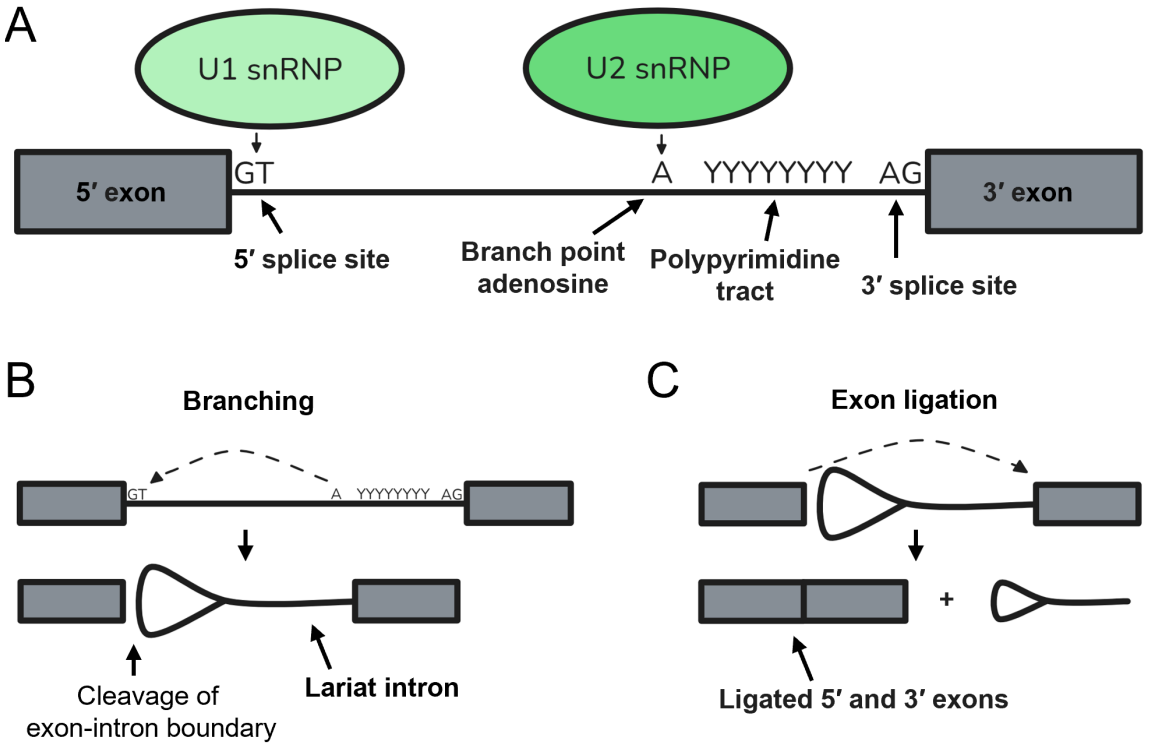

Figure 1.7: Intron-defining features and splicing. A) Features that define the intron: GT dinucleotide at the 5′ terminus (5′SS; which is bound by U1 snRNP); branch point adenosine (BP; which is bound by U2 snRNP); polypyrimidine tract; AG dinucleotide at the 3′ terminus. B) During branching the BP and 5′SS are brought together, cleaving the 5′ exon from the intron, and leading to formation of a lariat intron. C) During exon ligation the 5′ and 3′ exons are brought together and ligated, leading to excision of the lariat intron.

In addition to intron defining features (intron definition; as described above), exon definition can also drive splicing, whereby the binding of splicing factors to the exon define its 5′ and 3′ splice sites. This process involves interactions between exon-bound factors and both the 3′ and 5′ introns, highlighting a level of cross-talk between multiple splicing events (Rogalska et al., 2023).

1.4.1 Alternative splicing

In addition to removing introns, splicing enables the production of mRNA built from various combinations of introns and exons based on the splice sites selected and their relative strengths. For example, intron retention may occur where the intron is poorly

defined, or where the strength of its definition is modified in a context-dependent manner through binding of additional trans factors. By a similar mechanism, alternative 3′ splice site (A3SS) or alternative 5′ splice site (A5SS) selection occurs where multiple 5′ and 3′ splice sites exist, and they compete for occupancy of U1 and U2 snRNP respectively. Again, the selection of either site may be regulated in context-specific manner due to the binding of trans factors locally within the mRNP. These processes are collectively termed alternative splicing (AS) and occurs in roughly 95% of human multi-exon genes (Pan et al., 2008). Such widespread AS diversifies the transcriptome, enabling the production of distinct RNA transcripts from the same DNA sequence, potentially contributing to organismal complexity and diversity between species (Wright et al., 2022). This can change the proteins produced, through changing the CDS exon composition. AS can also occur in 5′UTRs and 3′UTRs, where it will not impact the amino acid sequence of the proteins produced, but instead modify the regulatory roles of such regions. AS is observed during development and in disease, as will be discussed herein.

1.4.2 Alternative splicing in development and stem cell differentiation

An interesting example of differentiation stage-specific exon inclusion is within the FOXP1 gene, which contains mutually exclusive exons 18 and 18b (Figure 1.8). In hESCs, exon 18b is selected; subsequently, FOXP1 drives expression of pluripotency factors OCT4 and NANOG in pluripotent cells. Whilst in differentiated cells, exon 18 is selected over exon 18b, changing the forkhead domain of FOXP1, meaning it no longer regulates OCT4 and NANOG (Gabut et al., 2011). FOXP1 AS is driven by MBLN1 and MBLN2, which are upregulated upon exit from pluripotency, leading to increased exon 18 inclusion (Han et al., 2013). MBLN1/2 regulate a network of ESC-specific AS events, and their knockdown leads to increased expression of pluripotency markers and cellular reprogramming. It was later shown that TRIM71 (LIN-41) negatively regulates MBLN1 expression in ESCs by binding its 3′UTR (Welte et al., 2019).

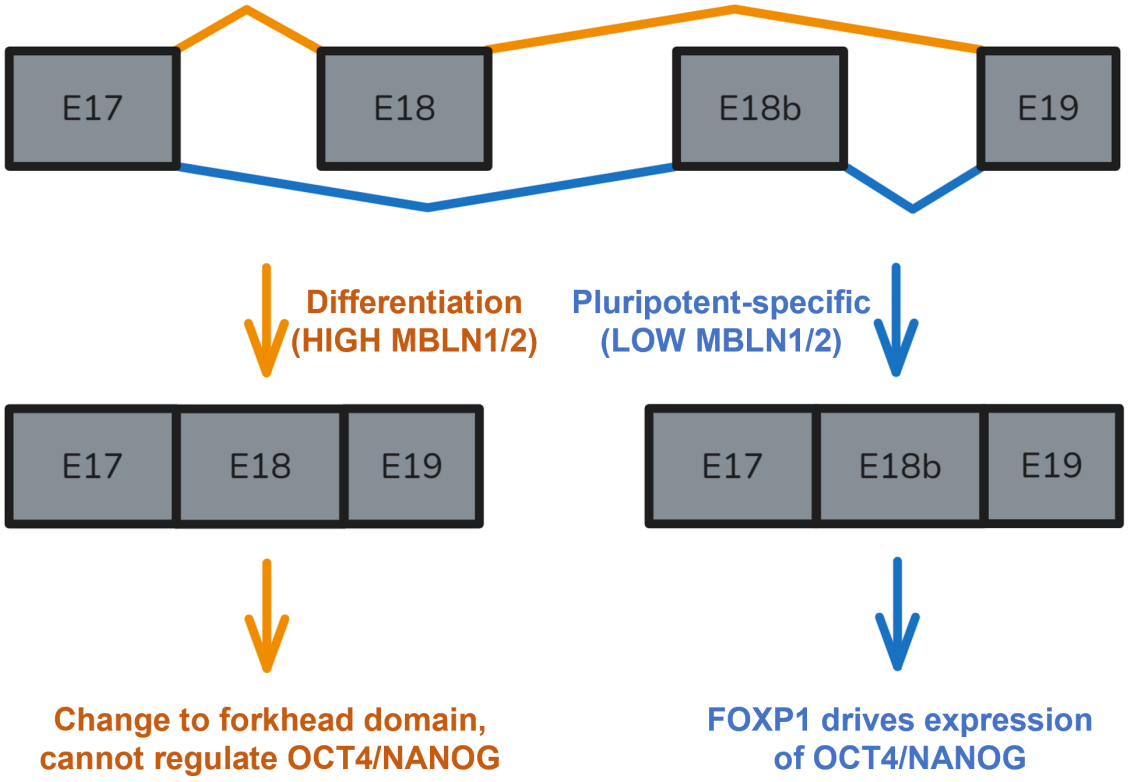

Figure 1.8: ESC-specific AS of FOXP1. FOXP1 contains mutually exclusive exons 18 and 18b. In ESCs exon 18b is selected (blue splicing pattern) meaning FOXP1 drives expression of pluripotency factors. As differentiation proceeds MBLN1/2 is upregulated, leading to selection of exon 18 (orange splicing pattern) over exon 18b. Subsequently the forkhead domain of FOXP1 is altered, and FOXP1 no longer drives expression of pluripotency factors.

AS can also lead to frameshifts and/or the introduction of premature termination codons, leading to transcript degradation via nonsense-mediated decay (NMD; Section 1.5). However, AS-linked to NMD (AS-NMD) can be actively utilised to regulate gene expression, including the autoregulation of various splicing factors during development. Large-scale transcriptomic analyses of RNAseq data from the Genotype-Tissue Expression (GTEx) project has revealed that the majority of variance between tissue types is explained by changes in gene expression level, where AS acts to complement this, and explain some of the variance observed between individuals (Melé et al., 2015). AS-NMD directly links AS to the regulation of gene expression, and several examples of regulation

in this manner are explored in Section 1.5.3.

1.4.3 Aberrant splicing in disease

Aberrant splicing is common in cancers, which are often characterized by widespread intron retention (Dvinge and Bradley, 2015). AS can be utilized by cancer cell populations to confer their own selective advantage, which is typified by an increased proliferative capacity and reduced levels of cell death, alongside other cancer hallmarks (Hanahan and Weinberg, 2011; Oltean and Bates, 2014). An example of this is exon 4 skipping within the EGFR gene, which leads to constitutive activity of the receptor independently of EGF binding, enhancing cell proliferation, colony formation, and invasion (Wang et al., 2011). Aberrant splicing is also common amongst neurodegenerative diseases, including Alzheimer's disease, Parkinson's disease, and Amyotrophic Lateral Sclerosis (Nikom and Zheng, 2023). A specific example of this is highlighted in Parkinson's disease, which is characterized by aggregation of α-synuclein. Wild type α-synuclein is made up of 140 amino acids (α-synuclein-140), however multiple splice variants exist where exons are skipped (Beyer et al., 2008). Where exon 5 is skipped, α-112 is produced with C-terminal truncation, leading to enhanced oligomerization compared to wild type α-synuclein, and subsequent defects in synaptic vesicle recycling (Soll et al., 2020).

Targeting AS also represents a potential therapeutic avenue, whereby antisense oligonucleotides (ASOs) can be used to target cis elements, thus preventing the binding of trans factors that regulate splicing. The ASO Nusinersen is used to treat patients with Spinal Muscular Atrophy (SMA) (Collotta et al., 2023). SMA is not caused by aberrant splicing, but instead by loss of SMN1 expression, either through deletion or mutation. Humans also have a paralogous gene SMN2, however over 80% of SMN2 transcripts skip exon 7, producing a truncated and unstable protein, and subsequently SMN2 cannot compensate for the loss of SMN1 expression (Cho and Dreyfuss, 2010; Singh and Singh, 2018). Nusinersen prevents SMN2 exon 7 skipping by binding to cis elements downstream of exon 7 that would otherwise bind trans factors that promote exon

skipping. Such treatment rescues expression of full length SMN2 and compensates for loss of SMN1, improving motor function and patient survival (Collotta et al., 2023).

1.5 Nonsense-mediated RNA decay (NMD)

Nonsense-mediated RNA decay (NMD) is an evolutionarily conserved RNA surveillance pathway common to all eukaryotes. NMD functions to detect and degrade transcripts containing premature termination codons (PTCs), thereby preventing expression of truncated proteins which may otherwise have deleterious effects on normal cellular function. NMD can be broadly categorized as having two arms: 1) the "EJC-dependent" arm responsible for detecting PTCs (Figure 1.9A); 2) the "EJC-independent" arm responsible for detecting transcripts with long 3'UTRs (Figure 1.9B). Both pathways will be discussed herein. As well as clearing PTC-containing transcripts, splicing introns in the 3'UTR (3UIs) is generally considered a signal to elicit NMD via the EJC-dependent arm (Bicknell et al., 2012). Therefore, transcripts which contain splice junctions within their 3'UTR may be NMD-sensitive despite encoding a full-length amino acid chain.

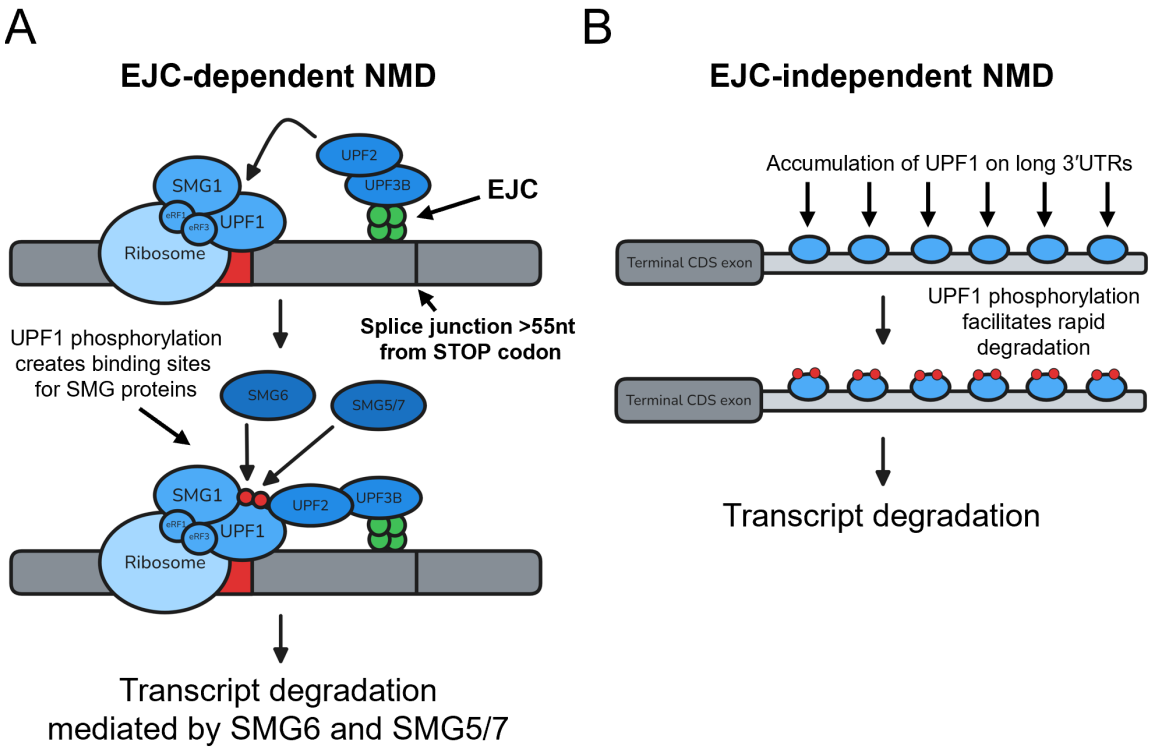

Figure 1.9: The two arms of nonsense-mediated decay. A) EJC-dependent arm. UPF2 and UPF3B are associated with EJCs (which persist if they are >55nt from the stop codon). SMG1-UPF1-eRF1-eRF3 (SURF complex) is associated with the ribosome. Interactions between UPF2-UPF3B-EJC and the SURF complex leads to UPF1 phosphorylation (red circles), enabling recruitment of SMG6 and SMG5/7, which subsequently degrade the transcript. B) EJC-independent NMD. UPF1 accumulates on long 3'UTRs, facilitating rapid degradation of transcripts with long 3'UTRs upon UPF1 phosphorylation.

1.5.1 EJC-dependent NMD

EJC-dependent NMD (Figure 1.9A) is translation-dependent. During the pioneer round of translation the ribosome strips EJCs that it encounters from the mRNA (Gehring et al., 2009). Therefore, EJCs persist if they are deposited downstream of the translation stop codon, such as in the case of CDS splice junctions downstream of a PTC, or splice junctions within the 3'UTR, presuming there is no translational readthrough. Importantly, the splice junction must be located more than 55nt from the stop codon to be considered NMD-sensitising (Nagy and Maquat, 1998). These persisting EJCs act as a platform that facilitates the interactions triggering NMD. Central to these interactions are those

between up-frameshift (UPF) proteins UPF2 and UPF3B, which bind the EJC (Le Hir et al., 2001), and the SURF complex, comprising SMG1, UPF1, eRF1 and eRF3 (Kashima et al., 2006). This interaction forms the decay-inducing (DECID) complex, wherein UPF1 is phosphorylated, producing binding sites for SMG6 and the SMG5-SMG7 complex (Ohnishi et al., 2003; Okada-Katsuhata et al., 2012). SMG6 functions directly as an endonuclease to cleave RNA bound by the DECID complex (Eberle et al., 2009). Meanwhile, the SMG5-SMG7 complex recruits CCR4-NOT, resulting in deadenylation and decapping, followed by mRNA decay (Loh et al., 2013). Additionally, the SMG5-SMG7 complex permits SMG6-mediated endonucleolytic activity, which was abolished in SMG5-SMG7 depleted cells (Boehm et al., 2021).

1.5.2 EJC-independent NMD

NMD can also occur in the absence of EJCs to degrade transcripts with long 3'UTRs (Figure 1.9B)(Muñoz et al., 2023). UPF1 accumulates on long 3'UTRs (Hogg and Goff, 2010). The exact reason for this enrichment remains unclear. UPF1 is an RNA helicase, therefore its enrichment may simply be due to non-specific RNA binding (Hogg and Goff, 2010), which would be enhanced where the UTR is longer. 3'UTR cis elements also contribute to UPF1 accumulation, where it was shown that GC-rich 3'UTR sequence elements were necessary to recruit UPF1 (Imamachi et al., 2017). 3'UTR cis elements can also recruit RBPs which lead to evasion of EJC-independent NMD (Section 1.5.5). UPF1 enrichment is proposed to prepare transcripts for degradation via NMD, where each bound UPF1 molecule acts as a platform to recruit SMG proteins following phosphorylation, facilitating decay (Muñoz et al., 2023). Therefore, shortening the 3'UTR via alternative polyadenylation (by selecting proximal PAS sites) or splicing out introns within the 3'UTR may represent a method to reduce UPF1 occupancy on the 3'UTR and alleviate EJC-independent NMD.

1.5.3 Regulation of gene expression by AS-NMD

NMD plays an important role in the regulation of gene expression besides its role in cleaning up processing errors, whereby AS coupled to NMD (AS-NMD) regulates the expression levels of various splicing factors. AS-NMD occurs when a transcript is alternatively spliced to introduce an NMD-sensitising element, such as a premature termination codon. The majority of serine-arginine-rich splicing factor (SRSF) proteins contain cassette exons which trigger NMD upon inclusion, these highly conserved elements are termed "poison exons" (Lareau et al., 2007). SRSF proteins regulate poison exon inclusion in their own transcripts, and also the transcripts of other SRSF proteins, creating a complex cross-regulatory network (Leclair et al., 2020). The inclusion of poison exons in these transcripts changes during differentiation (Leclair et al., 2020), thus tuning the SRSF network, which has widespread influence over alternative splicing across the transcriptome (Bradley et al., 2015). Poison exon inclusion is also utilized by HNRNPL to autoregulate its own expression (Rossbach et al., 2009). Exon skipping can also trigger NMD where it causes a frame shift. Polypyrimidine tract binding protein 1 (PTBP1) and PTBP2 promote PSD-95 exon 18 skipping in neuronal progenitor cells, sensitising these transcripts to NMD and reducing PSD-95 expression (Zheng et al., 2012). Downregulation of PTBP1 and PTBP2 in differentiated neurons subsequently releases PSD-95 from such regulation, and is essential for synaptic maturation. Additionally, NMD has been shown to regulate the expression of its own components, many of which have long 3'UTRs (Yepiskoposyan et al., 2011).

1.5.4 Variable NMD efficiency

Lou et al. (2016) observed differential expression of core NMD factors during the differentiation of hESCs into ectoderm, mesoderm and endoderm. Additionally, upon transfection of an NMD reporter construct into each cell type, they observed differential NMD efficiencies. They found that NMD was significantly more efficient in mesoderm

and ectoderm, and significantly less efficient in endoderm, compared to undifferentiated hESCs. Knocking down either UPF1 or UPF3B promoted endodermal fates, as determined by an increased expression of endodermal markers SOX17 and CXCR4. Together these results suggest that global NMD efficiency may play an important role in cell fate and lineage specification. In this regard, differing NMD efficiencies have been observed between tissues in both mouse models (Zetoune et al., 2008) and in human cohort studies (Rivas et al., 2015). Additionally, in a more recent study using a bi-directional NMD reporter in combination with flow cytometry, Sato and Singer (2021) were able to compare open reading frames side by side under the same promoter, including the comparison of wild-type and PTC-containing open reading frames. Whilst the PTC-containing open reading frame was generally expressed at a lower level compared to wild-type across the population as a whole, 27.9% of cells did not show this trend, despite being transfected with the same construct. This indicates that NMD efficiency can vary between cells from the same population, where some cells display reduced NMD efficiency, and in others, NMD is evaded entirely.

1.5.5 NMD evasion

Transcriptome-wide studies have shown that a substantial proportion of transcripts that are predicted to elicit NMD are able to evade it to some degree (Lappalainen et al., 2013; Lindeboom et al., 2019). However, the mechanisms behind NMD evasion are not completely understood. Early studies highlighted that the presence of downstream open reading frames enables NMD evasion due to translation reinitiation (Zhang and Maquat, 1997), where the intercistronic distance and size of the initial open reading frame were presented as potential determinant factors (Kozak, 1987; Neu-Yilik et al., 2011). NMD evasion is exploited by the Rous sarcoma virus, containing the *gag* gene, which has a 7kb 3'UTR which should be NMD-sensitising. A deletion mutagenesis screen of 3'UTR elements identified a 151nt RNA stability element (RSE) immediately downstream of the stop codon which imparts this NMD evading function, stabilizing *gag* mRNA (Weil and

Beemon, 2006; Withers and Beemon, 2010). A similar element was more recently described in Turnip crinkle virus (May et al., 2018). Such elements are proposed to act as binding sites for trans factors that facilitate NMD evasion. Ge et al. (2016) showed that the insertion of the RSE into the SMG5 gene, which contains a long NMD-sensitising 3'UTR, resulted in NMD evasion in human cells. Analysis of proteins enriched in the mRNP of constructs containing or lacking the RSE identified enrichment of PTBP1 and HNRNPL, as well as a reduction in UPF1 in RSE-containing 3'UTRs (Ge et al., 2016). Removal of PTBP1 binding sites within the RSE restored NMD in the SMG5 constructs, significantly reducing the half-life (Ge et al., 2016). Likewise, mutation or deletion of HNRNPL binding sites from the BCL2-IGH fusion mRNA 3'UTR reduced RNA stability in a UPF1-dependent manner (Kishor et al., 2019).

NMD evasion is also clinically relevant, which is highlighted by an abnormal mechanism of inheritance of β-thalassemia, impacting whether heterozygous carriers are symptomatic or asymptomatic. Nonsense mutations occurring in the β-globin gene lead to the production of truncated proteins which impart a dominant negative effect if not cleared. In the majority of cases, nonsense mutations in β-globin are NMD-sensitising, and as such, mRNA containing such mutations is degraded by NMD (Romão et al., 2000; Neu-Yilik et al., 2011). Whilst this prevents expression from the mutant allele, the presence of the wild-type allele is sufficient for normal function, meaning carriers of NMD-sensitising nonsense mutations in β-globin are asymptomatic, and their condition is inherited in a recessive manner. Where mutations lead to NMD escape, β-globin mRNA levels are not significantly reduced (Romão et al., 2000), the truncated protein imparts a dominant negative effect leading to the production of variant β-globin chains (Thein, 2013), meaning heterozygous carriers are symptomatic, and their condition is inherited in a dominant manner.

1.6 Examples of regulation by 3'UTR splicing

Whilst several examples of differential regulation due to APA were discussed in Section 1.3, the number of examples relating to 3'UTR splicing is substantially smaller. This is because splicing within the 3'UTR is predominantly viewed as NMD-sensitising, and may be considered "noise". However, this will not be the case where the splice junction is less than 55nt from the stop codon. Additionally, in some instances, such as autoregulation via AS-NMD, and in transcripts that need to be expressed transiently, triggering NMD via 3'UTR splicing may be functionally useful. Aside from NMD, 3'UTR splicing may also function to manipulate the mRNP composition, impacting the regulatory avenues described in Section 1.2. Here we present several examples of regulation at the mRNA level by 3'UTR splicing.

1.6.1 Autoregulation of AUF1/HNRNPD expression

The AUF1/HNRNPD 3'UTR usually contains a single 3UI, situated 31 nucleotides from the stop codon, which is therefore not predicted to sensitise the transcript to NMD (Figure 1.10). However, a cassette exon (exon 9) can also be included, which increases the distance between the terminal 3UI and the stop codon to >55 nucleotides, which is predicted to sensitise the transcript to NMD (Figure 1.10)(Wilson et al., 1999). Indeed, knockdown of UPF1 leads to an increased expression of these isoforms where intron 9 is also spliced (Banihashemi et al., 2006). Interestingly, AUF1 has been shown to bind its own 3'UTR at exon 9 (Wilson et al., 1999), and the overexpression of AUF1 leads to increased exon 9 inclusion in a Luciferase reporter system in HeLa cells (Kemmerer et al., 2018).

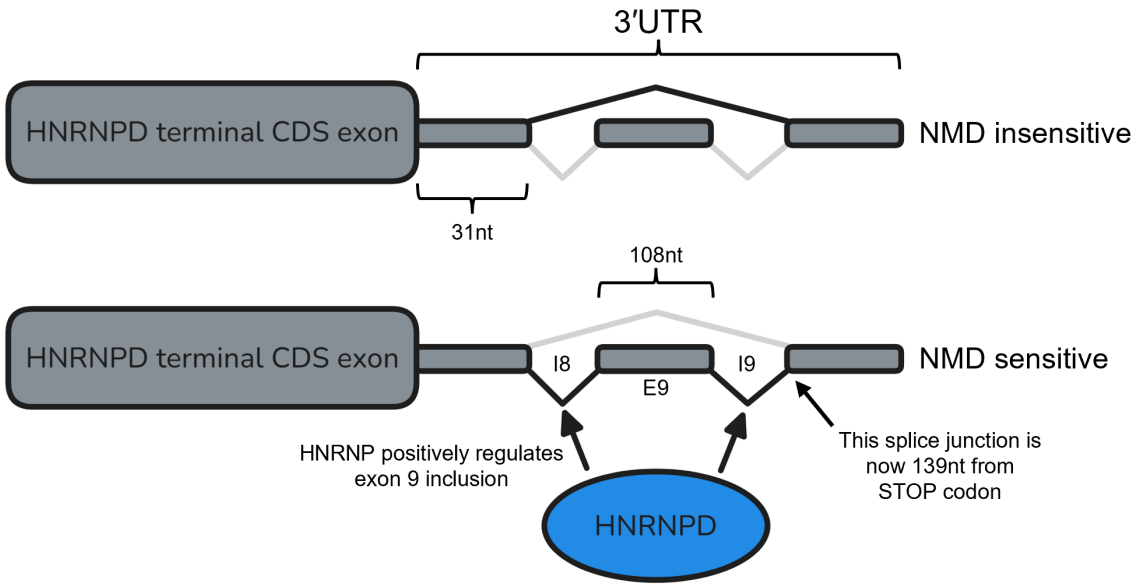

Figure 1.10: AUF1/HNRNPD autoregulation via AS-NMD. AUF1/HNRNPD transcripts with only a single 3UI (top) are NMD insensitive as the splice junction to stop codon distance is <55nt. Where AUF1/HNRNPD binds exon 9 (E9) it positively regulates its inclusion, producing two 3UIs (bottom). The splice junction to stop codon distance of the terminal 3UI in this context is >55nt, therefore these transcripts are NMD sensitive.

1.6.2 Dendritic localization of *Calm3* mRNA

Utilizing individual-nucleotide resolution crosslinking and immunoprecipitation (iCLIP) in the developing mouse brain, Sharangdhar et al. (2017) identified 356 neuronally expressed mRNAs where Staufen2 binds within the 3′UTR. 7.9% of these transcripts displayed Staufen2 binding within a retained intron in their 3′UTR, with the most significant enrichment being observed in *Calm3* mRNA. To understand more about the interaction between Staufen2 and the *Calm3* 3UI, Sharangdhar et al. (2017) created GFP reporter constructs upstream of either the full-length or spliced *Calm3* 3′UTR, or the intron sequence alone to act as a 3′UTR, and overexpressed these in rat hippocampal neurons. They observed a significant increase in dendritic localization of reporter mRNA containing the 3UI sequence (full-length and intron-only) compared with the 3′UTR spliced construct. It is important to note that the full-length construct would still be

subject to some level of 3'UTR splicing, which was presumably low given this result. However, this could have been accounted for by producing an extra construct where the 3'UTR splice site was mutated, thus forcing intron retention. Following transfection of neurons with shRNAs against Staufen2 they observed a significant reduction in dendritic *Calm3* mRNA localization, which was subsequently rescued by overexpression of RNAi-resistant Staufen2. These results indicate that Staufen2 mediates dendritic localization of *Calm3* mRNA when the 3UI is retained. Therefore 3'UTR splicing represents a mechanism to negatively regulate dendritic localization, and AS may regulate *Calm3* dendritic localization in a context-specific manner. In this regard, Sharangdhar et al. (2017) also showed that the intron-retaining isoform is progressively used more (as a % of total Calm3 expression) as neuronal development proceeds.

1.6.3 Fine tuning *Arc* mRNA expression

The Arc 3'UTR contains cis elements known as "dendritic targeting elements" that mediate its localization to neuronal dendrites (Kobayashi et al., 2005). Such localization is further enhanced where the 3'UTR is spliced endogenously, and is impaired where cells are transfected with "pre-spliced" cDNA-derived plasmids (Steward et al., 2018). To investigate how splicing the Arc 3'UTR regulates its expression, Paolantoni et al. (2018) created a series of luciferase plasmids containing the Arc 3'UTR derived from either the cDNA (i.e. already spliced; "pre-spliced plasmid") or gDNA (i.e. unspliced, but shown to be constitutively spliced upon transfection; "spliceable plasmid"). Upon transfection into rat cortical neurons, the pre-spliced plasmid was expressed significantly better than the spliceable plasmid at both the protein and mRNA levels, consistent with previous findings that splicing the Arc 3'UTR triggers NMD (Giorgi et al., 2007; Steward et al., 2018). Upon stimulation with brain-derived neurotrophic factor (BDNF) Paolantoni et al. (2018) observed no significant change in expression of either plasmids at the mRNA level compared to the untreated condition. However, a significant increase in expression at the protein level was observed only in the neurons transfected with the spliceable plasmid.

Additionally, polysome profiling revealed that BDNF stimulation increased translation efficiency in a splicing-dependent manner. These findings are in line with the role of Arc as an immediate early gene, and suggest that splicing its 3'UTR contributes to the rapid bursts of expression observed upon neuronal stimulation. For such rapid bursts to occur, Arc mRNA being very unstable and efficiently translated would be advantageous.

1.7 Human pluripotent stem cells as a model of development

Human pluripotent stem cells (hPSCs) are cells capable of indefinite self-renewal, and giving rise to all cell types within the human body. hPSCs encompass both human embryonic stem cells (hESCs) and human induced pluripotent stem cells (hiPSCs). hESCs are derived from the inner cell mass of the blastocyst-stage embryo (Thomson et al., 1998). hiPSCs are generated through the reprogramming of somatic cells, involving the overexpression of Oct4, Sox2, Klf4, c-Myc (OSKM) transcription factors (Takahashi et al., 2007). The use of hPSCs as a model system allows us to investigate biological questions in karyotypically normal human cells (unlike many cancer cell lines), which have not been subjected to immortalization for use in cell culture. Additionally, by applying our understanding of the signalling pathways that define distinct populations during development, we can differentiate hPSCs towards desired cells types through manipulating the media in which cells are grown (Murry and Keller, 2008; Tabar and Studer, 2014). This allows us to model stages of development which we currently cannot observe in human embryos in vivo due to strict regulation (see Hurlbut et al. (2017) for discussion on this "black box" period). Meanwhile, the use of hiPSCs allow us to investigate how development is impacted in disease-states by generating hiPSCs from patient groups of interest.

Unlike studying development using embryos, the use of cell lines allows us to conduct manipulations more easily. For example, we can assess the impact of individual genes on biological processes by knocking them out or over expressing them transiently. Such

knockout screens were used to identify genes responsible for hESC self-renewal and pluripotency (Chia et al., 2010). Additionally, a more recent screen in hESCs identified the impact of over-expressing hundreds of different transcription factors on the transcriptome (Nakatake et al., 2020). Additionally, using genome-editing technologies such as CRISPR/Cas9, mutations can be introduced into hPSC lines to study their effects (Bassett, 2017). Likewise, where mutations are known to be disease-linked, genome-editing of patient-derived hiPSCs can be conducted to correct these and study the impact on molecular processes. Of most relevance to this thesis, various RNA sequencing methods have been conducted on hPSCs at various stages of differentiation. Much of this data is publicly available from resources such as the Sequence Read Archive (SRA). Additionally, a large collection of RNAseq data from hiPSCs was produced and made available by the Human Induced Pluripotent Stem Cell Initiative (HipSci).

1.8 Aims and Objectives

3'UTRs have important functions in post-transcriptional regulation, including the regulation of transcript stability, localization, and translation efficiency. It has previously been shown that regulating the content of 3'UTRs can be achieved through APA, resulting in differences in functionality. Therefore, it is also feasible that splicing the 3'UTR can impact these regulatory modalities. In this regard, regulation of sequence content would not be limited to the 3' termini, instead any internal sequence contained within introns could be modulated. However, splicing 3'UTRs is generally considered to be a signal to elicit transcript degradation via NMD, hence 3'UTR-spliced transcripts are often considered to be "transcriptional noise". Despite this, many annotated transcripts contain introns in their 3'UTRs, including those with the highest transcript support level, and those annotated as the "canonical isoform". This suggests that many of these transcripts may be functional, and not noise. In many of these cases the 3UIs sit less than 55nt from the stop codon, and therefore are not predicted to be NMD-sensitive.

Additionally, NMD-sensitive transcripts may still be functional where having a short half-life is useful for function.

This study aims to investigate the extent of 3'UTR splicing across the transcriptome by using large existing RNAseq datasets such as from the HipSci consortium, The Cancer Genome Atlas (TCGA), and previously published papers. We will investigate the regulatory capacity of 3UIs by incorporating existing CLIP-seq and AGO-CLIP data to identify the trans factors that are regulated. Additionally, we will investigate how 3'UTR splicing changes in dynamic systems, such as across differentiation time courses, and in cancer transformation. As well as studying differences on steady state expression, through the establishment of a hESC differentiation model, and the use of metabolic labelling, we will investigate how expression dynamics change during differentiation. Additionally, through the development of novel analyses and pipelines, we will investigate the impact of 3'UTR splicing on RNA stability, including between 3UI-spliced/-retained pairs, and how this relationship changes during differentiation.

2. Materials and Methods

2.1 RNAseq analysis

High throughput RNA sequencing analysis was conducted through the creation and use of Ruffus pipelines (github.com/jjriley1 and github.com/sudlab) and the use of pre-existing Ruffus pipelines from the CGAT Flow suite (github.com/cgat-developers/cgat-flow). Scripts used in various analyses, and in the generation of figures for this thesis can be found in the thesis GitHub repository (github.com/jjriley1/Thesis).

A combination of new and pre-existing RNA sequencing data was used. Where new RNA sequencing data was produced, RNA was extracted as per Section 2.3.1 and sent to Novogene (Cambridge, UK) for library preparation (Poly(A) enrichment and cDNA synthesis) and sequencing (Illumina PE150). Where existing RNA sequencing data was used, this will be cited in the relevant results sections.

2.1.1 Read quality control

Raw RNA sequencing reads were quality checked using FastQC (Andrews, 2010) via pipeline_readqc from the CGAT Flow suite. FastQC results were compiled into final MultiQC reports to allow assessment and filtering of data where applicable. Samples were assessed on several FastQC metrics: read duplication rate < 70%, a consistent mean quality score > 20 across all bases of a read, the average GC content of reads following a normal distribution centered at approximately 50%. Samples exhibiting one or more poor FastQC metrics were filtered out. Poor FastQC metrics were observed in only 1 sample from the Human Induced Pluripotent Stem Cell Initiative RNAseq cohort.

2.1.2 Read mapping

RNA sequencing reads were mapped to the NCBI hg38 "analysis set" genome using STAR (Dobin et al., 2013). The following mapping options were supplied to STAR:

```
--outFilterMultimapnMax 20 --outFilterMatchNminOverLread 0.33
--alignIntronMax 500000 --outFilterScoreMinOverLread 0.33
--alignMatesGapMax 1000000 --alignSJDBoverhangMin 1
--twopassMode Basic --outSAMstrandField intronMotif
--outSAMattributes NH HI NM MD AS XS
```

2.1.3 3'UTR intron detection pipeline (pipeline_utrons)

In order to detect pre-existing and novel 3'UTR splicing events, and build novel transcript assemblies, pipeline_utrons was created by Ian Sudbery and Cristina Alexandru-Crivac (github.com/sudlab/pipeline_utrons). Pipeline_utrons.py is a bioinformatic pipeline written in Python using the CGAT Flow infrastructure.

2.1.3.1 Transcript assembly

Following read mapping with STAR (Section 2.1.2) individual .bam files were passed to portcullis alongside the `-b` (`--bam_filter`) argument to filter out alignments associated with splice junctions that are unlikely to be genuine. Novel transcripts where then assembled by passing each filtered .bam file to StringTie (Pertea et al., 2015) alongside the `-f 0.05` option to produce a transcript assembly (.gtf file) per sample. All samples for each dataset were then merged into a final transcriptome assembly (.gtf file) for that dataset by using the StringTie `--merge` argument alongside the `-F 0 -T 1` options. Therefore two final assemblies were produced, one for all samples from the TCGA cohort, and one for all samples from the HipSci cohort.

2.1.3.2 Detection and classification of 3'UTR introns

3'UTR intron detection was facilitated by the find_utrons.py python script within pipeline_utrons. Each transcript was matched to a reference transcript which shared the exact coding intron chain. The transcript also had to start before the reference transcript's start codon, and end after its stop codon. Introns found downstream of the stop codon

(3'UTR introns) were subsequently classified.

3'UTR introns were classified as "novel" if they did not share 5' exon-intron or 3' intron-exon boundaries with introns in the 3'UTRs of reference transcripts. Classification of 3'UTR introns as "partnered" occurred where they are found in a transcript that shares the exact CDS intron chain as a reference transcript, but differ only due the splicing of the 3'UTR intron. Finally, 3'UTR introns were classed as "nonPTC" if neither the 5' exon-intron boundary nor 3' intron-exon boundary overlapped with those from introns in the CDS of other reference transcripts. Coordinates of each event, for each classification, were outputted in .bed format.

2.1.3.3 Transcript quantification

Samples were quantified against their corresponding transcriptome using the Salmon package (Patro et al., 2017). Cancer samples were quantified against the TCGA assembly, whilst stem cell samples were quantified against the HipSci assembly. Firstly, a decoy-aware transcriptome index was generated using the `salmon index` command with the `-k 31` option. The hg38 genome was used as a decoy in order to reduce levels of invalid mapping (for example where a read that appears to be spliced in the transcriptome also maps entirely to a single unannotated genomic region). Samples were then quantified using the `salmon quant` command with the `--gcBias` option. This produced .sf files which were imported into R using tximport (Soneson et al., 2016). Subsequently, TPM values were extracted at both the transcript and gene levels, and used to calculate a "fraction expression" value representing the proportion each transcript contributes towards the expression of its gene in each sample.

2.1.4 Differential transcript/gene expression analysis

Differential transcript expression (DTE) and differential gene expression (DGE) analysis were conducted using DESeq2 (Love et al., 2014). Quality control was conducted in the

form of principle component analysis to check for clustering between replicates of the same condition. DTE and DGE were conducted for multiple comparisons, across different experiments. The corresponding scripts can be found in the thesis GitHub repository (github.com/jjriley1/Thesis).

2.1.5 Differential exon usage analysis

Differential exon usage analysis was conducted using rMATS turbo v4.1.2 (Shen et al., 2014). One benefit of the use of rMATS is the ability to test only a specific subset of splicing events through the use of the `--fixed-event-set` parameter. A fixed event set for intron retention was created through the use of the .gtf and .bed files generated in Sections 2.1.3.1 and 2.1.3.2. A custom script (github.com/jjriley1/Thesis/results _2_differential_analysis/fixed_event_sets) was used to accomplish this. Briefly, each intron retention event within the .bed file was compared against the .gtf file in order to obtain the start and stop positions of the upstream and downstream exons, which are required for the RMATS fixed event set file. Blank fixed event sets were created for skipped exon, alternative 5' splice site, alternative 3' splice site, and mutually exclusive exon events, in order to increase the speed of the pipeline.

Reads were mapped with STAR (Section 2.1.2) and passed to rMATS in .bam file format. rMATS was run in a two-step fashion, where a `--task prep` was run on every individual input file in parallel, followed by a single `--task post` at the end to collate results. For most use cases, the paired stats model provided by rMATS was not used as either a single cell line was used, or large cohorts of patients was used. However, in instances where this model was suitable the `--b1` and `--b2` inputs were ordered such that members of each pair were found at the same index in each input file, additionally the `--paired-stats` argument was provided at the `--task post` stage. Following the completion of the rMATS post task the "RI.MATS.JC.txt" file was filtered based on statistical significance, with events with $FDR \leq 5\%$ or $FDR \leq 10\%$ being considered statistically significant (dependent on use case; this will be specified in the results).

Where the results of rMATS were visualized as sashimi plots, the rmats2sashimiplot package (github.com/Xinglab/rmats2sashimiplot) was used. A grouping file (.gf) was generated to specify the comparison to be made based on the `--b1` and `--b2` inputs. The path of the grouping file was subsequently passed to rmats2sashimiplot by the `--grouping-info` parameter.

2.1.6 Differential transcript usage analysis

Differential transcript usage (DTU) was conducted using DEXseq (Anders et al., 2012) and DRIMseq (Nowicka and Robinson, 2016) using the workflow described in Love et al. (2018). FDR was controlled using stageR (Van den Berge et al., 2017). Where differentially used transcripts were clustered across differentiation, fraction expression values (from Section 2.1.3.3) were used for transcripts that had FDR<0.05. K-means clustering was conducted and heatmaps were plotted using ComplexHeatmap (Gu et al., 2016; Gu, 2022).

2.1.7 Differential alternative polyadenylation analysis

Alternative polyadenylation analysis was conducted with DaPars (Xia et al., 2014). DaPars outputs information on the difference in polyA site usage (PDUI) between groups, and the significance of this difference. Differences were considered statistically significant where adjusted P-value < 0.05. Where multiple PDUI values were generated per gene, due to multiple transcripts with different 3' ends, the weighted mean was taken, whereby each transcript's PDUI value was weighted by the TPM of that transcript. Distributions of PDUI values were subsequently plotted via density plot, and compared across differentiation via heatmap.

2.1.8 RBP and MRE enrichment analysis

RBP and MRE enrichment was carried out using CLIP-seq and AGO-CLIP data from starBase (Li et al., 2014). Enrichment analysis was conducted with the Genomic

Association Tester (GAT; github.com/AndreasHeger/gat). GAT allows enrichment of a subset of events to be tested against a background set. Where these comparisons are made both the test and background sets will be stated in the figure or results text. Adjusted P-values < 0.05 are considered statistically significant. However in some instances more stringent (Adjusted P<0.01) thresholds are used, this will be stated in the accompanying text where applicable.

2.2 Cell culture

2.2.1 Cell maintenance

2.2.1.1 Human colorectal carcinoma cells

Human colorectal carcinoma cell line HCT116 was maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% Fetal Bovine Serum (FBS) and Penicillin/Streptomycin. HCT116 cells were routinely maintained in T75 flasks at 37°C and 5% CO_2 and were passaged upon reaching >80% confluency. For cell passaging, media was aspirated and cells were washed with Phosphate Buffered Saline (PBS) before the addition of Trypsin-EDTA. Cells were incubated at 37°C and 5% CO_2 for 5 minutes. Trypsin-EDTA was neutralized by adding 5x volume culture media. Cells were reseeded at split ratios of 1:5 to 1:10 and culture media was replaced 24 hours following cell passaging.

2.2.1.2 Human embryonic stem cells

Human embryonic stem cell line H9 was maintained in thermostable E8 media (S8; prepared in house) at 37°C and 5% CO_2 in T12.5 flasks coated with Geltrex. S8 media consists of DMEM/F12 supplemented with 14µg/L sodium selenite, 19.4mg/L insulin, 40µg/L thermostable FGF2, 1383mg/L $NaHCO^3$, 10.7mg/L transferrin, 2µg/L TGFβ1 and 10ml/L GlutaMAX. For preparation of culture flasks and plates, hESC-qualified

LDEV-Free Geltrex (Gibco #1413302) was thawed on ice for 3 hours and diluted 1:100 in cold DMEM/F12. Diluted Geltrex was added to the flasks or plates and incubated at 37°C for 1 hour followed by at least 2 hours incubation at room temperature. For T12.5 flasks 1.5ml 1:100 Geltrex was used, for 6- and 12-well plates 1ml 1:100 Geltrex was used, and for 24 well plates 500µl 1:100 Geltrex was used. H9 cells were routinely passaged when they reached 60-80% confluency, as judged by microscopy, and were maintained for no more than 10 passages post-thawing to minimize the risk of karyotypic abnormalities arising. Cells were passaged as clumps by removing culture media and adding 1ml ReLeSR (STEMCELL Technologies #100-0484) for 30 seconds at room temperature. Following incubation, ReLeSR was aspirated and cells were incubated for 4 minutes at room temperature. S8 was then added followed by gentle tapping to lift cell clumps and gentle trituration with a 5ml serological pipette to break up clumps to optimal sizes (50 - 200µm). Clumps were subsequently replated at split ratios of 1:3 to 1:5.

For some applications H9s were lifted/seeded as single cells. To achieve this, culture media was removed and cells were washed with PBS. TrypLE (Gibco #12604021) was added and cells were incubated at 37°C for 4 minutes. TrypLE was neutralized by adding 10x volume S8. Cells were then centrifuged at 300g, media was aspirated, and the cell pellet was resuspended in a suitable (application-specific) volume of culture media supplemented with 10µM Rho kinase inhibitor Y-27632 (Tocris #1254).

2.2.2 Plasmid transfection

2.2.2.1 HCT116 cells

HCT116 cells were seeded into 24 well plates at 100,000 cells/well 24 hours prior to transfection. On the day of transfection, culture media was replenished with 450µl fresh culture media. Transfections mixes were made by up mixing polyethylenimine (PEI) with the plasmid DNA at a 3:1 ratio (PEI:pDNA) in 50µl warm Opti-MEM (Gibco #31985062). Transfection mixes were vortexed for 30 seconds then incubated at room temperature for

10 minutes before being added to each well. As a standard, 500ng pDNA and 1,500ng PEI was used for transfections, unless specified otherwise.

2.2.2.2 H9 cells

H9 cells were seeded as single cells into 24 well plates at 50,000 cells/well in 500µl mTeSR-Plus (STEMCELL Technologies #100-0276) supplemented with 10µM Y-27632. Media was replaced with 450µl mTeSR-Plus without Y-27632 24 hours after seeding. 48 hours following media change transfection mixes were made up by mixing 2µl Lipofectamine Stem (Invitrogen #STEM00001) with 500ng pDNA in 50µl warm Opti-MEM. Transfection mixes were vortexed for 30 seconds and incubated at room temperature for 10 minutes before being added to each well.

2.2.3 siRNA transfection (UPF1 knockdown)

2.2.3.1 HCT116 cells

HCT116 cells were seeded into 24 well plates at 100,000 cells/well 24 hours prior to transfection. Media was replaced with 450µl fresh media on the day of transfection. Transfection was conducted using 1.5µl Lipofectamine RNAiMAX (Invitrogen #13778075) and 30nM siRNA in 50µl warm Opti-MEM. Lipofectamine complexes were allowed to form for 10 minutes at room temperature before being added to each well. Media was replenished 24 hours following transfection. A second-hit was conducted a further 24 hours later. Cells were lysed at 72 hours post-transfection to assess the knockdown at the protein and RNA expression level, RNA-seq was also conducted at this time point.

2.2.3.2 H9 cells

H9 cells were transfected via reverse-transfection using DharmaFECT 1 (Horizon Discovery #T-2001-03). 20nM siRNA and 2µl Dharmafect 1 were added to 50µl Opti-MEM, vortexed, and incubated at room temperature for 20 minutes. 50µl

transfection mix was added to 100,000 cells suspended in 450µl mTeSR-Plus and 10µM Y-27632 and seeded into 24 well plates. Media was replaced with 500µl mTeSR-Plus without Y-27632 24 hours post-transfection. Cells were lysed at 48 hours post-transfection to assess the knockdown at the protein and RNA expression levels.

2.3 General molecular biology

2.3.1 RNA extraction

RNA extraction was conducted using either TRIzol (Sigma-Aldrich #T9424) or Total RNA Purification Plus Kit (Norgen Biotek Corp #48300). Both methods are described below.

2.3.1.1 TRIzol nucleic acid extraction

For TRIzol RNA extraction samples were processed following the manufacturer's instructions, with the addition of a DNAse treament step. Cells were lysed in 750µl (6 or 12 well plate setup) or 350µl (24 well plate setup) TRIzol. Plates were incubated at room temperature for 5 minutes followed by scraping each well with a P1000 pipette and pipetting up and down to ensure detachment and thorough lysis. Samples were transferred to 1.5ml Eppendorf tubes and incubated at room temperature for a further 5 minutes. 200µl chloroform was added per 750µl TRIzol and samples were shaken vigorously for 30 seconds, followed by a 10 minute incubation at room temperature. Samples were centrifuged at 12,000g for 15 minutes at 4°C. The aqueous upper phase of each sample was transferred to a fresh Eppendorf tube containing 1µl GlycoBlue (5mg/ml; Invitrogen #AM9515 diluted 1:3 in ddH_2O) and equal volume (relative to the aqueous upper phase) of isopropanol. Samples were thoroughly mixed and incubated at room temperature for 10 minutes before being centrifuged at 12,000g for 10 minutes at 4°C. Supernatants were removed and each pellet was washed with 1.2ml 75% EtOH before a further 12,000g spin for 10 minutes at °C. EtOH was removed and pellets air

dried at room temperature for 15 minutes.

2.3.1.2 DNAse treatment and retrieval of RNA

Pellets were resuspended in 43µl ddH_2O before 1µl RiboSafe RNAse inhibitor (Bioline #BIO65028), 1µl TURBO^{TM} DNAse and 5µl 10X TURBO^{TM} DNAse buffer (Invitrogen #AM2238) were added. Samples were incubated for 1 hour at 37°C and 450rpm in an Eppendorf Thermomixer. After DNAse treatment, 50µl ddH_2O was added to each sample followed by 100µl acidic phenol-chloroform (pH 4.5). Samples were shaken vigorously for 30 seconds followed by a 5 minute incubation at room temperature. Samples were then centrifuged at 12,000g for 5 minutes at 4°C before 90µl aqueous upper phase was transferred to a fresh Eppendorf tube containing 1µl GlycoBlue, 10µl NaAc (3M, pH 5.8) and 250µl 100% EtOH. Samples were thoroughly mixed and incubated overnight at -80°C. RNA was pelleted via centrifugation at 17,000g at 4°C for 30 minutes, washed with 75% ethanol, and resuspended in ddH_2O.

2.3.1.3 RNA extraction using Total RNA Purification Plus Kit

For Total RNA Purification Plus Kit, RNA extraction was carried out as per the manufacturer's instructions. Lysates from samples with >10^6 cells were passed through a 25 gauge needle 10 times. Both gDNA removal and RNA binding columns were used. All centrifugation steps were carried out at 4°C. For final elution, buffer was left on the column membrane for 3 minutes to maximise RNA yield.

2.3.1.4 Measuring RNA quality and quantity

Following RNA extractions, RNA concentration and A260/A280 ratios were obtained using a NanoDrop^{TM} Lite Spectrophotometer. In some instances, RNA concentration was further confirmed using Qubit^{TM} RNA Broad-Range Assay Kit (Invitrogen #Q10210) with the Qubit 4 Fluorometer (Invitrogen).

2.3.2 cDNA synthesis

cDNA synthesis was conducted using High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems^{TM} #4368814). 1µg of RNA was diluted in total volume of 14.2µl ddH_2O before 0.8µl 100mM dNTP mix and 2µl 10X Random Primers were added. Samples were incubated at 70°C for 5 minutes followed by 2 minutes incubation at 4°C. Subsequently 2µl 10X RT Buffer, 0.5µl MultiScribe^{TM} Reverse Transcriptase and 0.5µl RiboSafe RNAse inhibitor was added. ProFlex PCR System (Applied Biosystems^{TM}) was programmed as follows: 1) 10 minutes at 25°C; 2) 120 minutes at 37°C; 3) 5 minutes at 85°C; 4) Hold at 4°C. Following reverse transcription cDNA was diluted 5-fold with ddH_2O to a final volume of 100µl.

2.3.3 DNA extraction

For DNA extraction, cells were washed with PBS, lifted, and spun at 300g for 5 minutes. Cell pellets were then resuspended in 200µl extraction buffer containing 0.05% sodium dodecyl sulfate and 0.05mg/ml proteinase K, in TE buffer. Samples were incubated for 4 hours at 60°C and 450rpm in an Eppendorf Thermomixer. After 4 hours 200µl phenol-chloroform (pH 7.9) was added to each sample. Each sample was vigorously shaken for 30 seconds followed by a 5 minute incubation at room temperature. Samples were then centrifuged at 12,000g for 5 minutes at 4°C. 180µl upper aqueous phase was then transferred to an Eppendorf tube containing 1µl GlycoBlue, 20µl NaAc (3M, pH 5.8) and 500µl 100% EtOH. Samples were incubated at -80°C overnight. DNA was pelleted via centrifugation at 17,000g at 4°C for 30 minutes, washed with 75% ethanol, and resuspended in ddH_2O. DNA concentration and A260/A280 ratios were obtained using a NanoDrop^{TM} Lite Spectrophotometer.

2.3.4 PCR and qPCR

PCR was conducted on ProFlex PCR System (Applied Biosystems^{TM}). For validation of 3'UTR splicing using flanking primers Quick-Load Taq 2X Master Mix (New England Biolabs #M0271L) was used with both cDNA and gDNA as inputs. To validate whether the PCR reaction was successful, agarose gel (typically 0.8-2% agarose in TBE, plus ethidium bromide) electrophoresis was conducted and imaged on Bio-Rad ChemiDoc. For more complex applications such as molecular cloning, Q5® High-Fidelity DNA Polymerase (New England Biolabs #M0491L) was used as per the manufacturer's instructions.

qPCR was conducted in 10µl reactions made us as follows: 5µl 2X SensiMix^{TM} SYBR® Hi-ROX (Bioline #QT605-05); 0.5µl 10µM Forward Primer; 0.5µl 10µM Reverse Primer; 2µl ddH_2O; 2µl cDNA. For most amplifications of protein coding genes 20ng cDNA was used as input (2µl from cDNA synthesis). For highly expressed housekeeping genes such as 18S, 0.67ng cDNA was used, requiring a further 1:30 dilution of cDNA. qPCR was conducted using Rotor-Gene Q (QIAGEN) and relative abundances were calculated using the $2^{-\Delta\Delta Ct}$ method.

2.3.5 SDS-PAGE & Western blotting

Western blotting was conducted primarily to assess the effectiveness of protein knockdown following siRNA transfection (as described in 2.2.3). Culture media was aspirated from cells followed by 1X PBS wash. Following the removal of PBS, cells were lifted (as described in 2.2.1), transferred to a 1.5ml Eppendorf tube, and centrifuged at 300g for 4 minutes to pellet the cells. Following the aspiration of media, an appropriate amount of lysis buffer (50mM HEPES-NaOH, 100mM NaCl, 1mM EDTA, 0.5% Triton X-100, 10% Gylcerol, 1mM DTT, protease inhibitors) was used to lyse the cell pellet (usually 100µl unless cell number was low, in which case 30µl lysis buffer was used).

Tubes were incubated on ice for 30 minutes, with agitation via pipetting every 5 minutes, then centrifuged at 17,000g for 10 minutes at 4°C to separate cell debris from protein lysate. Supernatants were transferred to fresh Eppendorf tubes and kept on ice. Protein concentration was obtained via Bradford Protein Assay (Bio-Rad). Briefly, 799µl ddH_2O, 1µl protein lysate, and 200µl 5X Bradford reagent were added to semi-micro cuvettes (VWR #634-0676) and mixed thoroughly. Concentrations were obtained using the Jenway Genova Plus Spectrophotometer and pre-programmed Bradford method.

Polyacrylamide gels were cast at the desired percentage (8% resolving gels were used for UPF1 western blots) with a 5% upper stacking gel, as below:

Table 2.1: Composition of SDS-polyacrylamide gels (10ml)

Component	**Resolving Gel**	**5% Stacking Gel**
Protogel (Geneflow #EC-890)	variable	1.2ml
4X Resolving Buffer (1.5M Tris-HCl, pH 8.8, 0.15% SDS)	2.5ml	-
4X Stacking Buffer (0.5M Tris-HCl, pH 6.8, 0.15% SDS)	-	2.5ml
H_2O	variable (to 10ml)	6.3ml
10% APS	150µl	110µl
TEMED	20µl	20µl

20µg protein lysate was diluted to 22.5µl with ddH_2O. In cases where less than 20µg of protein was obtained, or samples were too dilute, then 22.5µl of the lowest concentration sample was used, and an appropriate volume of the other samples was used, and diluted to 22.5µl with ddH_2O, so that the same amount of protein was added to each well. 7.5µl 4X protein loading buffer (200mM Tris-HCl, 1% Bromophenol blue, 10% SDS, 50% Glycerol) was added to each sample. Samples were then boiled at 95°C for 5 minutes. Gels were loaded into a Mini-PROTEAN vertical electrophoresis cell (Bio-Rad) and reservoirs were filled with 1X running buffer (25mM Tris, 250mM Glycine, 0.1% SDS). Samples were loaded alongside 3µl PageRuler^{TM} Plus Prestained Protein Ladder (Thermo Scientific #26619). Samples were run at 25mA (per gel) for 15 minutes then increased to 35mA (per gel) until sufficient separation of ladder bands was observed. Gels were

transferred onto nitrocellulose membranes (Serva #71224.01) at 25V for 20 minutes in transfer buffer (25mM Tris, 250mM Glycine, 0.1% SDS, 20% methanol) using the Trans-Blot Turbo^{TM} (Bio-Rad). Successful transfer was checked using Ponceau S solution (Sigma #P7170-1L), and membranes were cut to desired size ranges.

Membranes were blocked with 5% skim milk blocking buffer (5% skim milk powder in 1X TBST; 1X TBST is composed of 20mM Tris-HCl, 137mM NaCl, and 0.1% Tween-20) for 1 hour at room temperature or overnight at 4°C. Membranes were then incubated with primary antibody against the protein of interest diluted in 5% skim milk blocking buffer (1:1000 for mouse α-UPF1 (Proteintech #66898); 1:10,000 for mouse α-Tubulin (Sigma-Aldrich #T6199)) for 2 hours at room temperature on a shaker. After 2 hours, membranes were washed 3X with TBST for 5 minutes each. Membranes were then incubated with HRP-conjugated secondary antibody diluted in 5% skim milk blocking buffer (1:10,000 for goat α-Mouse (Promega #W402B)) for 1 hour at room temperature on a shaker. Membranes were subsequently washed 3X with TBST for 5 minutes each. Signal was developed by adding equal volumes of ECL1 (2.5mM Luminol, 100mM Tris-HCl pH 8.5, 400µl p-coumaric acid) and ECL2 (100mM Tris-HCl pH 8.5, 5.3mM H_2O_2) solutions to the membrane and shaking for 30 seconds. Membranes were then exposed using the Bio-Rad ChemiDoc and auto-exposure setting.

2.3.6 Molecular cloning

2.3.6.1 PCR

For molecular cloning, Q5® High-Fidelity DNA Polymerase (New England Biolabs #M0491L) was used. PCR primers were designed such that they had an 20-30 nucleotide overhang which was complementary to the sequence it would be ligated to. Additionally PCR primer pairs were designed to have an annealing temperature of no more than 72°C. PCRs were conducted in 25µl reactions using either 10ng plasmid DNA, 20ng cDNA, or 50ng gDNA as a template. The PCR reaction consists of an initial denaturation at 98°C for

30 seconds followed by 30 cycles of: denaturation at 98°C for 10 seconds; annealing at 72°C for 30 seconds; and extension at 72°C for 40 seconds per kilobase of template. A final extension for 2 minutes at 72°C proceeds the final cycle followed by an indefinite hold at 4°C. Where plasmid DNA was used as a template, following PCR, 1µl DpnI (New England Biolabs #R0176L) was added. Each tube was then incubated at 37°C for 1 hour followed by heat inactivation at 80°C for 20 minutes. 5µl 6X Purple DNA Gel Loading Dye (New England Biolabs #B7024S) was then added to each tube and gel electrophoresis was conducted (as previously described in Section 2.3.4), followed by gel extraction (described in Section 2.3.6.3) to retrieve the fragment of interest.

2.3.6.2 Restriction enzyme digestion

Restriction enzymes purchased from New England Biolabs were used for molecular cloning, utilizing rCutSmart buffer (New England Biolabs #B6004). Restriction enzyme digests were conducted in 25µl reactions consisting of: 1µl restriction enzyme; 2.5µl 10X rCutSmart buffer; template DNA; ddH_2O up to 25µl. Double restriction enzyme digests were conducted in a single reaction, using 1µl of each restriction enzyme (where the optimal incubation temperatures matched). Reactions were incubated at 37°C in the ProFlex PCR System (Applied Biosystems^{TM}) for 1 hour, followed by heat inactivation at 80°C for 20 minutes.

2.3.6.3 Gel extraction

Gel extraction was conducted using the QIAquick Gel Extraction Kit (QIAGEN #28706). Following agarose gel electrophoresis, the gel was placed onto the Bio-Rad Chemidoc and bands were visualized via the UV transilluminator. A scalpel blade was used to excise the band of interest, which was then placed into a 1.5ml Eppendorf tube. The band was then weighed. The band was then dissolved in 3X volume Buffer QG (i.e. 300ul per 100mg of gel) at 50°C and 450rpm for 10 minutes on an Eppendorf Thermomixer. Once the gel was dissolved, isopropanol equal to the gel volume (i.e. 100µl per 100mg of original gel) was

added and thoroughly mixed. Each sample was then placed into a QIAquick spin column and centrifuged at 12,000g for 1 minute. The column was subsequently washed with 500µl Buffer QG and centrifuged at 12,000g for 1 minute. The column was also washed with 750µl Buffer PE, allowed to stand for 1 minute, and centrifuged at 12,000g for 2 minutes. DNA was eluted through the addition of 20µl ddH_2O, incubation at room temperature for 5 minutes, and centrifugation at 12,000g for 1 minute. DNA concentration and A260/A280 ratios were obtained using a NanoDrop^{TM} Lite Spectrophotometer.

2.3.6.4 Gibson assembly

Gibson assembly was conducted in 10µl reactions consisting of 5 µl 2X NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs #E2621L) and 5µl vector and insert fragments in ddH_2O. The amount of vector and insert was determined by their DNA molar ratios using the NEBioCalculator (nebiocalculator.neb.com/#!/ligation). Where only one insert was to be inserted into a vector, a vector:insert ratio of 1:2 was used. A maximum of 100ng vector was used. However, where DNA concentrations from PCR and gel extraction were substantially lower, the maximum amount of DNA (for vector and insert) that could be used in a final volume of 5µl was used. Reactions were incubated at 50°C in the ProFlex PCR System (Applied Biosystems^{TM}) for 1 hour before proceeding to bacterial transformation.

2.3.6.5 Bacterial culture

DH5-α E. coli cells were used for all bacterial culture experiments. Cells were thawed at 4°C for 20 minutes from the lab's communal bank (stored at -80°C). 10ng of plasmid DNA, or 10µl products of a Gibson assembly reaction were added to 100µl DH5-α cells and incubated on ice for 30 minutes to allow thorough diffusion of the plasmid throughout the mixture of cells. Cells were then heat shocked for 35 seconds at 42°C before being placed back onto ice for 2 minutes. 900µl LB media without antibiotics was then added and cells were incubated at 37°C and 300rpm for 1 hour on an Eppendorf

Thermomixer to allow recovery and expression of the antibiotic resistance gene. Cells were then centrifuged for 3 minutes at 3,000g to pellet them before 900µl media was removed. Cells were resuspended in the remaining media and transferred to an LB agar plate containing the antibiotic matching the plasmid resistance cassette (100µg/ml for ampicillin; 50µg/ml for kanamycin). Cells were distributed across the agar plate using a sterile glass spreader, and incubated overnight at 37°C. Individual colonies were then picked with a 10µl pipette tip and inoculated into LB media supplemented with the antibiotic being selected against (100µg/ml for ampicillin; 50µg/ml for kanamycin). Where the creation of plasmids needed to be validated, bacteria were inoculated into 7ml LB media supplemented with antibiotics in 15ml sterile polypropylene tubes. Where validated plasmids were expanded for use in mammalian cell transfection experiments, bacteria were inoculated into 50ml LB media supplemented with antibiotics in 250ml sterile conical flasks. Cultures were incubated for 18 hours at 37°C in a shaking incubator.

2.3.6.6 Extracting plasmid DNA

For extraction of plasmid DNA to validate the creation of specific plasmids, plasmid DNA was isolated using QIAprep Spin MiniPrep Kit (QIAGEN #27104). For extraction of plasmid DNA for use in mammalian transfection experiments, plasmid DNA was isolated using Plasmid Plus Midi Kit (QIAGEN #12945) due to the presence of an endotoxin removal step. Both kits were used in accordance with the manufacturer's instructions. For extraction using the Plasmid Plus Midi Kit, the high-yield protocol was followed. In both cases, final elution was done using ddH_2O instead of Buffer EB, and was incubated on the column membrane for 5 minutes at room temperature before centrifugation.

2.3.7 Primer list

Table 2.2: List of primers and their usage

Name	Sequence	Dir	Usage
HNRNPDL (ret)	TGTCACCCTGACCACGTCTA	FW	RT-qPCR
HNRNPDL (ret)	CGCTGGTACATGAAGTTGGA	RV	RT-qPCR
HNRNPDL (spl)	TGGAGAAAACAGGAGGAGATGT	FW	RT-qPCR
HNRNPDL (spl)	GGCAGCTATATACAGTTGGACACA	RV	RT-qPCR
DUSP28 (ret)	GACCACGGATGCTTCTGTTT	FW	RT-qPCR
DUSP28 (ret)	TCCGAACTCCTATGGGACAG	RV	RT-qPCR
DUSP28 (spl)	AAGGATGTCCCTGCACTGAT	FW	RT-qPCR
DUSP28 (spl)	GAGCTCCTTCCTGCAGACAC	RV	RT-qPCR
JUP (ret)	GCAGGCTTTTCCTCCTCTCT	FW	RT-qPCR
JUP (ret)	GAAGCTCAGCAGCAAAGGA	RV	RT-qPCR
JUP (spl)	CCCCAGTGCGGTTCCTCAT	FW	RT-qPCR
JUP (spl)	ACTTTTCTGTCTTGCCCCATC	RV	RT-qPCR
CTNNB1 (ret)	GTTGGTAGGGTGGGAGTGG	FW	RT-qPCR
CTNNB1 (ret)	GCAAAAGCTGTGGCAAAAA	RV	RT-qPCR
CTNNB1 (short_spl)	TGACCTGTAAATCATCCTTTA GCTGTATT	FW	RT-qPCR
CTNNB1 (short_spl)	CCTCGACCAAAAAGGACCAGA	RV	RT-qPCR
HRAS (ret)	CGCAGGTGAGGGGGACT	FW	RT-qPCR
HRAS (ret)	ATGTCCTGAGCTTGTGCTGG	RV	RT-qPCR
HRAS (spl)	TCTCCTGACGCAGCACAAG	FW	RT-qPCR
HRAS (spl)	TGACTGGGCTCCAGCAG	RV	RT-qPCR
HRAS flank	GGAGTGGAGGATGCCTTCTACA	FW	RT-PCR
HRAS flank	TGACTGGGCTCCAGCAG	RV	RT-PCR
CTNNB1 flank	GGCCTGGTTTGATACTGACC	FW	RT-PCR
CTNNB1 flank	CGACCAAAAAGGACCAGAAC	RV	RT-PCR
SRSF8 flank	GGACAGATGTCCTCTTAAGAAAATG	FW	RT-PCR
SRSF8 flank	GTGTGATTATCGTCAGCTAGATGTG	RV	RT-PCR
Luc-CTNNB1 into pCI-neo	CACTATAGGCTAGCCTCGAGCTTCA CCATGGAAGATGC	FW	Gibson
Luc-CTNNB1 into pCI-neo	GCCGCCCGGGTCGACTGAATGAAT TAAAAGTTTAATTCTGAACC	Rv	Gibson
Luc-HRAS into pCI-neo	CACTATAGGCTAGCCTCGAGCTTCA CCATGGAAGATGC	FW	Gibson

Luc-HRAS into pCI-neo	GCCGCCCGGGTCGACTGGAGGGTC TGCAGTCACCT	Rv	Gibson
CTNNB1 3′UTR	ATCATCCTTTAGGTAAGAAGTTTTA	FW	Gibson
CTNNB1 3′UTR	GAATGAATTAAAAGTTTAATTCTG AACC	RV	Gibson
CTNNB1 ΔI(S)	CTGTATTGTCTGAACTTGCATTGTG ATTG	FW	Mutagenesis
CTNNB1 ΔI(S)	CTAAAGGATGATTCAATGGTGATG GTGA	RV	Mutagenesis
CTNNB1 ΔI(L)	TGAATCATCCTTTAGGAGTAACAATA CAAATGGATTTTGGGAGT	FW	Mutagenesis
CTNNB1 ΔI(L)	GTTCAGACAATACAGCTAAAGGATG ATTCAATGGTGATGGTGA	RV	Mutagenesis
CTNNB1 5′SS	ATCATCCTTTAGGGAAGAAGTTTTA	FW	Mutagenesis
CTNNB1 5′SS	TCAATGGTGATGGTGATGATGACCG	RV	Mutagenesis
CTNNB1 3′SS(S)	TACTGCTGTATTGTCTGAACTTGCAT	FW	Mutagenesis
CTNNB1 3′SS(S)	TAAAGAGAGAAAGGCTGCTATTA ATGTT	RV	Mutagenesis
CTNNB1 3′SS(L)	TTGGTCGTGGAGTAACAATACAA ATG	FW	Mutagenesis
CTNNB1 3′SS(L)	AAAGGACCAGAACAAAAAGTTT ACTTC	RV	Mutagenesis
HRAS 5′SS	CAGGAGAGGGGGACTCCCAGGG CGGC	FW	Mutagenesis
HRAS 5′SS	GTCCCCCTCTCCTGCGTCAATGG TGATGG	RV	Mutagenesis

2.4 Immunolabelling

2.4.1 Immunocytochemistry

Culture media was removed from cells followed by 2X washes with PBS. Cells were fixed by adding 4% Paraformaldehyde (PFA) in PBS with 10% FBS (ICC buffer) for 10 minutes. After the buffer was removed, cells were permeabilized by adding 0.2% Triton X-100 in ICC buffer for 10 minutes. Cells were then washed with and incubated in ICC buffer for 1

hour at room temperature. ICC buffer was removed and primary antibodies (in ICC buffer) were added: mouse α-MYH4 mAb (Invitrogen #14-6503-82) at 1:100; rabbit α-CTNT mAb (Invitrogen #701620) at 1:100. Samples were incubated overnight at 4°C. Samples were washed 3X with ICC buffer. Secondary antibodies (in ICC buffer) were added: Alexa Fluor 488 conjugated Goat α-Mouse (Jackson ImmunoResearch Laboratories #115-545-044) at 1:200; Alexa Fluor 594 conjugated Goat α-Rabbit (Jackson ImmunoResearch Laboratories#111-585-003) at 1:200. Additionally, Hoechst 33342 (Tocris #5117) was added at 1:10,000. Samples were incubated for 2 hours at 4°C. Samples were washed 3X with ICC buffer. Samples were imaged using an IN Cell Analyzer 2200.

2.4.2 Flow cytometry

2.4.2.1 Sample preparation

Culture media was removed from cells followed by 1X wash with PBS. Cells were lifted with either TrypLE, or in the case of cardiomyocytes, using STEMdiff™ Cardiomyocyte Dissociation Kit (STEMCELL Technologies #05025). Dissociation reagent was inactivated by adding 10X volume DMEM/F12. To remove clumps, cells were passed through a 40µm cell strainer (Falcon™ #10737821). Cells were then pelleted by centrifugation at 300g for 5 minutes. The liquid was aspirated and cells were fixed through resuspension in 1ml 4% PFA in ICC buffer and incubation at room temperature for 10 minutes. 9ml ICC buffer was subsequently added and samples were centrifuged at 300g for 5 minutes. ICC buffer was aspirated and cells were resuspended in 0.2% Triton X-100 in ICC buffer for 10 minutes to permeabilize them. 9ml ICC buffer was subsequently added and samples were again centrifuged at 300g for 5 minutes. Pellets were then resuspended in 800µl ICC buffer and split into two 1.5ml Eppendorf tubes (400µl each). Primary antibody against the target of interest was added to one tube whilst the second tube acted as the "no primary" negative control. Mouse α-CTNT mAb (Invitrogen #MA5-12960) was added at 1:200. Samples were incubated for 2 hours at room temperature. Following incubation

cells were centrifuged at 300g for 3 minutes followed by resuspension in 1.2ml ICC buffer, this washing process was repeated for a total three times. Alexa Fluor 488 conjugated Goat α-Mouse secondary antibodies were then added at 1:300 in ICC buffer to all samples (including the "no primary" control). Samples were incubated at room temperature in the dark for 1 hour. Following incubation cells were washed 3X in 1.2ml ICC buffer. Fixed and stained cells were then resuspended in 1ml ICC buffer, transferred to 5ml round base polystyrene flow cytometry tubes, and topped up to 3ml with ICC buffer.

2.4.2.2 Using the flow cytometer

Flow cytometry was conducted using the BD FACSJazz^{TM} Cell Sorter (BD Biosciences). A series of gating options were used to identify the populations of interest, the cells included in a higher level gate were used as the input for the following gating step. Gating options are configured using the "no primary" control as input. Firstly, a dot plot of forward (FSC) vs side scatter (SSC) was generated and used to gate for the cell population of interest and to exclude debris. Next, a dot plot of SSC vs trigger pulse width was generated and used to filter out doublets or clumps of cells. Finally a histogram was generated to visualize the number of cells based on their fluorescence levels in the 530/40 (488) channel, the "no primary" negative control was used as input to determine background fluorescence and determine a "detection gate". Primary-stained cells are then used as input and this detection gate represents the proportion of cells that are positive for expression of the protein of interest in the sample.

2.5 Differentiation of hESCs

2.5.1 Cardiomyocyte differentiation

H9 hESCs were used for cardiomyocyte differentiation. H9 cells were used no earlier than three passages post-thawing to allow sufficient recovery, and no later than 10 passages post-thawing to minimize risk of karyotypic abnormalities arising. Cells were

differentiated using STEMdiff™ Ventricular Cardiomyocyte Differentiation Kit (STEMCELL Technologies #05010). hESC-Qualified Corning™ Matrigel ™ (Corning #354277) was diluted 1:100 in DMEM/F12 and 1ml was used to coat each well of a 12 well plate. Plates were incubated at 37°C and 5% CO_2 for 1 hour, followed by 2 hours incubation at room temperature. Matrigel was aspirated immediately prior to seeding cells, and wells were not allowed to dry. H9 cells were lifted as single cells using TrypLe and seeded at a density of 600,000 cells per well in 1ml mTeSR Plus supplemented with 10µM Y-27632. Cells were incubated at 37°C and 5% CO_2 for 24 hours. Media was then replenished with 1ml fresh mTeSR Plus without Y-27632 and cells were incubated for a further 24 hours at 37°C and 5% CO_2. Confluency of the cells was then assessed and was required to be >95%. 20µl undiluted Matrigel was added to 2ml Cardiomyocyte Differentiation Medium A, mTeSR Plus was aspirated, and replaced with the differentiation media. This was considered day 0 of differentiation. Cells were cultured at 37°C and 5% CO_2 for 48 hours. On day 2 media was replaced with 2ml Cardiomyocyte Differentiation Medium B. On day 4 media was replaced with 2ml Cardiomyocyte Differentiation Medium C. On day 6 media was replaced with a second dose of 2ml Cardiomyocyte Differentiation Medium C. On day 8 media was replaced with 2ml Cardiomyocyte Maintenance Medium, which was subsequently refreshed every 2 days. Differentiation efficiency was assessed at day 15 via immunocytochemistry (MYH4 and CTNT) and flow cytometry (CTNT).

2.6 RNA stability assays

2.6.1 Low throughput - Actinomycin D treatment

To assess relative RNA stability in colorectal carcinoma cells, HCT116 cells were seeded at 300,000 cells per well into 6-well plates in DMEM supplemented with 10% FBS and Penicillin/Streptomycin. After 48 hours, culture media was supplemented with 5µg/ml actinomycin D. RNA was extracted at the following time points: 0 hours; 0.5 hours; 1 hour;

3 hours; 6 hours; and 12 hours.

To assess the relative RNA stability in human embryonic stem cells, H9 cells were seeded at 300,000 cells per well into 6-well plates in S8 supplemented with 10µM Y-27632. After 24 hours, culture media was replaced with S8 without Y-27632. After a further 24 hours, culture media was supplemented with 5µg/ml actinomycin D. RNA was extracted at the following time points: 0 hours; 0.5 hours; 1 hour; 3 hours; and 6 hours.

Following RNA extraction, cDNA was synthesised and qPCR was performed. Expression was normalized to the most stable isoform within each splice pair (ΔCt) and then to the 0 hour time point ($\Delta\Delta$Ct) before being plotted. The processing of raw Ct values into RNA stability plots was automated using the process_qPCRs pipeline (github.com/jjriley1/process_qPCRs)

2.6.2 High throughput - SLAMseq

2.6.2.1 Cell viability optimization

The effect of culturing H9 cells with increasing concentrations of 4sU was initially assessed at the day 0 (undifferentiated) stage. 15 wells (in 12 well plates) were seeded at day -2 with 600,000 cells per well in 1ml mTeSR Plus supplemented with 10µM Y-27632. 24 hours later, media was replenished with 1ml fresh mTeSR Plus without Y-27632. After a further 24 hours, media was replaced with 1ml mTeSR Plus supplemented with increasing concentrations (0µM, 25µM, 50µM, 75µM and 100µM) of 4sU. Importantly, 4sU media and 4sU treated samples were kept in the dark at all times. Three biological replicates were considered for each concentration. Media was replaced with fresh 4sU supplemented mTeSR Plus every 3 hours for a total of 12 hours. After 12 hours cell viability was assessed by counting cells with a hemocytometer. Proportional viability was calculated as the number of cells per well divided by the average number of cells in the 0µM treated samples.

2.6.2.2 Testing incorporation efficiency

Cells were prepared such that at day 0, 2 and 15 they could be treated with 4sU in parallel. As above, cells were cultured in mTeSR Plus in the presence of 4sU for 12 hours in the dark with media changes every 3 hours. For day 0, a 0µM 4sU control was also considered to allow comparison between treated and untreated, thus facilitating a comparison against the background level of T->C conversions, potentially caused by IAA treatment. After 12 hours of 4sU treatment, RNA was extracted with Trizol. For day 0 65µM 4sU treated, a second replicate was conducted and RNA extracted using Total RNA Purification Plus Kit (Norgen Biotek Corp #48300) to compare the effects of RNA processing on relative 4sU-containing RNA yield.

For Trizol RNA extraction, the procedure described in 2.3.1 was slightly modified to maintain reducing conditions, through supplementation with Dithiothreitol (DTT). Additionally, all stages of RNA extraction were carried out in the dark. During isopropanol and ethanol precipitation steps, each sample was supplemented with 1/100 total volume 10mM DTT (for a final concentration of 100 µM). 100µM DTT was also required in each 75% EtOH wash. Importantly, DTT was not added during the DNAse treatment step. Following phenol-chloroform extraction and ethanol precipitation, RNA was resuspended in ddH_2O supplemented with 1mM DTT. Where the Total RNA Purification Plus Kit was used, each wash buffer was supplemented with 1/100 total volume 10mM DTT (100µM final concentration) and the final elution buffer was supplemented with 1/10 total volume 10mM DTT (1mM final concentration).

Each sample was then treated with iodoacetamide (IAA) to facilitate thiol modification. Importantly, fresh IAA stock was made on the day. IAA treatment reactions were made up as follows:

Table 2.3: Composition of iodoacetamide treatment reaction

Component	**Amount**
RNA	900ng
H_2O	variable
IAA (100mM in EtOH)	5µl
$NaPO_4$ (500mM; pH 8)	5µl
DMSO	25µl
Final Volume	50µl

Reactions were incubated at 50°C for 15 minutes before being quenched with 1µl 1M DTT. RNA was precipitated overnight at -80°C in 125µl EtOH, 5µl NaAc (3M, pH 5.8) and 1µl GlycoBlue (5mg/ml; Invitrogen #AM9515 diluted 1:3 in ddH_2O). Samples were centrifuged at 17,000g for 30 minutes at 4 °C before supernatant was removed and RNA pellets washed with 1.2ml 75% EtOH. Samples were centrifuged for a further 15 minutes at 17,000g at 4°C before EtOH was removed and pellets were air dried for 15 minutes at room temperature. Pellets were resuspended in 30µl ddH_2O. RNA was quantified using a NanoDrop™ Lite Spectrophotometer before samples were sent to Novogene (Cambridge, UK) for library preparation (Poly(A) enrichment and cDNA synthesis) and RNA-sequencing (Illumina PE150) at a relatively low depth (7M - 10M read pairs).

Raw sequencing reads were mapped to hg38 using STAR alongside the settings stated in 2.1.2. To determine the incorporation efficiency of each 4sU-treated sample, mapped reads in .bam format were passed through pipeline_slam_3UIs (github.com/jjriley1/slam_3UIs). Pipeline_slam_3UIs contains multiple python scripts that utilise the pysam package to interrogate the mapped reads (.bam) files. Multiple scripts were developed, with increasing levels of complexity (as will be examined in Section 5). The most complex of these scripts was the stranded_conversions_ per_pair_no_snp.py script, which will now be broken down. Prior to .bam files being passed to the stranded_conversions_per_pair_no_snp.py script, they were first sorted with samtools (Danecek et al., 2021) using the `-n` option to specify sorting by read name, thus allowing the processing of read pairs via pysam. Next, the sorted .bam files were

passed to the featureCounts package (Liao et al., 2014) to allow us to determine which transcripts each read could correspond to, but more importantly, which strand these transcripts were transcribed from. This was necessary in order to determine whether to look for T>C or A>G conversions on the forward read, and vice-versa for the reverse read, in each pair. To ensure that observed T>C or A>G conversions were not the result of single nucleotide polymorphisms (SNPs) or condition-specific RNA editing, the .bam file from the no4sU control (not treated with 4sU) was passed to the varscan package (Koboldt et al., 2012) `mpileup2snp` function alongside the `--variants 1 --output-vcf 1` options to return a .vcf file containing the positions of all SNPs (including non T>C SNPs).

The stranded_conversions_per_pair_no_snp.py script was then provided with the name-sorted featureCounts-assigned .bam file for each sample, alongside the SNP .vcf file generated from the no4sU control. For each base in each read, the script checks whether there is a T>C or A>G conversion, depending on whether the read is a forward or reverse read, and whether the transcript is sense or antisense stranded. At positions where a conversion is observed, the script then checks whether there is a SNP at this location, and if so, whether the SNP matches the conversion observed. Where this is the case, the base is skipped. Where a T>C or A>G conversion is observed and it is not present within the SNP .vcf file, then the read pair is assigned +1 to the `conversions` variable. The total number of conversions per read pair was then processed in R to determine how many read pairs have 0, 1+, 2+, or 3+ conversions for each sample.

2.6.2.3 SLAMseq in a cardiomyocyte differentiation time course

Day 0, 2 and 15 cells were prepared in the same manner as when incorporation efficiency was tested; however, more biological replicates and controls were considered. For each stage of differentiation 13 wells were required: one no-4sU control, four samples for 0 hour chase, four samples for 3 hour chase, and four samples for 12 hour chase. Both 4sU pulse and uridine chase were conducted in the dark.

Cells were cultured in mTeSR Plus in the presence of 4sU for 12 hours in the dark with media changes every 3 hours. For the no-4sU controls, mTeSR Plus was still changed every 3 hours. Following 4sU pulse, cells were washed 2X with PBS and media was changed to fresh mTeSR Plus containing non-thiol-containing uridine at 100X the concentration of 4sU (i.e. where 65µM 4sU was used for the pulse, 6.5mM uridine was used for the uridine chase). This media change was considered time-point 0hr for the uridine chase. Cells were incubated at 37°C and 5% CO_2 until they were harvested at the respective time-points. Non-thiol-containing uridine-supplemented media was not required to be changed every 3 hours. RNA extraction and IAA treatment was conducted as previously described in 2.6.2.2. Samples were sent to Novogene (Cambridge, UK) for library preparation (Poly(A) enrichment and cDNA synthesis) and sequencing (Illumina PE150) at a high sequencing depth (50M - 80M read pairs).

2.6.2.4 Data analysis

Raw sequencing reads were mapped to hg38 using STAR (Section 2.1.2). Mapped reads in .bam format were processed as previously described in 2.6.2.2. However, instead of .bam files being passed through the stranded_conversions_per_pair_no_snp.py script, they were passed through either the 3UI_spliced_counts_and_info.py script (for analysis of 3'UTR spliced transcript half-lives) or to the gene_level_counts_and_info.py script (for analysis of gene-level half-lives). The corresponding SNP .vcf file for that day of differentiation was also supplied to the script. Whether a read pair contained T>C or A>G conversions was determined in the same manner as described in 2.6.2.2. For analysis of 3'UTR spliced transcript half-lives, each read within each read pair was compared to a .bed file containing the coordinates of all 3UIs. The script then determines whether the read pair should be assigned to an event, and if so, whether the intron is spliced or retained. For gene-level analysis, the transcript assignment from featureCounts was compared to a transcript-to-gene lookup table. In addition to the read assignment (3'UTR spliced event or gene), and number of conversions within the read, additional metadata

relating to the total coverage by the read pair (to account for cases where the sequenced fragment is less than 300bp), and the total "convertible sequence" (the number of Ts for T>C conversions and the number of As for A>G conversions), was collected for each read pair. The output of these scripts is a .tsv file where each line is a read pair followed by information on its assignment, conversions, coverage, and convertible sequence.

For each sample (except no4sU controls), all assigned read pairs for each event were then summarised in R to determine the percentage of read pairs that were converted (had at least one T>C or A>G conversion) out of the total read pairs assigned to that event. This allowed us to observe the decrease in converted percentage (y-axis) across the uridine chase time course (time on the x-axis), at each differentiation time point. For each event, and at each differentiation time point, the data points were normalized to the 0 hour chase time point, and then fit to an exponential decay curve with the formula $y = exp(b \times time)$ where the `nls` function was used to estimate `b` (steepness of the slope). Half-life could subsequently be calculated as follows: $HL = \frac{log(2)}{-b}$. Half-lives were filtered out where they were >24 hours or <30 minutes.

To determine whether there were significantly different fits between spliced and retained isoforms, or significant differences over time, an ANOVA test was conducted with two models. The first model (referred to as the shared parameters model) fits all the data into a single grand fit, and therefore only has one `b` coefficient. The second model (referred to as the combined parameters model) fits all the data into a single model but has multiple `b` coefficients (two for spliced vs retained, three for day 0 vs day 2 vs day 16). Where the multiple `b` coefficients in the combined parameters model are very similar, they are unlikely to be significantly different from the single `b` coefficient in the shared parameters model, and the P-value from the ANOVA test will reflect this. Where they are different, the combined parameters model will fit the data better than the shared parameters model, and as such the P-value from the ANOVA test will reflect this. Where we tested for significance in the interaction over time (i.e. whether the spliced/retained ratio changes

across differentiation) the shared parameters model has three `b` values (day 0, day 2, day 16) and one `I` value (interaction; which represents the average difference between the spliced and retained isoforms across all three differentiation time points). The combined parameters model has three `b` values (day 0, day 2, day 16) and three `I` values (day 0, day 2, day 16). Schematic representations of these models are shown alongside the relevant results in Chapters 5 and 6. All P-values generated in this manner were adjusted for multiple hypothesis testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Adjusted P-values < 0.05 were considered statistically significant.

2.7 Luciferase assays

2.7.1 Creating Luciferase Plasmids

2.7.1.1 Using a pre-existing Luciferase plasmid

Full-length and ΔIntron HRAS Luciferase plasmids were originally generated by Cristina Alexandru in pcDNA4/TO/myc-His (Invitrogen #K103002). This plasmid was originally used as the backbone for our FL-CTNNB1, ΔI-CTNNB1 and 3'SS(S)-CTNNB1 Luciferase plasmids. PCR primers were designed to amplify the plasmid backbone, including the expression cassette and Luciferase2 gene. These primers contained overhangs matching the CTNNB1 3'UTR. Following PCR, the product was exposed to a DpnI digestion for 1 hour (as previously described in Section 2.3) to digest template plasmid. PCR primers were also designed to amplify the whole CTNNB1 3'UTR from gDNA so that splice sites and intronic sequence was included. These primers had overhangs with the plasmid backbone. Products of both PCR reaction were run on agarose gels to verify correct product sizes, and retrieved by gel extraction (as previously described in Section 2.3). A Gibson assembly reaction was then conducted (as previously described in Section 2.3) with a 1:2 DNA molar ratio of vector:insert. Following transformation into DH5-α E. coli cells, recovery, plating, inoculation, and pDNA extraction (all previously described in

Section 2.3), incorporation of the CTNNB1 3'UTR into the plasmid was verified by PCR and gel electrophoresis, and Sanger sequencing (described in Section 2.3). To generate intronless- and splice-site- mutants the Q5 Site-Direction Mutagenesis Kit (New England Biolabs #E0554S) was used. Successful site directed mutagenesis was assessed via PCR and gel electrophoresis (for intronless mutant constructs) and Sanger sequencing (for all mutant constructs).

2.7.1.2 Using pCI-neo

Due to the lack of introns within the expression cassette besides those found within the full-length 3'UTR constructs, it is possible that differences reflect intron-containing vs intronless expression. If this were the case then such results would be less representative of the endogenous context given that introns would usually be found within the protein coding sequence of the genes of interest. As such, a second set of plasmids were created utilizing the pCI-neo plasmid (Promega #E1841), which contains a chimeric intron (from β-globin and IgG) downstream of the cytomegalovirus immediate-early enhancer/promoter.

PCR primers were designed to amplify the Luciferase2 gene and the full-length 3'UTR from the plasmids created in Section 2.7.1.1. These primers were designed such that the forward primer had an overhang matching the sequence immediately upstream of an EcoRI digestion site in pCI-neo, and the reverse primer had an overhang matching the sequence immediately downstream of an XbaI digestion site in pCI-neo. Following PCR, a DpnI digest was conducted to remove the template plasmid from Section 2.7.1.1, followed by gel electrophoresis and gel extraction. The pCI-neo plasmid was digested with both EcoRI and XbaI. A Gibson assembly reaction was then conducted (as previously described in Section 2.3) with a 1:2 DNA molar ratio of vector:insert. Following transformation into DH5-α E. coli cells, recovery, plating, inoculation, and pDNA extraction (all previously described in Section 2.3), transfer of the Luciferase2 gene and 3'UTR of interest into the pCI-neo was verified by PCR and gel electrophoresis, and

Sanger sequencing (described in Section 2.3). Site directed mutagenesis was conducted and verified as described in Section 2.7.1.1.

2.7.2 Obtaining relative luminescence measurements

Luciferase assays were conducted using the Dual-Luciferase® Reporter Assay System Kit (Promega #E1910). Cells were transfected in 24 well plates (as previously described) with 500ng of Firefly Luciferase-3'UTR constructs and 5ng of hRLuc (Renilla Luciferase) normalizer. 72 hours post transfection, cells were lysed in 100µl Passive Lysis Buffer for 20 minutes at room temperature on a shaker. Luminescence readings were obtained using a Berthold Sirius Single Tube Luminometer (v3.1), with a 1 second delay and a measurement time of 5 seconds. Assays were carried out manually. 25µl LAR II was added to a 5ml polystyrene tube before 5µl sample was added. The tube was swirled for precisely 10 seconds before being placed in the luminometer to measure luminescence from the firefly lucifease. Following measurement, the tube was removed and 25µl Stop & Glo® Reagent was added. The tube was again swirled for precisely 10 seconds before being placed in the luminometer to measure luminescence from the Renilla luciferase. For each sample relative luminescence was calculated as the observed value from the firefly luciferase (3'UTR constructs) divided by that observed from the Renilla luciferase (normalizer).

3. Characterizing 3'UTR splicing events

Introns within 3'UTRs (3UIs) are generally considered to be a signal for transcript degradation via the nonsense-mediated RNA decay (NMD) pathway (Bicknell et al., 2012); therefore, transcripts harboring 3UIs could be considered transcriptional noise. Prior to starting this project, Ian Sudbery and Chirstina Alexandru developed a bioinformatic pipeline (pipeline_utrons) to detect 3'UTR splicing events, and found that not only were 3UIs detectable in highly expressed transcripts in cancer cells, but that this phenomena was widespread throughout the transcriptome. However, it was unclear whether this was a feature unique to cancer. We hypothesised that 3'UTR splicing could also play an important role in non-cancer biology, and wasn't simply representing transcriptional noise. To interrogate this in a non-cancer setting, this project set out to study 3'UTR splicing events in human pluripotent stem cells (hPSCs). As well as representing a non-cancer model, hPSCs can be used to model developmental processes such as organogenesis, allowing us to study differential splicing across cellular differentiation (which will be addressed in Chapter 4).

This chapter aims to determine the extent of 3'UTR splicing through the hPSC transcriptome, and characterize the transcripts harboring these events (including expression levels, nuclear export, and NMD sensitivity), as well as the splice site and intronic composition (including conservation, RBP/miRNA enrichment, and effect on expression upon mutation). This was achieved by using large RNAseq cohorts such as The Human Induced Pluripotent Stem Cell Initiative (HipSci), in combination with H9 hESCs, with which RNAi and Luciferase assays were conducted. To facilitate a comparison between cancer and hPSCs, we also utilized RNAseq data from The Cancer Genome Atlas (TCGA), and HCT116 colorectal carcinoma cells as a wet-lab model.

3.1 3'UTR splicing is widespread

To determine the extent of 3'UTR splicing throughout the transcriptome, and between individuals in a population, our 3UI detection pipeline (pipeline_utrons) was deployed on two large RNA-seq cohorts to detect, classify, collate, and quantify 3'UTR splicing events. The results in this chapter predominantly focus on the characterisation of 3'UTR splicing in hPSCs, using RNA-seq samples from the HipSci cohort. Our 3UI detection pipeline had previously been run on RNA-seq samples from the TCGA cohort by Ian Sudbery and Cristina Alexandru-Crivac. As such, characterization of both transcript assemblies was conducted in order to facilitate a comparison between 3'UTR splicing in the stem cell and cancer settings.

3.1.1 Detection of 3'UTR splicing events

As part of pipeline_utrons, all novel transcripts from 382 RNAseq samples from HipSci were compiled into a master HipSci transcriptome assembly. A transcriptome assembly built from TCGA samples had already been compiled (as discussed in 3.1) using RNA-seq data from 7,897 primary tissue samples of 16 solid tumour types. In both instances, novel transcripts were assembled on a sample-by-sample basis and then merged into a master assembly, facilitating parallelization. An alternative approach could have seen all reads merged into a single sample followed by novel transcript assembly. This approach would be more sensitive in its ability to detect lowly expressed junctions; however, due to the high computational requirement for a single task, the latter approach was not conducted.

Following assembly, each transcriptome was passed to a script to detect 3'UTR splicing events and classify them as "novel", "nonPTC" or "pPTC" (see 2.1.3.2). As shown in Figure 3.1A, 3UIs are classed as "nonPTC" where they do not share a splice donor or splice acceptor with introns that are found in the CDS of other transcripts, i.e. they are only ever seen in 3'UTRs. 3UIs are classed as "pPTC" where they overlap CDS introns. pPTC 3UIs are found in transcripts that have a 3'UTR annotation that starts early, potentially due to

presence of a premature termination codon (hence the term pPTC), and as such pPTC 3UIs are never novel. 3UIs are classed as "novel" where they are found in the 3'UTR, do not overlap the CDS of other transcripts, and are not found within the reference annotation.

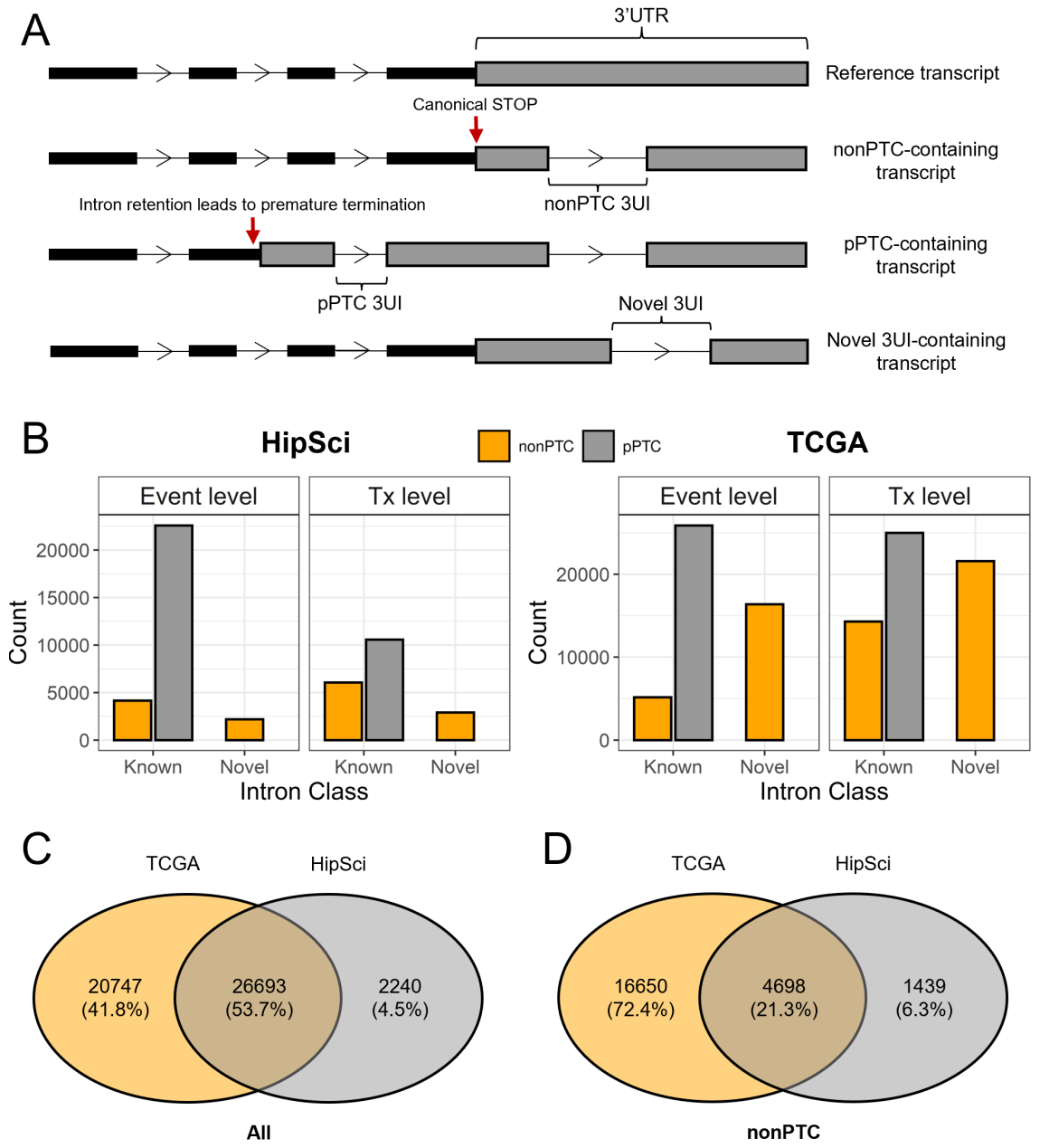

Figure 3.1: Detection of 3'UTR splicing events. A) Schematic representation of transcript classifications. nonPTC 3UIs are introns that are only ever found in the 3'UTR and have no overlap with a coding domain intron in the reference. pPTC 3UIs overlap coding domain introns and are present in the 3'UTR due to premature termination. Novel 3UIs are nonPTC 3UIs that are not seen in the reference. B) Breakdown of 3UI detection at both the individual event level, and transcript level (i.e. number of transcripts that contain 3UIs) in both the HipSci and TCGA assemblies. C+D) Comparison of 3UI event detection between TCGA and HipSci assemblies for: C) All 3UIs; D) nonPTC 3UIs.

From the HipSci assembly, we detected expression of 4,154 nonPTC 3UIs (in 1715 genes) and 22,596 pPTC 3UIs (in 3299 genes) which were already present within the reference annotation (known), in addition to 2,183 novel 3UIs (in 1218 genes; Figure 3.1B; left panels). The 4,154 nonPTC 3UIs were detected in 6,062 transcripts (an average of 1.46 transcripts per event), whilst the 22,586 pPTC 3UIs were detected in 10,579 transcripts (an average of 2.13 events per transcript). The reason behind the increased event to transcript ratio in pPTC 3UIs is that premature termination, and therefore early 3'UTR annotation, often leads to the multiple CDS introns now being annotated as within the 3'UTR. In some instances the majority of the CDS intron chain can be annotated within the 3'UTR if a premature termination codon occurs very early within the transcript. On the other hand, a decreased event to transcript ratio is observed for nonPTC 3UIs as the same 3UI is often detected in multiple transcripts. The majority (55.6%) of nonPTC 3UI-containing transcript only contain a single 3UI, whilst 44.4% contain 2+, 21.5% contain 3+, and 12.6% contain 4+ 3UIs.

In comparison, for known introns from the TCGA assembly, we detected expression of 1,009 more nonPTC 3UIs compared to HipSci, and 3,296 more pPTC 3UIs (Figure 3.1B; right panels). We detected 16,385 novel nonPTC 3UIs (vs 2,183 in HipSci), representing a 7.5-fold increase. When comparing individual events between HipSci and TCGA assemblies, we found that 53.7% of all 3UIs were detectable in both, whilst only 4.5% were HipSci-specific (Figure 3.1C). In terms of nonPTC 3UIs, those which are only ever found in 3'UTRs, and the main focus of this thesis, we found that 21.3% were found in both assemblies, whilst 6.3% are HipSci-specific, and 72.4% were TCGA-specific (Figure 3.1D).

Increased 3UI detection in TCGA could stem from heterogeneity between individuals, and could be explained by the size difference of the cohorts (n=382 for HipSci vs n=7897 for TCGA). Alternatively, it could stem from increased heterogeneity within the cancer setting, where aberrant splicing is known to occur commonly (Oltean and Bates, 2014). In

order to distinguish these explanations, we measured 3UI detection using an increasing number of samples from either the HipSci cohort or from the TCGA-COAD (colorectal carcinoma) subgroup via pipeline_3UI_saturation (github.com/jjriley1/pipeline_3UI_saturation). Following novel transcript assembly for each sample, we merged increasing numbers of individual assemblies, producing transcript assemblies built from 1 up to 300 samples for HipSci samples, or 1 up to 450 samples for TCGA-COAD samples. 3UI detection was saturated at approximately 100 individual samples using HipSci data (Figure 3.2A). Whilst for TCGA-COAD, we saw that at the same number of samples only 87% of total 3UIs were detected, and the level of detection did not appear to saturate even at the maximal sample size (Figure 3.2B). This lack of saturation in TCGA-COAD suggests that the difference in 3UI detection observed between the HipSci and TCGA assemblies is not due to sample size differences, but is due to additional patient-specific 3UI events being detected with the addition of each cancer sample.

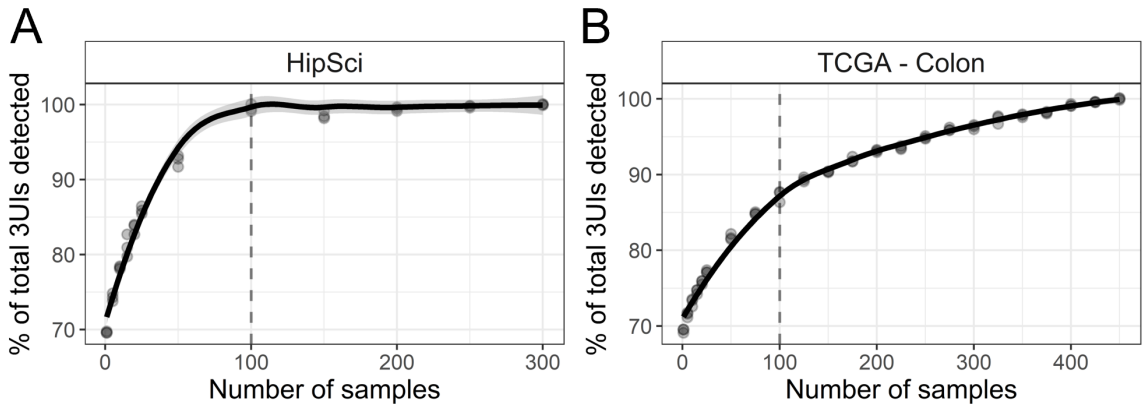

Figure 3.2: Saturation of 3'UTR intron detection. Saturation of 3UI detection by randomly sampling an increasing number of: A) hiPSC samples from HipSci; B) colon cancer samples from TCGA. Following sampling, transcript assemblies were merged and passed to our 3UI detection script. The percentage of total 3UIs detected is plotted for transcript assemblies built from an increasing number of samples. n=3 per sampling size.

3.1.2 Many 3'UTR spliced transcripts are highly expressed

Next, we determined how many of these transcripts are highly expressed within each sample, and across the cohort. We found that 90.1% of nonPTC 3UI-containing transcripts are expressed with a transcripts per million (TPM) value > 1 in at least one individual within the cohort (Figure 3.3A orange). A TPM > 1 means the transcript is expressed in the top 8.7% of all transcripts within the HipSci assembly. Meanwhile 31.5% of nonPTC 3UI-containing transcripts are expressed with a TPM > 5 (top 3.7% of transcripts) in at least one individual within the cohort (Figure 3.3A grey). Turning towards how commonly these transcripts are expressed throughout the cohort, we found that 2,069 nonPTC 3UI-containing transcripts are expressed > 1 TPM in $> 10\%$ of the cohort, 1,505 in $> 25\%$, 1,124 in $> 50\%$, and 131 in all HipSci samples (Figure 3.3A).

Whilst a given transcript having a TPM value > 1 indicates a relatively high level of expression relative to all transcripts, it does not necessarily indicate a significant contribution towards its total gene output. To account for this, we calculated a "transcript to gene ratio" (Tx:G) and used this to determine how many transcripts contribute at least 25% (Tx:G>0.25) or 50% (Tx:G>0.5) of the output of the gene they are from (Figure 3.3B). We found that 3,553 of the 6,062 nonPTC 3UI-containing transcripts had a Tx:G > 0.25 in at least one individual within the HipSci cohort, whilst 1,283 had a Tx:G > 0.25 in 10% of samples, and 146 had a Tx:G > 0.25 in all HipSci samples (Figure 3.3B).

By combining these metrics, we filtered our nonPTC 3UI-containing transcripts to those that had > 1 TPM and > 0.25 Tx:G in at least 10% of HipSci samples, and termed these "broadly expressed". We find 2,490 nonPTC 3UI transcripts with > 1 TPM and > 0.25 Tx:G in at least one sample within the HipSci cohort, whilst 766 are broadly expressed (Figure 3.3C; intersection of orange curved line and dashed grey vertical line). Of these broadly expressed 3UI transcripts, 567 contain 3UIs that were previously annotated, whilst 199 3UIs are novel (Figure 3.3D).

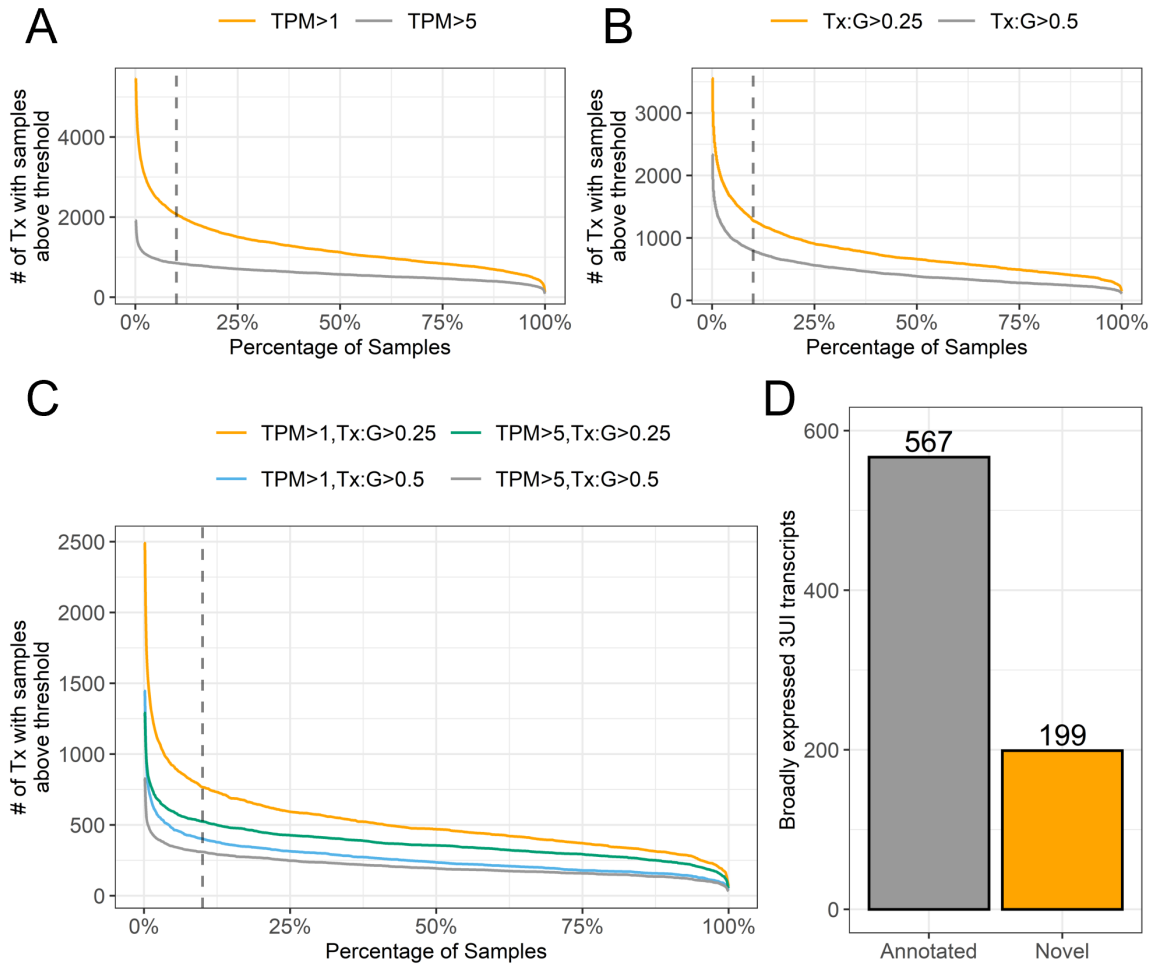

Figure 3.3: Expression of 3'UTR spliced transcripts in HipSci. A) The number of nonPTC 3UI-containing transcripts that meet the criteria of either TPM>1 or TPM>5 in a given percentage of samples from HipSci. B) The number of nonPTC 3UI-containing transcripts that meets the criteria of either Transcript/Gene fraction (Tx:G) > 0.25 or Tx:G > 0.5 in a given percentage of samples from HipSci. C) The number of nonPTC 3UI-containing transcripts that meet increasingly stringent expression criteria in a given percentage of samples from HipSci. D) Number of annotated or novel transcripts that are broadly expressed (TPM>1; Tx:G>0.25 in >10% of samples) across the HipSci cohort.

To gain insight into the potential functions of genes with transcripts that contain broadly expressed 3UIs in HipSci samples, we conducted gene ontology analysis against the GO:BP (biological processes) and GO:KEGG (KEGG pathways) categories. We observed a significant enrichment of broadly expressed 3UIs in genes related to "mRNA 3′ end

processing", "mRNA catabolic process", "mRNA splicing", "protein-containing complex localization", and "regulation of translation" (Figure 3.4A). Significant enrichment of broadly expressed 3UIs in genes related to neurotrophin, chemokine, and Wnt signalling pathways was also observed (Figure 3.4B). In addition to the top GO:BP terms shown in Figure 3.4A, we also observed significant enrichment of "negative regulation of cell development" (Padj=0.01; Hits=6.5%), "regulation of Wnt signaling pathway" (Padj=0.01; Hits=6.2%), "regulation of mRNA stability" (Padj=0.012; Hits=7.5%), "stem cell differentiation" (Padj=0.012; Hits=6.9%), and "stem cell population maintenance" (Padj=0.02; Hits=7.4%).

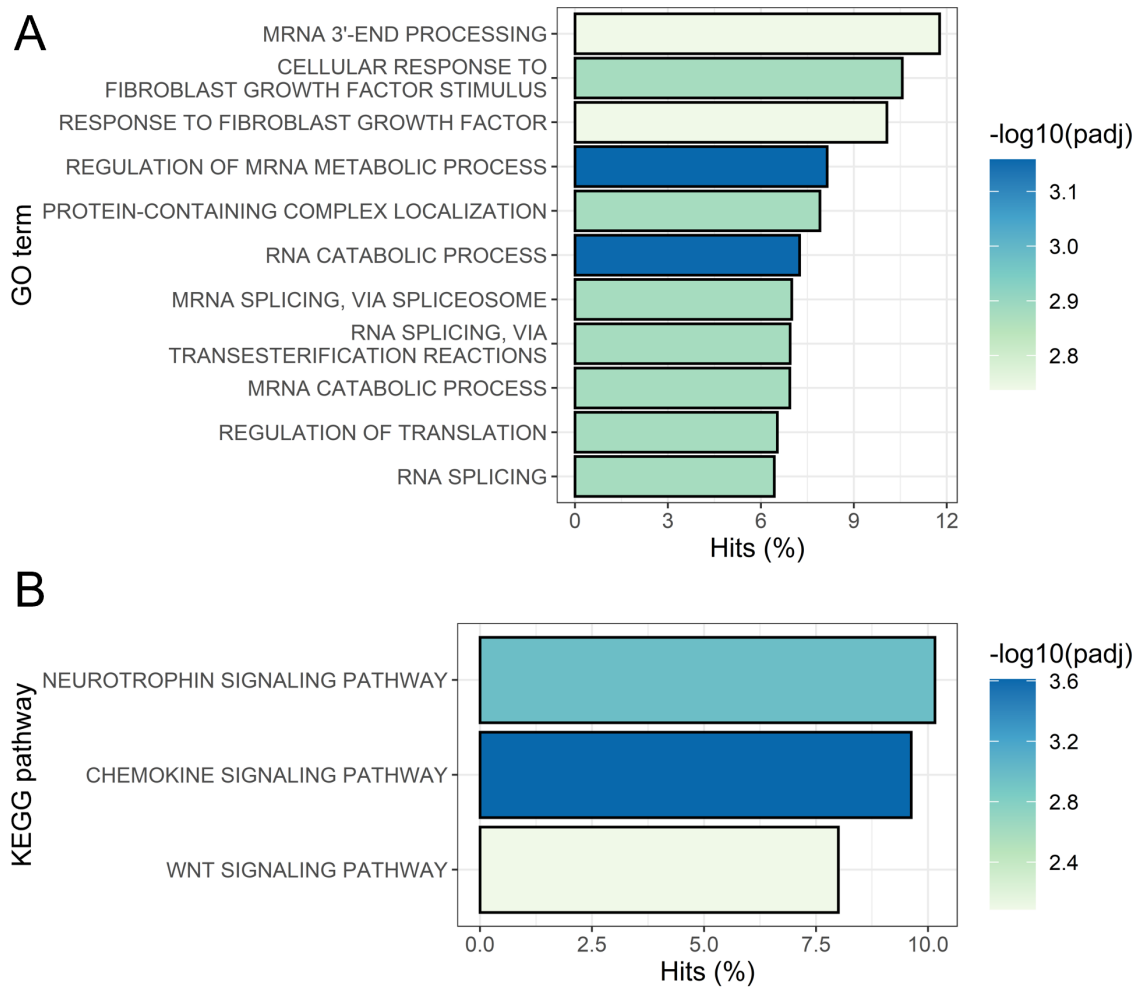

Figure 3.4: Gene ontology analysis of highly expressed 3UI transcripts in HipSci. A) Enrichment of biological processes gene ontology terms (GO:BP) for genes that have broadly expressed 3UI-containing transcripts in HipSci samples. B) Enrichment of KEGG pathways for genes that have broadly expressed 3UI-containing transcripts in HipSci samples.

3.2 Validating existence/expression of 3'UTR splicing events

3.2.1 3'UTR splicing is evolutionarily conserved

To determine whether nonPTC 3UIs resembled normal introns, we first interrogated their splice site composition. We found that 91.9% of nonPTC 3UIs utilised canonical GT-AG splice sites. The most common non-canonical splice site used was GC-AG in 5.8% of

cases, whilst the second most common was AT-AC in 0.3% of cases. We also found that the majority of these non-canonical splice site utilising events were novel (73.1%). In comparison, when interrogating the entire 3UI set (nonPTC and pPTC/coding) we found that 96.3% of introns used the canonical GT-AG splice site, 3.7% used non-canonical sites, of which 2.8% were GC-AG sites.

Given that many GT and AG dinucleotides do not act as splice sites, we next examined the 3'UTR splice sites in the context of their genomic position, and interrogated their conservation scores. First, by looking at all 3UI events, we found that the 5'SS and 3'SS are more highly conserved than their surrounding intronic and non-intronic sequence (Figure 3.5; Wilcoxon P<0.001). Next, we found that the intronic sequence is substantially less well conserved than the splice sites and surrounding non-intronic sequence (Wilcoxon P<0.001). We observed the same trends when focusing solely on nonPTC 3UIs; however, the phyloP scores are lower across the track (Figure 3.5). This indicates that introns within the coding sequence are more highly conserved than those found within the 3'UTR. Finally, novel nonPTC 3UIs do not appear to have conserved 5'SS, although a slight increase in conservation is observed in the terminal 3'SS base (Figure 3.5), reinforcing our previous findings that many of these novel events are found in a sample-specific manner.

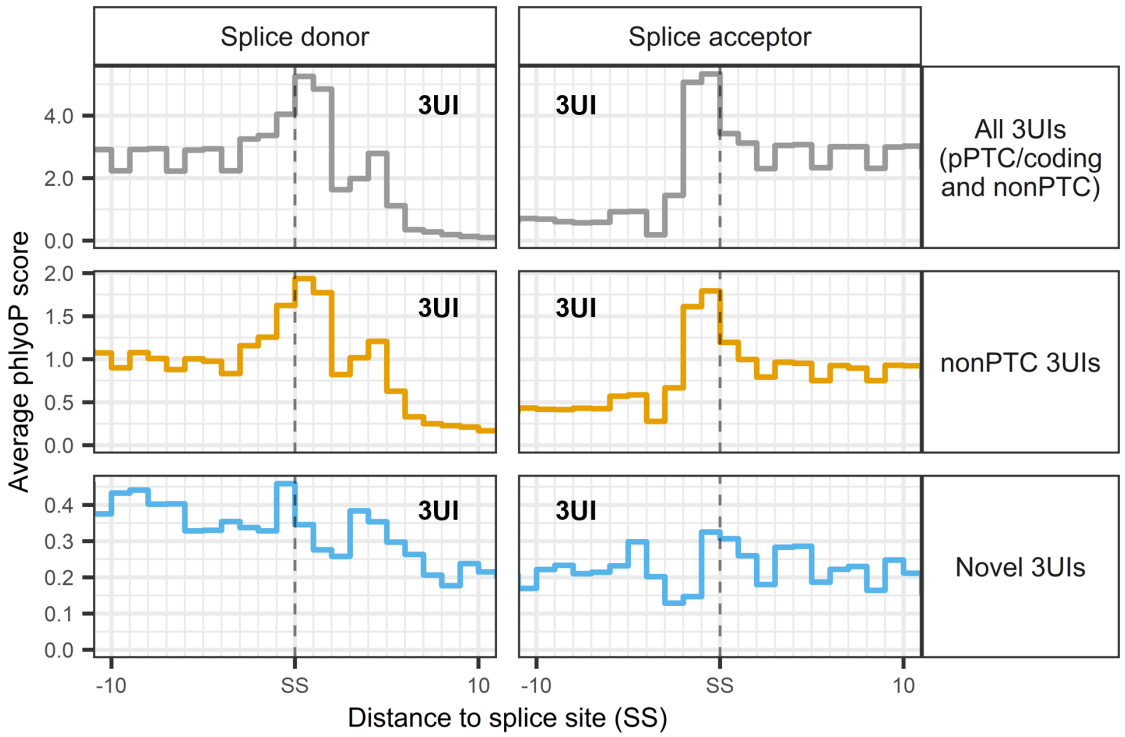

Figure 3.5: Conservation of 3'UTR splice sites between 3UI classifications. Average PhyloP score at each position relative to the splice site (SS) is shown, broken down by 3UI classification (rows). All 3UIs facet includes both nonPTC 3UIs (3'UTR only) and pPTC 3UIs (overlap CDS introns).

Given our previous finding that many nonPTC 3UIs within the TCGA cohort are sample-specific, and with the knowledge that splicing is often dysregulated in cancers, we compared conservation scores between all nonPTC 3UIs in the TCGA cohort versus the HipSci cohort. We found that conservation within the HipSci cohort was substantially higher than within the TCGA cohort when looking at all nonPTC events (Figure 3.6A; Wilcoxon P=0.001). However, where we limit this comparison to events which are broadly expressed in HipSci versus those which are broadly expressed in any cancer type from the TCGA cohort, we find that this difference is smaller, however the HipSci nonPTC 3UIs are still slightly more conserved (Figure 3.6B; Wilcoxon P=0.04).

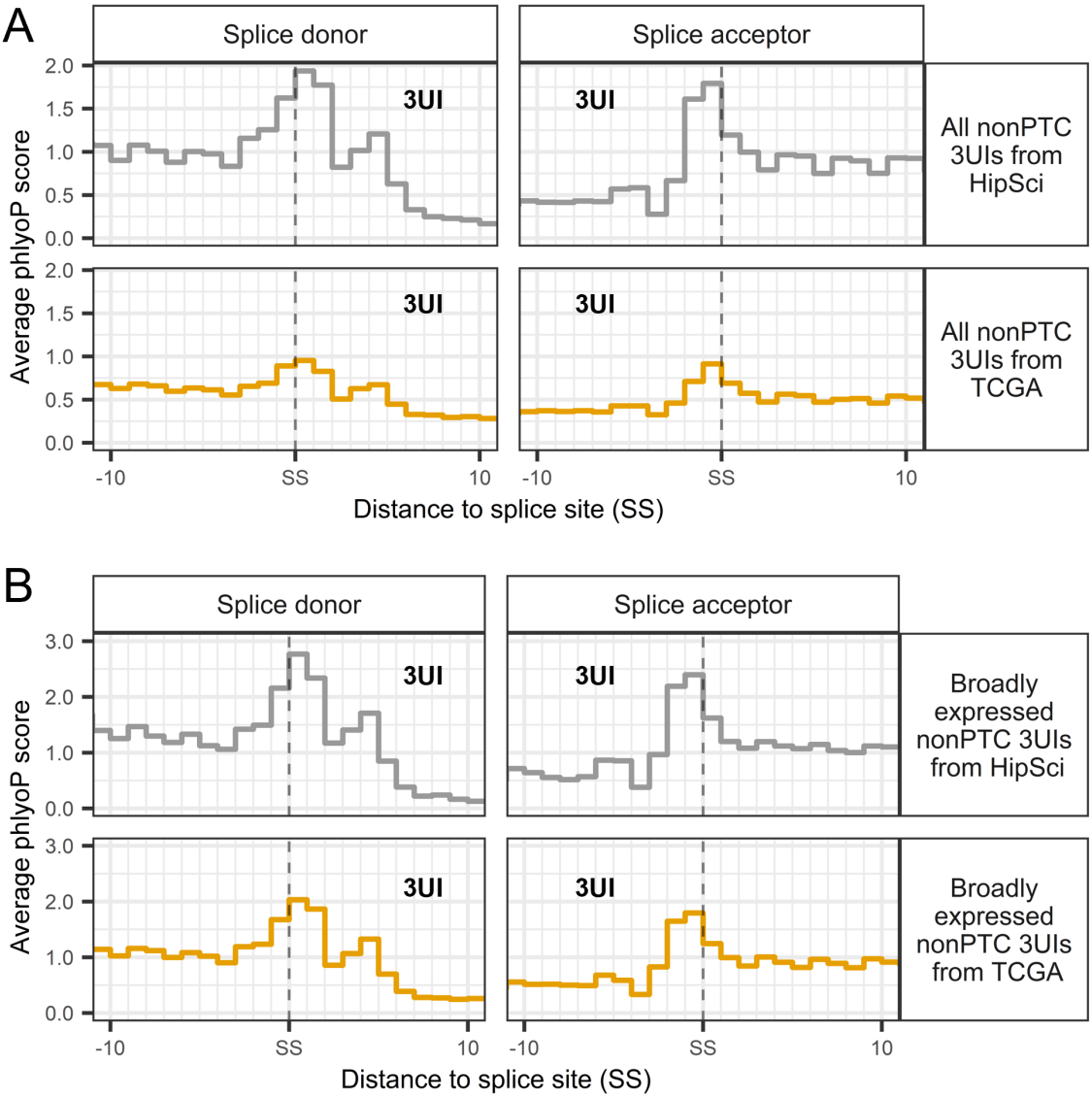

Figure 3.6: Conservation of 3'UTR splice sites between HipSci and TCGA. Average PhyloP score at each position relative to the splice site (SS) is compared between HipSci and TCGA events. A) All nonPTC 3UI events. B) Broadly expressed 3UI events (TPM>1 & Transcript/Gene ratio > 0.25 in >10% of samples). Transcripts are classed as broadly expressed from TCGA where they are broadly expressed in at least 1 tissue type.

Interestingly, in all cases it appears that the base immediately upstream of the 5'SS is more highly conserved than the adjacent upstream sequence. Additionally, when looking at the "all 3UIs" set (which includes events that overlap coding introns), we notice a

trinucleotide pattern where the third nucleotide appears to be less conserved than positions 1 and 2. This is consistent with wobble base pairing. Surprisingly, we also observe a smaller, yet noticeable, degree of this within the nonPTC 3UI set (which are only ever found in the 3'UTR). Whilst this could be explained by an overrepresentation of nonPTC 3UIs "in-frame" within the preserved open reading frame downstream of the stop codon, analysis of the position of the 3UIs within the 3'UTR revealed an even distribution of introns in the 0, +1 and +2 reading frames.

3.2.2 Our detected 3UIs are a result of splicing

In order to determine whether the 3'UTR splicing events we have presented were expressed in our cell line within the lab (H9 hESCs), as well as to rule out the possibility that these events were detected as a result of reads mismapping to regions of the genome that match our transcripts sequence but lack the intron, we set out to amplify these events from both cDNA and gDNA for several nonPTC 3UI examples. By designing primers that flanked the 3'UTR intron (Figure 3.7A; green) we were able to amplify both retained and spliced isoforms only in the cDNA, whilst amplification from gDNA produced only the retained isoforms (Figure 3.7B). Additionally, we also designed primer pairs where primers either spanned the 3UI splice site (Figure 3.7A; orange), or were entirely contained within the 3UI (Figure 3.7A; blue), and found that amplification of the 3UI-spliced isoforms was only possible from cDNA and not gDNA (Figure 3.7C). This indicates that these events are indeed the result of splicing.

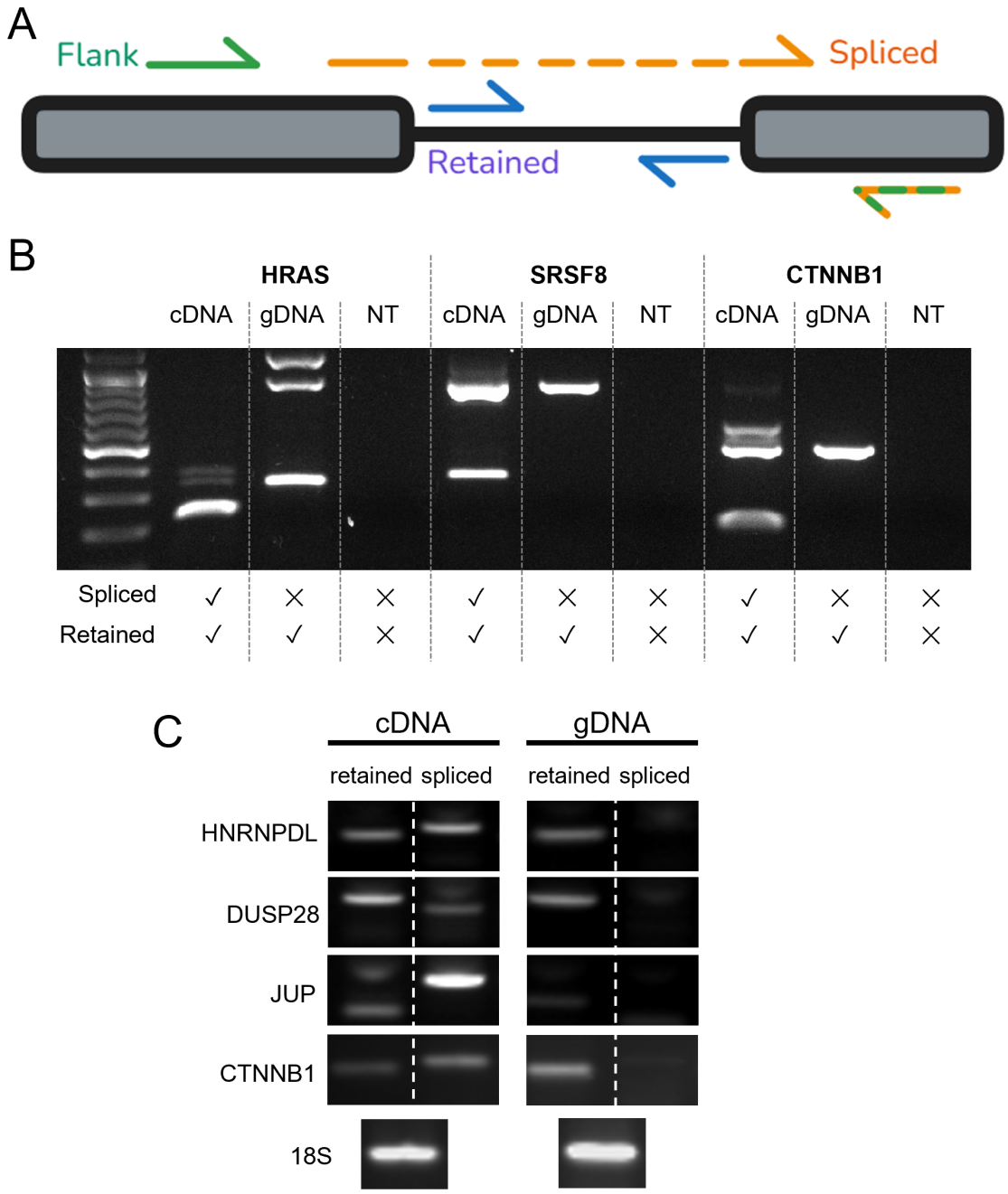

Figure 3.7: Validation of example 3UI events via RT-PCR. A) Primer design: flanking primers either side of the intron can amplify both spliced and retained isoforms; splice-specific primers span the splice junction; retained-specific primers are within the 3UI. B) RT-PCR conducted with flanking primers using either cDNA or gDNA as template, or a no template (NT) control. C) RT-PCR conducted with spliced-specific or retained-specific primers using either cDNA and gDNA as template.

3.2.3 3'UTR spliced transcripts are effectively exported from the nucleus

By using existing RNA sequencing data produced by the Wilson lab, where nuclear-cytosolic fractionation had been conducted in HCT116 colorectal carcinoma cells, we were able to analyse the nuclear export of 3UI-containing transcripts. We quantified the RNAseq data using our TCGA transcriptome assembly and performed differential transcript expression analysis between cytosolic and nuclear fractions. We then compared nonPTC 3UI-containing transcripts with all protein coding transcripts via a Kolmogorov-Smirnov test. We found that 3UI-containing transcripts are successfully exported from the nucleus, and are exported slightly better than protein coding transcripts in HCT116 cells (Figure 3.8).

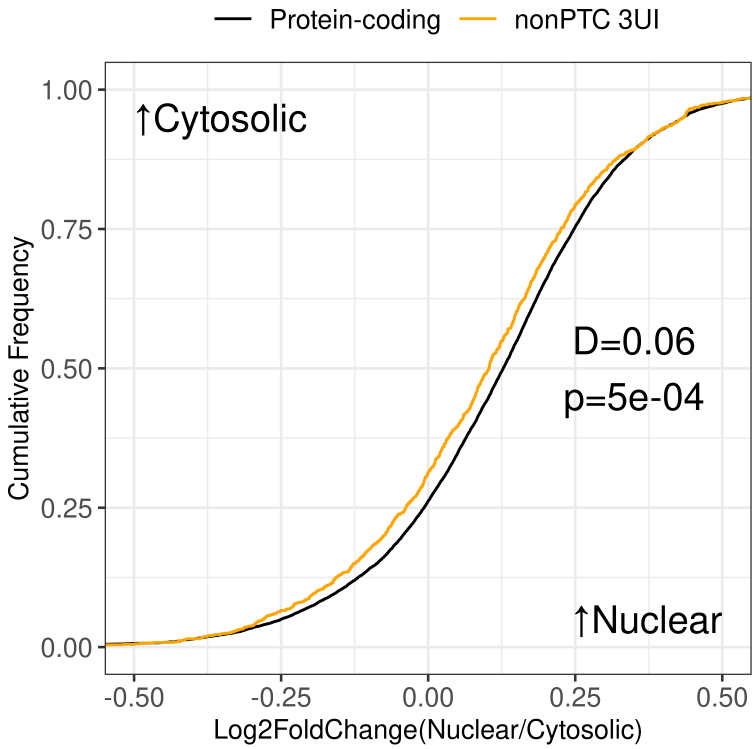

Figure 3.8: 3'UTR spliced transcripts are exported from the nucleus. Empirical Cumulative Distribution Function (ECDF) plot comparing nuclear/cytosolic ratios of nonPTC 3UI-containing transcripts (orange line) vs all protein-coding transcripts (black line). Statistical significance was tested via the Kolmogorov-Smirnov test. D value = distance between the two populations. P-value = probability that the null hypothesis (both populations are the same) is true.

3.3 Impact of 3'UTR splicing on mRNP composition

Previous studies have shown that the RNA binding protein composition within the mRNP can be regulated by alternative polyadenylation, where long and short 3'UTR isoforms have different mRNP compositions, and potentially different functions (Section 1.3). In order to determine the impact of 3'UTR splicing on mRNP composition, and gain a potential insight into its regulatory effects, we conducted multiple analyses involving both predictive approaches and the use of existing CLIP-seq and AGO-CLIP data from ENCORI (Li et al., 2014). First, we looked at the distribution of 3UI sizes between 3UI

classes and between the TCGA and HipSci assemblies, hinting at the extent of sequence modulation. The average size of annotated 3'UTR introns in the TCGA assembly is 1,128nt, and in the HipSci assembly is 1,237nt (Figure 3.9). Novel 3UIs tend to be smaller, at 101nt in the TCGA assembly, and 326nt in the HipSci assembly (Wilcoxon P<0.001 for both comparisons). This suggests that novel 3UIs would be less likely to modulate the inclusion/exclusion of RBPs and MREs based on the fact that they are smaller.

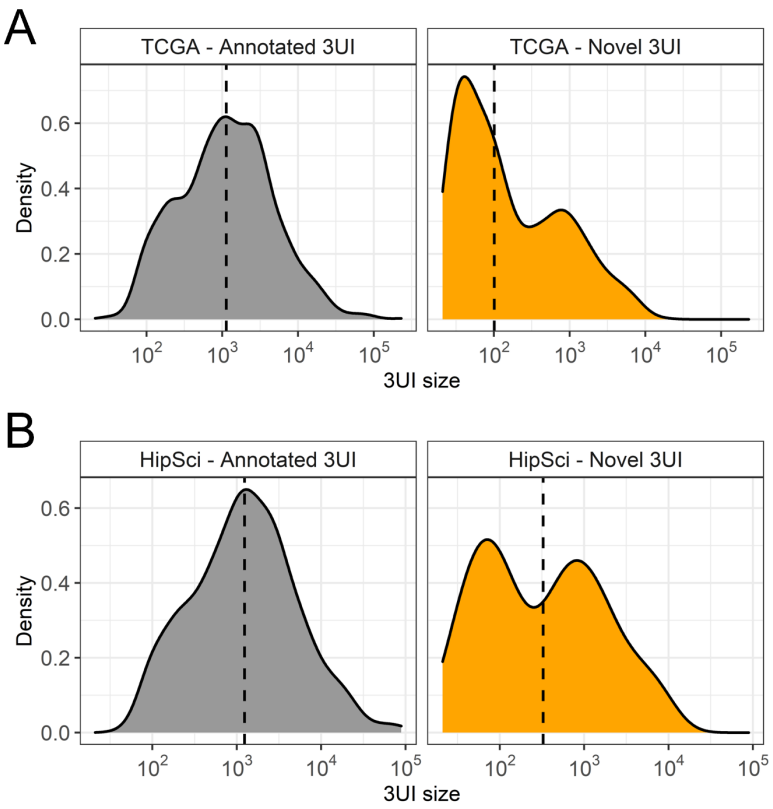

Figure 3.9: Distribution of 3UI sizes. For annotated (grey) or novel 3UIs (orange) found in transcripts that are broadly expressed in: A) any cancer type from TCGA; B) HipSci cohort. Dashed line indicates the median 3UI size for each facet.

Using existing CLIP-seq and AGO-CLIP data, we next looked for overlaps between our 3'UTR introns and the binding of various RBPs, or of Argonaute. Where Argonaute binding was observed it was matched to known miRNA seed sequences to predict their binding.

3.3.1 Analysis of CLIPseq data reveals RBP enrichment in 3UI subsets

First, by comparing our nonPTC 3UIs against the entire 3'UTR sequence as background, we observed significant enrichment (Padj < 0.05) of 40 RBPs (top 25 shown in Figure 3.10A). Many of these RBPs are known to be involved in RNA splicing and pre-mRNA processing, for example SF3B4, SF3A3, and PRPF8. This further supports our previous evidence that 3'UTR introns are characterized by similar features to coding domain introns. Multiple heterogeneous nuclear ribonucleoproteins (hnRNPs) are also enriched, including HNRNPA1, HNRNPUL1, HNRNPU, HNRNPK, and HNRNPA2B1. Interestingly, we observed a significant enrichment of m6A eraser FTO, and a significant under-representation of m6A readers YTHDF1 (18.7% under-represented; Padj=0.033) and YTHDF2 (38.0% under-represented; Padj=0.037) in nonPTC 3UIs compared to the total 3'UTR background, despite there being no significant difference in the number of DRACH motifs (Section 1.2.1.3) between the two sequence sets.

Next, we compared broadly expressed nonPTC 3UIs with all nonPTC 3UIs to identify the RBP binding sites that are commonly spliced out in highly and broadly expressed transcripts. We observed significant enrichment of 105 RBPs (top 25 shown in Figure 3.10B). Again, we observed significant enrichment of RBPs involved in RNA splicing including SF3A3, SF3B4, PRPF8, U2AF1 and U2AF2 (P<0.001), which could indicate that these introns have stronger splice sites than the "all nonPTC 3UIs" background. We also observed significant enrichment of many RBPs known to regulate RNA stability, including AU-rich binding proteins ELAV1, HNRNPA1 and TIA1 (P<0.001). Therefore the exclusion of these RBPs via splicing may contribute to the broad expression of these 3UI containing transcripts compared to the "all nonPTC 3UIs" background. Additionally, we observed significant enrichment of several Argonaute proteins including AGO1 (2.4-fold enrichment; Padj=0.001), AGO3 (2.1-fold enrichment; Padj=7.6e-5), and AGO4 (2.1-fold enrichment; Padj=0.01). Subsequently AGO-CLIP data was examined to determine which miRNAs may be binding to these regions.

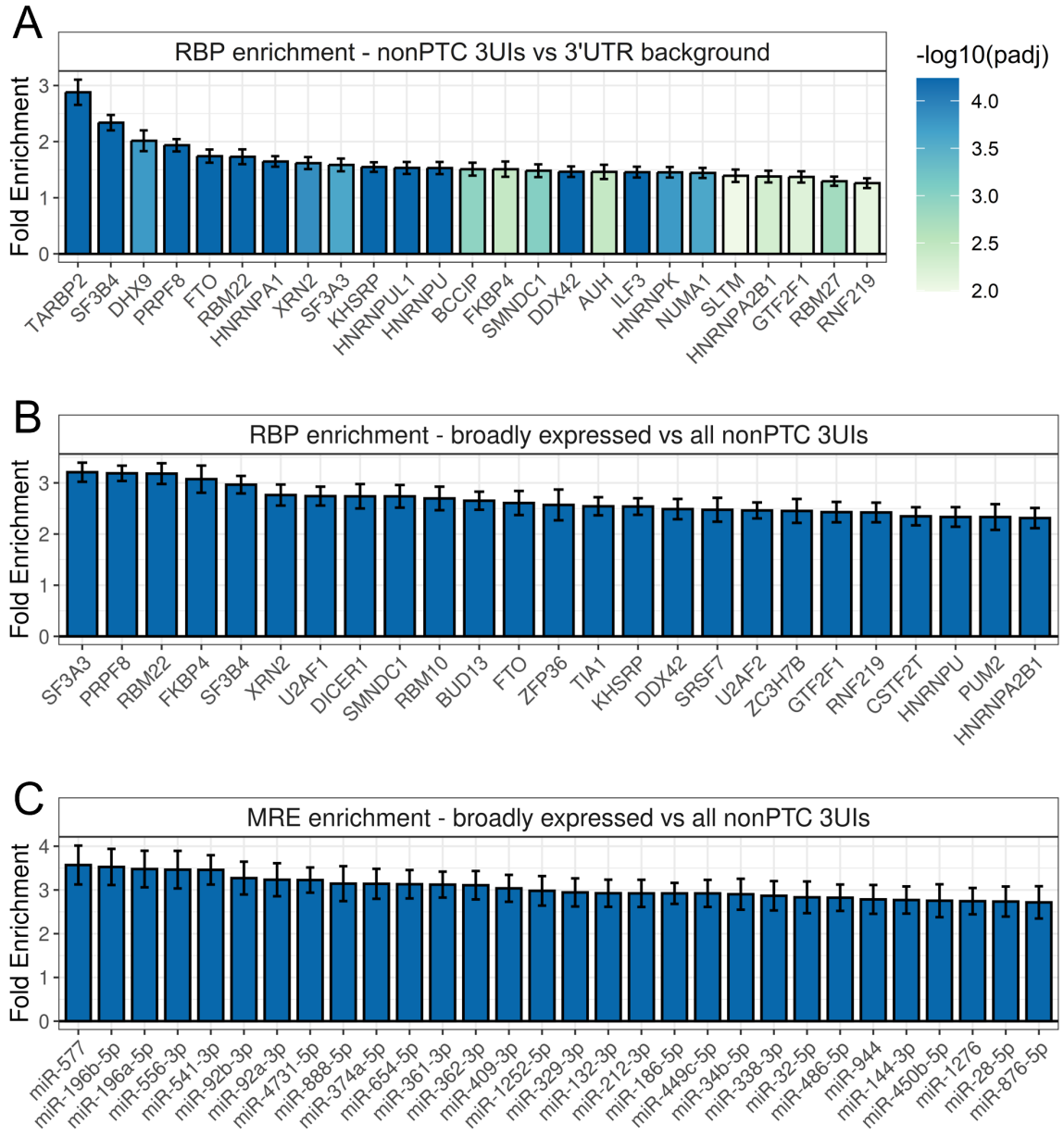

Figure 3.10: Enrichment of RBPs and MREs in 3'UTR introns. A) Enrichment of RBPs in the 3UI compared to the whole 3'UTR sequence as background. B) Enrichment of RBPs in broadly expressed nonPTC 3UI-containing transcripts compared to all nonPTC 3UI-containing transcripts. All adjusted P-values < 0.001. C) Enrichment of MREs in broadly expressed nonPTC 3UI-containing transcripts compared to all nonPTC 3UI-containing transcripts. All adjusted P-values < 0.001.

3.3.2 Analysis of AGO-CLIP data reveals miRNA exclusion in broadly expressed 3UI transcripts

Examining broadly expressed 3UIs vs all nonPTC 3UIs as background, we observed significant enrichment of 379 miRNAs (top 30 shown in Figure 3.10C). Given that miRNA binding leads to transcript degradation, the removal of these MREs via splicing would produce an overall positive effect on transcript expression. However, it is important to note that the relative level of expression of each miRNA, as well as how "diluted" it is amongst its mRNA targets, would dictate the overall repressive effect of individual miRNAs, and is likely highly cell-type and cell-state specific. Several of these enriched miRNAs have been linked to promoting hESC self-renewal (Zhang et al., 2015), including miR-302b (1.8-fold enrichment; Padj=0.015) and miR-367 (2.1-fold enrichment; Padj=0.008).

3.4 Interplay with NMD

The presence of introns in 3'UTRs is generally considered to be a signal to elicit transcript degradation by the nonsense mediated decay (NMD) pathway (Section 1.5). By knocking down UPF1, a core NMD component, we evaluated the impact of splicing 3'UTRs on transcript expression in both HCT116 colorectal carcinoma cells and H9 hESCs.

3.4.1 Predictive approaches

Given that the threshold for NMD had previously been established as 55nt downstream of the stop codon, we first interrogated our events to determine the distance between the 5'SS and the stop codon. In instances where a transcript had multiple introns downstream of the stop codon, only the distance of the terminal intron was considered, as this would be the NMD-sensitising event. We found that the majority of 3UIs were situated less than 55nt downstream of the stop codon, suggesting that they would not be NMD-sensitising. This was the case for both the TCGA set (Figure 3.11A) and the HipSci

set (Figure 3.11B). Additionally, this observation is exaggerated in transcripts which are broadly expressed. Interestingly, novel 3UIs tend to be further than 55nt from the stop codon in both assemblies, including those which are found in broadly expressed transcripts. Together with our previous finding in Figure 3.9 that novel 3UIs tend to be smaller, this could suggest that novel 3UIs are more likely to act through NMD as opposed to sequence (RBP/miRNA binding site) modulation.

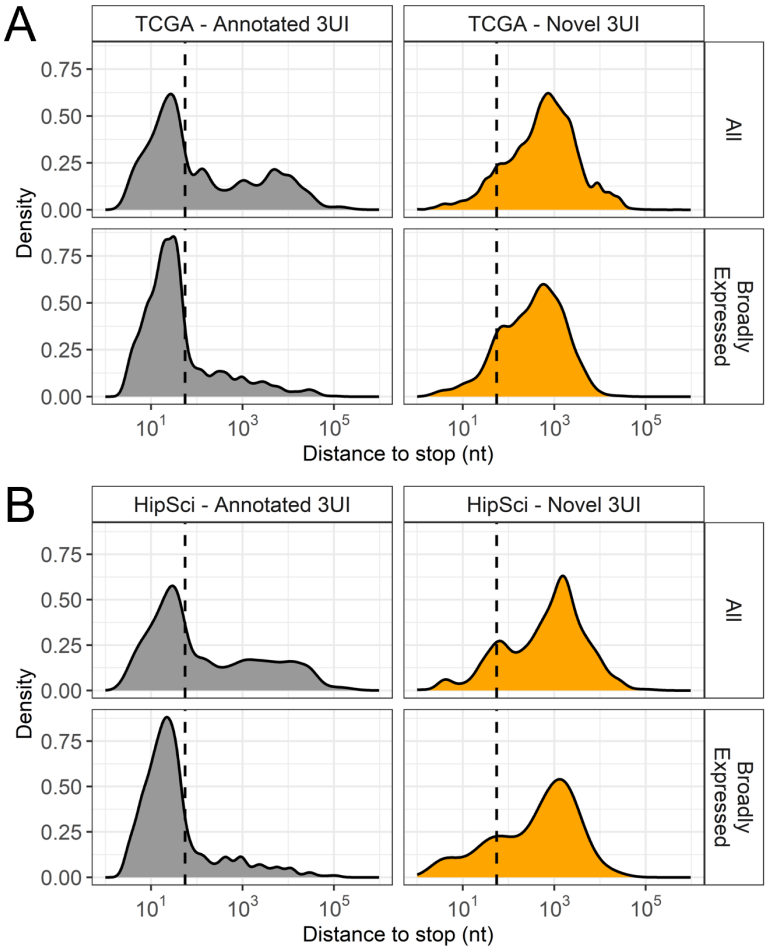

Figure 3.11: Distribution of distances between terminal 3UI and stop codon. For annotated or novel 3UIs that are found in transcripts that are broadly expressed (bottom rows) compared to total detected (top rows) in: A) any cancer type from TCGA; B) HipSci cohort. Dashed line indicates the 55nt marker to separate transcripts which would be predicted to be NMD sensitive (>55nt) or NMD insensitive (<55nt) based on the "55nt rule".

3.4.2 UPF1 knockdowns

To address these predictions in the cancer setting, UPF1 was knocked down in HCT116 colorectal carcinoma cells in an experiment designed and executed by Cristina Alexandru. Validation of knockdown efficiency and bioinformatic analysis was performed by the Jack Riley. In the stem cell setting, UPF1 was knocked down in H9 hESCs in an experiment designed, executed, and analysed by the Jack Riley.

3.4.2.1 In colorectal carcinoma cells

HCT116 cells were transfected with 30nM siRNA against UPF1 (with two different siRNAs) or against DsRed. Knockdown efficiency was determined by western blotting (Figure 3.12A) and RT-qPCR (Figure 3.12B). In HCT116 cells knockdown of UPF1 did not lead to increased cell death beyond that observed in the control knockdown condition. In both protein- and RNA-level analyses the siUPF1_2 duplex was the most effective at knocking down UPF1. However, for the RNA-sequencing experiment in HCT116 cells conducted by Cristina Alexandru, the siUPF1_1 duplex (83% knockdown at the protein level) was chosen. HCT116 cells were also treated with either DMSO, or NMDI14, a small molecule inhibitor of the UPF1-SMG7 interaction, to inhibit NMD. Cluster analysis, and interrogation of known NMD targets revealed a negligible impact of using NMDI14 in HCT116 cells (very few differentially expressed genes between NMDI14- and DMSO-treated samples in siDsRed transfected cells, and no clustering between DMSO and NMDi in Figure 3.12C "Inhib" annotation). Nevertheless the treatment with NMDI14 was taken into account in the design formula for differential transcript expression. To validate that UPF1 knockdown was inhibiting NMD, we randomly sampled 100 known NMD-sensitive transcripts and found that the majority of these were upregulated upon UPF1 knockdown (Figure 3.12C).

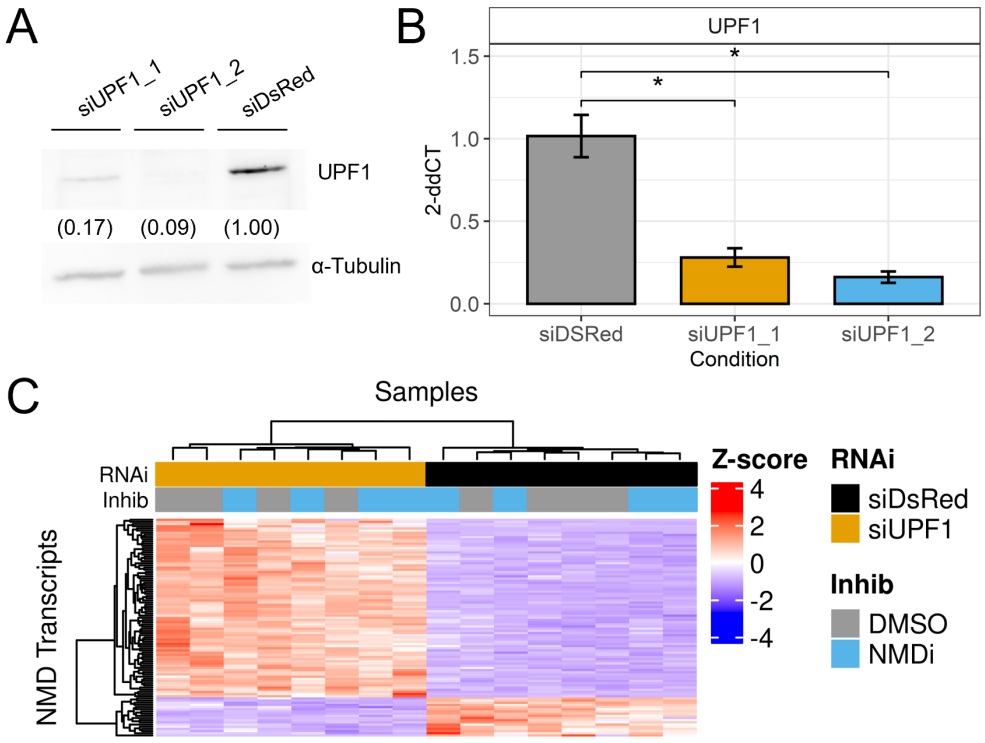

Figure 3.12: Validation of UPF1 knockdown in HCT116 cells. A) Western blot showing the expression of UPF1 protein following RNAi against UPF1 or DsRed (control) in HCT116 cells. α-Tubulin was used as a loading control. Signal quantification (in brackets under UPF1 bands) was normalized against the loading control and the siDsRed condition. B) qPCR showing UPF1 mRNA expression following RNAi against UPF1 or DsRed (control) in HCT116 cells. *P<0.05. C) Differential transcript expression analysis of 100 transcripts annotated as "nonsense_mediated_decay" in the reference shows that the majority are upregulated upon UPF1 knockdown. Heatmap annotations indicate whether cells were transfected with the control (siDsRed) or UPF1 siRNA, and whether they were treated with a small molecule NMD inhibitor (NMDi) or control (DMSO).

Next we performed differential transcript usage (DTU) analysis using DEXseq (Anders et al., 2012) to identify 3UI-containing transcripts which were significantly regulated by UPF1 knockdown. DTU analysis incorporates changes at both the transcript and gene level to identify instances where transcript expression changes disproportionately to gene expression. This allows us to focus on transcripts which are subjected to differential post-transcriptional regulation, as opposed to differential transcriptional output. From our analysis we identified 660 3UI-containing transcripts displaying significant DTU

upon UPF1 knockdown (Figure 3.13A). Whilst we expected to see a bias towards upregulation of 3UI-spliced transcripts upon UPF1 knockdown, we observed a roughly equal split of transcripts being used more or less upon UPF1 knockdown (Figure 3.13A). Examining the entire set of transcripts regardless of individual levels of significance with DESeq2, we found that the distribution of log2FoldChange values for nonPTC 3UIs between UPF1 knockdown and non-knockdown conditions was normally distributed and centered at log2FoldChange=0 (Figure 3.13B). Statistical testing via a KS test indicated no significant difference between nonPTC 3UI-containing transcripts and all protein-coding transcripts. However, for PTC 3UI-containing transcripts, and those annotated as being nonsense-mediated decay sensitive (grouped together in Figure 3.13B as "NMD"), we find a significant shift towards upregulation upon UPF1 knockdown as expected (KS test $P<0.001$; Figure 3.13B).

We hypothesised that the distribution observed for nonPTC 3UIs in Figure 3.13B was due to a mixture of nonPTC 3UIs being either more or less than 55 nucleotides from the stop codon (the threshold for NMD sensitivity, see Section 1.5.1). Therefore we subsetted these events and compared the two subclassifications via a KS test, finding no significant difference (Figure 3.13C). This suggests that introns which are only ever found within the 3'UTR behave qualitatively differently from those arising from premature termination, and do not appear to sensitise transcripts to NMD in the colorectal carcinoma setting.

We also hypothesised that splicing within the 3'UTR may not act in an "all or nothing" manner with regards to transcript decay, and may instead act as a form of decay enhancer. If this were the case then we would expect the presence of multiple introns within the 3'UTR to produce a more profound effect upon UPF1 knockdown. However upon classifying transcript by the number of 3UIs they contained, and conducting a KS test between the subset that contained a single 3UI vs the subset that contained multiple 3UIs, we observed no significant difference (Figure 3.13D).

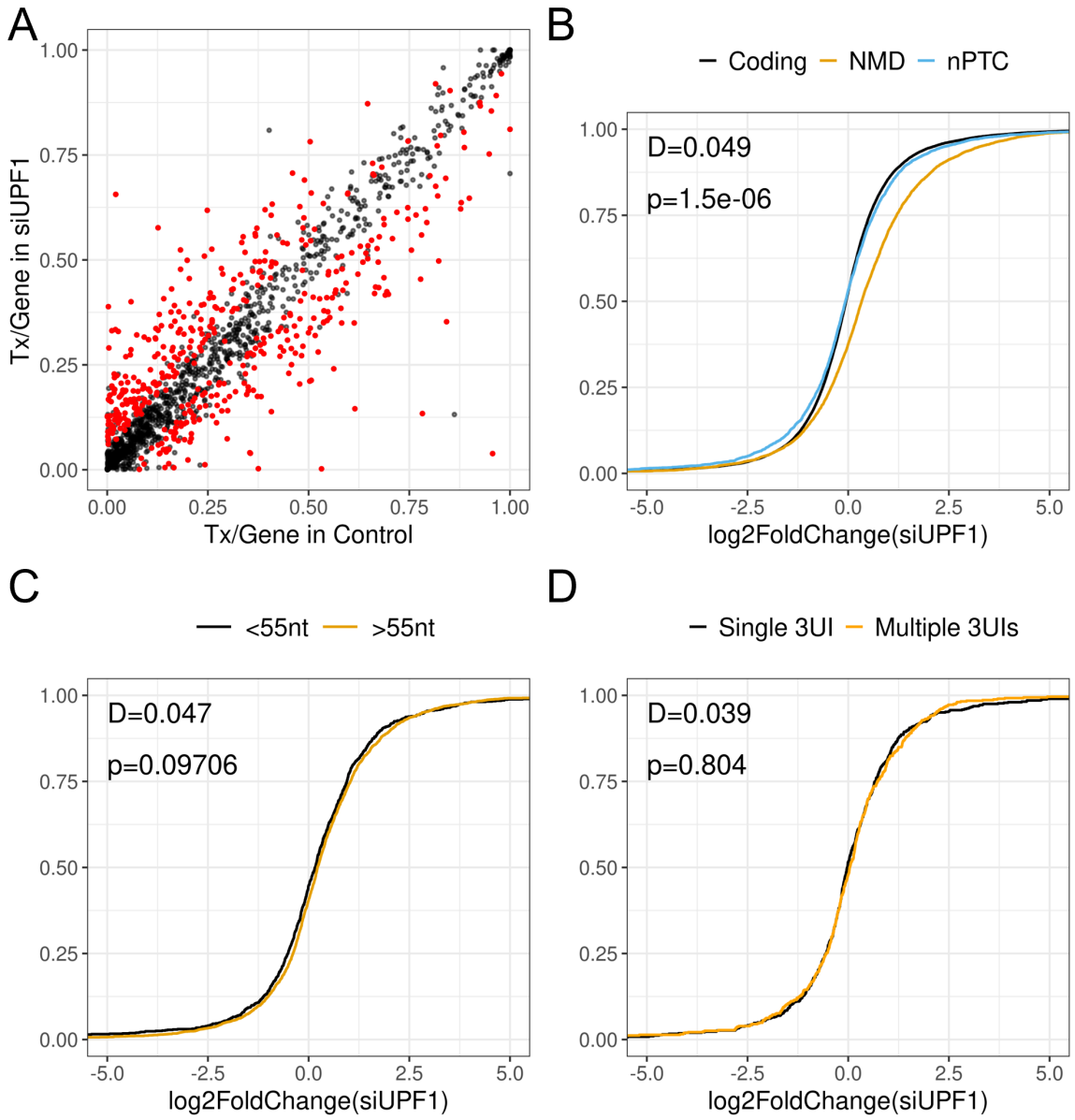

Figure 3.13: Impact of UPF1 knockdown on 3UI transcript expression. A) Fraction expression (transcript expression / gene expression) of 3UI transcripts in control- (siDsRed) vs siUPF1-treated HCT116 cells. Each black dot represents a 3UI transcript. Red dots represent 3UI transcripts that show significant differential transcript usage (P<0.05 and absolute effect size > 0.05). B-D) ECDF plots comparing log2FoldChanges upon UPF1 knockdown compared to negative control. B) Comparison between protein coding transcripts (black line), NMD-sensitive transcripts (pPTC 3UI-containing transcripts, and transcripts annotated as "nonsense_mediated_decay" in the reference; orange line), and nonPTC 3UI-containing transcripts (blue line). C) Comparison between 3UI transcripts where the terminal 3UI is less than 55nt from the stop codon (black line) or more than 55nt from the stop codon (orange line). D) Comparison between 3UI transcripts that only have a single 3UI (black line) or multiple 3UIs (orange line).

Additionally, we hypothesised that the differential transcript expression effect size upon UPF1 knockdown may be a function of the size of the intron, perhaps due to the removal of synergistic trans factors, or due to substantial shortening of long 3'UTRs, which could rescue them from EJC-independent NMD. Upon categorizing transcripts based on the total size of their 3UIs relative to the total size of their 3'UTR (into bins of 0-25%, 25-50%, 50-75%, and 75-100% spliced out), we observed no significant difference in expression between classes upon UPF1 knockdown (data not shown).

To explore the idea that 3'UTR splicing may be able to rescue transcripts from NMD, we identified a subset of nonPTC 3UI-spliced transcripts which were significantly downregulated upon UPF1 knockdown, whilst their 3UI-retaining partner transcripts were significantly upregulated (see HRAS example in Figure 3.14, and red dots in Figure 3.15A). Members of this subset of transcript included HRAS, CTNNB1, SRSF2, and RBM8A. In these instances, and using HRAS as an example, we find that the intron retaining partner is sensitive to UPF1 knockdown (potentially via EJC-independent NMD), and splicing the 3'UTR rescues the transcript from degradation (Figure 3.14).

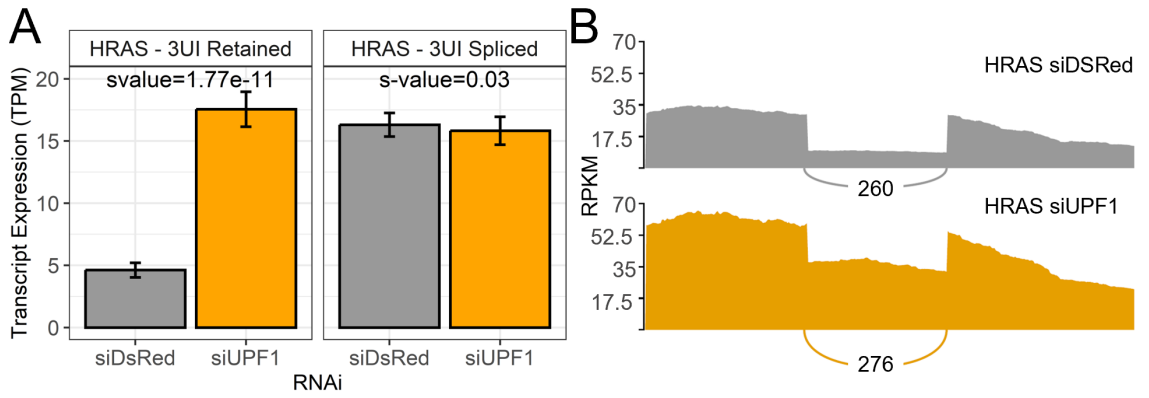

Figure 3.14: HRAS 3'UTR splicing rescues the transcript from NMD. Example of an NMD-rescuing event in the HRAS 3'UTR. A) Transcript level quantification (TPM values) and s-values from differential transcript expression analysis. The s-value represents the probability that the sign (i.e. direction of change) is incorrect, therefore the lower the s-value the higher the probability that the sign is correct. B) Sashimi plot showing event-level quantification and splice-spanning read count.

We first hypothesised that these may be instances where splicing the 3'UTR substantially reduces its length, thus releasing such transcripts from long 3'UTR (EJC-independent) NMD. Unlike in *S. cerevisiae* and *D. melanogaster*, no exact 3'UTR length threshold has been established for EJC-independent NMD in mammals (Muñoz et al., 2023), although longer 3'UTRs have been associated with increased UPF1 accumulation (Hogg and Goff, 2010). We found that the average 3'UTR size of these transcripts was 1114 nucleotides, whilst the average 3UI size was 151 nucleotides (Figure 3.15B). Therefore it appears unlikely that regulation of length alone is the primary mechanism of action in this instance.

A potential explanation for NMD rescue due to 3'UTR splicing by this subset of transcripts is that they may harbour cis elements that bind UPF1-interacting trans factors. The removal of these elements could rescue transcripts from UPF1-mediated regulation. To address this, we returned to the CLIPseq data used in Section 3.3. In the NMD rescuing nonPTC 3UIs we observed significant enrichment of 18 RBPs (Figure 3.15C), including RNA-induced silencing complex members AGO1, AGO3, and AGO4. We observed the highest fold enrichment in ALKBH5, an m6A demethylase, whilst we observed the most significant enrichment in LARP4B, RBM27, ILF3, FAM120A and CNBP. Increased GC content has previously been associated with EJC-independent NMD (Imamachi et al., 2017), therefore it is interesting that we observed significant enrichment of PCBP2, which binds poly(RC) motifs, and CNBP, which binds G-rich motifs.

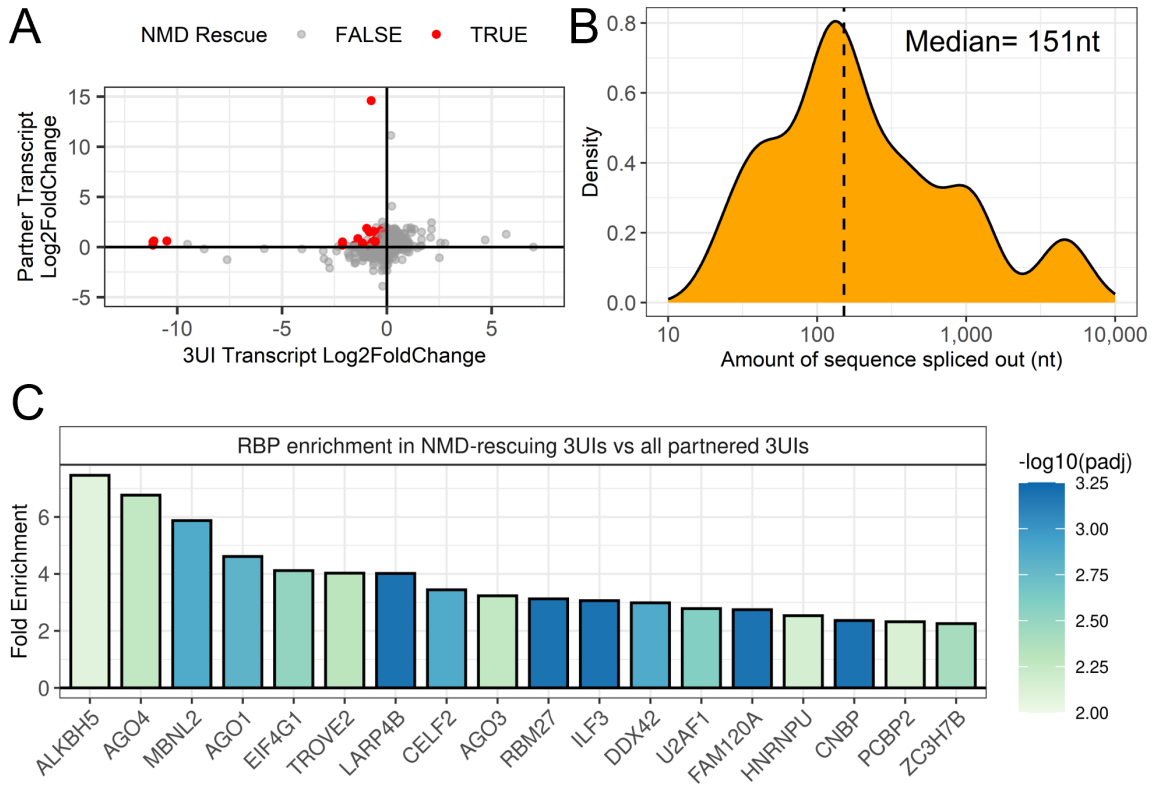

Figure 3.15: Identification of a subset of NMD-rescuing events. A) Log2FoldChanges of 3UI-spliced transcripts and matched 3UI-retaining partner transcripts in siUPF1 compared to siDsRed. Red dots indicate pairs where the 3UI-spliced transcript is downregulated upon UPF1 knockdown, the 3UI-retaining partner transcript is upregulated upon UPF1 knockdown, and both s-values < 0.05. B) Distribution of 3UI sizes, representing the amount of sequence spliced out between partners. Dotted line represents the median size. C) Enrichment of RBPs (from RBP-CLIPseq data) in NMD-rescuing 3UI events (red dots in A) compared to all partnered 3UI events (all dots in A).

3.4.2.2 In stem cells

To determine whether 3'UTR spliced transcripts in non-cancer cells are sensitive to NMD, we proceeded to knockdown UPF1 in H9 hESCs. In order to facilitate a side-by-side comparison, the same siRNA duplexes and concentrations used in the HCT116 knockdown were also used in the knockdown in H9 cells. Using Lipofectamine RNAiMAX alongside 30nM siUPF1_1 or siUPF1_2 in H9 hESCs led to widespread cell death within 72 hours of transfection; therefore, we proceeded with optimizing other

transfection reagents, concentrations, and transfection times, including PEI and Dharmafect 1 for 24, 48, and 72 hours. Following optimization, we found that reverse transfection with Dharmafect 1 resulted in highest levels of cell viability. Transfection of 20nM siUPF1 into H9 hESCs for 48 hours resulted in a 64% knockdown at the protein level for the siUPF1_1 duplex, and a 90% knockdown at the protein level for the siUPF1_2 duplex (Figure 3.16A). This was consistent with the knockdown in HCT116 cells where we also found that the siUPF1_2 duplex produced a more effective UPF1 knockdown compared to the siUPF1_1 duplex. UPF1 knockdown was conducted for 48 hours as opposed to 72 hours due to effects on cell viability. Even following optimization there was a greater degree of cell death observed in H9 hESCs compared to HCT116 cells after 24 hours. The level of cell death was workable after 48 hours, where RNA extraction could be conducted to produce enough RNA for sequencing, whilst after 72 hours the level of cell death was too high to facilitate downstream analysis. Unlike the knockdown in HCT116 cells, an NMD small molecule inhibitor was not considered, and instead RNA sequencing was conducted on RNA extracted from cells transfected with both the UPF1_1 duplex, siUPF1_2 duplex, alongside negative control, with four biological replicates for each condition (12 total samples; whereas for HCT116 cells only the siUPF1_1 and negative control were sequenced).

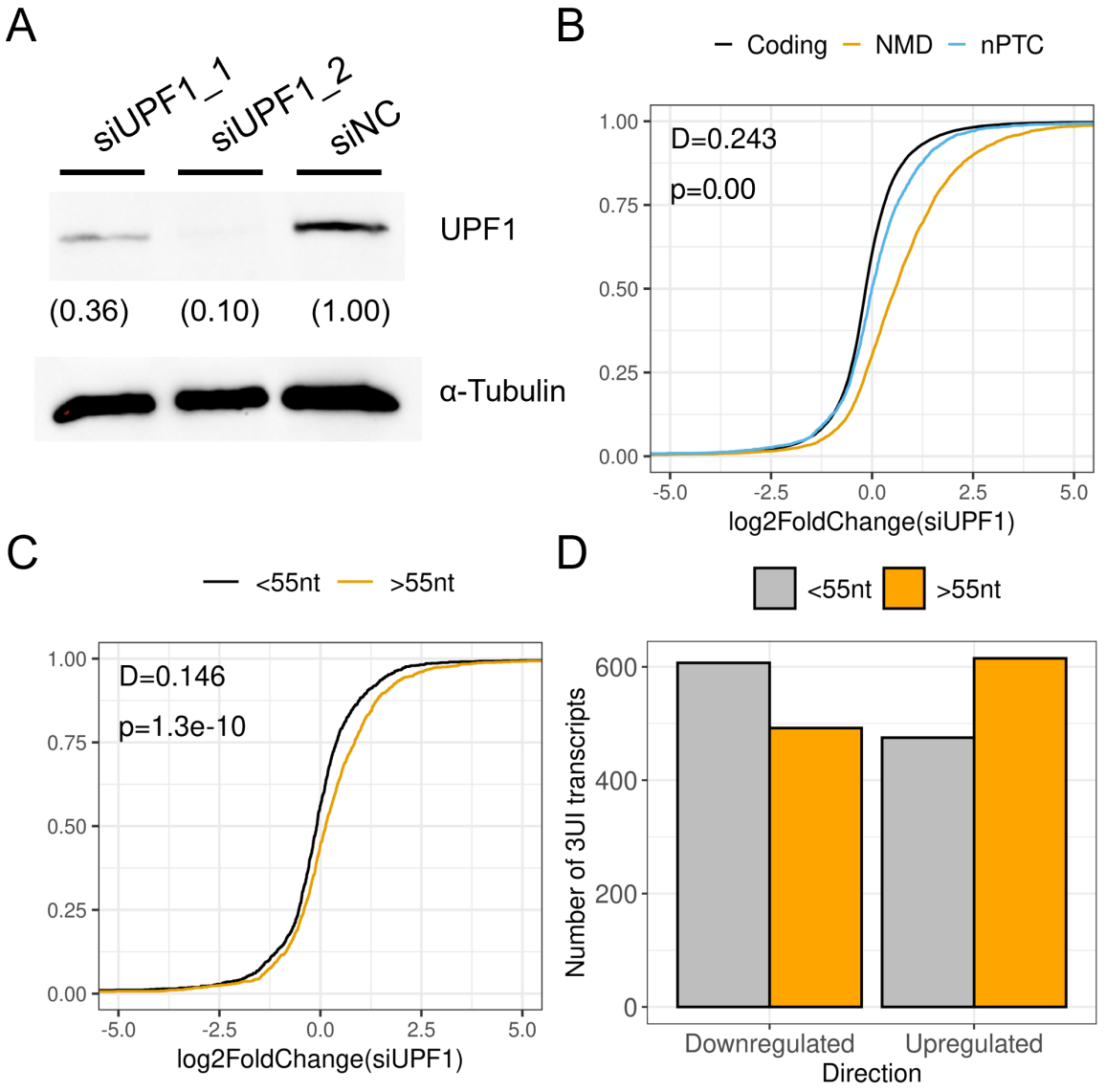

Figure 3.16: Knockdown of UPF1 in H9 hESCs reveals effect of 3'UTR splicing on NMD sensitivity. A) Validation of UPF1 knockdown in H9 hESCs by Western blotting following siRNA transfection. B+C) ECDF plots showing distribution of log2FoldChange upon UPF1 knockdown, compared between: B) All protein coding (black), transcripts annotated as "nonsense_mediated_decay" or pPTC 3UIs (orange), nonPTC 3UIs (blue); C) nonPTC 3UI-containing transcripts that have their terminal 3UI less than 55nt from the stop codon (black) or more than 55nt from the stop codon (orange). D) Breakdown of number of nonPTC 3UI transcripts that have a terminal 3UI >55nt from the stop codon (orange) or <55nt from the stop codon (grey) that are either upregulated or downregulated upon UPF1 knockdown.

Despite reduced levels of total RNA extracted in the siUPF1_1 and siUPF1_2 conditions due to cell death, sample quality control conducted by Novogene indicated that RNA was high quality (average RIN = 9.4). Subsequently RNA was sequenced. Reads were mapped and quantified against the HipSci transcriptome assembly before differential transcript expression analysis was conducted. UPF1 knockdown in H9 hESCs resulted in significant upregulation of transcripts harbouring a pPTC 3UI and those annotated as "NMD sensitive" in the reference (D=0.243; P<2.2e-16; Figure 3.16B orange line). However, we also observed a significant, but smaller, upregulation of transcripts harbouring nonPTC 3UIs (D=0.118; P<2.2e-16; Figure 3.16B blue line). Additionally, we found that transcripts where the terminal nonPTC 3UI was >55nt from the stop codon were upregulated more often than those where the terminal 3UI was <55nt from the stop codon (Figure 3.16C). However, there is still a substantial proportion of transcripts that do not follow this trend (Figure 3.16D). We find 481 transcripts (in 328 genes) with a terminal 3UI <55nt (predicted to be NMD-insensitive) that are upregulated upon UPF1 knockdown, as well as 499 transcripts (in 402 genes) that are downregulated despite having a terminal 3UI >55nt (Figure 3.16D). This latter subset may represent transcripts that are evading NMD. To determine whether this was specific to a subset of biological functions, we conducted gene ontology analysis on the 402 genes; however, we observed no significant hits.

Similarly to the "NMD rescue" analysis conducted on HCT116 cells, we examined the 3UI splicing events that were downregulated upon UPF1 knockdown to determine whether there was an opposing upregulation in their 3UI-retaining partner. Whereas in HCT116 cells we found 48 instances of "NMD rescue", for H9 cells we only found 10. Interestingly, splicing the 3'UTR of HNRNPDL and SRSF2 was found to produce such a phenomena in H9 hESCs; however, unlike in HCT116 cells, HRAS and CTNNB1 were not found within this subset.

3.5 Impact on post-transcriptional regulatory roles of 3'UTR

To gain a more mechanistic insight into the effect that splicing the 3'UTR has on post-transcriptional regulation of gene expression, we conducted a series of Luciferase assays in HCT116 colorectal carcinoma cells. For these assays we used CTNNB1 and HRAS as model genes as they are both "broadly expressed" in colon cancer. The use of the Luciferase system in combination with site directed mutagenesis allowed the interrogation of the post-transcriptional regulatory mechanisms acting on each isoform, as well as the effect of preventing splicing (through mutating the splice sites), or forcing intron removal (through cloning out the introns). However, given that the output of the Luciferase assay is relative luminescence, it represents the net sum of post-transcriptional output (including transcription, RNA export, RNA stability, and translation efficiency). Therefore in order to break the results of the Luciferase assays down further we also treated cells with Actinomycin D (to determine the effect of 3'UTR splicing on RNA stability) and conducted polysome profiling (to determine relative translation efficiency of 3'UTR intron spliced/retained pairs).

3.5.1 Luciferase assays highlight net effect of 3'UTR splicing on post-transcriptional regulation

The HRAS and CTNNB1 3'UTRs were cloned downstream of the Luciferase coding sequence in the pCI-neo plasmid and subjected to site-directed mutagenesis either to mutate the splice site, and therefore prevent splicing, or to clone out the intron (Figure 3.17A-B). An important feature of these Luciferase plasmids is that they contain a chimeric intron immediately downstream of the CMV promoter, therefore even where the 3'UTR intron is cloned out, the expression construct is not intronless, therefore any comparison made does not reflect intron-containing vs intron-less expression.

Whilst the HRAS 3'UTR contains only a single intron, and was therefore only subjected to 5'SS mutagenesis, the intron in CTNNB1 has two possible 3'SS. Therefore for CTNNB1,

three splice site mutant constructs were generated: a 5′SS mutant (which prevents splicing of both intron-containing isoforms); a 3′SS mutant targeting the proximal 3′SS (preventing production of the short-intron-spliced isoform); and a 3′SS mutant targeting the distal 3′SS (preventing production of the long-intron-spliced isoform)(Figure 3.18A-B).

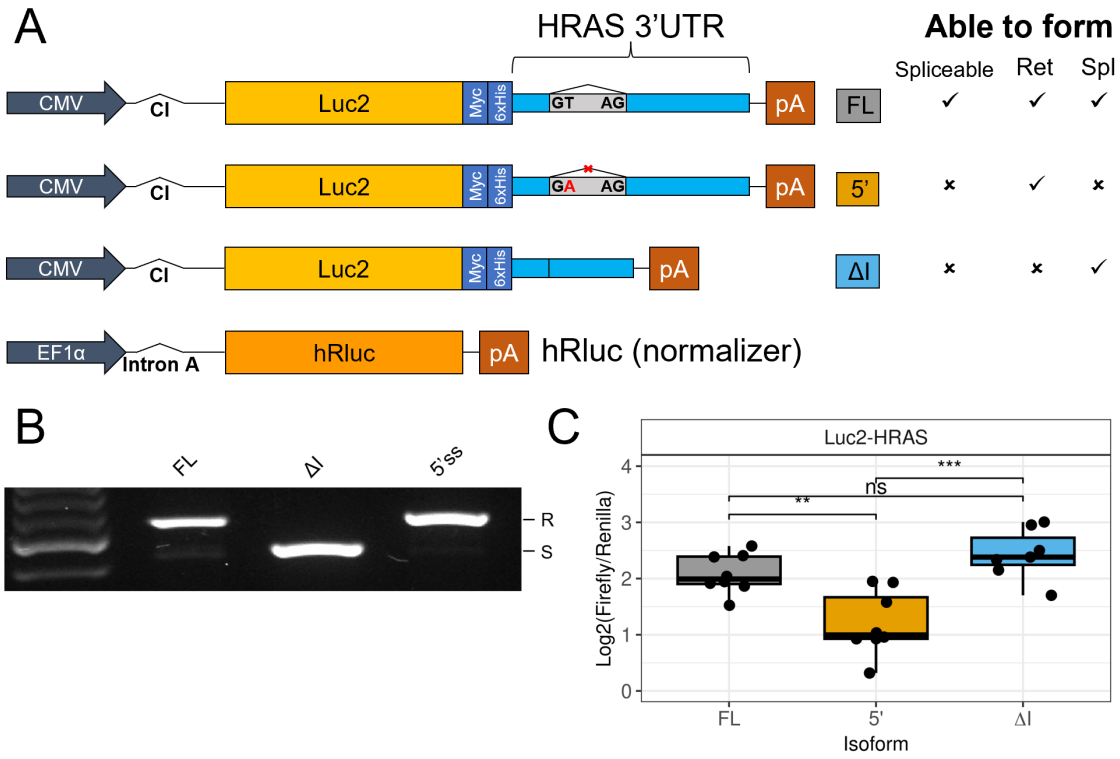

Figure 3.17: Effect of HRAS 3′UTR splicing on Luciferase expression. A) Schematic overview of Luc2-HRAS constructs and the isoforms they can produce. B) RT-PCR showing the production of each isoform from the corresponding plasmid following its transfection into HCT116 cells. C) Relative luminescence observed upon transfection of each Luc2-HRAS construct into HCT116 cells.

By mutating the 5′SS in the HRAS 3′UTR, we observed a significant decrease in relative luminescence compared to the full-length HRAS 3′UTR, whilst cloning out the HRAS 3′UTR intron did not change relative luminescence significantly (Figure 3.17C). However, when comparing directly between the 5′SS mutant (which is forced to retain the full 3′UTR intron sequence), and the Δintron construct (where the intron has been removed, but due to cloning, not endogenous splicing), we observed a significant difference (Figure

3.17C). Together these findings indicate that the modulation of sequence within the intron, as opposed to the physical act of splicing itself (e.g. though regulation of mRNP composition due to spliceosome-dependent interactions) are most important to overall expression of this construct, as neither the 5′SS construct nor the Δintron construct are spliced within the cell.

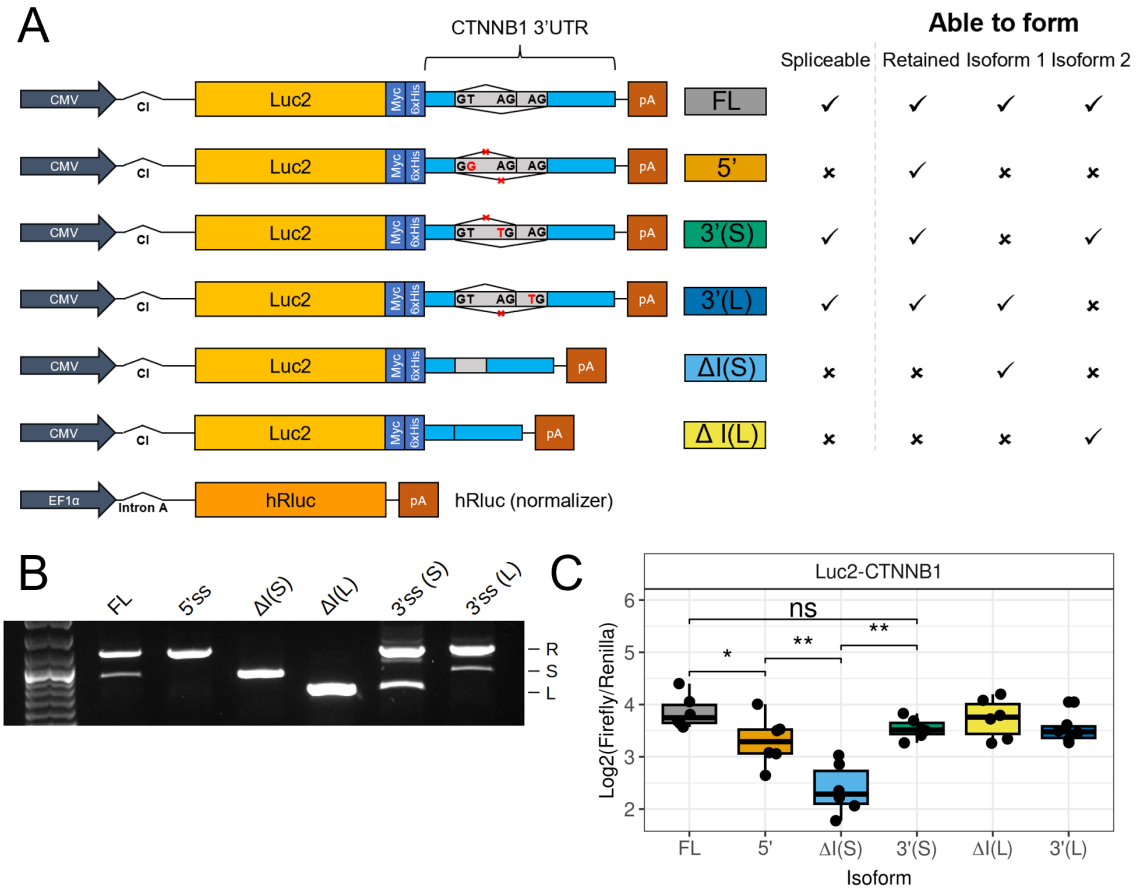

Figure 3.18: Effect of CTNNB1 3′UTR splicing on Luciferase expression. A) Schematic overview of Luc2-CTNNB1 constructs and the isoforms they can produce. B) RT-PCR showing the production of each isoform from the corresponding plasmid following its transfection into HCT116 cells. C) Relative luminescence observed upon transfection of each Luc2-CTNNB1 construct into HCT116 cells.

When we mutated the 5′SS of CTNNB1 to force retention of both introns, we observed a significant reduction in relative luminescence (Figure 3.18C). However, where either of the 3′SS were mutated (3′(S) - to prevent splicing of the short 3′UTR introns; 3′(L) - to

prevent splicing of the long 3'UTR intron) there was no significant reduction in expression compared to the full-length CTNNB1 3'UTR (Figure 3.18C). This suggests that there may be some degree of compensation occurring, i.e. when either isoform is mutated the other isoform is spliced more. By performing PCR amplification of the entire 3'UTR from cDNA extracted from HCT116 cells transfected with each construct (thus representing the endogenous splicing profiles produced from these transfected constructs), we observed this compensatory phenomena (Figure 3.18B). Cloning out the short intron produced an even larger reduction in relative luminescence, whilst cloning out the long intron did not (Figure 3.18C). Potential explanations for this finding are discussed in the chapter summary.

3.5.2 Effects of 3'UTR splicing on RNA stability and translation efficiency

Whilst the Luciferase assays conducted in Section 3.5.1 informed us of the impact of splicing or removing 3'UTR introns on expression output, they did not shed light on the specific post-transcriptional regulatory effects of such mutations. Instead they represent the net sum of regulation. To determine the impact of splicing 3'UTRs on RNA stability, we conducted actinomycin D RNA stability assays in HCT116 cells (to allow comparison with Luciferase assays) and H9 hESCs. Actinomycin D treatment shuts down transcription within the cells, RNA is subsequently extracted at various time points thereafter, allowing us to quantify how much RNA remains for 3UI-spliced and 3UI-retained isoforms. With the help of Chiara Galloni, who generated polysome gradients, we also conducted polysome profiling to assess the association of 3UI-spliced and 3UI-retained isoforms with monosomes and polysomes, allowing us to estimate relative translation efficiencies.

3.5.2.1 RNA stability

HCT116 cells were treated with Actinomycin D for 12 hours, and RNA was extracted at 0hr, 3hr, and 12hr of treatment. Following qPCR, Ct values were normalized within spliced-retained pairs for each biological replicate, to account for different transcript expression levels between samples. For both HRAS and CTNNB1, we found that the spliced isoform was more stable in HCT116 cells, therefore both Ct values were normalized against the spliced isoform, as shown in Figure 3.19A-B. The finding that spliced isoforms of CTNNB1 and HRAS were more stable than their retaining partners supports the results previously described in Section 3.4.2.1 that their retained partners were upregulated upon UPF1 knockdown, and that splicing of these transcripts may rescue them from UPF1-mediated degradation. However, Actinomycin D treatment in combination with UPF1 knockdown would need to be conducted to test this hypothesis further. The finding from the Luciferase assay that removing the short intron of CTNNB1 led to a significant decrease in relative luminescence (where the intron was removed via cloning; ΔIntron(S) construct in Figure 3.18C), whilst splicing the intron out increases stability (Figure 3.19A) and relative luminescence (FL construct in Figure 3.18C), suggests that the difference in relative luminescence between the two constructs is in part due to a splicing-dependent effect on RNA stability.

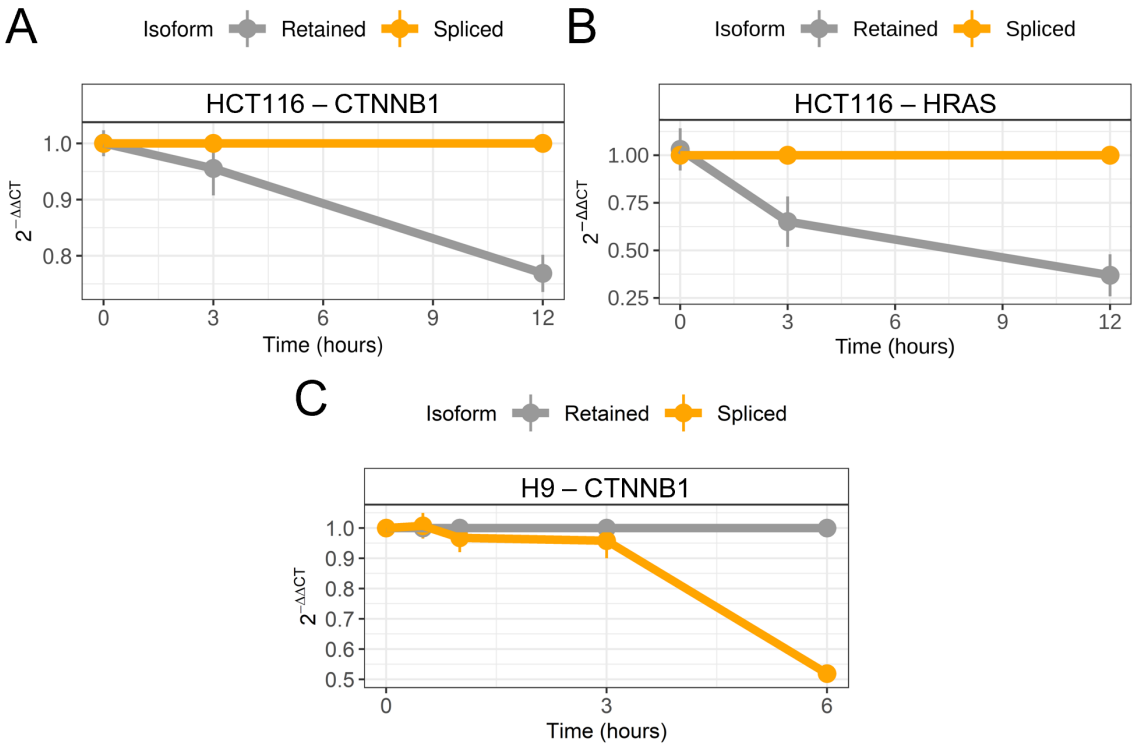

Figure 3.19: Impact of 3'UTR splicing on CTNNB1 and HRAS RNA stability. Relative expression of spliced (orange lines) vs retained isoforms (grey lines) upon Actinomycin D mediated transcriptional shutdown over time for: A) CTNNB1 in HCT116; B) HRAS in HCT116; C) CTNNB1 in H9.

Actinomycin D treatment of H9 hESCs produced a much greater effect on cell viability, where over 50% of cells had died by 6 hours, and less than 10% survived 12 hours of treatment. Therefore H9 hESCs were only treated with Actinomycin D for 6 hours, and RNA was extracted at 0hr, 3hr, and 6hr of treatment. Interestingly in H9 hESCs, we observed that the 3UI-spliced isoform of CTNNB1 was less stable than its 3UI-retaining partner (Figure 3.19C), which is the opposite of the result seen in HCT116 cells (Figure 3.19A). However, the impact of UPF1 knockdown on expression of the CTNNB1 spliced isoforms was also different from that observed in HCT116, where splicing was shown to rescue transcripts from NMD. In H9 hESCs neither 3UI-spliced isoform was significantly impacted by UPF1 knockdown, whilst the retained isoform was significantly downregulated. Given that the intron within the CTNNB1 3'UTR is <55nt from the stop

codon, and was not shown to be sensitive to UPF1 knockdown in H9 hESCs, the finding that splicing the CTNNB1 3'UTR decreases RNA stability suggests that either: 1) the CTNNB1 3'UTR intron contains stabilizing elements, or 2) there is a splicing-dependent destabilising effect independent of the NMD pathway, or a combination of both.

Whilst RNA stability has been investigated here for only a single gene of interest in H9 hESCs via Actinomycin D treatment, an orthologous transcriptome-wide approach is conducted in Chapters 5 and 6 across a cardiomyocyte differentiation time course. Such an approach later allows us to investigate how RNA stability changes during hESC differentiation (Chapter 5), and how 3'UTR splicing impacts RNA stability at each stage, including differentiation stage-specific regulation (Chapter 6).

3.5.2.2 Translation efficiency

To determine whether splicing the 3'UTR of CTNNB1 and HRAS impacted translation efficiency, we next conducted qPCRs on polysome gradient fractions generated by Chiara Galloni using HCT116 cells. An example polysome profile is shown in Figure 3.20A. For our qPCRs, only the 80S monosomal, LO polysomal, and HI polysomal fractions were used as input, to represent ribosome-loaded, lowly-translated, and highly-translated RNA pools. Conducting qPCR on each pool indicated that the 3UI-spliced isoform of HRAS was more abundant than the 3UI-retaining isoform in all three fractions (Figure 3.20B). Whilst for CTNNB1 the 3UI-retaining isoform was more abundant across all three fractions (Figure 3.20D). However, these relative abundances also reflect the overall expression ratio (spliced vs retained) observed across the whole cell. Therefore, in order to determine the relative translation efficiency of each isoform, we normalized the amount of RNA detected in the LO and HI polysomal fractions by the amount detected in the 80S monosomal fraction. As such the values presented in Figure 3.20C+E represent the relative amount of ribosomally-associated RNA for HRAS (Figure 3.20C) and CTNNB1 (3.20E) that have several (LO) or many (HI) ribosomes associated. For HRAS, whilst there was an slight reduction in average translation efficiency for 3UI-spliced

RNA, this difference was not significant (Figure 3.20C). Similarly for CTNNB1, there was a slight, but non-significant, increase in translation efficiency of the 3UI-spliced isoform compared to the 3UI-retaining isoform (Figure 3.20E).

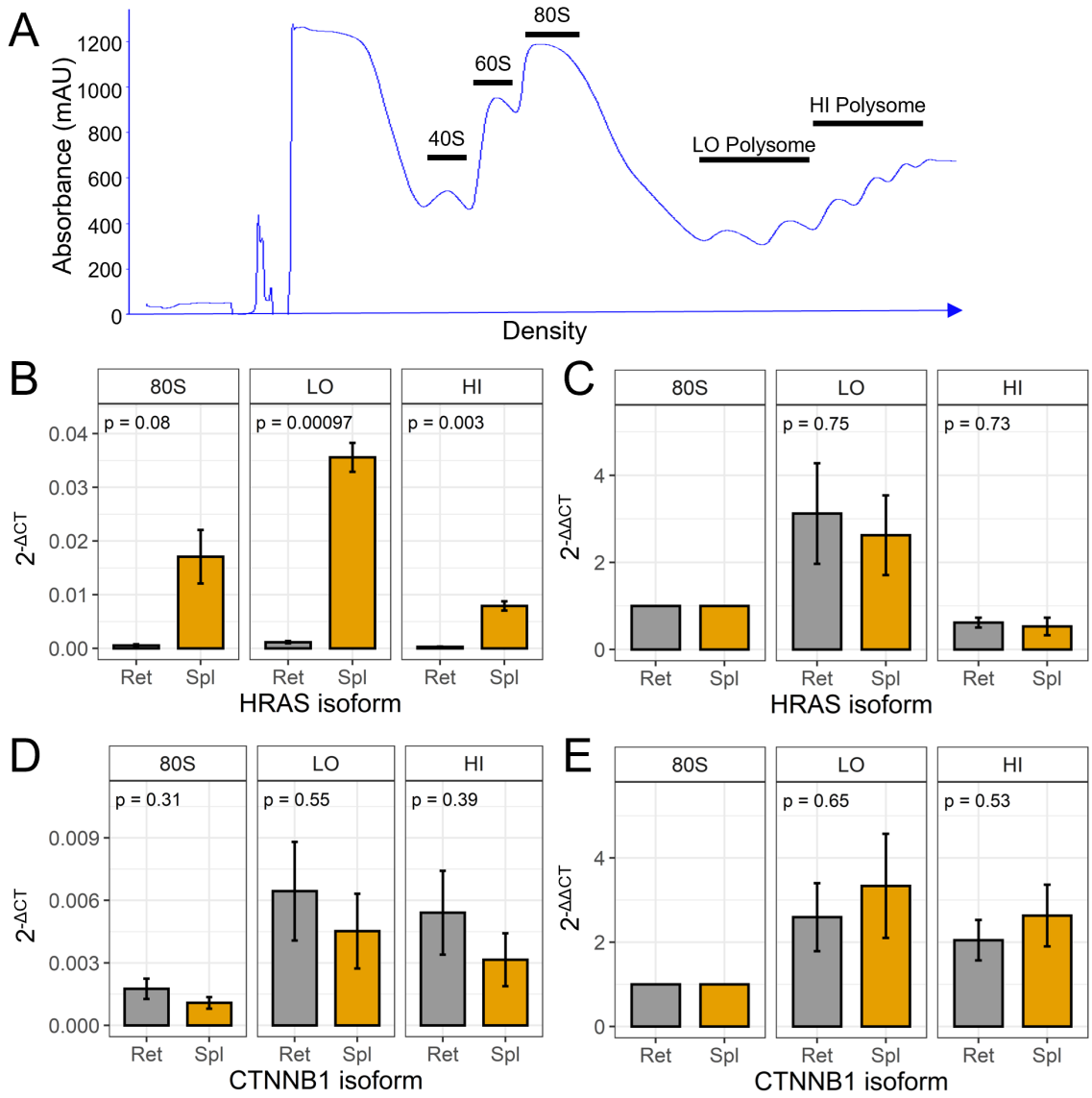

Figure 3.20: Polysome profiling and qPCRs. A) Example polysome gradient from one of four biological replicates ran on sucrose gradient. B+D) Relative expression of B) HRAS and D) CTNNB1 spliced and retained isoforms in 80S, LO polysome and HI polysome fractions, normalized to RPL30. C+E) Relative translation efficiency (where expression in each fraction has been normalized to the corresponding isoform in 80S fraction) for spliced and retained isoforms of C) HRAS and E) CTNNB1.

3.6 Summary

In this chapter we detected several thousand 3UIs and 3UI-containing transcript and found that many of these are highly and broadly expressed. These 3UIs contribute substantially towards the expression of the genes they are from, which suggests that their presence may be important in maintaining a specific level of expression for genes belonging to the biological processes and pathways shown in Figure 3.4. Regardless of whether the presence of a 3UI increases or decreases expression, without the 3UI the expression of these genes may be disrupted, which could in turn lead to dysregulation of the biological processes in Figure 3.4A, and signalling pathways in Figure 3.4B. Additionally, condition-specific modulation of the inclusion of these 3UIs via alternative splicing, for example during cell differentiation, may lead to the regulation of these pathways and processes. These regulatory avenues are explored further in the next chapter.

We validated that these events were the result of genuine splicing, demonstrated that their splice sites consist of the canonical GT-AG dinucleotides in the majority of cases, and are more conserved than the surrounding sequence. We also showed that 3UI-spliced transcripts are exported from the nucleus, therefore they are likely to be translated at least once regardless of NMD-sensitivity. To address whether 3UI-spliced transcripts are subject to NMD, we knocked out UPF1 in both HCT116 colorectal carcinoma cells and H9 hESCs, allowing us to compare 3UI-mediated NMD between cancer and non-cancer settings. We found that in the cancer setting the nonPTC 3UIs, those only ever observed in the 3'UTR, generally did not elicit NMD, regardless of their position relative to the 55nt rule for NMD. We instead identified a subset of 3UI-spliced transcripts where splicing out the 3UI rescued the transcripts from UPF1-mediated degradation. Given that the 3UI-retaining isoform displayed sensitivity to UPF1 knockdown, this indicated that removing the intron via splicing rescued the transcript from EJC-independent NMD. In contrast, by knocking out UPF1 in H9 hESCs, we observed the 55nt rule in action, where

3UIs situated >55nt from the stop codon were upregulated more commonly upon UPF1 knockdown than those situated <55nt from the stop codon. However, we still observed a substantial pool of 3UI-spliced transcripts that violated this rule, suggesting that these transcripts may be evading EJC-dependent NMD. It is important to note that the level of cell death in H9 hESCs was substantially higher than in HCT116 cells upon knocking down UPF1 for 72 hours, suggesting a greater sensitivity to global transcriptome disruption in H9 hESCs. This led to both reduced total RNA yield and reduced RNA quality following RNA extraction. This may be directly related to increased levels of apoptosis (Thomas et al., 2015), or potentially due to an increased level of RNA-degrading contaminants accompanying cell debris in culture. To account for this, we ensured that dead cells and cell debris were cleared via a PBS wash prior to RNA extraction. Additionally, we used a lower siRNA concentration and incubation time (as discussed in 3.4.2.2) to reduce levels of cell death, which substantially increased yield and RNA quality.

We found that 3UIs are significantly enriched for many RBPs, including those involved in RNA splicing. Additionally, we find that broadly expressed 3UIs are significantly enriched for many RBPs and miRNA binding sites compared to an all nonPTC 3UI background. This suggests that the removal of these RBP and miRNA binding sites contributes to the higher levels of expression and usage of these broadly expressed 3UIs.

To gain a more mechanistic insight into the effect of 3'UTR splicing on gene expression, we cloned the 3'UTRs of HRAS and CTNNB1 downstream of the Luciferase coding sequence and conducted a series of mutations to either remove each intron, or force their retention. Subsequently we were able to compare the relative Luciferase output of each construct in transfected HCT116 cells. These assays were also attempted in H9 hESCs, although due to poor transfection efficiency a high degree of variability was observed. For the HRAS 3'UTR we found that removing the intron significantly increased Luciferase output, whilst forcing intron retention via splice site mutation significantly

reduced Luciferase output. We concluded that this was predominantly through sequence modulation (e.g. removal/inclusion of destabilizing element), and not due to the physical act of splicing, as the major effect was observed between unspliced constructs. Where the act of splicing also contributed to increased expression, we would expect the full length HRAS 3'UTR construct to have the highest expression given that it is spliced to a high degree endogenously, and therefore we may have expected to see a synergistic relationship between both splice-dependent and splice-independent forms of regulation.

In contrast, we found that removing the short intron from the CTNNB1 3'UTR significantly reduced expression. However, removing the long intron did not. This is surprising given that the long intron encompasses the entire short intronic sequence, plus an extra 159 nucleotides. When the short intron was cloned out it may have removed stabilizing sequence elements that positively regulate expression, and whilst these sequence elements were present in the short intron, the long intron may have also contained additional destabilizing elements which masked this. Cloning out the short intron could have also produced a destabilizing sequence element by bringing together sequences from the 5' and 3' of the intron, whilst the sequence 3' of the long intron may not have contributed to the production of such an element upon cloning out the long intron (Figure 3.21A). Given that this sequence was removed via cloning, i.e. no splicing involvement, then there is no EJC, or EJC-interacting proteins in the nearby vicinity to prevent binding of destabilizing RBPs or miRNAs to this newly created sequence element (Figure 3.21B). Where splicing occurs, e.g. in the case of splicing in the 3'(L) construct, which can only produce the short-spliced isoform, the composition of the mRNP is modulated, which could lead to an EJC-mediated upregulation (relative to ΔIntron(short)) due to the binding of trans-factors (Figure 3.21C).

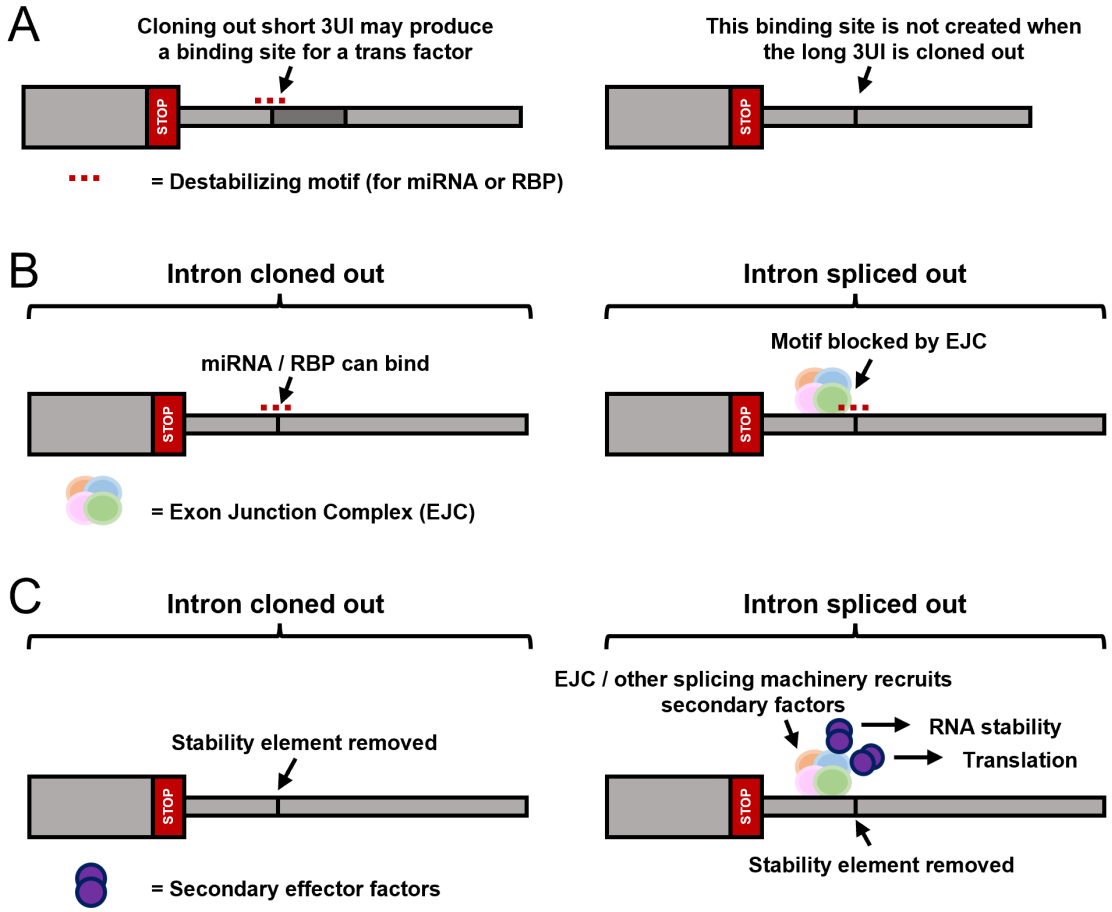

Figure 3.21: Potential explanations for Luc-CTNNB1 results. A) Cloning out the short 3UI may produce a binding site for trans factors by joining together sequences 5′ and 3′ of the intron. This binding site may not been created when the long 3UI is cloned out. B) Where the intron is cloned out, it may produce binding sites for miRNAs/RBPs. These binding sites may be blocked by the EJC or other spliceosome-associated RBPs when the intron is removed via splicing. C) Where the intron is cloned out, stability elements may be excluded. This also occurs when the intron is spliced out; however, there may be additional mRNP modulation and involvement of trans regulators which influence how the construct is expressed.

Given that differences in Luciferase output represent the net sum of regulation, we conducted Actinomycin D RNA stability assays and polysome profiling. This led us to observe an increase in RNA stability by 3′UTR splicing for both CTNNB1 and HRAS in HCT116, whilst we observed the opposite for CTNNB1 in H9 hESCs. Together these results highlight the importance of 3′UTR splicing in being able to regulate the inclusion

or exclusion of sequence elements that increase or decrease expression, including stability elements. Additionally, given that cloning out the CTNNB1 3'UTR intron reduced expression, whilst splicing it out did not, highlights the impact that the act of splicing itself has on expression, possibly due to the spliceosome modifying the mRNP composition.

Whilst this chapter has focused on characterizing the extent and impact of 3'UTR splicing in colorectal carcinoma cells and undifferentiated hESCs, these introns may be subject to alternative splicing in condition- and cell-type-specific contexts. As such, Chapter 4 will address how 3'UTR splicing changes between non-cancer and cancer samples, and during stem cell differentiation, whilst Chapters 5 and 6 will address how RNA stability changes during the differentiation of H9 hESCs into cardiomyocytes, and how 3'UTR splicing impacts this.

4. Alternative 3'UTR splicing in stem cell differentiation and cancer transformation

As discussed in Chapter 3, 3'UTR splicing is widespread (Figure 3.1), many 3UI-containing transcripts are highly expressed (Figure 3.3), and are found in genes with functions in important biological processes and signalling pathways (Figure 3.4). Therefore, we hypothesised that the levels of 3'UTR splicing would differ between biological contexts, for example during cancer transformation, and in differentiation. Modifying the level of 3'UTR splicing between cell-states (e.g. during differentiation) through alternative splicing may represent a mechanisms to regulate gene expression through the regulation of NMD sensitivity (Section 1.5.3), miRNA and RBP binding composition, as well as the dynamics of expression (Chapters 5 and 6). This chapter aims to explore the extent of alternative 3'UTR splicing to gain insight into potential context-specific functions of 3'UTR spliced transcripts.

In Chapter 3 we explored how the effects of 3'UTR splicing differed between the cancer and non-cancer setting, primarily focusing on differences in NMD sensitivity. To further investigate how 3'UTR splicing differs between cancer and normal samples, in Section 4.1 we utilize the TCGA dataset to interrogate the extent of splicing changes between cancer and matched normal samples. By using HCT116 cells as a colorectal carcinoma model, we investigate how changes in 3'UTR splicing may relate to predicted NMD sensitivity. Additionally, we investigate how the Wnt signalling pathway, which is commonly hyper-activated in colon cancer (The Cancer Genome Atlas Network, 2012), regulates expression of 3'UTR spliced transcripts.

We also investigate how 3'UTR splicing changes during stem cell differentiation. In Section 4.2 we utilize a high resolution cardiomyocyte differentiation dataset (Strober et al., 2019) to investigate how the usage of 3'UTR spliced transcripts changes during differentiation, including 3UIs within the Wnt pathway, and its central regulator CTNNB1.

4.1 3'UTR splicing is altered between healthy and solid tumour samples

In the first section of this chapter, the results of differential analyses conducted in contribution to a recent publication (Riley et al., 2024) will be discussed. Here, differential exon usage (DEU) was conducted between normal and cancer samples for 15 solid tumour types from TCGA, using RMATS. We first interrogated all pPTC and nonPTC 3UI splicing events contained within our TCGA transcriptome assembly, regardless of each individual event's significance level, and as such representing an overall splicing signature. We found that the majority of pPTC (CDS-overlapping) 3UIs, tend to display an increased level of intron retention in cancer samples compared to matched normals, except in tumours from the brain, breast and thyroid, which display an equal split between increased retention and increased splicing (Figure 4.1A). In the case of liver and pancreatic tumours, the majority of pPTC 3UIs are spliced more in cancer compared to matched normals (Figure 4.1A). Turning to nonPTC 3UIs, those only ever found in 3'UTRs, we find these are generally spliced more in cancer than in matched normals in all tumour types except liver cancer (Figure 4.1A). However, the effect size of these changes is relative small, typically representing changes of less than 2% (Figure 4.1B). This paints the picture of a system in which global splicing is slightly out of tune, and the differences in 3'UTR splicing reflects this. However, it is interesting that the direction of change for pPTC and nonPTC 3UIs differs, suggesting that these introns may be regulated differently.

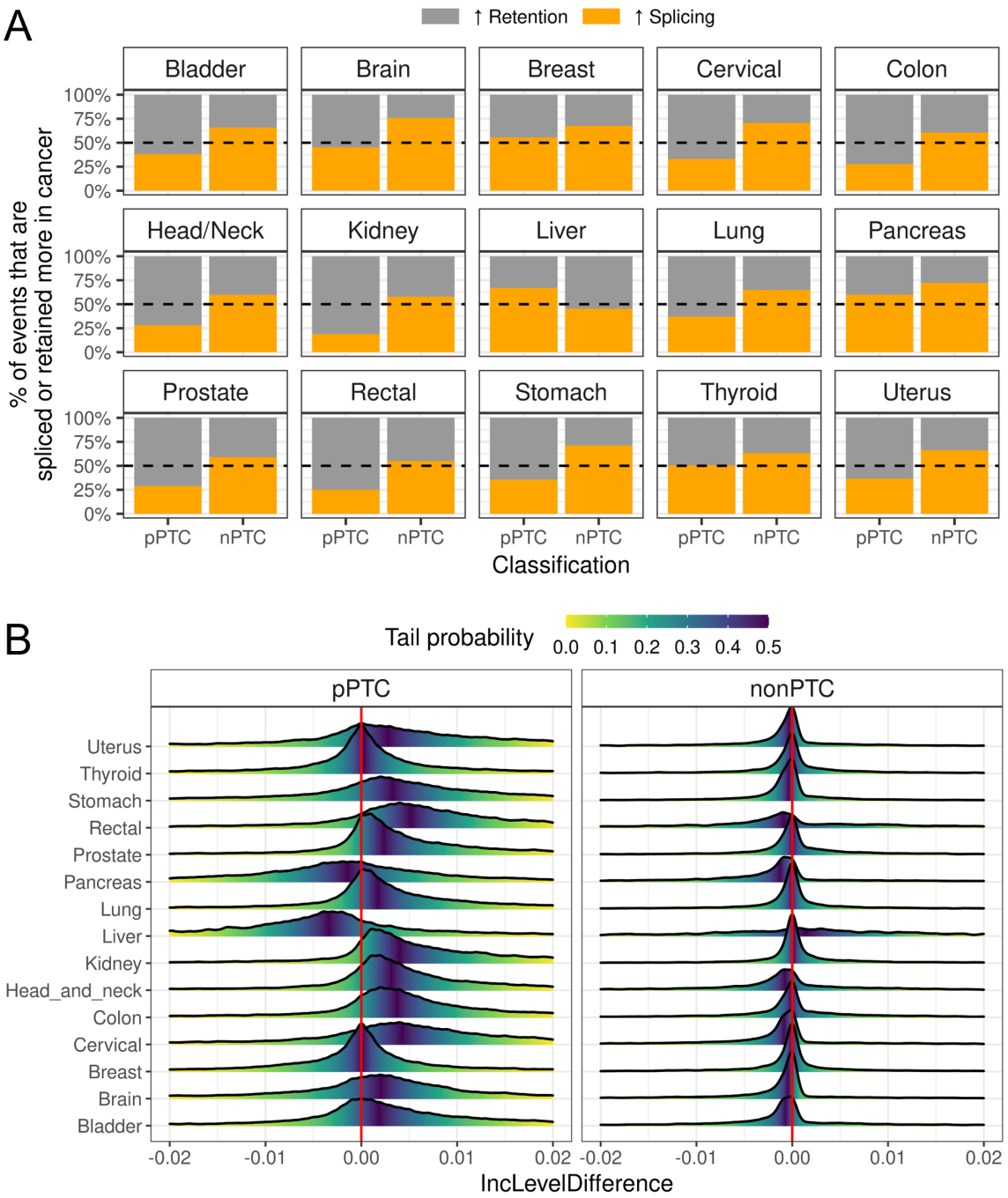

Figure 4.1: 3'UTR splicing is dysregulated between healthy and cancer samples across multiple tissues. A) Percentage of 3'UTR splicing events that are spliced more (orange) or retained more (grey) in cancer compared to non-cancer samples, faceted by cancer type. Each facet has two bars, one for pPTC 3UIs (overlap the CDS), one for nonPTC 3UIs (only found in the 3'UTR). B) Ridgeline plot showing the distribution of IncLevelDifference between cancer and normal samples for each cancer type. Negative IncLevelDifference = more spliced in cancer compared to non-cancer; positive IncLevelDifference = more retained in cancer compared to non-cancer.

4.1.1 Investigating significant differences in 3'UTR splicing in colorectal carcinoma cells

To investigate more substantial changes in 3'UTR splicing, i.e. those that display statistically significant differences, we focused on nonPTC 3UI splicing in colorectal carcinoma. We focused on this specific cancer type because HCT116 cells were available within the laboratory, and UPF1 knockdown had previously been conducted with this cell type, facilitating comparison of 3'UTR splicing differences and NMD sensitivity. We found that 340 nonPTC 3UIs were significantly differentially spliced (FDR<0.05) in colon tumour samples compared to normal samples. These 340 nonPTC 3UIs were found in 235 genes. To identify any functional enrichment, we ranked them from most over-spliced in colon cancer (compared to non-cancer controls) to most over-retained, and performed GSEA. We observed significant enrichment of gene sets related to apoptosis (NES=1.791), DNA repair (NES=1.735), Wnt signalling targets (NES=1.455), and Myc targets (NES=1.441). Next, to investigate this on an individual sample level, we calculated the percent spliced out (PSO) for each significant nonPTC 3UI event in each sample, including normal samples. The average PSO for each nonPTC 3UI event in normal samples was then deducted from the PSO of each cancer sample for each event, producing a ΔPSO value. These ΔPSO values were subsequently clustered via K-means clustering, resulting in four clusters of events (Figure 4.2A). Clusters A and C were predominantly made up of nonPTC 3UI splicing events that displayed more intron retention (negative ΔPSO) in the cancer samples compared to normal samples. Cluster B contained nonPTC 3UI splicing events that were spliced more in cancer samples (positive ΔPSO) compared to normal samples. Cluster D contained nonPTC 3UI splicing events that were both spliced more and retained more.

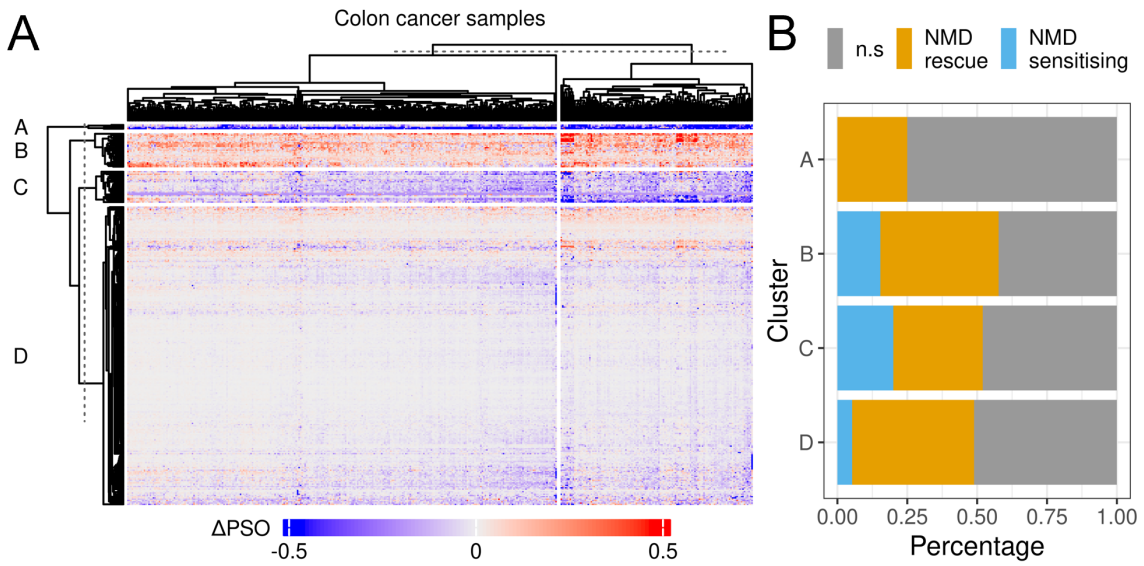

Figure 4.2: Comparing significant differentially spliced events with NMD sensitivity. A) Heatmap of events which are spliced significantly more or less between normal and cancer samples. Each row represents an event. Each column represents a colon cancer sample. The color represents the ΔPSO value (PSO for that event in that sample minus the average PSO for that event in all normal samples). A negative ΔPSO value indicates more 3'UTR intron retention in cancer compared to normal samples. A positive ΔPSO value indicates more 3'UTR intron splicing in cancer compared to normal samples. B) Percentage of events from each cluster in A that are regulated by UPF1 knockdown in HCT116 cells. Events are classed as: NMD sensitising if they are significantly upregulated upon UPF1 knockdown; NMD rescusing if they are significantly downregulated upon UPF1 knockdown; or n.s. if FDR>0.1.

Next, we compared the events in each cluster to the results of UPF1 knockdown (Section 3.4.2.1) to investigate whether the differences in PSO observed were primarily due to differential regulation by NMD. We hypothesised that increased 3UI retention (as observed in clusters A and C) may be acting as a mechanism to evade NMD where 3UI splicing is NMD-sensitising. Whilst increased 3UI splicing (as observed in cluster B) may be acting as a mechanism to rescue transcripts from NMD, where 3UI splicing is "NMD-rescuing". Whilst we expected to see enrichment of NMD-sensitising events in clusters A and C, we found that for cluster A most events were not significantly regulated by UPF1 knockdown, and those that were resulted in downregulation upon UPF1 knockdown (NMD rescue; Figure 4.2B). Additionally, for events in cluster C, whilst we

observed NMD-sensitising events, we observed more NMD-rescuing events (Figure 4.2B). Our "NMD rescue" classification is imparted by splicing, therefore increased intron retention of these cluster A and C events leads to the production of more NMD-sensitive isoforms in cancer compared to normal.

Conversely, and in line with our hypothesis, for events in cluster B (those spliced more in cancer) we observed selection of NMD-rescuing events (Figure 4.2B), which suggests that the increased level of splicing of these events leads a reduction in NMD sensitivity, perhaps as a mechanism to reinforce expression. However, roughly 15% of events within this cluster were also NMD-sensitising. Together these results suggest that the regulation of NMD sensitivity may partially explain some of the nonPTC 3UI splicing differences observed between normal and cancer samples; however, additional regulatory pathways are likely to be involved in such cancer-related changes in 3UI splicing.

Of the 340 nonPTC 3UI splicing events that were displayed significant differential splicing between normal and colorectal cancer samples, 67 of these displayed an absolute change in intron retention greater than 5% (Figure 4.3A; highlighted in orange), whilst 24 of these displayed an absolute change greater than 10% (Figure 4.3B). Of these 24 nonPTC 3UI events, 10 were spliced more in colon cancer, whilst 14 were retained more in colon cancer, compared to normal samples. Of the 10 genes that contain nonPTC 3UIs that are spliced more in cancer, seven of them displayed an upregulation in gene expression in colon cancer compared to normal samples (FBXL8, TAF1D, CTNNB1, KIFC2, LRATD2, SOD2, CPNE1), whilst three were downregulated (ACADVL, CASP1, NRBP1). Whilst of the 12 genes that contain the 14 nonPTC 3UIs that are retained more in colon cancer (SRSF2 and RPS24 have two nonPTC 3UIs each), eight of these were upregulated (MESD, RPS24, ALDH16A1, RPS3, SRSF2, CBX5, PABPC1, MARK3), whilst four were downregulated (ID3, UQCR10, HUS1, GDA) in colon cancer compared to normal samples.

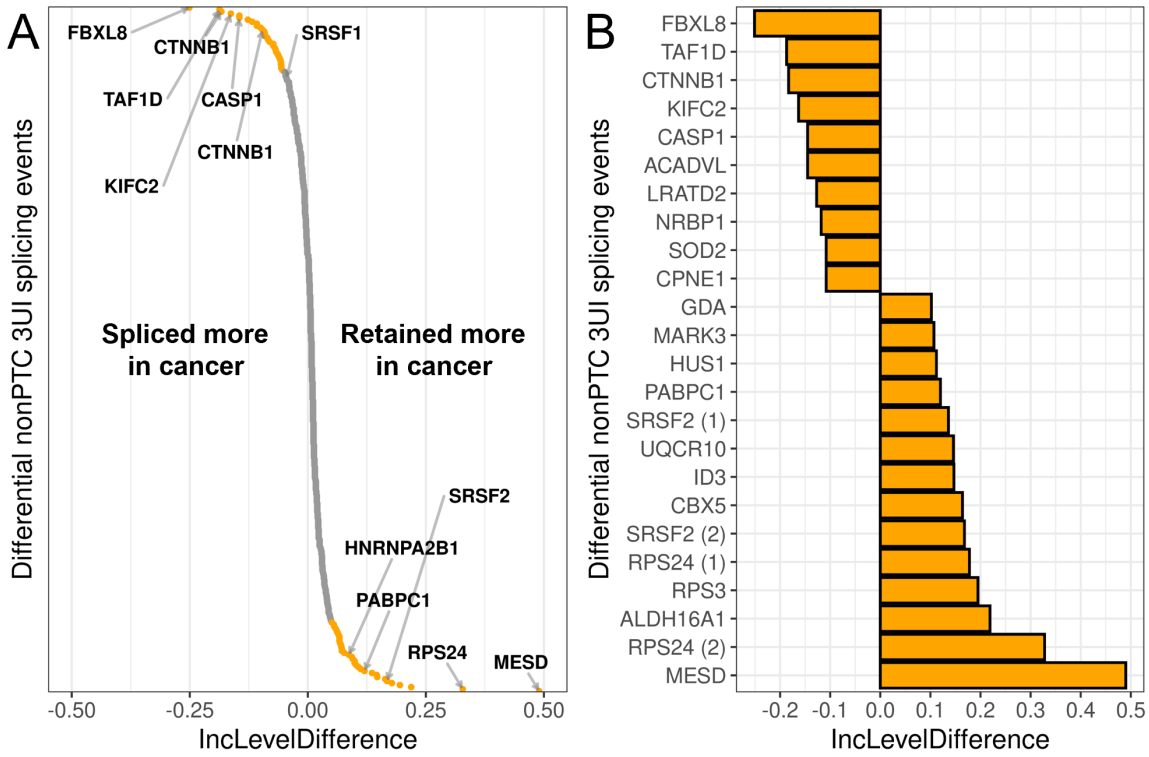

Figure 4.3: Significant differentially spliced nonPTC 3UIs in colorectal carcinoma. IncLevelDifference is the level of intron inclusion in cancer samples minus the level of intron inclusion in normal samples. Therefore a negative IncLevelDifference means the 3UI is spliced more in cancer compared to normal, whilst a positive IncLevelDifference means more intron retention is observed in cancer compared to normal. A) All significant (FDR<0.05) differentially spliced nonPTC 3UI events. Dots are coloured orange if the absolute IncLevelDifference $\geq$ 0.1, or grey if <0.1. B) Significant (FDR<0.05) differentially spliced nonPTC 3UI events with an absolute IncLevelDifference $\geq$ 0.1.

Interestingly, CTNNB1 contains one of the most over-spliced nonPTC 3UIs in colon cancer. This is important in the context of colon cancer, as CTNNB1 is the central regulator of the Wnt signalling pathway, which was previously found to be altered in roughly 93% of the TCGA colon cancer samples (The Cancer Genome Atlas Network, 2012). The major sources of altered Wnt signalling in these samples was attributed to to inactivation of APC, or over-activation of CTNNB1 (The Cancer Genome Atlas Network, 2012). We annotated each sample with its CTNNB1 and APC mutation status, and performed K-means clustering of nonPTC 3UI PSO values in line with the analysis conducted in Figure 4.2, to determine

whether the splicing signatures observed would cluster by mutation status. However, this was not found to be the case.

Whilst the CTNNB1 long 3UI is the most over-spliced, the short 3UI is also spliced more in colon cancer compared to normal samples, as demonstrated by the density plots in Figure 4.4A. Additionally, the short 3UI spliced isoform represents the major isoform in more samples than the long 3UI spliced isoform. Given the over-splicing of CTNNB1 observed in colon cancer, we hypothesised that other Wnt pathway components may also be over-spliced. To test this we ranked all our nonPTC 3UI RMATS outputs by IncLevelDifference, producing a list of genes ranked from most over-spliced to most over-retained in colon cancer. We then conducted GSEA against the "GOBP_CANONICAL_WNT_SIGNALING_PATHWAY" C5 ontology set, which contains genes that make up the canonical Wnt signaling pathway, and observed significant enrichment (NES=1.678; FDR=0.001; Figure 4.4B).

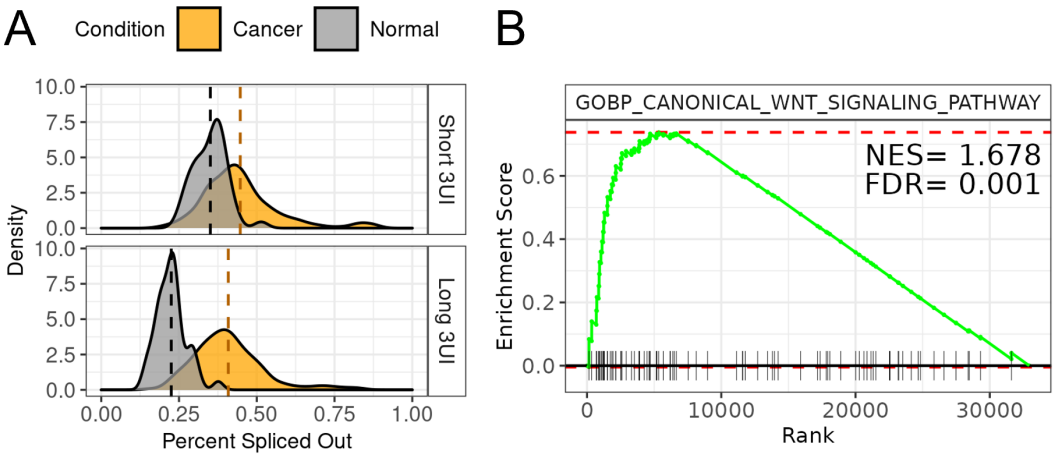

Figure 4.4: Splicing of Wnt signalling component 3UIs is dysregulated in colorectal carcinoma. A) Splicing of the CTNNB1 3'UTR is increased in colorectal carcinoma compared to normal colon samples, for both 3UI isoforms. Dashed lines represent median percent spliced out values. B) Gene set enrichment analysis of all nonPTC 3UI events ranked by IncLevelDifference from most over-spliced in cancer to most over-retained in cancer shows enrichment of the GO:BP canonical Wnt signalling pathway.

4.1.2 Manipulating the Wnt signalling pathway alters 3'UTR splicing in colorectal carcinoma cells

Given the over-splicing of Wnt signalling pathway components in colon cancer, including central regulator CTNNB1, we hypothesised that this may be in part be due to hyperactive Wnt signalling. To test this we exposed HCT116 cells to increasing concentrations of CHIR99021 over 24 hours. CHIR99021 is a GSK3β inhibitor, the presence of which prevents degradation of CTNNB1, leading to accumulation within the nucleus, and subsequent activation of the Wnt signalling pathway. We found that increasing CHIR99021 concentrations led to a significant decrease in expression of the CTNNB1 short 3UI-spliced isoform ($R=-0.77$, $p<0.001$), whilst no significant correlation was observed for the 3UI-retaining isoform (Figure 4.5). This suggests that increasing levels of CHIR99021 may be leading to isoform-specific regulation of the short 3UI-spliced isoform, as opposed to differences in splicing, where we would expect to see the level of 3UI retention increase in a compensatory manner. To confirm that this was due to Wnt signalling and not a secondary effect of GSK3β inhibition, we next exposed HCT116 cells to increasing concentrations of Wnt pathway inhibitor IWR-1 over 48 hours. IWR-1 stabilizes Axin, a scaffold protein central to the CTNNB1 degradation complex (which also contains GSK3β), as such CTNNB1 is degraded more, reducing levels of Wnt signalling. Here we find the opposite trend compared to Wnt activation, where increasing the IWR-1 concentration leads to a significant increase in CTNNB1 short 3UI-spliced isoform expression ($R=0.65$; $p=0.002$) in an isoform-specific manner (Figure 4.5). Unfortunately we were unable to measure expression of the CTNNB1 long 3UI-spliced isoform via qPCR due to similarity between the sequence immediately downstream of the long 3UI 3'SS, and the sequence immediately downstream of the shared 5'SS. Where qPCR primers were tested for the long 3UI-spliced isoform, they were also able to amplify from unspliced plasmid DNA. This was not the case for the short 3UI-spliced isoform-specific primers, which were successfully validated (Figure 3.7C).

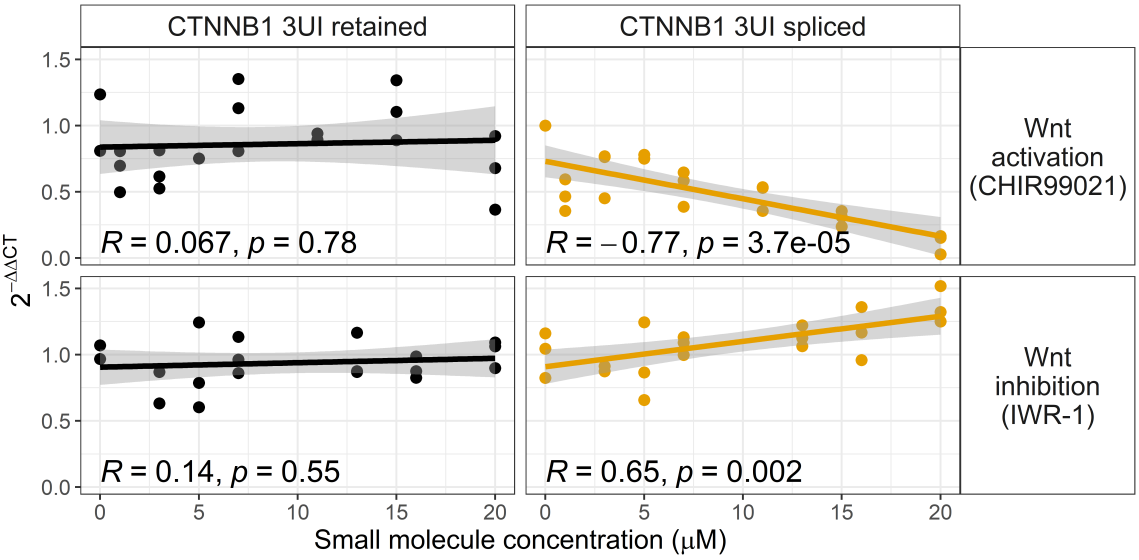

Figure 4.5: Wnt signalling pathway manipulation in HCT116 cells alters expression of CTNNB1 3UI spliced isoform Treatment of HCT116 cells with increasing concentrations of CHIR99021 (Wnt signalling pathway activator) and IWR-1 (Wnt signalling pathway inhibitor) lead to isoform-specific regulation of the CTNNB1 3UI spliced isoform. Spearman's rank correlation coefficient (R) was calculated to test the strength of the relationships. P<0.05 was considered statistically significant.

To test the effects of Wnt pathway activation on the splicing of other Wnt pathway component 3'UTRs, we treated HCT116 cells with either 0μM or 20μM CHIR99021 for 24 hours, and conducted RNAseq. We next conducted DEU analysis with RMATS to investigate the global trend in 3'UTR splicing between CHIR99021 treated cells and control cells. We found that roughly equal proportions of nonPTC 3UIs were spliced more or spliced less in CHIR99021-treated HCT116 cells compared to those treated with 0μM, whilst pPTC 3UIs were retained more in the majority of instances (Figure 4.6; right facet). This is partially comparable with the results of DEU analysis conducted between normal colon and colon cancer TCGA samples (Figure 4.6; left facet).

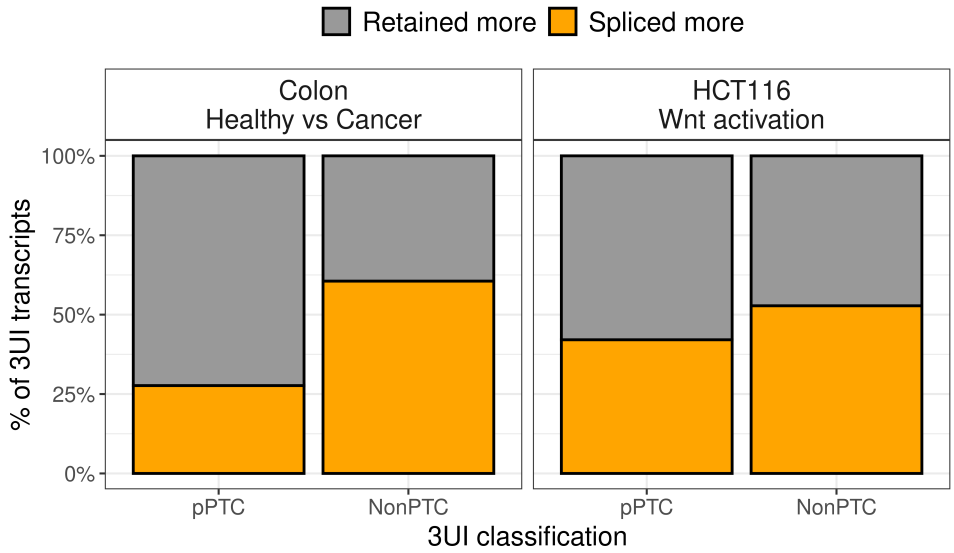

Figure 4.6: Differential exon usage analysis compared between TCGA and Wnt hyperactivation in HCT116 cells Left facet = Differential exon usage analysis for normal vs colorectal carcinoma samples from TCGA. Right facet = Differential exon usage analysis for untreated HCT116 cells vs HCT116 cells cultured with 20µM CHIR99021 (Wnt signalling pathway activator). The amount of the bar filled represents the proportion of events that were spliced more in cancer or with Wnt treatment (orange), or were retained more (grey).

As was the case for normal colon vs colon cancer TCGA samples, GSEA conducted on differentially spliced nonPTC 3UIs following CHIR99201 activation in HCT116 cells revealed significant enrichment of Wnt signalling pathway components (Figure 4.7A). Interestingly, the long 3UI-spliced isoform of CTNNB1 was found to be one of the most differentially spliced Wnt pathway components upon Wnt activation (Figure 4.7B). A breakdown of the Wnt pathway components, alongside whether they were retained more or spliced more upon CHIR99021 treatment is depicted in Figure 4.7C. Visualization in this manner highlights that there does not appear to be a skew towards increased splicing or increased retention for Wnt pathway members that are either activators or inhibitors of the pathway.

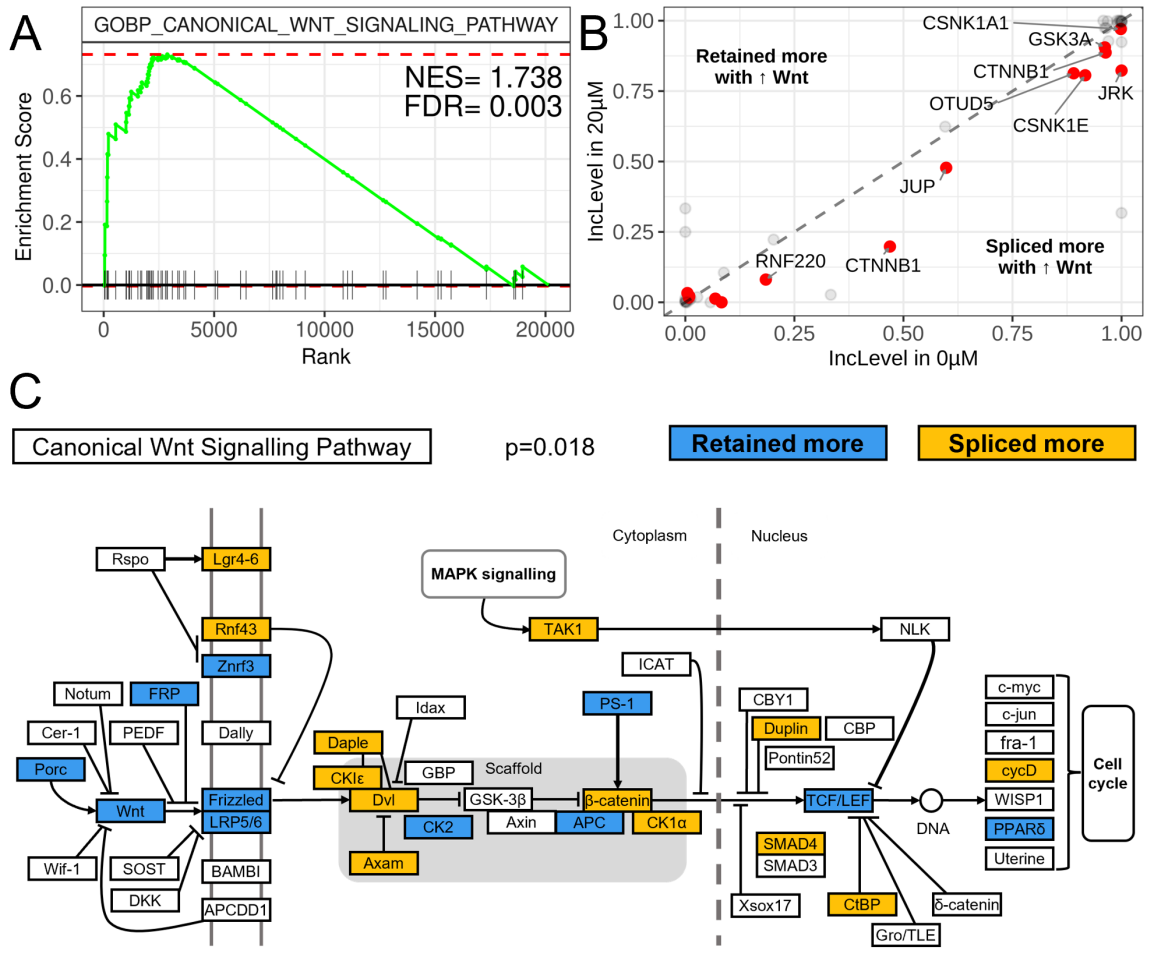

Figure 4.7: Wnt signalling pathway manipulation in HCT116 cells alters splicing of Wnt component 3UIs A) Gene set enrichment analysis of all nonPTC 3UI events ranked by IncLevelDifference from most over-spliced in CHIR99021 treated vs untreated, to most over-retained, shows enrichment of GO:BP canonical Wnt signalling pathway. B) Dot plot comparing IncLevelDifference for canonical Wnt signalling pathway components in untreated (X-axis) vs CHIR99021-treated (Y-axis) HCT116 cells. Grey dots represent non-significant events. Red dots represent significant events (FDR<0.05). C) Pathway plot showing distribution of 3UI-containing transcripts throughout the KEGG canonical Wnt signalling pathway alongside whether they are spliced more (orange) or retained more (blue) upon Wnt signalling activation. The P-value displayed represents significance of enrichment of the KEGG Wnt signalling pathway via gene ontology analysis.

4.2 Alternative splicing of 3'UTRs is observed in stem cells

In the second section of this chapter, we turn away from the cancer setting and here aim to determine the extent of alternative splicing of 3'UTRs in stem cells and cell differentiation. Firstly, we test whether 3'UTRs are subject to alternative splicing during differentiation by using existing data from Strober et al. (2019). In the Strober et al. (2019) study, 19 hiPSC cell lines were differentiated in cardiomyocytes over 15 days, with RNAseq conducted at each day. Therefore this dataset has both a large sample size, and a high temporal resolution. In addition, at the time of analysis, cardiomyocyte differentiation was actively being performed by Theodore Wing in the Barbaric lab, thus facilitating the establishment of our own model to further investigate any findings of our analysis of this dataset. Finally, we investigate how 3'UTR splicing varies in disease outside of the cancer setting, by using RNAseq data from patient-derived iPSCs from the HipSci consortium.

4.2.1 Differential 3UI splicing during cardiomyocyte differentiation

We quantified raw RNAseq reads from Strober et al. (2019) against our HipSci transcriptome assembly and performed differential transcript usage analysis (DTU) to interrogate whether 3'UTR-spliced isoforms were used more or used less as differentiation proceeds. We found that 180 3'UTR spliced transcripts (from 161 genes) displayed significant DTU during differentiation. Using the more sensitive DEU approach (as in Section 4.1) between day 0 and day 15 we detected 219 significant events (from 128 genes). We next performed K-means clustering of the fraction expression of each isoform that displayed significant DTU, both between events (rows in Figure 4.8), and between samples (columns in Figure 4.8). Firstly, we found that samples clustered by their differentiation stage. As such this heatmap can be viewed as progressing in differentiation by scanning between each sample cluster from left to right. We also found two distinct clusters of 3UI-containing transcripts in Figure 4.8, the top cluster (becoming a darker red from left to right) contains 106 transcripts that are used more as

differentiation progresses. Meanwhile, the bottom cluster (becoming a darker blue from left to right) contains 74 transcripts that are used less as differentiation progresses.

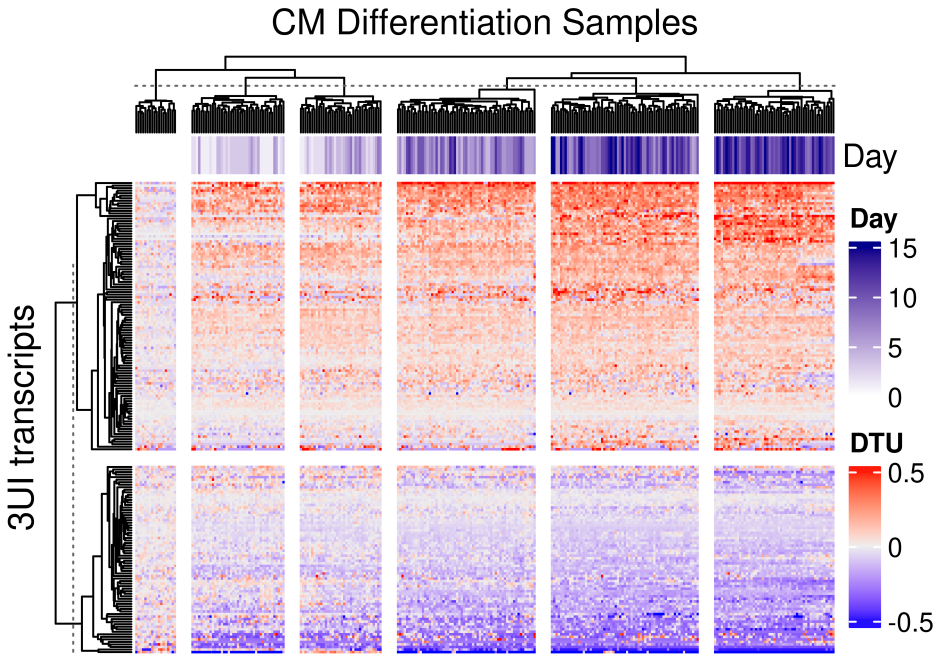

Figure 4.8: Differential transcript usage of 3UI-containing transcripts during cardiomyocyte differentiation Heatmap showing the trends of transcript usage for 3UI-containing transcripts that display significant differential transcript usage (adjusted P-value<0.05). Each row represents a transcript. Each column represents an RNAseq sample from the Strober hiPSC cohort (Strober et al., 2019), colour coded by day of differentiation (white = day 0, dark blue = day 15). Each cell represents the fraction expression (transcript expression / gene expression) for each transcript in each sample, minus the average fraction expression for all day 0 samples for that transcript. Therefore the colour represents whether the transcript is used more (red) or used less (blue) as differentiation proceeds.

Cardiomyocyte differentiation is conducted through manipulation of the Wnt signalling pathway, therefore we investigated whether any Wnt pathway components were within either cluster. Interestingly, we found that the short 3UI-spliced isoform of CTNNB1 (central Wnt regulator) was in the second cluster and was therefore used less as differentiation proceeds. Expanding on this, we conducted gene ontology analysis of the genes found in each cluster, and tested for enrichment of the KEGG Wnt signalling pathway, representing the core Wnt pathway components. We found significant

enrichment amongst 3'UTR spliced transcripts that are used more as differentiation proceeds (P=0.008), although this was not significant amongst those used less as differentiation proceeds (P=0.0765). Turning to those regulated in any capacity, i.e. any that displayed significant DTU in either direction, we observed significant enrichment of the KEGG Wnt signalling pathway (DE/totalInCategory = 6/150; P=0.0009). This enrichment represented only six genes (CTNNB1, WNT3, PLCB3, CSNK1E, CSNK2A2, PRKX) out of a total 161 genes that displayed DTU (whilst there were 180 3'UTR spliced transcripts showing DTU, some of these were found within the same gene). However, this still represented a statistically significant over-representation of Wnt pathway genes compared to the expected number across all genes. We also observed a significant enrichment of the larger "GOBP Canonical Wnt Signaling Pathway" ontology (DE/totalInCategory = 8/330; P=0.006). We next widened our approach to be more hypothesis-generating as opposed to hypothesis-driven, and asked whether any other pathways or gene ontology terms were enriched within this dataset; however, following P-value adjustment, no significant enrichment was found.

Given that CTNNB1 has two possible 3UIs, as well as a 3UI-retaining isoform, we next asked what happens to each of these isoforms. As depicted in Figure 4.9A+B, as differentiation proceeds there is an increased usage of the 3UI-retaining isoform, whilst the usage of the short 3UI-spliced isoform is reduced. The usage of the long 3UI-spliced transcript appears to differ between samples, including those of the same differentiation stage, and does not display a selection towards increased splicing or increased retention as differentiation proceeds (Figure 4.9A).

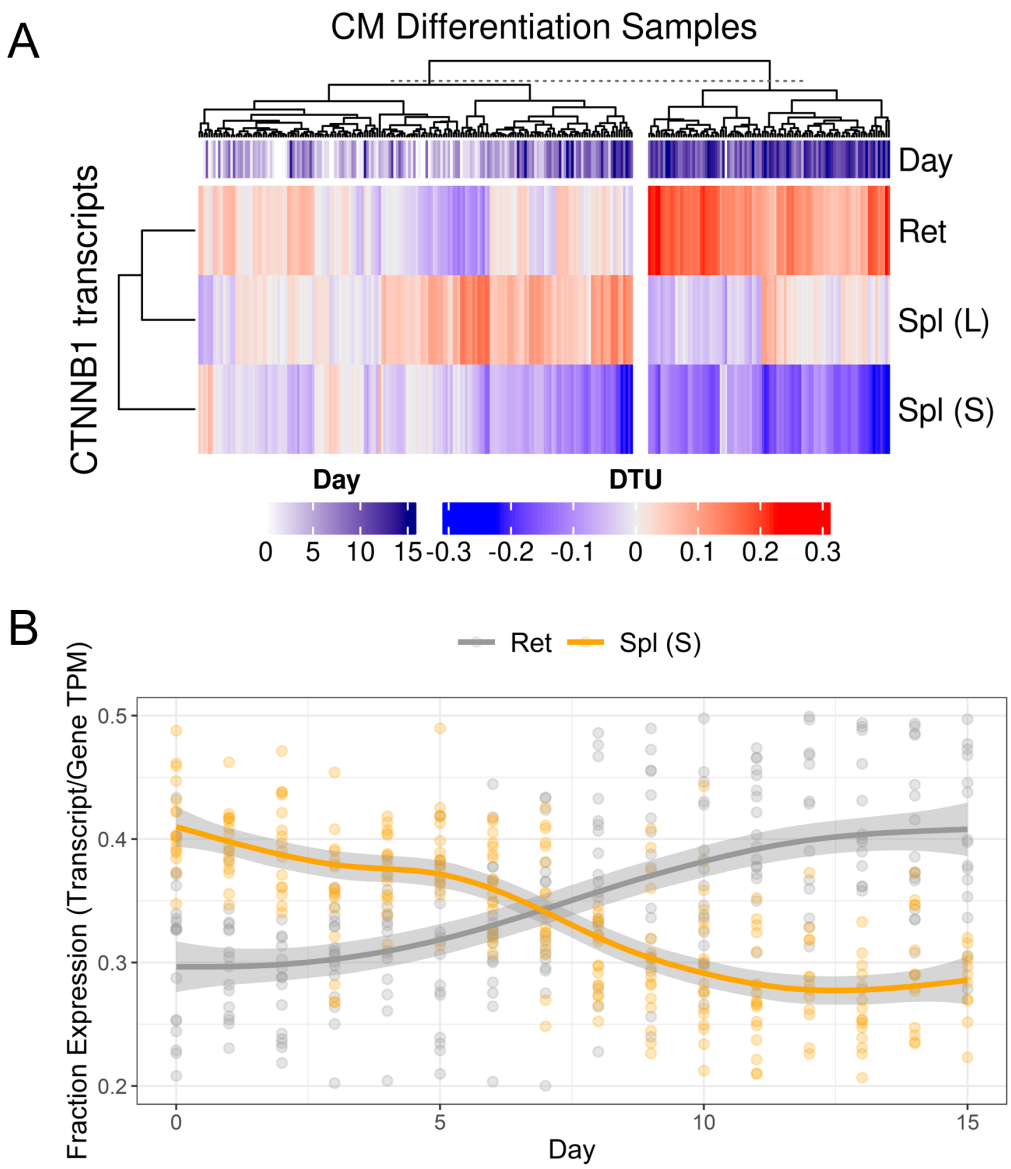

Figure 4.9: Differential transcript usage of CTNNB1 3UI isoforms during cardiomyocyte differentiation A) Heatmap showing the trends of CTNNB1 transcript usage over differentiation. Each row represents a CTNNB1 3UI isoform. Each column represents an RNAseq sample from the Strober hiPSC cohort(Strober et al., 2019), colour coded by day of differentiation (white = day 0, dark blue = day 15). Each cell represents the fraction expression (transcript expression / gene expression) for each isoform in each sample, minus the average fraction expression for all day 0 samples for that isoform. Therefore the colour represents whether the transcript is used more (red) or used less (blue) as differentiation proceeds. B) Line graph of CTNNB1 3UI-retaining and 3UI-spliced (short intron) isoforms over differentiation.

4.2.2 Dysregulation of 3'UTR splicing in HipSci patient cohorts

To gain insight into whether 3'UTR splicing may be dysregulated in diseases outside of the cancer setting, we utilized RNAseq data from iPSCs generated from patient cohorts collected by HipSci. Samples were quantified against our HipSci transcriptome assembly, and DEU analysis was conducted with RMATS to compare each cohort to the "normals" cohort from which the HipSci transcriptome assembly was initially built (Figure 4.10).

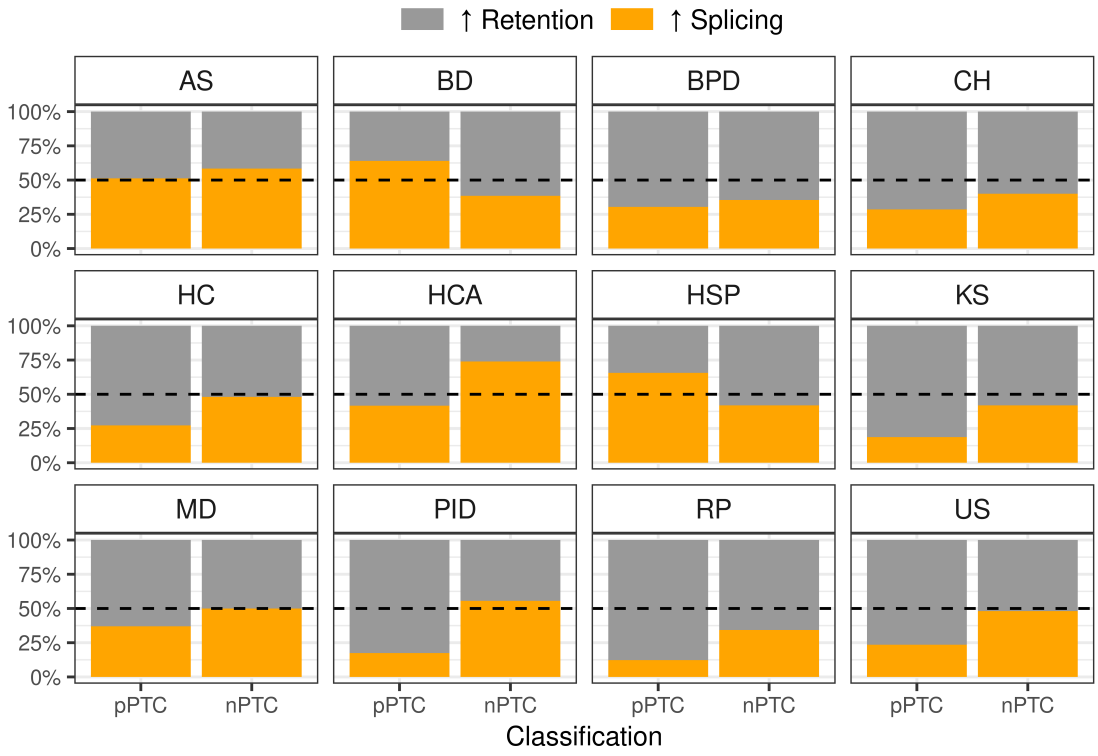

Figure 4.10: Dysregulation of 3'UTR splicing in HipSci patient cohorts Percentage of significant differentially spliced events (FDR<0.05; absolute IncLevelDifference > 0.05) that are spliced more (orange) or retained more (grey) in each cohort (each facet) compared to normals. Cohorts: AS = Alport Syndrome; BD = Batten Disease; BPD = Bleeding and Platelet Disorders; CH = Congenital Hyperinsulinism; HCA = Hereditary Cerebellar Ataxia; HSP = Hereditary Spastic Paraplegia; HC = Hypertrophic Cardiomyopathy; KS = Kabuki Syndrome; MD = Macular Dystrophy; PID = Primary Immune Deficiency; RP = Retinitis Pigmentosa; US = Usher Syndrome.

Figure 4.10 depicts the percentage of significant differential splicing events (FDR<0.05; effect size > 5%) that are spliced more or retained more in each cohort compared to normals. Once again we observed different patterns between pPTC and nonPTC 3UIs, across all cohorts. Increased nonPTC 3UI splicing was observed in Alport Syndrome, Hereditary Cerebellar Ataxia and Primary Immune Deficiency cohorts compared to normals. Whilst for Batten Disease, Bleeding and Platelet Disorders, Congenital Hyperinsulinism, Hereditary Spastic Paraplegia, Kabuki Syndrome and Retinitis Pigmentosa, nonPTC 3UIs tend to be retained more.

4.3 Summary

In this chapter we have shown that 3'UTR splicing is dysregulated in many cancer types. The finding that pPTC 3UIs are retained more in cancer is consistent with existing literature (Dvinge and Bradley, 2015), given that these introns overlap coding domain introns, and would be classic "retained introns". In contrast we found that nonPTC 3UIs are generally spliced more in cancer compared to matched normals, suggesting that these introns are regulated differently from the pPTC 3UIs. By focusing on colon cancer, and comparing the usage of 3'UTR spliced isoforms to their change in expression upon UPF1 knockdown in HCT116 cells, we found that of the transcripts that are spliced more in cancer compared to normals, more are downregulated upon UPF1 knockdown than are upregulated, suggesting that splicing these introns reduces NMD sensitivity. Whilst transcripts that show an increased level of 3'UTR intron retention are more likely to be NMD sensitive. It is important to note that the predicted NMD sensitivity in this comparison is based upon UPF1 knockdown in HCT116 cells. NMD efficiency is known to vary between cell types (Linde et al., 2007), and even cells within the same population (Sato and Singer, 2021), therefore confirmation of predicted NMD sensitivity by knockdown of UPF1 in multiple colon cancer cell lines would ideally be conducted in the future to validate these findings.

We found that an intron within the 3'UTR of CTNNB1, the central regulator of the Wnt signalling pathway, was one of the most over-spliced 3UIs in colon cancer compared to normal colon samples, which led us to investigate other Wnt components. We found that many Wnt pathway components contain 3UIs, and that the pathway as a whole was significantly enriched for 3'UTR splicing in colon cancer compared to normal. Given that the colon cancer samples from TCGA often contain mutations to the Wnt pathway (The Cancer Genome Atlas Network, 2012), we manipulated the Wnt pathway in HCT116 cells to investigate the effect on 3'UTR spliced isoforms. However, it is important to note that Wnt activation with CHIR99201 was conducted in a cancer cell line, which may have already undergone such Wnt-dependent changes in 3'UTR splicing, as such we may just be "pushing these further". We observed Wnt-mediated regulation of 3UI-spliced isoform expression, including CTNNB1, and significant enrichment of the Wnt pathway again. This suggests that 3'UTR splicing of Wnt pathway components may function as part of an auto-regulatory mechanism to control expression, or expression dynamics, of its component upon Wnt signalling activation, at least in the colon cancer setting. To validate these findings, an antagonistic approach should be taken on the transcriptome-wide scale, as was conducted for the CTNNB1 example, which was exposed to both CHIR99021 (activator) and IWR-1 (inhibitor). Alternatively, an endogenously expressed activator such as Wnt3A could have be used instead of, or in addition to, CHIR99021 (a small molecule GSK3β-inhibitor) to better reflect the endogenous biological context.

We also investigated 3'UTR splicing in the stem cell setting by using cardiomyocyte differentiation data from Strober et al. (2019). We found that 3'UTR spliced isoforms are subject to differential transcript usage during differentiation, where two distinct subsets of transcripts were progressively used more or used less as differentiation proceeded. CTNNB1 was one of the latter transcripts, which is interesting given the importance of Wnt signalling in cardiomyocyte differentiation. This highlights the importance of 3'UTR splicing outside of the cancer context. Here we show that the cells have specifically selected for, or selected against, preferential expression of 3UI-spliced transcripts, further

supporting our claim that 3'UTR splicing is more than transcriptional noise. We previously showed that 3'UTR splicing in stem cells can elicit transcript degradation by NMD. As such, alternative splicing of 3'UTRs may represent a method to modulate NMD sensitivity during differentiation through AS-NMD (Section 1.5.3). This is addressed further in Chapters 5 and 6 where we establish our own cardiomyocyte differentiation model and conduct a global RNA stability assay, matching isoform-level half-life estimates to predicted NMD sensitivity through a differentiation time course experiment.

Finally, we showed that 3'UTR splicing is dysregulated in patient cohorts from the HipSci consortium. This represents a very preliminary inquiry into the role of 3'UTR splicing in disease outside of the cancer context. It is important to note that in this dataset RNAseq was conducted on iPSCs derived from skin biopsies of patients within each cohort. As such the results might not reflect those observed in the specific tissue of interest. For example, it may be more relevant to study hypertrophic cardiomyopathy in cardiomyocytes as opposed to iPSCs. In this regard, the cell lines used to generate this data are available to purchase from HipSci upon request which, in combination with directed cell differentiation into the tissues of interest, facilitates future study into this phenomena.

5. Global RNA stability changes during stem cell differentiation

In Chapter 4 we showed that many 3'UTR spliced transcripts displayed differential transcript usage during cardiomyocyte differentiation by using data from Strober et al. (2019). In order to gain a more mechanistic insight into the post-transcriptional regulatory effects of 3'UTR splicing, we set out to conduct a global RNA stability assay (SLAMseq) at the transcript level. Whilst the results of such experiments are presented in Chapter 6, in order to achieve this, this chapter describes how we established a cardiomyocyte differentiation model (Section 5.1), ensured it was compatible with SLAMseq, and developed bioinformatic tools to analyse the data in this format (Section 5.2). In addition, and prior to addressing changes in stability caused by 3'UTR splicing across differentiation, this chapter investigates global gene-level changes in RNA stability across cardiomyocyte differentiation (Section 5.3).

Understanding how gene expression changes during hESC differentiation is essential to study the underlying regulatory mechanisms governing self-renewal vs cell fate determination. However, this is often achieved by comparing steady-state expression in different cell types. Whilst this informs us of the changes that have occurred in cell type A vs cell type B, steady-state approaches do not provide information on how quickly the transcriptome changes. Understanding the rate of RNA turnover is highly important, especially in pluripotent cells such as hESCs, and multipotent cell populations such as early mesodermal cells, which are capable of differentiating into all (for hESCs) or many (for mesoderm) cell types within the body, depending on the signalling factors they encounter. By studying this we can address how quickly cell fate decisions are made during differentiation, and the speed with this these decisions are propagated across the transcriptome. We hypothesised that turnover rate is likely to change during differentiation, potentially matching the "decision burden" of the cells at each stage of differentiation. To address this, we utilized SLAMseq across a cardiomyocyte

differentiation time course. SLAMseq is metabolic labelling technique developed by Herzog et al. (2017). In SLAMseq 4-Thiouridine (4sU) is incorporated into RNA in the place of uracil, which upon RNA extraction, iodoacetamide treatment, and sequencing, allows us to distinguish old RNA from new RNA via the detection of T>C conversions (Figure 5.1A). Subsequently, we can calculate RNA decay rates (Figure 5.1B).

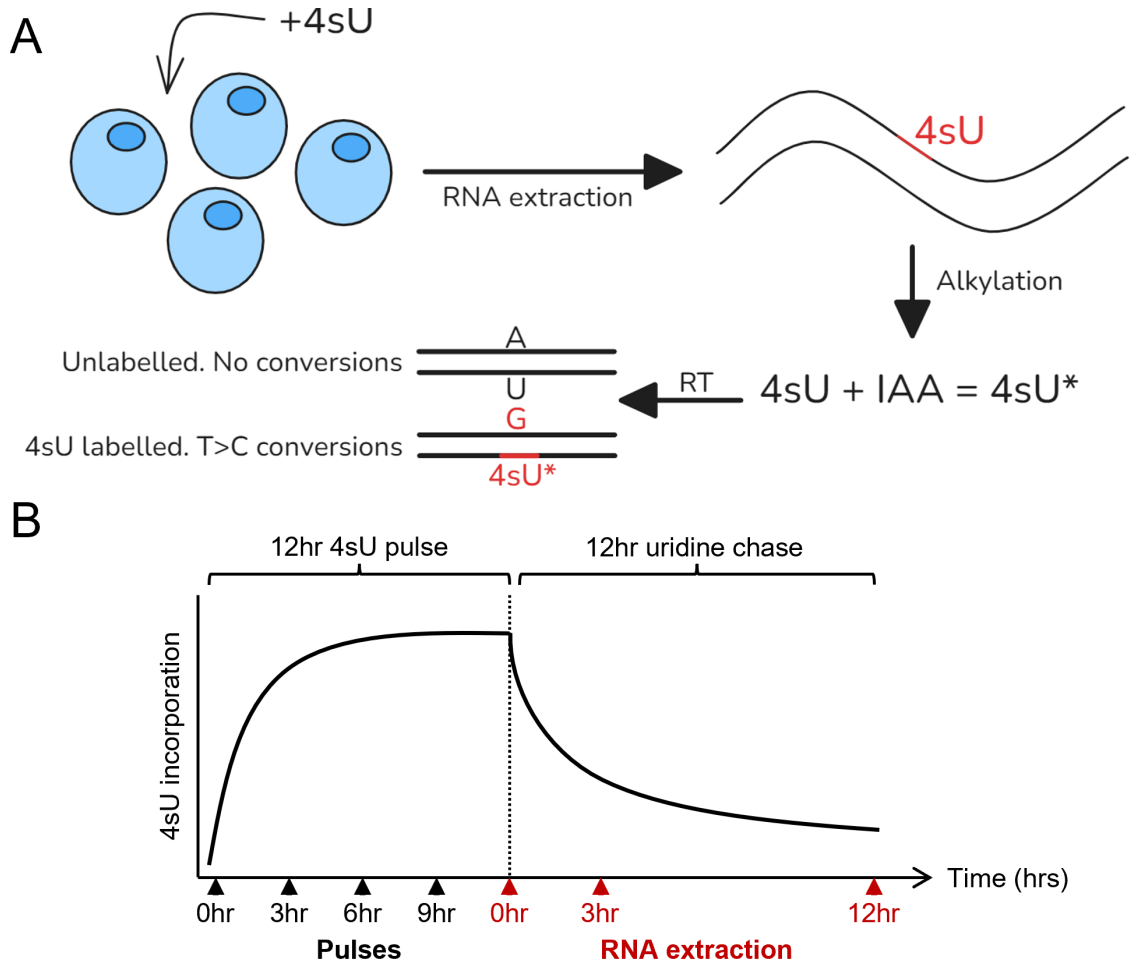

Figure 5.1: Overview of SLAMseq. A) 4sU is supplemented into culture media and incorporates into RNA in place of uricil, generating a pool of labelled and unlabelled RNA. Upon RNA extraction, RNA is treated with iodoacetamide (IAA), alkylating the 4sU thiol group. Upon reverse transcription (RT) during cDNA library preparation, the presence of 4sU leads to T>C conversions which can be detected within the sequencing data. B) 4sU is pulsed for 12 hours, with media changes every three hours. Following the 4sU pulse, RNA is extracted at various time points, treated, and sequenced, allowing decay rates to be calculated.

5.1 Establishing a cardiomyocyte differentiation model with H9 hESCs

Cardiomyocyte differentiation was considered for several reasons. First, we find that 3UIs are enriched in Wnt signaling pathway components (Figure 3.4B), and Wnt signaling is manipulated during cardiomyocyte differentiation. Second, we had previously conducted DTU analysis using the Strober et al. (2019) hiPSC differentiation time course dataset. Third, there are multiple established cardiomyocyte differentiation protocols (Lyra-Leite et al., 2022), including a microcarrier-based approach (Ting et al., 2014) that was actively being explored in the Barbaric lab by Theo Wing. Finally, cardiomyocyte differentiation is relatively fast (8-15 days) compared to other cell types (e.g. functional motor neurons; roughly 28 days; Yang et al. (2023)), and the success of differentiation can be easily observed through the presence of beating cells prior to flow cytometry and transcriptomic interrogation.

5.1.1 Suspension culture on microcarriers

Our initial attempts at cardiomyocyte differentiation utilised a microcarrier-based protocol (Ting et al., 2014; Laco et al., 2020). H9 hESCs were seeded onto and expanded on vitronectin-coated microcarrier beads for up to one week prior to differentiation. Cardiomyocyte differentiation was then conducted through manipulation of the Wnt signalling pathway. First, aggregates were treated with Wnt agonist CHIR99021 for 48 hours in RPMI+B27, followed by unsupplemented RPMI+B27 for a further 48 hours. Media was subsequently changed to RPMI+B27 supplemented with Wnt antagonist IWR-1 for 48 hours, followed by unsupplemented RPMI+B27 for a further 48 hours. After eight days, beating was observed in some aggregates; however, the amount of beating between aggregates was not consistent, and varied greatly depending on aggregation density and clump size. Whilst previous studies had shown that this protocol produced >80% CTNT+ cells, flow cytometry of cells differentiated in our hands yielded 30-50%

CTNT+ cells, with a high degree of variability between biological replicates, primarily based on aggregate size and density. As such we turned to 2D approaches.

5.1.2 2D differentiation

We next turned to a 2D approach described by Burridge et al. (2014), which utilized a minimal chemically defined media, CDM3, consisting of RPMI 1640, L-ascorbic acid 2-phosphate, and human albumin. The theory behind this differentiation was similar to that considered in Section 5.1.1, and revolved around the growth of H9 hESCs on vitronectin coated plates, followed by manipulation of the Wnt signalling pathway with CHIR99021 and IWR-1. However, in our hands we observed widespread cell death by day 4 of differentiation, which was not resolved through optimization of seeding density or CHIR99021 concentration. It is important to note that these differentiation attempts coincided with high levels of batch-to-batch variability in vitronectin quality experienced by the lab. This led us to turn to a Matrigel-based approach.

H9 hESCs were seeded into Matrigel-coated 12-well plates and differentiated using the STEMCELL Technologies Ventricular Cardiomyocyte Differentiation Kit (Figure 5.2A; for more detailed information related to differentiation, see Materials & Methods Section 2.5.1.). An important difference between this protocol and previously described 2D differentiations (e.g. Burridge et al. (2014)) is the addition of extra Matrigel at the beginning of differentiation, to create a "matrigel sandwich" of sorts (Figure 5.2B). Using this kit, some beating was observed starting at day 8 of differentiation, but was more profound and consistent between biological replicates by day 10 of differentiation.

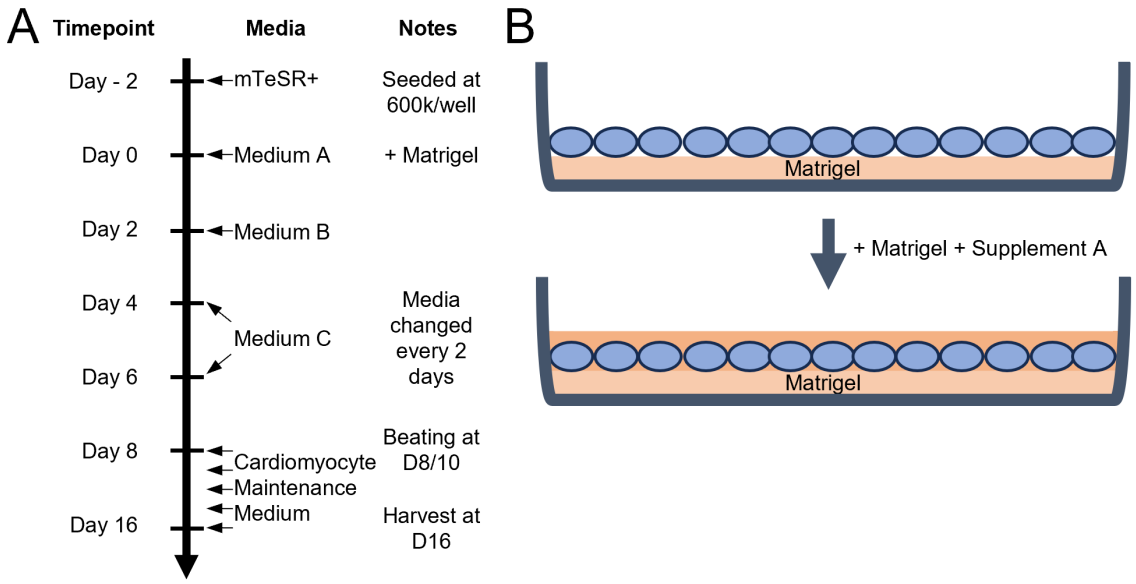

Figure 5.2: Schematic of 2D differentiation protocol. A) Flow diagram of 2D differentiation protocol, showing the media change steps, alongside comments. A detailed breakdown can be found in Section 2.5.1. B) Schematic of "matrigel sandwich" style approach, blue circles represent hESCs.

At day 16 of differentiation, cells were dissociated from the culture well by treatment with TryplE for 30 minutes. Whilst this should have been sufficient for single cell dissociation, clumps were still present, and therefore had to be filtered with a cell strainer to prevent blockage of the flow cytometer. Flow cytometry indicated that our population of analyzed cells were on average 57% CTNT-positive (Figure 5.3A-C). However, if a disproportionate amount of cardiomyocytes were present within clumps, then the level of CTNT+ cells relative to total cells analyzed by flow cytometry may underestimate the true value. Immunocytochemistry against MYH4 and CTNT, which did not require cell dissociation, also supports this (Figure 5.3D-E). Nevertheless, the levels of CTNT+ cells between biological replicates in 2D differentiation was much more consistent than observed with microcarrier-based cardiomyocyte differentiation.

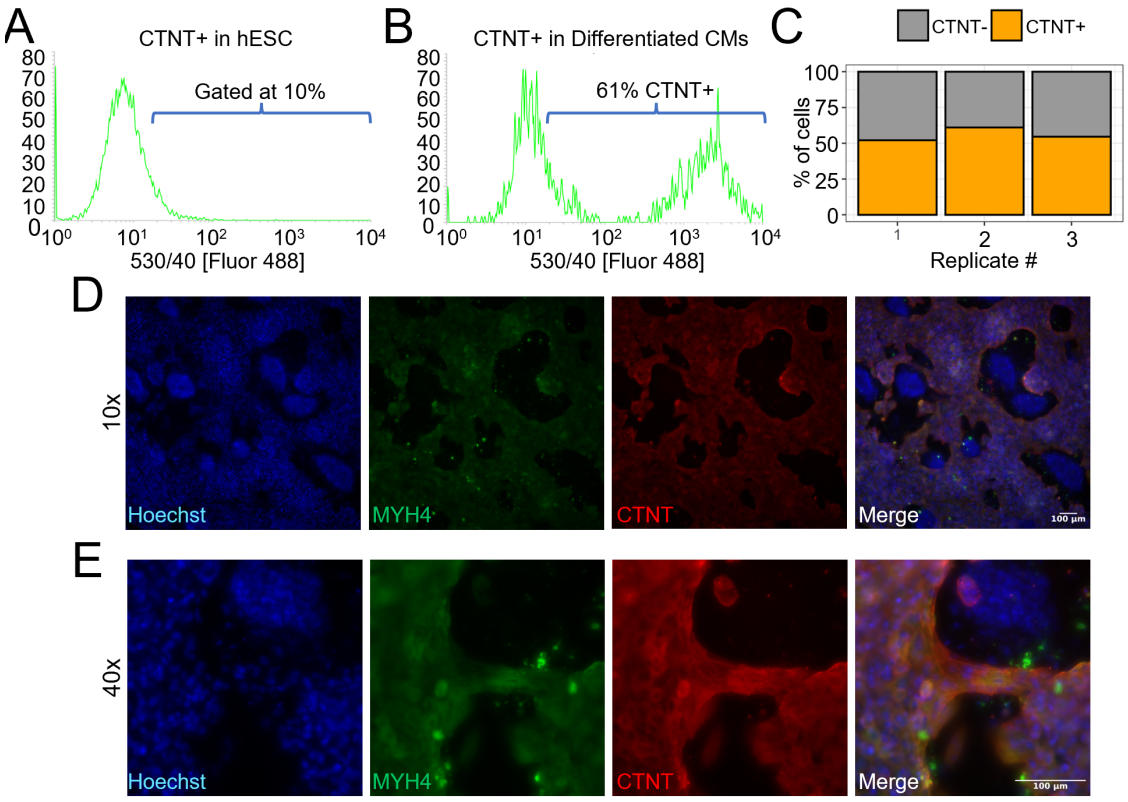

Figure 5.3: Assessing differentiation efficiency via CTNT flow cytometry. A) Flow cytometry showing detection of CTNT in undifferentiated H9 hESCs stained with CTNT primary antibody and Alexa-Fluor 488-conjugated secondary antibody. B) Flow cytometry showing detection of CTNT in H9-derived cardiomyocytes stained with CTNT primary antibody and Alexa-Fluor 488-conjugated secondary antibody. C) Breakdown of CTNT expression following differentiation of three cell populations. (D-E) H9-derived cardiomyocytes were immunolabelled for MYH4 (Myosin 4; green) and CTNT (Cardiac Troponin; red) and visualized via fluorescence microscopy at: D) 10x; E) 40x.

We also conducted differential gene expression analysis to further validate that the cell types produced at each stage of differentiation represented those expected. Principal component analysis revealed three distinct clusters of samples, clearly separated by day of differentiation (Figure 5.4B), indicating substantial differences in global gene expression profiles. Differential gene expression analysis between day 0 and day 2 revealed significant upregulation of mesodermal markers LHX1, TBXT, MIXL1, MESP1 and MESP2. Additionally, differential gene expression analysis between day 0 and day 16

revealed significant upregulation of cardiac markers MYBPC3, MYL2, MYH7, TBX5, and TNNT (Figure 5.4C), whilst there was significant downregulation of pluripotency genes FGF2, SOX2, and NANOG (Figure 5.4D). Finally, gene set enrichment analysis conducted on genes ranked from most upregulated at day 16 to most downregulated at day 16 revealed a significant enrichment of genes relating to "cardiac muscle tissue development", "regulation of heart contraction", and "cardiocyte differentiation" (Figure 5.4E).

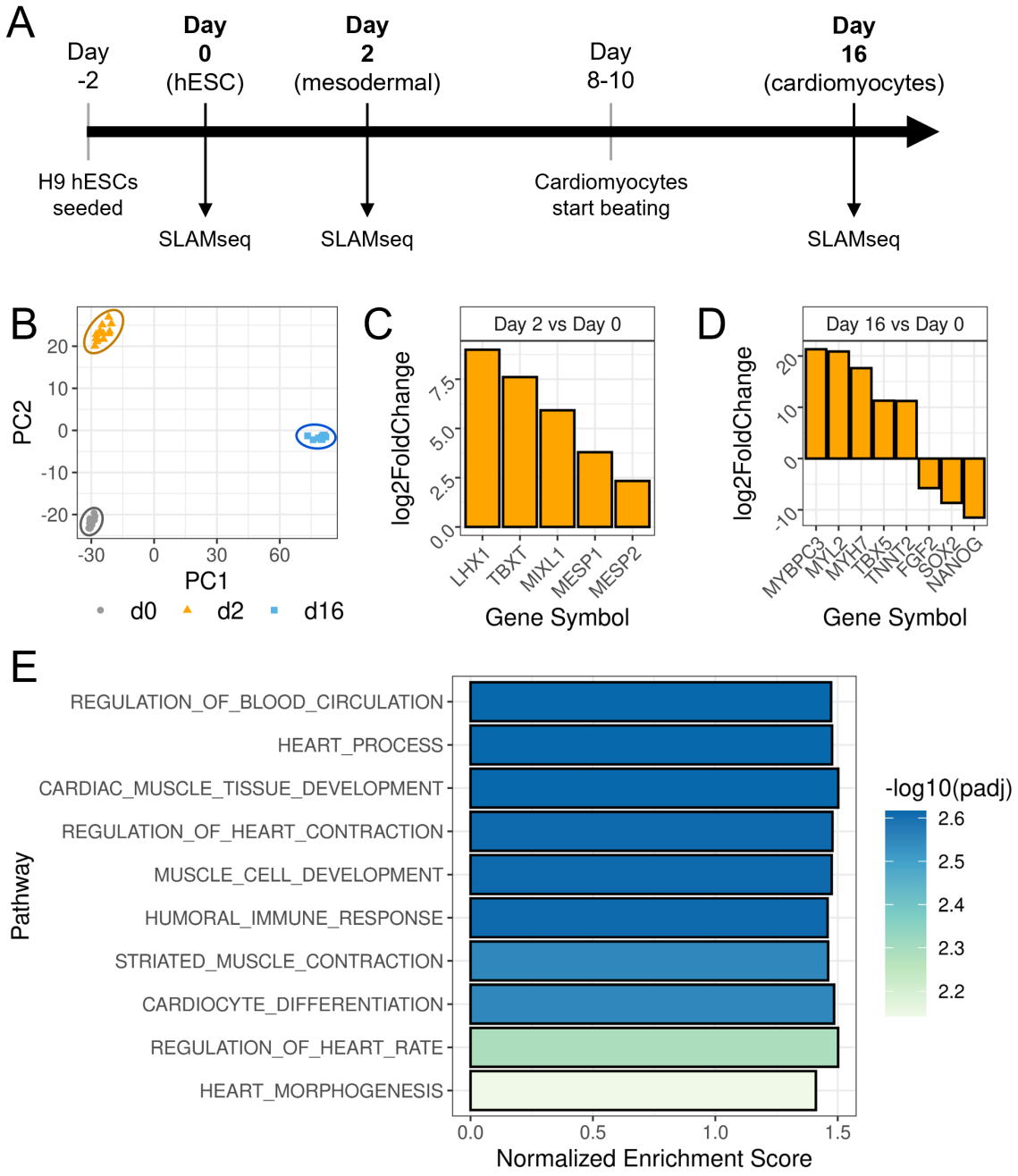

Figure 5.4: Differential gene expression analysis in cardiomyocyte differentiation. A) Overview of differentiation time course and plan for SLAMseq experiment. B) Principal Component Analysis (PCA) plot for day 0, day 2, and day 16 samples. C) Upregulation of mesodermal markers at day 2 compared to day 0. D) Upregulation of cardiac markers and downregulation of pluripotency markers at day 16 compared to day 0. E) Gene set enrichment analysis on genes ranked on their upregulation at day 16 compared to day 0.

5.2 Measuring global mRNA stability in a cardiomyocyte differentiation timecourse

Following the successful establishment of a cardiomyocyte differentiation time course, we planned to conduct SLAMseq at three time points during differentiation in order to produce transcriptome wide RNA stability estimates. Unlike tradition SLAMseq, which uses QuantSeq to sequence the 3' end of mRNA, we utilized high-depth paired-end RNAseq (>120M PE150 reads), which allowed us to conduct differential analyses at both the gene (Section 5.3) and transcript level (Section 6). We planned to conduct SLAMseq at day 0, day 2, and day 16 of differentiation, corresponding to hESCs, mesoderm, and cardiomyocytes (Figure 5.4A). However, prior to conducting SLAMseq we first had to optimize the 4sU incorporation concentration, determine whether incorporation levels were sufficient to proceed, and develop software to analyse data in this format. These optimization and developments are discussed herein.

5.2.1 Optimizing 4sU concentration

Prior to conducting SLAMseq with multiple biological replicates and across multiple differentiation time points, we optimized the 4sU-incorporation step to ensure workable levels of cell viability and 4sU incorporation in hESCs. Whilst it was originally claimed that 4sU is minimally toxic to cells, as shown in mouse embronic stem cells (Herzog et al., 2017), multiple studies have demonstrated that culturing cells with media containing 4sU for prolonged periods of time results in cell death. Therefore to assess the impact of 4sU on the viability in hESCs, we cultured our H9 hESCs in conditions mimicking day 0 of differentiation (growth on Matrigel for two days in mTeSR-Plus following seeding at 600,000 cells/well), but with increasing concentrations of 4sU supplemented into the media, for a total of 12 hours. After 12 hours, we assessed cell viability via cell counting. We found that at the highest concentrations there was a profound effect on cell viability (Figure 5.5). In line with a previous SLAMseq experiment conducted in the lab which

found that 4sU concentrations resulting in roughly 60% viability led to workable 4sU incorporation rates, we opted to use 65µM as our working 4sU concentration (Figure 5.5).

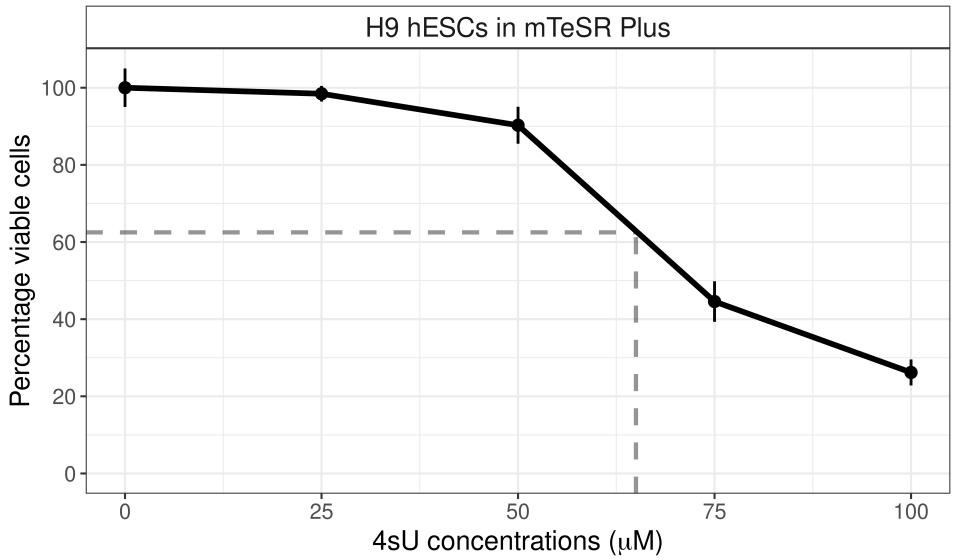

Figure 5.5: Effect of 4sU treatment on cell viability. H9 hESCs (day 0 of differentiation time course) were treated with increasing concentrations (0-100µM) of 4sU in mTeSR Plus medium for 12 hours, with media replacement every three hours. Points represent the mean viability of three biological replicates ± 1 standard error of the mean. Dashed line from 65µM indicates the relative viability at the concentration selected for further usage.

5.2.2 Analysis of 4sU incorporation rates

To determine the 4sU incorporation rates at each time point in differentiation, we cultured cells at day 0, day 2, and day 16, in the relevant differentiation media supplemented with 65µM 4sU for 12 hours, then extracted RNA, performed iodoacetamide treatment to alkylate 4sU, and sent this for library preparation and sequencing at Novogene. Upon library preparation, alkylated 4sU would cause T>C conversions within the sequencing inserts at the reverse transcription stage, these would subsequently be detected through the development of a bioinformatic pipeline.

Sequencing for SLAMseq is usually conducted with QuantSeq, therefore there is high

read coverage at the 3′ end of each gene. As such, analysis relies around interrogating T>C conversion rates at the level of bases. Given that we have used "normal" RNAseq, our reads are not limited to the 3′ termini of each gene, and are instead distributed through the entire transcript body. Therefore we are unlikely to have a high enough incorporation rate to interrogate changes at the base level. Instead we can look at the rate of converted reads and see how this changes over time. This can be conceptualized as an "in silico pulldown". Similarly to how 4sU-containing RNA can be pulled down following biotinylation, instead of biotinylation and physical pull down, here we sequence all the RNA following iodoacetamide treatment and pull down based on the presence of conversions.

Figure 5.6 presents an overview of pipeline development where analyses of increasing complexity were developed over time. Prior to determining the number of 4sU-related nucleotide conversions, raw reads from the pilot sequencing experiment were mapped with STAR. Unlike the SlamDunk software that accompanies commercially available SLAMseq kits, we used the STAR algorithm to align reads, as opposed to NextGenMap (NGM; used by SlamDunk) as NGM does not map spliced reads, whilst STAR is a splice-aware alignment tool and we eventually wish to assess isoform-specific stability. Mapped reads in .bam format were then interrogated to determine conversions. Each step of the flow diagram in Figure 5.6A represents an increasingly complex approach to determining converted bases. Each analysis builds upon the last to either increase detection rate, or to reduce false positive rate.

The first and simplest script iterates through the .bam file (mapped reads) read-by-read and determines whether there are any nucleotide mismatches between the aligned reads and the reference genome. Where there were mismatches, the script checks whether these represent T>C or A>G conversions. Where this was the case, the conversion count was increased by one per mismatch. This allowed us to determine the distribution of conversions across all reads per sample. In hESCs, mesoderm and cariomyocyte samples,

39.5%, 56.5% and 54.4% of reads had at least one conversion, respectively (Figure 5.6B; panels 1-3). However, 7.1% of reads in the no 4sU negative control sample also contained at least one conversion (Figure 5.6B; panel 4); this represents the false positive rate.

Given that we conducted paired-end sequencing, instead of looking at individual reads we were able to look at pairs, thus allowing us to determine the number of conversions per sequencing insert as opposed to per read. This would mean that for a pair to be classed as converted, it would only require 1 in 300 nucleotides to be converted, as opposed to 1 in 150. We achieved this by name sorting the alignment files so that reads from the same pair would appear sequentially within the .bam file. Subsequently, we iterated over read pairs as opposed to individual reads. Conversions were detected in the same manner as in the first script; however, instead of outputting the conversions per read, the sum of both reads in each pair was taken to give a "conversions per pair" value. Plotting the distribution of conversions per pair, we found that 64.3%, 75.3% and 70.7% of read pairs contained at least one T>C or A>G conversion in hESCs, mesoderm, and cardiomyocyte samples, respectively (Figure 5.6C). Given that this script upgrade increases the sensitivity of detection, the increase in conversion rate was also accompanied by an increase in false positive rate, amounting to 11.1% in the no 4sU control (Figure 5.6C). However, it is important to note that the sensitivity increased more than the accuracy decreased.

So far, we were detecting all T>C and A>G conversions regardless of strand, which does not take into the account the underlying biology. 4sU incorporation would cause T>C conversions in the sense strand relative to the direction of transcription, and subsequently produce A>G conversions in the antisense strand. Given that these previous two scripts were unaware of the direction of transcription, all conversions observed were captured. To fix this issue, we used featureCounts to assign reads to genes. We used this information to determine if each read was in the forward or reverse direction relative to transcription. On the forward read we would only consider T>C conversions, whilst the reverse read would

only consider A>G conversions. This upgrade reduced the false positive rate from 11.1% to 6.3% in the negative control (Figure 5.6D). This was also accompanied by a reduction in false positive rate in our 4sU-treated samples, where we found that 50.4%, 60.8% and 56.7% of read pairs contained at least one conversion in hESCs, mesoderm, and cardiomyocytes, respectively.

To reduce the false positive rate further, we next accounted for the potential presence of single nucleotide polymorphisms (SNPs). Where SNPs are present in the genome of our H9 cells relative to the reference genome, if these corresponded to T>C or A>G conversions, these could be mistaken for 4sU-linked conversions. Additionally, where RNA editing occurs it is possible that nucleotide conversion may occur independently of 4sU incorporation. To account for this, we used the no 4sU negative control file to identify consistent differences from the reference genome using VarScan, producing a VCF file with the location of each SNP. In the next version developed, where a conversion is called, it is subsequently checked against the SNP VCF file, and if it matches it is disregarded. For the pilot sequencing experiment, only one negative control was sequenced; therefore, the hESC no 4sU control was compared to all 4sU-treated samples. However, for the main SLAMseq experiment (with multiple biological replicates, and a chase time course), a no 4sU control would be conducted and sequenced at each stage of differentiation, allowing direct comparison. This would account for capture of transcripts which are only expressed at a single stage of differentiation. By implementing these changes, we further reduced the false positive rate from 6.3% to 4.2% in the no 4sU control (Figure 5.6D). We found that 49.4%, 59.9% and 55.7% of read pairs show at least one conversion in hESCs, mesoderm, and cardiomyocytes respectively (Figure 5.6E).

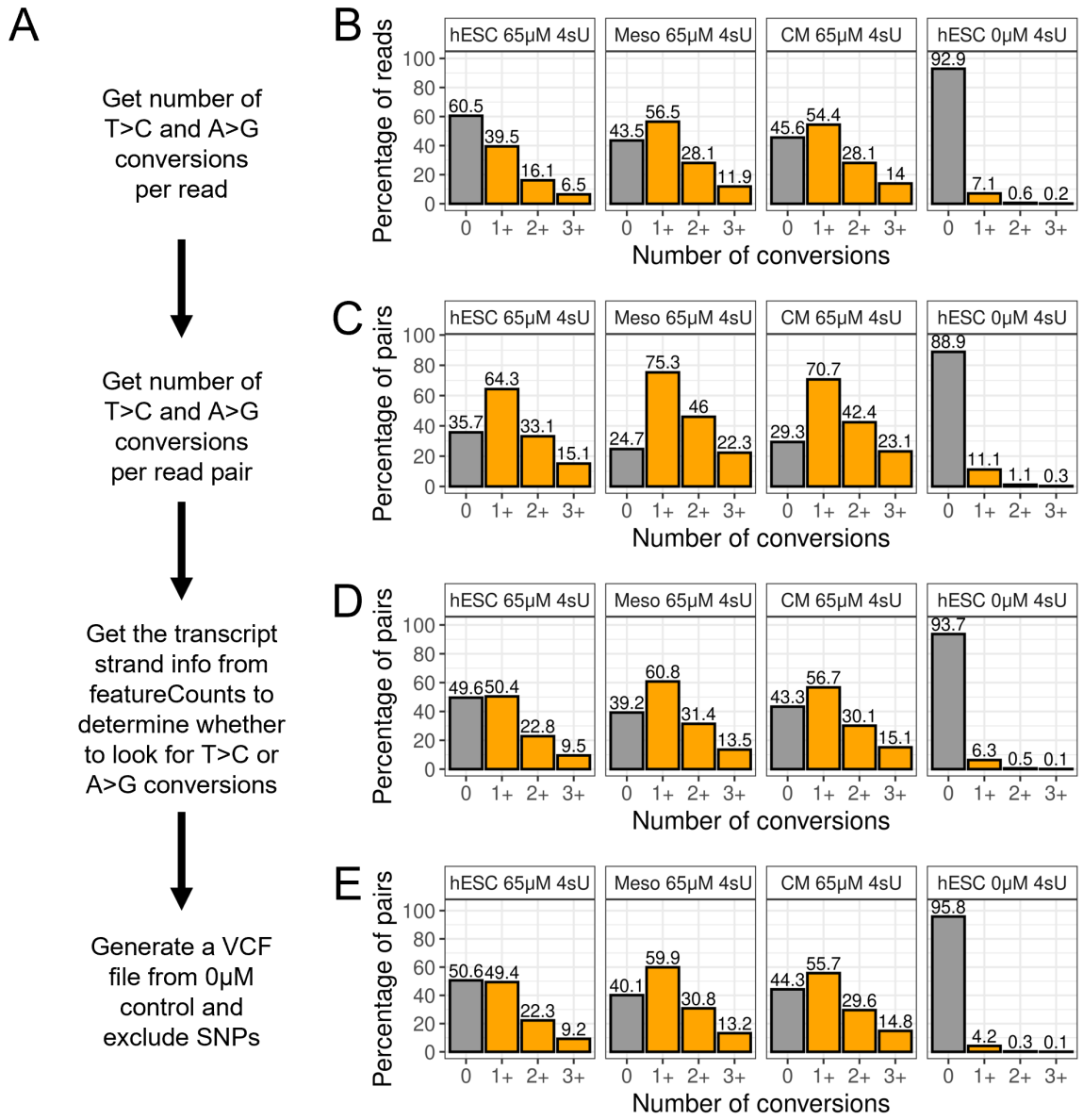

Figure 5.6: Percentage nucleotide conversion during script development. Data from the pilot RNAseq experiment with 5 million PE150 read pairs. A) Schematic of script development. B) Percentage of reads showing X T>C or A>G conversions. C-E) Percentage of read pairs showing X T>C or A>G conversions. D) Transcript strand information from featureCounts was taken into account. If the transcript was from the sense strand then T>C conversions were counted on the forward read, and A>G conversions counted on the reverse read. If the transcript was from the antisense strand then A>G conversions were counted on the forward read, and T>C conversions counted on the reverse read. E) A variant call format (VCF) file was generated from the no 4sU control to capture the position of SNPs, which were subsequently excluded from conversion counting.

We subsequently conducted SLAMseq with multiple biological replicates across differentiation, with chase time points at 0hr, 3hr, and 12hr post-pulse. An overview of 4sU conversion rates in each biological replicate is shown in Figure 5.7. Importantly, there is a high degree of consistency between biological replicates per chase time point at each stage of differentiation, and we can clearly see a decrease in "pull down" as the chase proceeds and old RNA is degraded. The high level of consistency indicates that our 4sU incorporation rate is consistent between biological replicates. Alkylation rate (i.e. the proportion of 4sU-incorporated nucleotides that successfully lead to the observation of T>C conversions following IAA treatment) is also a potential source of intra-timepoint variability, and whilst we did not conduct a "no IAA" control, the low level of variability between biological replicates of the same time point suggests that our conversion rate is high in line with the original SLAMseq methods paper (Herzog et al., 2017).

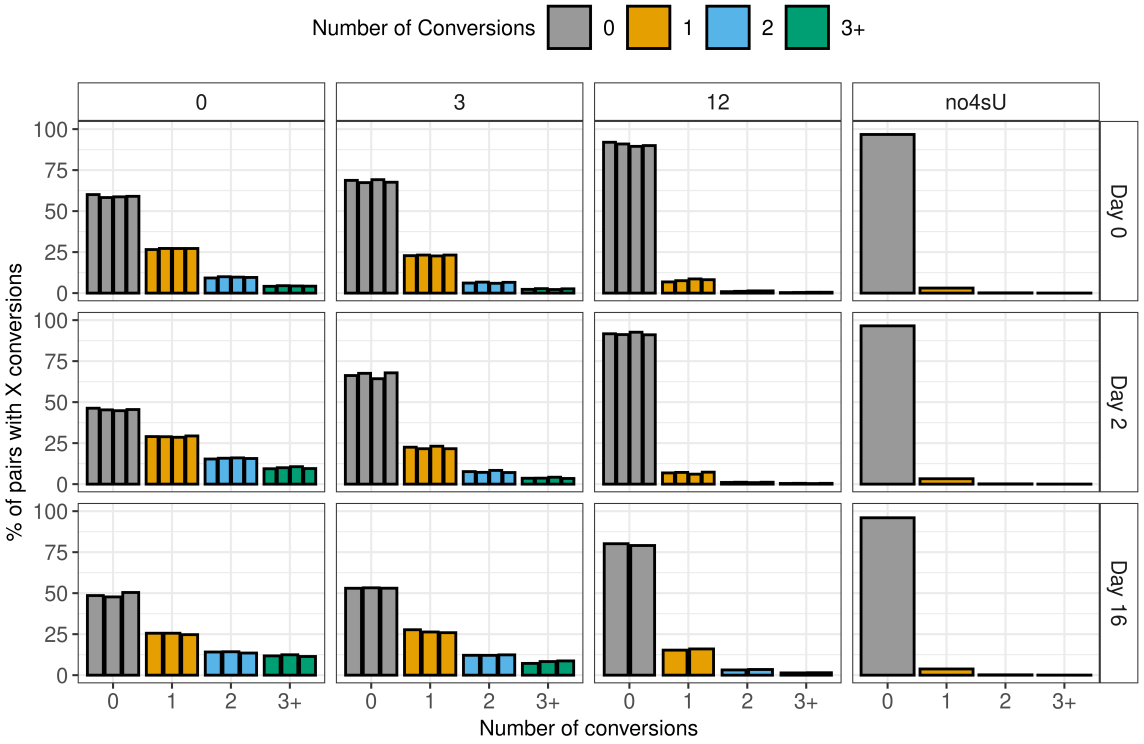

Figure 5.7: SLAMseq per-sample conversion rates. Each column represents the conversion rates across the chase time course, from 0 hours across to 12 hours post-pulse. Each row represents a different stage of differentiation. Day 0 = hESCs, day 2 = mesoderm, day 16 = cardiomyocytes.

5.2.3 Generation of decay curves for half-life estimation

Whilst these percentages represent the total 4sU-related nucleotide conversion rates, in order to generate decay curves for individual genes, our final script outputs metadata from each read pair, including: the number of conversions, the total coverage, the total T/A content (relative to the read, strand, and direction of transcription; effectively giving us a "total convertible sequence" value), and the gene assignment. In order to calculate gene-level half-life estimates, we imported the metadata output from our python script into R. By summarizing over each gene at each chase time point, and at each stage of differentiation, we were able to determine how many read pairs contained at least one conversion. By comparing this to the total number of those that did not contain a

conversion, we were able to calculate a "percent converted" for each gene, corresponding to the percentage of old RNA (RNA from the 4sU pulse) that remains. As the time following on from the 4sU pulse increased, RNA decays, and is replaced by RNA that does not contain 4sU, therefore the percent converted decreases over time (this is visible in Figure 5.7 by looking across the coloured bars in each row). For each gene, and at each stage of differentiation, we fit the decrease in percentage conversion over time to an exponential decay model using nls from the stats package in R. The following equation was used:

Where e = exponential function; b = decay rate; t = chase time

$$Remaining\ old\ RNA\ (Percentage\ Conversion) = e^{bt}$$

The nls package uses non-linear least squares regression to estimate the b value (decay rate) based on the data points provided (percentage conversion and time). This b value can be used to directly calculate the half-life, as follows:

$$Half\ Life = \frac{log(2)}{-b}$$

5.2.4 Validating half-life estimates against expected biological function

To determine whether our half-life estimates related to what we would expect from the underlying biology, we took our half-life estimates from the day 0 hESCs, ranked them from most stable to least stable, and conducted GSEA to assess the enrichment of genes related to individual biological processes. We found that genes with housekeeping functions, such as "aerobic respiration", "oxidative phosphorylation", "ATP metabolic process", "cytoplasmic translation", and "mitochondrial transmembrane transport", had the most stable RNA (Figure 5.8). Meanwhile genes with functions in "embryonic organ development", "cell fate commitment", and multiple transcriptional regulation ontologies, had the least stable RNA (Figure 5.8).

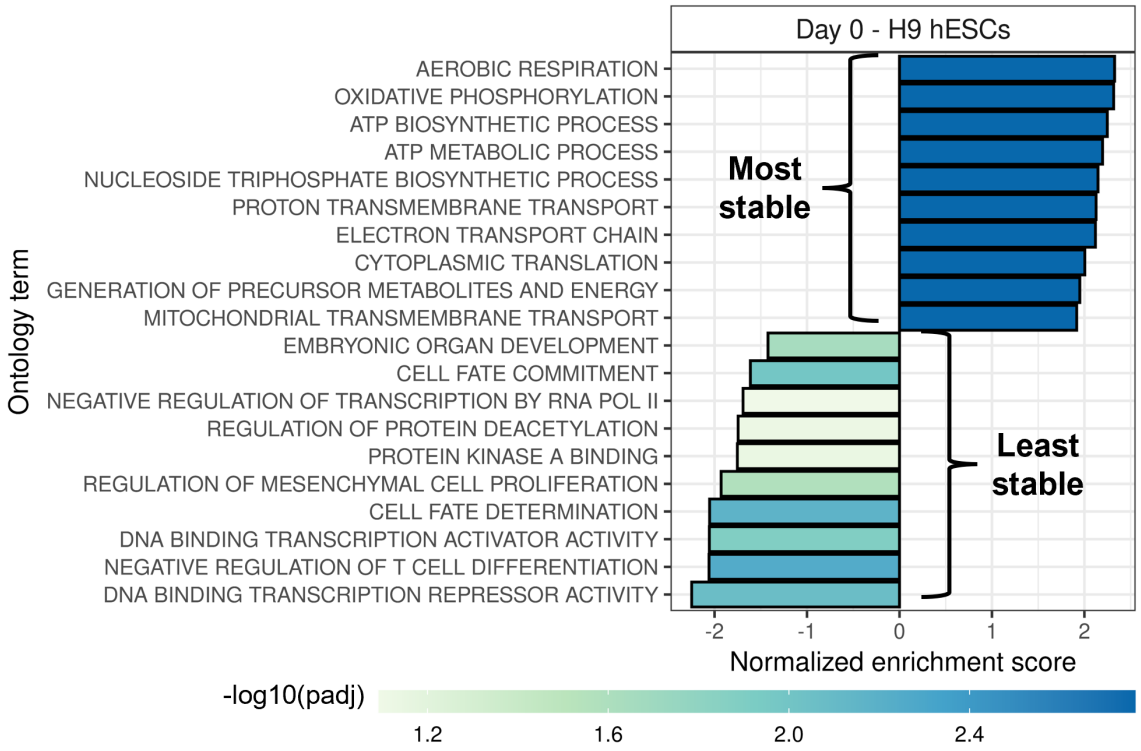

Figure 5.8: Half-life estimates correlate with expected biological function. Gene set enrichment analysis conducted on half-life estimates from day 0 hESCs ranked from most stable to least stable.

5.2.5 Assessing significance of half-life differences

Whilst having decay rates at each stage of differentiation allows us to assess whether there are differences in half-lives across differentiation, it does not provide any assessment of the level of statistical certainty of any trend between conditions (e.g. day of differentiation). In order to assess the statistical significance of differential decay rates during differentiation, for each gene we fit two models as outlined in Figure 5.9A. The first model, termed the "shared parameters" model, represents a "grand fit", where one decay rate is produced by fitting a single exponential decay curve to all data points from all three differentiation stages. The second model, termed the "combined parameters" model, fits three decay rates, one per differentiation stage. These two models can then be compared via an ANOVA test to determine whether the more complex "combined

parameters" model fits the data significantly better. Where the three fits from the "combined parameters" model are very similar, they will not be significantly different from the "shared parameters" fit, and the P-value obtained from the ANOVA test will reflect this. However, where all fits are substantially different from each other, or where one fit is substantially different from the other two, the ANOVA test will reflect this with a low P-value, based on the significance of difference. Together this allows us to determine the half-life of each gene at each stage of differentiation, and produce a P-value reflecting whether there is a significant change in half-lives across differentiation. As demonstrated by the example in Figure 5.9B, HNRNPD shows a significant difference in decay rates across differentiation, where the decay rate at day 2 is faster than days 0 and 16, indicating a shorter half life at day 2, whilst half-lives at day 0 and day 16 are longer and more similar to each other.

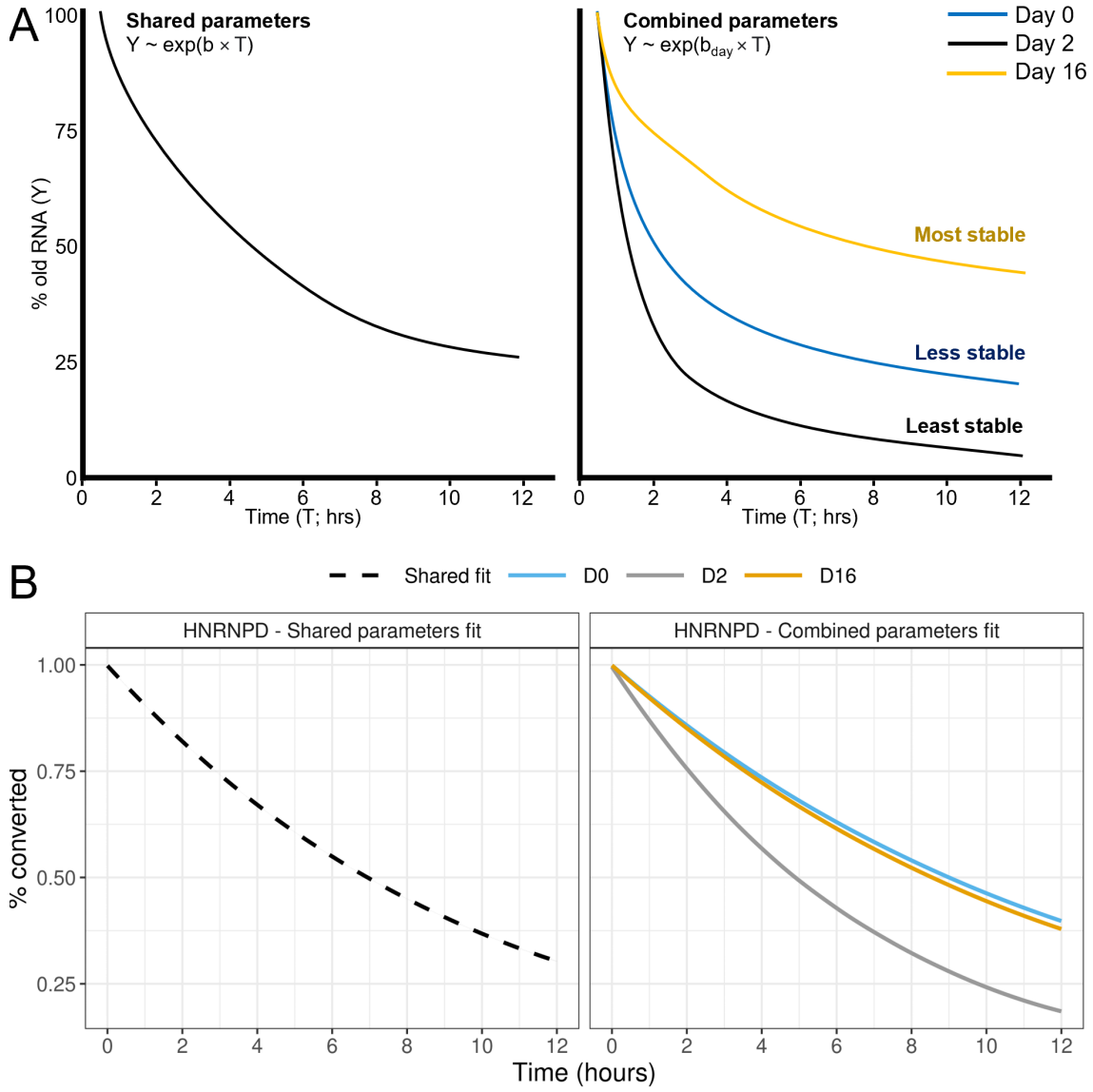

Figure 5.9: Schematic depicting the calculation of P-values for changes in RNA stability across differentiation. A) In the shared parameters model the b value (strength of slope) is calculated using data points from all differentiation time points (i.e. it represents an average decay rate). In the combined parameters model the b value is calculated for each day of differentiation. P-values are calculated by conducting an ANOVA test between the two models. B) Example usage in HNRNPD.

5.3 Global gene-level half-life estimates vary across differentiation

Using the final script presented in Section 5.2, we next attempted to estimate the half-lives for all genes at day 0, day 2, and day 16 of differentiation. The gene-level half-life estimates presented in this section represent those that could be fit at all three time points of differentiation. Where gene expression was very low or zero, at any given time point of differentiation, a decay curve would not be fit, due to lack of counts. Additionally, given that our chase time points were at 0, 3, and 12 hours, we filtered out half-life estimates that were less than 30 minutes and more than 24 hours. After specifying these criteria, we were able to fit decay curves at all three time points of differentiation for 13,395 genes, whilst also obtaining adjusted P-values representing the significance in half-life difference across differentiation.

To identify global trends in how RNA stability changes during differentiation, we first generated a density plot to visualize the distributions of half lives. In the density plot each half-life was weighted by its gene expression level (TPM value), therefore the distributions presented in Figure 5.10A represent the distribution of all RNA in the cell. In other words, if a random RNA molecule was to be picked from a cell and examined, the likelihood of it having a given half-life is shown by Figure 5.10A. We find that the average expression-weighted half-life at day 0 is 6.68 hours, at day 2 is 4.31 hours, and at day 16 is 10.01 hours. This indicates that there is a transient destabilization of RNA globally at day 2, followed by stabilization by day 16.

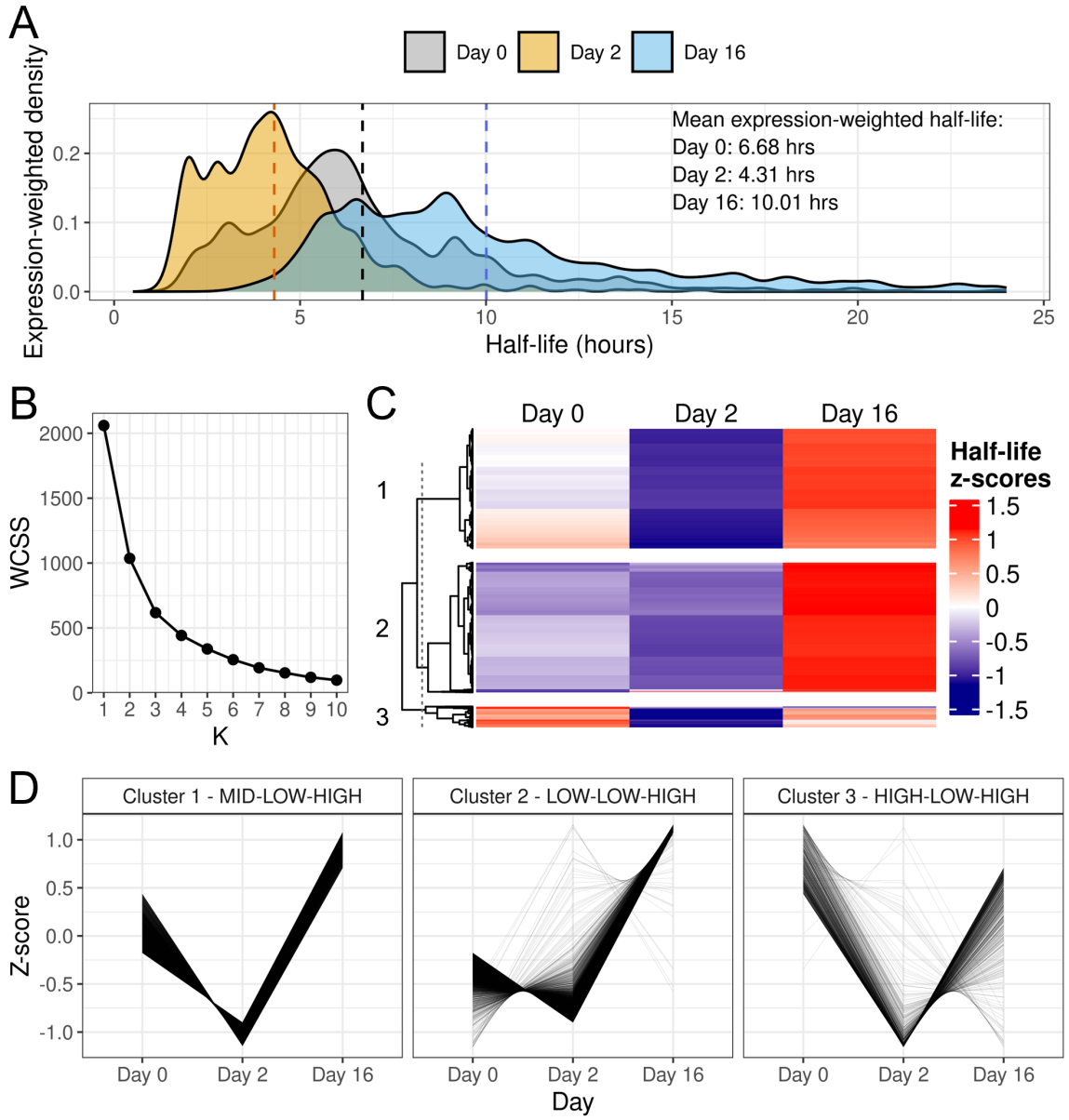

Figure 5.10: Comparison of gene level half-life estimates between hESCs and differentiated cells. A) Gene-level half life estimates weighted by gene expression at day 0 (hESC), day 2 (Mesodermal), and day 16 (Cardiomyocytes). B) Elbow plot showing the effect of K value selection on the within-cluster sum of squares (WCSS) value. The value of k is selected based on the asymptote (here k=3). C) Heatmap where half-life z-scores have been clustered via k-means clustering to produce three clusters. Cluster 1 = MID-LOW-HIGH. Cluster 2 = LOW-LOW-HIGH. Cluster 3 = HIGH-LOW-HIGH. D) Line graph showing the z-score trends of clusters 1-3.

Next, instead of looking at all genes, we focused on those that displayed a significant half-life difference across differentiation (Padj<0.05). Of the 13,395 genes we could generate a half-life estimate for, 8,781 of these showed a significant difference across differentiation. To determine whether all of these genes showed the same trend, or whether individual groups of genes displayed different patterns, we conducted k-means clustering. First, we conducted k-means clustering with an increasing K value, calculated the within cluster sum of squares (WCSS), and plotted this as an elbow plot in Figure 5.10B. We found that the elbow corresponded to k=3, therefore we proceeded to conduct k-means clustering with three clusters, and produced the heat map in Figure 5.10C. All three clusters showed a transient destabilization at day 2, followed by stabilization at day 16. The main difference between the three clusters was the half-life at day 0, being either MID (0.25 $>$ z-score $>$ -0.25), LOW (z-score $<$ -0.25), or HIGH (z-score $>$ 0.25) in clusters 1, 2, and 3 respectively. A secondary visualization of these trends across differentiation between clusters is shown as a line graph in Figure 5.10D.

To gain more insight into the genes in each cluster, and the potential biological functions they contribute towards, we conducted gene ontology analysis on each cluster. We found that genes in cluster 1 were enriched for biological processes related to RNA localization, nuclear export, regulation of mRNA metabolic process, and the cell cycle (Figure 5.11A). Meanwhile genes in cluster 2 related to glycoprotein function, endoplasmic reticulum stress, autophagy, and the JAK-STAT pathway (Figure 5.11B). No enrichment of ontology terms was observed for genes in cluster 3.

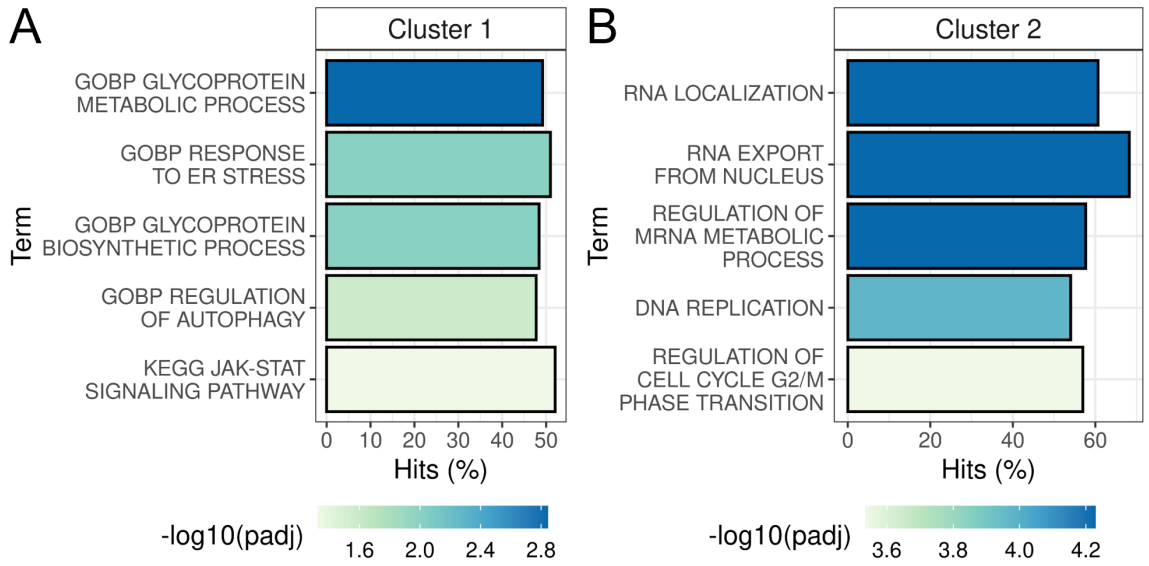

Figure 5.11: Gene ontology enrichment for gene-level half-life clusters. A) GO term enrichment for genes in cluster 1 (MID-LOW-HIGH). B) GO term enrichment for genes in cluster 2 (LOW-LOW-HIGH).

To determine whether changes in RNA stability are accompanied by changes in gene expression levels, we conducted differential gene expression analysis between day 0 and day 2, and between day 2 and day 16. Changes in RNA half-life, and gene expression changes, between day 0 and day 2 were correlated in Figure 5.12A, producing four quadrants (Q1-4). These two variables do not appear to correlate, and the major effect appears to be destabilization, which occurs independently of changes in gene expression. 24 genes were found in Q1 where RNA was stabilized despite downregulation of gene expression, no gene ontology terms or KEGG pathways were found to be enriched in Q1. Q2 contained 108 genes, which displayed RNA stabilization accompanying upregulation of gene expression. Again, no gene ontology terms or KEGG pathways were enriched in Q2. Q3 contained 1295 genes, these genes displayed RNA destabilization accompanied by downregulated gene expression. Genes in Q3 were enriched for KEGG pathways relating to the "cell cycle", "ribosome biogenesis in eukaryotes", "DNA replication", "purine metabolism", and "RNA transport" (Figure 5.12B). Finally there were 1067 genes in Q4. Interesting these genes displayed RNA destabilization despite an increase in gene

expression. Genes in Q4 were enriched for KEGG pathways including "cell adhesion molecules" and "endocytosis" (Figure 5.12C).

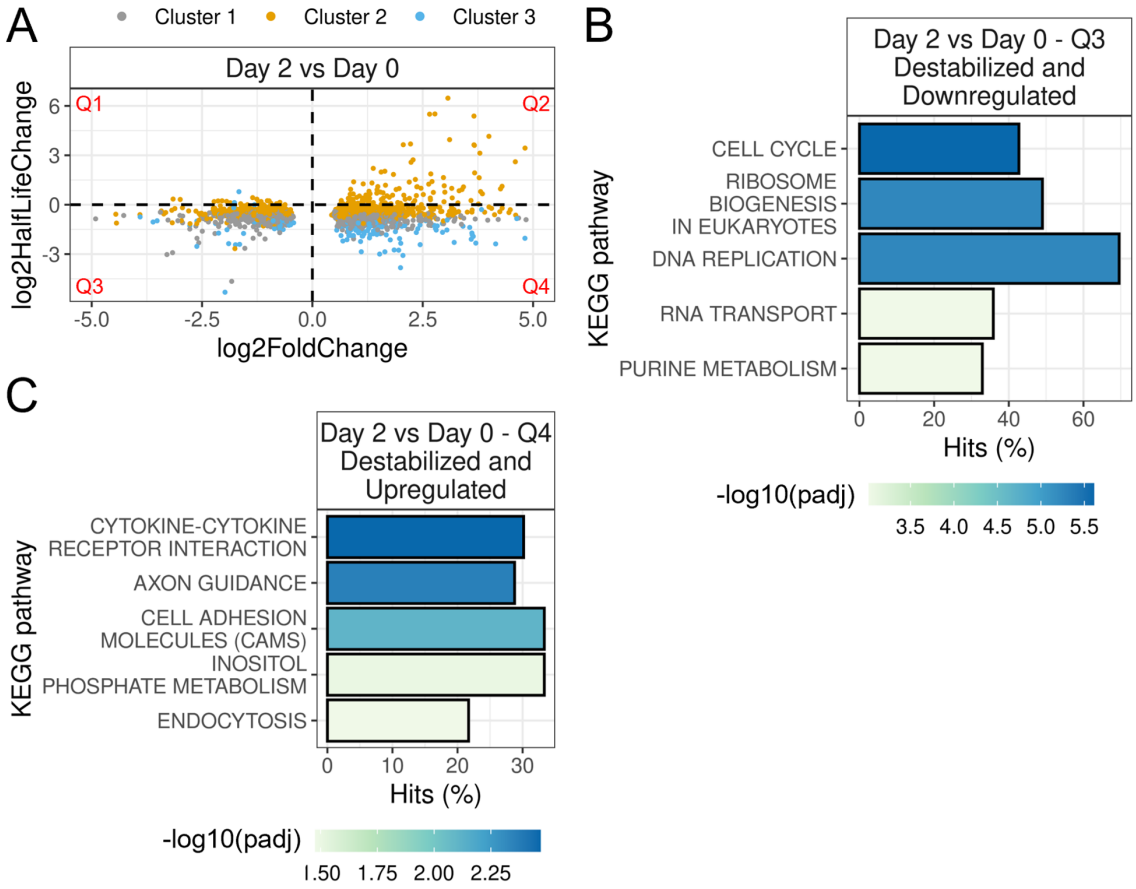

Figure 5.12: Comparison of half-life changes and gene expression changes between day 0 and day 2. A) Gene expression changes (Log2FoldChange values) are plotted on the X-axis, stability changes (Log2HalfLifeChange values) are plotted on the Y-axis. Log2FoldChange > 0 = upregulation at day 2. Log2HalfLifeChange > 0 = stabilization at day 2. (B-C) KEGG pathway enrichment for: Q3 (B) and Q4 (C).

A similar analysis was also conducted between day 2 and day 16, as shown in Figure 5.13A. The major effect appeared to be stabilization, again independently of expression changes. There were 1630 genes in Q1 and 1676 genes in Q2. Genes in Q1 displayed an increase in RNA stability despite a decrease in gene expression. These genes were enriched for KEGG pathways relating to the "splicesome", "cell cycle", "ribosome

biogenesis in eukaryotes", "RNA transport" and interestingly "RNA degradation" (Figure 5.13B). Genes in Q2 were stabilized and upregulated, and were enriched for KEGG pathways relating to "oxidative phosphorylation", "cardiac muscle contraction", "ECM-receptor interactions", "focal adhesion", and "calcium signaling pathway" (Figure 5.13C).

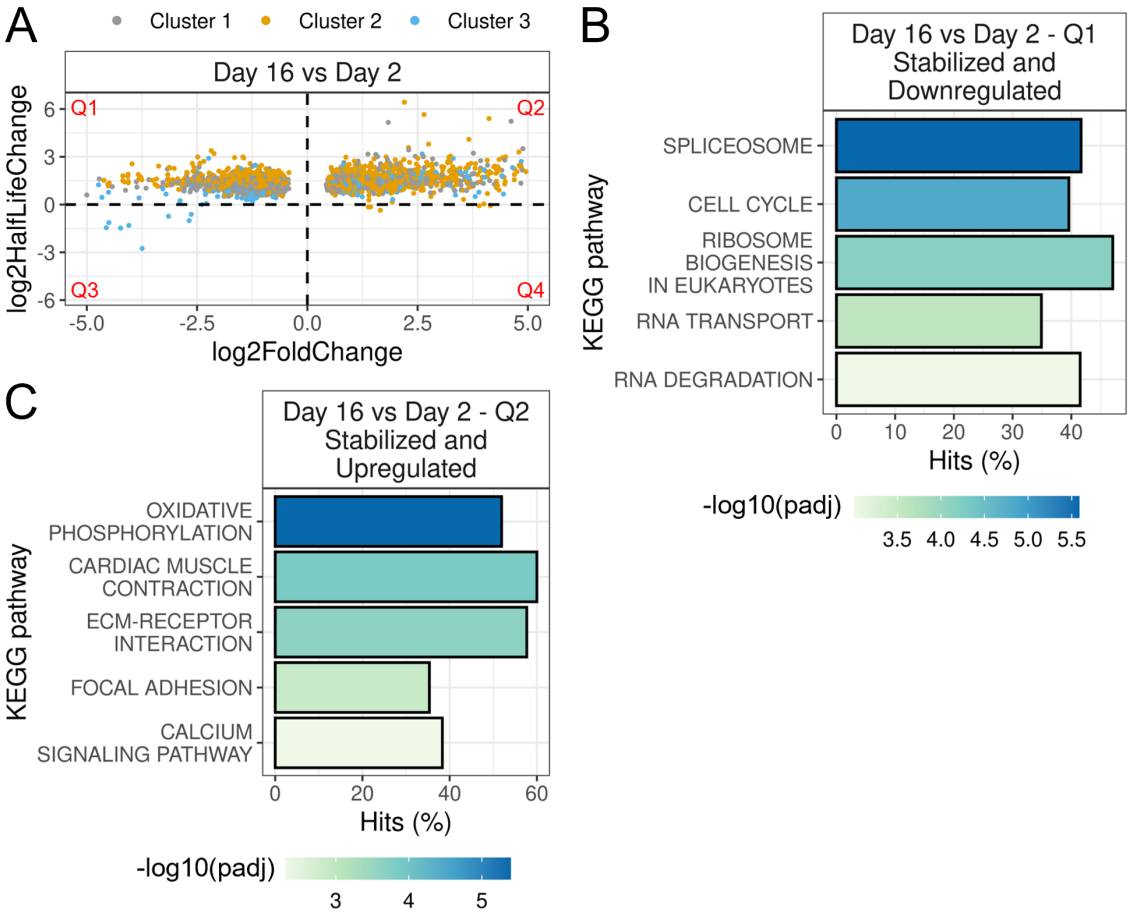

Figure 5.13: Comparison of half-life changes and gene expression changes between day 2 and day 16. A) Gene expression changes (Log2FoldChange values) are plotted on the X-axis, stability changes (Log2HalfLifeChange values) are plotted on the Y-axis. Log2FoldChange > 0 = upregulation at day 16. Log2HalfLifeChange > 0 = stabilization at day 16. (B-C) KEGG pathway enrichment for: Q1 (B) and Q2 (C).

To investigate the features of genes that contribute to their observed RNA stability estimates, and how these interact with expression levels, we conducted a series of

correlations. By comparing between genes, and across differentiation, we find that genes with longer 3'UTRs are generally expressed at lower levels (Figure 5.14A). Additionally genes with longer 3'UTRs tend to have shorter half lives (Figure 5.14B). These correlations are observed across differentiation, they are statistically significant; however, the relationships are weak to moderate (R between -0.16 and -0.31). We also find that the half-life positively correlates with gene expression between genes, and that this correlation becomes progressively stronger during differentiation (Figure 5.14C). Additionally, we find that 3'UTR GC content positively correlates with half-life, and that this relationship becomes progressively weaker during differentiation (Figure 5.14D).

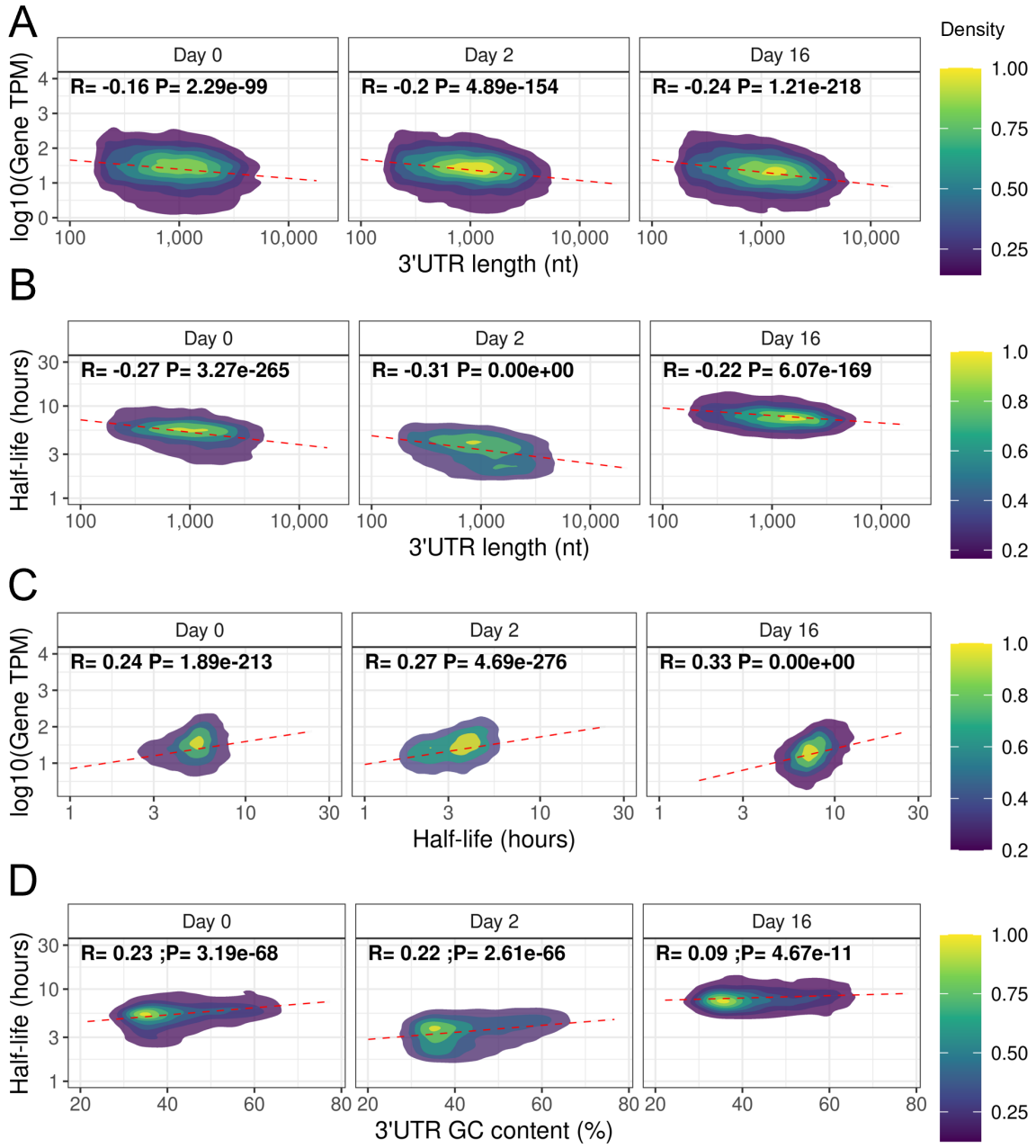

Figure 5.14: Relationship between 3'UTR length, half-life estimates, and gene expression. A) Correlation between 3'UTR length and gene expression. B) Correlation between 3'UTR length and gene-level half-life estimates. C) Correlation between gene-level half-life estimates and gene expression. D) Correlation between 3'UTR GC content and gene-level half-life. R and P values relate to Spearman's rank-order correlation. R = strength of the relationship. P = significance level. P<0.05 was considered statistically significant.

Whilst these correlations represent the differences between genes, they do not take into account that the 3'UTR length of individual genes can change during differentiation due to alternative polyadenylation. 3'UTR shortening has previously been observed during the early stages of cell differentiation cascades, where highly proliferative cell populations arise, as well as in cancers (see Section 1.3). To assess 3'UTR length changes for individual genes across differentiation, we conducted differential alternative polyadenylation analysis. This allowed us to determine differences in distal versus proximal polyA site usage, and how this changes during differentiation. We find that between day 0 and day 2, 2,559 3'UTRs shorten significantly, while 1,657 3'UTRs lengthen significantly (Figure 5.15A). Inversely, between day 2 and day 16, 2,635 3'UTRs lengthen significantly, whilst 1,128 3'UTRs shorten significantly (Figure 5.15B). Across the total differentiation time course we find that roughly equal proportions of genes show significant shortening and lengthening as shown by the 0 vs day 16 comparison (Figure 5.15C). Besides global trends being observed, k-means clustering was conducted to identify clusters of genes that behaved similarly. This produced three clusters: cluster 1 displayed a progressive lengthening of 3'UTRs across the differentiation; cluster 2 displayed a progressive shortening of 3'UTRs during differentiation; and cluster 3 displayed an intermediate phenomena, where 3'UTRs were longest at day 0, shortened by day 2, and lengthened by day 16 (Figure 5.15D). Gene ontology and KEGG pathway analysis was conducted on all three clusters, however significant hits were only observed in cluster 2 (progressive shortening), indicating a significant enrichment of genes related to mRNA metabolism, the splicesome, and several signaling pathways, including the Wnt signal pathway (Figure 5.15E).

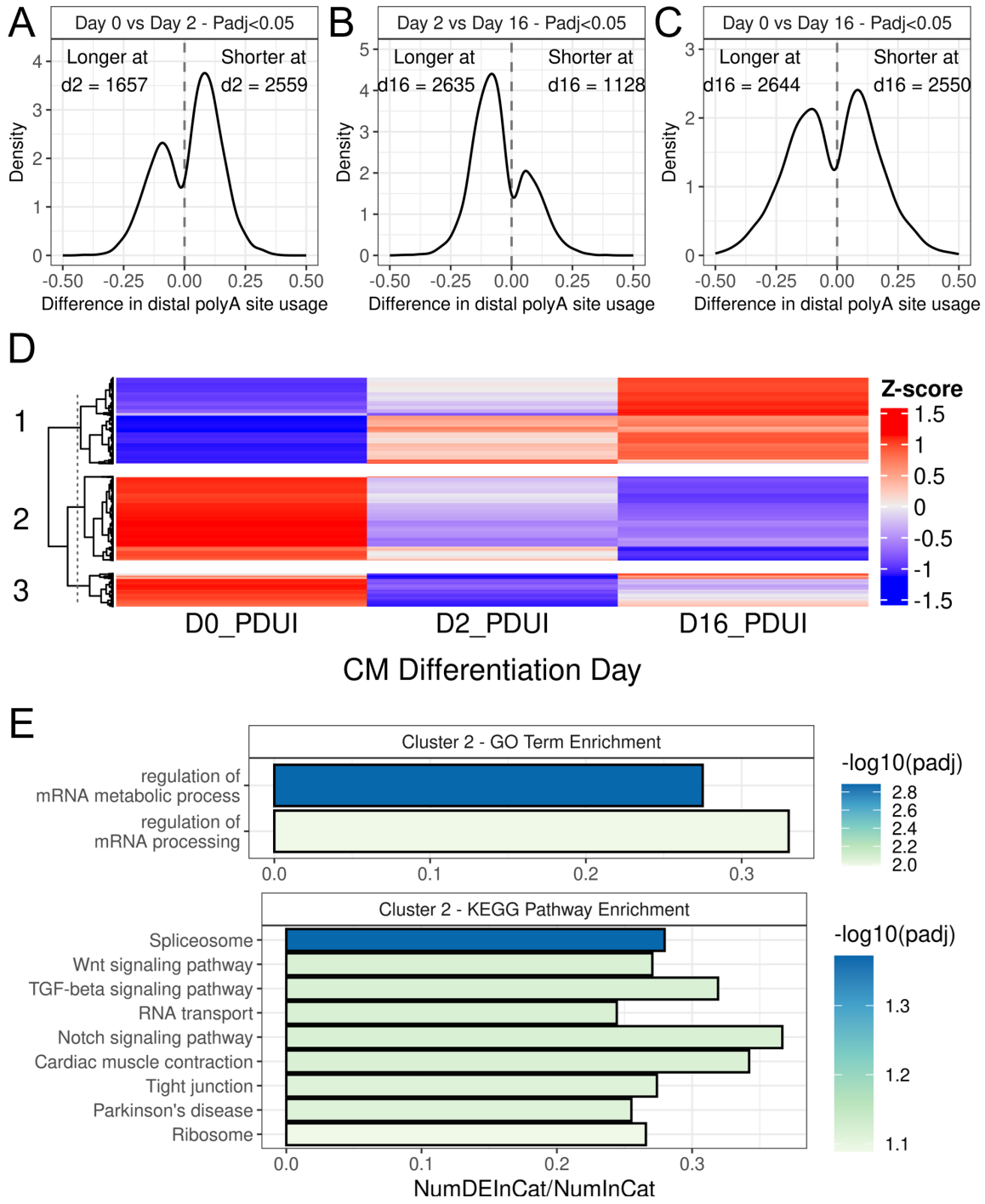

Figure 5.15: Differential alternative polyadenylation during cardiomyocyte differentiation. A-C) Difference in distal polyA site usage (PDUI) for significant events (padj<0.05). D) K-means clustering of PDUI z-scores across cardiomyocyte differentiation reveals three clusters: 1) progressive 3′UTR lengthening; 2) progressive 3′UTR shortening; 3) intermediate 3′UTR shortening. E) GOBP and KEGG enrichment analysis of cluster 2 genes (progressive 3′UTR shortening during differentiation).

We next asked whether changes in the 3'UTR length of individual genes across differentiation, as a result of APA, were accompanied by changes in RNA stability. Whilst we had previously observed that an increasing 3'UTR length correlates with reduced RNA stability between genes, for genes that display differential APA we find that the amount of 3'UTR lengthening does not correlate with changes in RNA stability between either the day 0 vs day 2 comparison (Figure 5.16A; R=0.01; P=0.533), or the day 2 vs day 16 comparison (Figure 5.16B; R=-0.02; P=0.258).

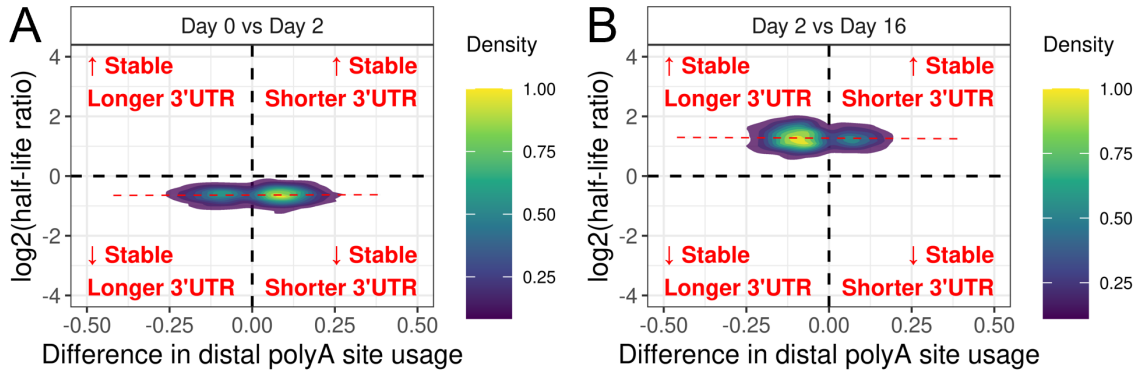

Figure 5.16: Lack of correlation between differential polyA site usage and half-life differences. 2D density plots represent the distribution of genes across the two variables. Comparison of differences in 3'UTR length (distal polyA site usage) versus differences in RNA stability between: A) Day 0 and Day 2; B) Day 2 and Day 16.

5.4 Summary

In this chapter we first established a cardiomyocyte differentiation model by trialing several protocols. Our initial attempts to establish a differentiation model using suspension culture on microcarriers, and with CDM3 media, resulted in high variability in differentiation efficiency and high levels of cell death. However, through the use of the STEMCELL Technologies Ventricular Cardiomyocyte Differentiation Kit, we were able to obtain a high proportion of CTNT+ cells at day 16. We also showed that these cells were enriched for multiple cardiac markers, whilst at day 2 cells were enriched for mesodermal markers.

Using this cardiomyocyte differentiation model, we subsequently conducted SLAMseq at day 0 (hESCs), day 2 (mesodermal), and day 16 (cardiomyocytes) using high-depth paired-end RNAseq. This allowed us to develop a bioinformatic pipeline to estimate gene-level half-lives at each time point, and conduct statistical testing for differences in decay rates between time points across cardiomyocyte differentiation. Whilst gene-level half-life estimation could have been conducted through the use of QuantSeq as part of the standard SLAMseq workflow, the use of high-depth paired-end RNAseq facilitates the scaling of our pipeline to the individual isoform level, which will be presented in Chapter 6.

By producing gene-level half-life estimates, we were able to group our estimates by the biological processes they contribute towards by conducting GSEA. This revealed that genes with housekeeping functions generally have more stable RNA, whilst genes with functions in transcriptional regulation and cell fate determination generally have less stable RNA. By comparing half-life estimates across differentiation, we observed a transient destabilization of RNA at day 2 of differentiation, followed by RNA stabilisation at day 16. A recent study by Mufteev et al. (2023) using RATE-seq showed RNA stabilization during the differentiation of hiPSCs to neuronal progenitor cells (NPCs), followed by RNA destabilization upon terminal differentiation into neurons. This highlights how cell-type specific RNA stability profiles are likely to be. In the case of neuronal cells, having less stable RNA may aid in the "bursts" of expression that are observed upon neuronal activation (see Arc example in Section 1.6.3). It is important to note that our day 2 mesodermal cells represent a substantially earlier cell-type than NPCs, which take 14 day to generate. In this regard, our day 16 cardiomyocytes may represent a similar level of maturity to these NPCs compared to the terminally differentiated neurons, which were differentiated from NPCs over a further three weeks. As such the transient destabilization we observed at day 2 may be linked to the relative potency of the cell population.

Our day 2 cells express mesodermal markers, but also still maintain expression of the

OSN-group self-renewal markers (OCT4, SOX2 and NANOG), highlighting how early in differentiation this population truly is. At this stage, the day 2 cells are highly proliferative, and capable of differentiating into any mesodermal lineage, depending on the growth factors they are exposed to. As such, having relatively unstable RNA would be a suitable mechanism to increase the "reaction time" of the cells. For example, where a new growth factor is encountered, which leads to a change in transcriptional output of the cell, it is important that this be reflected by the overall transcriptome rapidly, something that is made possible by having globally less stable RNA. Where RNA is very stable, these changes would take longer to be reflected, making the overall system more "laggy".

On the other hand, we showed that our day 16 cardiomyocytes have a reduced expression of genes with functions in RNA degradation, and have globally more stable RNA. These cells are terminally differentiated, although they may not be completely matured. As such, having a slower overall turnover of RNA makes the system more robust, whereby transient disturbances (e.g. due to stress) are less likely to have transcriptome-wide repercussions. Additionally, a slower turnover also reduces energy consumption due to less transcription and active degradation.

When we compared changes in RNA stability vs changes in gene expression between differentiation stages, we found that the major effect was on RNA stability, which appeared to change independently of gene expression (i.e. no correlation). It is important to note that in this comparison only the X-axis (gene expression) is normalized between conditions, as DESeq2 assumes that the average change in expression between samples is 0 (however individual genes can display change). Additionally, gene expression is relative to total RNA sequenced. On the other hand, half-life values are absolute, as they are calculated from the percentage of converted RNA (i.e. internally normalized per gene). Therefore, to determine whether increases or decreases in RNA stability lead to increased levels of each gene, we would need an absolute measure of gene

expression.

Finally, we showed that RNA half-life differences between genes are in part caused by differences in 3'UTR length and GC content. Additionally, genes that have more stable RNA tend to be expressed at higher levels. However, when we looked at how 3'UTR length changes during differentiation due to alternative polyadenylation, we found that the difference in 3'UTR length did not correlate with the difference in RNA stability observed between each differentiation state. This supports the recent findings of Fansler et al. (2024) who found that 3'UTR length changes independently of gene expression across several hundred cell types.

6. 3'UTR splicing regulates transcript stability during stem cell differentiation

In Chapter 5, we examined the global trend of RNA stability changes through hESC to cardiomyocyte differentiation at the gene-level. We identified a transient destabilization of RNA at day 2 of differentiation followed by RNA stabilization as differentiation proceeded. In this chapter we take this a step further by processing individual splicing events to investigate the effect of 3'UTR splicing on RNA stability by comparing between 3UI-spliced and 3UI-retaining reads. Additionally, we examine how the RNA stability of each isoform changes across differentiation. We hypothesised that the stability of 3UI-spliced and 3UI-retained isoforms may change at different rates during differentiation due to isoform-specific regulation, for example: differential expression of miRNAs that bind to retained 3UIs, or RBPs that bind to the 3UI-spliced isoforms. Therefore, we also examine the interaction between splicing and the stage of differentiation on RNA stability.

6.1 Estimating RNA half-lives at the individual event level

To assess the impact of 3'UTR splicing on RNA stability, we developed analyses similar to those presented in Section 5.2.2, except read pairs were assigned to individual splicing events as opposed to individual genes. Through our "read assignment" step using featureCounts, we already have information about which genes the reads map to. We next determined whether reads corresponded to our splicing events of interest, this was achieved through cross-referencing to the coordinates of 3'UTR introns. This step was necessary to ensure that we are only comparing reads that contain evidence of intron-splicing or intron-retention, and not those that map elsewhere within the transcript body. Reads were classed as spliced where they span the splice site, i.e. they start before the intron, end after the intron, and their mapped blocks match the intron coordinates. Reads were classed as retained where they contain the intronic sequence. This can be

present in multiple ways: 1) the read starts before the intron and ends within the intron; 2) the read starts within the intron and ends after the intron; 3) the read is entirely contained within the intron; 4) the intron is entirely contained within the read. As was the case for our gene-level script, where we have evidence of assignment to either the retained/spliced isoform for one read in the pair, the other read within the sequencing insert is also "pulled down" alongside, providing more sequence to detect T>C or A>G conversions. Our script subsequently outputs metadata on each read pair, the sample it is from, the events it is associated with, the isoform (spliced or retained), and the number of conversions observed.

Half-lives were subsequently calculated in R in the same manner as in Section 5.2.3, although this time on an individual event level. Here two half-lives were produced, one for the spliced isoform and one for the retained isoform. To determine whether 3'UTR splicing produced a significant change in half-life, we fit our data to the two models introduced in Section 5.2.5, except here in the combined parameters model the b value (decay rate) was set per isoform (Figure 6.1). This was only possible where there were a sufficient number of reads for both the spliced and retained isoforms, where this was not the case (e.g. in instance where introns were constitutively spliced, and there were few retained reads) then a decay curve would not be fit for both isoforms, and as such we could not conduct statistical interrogation. From the 6,337 nonPTC 3UI events in our transcript assembly, we were able to fit decay curves for both the spliced and retained isoforms for 826 events at day 0, 1,174 at day 2, and 805 at day 16. Of these 151, 260, and 129 displayed significant differences in decay rates between the 3UI-spliced and 3UI-retaining isoforms at day 0, 2, and 16 of differentiation, respectively.

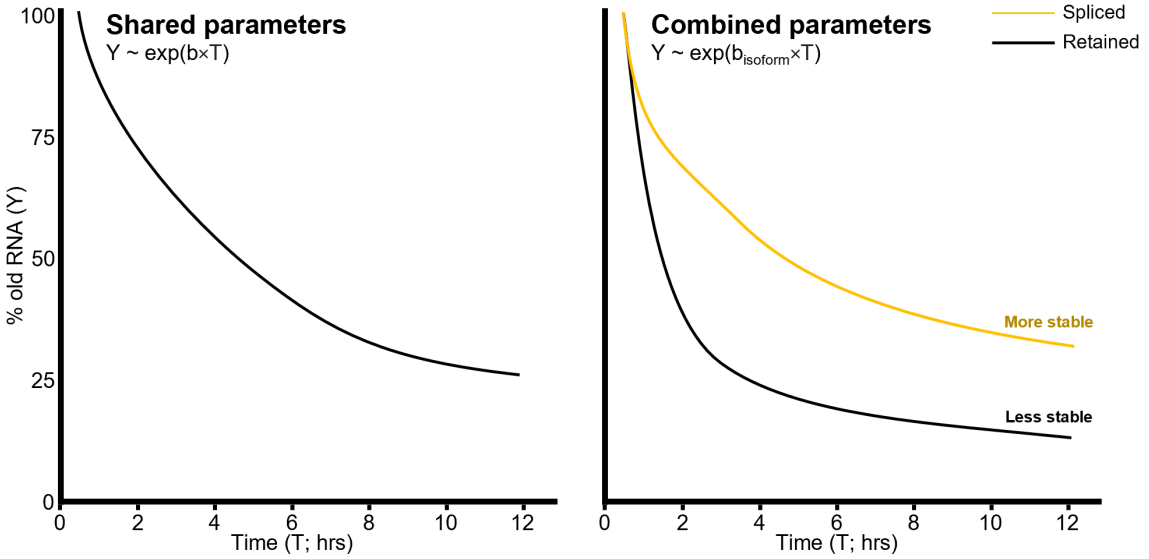

Figure 6.1: Schematic depicting the calculation of P-values for differential decay rates between splice partners. In the shared parameters model the b value (strength of the slope) is fit to all the data. In the combined parameters model two b values are generated b[spliced] and b[retained] effectively fitting two lines. P-values are calculated by conducting an ANOVA test between the two models.

In addition to calculating P-values to distinguish spliced and retained isoform, we also calculated the standard deviation (SD) of decay for each isoform. To do this we fit a normal distribution of the data points at each chase time point, randomly sample from this distribution, and produce thousands of bootstrap replicates, thus allowing us to calculate the SD of b. Examples of these statistical methods applied to two examples are shown in Figure 6.2. In Figure 6.2A we found that splicing the HNRNPA2B1 3'UTR results in a significant increase in RNA stability, where both the decay curves have a relatively low SD. In Figure 6.2B we found that splicing the SRSF2 3'UTR results in a significant decrease in RNA stability, where the decay curve of the spliced isoform appears to be more varied than the retained isoform. By correlating across all events and isoforms, we found that as isoforms decay faster (i.e. become less stable), that the associated SD generally increases ($R=0.67$; $P<0.0001$), in line with more variability between replicates at the later chase time points.

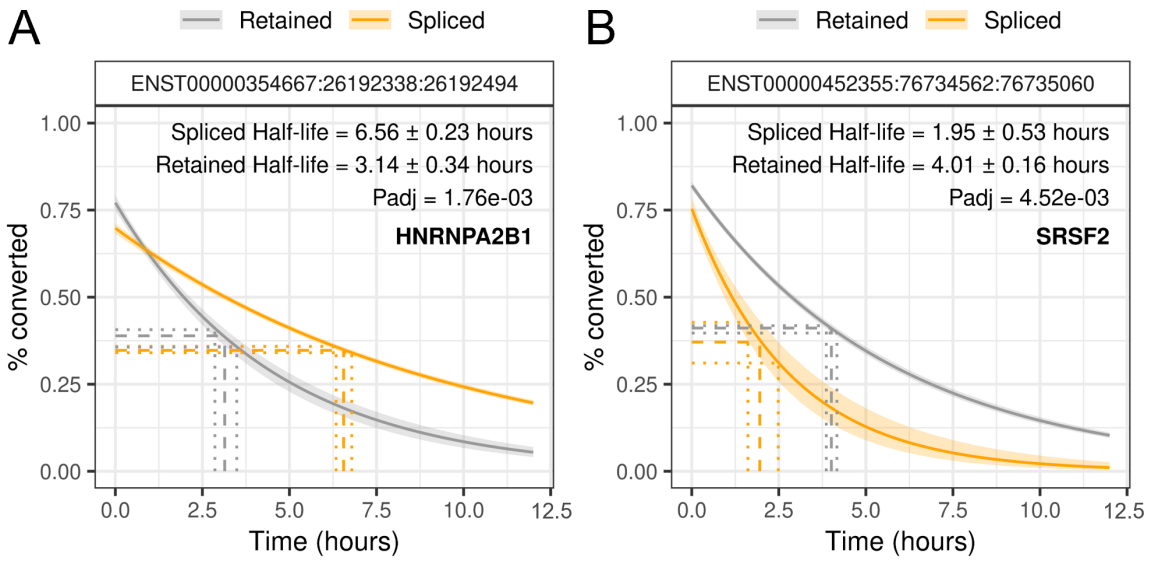

Figure 6.2: Examples of 3'UTR splicing impacting RNA stability. Comparison of 3UI-spliced vs 3UI-retained isoform decay curves at day 2 of differentiation for A) HNRNPA2B1, and B) SRSF2. In each panel the orange line represents the decay of the 3UI-spliced isoform, whilst the grey line represents the decay of the 3UI-retaining partner isoform. Where dashed lines intersect the x-axis represents the median half-life from bootstrapped estimates, dotted lines represents + or - 1 SD.

6.1.1 Splicing 3'UTR introns impacts RNA stability differently from CDS introns

In addition to calculating half-lives for 3'UTR introns, we also calculated half-lives for all introns in protein coding genes, allowing us to compare the effect of 3'UTR splicing vs CDS splicing on RNA stability. We calculated spliced to retained half-life ratios for events where splicing significantly changed RNA stability, allowing us to compare the distribution of ratios between intron groups (Figure 6.3). A positive $log_2(\frac{spliced\ half-life}{retained\ half-life})$ value indicates the spliced isoform is more stable, whilst a negative value indicates the retained isoform is more stable. We found the splicing CDS introns results in RNA stabilization in the majority of cases (day 0 = 92.1%; day 2 = 74.4%; day 16 = 89.9%), whilst intron retention is usually destabilizing (Figure 6.3A). On the other hand, splicing 3'UTR introns results in a more even split between stabilization (day 0 = 56.3%; day 2 = 51.2%; day 16 = 53.5%) and destabilization of RNA (Figure 6.3B). This near-even split between RNA stabilization

and destabilization by 3'UTR splicing is also demonstrated in Figure 6.4A by plotting the spliced half-life estimate directly against the retained half-life estimate for each event, the results of which are subsequently quantified in Figure 6.4B.

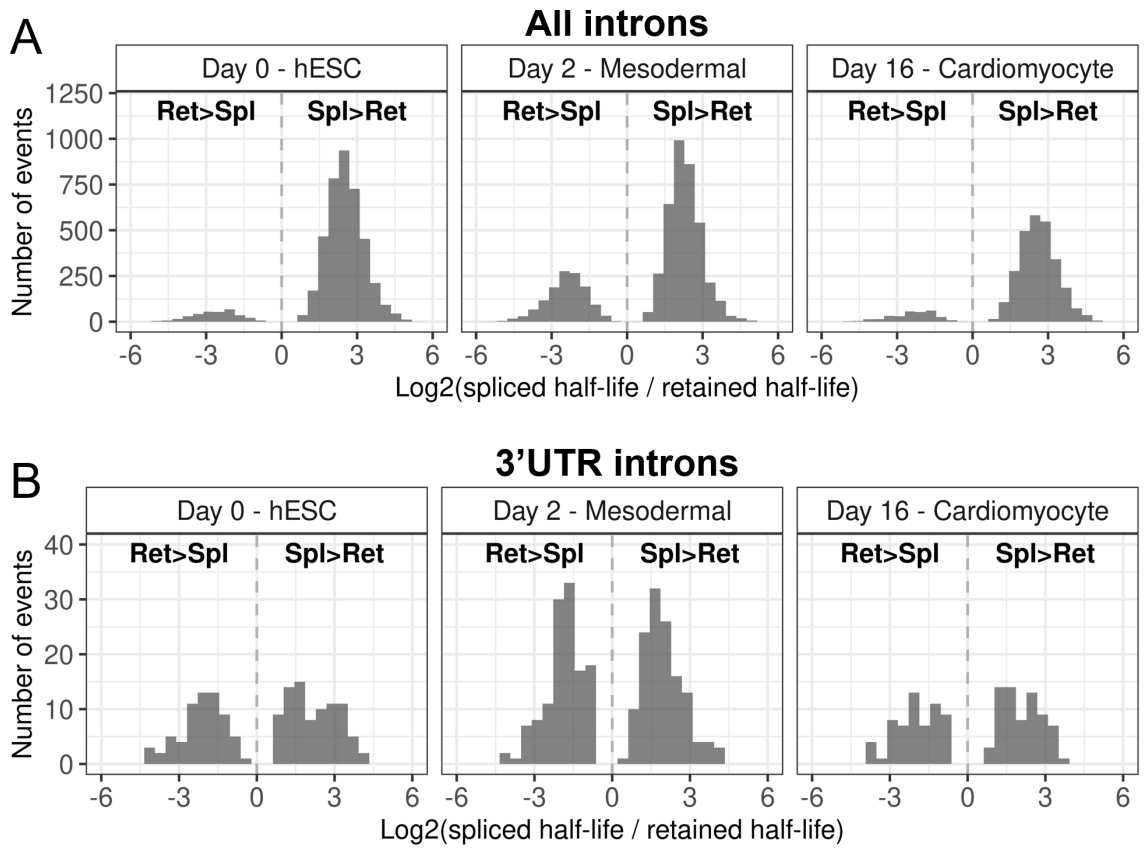

Figure 6.3: Effect of splicing 3'UTR introns vs all introns on RNA stability. Histograms plotting the distribution of intron-spliced to intron-retained half-life ratios for events that have significantly different decay-rates (adjusted P-value < 0.05) for: A) all introns; B) nonPTC 3'UTR introns. Log2(spliced half-life / retained half-life) values: >0 = spliced isoform is more stable than the retained isoform; <0 = retained isoform is more stable than the spliced isoform.

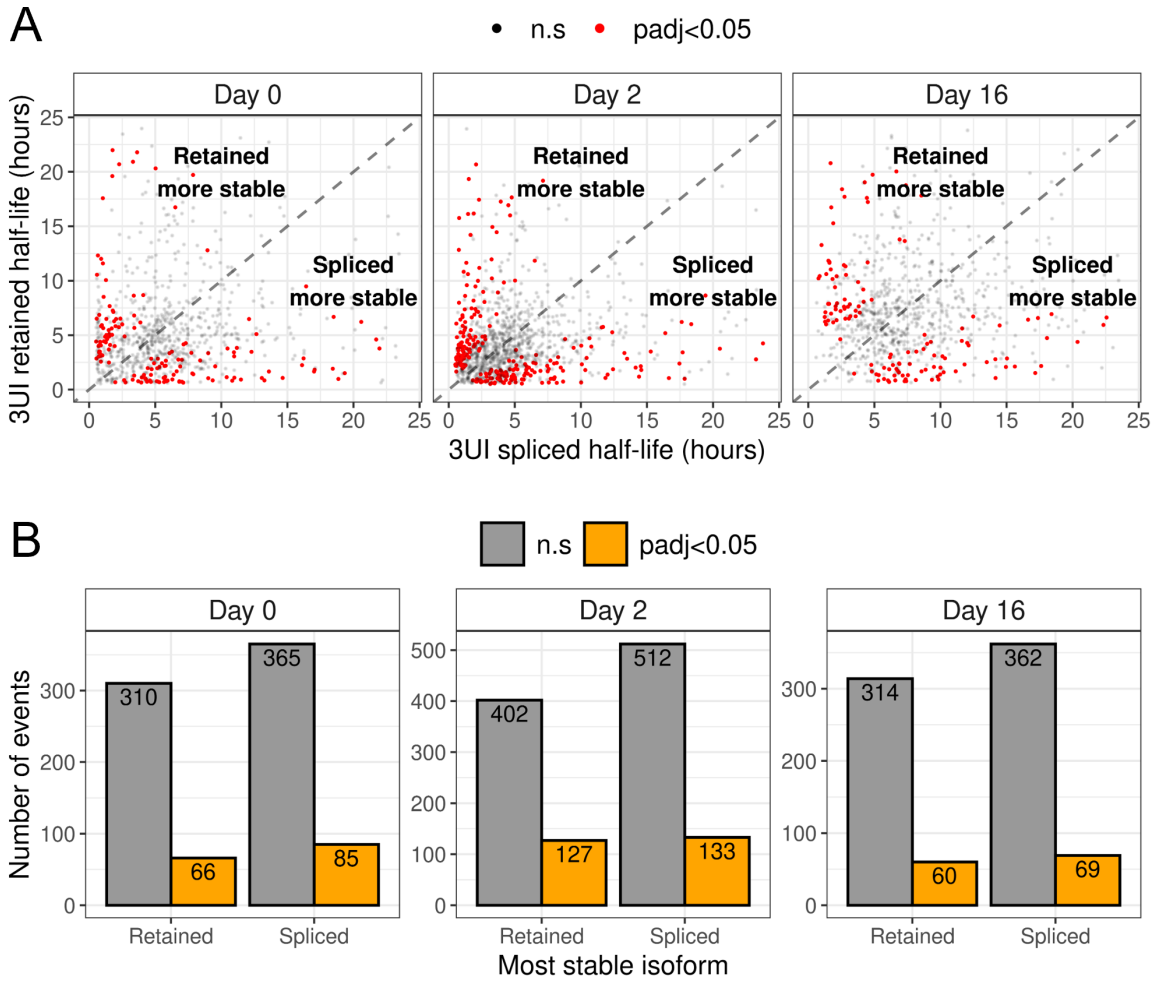

Figure 6.4: Comparison of 3UI-spliced vs 3UI-retaining half-lives. A) dot plots of 3UI-spliced isoform half-life estimates (X-axis) vs 3UI-retained isoforms (Y-axis) at day 0, day 2, and day 16 of differentiation. Red dots indicate events where the decay rates are significant different (adjusted P-value < 0.05), remaining non-significant differences are plotted as grey dots. B) Quantification of events in A.

6.1.2 Impact of 3'UTR splicing on RNA stability is partially explained by predicted NMD-sensitivity

We next investigated whether 3'UTR splicing having a stabilizing or destabilizing effect was due to the position of the splice site within the 3'UTR relative to the 55nt threshold for NMD-sensitivity. In line with our findings in Section 3.4.2.2 we hypothesised that 3'UTR splicing events more than 55nt from the stop codon would be NMD sensitive and

therefore result in RNA destabilization. In Figure 6.5 we plot the distribution of $log_2(\frac{spliced\,half-life}{retained\,half-life})$ values for events that are more than (orange) or less than 55nt (grey) from the stop codon. We found that 3'UTR splicing taking place less than 55nt from the stop codon is more often stabilizing than destabilizing (day 0 = 58.4%; day 2 = 56.7%; day 16 = 58.7%). On the other hand, and in line with our hypothesis, we found that 3'UTR splicing taking place more than 55nt from the stop codon is more often destabilizing than stabilizing (day 0 = 55.9%; day 2 = 51.9%; day 16 = 60.8%). To test the statistical significance of these trends we conducted Kolmogorov–Smirnov tests, which revealed a high degree of statistical significance at day 0 (D=0.156; P=0.0045) and day 16 (D=0.21; P=0.000039), but not at day 2 (D=0.091; P=0.117). This trend in significance matches the global half-life differences presented in Figure 5.10, where significance is lowest at day 2 where RNA is also least stable. This could suggest that where RNA is already unstable NMD may have less of an effect. Together these results indicate that 3'UTR splicing taking place more than 55nt from the stop codon is generally destabilizing. However, there is a substantial proportion of 3'UTR splicing events that do not follow this trend, further supporting our findings in Section 3.4.2.2 that many of these events may not be NMD sensitive.

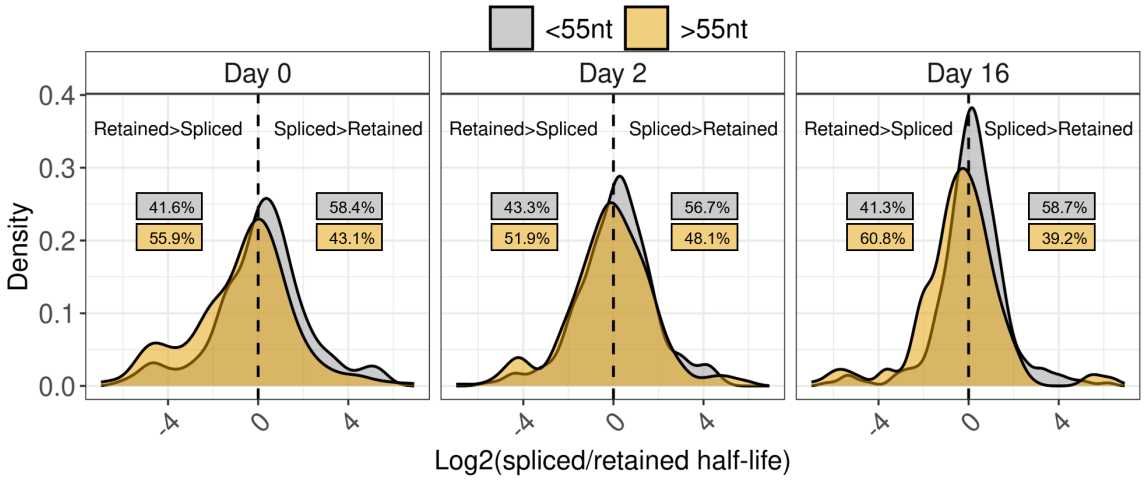

Figure 6.5: Differences in RNA stability are partially explained by distance from stop codon. Density plots showing the distribution of Log2(spliced half-life / retained half-life) values for nonPTC 3UI splicing events that occur less than 55nt from the stop codon (grey) and more than 55nt from the stop codon (orange). Percentages of each distribution that are above or below 0 are depicted in the boxes of the matching colour.

6.2 Stability of 3UI spliced and retained isoforms changes during differentiation

To investigate whether the stability of 3UI-spliced and 3UI-retaining isoforms changes during differentiation, we subjected our decay curves to statistical interrogation in the same manner as in Figure 5.8, except on the isoform-level as opposed to the gene-level. We were able to fit decay curves at all three differentiation time points for 5,947 isoforms, consisting of 3,406 3UI-retaining, and 2,541 3UI-spliced isoforms. We found that 1,322 3UI-retaining isoforms (in 836 genes), and 632 3UI-spliced isoforms (in 275 genes), displayed significant variation in RNA stability across differentiation at the Padj<0.05 cutoff. Applying a more stringent P-value threshold (Padj<0.01), we found that 867 3UI-retaining, and 347 3UI-spliced isoforms (in 580 and 163 genes, respectively), displayed significant variation in RNA stability across differentiation (Figure 6.6A). Interestingly, genes with transcripts displaying differential stability across differentiation were enriched for biological processes including "RNA 3'-end processing" (Padj=0.009),

"regulation of mRNA stability" (Padj=0.016), and "regulation of RNA splicing" (Padj=0.023).

Given that our gene-level analysis of RNA stability across differentiation revealed a transient destabilizing phenomena at day 2, we also expected to see this at the isoform-level. However, we were interested to determine whether the relative stability of 3UI-spliced and 3UI-retained isoforms would cluster during differentiation due to a bias in regulation of either isoform. For example, if the overall effect observed at the gene level were primarily due to differential miRNA expression, we would expect to see differences in the 3UI-retaining isoforms, but not in the 3UI-spliced isoforms. On the other hand, if the overall effect were due to splicing-related changes to the mRNP composition (e.g. spliceosomally deposited/EJC-associated), then we would expect to see differences in the 3UI-spliced isoforms, but not in the 3UI-retaining isoforms. Following K-means clustering, we did not observe a clear bias towards either isoform, as shown by the "bee-stripe"-like pattern in the heatmap annotation for isoforms in Figure 6.6A. Instead our isoform-level analysis showed similar clustering to our gene-level analysis, including transient destabilization at day 2 and stabilization at day 16 (Figure 6.6B).

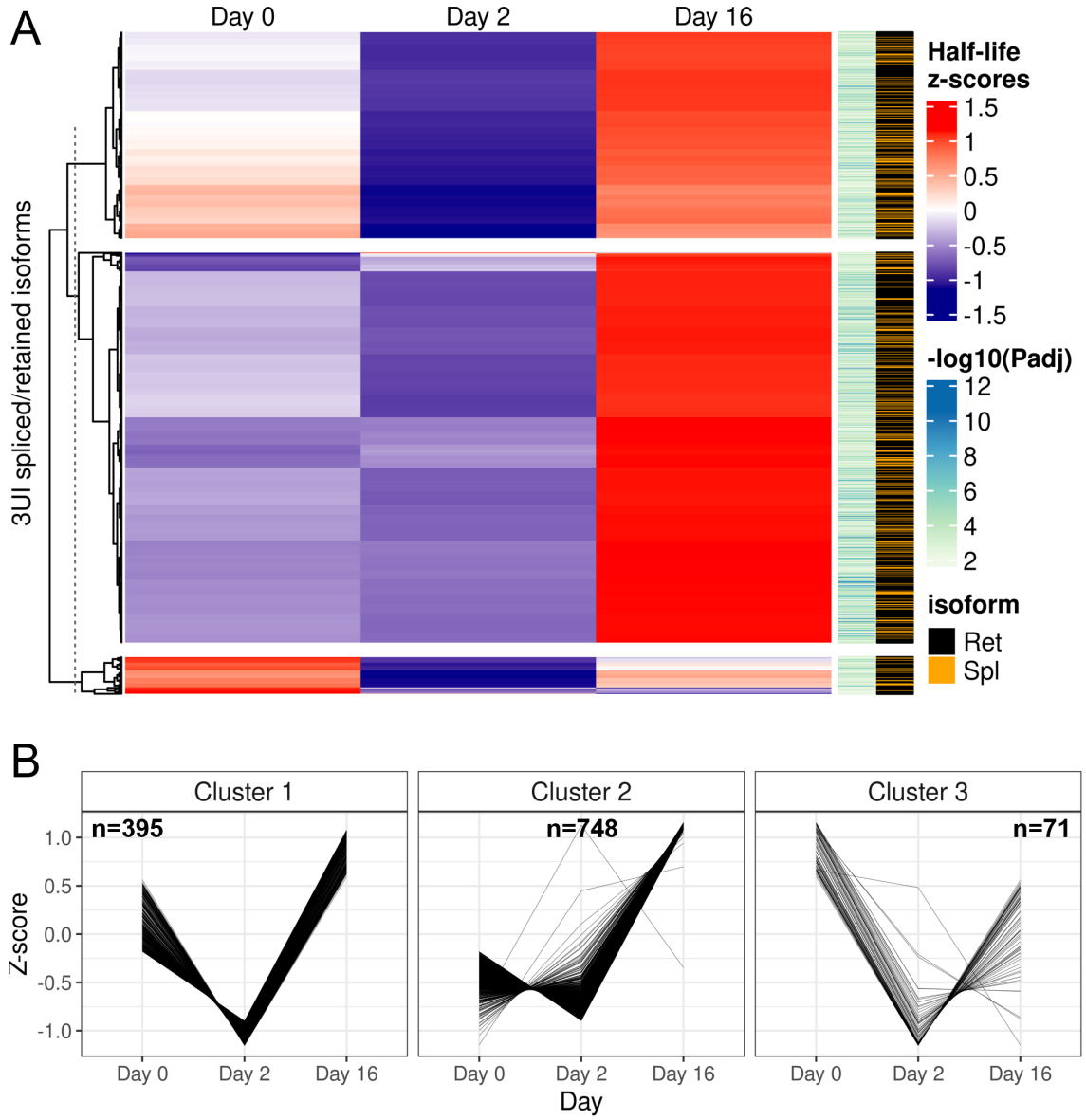

Figure 6.6: Half lives of 3UI spliced/retained isoforms changes during differentiation. A) Heatmap showing the z-scores of 3UI-spliced or 3UI-retained isoform half-lives at day 0, 2 and 16 of cardiomyocyte differentiation. If a cell is coloured red it indicates that half-life is the highest for that isoform at that time of differentiation, whilst blue is the lowest. Adjusted P-values (comparing decay rates between stages of differentiation) are represented in the first heatmap annotation on the right (light green to dark blue), where higher values (darker blue) indicate a smaller, more significant P-value. Isoforms are annotated in either black (3UI-retained) or orange (3UI-spliced) on the second heatmap annotation on the right. B) Line graphs showing the trends in z-scores for each cluster shown in A.

6.3 3'UTR splicing has a differential impact on RNA stability as differentiation proceeds

We next asked whether the act of 3'UTR splicing changes RNA stability between pairs across differentiation. As such this represents the interaction between stability differences and the stage of differentiation, and allows us to identify stage-specific regulation. For example, hypothetical "Gene A" has a 3UI, which when spliced out increases stability 2-fold at day 0, 2-fold at day 2, and 2-fold at day 16. Meanwhile, "Gene B" also has a 3UI, but when spliced out at day 0 produces a 2-fold increase, at day 2 produces a 3-fold increase, and at day 16 produces a 4-fold increase, i.e. the act of splicing is becoming progressively more stabilising as differentiation proceeds. Our approach aims to identify genes that behave like "Gene B". Additionally, through this approach we would also be able to identify genes where the effect of 3'UTR splicing on stability is progressively more destabilising, or where there is an intermediate effect. In its most basic form, we can investigate this through examining how spliced/retained half-life ratios change across differentiation (Section 6.3.1). We then develop analyses to interrogate the significance of this effect, allowing us to identify candidate events for potential future study (Section 6.3.2).

6.3.1 Identifying differentiation stage-specific regulation

We were able to fit decay curves for both 3UI-spliced and 3UI-retained isoforms across all three time points of differentiation for 629 events. Next we proceeded to conduct K-means clustering on the spliced/retained half-life ratios across differentiation. Producing an elbow plot of the within cluster sum of squares (WCSS) versus potential K values indicated that three clusters should be produced (Figure 6.7A). Subsequently, K was set as 3, and ratios were clustered accordingly (Figure 6.7B-C). This revealed instances where splicing the 3'UTR produced a disproportionate change in RNA stability at a given time point of differentiation compared to the whole time-course. For example,

events in Cluster 1 have the highest spliced/retained half-life ratio at day 2 of differentiation, whilst those in Cluster 2 are highest at day 16, and those in Cluster 3 are highest at day 0 (Figure 6.7B-C).

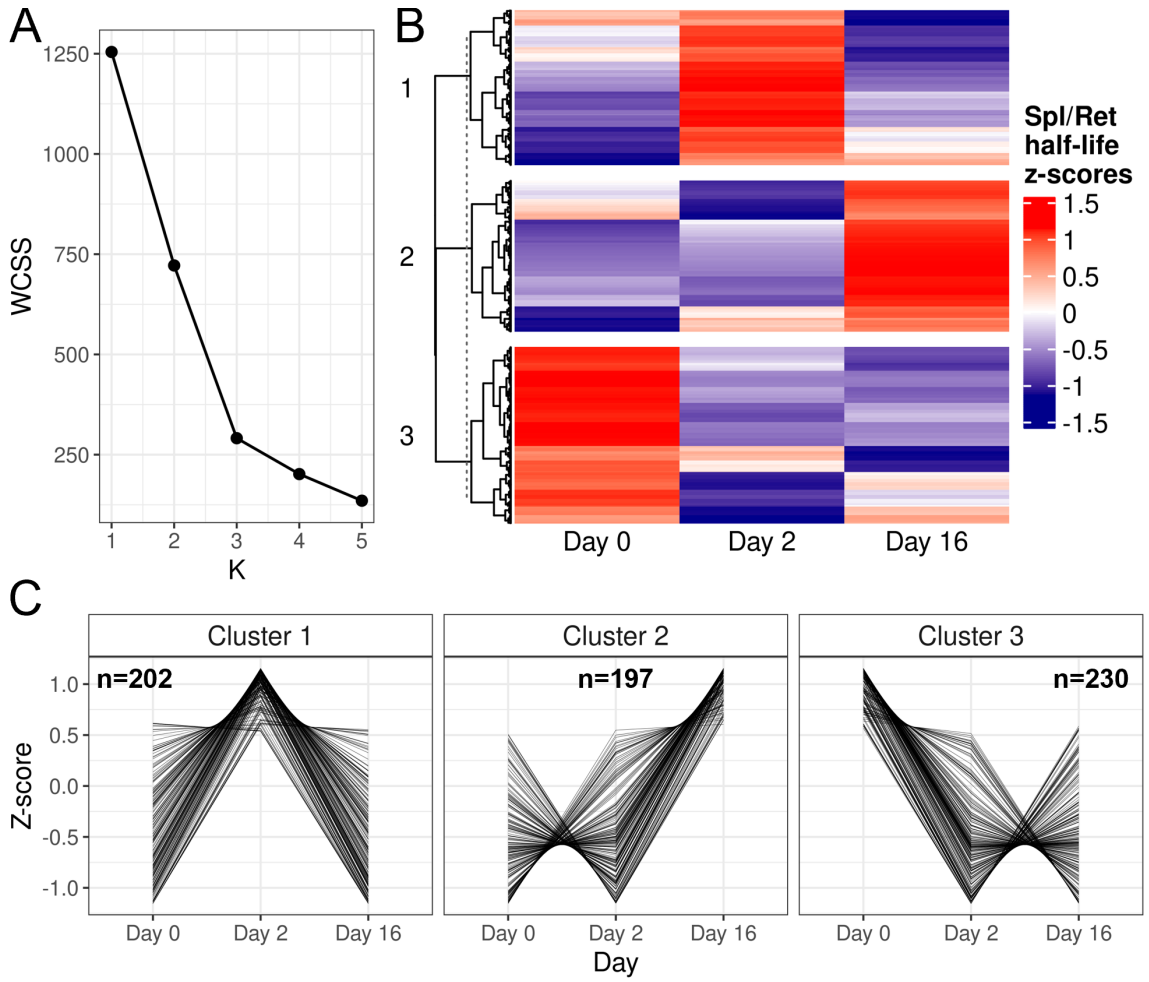

Figure 6.7: 3UI-spliced/3UI-retained half-life ratios change during differentiation. A) Elbow plot showing the effect of K value selection on the within-cluster sum of squares (WCSS) value. The value of k is selected based on the asymptote (here k=3). B) Heatmap showing the z-scores of 3UI-spliced/3UI-retained half-life ratios at day 0, 2 and 16 of cardiomyocyte differentiation. C) Line graphs showing the trends in Z-scores for each cluster shown in B.

The presence of these trends in spliced/retained half-life ratios suggests that the events in each cluster may be subjected to stage-specific regulation, potentially due to differential expression of trans factors such as miRNAs and RBPs. We hypothesised that such

stage-specific regulation may regulate important biological processes; however, upon conducting gene ontology analysis for each cluster we did not observe any significant ontology terms. To gain insight into the trans factors that may be regulating the stability of these events in a stage-specific manner, we conducted RBP and MRE enrichment analysis to identify trans factors that are enriched in the 3UIs within each cluster compared to all nonPTC 3UIs. This allowed us to identify the RBPs (Figure 6.8) and miRNA binding events (Figure 6.9) that are modulated due to 3'UTR splicing, and compare this to the trends in spliced/retained half-life ratios in Figure 6.7.

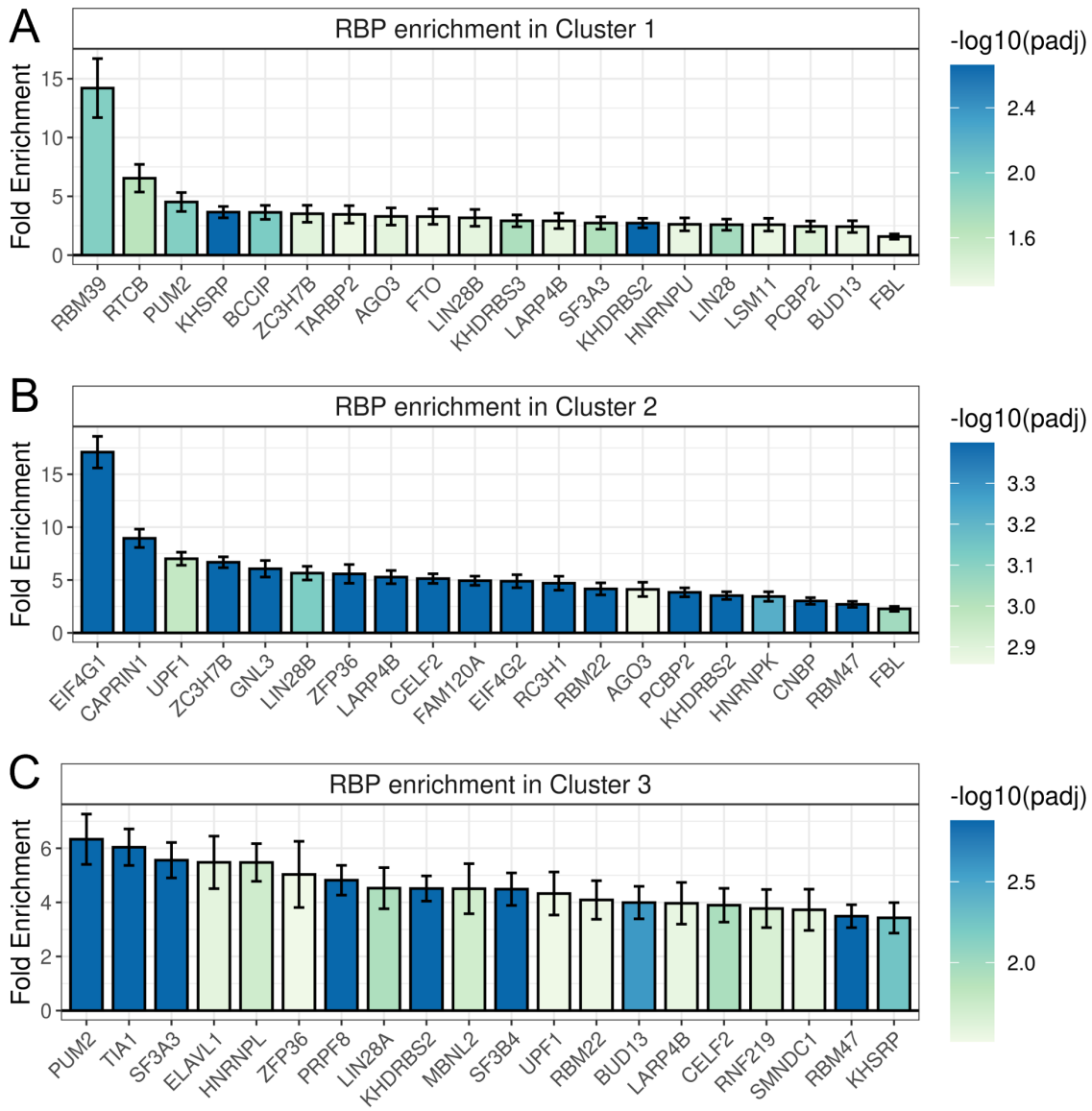

Figure 6.8: Enrichment of RBP binding to 3UIs in each interaction cluster. Enrichment of RNA binding protein binding (from ENCORI RBP-CLIPseq data) within the intron for the events clustered in Fig 6.7B: A) Cluster 1; B) Cluster 2; C) Cluster 3.

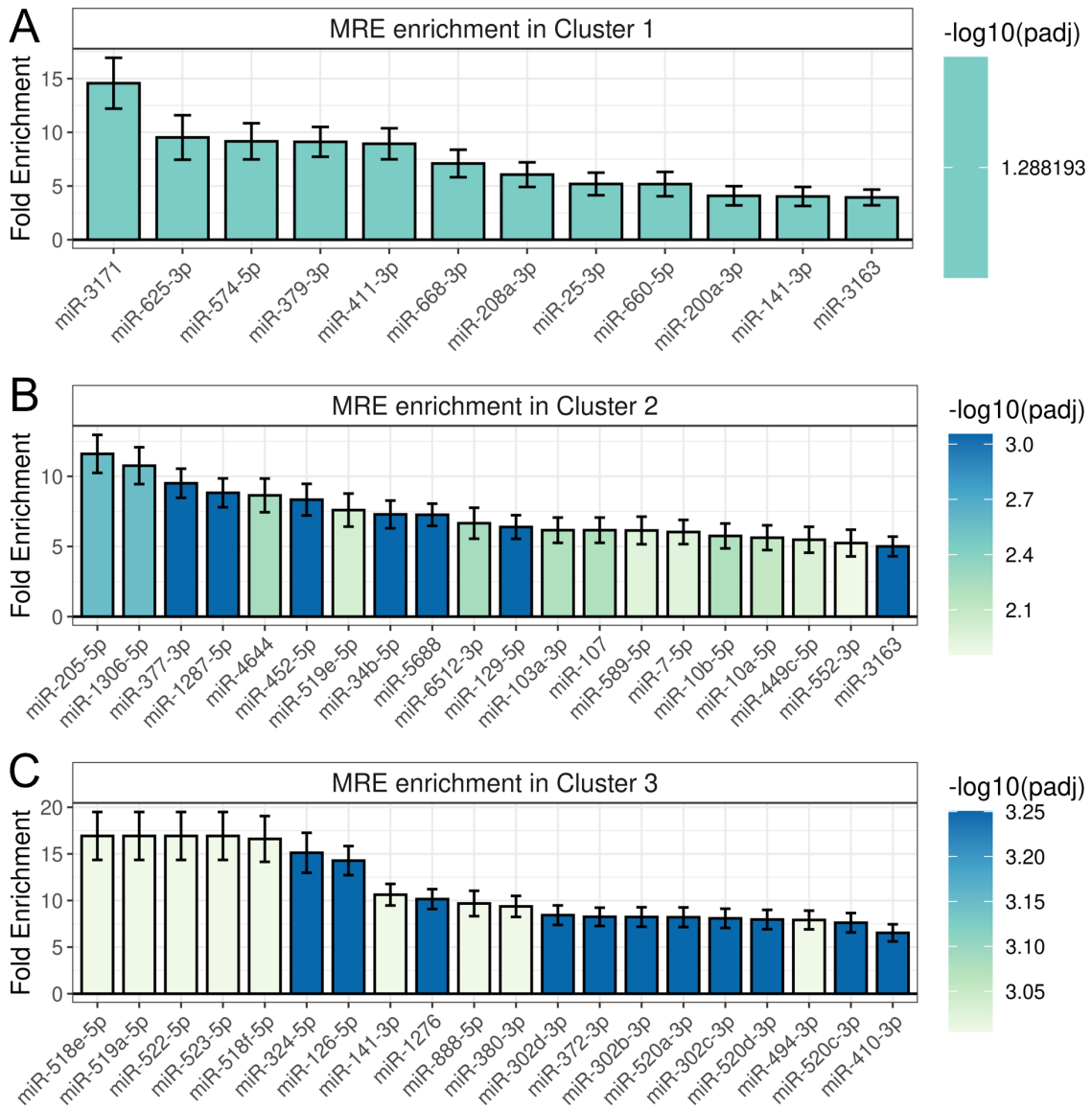

Figure 6.9: Enrichment of miRNA binding to 3UIs in each interaction cluster. Enrichment of miRNA binding (from ENCORI AGO-CLIPseq data) within the intron for the events clustered in Fig 6.7B: A) Cluster 1; B) Cluster 2; C) Cluster 3.

We observed significant enrichment of 20, 64, and 30 RBPs in Clusters 1, 2 and 3, respectively, the top 20 of which are presented in Figure 6.8. We also observed significant enrichment of 58 and 183 MREs in Clusters 2 and 3, respectively, whilst Cluster 1 had 12 candidate miRNAs with Padj=0.0515 (Figure 6.9). The most notable finding from this

analysis is that each cluster of events, subjected to differentiation stage-specific regulation, displays enrichment of a unique subset of RBPs and miRNAs. We hypothesised that these RBPs and miRNAs might display differential expression or binding efficiency to accompany the trends in Figure 6.7. Whilst we could not address this question for miRNAs without miRNA-seq data, we next explored how expression of each cluster's enriched RBPs changed during differentiation by interrogating gene-level quantification from our RNAseq data (Figure 6.10). We identified instances where the RBPs were most highly expressed at the same differentiation stage that the 3UI-spliced/3UI-retained half-life ratio was highest (Figure 6.10A-C red lines), highlighting a potential link. Interestingly, UPF1 was one such gene in Cluster 3, where the 3UI-spliced/3UI-retained half-life ratio was highest in undifferentiated hESCs (Figure 6.10C). Given that enrichment analysis was conducted on the 3UI sequence (i.e. enriched within the 3UI-retaining isoform), UPF1 occupancy would be linked to EJC-independent functions. HNRNPL, a gene with potential functions in NMD evasion (Section 1.5.5), was also identified in this subset, in addition to ELAVL1 (also known as HuR; Section 1.2.1.1).

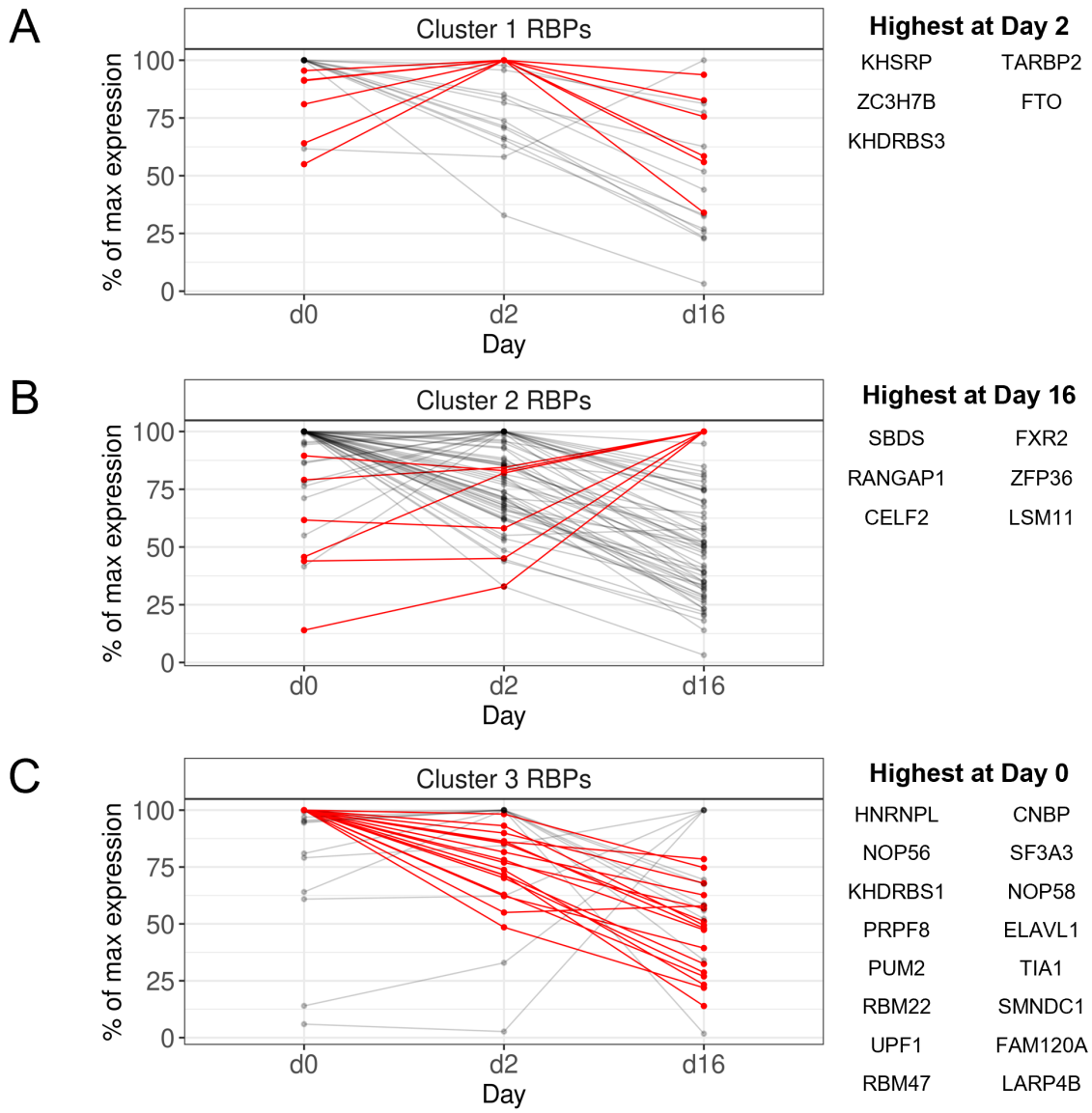

Figure 6.10: Differential expression of RBPs from each interaction cluster. Changes in gene expression for the RBPs in each cluster in Figure 6.8. RBPs which display their highest level of expression coinciding with the highest 3UI-spliced/3UI-retained half-life ratio from the matched cluster in Figure 6.7B are coloured red and listed next to each pane. A) Cluster 1 RBPs (where 3UI-spliced/3UI-retained half-life ratio was highest at day 2). B) Cluster 2 RBPs (where 3UI-spliced/3UI-retained half-life ratio was highest at day 16). C) Cluster 3 RBPs (where 3UI-spliced/3UI-retained half-life ratio was highest at day 0).

6.3.2 Testing significance of the interaction between differentiation stage and RNA stability differences

To identify instances where the interaction between the differentiation stage and RNA stability differences (between 3UI-spliced and 3UI-retaining isoforms) was significant, we developed a novel statistical approached based around the comparison of two non-linear models via ANOVA testing, building upon those introduced in Figure 5.8 (between differentiation days) and Figure 6.1 (between isoforms). Our reference model ("Fixed interaction" model; Figure 6.11A left) fits three different day-specific decay rates, and one fixed offset to represent the difference between the 3UI-spliced and 3UI-retaining isoforms. This fixed offset is based on the average difference in stability across all three differentiation time points. Our test model ("Free interaction" model; Figure 6.11A right) also fits three day-specific decay rates, however, it also fits a further three day-specific offsets representing the difference in stability between 3UI-spliced and 3UI-retaining isoforms at each stage of differentiation. Therefore where the difference of RNA stability differences is large, an ANOVA test between the the reference and test model will result in $P<0.05$. Where the difference of RNA stability differences is low, then the P value will be larger, and where it is null, i.e. where the spliced/retained stability ratio is exactly the same at all three differentiation time points, then P will equal 1.

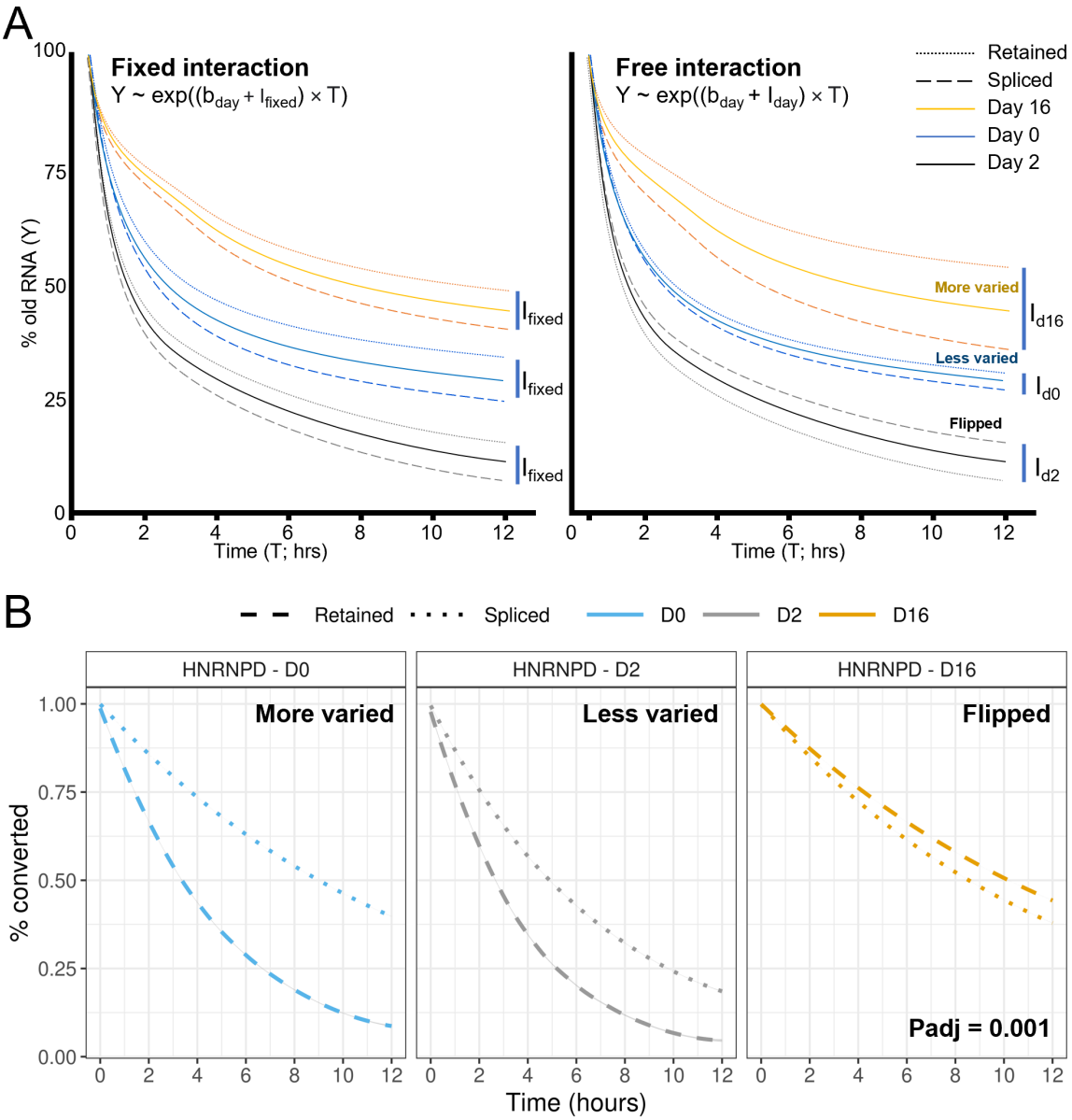

Figure 6.11: Schematic depicting the calculation of P-values for the interaction between 3'UTR splicing and differentiation time, on decay rate. A) In the fixed interaction model a b value (strength of slope) is calculated for each day of differentiation. A fixed interaction is then added to each b value, to represent the average difference between 3UI-spliced and 3UI-retained isoform decay rates, across differentiation. In the free interaction model a b value (strength of slope) is calculated for each day of differentiation. Individual interactions are then added to each b value, to represent the difference between 3UI-spliced and 3UI-retained isoform decay rates at each stage of differentiation. P-values are calculated by conducting an ANOVA test between the two models. B) Example usage in HNRNPD.

A specific example of a significant interaction is shown in Figure 6.11B for HNRNPD. At day 0 of differentiation the 3UI-spliced isoform is more stable than the 3UI-retaining isoform. The spliced isoform is still more stable than its retaining partner at day 2, although the difference is reduced. However, at day 16 this relationship is flipped, and whilst both isoforms are more stable than at day 0 and day 2, at day 16 the 3UI-retaining isoform is more stable the its 3UI-spliced partner. Applying this approach to all events where both 3UI-spliced and 3UI-retaining isoforms could be fit at all three differentiation time points, and following P-value adjustment, we observed 25 3'UTR splicing events with a significant interaction between half-life ratio and differentiation stage.

These events were subsequently clustered via K-means clustering into three clusters (Figure 6.12). Cluster 1 contained five 3'UTR splicing events (Figure 6.12A), which display the highest ratios at day 16 of differentiation (Figure 6.12B). This cluster included cyclin G1 (CCNG1), poly(ADP-ribose) glycohydrolase (PARG) and integrin subunit β3 binding protein (ITGB3BP). Cluster 2 contained eight 3'UTR splicing events, which display an intermediate effect, where the half-life ratios are highest at day 2 of differentiation (Figure 6.12B). This cluster included one of two splicing events in the 3'UTR of the splicing factor SRSF2, as well as two of three splicing events in the 3'UTR of TATA-box binding protein associated factor TAF1D. Finally, Cluster 3 contained 12 3'UTR splicing events, which had the highest half-life ratios at day 0 of differentiation (Figure 6.12B). This cluster included heterogeneous nuclear ribonucleoproteins A2B1 (HNRNPA2B1) and D (HNRNPD), SRSF2, TAF1D and nuclear export factor Exportin 1 (XPO1). A breakdown of the half-lives for both 3UI-spliced and 3UI-retaining isoforms across differentiation for all 25 of these significant interaction events is presented in Figure 6.13.

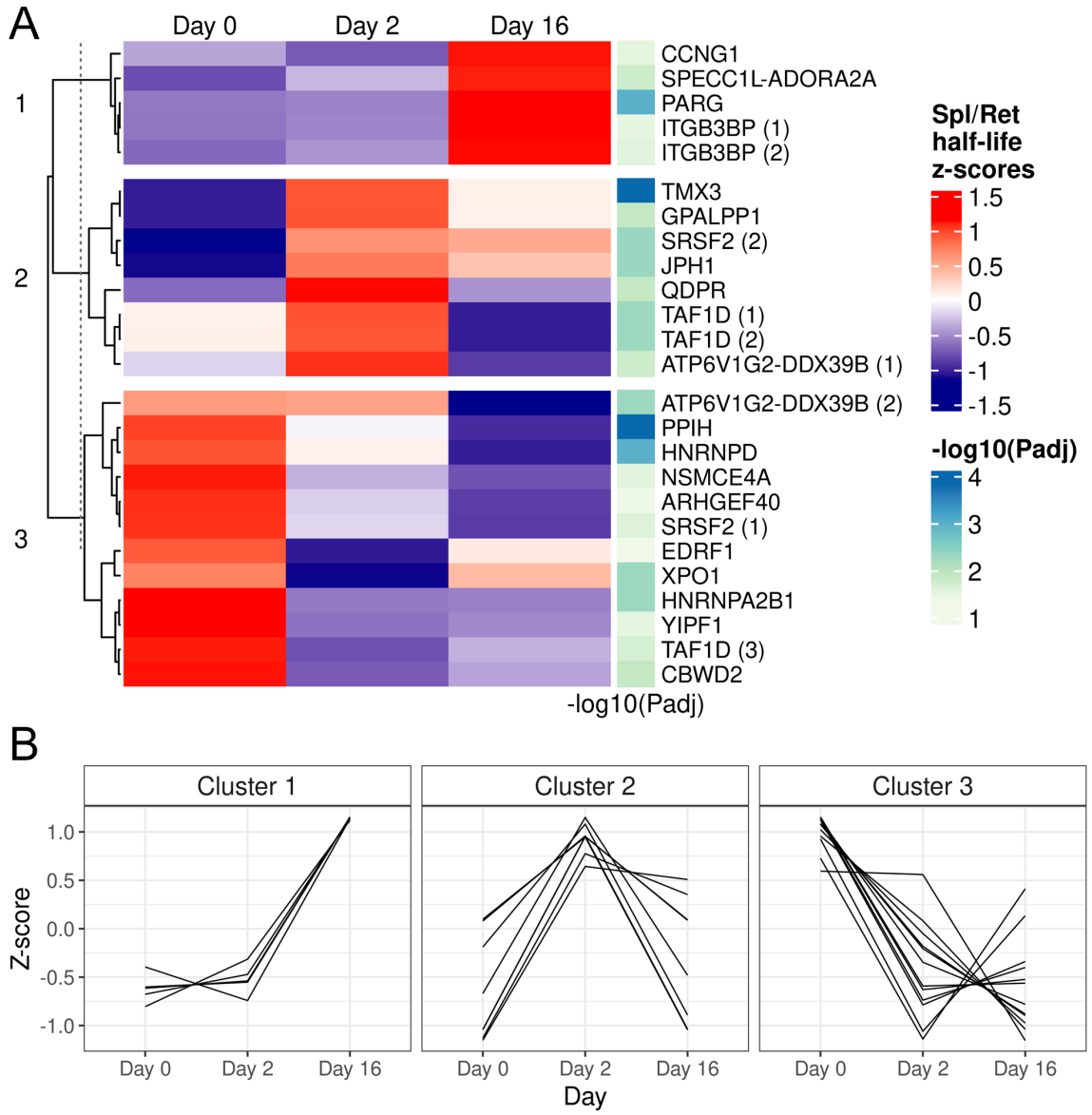

Figure 6.12: Events where there was significant interaction between 3'UTR splicing and differentiation stage, on decay rate. A) Heatmap show the z-scores of 3UI-spliced/3UI-retained half-life ratios at day 0, 2 and 16 of cardiomyocyte differentiation. Only events where adjusted P-value < 0.05 are shown. Where the interaction is more significant, the events will have a higher -log10(Padj) value, and are coloured darker blue on the heatmap annotation to the right. B) Line graphs showing the trends in z-scores for each cluster shown in B.

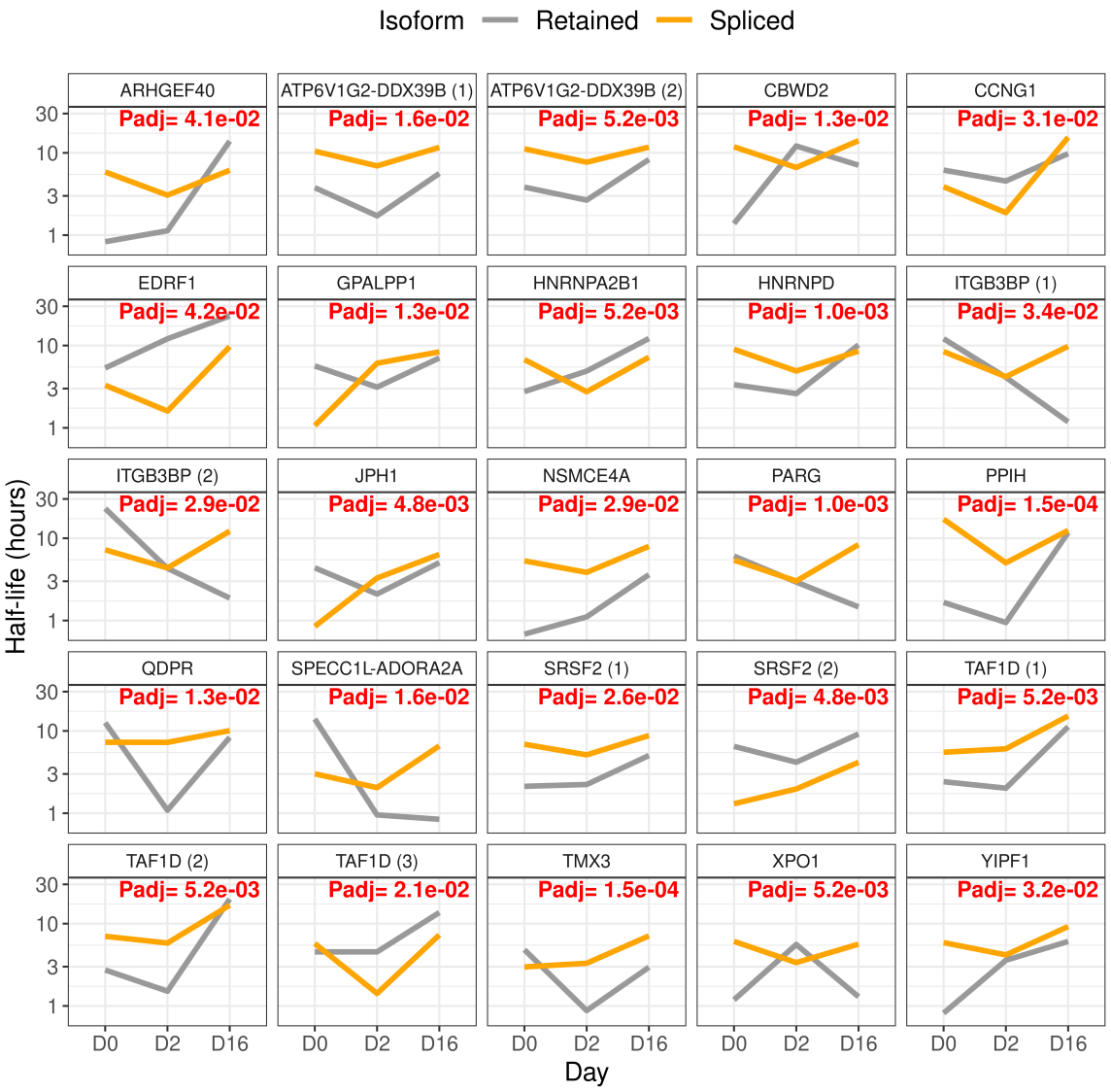

Figure 6.13: Change in half-life differences during differentiation for events with significant interaction. Only events which have a statistically significant interaction (adjusted P-value < 0.05) are shown, i.e. events that show a significant difference in 3UI-spliced/3UI-retained half-life ratio across differentiation. The Y-axis is plotted on a log10 scale.

Together these results highlight that 3'UTR splicing can have a varied impact on RNA stability, depending on the stage of differentiation. Whilst stage-specific regulation of RNA stability can be observed for a large number of 3'UTR splicing events, as demonstrated in

Figure 6.7, the most significant events presented in Figures 6.12 and 6.13 represent ideal candidates for future study of 3'UTR splicing-mediated stage-specific regulation of RNA stability.

6.4 Differential splicing and differential stability contribute to changes in exon usage

The results presented in Section 6.3 highlighted that the effect of 3'UTR splicing on RNA stability can change during differentiation. As such, differences in percent spliced out (PSO; opposite of percent spliced in (PSI)) observed between conditions could either be due to alternative splicing (i.e. different splicing rates) or differential RNA stability (i.e. different decay rates), or a combination of both. Through the use of event-level SLAMseq, we estimated how quickly the 3UI-spliced and 3UI-retained isoforms decay.

This can be used to calculate a decay ratio:

$$decay\ ratio = \frac{retained\ decay\ rate}{spliced\ decay\ rate}$$

We can also calculate the PSO and PSI as follows:

$$PSO = \frac{number\ of\ spliced\ reads}{number\ of\ spliced\ and\ retained\ reads} \quad PSI = \frac{number\ of\ retained\ reads}{number\ of\ spliced\ and\ retained\ reads}$$

And use these to calculate the splicing rate:

$$splicing\ rate = \frac{PSO}{PSO + (PSI \times decay\ ratio)}$$

Splicing rates were calculated for every event we could generate a decay ratio for, and at each stage of differentiation. By using these rates, we are able to determine the contribution of splicing and decay to the PSO values observed. For example, in Figure 6.14A we present the contribution of stability to the PSO values observed at each stage of differentiation. In all cases this resembles a normal distribution centred on 0. Where the X-axis value is 0, then the PSO is entirely driven by the splicing rate, as the 3UI-spliced and 3UI retained decay rates are equal. Whilst deviation from the midpoint indicates a positive or negative

contribution of stability towards the observed PSO.

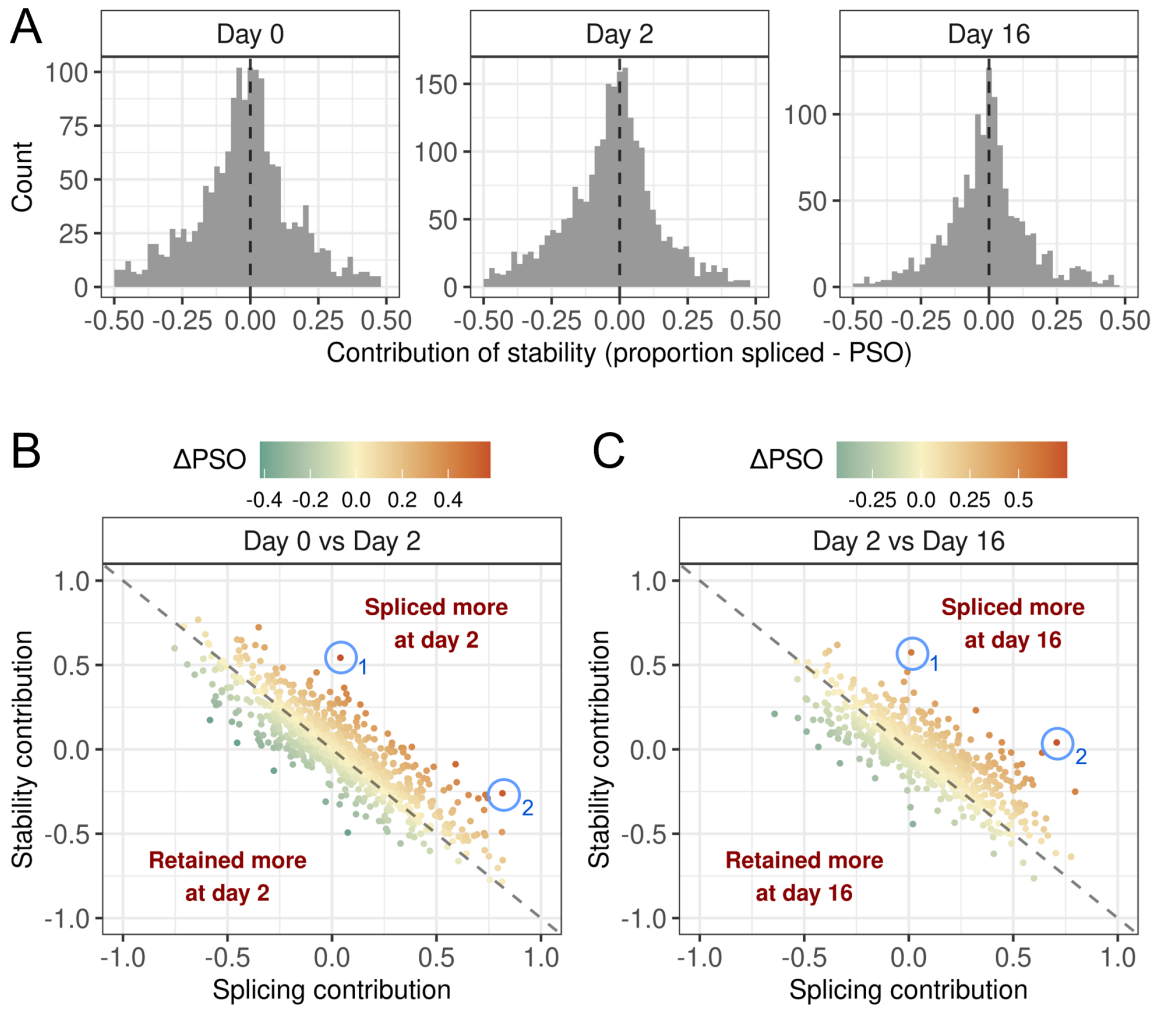

Figure 6.14: Contributions of splicing rates and RNA stability to observed PSO and PSO differences during differentiation. A) Contribution of differential RNA stability between 3UI-spliced and 3UI-retained isoforms towards the observed PSO. Dashed line intercepting the X-axis at 0 indicates no contribution of stability (and 100% due to splicing), whilst deviation from this line represents increasing contribution of stability. B+C) Predicted contributions of splicing rate differences (X-axis) and stability ratio differences (Y-axis) towards changes in observed PSO (colour gradient) between B) day 0 vs day 2; C) day 2 vs day 16. Each dot represents a splicing event. Diagonal dashed line indicates no observed PSO change between conditions, deviation indicates increased splicing/retention between conditions.

Next, we sought to predict the contribution of changes in splicing rate vs changes in decay rates to the observed changes in PSO during differentiation. We can predict the

contribution of stability changes by fixing the splicing rate value between conditions and recalculating the PSO value. This tells us the amount of difference that is solely produced by differential decay rates.

For the example of the day 0 vs day 2 comparison, we recalculate the PSO at day 2 using the splicing rate from day 0:

$$PSO_{fixed\_splice\_rate\_day2} = \frac{splicing\_rate\_D0 \times decay\_rate\_D2}{1 - splicing\_rate\_D0 + (splicing\_rate\_D0 \times decay\_rate\_D2)}$$

Therefore, the contribution of changes in RNA stability towards the differences in observed PSO is:

$$stability\ contribution = PSO_{fixed\_splice\_rate\_day2} - PSO_{day0}$$

Whilst the contribution of changes in splicing rate is the difference between the observed ΔPSO and the stability contribution:

$$\Delta\ PSO_{observed} = PSO_{day2} - PSO_{day0}$$

$$splicing\ contribution = \Delta\ PSO_{observed} - stability\ contribution$$

The contributions of both splicing changes and stability changes towards observed PSO differences was calculated for both the day 0 vs day 2, and day 2 vs day 16 comparisons. The results of which are shown in Figure 6.14B and 6.14C respectively. This highlights instances where the major driver of PSO changes is either stability changes (e.g. blue circle #1 in Figure 6.14B+C) or splicing rate changes (e.g. blue circle #2 in Figure 6.14B+C). Interestingly, there are many instances where changes in stability and splicing rate cancel each other out (those that sit on the diagonal mid-line; 41.7% in day 0 vs day 2; 39.1% in day 2 vs day 16). Additionally, these two predictors appear to counteract each other (top left and bottom right quadrants; 44.9% in day 0 vs day 2; 46.3% in day 2 vs day 16) more often than they synergise (top right and bottom left quadrants; 13.3% in day 0 vs day 2; 14.6% in day 2 vs day 16).

6.5 Summary

In this section we have demonstrated that SLAMseq can be conducted on the individual event level through the use of high-depth paired-end RNAseq, and the development of novel analysis pipelines and statistical methods. This allowed us to interrogate the impact of splicing on RNA stability by directly comparing intron-spliced and intron-retaining reads. We conducted this at each stage of differentiation and found that whilst splicing introns in the CDS results in a substantial bias towards RNA stabilization at all three differentiation time points, splicing introns in the 3'UTR resulted in a more even split of stabilization and destabilization. This further supports our hypothesis that 3'UTR introns may be regulated differently to CDS introns. In addition to examining differences between isoforms at each stage of differentiation, through the use of event-level SLAMseq we were able to examine how each isoform changes across the differentiation time course. Whilst we anticipated that this may highlight trends specific to 3UI-spliced or 3UI-retained isoforms, our results were mostly comparable to those presented in Chapter 5 at the gene-level.

Through examining how the ratio of spliced/retained half-lives changed across differentiation we were able to identify clusters of events where 3'UTR splicing produced the biggest impact on RNA stability at a specific stage of differentiation. The idea that this may be due to stage-specific regulation of RNA stability was further supported by our finding that each cluster was enriched for a unique signature of interacting RBPs and miRNAs. However, it is important to note that the CLIPseq and AGO-CLIP data utilized was generated in cancer cell lines and therefore only represents an estimate of potential RBP and miRNA binding. We do not have direct evidence of these interactions in H9 cells, nor through the differentiation time course. In order to address whether RBPs presented in Figure 6.8 were differentially expressed during differentiation, we examined their expression in our RNAseq data. This allowed us to identify a panel of candidates which A) show enriched binding in the CLIPseq data and B) were maximally expressed at

the same stage of differentiation that had the highest 3UI-spliced/3UI-retained half-life ratio for their cluster. However, it is important to note that gene expression may not directly reflect the RBPs occupancy on the 3UI. This may also be impacted by the expression of other targets of the RBP which compete for its occupancy. Additionally, changes at the RNA level do not necessarily represent changes at the protein level (Koussounadis et al., 2015; Cheng et al., 2016; Liu et al., 2016). Addressing the question of stage-specific regulation would require the integration of multiple omic approaches such as miRNA-seq and CLIP-seq conducted at each stage of differentiation. As such the findings in Figures 6.8 and 6.9 represent preliminary data and form the basis for potential future study. Future study in this regard could involve the validation of binding at each stage of differentiation, evidencing differential expression or binding during differentiation, functional validation through mutation of the RBP/miRNA motif in a target of interest, and determining the effect of mutation on RNA stability and gene expression.

In this regard, our development of statistical testing to identify the most significant changes in RNA stability caused by 3'UTR splicing across differentiation has revealed a potential pool of candidate events for future study. Our finding that only 25 events were statistically significant following P-value adjustment may be in part due to the high number of degrees of freedom in the "free interaction" model. In this model we have not only three b values (decay-rate) but also three i values (spliced/retained offset), this increased number of predictors decreases the overall power to detect significance of any individual effect, in this case the interaction between day and spliced/retained ratio. Nevertheless these events represent useful target candidates for future study, spanning a diverse range of biological functions. For example HNRNPD (also known as AUF1) has a well established role in RNA decay through binding AU-rich elements in 3'UTRs (Gratacós and Brewer, 2010) and also in promoting miRNA-mediated decay (Min et al., 2017), HNRNPA2B1 and SRSF2 both have roles in RNA splicing (Fu and Maniatis, 1992; Fu et al., 1992; Peng et al., 2021), and XPO1 is important for the nuclear export of both

proteins and RNAs (Azmi et al., 2021).

Finally, given that we observed differences in the effect of 3'UTR splicing on RNA stability across differentiation, we used this information to predict the contributions of stability changes and splicing rate changes towards observed PSO changes. This allowed us to identify examples of PSO changes that are driven primarily by stability changes or splicing rate changes (e.g. blue circles in Figure 6.14-C). More importantly, we were able to conduct this analysis for all events where we could fit decay curves for both 3UI-spliced and 3UI-retained isoforms (where both half-lives were between 0.5 and 24 hours), meaning we can lookup the predicted contributions for any event of interest (presuming it was fit). The implication of this is discussed further in Section 7.5.3.

7. Discussion

7.1 Widespread and differential 3'UTR splicing

Investigation into the effects of 3'UTR splicing has previously been limited to specific examples where individual genes were studied in static contexts (see Section 1.6). Taking a transcriptome-wide approach, we detected thousands of 3'UTR splicing events (Figure 3.1), found that many of these are broadly expressed across a hiPSC cohort (Figure 3.3), and are subject to differential usage across cardiomyocyte differentiation (Figure 4.8). Whilst this thesis focuses predominantly on 3'UTR splicing in hPSCs, we also investigated how 3'UTR splicing changes between normal and cancer samples across multiple solid tumour types (Figure 4.1). Whilst the majority of this work had already been conducted, prior to publication (currently pre-printed, see Riley et al. (2024)) an independent group published a similar analysis (Chan et al., 2022). The analysis conducted by Chan et al. (2022) detected splicing events solely downstream of the stop codon, therefore a notable difference between our analyses was that we also compiled 3'UTR splicing events that overlapped the CDS (pPTC 3UIs), allowing direct comparison between pPTC and nonPTC (3'UTR only) splicing events. Additionally, Chan et al. (2022) focused on hepatocellular carcinoma, whilst we focused on colorectal carcinoma. Both our studies identified that the majority of 3'UTR introns (nonPTC) are found less than 55nt from the stop codon, and are therefore unlikely to be NMD-sensitising. We validated this experimentally by knocking down UPF1, conducting RNAseq, and comparing: pPTC vs nonPTC 3UIs; and nonPTC 3UIs found more than or less than 55nt from the stop codon. Chan et al. (2022) also knocked down UPF1, however only focused on a panel of candidate genes. Interestingly, both studies independently identified CTNNB1 as a "model gene", where Chan et al. (2022) found its splicing correlated with poor prognosis, whilst we identified 3'UTR splicing changes driven by Wnt signalling, of which CTNNB1 is a central regulator.

By generating a transcriptome assembly using HipSci data, we were able to compare the extent of 3'UTR splicing in a "normal" population as opposed to a cancer cohort. Simulation of 3UI detection, by increasing the number of samples used to build each transcript assembly, revealed that we had saturated event detection by 100 samples using hiPSC RNAseq data, whilst using equivalent data from colon cancer samples we continued to detect novel events with each sample added. This is likely due to patient-specific splicing events encountered resulting from aberrant splicing in cancer (Oltean and Bates, 2014), which could also occur in the 3'UTR.

We also validated the existence of several example 3'UTR splicing events by attempting to amplify them from either gDNA or cDNA using splice site flanking primers or isoform-specific primers. We concluded that these events were indeed the result of splicing as we were only able to amplify spliced isoforms from cDNA (derived from RNA which was subjected to splicing), and not from gDNA (which cannot be spliced). To further validate the existence of our transcripts within the HipSci assembly we could have compared them to transcripts detected via long-read RNAseq. Such an analysis was conducted by Ian Sudbery in the Riley et al. (2024) manuscript for the TCGA assembly, and found that 73% of nonPTC 3UIs expressed over 1TPM in HCT116 cells were also present in long-read RNAseq data. However, a comparable analysis for the HipSci assembly has not yet been conducted.

7.2 3'UTR splicing and NMD

We found that 3'UTR splicing was generally NMD-sensitising in hESCs, and the position of the splice site relative to the 55nt threshold was a significant contributor. However, in colorectal carcinoma cells, 3'UTR splicing was not significantly effected by UPF1 knockdown, including events where the splice junction was >55nt from the stop codon. This could suggest that the NMD pathway is faulty in HCT116 cells, however, we still observed upregulation of pPTC (CDS-overlapping) 3UIs upon UPF1 knockdown,

meaning such faults would be restricted to splice junctions solely within the 3'UTR. Differences in the regulation of pPTC and nonPTC 3UIs upon UPF1 knockdown were observed in both H9 hESCs and HCT116 cells, which suggests that splice junctions that lead to premature termination vs those that occur within the 3'UTR are regulated differently. This could potentially be due to an additional, and as yet unexplored sub-branch of EJC-dependent NMD, or perhaps due to interactions between the NMD machinery with additional 3'UTR-specific trans factors.

We identified a subset of transcripts that contained 3UIs >55nt from the stop codon yet were not upregulated upon UPF1 knockdown, and in some instances were significantly downregulated. This subset may represent NMD evading transcripts, given that they are also often highly expressed. It is possible that true NMD targets, for example those resulting from splicing errors, are so efficiently degraded that they were not detected during our transcriptome assembly, and as such were never quantified against in the siUPF1 condition. This could be rectified by creating a UPF1 knockdown-specific transcriptome with StringTie and comparing this against a control-specific transcriptome. However, as shown in Figure 3.2, we do not have enough samples to detect all events, and therefore detection by these means would likely be under-powered.

Knockdown of UPF1 in H9 hESCs resulted in substantial levels of cell death within 72 hours of siRNA transfection, likely a secondary effect of widespread expression changes throughout the transcriptome. To account for this, and to reduce the capture of gene expression changes resulting from secondary effects (i.e. changes in gene B (which is not NMD-sensitive) driven by changes in gene A (which is NMD-sensitive and upregulated)), it would be beneficial to observe the effects of NMD factor depletion within a shorter time frame. Therefore future experiments should involve rapid depletion of these factors at the protein-level, as opposed to RNA-level interference. Technologies such as the auxin-degron 2 system (Yesbolatova et al., 2020) or the dTAG system (Nabet et al., 2018) could be used to create cell lines in which UPF1 protein could be depleted

rapidly and in an inducible manner. Additionally, in order to conclude that transcript degradation is driven by NMD, other NMD factors such as SMG5 and SMG6 could be depleted in addition to or in combination with UPF1. Doing so would allow us to distinguish NMD from other forms of UPF1-mediated degradation.

The creation of these stable cell lines would allow us to investigate how 3'UTR splicing interacts with NMD across differentiation. Cells could be differentiated as previously demonstrated and NMD factors could be depleted prior to RNA extraction. Given the rapid degradation caused by the AUX2/dTAG systems, RNA could be extracted within hours of NMD factor depletion, as opposed to within days when using RNAi. This would allow greater precision over which stages of differentiation observations are taken at. Additionally, NMD factor depletion via AUX2/dTAG is less likely to impact lineage specification, as was shown by UPF1 depletion via RNAi in H9 hESCs (Lou et al., 2016), given that RNA extraction could be conducted within hours of induction, minimising such secondary effects.

7.3 3'UTR splicing and trans factor binding

Besides NMD-sensitisation as a mechanism of action of 3'UTR splicing, we also explored potential EJC-independent forms of regulation. The most obvious route by which 3'UTR splicing may impact the post-transcriptional regulation of mRNA is through removing binding sites for various trans factors. Through utilizing existing CLIP-seq and AGO-CLIP data, we identified a subset of RBPs and miRNAs that are enriched in 3UIs compared to the rest of the 3'UTR. The modulation of these interactions represents an additional layer of regulatory complexity, meaning distinct RNA populations can arise which may be characterized by different properties (e.g. stability, translation efficiency, subcellular localization), despite encoding for the same amino acid chain.

However, it is important to note that our RBP and miRNA enrichment results are highly preliminary, given that the dataset used was collated from many RBP CLIP-seq and

AGO-CLIP experiments, which were predominantly conducted in HepG2 and K562 cancer cell lines. Therefore whilst evidence of RBP and miRNA binding at these positions has been experimentally validated in these cancer cell lines, whether these trans factors also bind in the hESC setting has not yet been experimentally validated. However, we deemed this approach to be more informative than entirely predictive approaches, such as the use of MEME for sequence motif detection followed by the use of TOMTOM to match these motifs to RNA binding protein binding motifs. We could overcome this issue in the future by conducting CLIP-seq and AGO-CLIP specifically in hESC lines, or alternatively, RNA immunoprecipitation could be conducted to validate specific RBP-3'UTR interactions.

In the future, functional validation of the effect of these RBP-3'UTR interactions could be conducted. A relatively straightforward approach would be to knock down expression of an RBP of interest and probe the effect on 3'UTR splicing; however, this would likely have a global effect and therefore transcripts may be regulated directly (by knockdown of the RBP) and indirectly (by differential expression of other trans factors which themselves are regulated by said RBP). A more targeted approach could involve targeting RBP binding motifs with transcript-specific ASOs, or mutating the RBP binding motif within a specific target of interest via CRISPR.

7.4 RNA stability changes during differentiation

SLAMseq has recently been used to study RNA stability in response to gene knockdowns (Herzog et al., 2017; Feng et al., 2021), gene over-expression (Shi et al., 2023), and in mouse embryonic stem cell differentiation (Viegas et al., 2022). However these approaches represent A vs B comparisons. The use of SLAMseq was also reported as part of a time course in early zebrafish development (Bhat et al., 2023), however this was conducted in the anabolic SLAMseq configuration where 1-cell stage embryos were injected with 4sU and then chased at various time points following fertilization, thus

directly measuring the rate of transcription, not RNA degradation. The use of SLAMseq as part of a hESC differentiation time course has not been reported. Therefore, our use of SLAMseq at the event level as part of a hESC differentiation time course to measure RNA degradation represents a novel application on two fronts.

This approach allowed us to directly compare decay rates of 3UI-spliced and 3UI-retaining isoforms, and determine how these ratios changed across differentiation (discussed further in Section 7.5.2). However, prior to event-level analysis we looked broadly at how RNA stability changed across differentiation by interrogating our RNAseq data at the gene-level. In doing so, we identified a transient global destabilization of RNA at day 2 of differentiation, followed by stabilization at day 16. A model highlighting this is presented in Figure 7.1A, whilst the potential impacts associated with such changes in RNA stability is presented in Figure 7.1B.

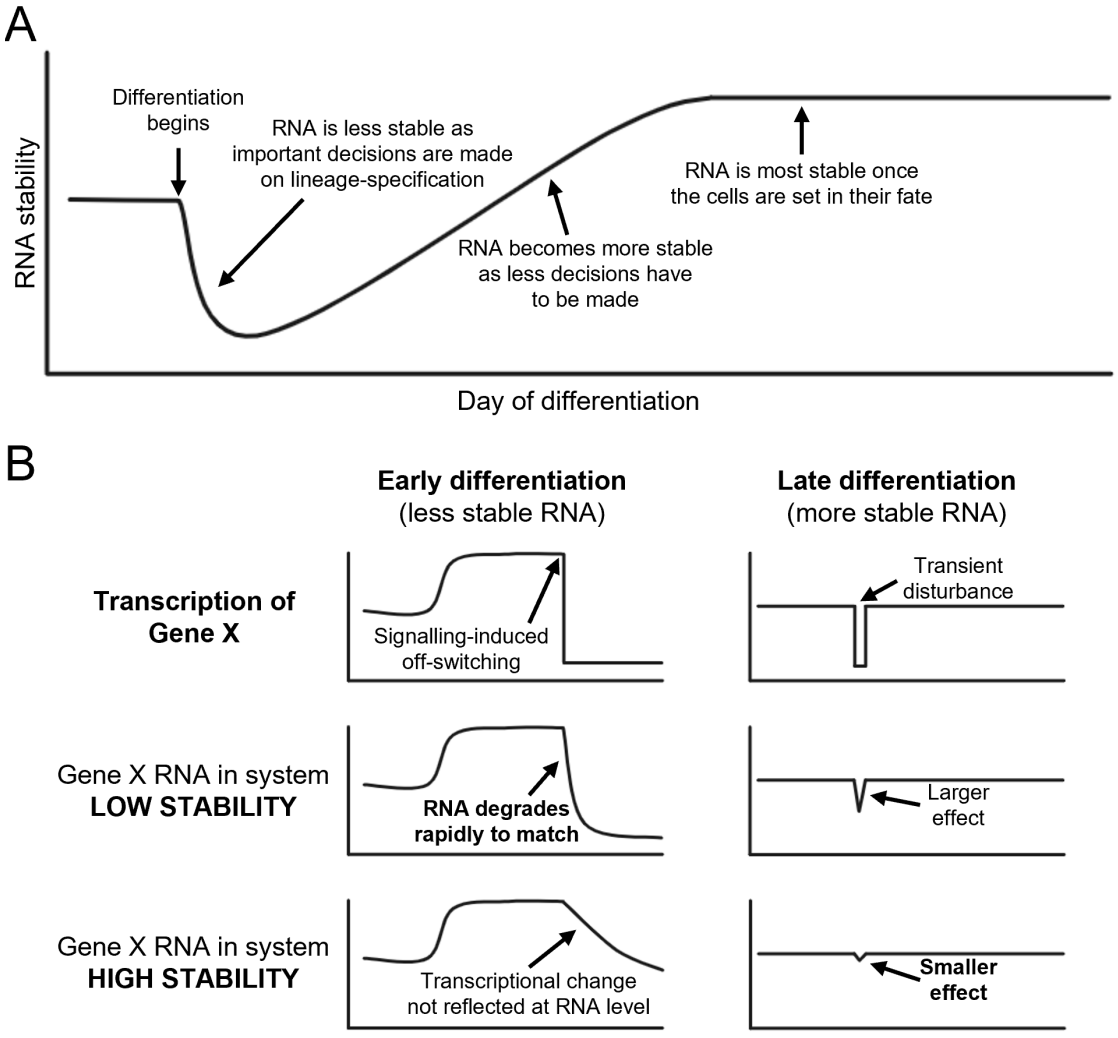

Figure 7.1: Impact of RNA stability on transcriptome changes. A) Potential model of how RNA stability changes during cardiomyocyte differentiation. B) In early differentiation RNA is less stable, whilst in late differentiation RNA is more stable. The top row indicates two examples of transcriptional changes: left) signalling-induced transcriptional shutdown; right) transient transcriptional shutoff. The rows below highlight what the response at the transcriptome-level would be if the RNA is less stable or more stable.

We propose that RNA destabilization early in differentiation increases the "reaction time" within the system, allowing changes at the level of transcription to quickly be reflected across the transcriptome in response to the signals the cells encounter during differentiation. This is especially important early in differentiation as the cells are highly

potent, encounter many cell fate-determining decisions relatively quickly, and have to respond to these accordingly. In more differentiated cell types (day 16), RNA being more stable increases the robustness of the system, where transient changes in transcription due to stresses are less likely to lead to global changes in the transcriptome. We could further validate the model in Figure 7.1A by conducting SLAMseq at additional time points in differentiation to increase the temporal resolution. Additionally, to determine whether RNA stability changes are necessary for differentiation, we could overexpress or knockdown factors involved in RNA degradation, and observe the effect on cell differentiation.

7.5 Event-level SLAMseq analysis

By using standard RNAseq as opposed to QuantSeq, we were able to conduct SLAMseq on the individual event level. An additional benefit of this approach was that we were not restricted to studying the 3' end of the RNA. As such, we were able to compare half-life differences caused by intron retention in both the CDS and the 3'UTR. We found that intron retention is generally destabilizing. This is likely due to the introduction of premature termination codons, either directly or due to frameshifting. However, where we limited our analysis to the 3'UTR, we found that intron retention was producing roughly equal proportions of stabilization and destabilization.

7.5.1 Stability vs predicted NMD sensitivity

By comparing the 3UI-spliced/3UI-retained half-life ratios between events situated more than or less than 55nt from the stop codon, we found that splicing generally reduced stability when the event was found more than 55nt from the stop codon. This is consistent with splicing inducing NMD. However, all the events analysed had half-lives greater than 30 minutes, and the majority were several hours. Additionally, there was a substantial proportion of 3UI-spliced isoforms that were more stable than their 3UI-retaining

partners despite being more than 55nt from the stop codon. As discussed in Section 7.2, it is possible that many of the events we are examining here may be partially or entirely resistant to NMD. Where transcripts are truly NMD sensitive, they may be too lowly expressed, or their half-lives so short, that they are filtered out of our analysis (Section 2.6.2.4). To determine whether this is the case, we could combine event-level SLAMseq with inducible knockdown of NMD factors, as discussed in Section 7.2. Where an event is NMD-sensitising we would expect it to be upregulated upon NMD factor depletion; however, this occurs secondarily to stabilization, which would be the direct effect of preventing their active degradation. Through this method we would be able to examine changes at both the levels of stability and expression using the same RNAseq data.

7.5.2 Differentiation-stage specific regulation by 3'UTR splicing

During differentiation the cellular composition of trans factors such as RBPs and miRNAs is subject to change (Zandhuis et al., 2021; Fedorova et al., 2023). Therefore we hypothesised that the effect of 3'UTR splicing on RNA stability would also change during differentiation, in light of increased or decreased levels of trans factor binding to the 3UI. By looking at global trends in spliced/retained half-life ratios we identified clusters of events showing this effect. Additionally, we found that 3UIs in each of these clusters were enriched for unique subsets of RBPs and MREs. These trans factors represent potential regulators of these differentiation-stage-specific RNA stability changes, and as such, potential targets for future study. As discussed in Section 7.3, these RBPs could be depleted or their targets could be targeted with ASOs or mutated via CRISPR to prevent RBP binding. Subsequently, assays could be conducted at both the cellular (e.g. ability to differentiate, impact on lineage-specification) and molecular (e.g. impact on expression / RNA stability) levels.

By testing the interaction between 3UI-spliced and 3UI-retained decay rates across differentiation (Figure 6.11), we identified a pool of events which had the most significant spliced/retained half-life ratio changes across differentiation. These events would serve

as ideal candidates to model differentiation-stage-specific regulation by 3'UTR splicing. Where specific RBP-3'UTR interactions are to be studied, these would represent the candidates that CRISPR mutations could be conducted on. HNRNPD (AUF1) is an appealing candidate given its roles in RNA stability regulation and splicing (Brewer, 1991; Kemmerer et al., 2018). As discussed in Section 1.6.1, HNRNPD can be subjected to alternative splicing that changes its NMD-sensitivity (Wilson et al., 1999; Banihashemi et al., 2006). However, the HNRNPD 3'UTR splicing event shown in Figure 6.12 is from the NMD-insensitive isoform (first isoform in Figure 1.10). This isoform shows a progressive decrease in 3UI-spliced/3UI-retained half-life ratio as differentiation progresses (Figure 6.12), whereby the 3UI-spliced isoform is more stable than its 3UI-retaining partner at day 0, less so at day 2, and by day 16 this relationship is flipped. Given that the splice junction is <55nt from the stop codon, it is unlikely that the differential effect of 3'UTR splicing on RNA stability is driven by an EJC-dependent process (e.g. NMD) given that the EJC should be stripped during the pioneer round of translation. Instead, the change in 3UI-spliced/3UI-retained half-life ratio could be explained by a change in the composition of trans factors that bind to the 3UI, i.e. either a decrease in destabilizing factors, or an increase in stabilizing factors, as differentiation proceeds. Identification of the factor(s) responsible could serve as a future research goal. This could be accomplished by using the TREX (targeted RNase H-mediated extraction of crosslinked RBPs) approach (Dodel et al., 2024) to compare RBP occupancy on the 3UI across differentiation.

7.5.3 Calculating splicing rates

By combining PSO values and decay rate ratios (between 3UI-retained and 3UI-spliced isoforms), we were able to calculate the splicing rate for each event. Subsequently, this allowed us to determine the contribution of splicing rate and decay rate changes towards differences in PSO observed during differentiation. This is important as differential exon usage analysis is often conducted under the premise of "alternative splicing analysis";

however, it is possible that differences in PSO are observed between conditions despite there being no change in splicing rate, as shown in Figure 7.2.

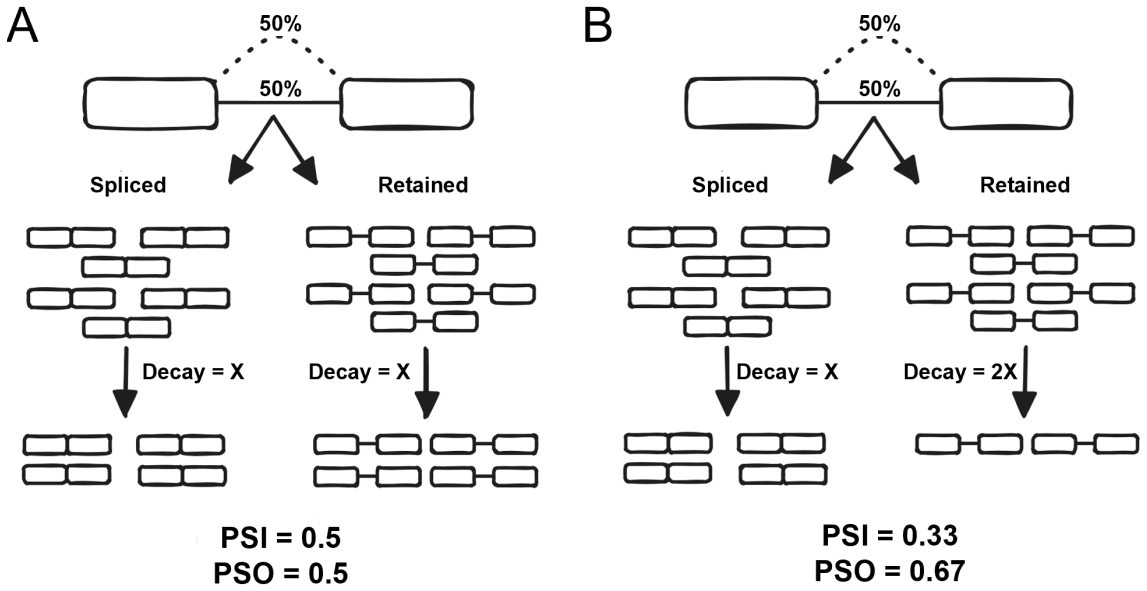

Figure 7.2: Contribution of splicing and decay rates to PSO. An example of differential PSO between conditions A and B despite there being no change in the splicing rate. In A the intron is spliced 50% of the time, and spliced and retained isoforms are degraded at equal rates. Therefore the PSO is 0.5. In B the intron is spliced at the same rate as in A; however, here the retained isoform is degraded twice as quickly as the spliced isoform. Therefore the PSO is 0.67. Delta PSO = 0.17 despite no change in splicing rate.

By determining the contribution of splicing changes and stability changes towards differences in PSO, it allows us to make more informed decisions on the direction of future study. For example, if we wanted to identify the trans factor(s) responsible for differential exon usage of a given event across differentiation, knowing whether the major driver of change was stability or splicing would allow us to focus our efforts towards studying either stability factors or splicing factors, respectively.

7.6 Future directions

7.6.1 Implementation of a Bayesian inference model

A Bayesian inference model could be utilized to take into account the amount of convertible nucleotides versus those that are observed as being converted. Our script currently outputs a "convertible sequence" variable into the metadata, representing the number of As and Ts present within the read pair (the pool of nucleotides that could be converted into Cs or Gs due to 4sU incorporation and alkylation). This variable is not used in the current version of the script; however, it was calculated to facilitate future implementation of a Bayesian model. As depicted in Figure 7.3A where there are many reads covering a given event/gene, then the distribution of reads over the loci should be relatively even. As such, where there is a more AT-rich region (where we would be more likely to observe conversions), we could assume reads in different conditions are equally likely to have coverage over these regions. For those which are lowly expressed (Figure 7.3B) read coverage is likely to be more varied between replicates and between conditions. Therefore the chance of observing T>C or A>G conversions may also change.

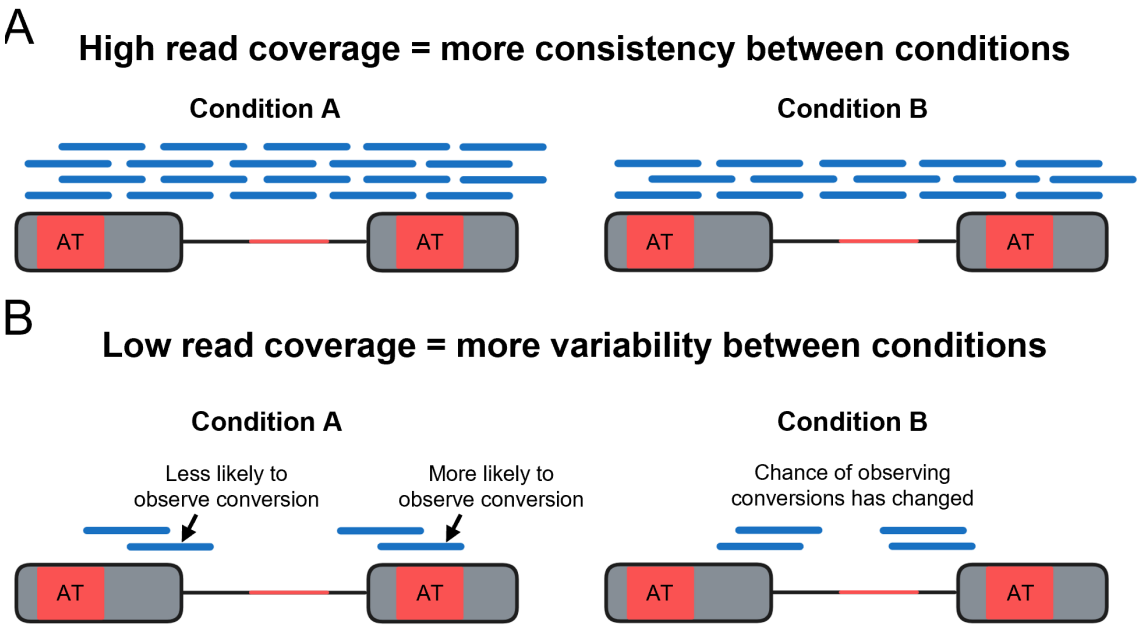

Figure 7.3: Read coverage impacts chance of observing conversions between replicates and conditions. Blue lines = RNAseq reads, red regions = higher AT content. (A) Where read coverage is high then the distribution of reads in different conditions will be relatively even. Therefore the chance of detecting conversions is similar. (B) Where read coverage is low, then the distribution of reads in different conditions or replicates is likely to vary. Therefore the chance of detecting conversions can also change depending on the AT content across the loci.

We could improve the discovery rate by factoring in the potential sequence that can be converted compared to the number of conversions observed. For example, if 2% of Ts were converted to Cs (or As to Gs) then if a read pair (of 300 bases) contains 150 convertible bases, we would expect to see, on average, 3 conversions. However, if a read pair only contains 50 convertible bases due to its position across the loci, we are less likely to observe conversions. Therefore, conversions observed in this read should be assigned a higher power compared to those where the chance is greater. This would account for differences in AT content observed across the loci, improve consistency between replicates, and improve the ability of the script to compare between conditions for lowly expressed events.

7.6.2 Incorporation of additional event types

Our pipeline was developed for the purpose of comparing 3UI-spliced vs 3UI-retaining isoforms, and as such is currently limited to studying intron retention. Event assignment could be expanded to allow comparison of additional forms of alternative splicing, similarly to RMATS (Shen et al., 2014). As such, decay rates could be compared in a pairwise manner between: mutually exclusive exons, skipped vs non-skipped exons, alternative 5′ splice sites, alternative 3′ splice sites. Implementation of these features would make the script of greater use to the wider field by addressing questions related to other forms of alternative splicing beyond intron retention.

7.6.3 Transcript level analysis

Whilst we have demonstrated our ability to calculate decay rates and half-life estimates at the gene-level and individual event-level, our pipeline does not produce transcript-level estimates. One possible way of dealing with this could be to incorporate additional splicing events into our pipeline, and also process read pairs mapping to each exon. We could then estimate decay rates for each transcript feature (e.g. exons and splicing events), and average these to produce transcript-level estimates. Such calculation would require the weighting of each feature based on its likelihood to contribute to a given transcript. A benefit to this approach is that it would allow the user to estimate transcript-level half-lives, and also interrogate the individual features that make up each transcript. Alternatively, instead of assigning read pairs to individual events, we could instead assign them to transcripts. This would be relatively straightforward in instances where the read pair maps to only a single transcript. However, in instances where reads map to multiple transcripts, assignment would have to be biased based on the fraction each transcript contributes to overall gene expression, similarly to how ambiguity is dealt with by the Salmon package (Patro et al., 2017).

A more direct approach could involve the use of long read RNA sequencing. This would be accompanied by the detection of novel differentiation stage-specific isoforms. A 4sU pulse-chase followed by IAA treatment could be conducted in the same manner described in Section 5; however, given that reads would be substantially longer we would have to take a more sophisticated approach to setting the "pull down" threshold. Long read RNA sequencing is used in the nano-ID approach (Maier et al., 2020). However, instead of 4sU incorporation followed by IAA treatment, nano-ID involves 5-ethynyluridine (5EU) incorporation, which is detected by direct RNA nanopore sequencing (Garalde et al., 2018). Additionally, a pulse-chase is not performed, instead cells are pulsed for 1 hour and a machine learning algorithm is used to estimate both transcription and decay rates (i.e. overall turnover). Using this technology, Maier et al. (2020) were only able to assess 2068 isoforms across all samples and conditions. This could reflect the lower throughput of nanopore sequencing, compared to short read sequencing approaches, although this continues to improve (Wang et al., 2021).

7.6.4 Improving general usability

Our pipeline currently utilizes CGAT-Ruffus infrastructure to process files in a step-wise and reproducible manner from inputs to final output. This allows parallelization of tasks such as read mapping, feature assignment, and detection of nucleotide conversions on a sample-by-sample basis simultaneously on HPC systems. CGAT-Ruffus is less well adopted than alternative workflow managers such as Nextflow (Di Tommaso et al., 2017) and Snakemake (Mölder et al., 2021), therefore the porting of our pipeline to one of these workflow managers may accommodate a greater uptake of our software. However, a more attractive approach would see our software compiled into a self-contained executable package, similar to the existing SLAMseq software SLAM-DUNK (Neumann et al., 2019). In this regard we were recently successful in obtaining support from a Research Software Engineer at the University of Sheffield from October 2024 to April 2025 to conduct this work. Given the size of the input files required to conduct event-level

SLAMseq (approximately 10Gb per sample for 100M paired-end reads) the software would still be intended for use on HPC systems, therefore we intend for it to be run broadly in two stages, with a "prep" and "post" stage, similar to RMATS (Shen et al., 2014). In the "prep" stage, each sample could be processed individually in parallel on a HPC system. In the "post" stage the intermediate outputs would be summarized and statistical analysis (calculation of decay rates and half-lives; comparison of differential decay rates and interactions) would be conducted as a single final process. This standalone software would subsequently be released for use by the community via a software publication and accompanying documentation for use.

7.6.5 Investigating the effect of 3'UTR splicing on translation efficiency

Here we have focused heavily on the regulation of RNA stability through 3'UTR splicing, using both low throughput (Actinomycin D treatment and chase) and high throughput (event-level SLAMseq) approaches. Whilst this sheds light on the effect of 3'UTR splicing on RNA expression dynamics, it does not inform us of the full biological impact, including protein level expression, which would be impacted by translation efficiency. For example, a 3'UTR splicing event may stabilize a transcript, but make it translate very poorly. Conversely, transcript destabilization may be accompanied by an increase in translation efficiency (as was the case for Arc mRNA; Section 1.6.3). When we addressed translation efficiency for 2 examples (CTNNB1 and HRAS) in Section 3.5.2.2, we observed no significant impact of 3'UTR splicing. However, to address this on a transcriptome-wide scale we could conduct TrIP-seq (Floor and Doudna, 2016) to directly compare translation efficiency of 3UI-spliced and 3UI-retained isoforms. TrIP-seq could also be conducted across our differentiation time course. Subsequently, we could compare the effects of 3'UTR splicing on transcript stability and translation efficiency to determine how these two modalities interact in contribution to expression.

7.7 Concluding remarks

3'UTRs play important roles in post-transcriptional regulation by acting as binding platforms for various trans factors which can have downstream effects on RNA stability, localization and translation efficiency. Despite this, 3'UTRs are often overlooked in favour of studying the coding domain directly. This thesis has shed light on the extent and impacts of 3'UTR splicing by taking a transcriptome-wide approach. 3'UTR splicing >55nt from the stop codon is generally considered a signal to elicit NMD. Whilst we confirm this partially, we also identify a subset of transcripts that appear to be NMD-insensitive. These events, alongside those that are <55nt from the stop codon, are predicted to modulate trans factor binding via alternative splicing.

By interrogating RNA stability at the individual event level, we found that splicing within the 3'UTR can both stabilize and destabilize RNA, whilst splicing within the coding region is generally stabilizing. We also found that RNA stability changed during differentiation, where we observed a transient global destabilization of RNA early in differentiation, followed by stabilization as differentiation proceeded. Additionally, the interaction between 3'UTR splicing and RNA stability also changed during differentiation, potentially due to differential regulation of trans factors. In this regard, we identified the most significant instances of differentiation-stage-specific regulation, and accompanying RBP enrichment signatures, representing potential candidates for future study.

References

Ainger, K., Avossa, D., Diana, A.S. et al (1997). Transport and Localization Elements in Myelin Basic Protein mRNA. *The Journal of Cell Biology*, 138(5):1077–1087.

Akinyi, M.V. and Frilander, M.J. (2021). At the Intersection of Major and Minor Spliceosomes: Crosstalk Mechanisms and Their Impact on Gene Expression. *Frontiers in Genetics*, 12.

Alles, J., Fehlmann, T., Fischer, U. et al (2019). An estimate of the total number of true human miRNAs. *Nucleic Acids Research*, 47(7):3353–3364.

Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, New York.

Anders, S., Reyes, A. and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017.

Andrews, S. (2010). Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.

Anokye-Danso, F., Trivedi, C.M., Juhr, D. et al (2011). Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell stem cell*, 8(4):376–388.

Aslanzadeh, V., Huang, Y., Sanguinetti, G. et al (2018). Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast. *Genome Research*, 28(2):203–213.

Azmi, A.S., Uddin, M.H. and Mohammad, R.M. (2021). The nuclear export protein XPO1 — from biology to targeted therapy. *Nature Reviews Clinical Oncology*, 18(3):152–169.

Bae, H. and Coller, J. (2022). Codon Optimality-Mediated mRNA degradation (COMD): Linking Translational Elongation to mRNA Stability. *Molecular cell*, 82(8):1467–1476.

Banihashemi, L., Wilson, G.M., Das, N. et al (2006). Upf1/Upf2 Regulation of 3

Untranslated Region Splice Variants of AUF1 Links Nonsense-Mediated and A+U-Rich Element-Mediated mRNA Decay. *Molecular and Cellular Biology*, 26(23):8743–8754.

Bar, M., Wyman, S.K., Fritz, B.R. et al (2008). MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. *Stem cells (Dayton, Ohio)*, 26(10):2496–2505.

Bassett, A.R. (2017). Editing the genome of hiPSC with CRISPR/Cas9: disease models. *Mammalian Genome*, 28(7):348–364.

Bawankar, P., Loh, B., Wohlbold, L. et al (2013). NOT10 and C2orf29/NOT11 form a conserved module of the CCR4-NOT complex that docks onto the NOT1 N-terminal domain. *RNA biology*, 10(2):228–244.

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):289–300.

Berkovits, B.D. and Mayr, C. (2015). Alternative 3'UTRs act as scaffolds to regulate membrane protein localization. *Nature*, 522(7556):363–367.

Bertero, A., Brown, S., Madrigal, P. et al (2018). The SMAD2/3 interactome reveals that TGFβ controls m6A mRNA methylation in pluripotency. *Nature*, 555(7695):256–259.

Beyer, K., Domingo-Sábat, M., Lao, J.I. et al (2008). Identification and characterization of a new alpha-synuclein isoform and its role in Lewy body diseases. *Neurogenetics*, 9(1):15–23.

Bhat, P., Cabrera-Quio, L.E., Herzog, V.A. et al (2023). SLAMseq resolves the kinetics of maternal and zygotic gene expression during early zebrafish embryogenesis. *Cell Reports*, 42(2):112070.

Bhattacharyya, S.N., Habermacher, R., Martine, U. et al (2006). Relief of microRNA-

mediated translational repression in human cells subjected to stress. *Cell*, 125(6):1111–1124.

Bicknell, A.A., Cenik, C., Chua, H.N. et al (2012). Introns in UTRs: why we should stop ignoring them. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 34(12):1025–1034.

Boehm, V., Kueckelmann, S., Gerbracht, J.V. et al (2021). SMG5-SMG7 authorize nonsense-mediated mRNA decay by enabling SMG6 endonucleolytic activity. *Nature Communications*, 12(1):3965.

Borman, A.M., Michel, Y.M. and Kean, K.M. (2000). Biochemical characterisation of cap–poly(A) synergy in rabbit reticulocyte lysates: the eIF4G–PABP interaction increases the functional affinity of eIF4E for the capped mRNA 5-end. *Nucleic Acids Research*, 28(21):4068–4075.

Bradley, T., Cook, M.E. and Blanchette, M. (2015). SR proteins control a complex network of RNA-processing events. *RNA*, 21(1):75–92.

Brewer, G. (1991). An A + U-rich element RNA-binding factor regulates c-myc mRNA stability in vitro. *Molecular and Cellular Biology*, 11(5):2460–2466.

Burridge, P.W., Matsa, E., Shukla, P. et al (2014). Chemically Defined and Small Molecule-Based Generation of Human Cardiomyocytes. *Nature methods*, 11(8):855–860.

Caizzi, L., Monteiro-Martins, S., Schwalb, B. et al (2021). Efficient RNA polymerase II pause release requires U2 snRNP function. *Molecular Cell*, 81(9):1920–1934.e9.

Calvo, S.E., Pagliarini, D.J. and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proceedings of the National Academy of Sciences*, 106(18):7507–7512.

Card, D.A., Hebbar, P.B., Li, L. et al (2008). Oct4/Sox2-Regulated miR-302 Targets Cyclin D1 in Human Embryonic Stem Cells. *Molecular and Cellular Biology*, 28(20):6426–6438.

Chan, J.J., Zhang, B., Chew, X.H. et al (2022). Pan-cancer pervasive upregulation of 3 UTR splicing drives tumourigenesis. *Nature Cell Biology*, 24(6):928–939.

Cheloufi, S., Dos Santos, C.O., Chong, M.M.W. et al (2010). A Dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*, 465(7298):584–589.

Chen, C.Y., Chen, S.T., Juan, H.F. et al (2012). Lengthening of 3UTR increases with morphological complexity in animal evolution. *Bioinformatics*, 28(24):3178–3181.

Chen, M., Lyu, G., Han, M. et al (2018). 3' UTR lengthening as a novel mechanism in regulating cellular senescence. *Genome Research*, 28(3):285.

Cheng, Z., Teo, G., Krueger, S. et al (2016). Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Molecular Systems Biology*, 12(1):855.

Chia, N.Y., Chan, Y.S., Feng, B. et al (2010). A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature*, 468(7321):316–320.

Cho, E.J., Takagi, T., Moore, C.R. et al (1997). mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminaldomain. *Genes & Development*, 11(24):3319–3326.

Cho, S. and Dreyfuss, G. (2010). A degron created by SMN2 exon 7 skipping is a principal contributor to spinal muscular atrophy severity. *Genes & Development*, 24(5):438–442.

Cifuentes, D., Xue, H., Taylor, D.W. et al (2010). A Novel miRNA Processing Pathway Independent of Dicer Requires Argonaute2 Catalytic Activity. *Science*, 328(5986):1694–1698.

Collignon, E., Cho, B., Fothergill-Robinson, J. et al (2023). m 6 A RNA methylation orchestrates transcriptional dormancy during developmental pausing. *bioRxiv: The Preprint Server for Biology*, page 2023.01.30.526234.

Collotta, D., Bertocchi, I., Chiapello, E. et al (2023). Antisense oligonucleotides: a novel Frontier in pharmacological strategy. *Frontiers in Pharmacology*, 14:1304342.

Cooke, A., Prigge, A. and Wickens, M. (2010). Translational repression by deadenylases. *The Journal of Biological Chemistry*, 285(37):28506–28513.

Danecek, P., Bonfield, J.K., Liddle, J. et al (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008.

Davis, A.G., Johnson, D.T., Zheng, D. et al (2022). Alternative polyadenylation dysregulation contributes to the differentiation block of acute myeloid leukemia. *Blood*, 139(3):424–438.

DeMaria, C.T. and Brewer, G. (1996). AUF1 Binding Affinity to A+U-rich Elements Correlates with Rapid mRNA Degradation (). *Journal of Biological Chemistry*, 271(21):12179–12184.

Di Tommaso, P., Chatzou, M., Floden, E.W. et al (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319.

Dobin, A., Davis, C.A., Schlesinger, F. et al (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Dodel, M., Guiducci, G., Dermit, M. et al (2024). TREX reveals proteins that bind to specific RNA regions in living cells. *Nature Methods*, 21(3):423–434.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S. et al (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, 485(7397):201–206.

Driever, W. and Nüsslein-Volhard, C. (1988). A gradient of *bicoid* protein in Drosophila embryos. *Cell*, 54(1):83–93.

Du, H., Zhao, Y., He, J. et al (2016). YTHDF2 destabilizes m6A-containing RNA through direct recruitment of the CCR4–NOT deadenylase complex. *Nature Communications*, 7:12626.

Dvinge, H. and Bradley, R.K. (2015). Widespread intron retention diversifies most cancer transcriptomes. *Genome Medicine*, 7(1):45.

Eberle, A.B., Lykke-Andersen, S., Mühlemann, O. et al (2009). SMG6 promotes endonucleolytic cleavage of nonsense mRNA in human cells. *Nature Structural & Molecular Biology*, 16(1):49–55.

Eichhorn, S.W., Guo, H., McGeary, S.E. et al (2014). mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. *Molecular Cell*, 56(1):104–115.

Ezkurdia, I., Juan, D., Rodriguez, J.M. et al (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, 23(22):5866–5878.

Fabian, M.R., Cieplak, M.K., Frank, F. et al (2011). miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4–NOT. *Nature Structural & Molecular Biology*, 18(11):1211–1217.

Fan, X.C. and Steitz, J.A. (1998). Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs. *The EMBO journal*, 17(12):3448–3460.

Fansler, M.M., Mitschka, S. and Mayr, C. (2024). Quantifying 3UTR length from scRNA-seq data reveals changes independent of gene expression. *Nature Communications*, 15(1):4050.

Farina, K.L., Hüttelmaier, S., Musunuru, K. et al (2003). Two ZBP1 KH domains facilitate β-actin mRNA localization, granule formation, and cytoskeletal attachment. *The Journal of Cell Biology*, 160(1):77–87.

Fedorova, V., Amruz Cerna, K., Oppelt, J. et al (2023). MicroRNA Profiling of Self-Renewing Human Neural Stem Cells Reveals Novel Sets of Differentially Expressed

microRNAs During Neural Differentiation In Vitro. *Stem Cell Reviews and Reports*, 19(5):1524–1539.

Feng, M., Xie, X., Han, G. et al (2021). YBX1 is required for maintaining myeloid leukemia cell survival by regulating BCL2 stability in an m6A-dependent manner. *Blood*, 138(1):71–85.

Ferrandon, D., Elphick, L., Nüsslein-Volhard, C. et al (1994). Staufen protein associates with the 3'UTR of bicoid mRNA to form particles that move in a microtubule-dependent manner. *Cell*, 79(7):1221–1232.

Floor, S.N. and Doudna, J.A. (2016). Tunable protein synthesis by transcript isoforms in human cells. *eLife*, 5:e10921.

Friedman, R.C., Farh, K.K.H., Burge, C.B. et al (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105.

Fu, X.D. and Maniatis, T. (1992). The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site. *Proceedings of the National Academy of Sciences of the United States of America*, 89(5):1725–1729.

Fu, X.D., Mayeda, A., Maniatis, T. et al (1992). General splicing factors SF2 and SC35 have equivalent activities in vitro, and both affect alternative 5' and 3' splice site selection. *Proceedings of the National Academy of Sciences of the United States of America*, 89(23):11224–11228.

Gabut, M., Samavarchi-Tehrani, P., Wang, X. et al (2011). An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming. *Cell*, 147(1):132–146.

Garalde, D.R., Snell, E.A., Jachimowicz, D. et al (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, 15(3):201–206.

Garcia, D.M., Baek, D., Shin, C. et al (2011). Weak Seed-Pairing Stability and High Target-

Site Abundance Decrease the Proficiency of lsy-6 and Other miRNAs. *Nature structural & molecular biology*, 18(10):1139–1146.

Ge, Z., Quek, B.L., Beemon, K.L. et al (2016). Polypyrimidine tract binding protein 1 protects mRNAs from recognition by the nonsense-mediated mRNA decay pathway. *eLife*, 5:e11155.

Gehring, N.H., Lamprinaki, S., Kulozik, A.E. et al (2009). Disassembly of Exon Junction Complexes by PYM. *Cell*, 137(3):536–548.

Gerstberger, S., Hafner, M. and Tuschl, T. (2014). A census of human RNA-binding proteins. *Nature Reviews Genetics*, 15(12):829–845.

Giammartino, D.C.D., Nishida, K. and Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. *Molecular cell*, 43(6):853.

Giorgi, C., Yeo, G.W., Stone, M.E. et al (2007). The EJC factor eIF4AIII modulates synaptic strength and neuronal protein expression. *Cell*, 130(1):179–191.

Gratacós, F.M. and Brewer, G. (2010). The role of AUF1 in regulated mRNA decay. *Wiley interdisciplinary reviews. RNA*, 1(3):457–473.

Grimson, A., Farh, K.K.H., Johnston, W.K. et al (2007). MicroRNA Targeting Specificity in Mammals: Determinants Beyond Seed Pairing. *Molecular cell*, 27(1):91–105.

Gromadzka, A.M., Steckelberg, A.L., Singh, K.K. et al (2016). A short conserved motif in ALYREF directs cap- and EJC-dependent assembly of export complexes on spliced mRNAs. *Nucleic Acids Research*, 44(5):2348–2361.

Gruber, A.J., Schmidt, R., Gruber, A.R. et al (2016). A comprehensive analysis of 3 end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Research*, 26(8):1145.

Gu, Z. (2022). Complex heatmap visualization. *iMeta*, 1(3):e43.

Gu, Z., Eils, R. and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849.

Han, H., Irimia, M., Ross, P.J. et al (2013). MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature*, 498(7453):241–245.

Hanahan, D. and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674.

He, P.C. and He, C. (2021). m6A RNA methylation: from mechanisms to therapeutic potential. *The EMBO Journal*, 40(3):e105977.

Herzog, V.A., Reichholf, B., Neumann, T. et al (2017). Thiol-linked alkylation of RNA to assess expression dynamics. *Nature Methods*, 14(12):1198–1204.

Hillier, L.W., Coulson, A., Murray, J.I. et al (2005). Genomics in C. elegans: So many genes, such a little worm. *Genome Research*, 15(12):1651–1660.

Ho, C.K. and Shuman, S. (1999). Distinct Roles for CTD Ser-2 and Ser-5 Phosphorylation in the Recruitment and Allosteric Activation of Mammalian mRNA Capping Enzyme. *Molecular Cell*, 3(3):405–411.

Hoffman, Y., Bublik, D.R., Ugalde, A.P. et al (2016). 3'UTR Shortening Potentiates MicroRNA-Based Repression of Pro-differentiation Genes in Proliferating Human Cells. *PLoS Genetics*, 12(2).

Hogg, J.R. and Goff, S.P. (2010). Upf1 Senses 3UTR Length to Potentiate mRNA Decay. *Cell*, 143(3):379–389.

Holt, C.E., Martin, K.C. and Schuman, E.M. (2019). Local translation in neurons: visualization and function. *Nature Structural & Molecular Biology*, 26(7):557–566.

Huang, H., Weng, H., Sun, W. et al (2018). Recognition of RNA N6-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nature Cell Biology*, 20(3):285–295.

Hurlbut, J.B., Hyun, I., Levine, A.D. et al (2017). Revisiting the Warnock rule. *Nature Biotechnology*, 35(11):1029–1042.

Imamachi, N., Salam, K.A., Suzuki, Y. et al (2017). A GC-rich sequence feature in the 3 UTR directs UPF1-dependent mRNA decay in mammalian cells. *Genome Research*, 27(3):407–418.

Irion, U. and St Johnston, D. (2007). bicoid RNA localization requires specific binding of an endosomal sorting complex. *Nature*, 445(7127):554–558.

Ji, Z., Lee, J.Y., Pan, Z. et al (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences*, 106(17):7028–7033.

Jia, Q., Xie, B., Zhao, Z. et al (2022). Lung cancer cells expressing a shortened CDK16 3'UTR escape senescence through impaired miR-485-5p targeting. *Molecular Oncology*, 16(6):1347.

Jorquera, R., Ortiz, R., Ossandon, F. et al (2016). SinEx DB: a database for single exon coding sequences in mammalian genomes. *Database*, 2016:baw095.

Kahvejian, A., Svitkin, Y.V., Sukarieh, R. et al (2005). Mammalian poly(A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms. *Genes & Development*, 19(1):104–113.

Kashima, I., Yamashita, A., Izumi, N. et al (2006). Binding of a novel SMG-1–Upf1–eRF1–eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. *Genes & Development*, 20(3):355–367.

Keller, W., Bienroth, S., Lang, K. et al (1991). Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3 processing signal AAUAAA. *The EMBO Journal*, 10(13):4241–4249.

Kemmerer, K., Fischer, S. and Weigand, J.E. (2018). Auto- and cross-regulation of the hnRNPs D and DL. *RNA*, 24(3):324–331.

Khoroshkin, M., Buyan, A., Dodel, M. et al (2024). Systematic identification of post-transcriptional regulatory modules. *Nature Communications*, 15(1):7872.

Kim, H.H., Kuwano, Y., Srikantan, S. et al (2009). HuR recruits let-7/RISC to repress c-Myc expression. *Genes & Development*, 23(15):1743–1748.

Kim, J., Park, R.Y., Kee, Y. et al (2022). Splicing factor SRSF3 represses translation of p21cip1/waf1 mRNA. *Cell Death & Disease*, 13(11):1–11.

Kishor, A., Ge, Z. and Hogg, J.R. (2019). hnRNP L-dependent protection of normal mRNAs from NMD subverts quality control in B cell lymphoma. *The EMBO journal*, 38(3):e99128.

Kislauskis, E.H., Zhu, X. and Singer, R.H. (1994). Sequences responsible for intracellular localization of beta-actin messenger RNA also affect cell phenotype. *The Journal of Cell Biology*, 127(2):441–451.

Kobayashi, H., Yamamoto, S., Maruo, T. et al (2005). Identification of a cis-acting element required for dendritic targeting of activity-regulated cytoskeleton-associated protein mRNA. *The European Journal of Neuroscience*, 22(12):2977–2984.

Koboldt, D.C., Zhang, Q., Larson, D.E. et al (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576.

Koussounadis, A., Langdon, S.P., Um, I.H. et al (2015). Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Scientific Reports*, 5(1):10775.

Kozak, M. (1987). Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Molecular and Cellular Biology*, 7(10):3438–3445.

Kwon, Y.S. and Song, H. (2016). Analysis of microRNAs in a knock-in hESC line expressing epitope-tagged AGO2. *Animal Cells and Systems*, 20(1):24–30.

Laco, F., Lam, A.T.L., Woo, T.L. et al (2020). Selection of human induced pluripotent stem cells lines optimization of cardiomyocytes differentiation in an integrated suspension microcarrier bioreactor. *Stem Cell Research & Therapy*, 11:118.

Lappalainen, T., Sammeth, M., Friedländer, M.R. et al (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.

Lareau, L.F., Inada, M., Green, R.E. et al (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, 446(7138):926–929.

Le Hir, H., Gatfield, D., Izaurralde, E. et al (2001). The exon–exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *The EMBO Journal*, 20(17):4987–4997. Num Pages: 4997 Publisher: John Wiley & Sons, Ltd.

Lebedeva, S., Jens, M., Theil, K. et al (2011). Transcriptome-wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR. *Molecular Cell*, 43(3):340–352.

Leclair, N.K., Brugiolo, M., Urbanski, L. et al (2020). Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis. *Molecular Cell*, 80(4):648–665.e9.

Lee, S.H. and Mayr, C. (2019). Gain of Additional BIRC3 Protein Functions through 3-UTR-Mediated Protein Complex Formation. *Molecular Cell*, 74(4):701–712.e9.

Lewis, B.P., Burge, C.B. and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20.

Li, J.H., Liu, S., Zhou, H. et al (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-

ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, 42(D1):D92–D97.

Li, Y., Sun, Y., Fu, Y. et al (2012). Dynamic landscape of tandem 3′ UTRs during zebrafish development. *Genome Research*, 22(10):1899.

Liao, Y., Smyth, G.K. and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Linde, L., Boelz, S., Neu-Yilik, G. et al (2007). The efficiency of nonsense-mediated mRNA decay is an inherent character and varies among different cells. *European Journal of Human Genetics*, 15(11):1156–1162.

Lindeboom, R.G.H., Vermeulen, M., Lehner, B. et al (2019). The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. *Nature Genetics*, 51(11):1645–1651.

Linder, B., Grozhik, A.V., Olarerin-George, A.O. et al (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nature Methods*, 12(8):767–772.

Liu, Y., Beyer, A. and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165(3):535–550.

Loh, B., Jonas, S. and Izaurralde, E. (2013). The SMG5–SMG7 heterodimer directly recruits the CCR4–NOT deadenylase complex to mRNAs containing nonsense codons via interaction with POP2. *Genes & Development*, 27(19):2125–2138.

Long, D., Lee, R., Williams, P. et al (2007). Potent effect of target structure on microRNA function. *Nature Structural & Molecular Biology*, 14(4):287–294.

Lou, C.H., Dumdie, J., Goetz, A. et al (2016). Nonsense-Mediated RNA Decay Influences Human Embryonic Stem Cell Fate. *Stem Cell Reports*, 6(6):844–857.

Love, M., Soneson, C. and Patro, R. (2018). Swimming downstream: statistical analysis of

differential transcript usage following Salmon quantification [version 3; peer review: 3 approved]. *F1000Research*, 7(952).

Love, M.I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

Lyra-Leite, D.M., Gutiérrez-Gutiérrez, , Wang, M. et al (2022). A review of protocols for human iPSC culture, cardiac differentiation, subtype-specification, maturation, and direct reprogramming. *STAR Protocols*, 3(3):101560.

Macdonald, P.M. and Struhl, G. (1988). Cis- acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. *Nature*, 336(6199):595–598.

Maier, K.C., Gressel, S., Cramer, P. et al (2020). Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms. *Genome Research*, 30(9):1332.

Marson, A., Levine, S.S., Cole, M.F. et al (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, 134(3):521–533.

Mathonnet, G., Fabian, M.R., Svitkin, Y.V. et al (2007). MicroRNA Inhibition of Translation Initiation in Vitro by Targeting the Cap-Binding Complex eIF4F. *Science*, 317(5845):1764–1767.

May, J.P., Yuan, X., Sawicki, E. et al (2018). RNA virus evasion of nonsense-mediated decay. *PLOS Pathogens*, 14(11):e1007459.

Mayr, C. (2016). Evolution and Biological Roles of Alternative 3UTRs. *Trends in Cell Biology*, 26(3):227–237.

Mayr, C. (2019). What Are 3 UTRs Doing? *Cold Spring Harbor Perspectives in Biology*, 11(10):a034728.

Mayr, C. and Bartel, D.P. (2009). Widespread Shortening of 3UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell*, 138(4):673–684.

Meister, G., Landthaler, M., Patkaniowska, A. et al (2004). Human Argonaute2 Mediates RNA Cleavage Targeted by miRNAs and siRNAs. *Molecular Cell*, 15(2):185–197.

Melé, M., Ferreira, P.G., Reverter, F. et al (2015). The human transcriptome across tissues and individuals. *Science (New York, N.Y.)*, 348(6235):660–665.

Meyer, K.D., Patil, D.P., Zhou, J. et al (2015). 5 UTR m6A Promotes Cap-Independent Translation. *Cell*, 163(4):999–1010.

Meyer, K.D., Saletore, Y., Zumbo, P. et al (2012). Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3' UTRs and Near Stop Codons. *Cell*, 149(7):1635–1646.

Miller, A.D., Curran, T. and Verma, I.M. (1984). *c-fos* protein can induce cellular transformation: A novel mechanism of activation of a cellular oncogene. *Cell*, 36(1):51–60.

Min, K.W., Jo, M.H., Shin, S. et al (2017). AUF1 facilitates microRNA-mediated gene silencing. *Nucleic Acids Research*, 45(10):6064–6073.

Miura, P., Shenker, S., Andreu-Agullo, C. et al (2013). Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Research*, 23(5):812–825.

Mueller, A.A., Cheung, T.H. and Rando, T.A. (2013). All's Well that Ends Well: Alternative Polyadenylation and its Implications for Stem Cell Biology. *Current opinion in cell biology*, 25(2):222–232.

Mufteev, M., Rodrigues, D.C., Yuki, K.E. et al (2023). Transcriptional buffering and 3UTR lengthening are shaped during human neurodevelopment by shifts in mRNA stability and microRNA load.

Mukherjee, N., Corcoran, D.L., Nusbaum, J.D. et al (2011). Integrative regulatory mapping indicates that the RNA-binding protein HuR (ELAVL1) couples pre-mRNA processing and mRNA stability. *Molecular cell*, 43(3):327–339.

Murry, C.E. and Keller, G. (2008). Differentiation of Embryonic Stem Cells to Clinically Relevant Populations: Lessons from Embryonic Development. *Cell*, 132(4):661–680.

Muñoz, O., Lore, M. and Jagannathan, S. (2023). The long and short of EJC-independent nonsense-mediated RNA decay. *Biochemical Society Transactions*, 51(3):1121–1129.

Mölder, F., Jablonski, K.P., Letcher, B. et al (2021). Sustainable data analysis with Snakemake.

Nabet, B., Roberts, J.M., Buckley, D.L. et al (2018). The dTAG system for immediate and target-specific protein degradation. *Nature Chemical Biology*, 14(5):431–441.

Naeli, P., Winter, T., Hackett, A.P. et al (2023). The intricate balance between microRNA-induced mRNA decay and translational repression. *The FEBS Journal*, 290(10):2508–2524.

Nagy, E. and Maquat, L.E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends in Biochemical Sciences*, 23(6):198–199.

Nakatake, Y., Ko, S.B.H., Sharov, A.A. et al (2020). Generation and Profiling of 2,135 Human ESC Lines for the Systematic Analyses of Cell States Perturbed by Inducing Single Transcription Factors. *Cell Reports*, 31(7):107655.

Nam, J.W., Rissland, O.S., Koppstein, D. et al (2014). Global analyses of the effect of different cellular contexts on microRNA targeting. *Molecular cell*, 53(6):1031–1043.

Neu-Yilik, G., Amthor, B., Gehring, N.H. et al (2011). Mechanism of escape from nonsense-mediated mRNA decay of human β-globin transcripts with nonsense mutations in the first exon. *RNA*, 17(5):843.

Neumann, T., Herzog, V.A., Muhar, M. et al (2019). Quantification of experimentally induced nucleotide conversions in high-throughput sequencing datasets. *BMC Bioinformatics*, 20(1):258.

Nguyen, T.H.D., Galej, W.P., Bai, X.c. et al (2015). The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature*, 523(7558):47–52.

Nikom, D. and Zheng, S. (2023). Alternative splicing in neurodegenerative disease and the promise of RNA therapies. *Nature Reviews Neuroscience*, 24(8):457–473.

Nowicka, M. and Robinson, M.D. (2016). DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, 5:1356.

Ohnishi, T., Yamashita, A., Kashima, I. et al (2003). Phosphorylation of hUPF1 induces formation of mRNA surveillance complexes containing hSMG-5 and hSMG-7. *Molecular Cell*, 12(5):1187–1200.

Okada-Katsuhata, Y., Yamashita, A., Kutsuzawa, K. et al (2012). N- and C-terminal Upf1 phosphorylations create binding platforms for SMG-6 and SMG-5:SMG-7 during NMD. *Nucleic Acids Research*, 40(3):1251–1266.

Oleynikov, Y. and Singer, R.H. (2003). Real-time visualization of ZBP1 association with beta-actin mRNA during transcription and localization. *Current biology: CB*, 13(3):199–207.

Oltean, S. and Bates, D.O. (2014). Hallmarks of alternative splicing in cancer. *Oncogene*, 33(46):5311–5318.

Pan, Q., Shai, O., Lee, L.J. et al (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415.

Paolantoni, C., Ricciardi, S., De Paolis, V. et al (2018). Arc 3′ UTR Splicing Leads to Dual and Antagonistic Effects in Fine-Tuning Arc Expression Upon BDNF Signaling. *Frontiers in Molecular Neuroscience*, 11.

Patro, R., Duggal, G., Love, M.I. et al (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419.

Pendleton, K.E., Chen, B., Liu, K. et al (2017). The U6 snRNA m6A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention. *Cell*, 169(5):824–835.e14.

Peng, S.S., Chen, C.Y., Xu, N. et al (1998). RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein. *The EMBO Journal*, 17(12):3461–3470.

Peng, W.z., Zhao, J., Liu, X. et al (2021). hnRNPA2B1 regulates the alternative splicing of BIRC5 to promote gastric cancer progression. *Cancer Cell International*, 21(1):281.

Pertea, M., Pertea, G.M., Antonescu, C.M. et al (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295.

Popovitchenko, T., Thompson, K., Viljetic, B. et al (2016). The RNA binding protein HuR determines the differential translation of autism-associated FoxP subfamily members in the developing neocortex. *Scientific Reports*, 6(1):28998.

Proudfoot, N. (1991). Poly(A) signals. *Cell*, 64(4):671–674.

Proudfoot, N.J. and Brownlee, G.G. (1976). 3 Non-coding region sequences in eukaryotic messenger RNA. *Nature*, 263(5574):211–214.

Raineri, I., Wegmueller, D., Gross, B. et al (2004). Roles of AUF1 isoforms, HuR and BRF1 in ARE-dependent mRNA turnover studied by RNA interference. *Nucleic Acids Research*, 32(4):1279–1288.

Riley, J., Alexandru, C., Bryce-Smith, S. et al (2024). Widespread 3' UTR splicing regulates expression of oncogene transcripts in sequence-dependent and independent manners. *bioRxiv: The Preprint Server for Biology*.

Rivas, M.A., Pirinen, M., Conrad, D.F. et al (2015). Impact of predicted protein-truncating genetic variants on the human transcriptome. *Science (New York, N.Y.)*, 348(6235):666–669.

Rogalska, M.E., Vivori, C. and Valcárcel, J. (2023). Regulation of pre-mRNA splicing: roles

in physiology and disease, and therapeutic prospects. *Nature Reviews Genetics*, 24(4):251–269.

Romão, L., Inácio, A., Santos, S. et al (2000). Nonsense mutations in the human beta-globin gene lead to unexpected levels of cytoplasmic mRNA accumulation. *Blood*, 96(8):2895–2901.

Rossbach, O., Hung, L.H., Schreiner, S. et al (2009). Auto- and Cross-Regulation of the hnRNP L Proteins by Alternative Splicing. *Molecular and Cellular Biology*, 29(6):1442–1451.

Roundtree, I.A., Evans, M.E., Pan, T. et al (2017). Dynamic RNA modifications in gene expression regulation. *Cell*, 169(7):1187–1200.

Sandberg, R., Neilson, J.R., Sarma, A. et al (2008). Proliferating Cells Express mRNAs with Shortened 3′ Untranslated Regions and Fewer MicroRNA Target Sites. *Science*, 320(5883):1643–1647.

Sato, H. and Singer, R.H. (2021). Cellular variability of nonsense-mediated mRNA decay. *Nature Communications*, 12(1):7203.

Sharangdhar, T., Sugimoto, Y., Heraud-Farlow, J. et al (2017). A retained intron in the 3′-UTR of Calm3 mRNA mediates its Staufen2- and activity-dependent localization to neuronal dendrites. *EMBO reports*, 18(10):1762–1774.

Shaw, G. and Kamen, R. (1986). A conserved AU sequence from the 3 untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell*, 46(5):659–667.

Shen, S., Park, J.W., Lu, Z.x. et al (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, 111(51):E5593–E5601.

Shi, J., Yang, C., Zhang, J. et al (2023). NAT10 Is Involved in Cardiac Remodeling Through ac4C-Mediated Transcriptomic Regulation. *Circulation Research*, 133(12):989–1002.

Singh, R.N. and Singh, N.N. (2018). Mechanism of Splicing Regulation of Spinal Muscular Atrophy Genes. *Advances in neurobiology*, 20:31–61.

Soll, L.G., Eisen, J.N., Vargas, K.J. et al (2020). α-Synuclein-112 Impairs Synaptic Vesicle Recycling Consistent With Its Enhanced Membrane Binding Properties. *Frontiers in Cell and Developmental Biology*, 8:405.

Sommerkamp, P., Altamura, S., Renders, S. et al (2020). Differential Alternative Polyadenylation Landscapes Mediate Hematopoietic Stem Cell Activation and Regulate Glutamine Metabolism. *Cell Stem Cell*, 26(5):722–738.e7.

Soneson, C., Love, M.I. and Robinson, M.D. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences.

Spirov, A., Fahmy, K., Schneider, M. et al (2009). Formation of the bicoid morphogen gradient: an mRNA gradient dictates the protein gradient. *Development (Cambridge, England)*, 136(4):605–614.

St Johnston, D., Beuchle, D. and Nüsslein-Volhard, C. (1991). Staufen, a gene required to localize maternal RNAs in the Drosophila egg. *Cell*, 66(1):51–63.

Steward, O., Matsudaira Yee, K., Farris, S. et al (2018). Delayed Degradation and Impaired Dendritic Delivery of Intron-Lacking EGFP-Arc/Arg3.1 mRNA in EGFP-Arc Transgenic Mice. *Frontiers in Molecular Neuroscience*, 10:435.

Strober, B.J., Elorbany, R., Rhodes, K. et al (2019). Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290.

Tabar, V. and Studer, L. (2014). Pluripotent stem cells in regenerative medicine: challenges and recent progress. *Nature reviews. Genetics*, 15(2):82–92.

Takagaki, Y. and Manley, J.L. (1997). RNA recognition by the human polyadenylation factor CstF. *Molecular and Cellular Biology*, 17(7):3907.

Takahashi, K., Tanabe, K., Ohnuki, M. et al (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–872.

The Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337.

Thein, S.L. (2013). The Molecular Basis of β-Thalassemia. *Cold Spring Harbor Perspectives in Medicine*, 3(5):a011700.

Thomas, M.P., Liu, X., Whangbo, J. et al (2015). Apoptosis Triggers Specific, Rapid, and Global mRNA Decay with 3 Uridylated Intermediates Degraded by DIS3L2. *Cell reports*, 11(7):1079–1089.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S. et al (1998). Embryonic stem cell lines derived from human blastocysts. *Science (New York, N.Y.)*, 282(5391):1145–1147.

Tian, B., Hu, J., Zhang, H. et al (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33(1):201.

Ting, S., Chen, A., Reuveny, S. et al (2014). An intermittent rocking platform for integrated expansion and differentiation of human pluripotent stem cells to cardiomyocytes in suspended microcarrier cultures. *Stem Cell Research*, 13(2):202–213.

Treisman, R. (1985). Transient accumulation of c-*fos* RNA following serum stimulation requires a conserved 5 element and c-*fos* 3 sequences. *Cell*, 42(3):889–902.

Turunen, J.J., Niemelä, E.H., Verma, B. et al (2013). The significant other: splicing by the minor spliceosome. *WIREs RNA*, 4(1):61–76.

Uzonyi, A., Dierks, D., Nir, R. et al (2023). Exclusion of m6A from splice-site proximal regions by the exon junction complex dictates m6A topologies and mRNA stability. *Molecular Cell*, 83(2):237–251.e7.

Van den Berge, K., Soneson, C., Robinson, M.D. et al (2017). stageR: a general stage-wise

method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biology*, 18(1):151.

Van Nostrand, E.L., Freese, P., Pratt, G.A. et al (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, 583(7818):711–719.

Vicens, Q., Kieft, J.S. and Rissland, O.S. (2018). Revisiting the closed loop model and the nature of mRNA 5–3 communication. *Molecular cell*, 72(5):805–812.

Viegas, J.O., Azad, G.K., Lv, Y. et al (2022). RNA degradation eliminates developmental transcripts during murine embryonic stem cell differentiation via CAPRIN1-XRN2. *Developmental Cell*, 57(24):2731–2744.e5.

Viphakone, N., Sudbery, I., Griffith, L. et al (2019). Co-transcriptional Loading of RNA Export Factors Shapes the Human Transcriptome. *Molecular Cell*, 75(2):310–323.e8.

Wang, H., Zhou, M., Shi, B. et al (2011). Identification of an Exon 4-Deletion Variant of Epidermal Growth Factor Receptor with Increased Metastasis-Promoting Capacity. *Neoplasia (New York, N.Y.)*, 13(5):461–471.

Wang, L., Chen, M., Fu, H. et al (2020). Tempo-spatial alternative polyadenylation analysis reveals that 3' UTR lengthening of Mdm2 regulates p53 expression and cellular senescence in aged rat testis. *Biochemical and biophysical research communications*, 523(4).

Wang, S., Lv, W., Li, T. et al (2022). Dynamic regulation and functions of mRNA m6A modification. *Cancer Cell International*, 22(1):48.

Wang, X., Lu, Z., Gomez, A. et al (2014). m6A-dependent regulation of messenger RNA stability. *Nature*, 505(7481):117–120.

Wang, Y., Zhao, Y., Bollas, A. et al (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, 39(11):1348–1365.

Weil, J.E. and Beemon, K.L. (2006). A 3 UTR sequence stabilizes termination codons in the unspliced RNA of Rous sarcoma virus. *RNA*, 12(1):102–110.

Wells, S.E., Hillner, P.E., Vale, R.D. et al (1998). Circularization of mRNA by Eukaryotic Translation Initiation Factors. *Molecular Cell*, 2(1):135–140.

Welte, T., Tuck, A.C., Papasaikas, P. et al (2019). The RNA hairpin binder TRIM71 modulates alternative splicing by repressing MBNL1. *Genes & Development*, 33(17-18):1221–1235.

Wilkinson, M.E., Charenton, C. and Nagai, K. (2020). RNA Splicing by the Spliceosome. *Annual Review of Biochemistry*, 89(Volume 89, 2020):359–388.

Wilson, G.M., Sun, Y., Sellers, J. et al (1999). Regulation of AUF1 Expression via Conserved Alternatively Spliced Elements in the 3 Untranslated Region. *Molecular and Cellular Biology*, 19(6):4056–4064.

Withers, J.B. and Beemon, K.L. (2010). Structural features in the Rous sarcoma virus RNA stability element are necessary for sensing the correct termination codon. *Retrovirology*, 7:65.

Wright, C.J., Smith, C.W.J. and Jiggins, C.D. (2022). Alternative splicing as a source of phenotypic diversity. *Nature Reviews Genetics*, 23(11):697–710.

Wu, Q., Ferry, Q.R.V., Baeumler, T.A. et al (2017). In situ functional dissection of RNA cis-regulatory elements by multiplex CRISPR-Cas9 genome engineering. *Nature Communications*, 8:2109.

Wu, R., Li, A., Sun, B. et al (2019). A novel m6A reader Prrc2a controls oligodendroglial specification and myelination. *Cell Research*, 29(1):23–41.

Xia, Z., Donehower, L.A., Cooper, T.A. et al (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3-UTR landscape across seven tumour types. *Nature Communications*, 5(1):5274.

Yang, C., Hu, Y., Zhou, B. et al (2020). The role of m6A modification in physiology and disease. *Cell Death & Disease*, 11(11):960.

Yang, M., Liu, M., Sánchez, Y.F. et al (2023). A novel protocol to derive cervical motor neurons from induced pluripotent stem cells for amyotrophic lateral sclerosis. *Stem Cell Reports*, 18(9):1870–1883.

Yang, Q., Gilmartin, G.M. and Doublié, S. (2010). Structural basis of UGUA recognition by the Nudix protein CFIm25 and implications for a regulatory role in mRNA 3 processing. *Proceedings of the National Academy of Sciences of the United States of America*, 107(22):10062–10067.

Yang, X., Triboulet, R., Liu, Q. et al (2022). Exon junction complex shapes the m6A epitranscriptome. *Nature Communications*, 13(1):7904.

Yepiskoposyan, H., Aeschimann, F., Nilsson, D. et al (2011). Autoregulation of the nonsense-mediated mRNA decay pathway in human cells. *RNA*, 17(12):2108–2118.

Yergert, K.M., Doll, C.A., O'Rouke, R. et al (2021). Identification of 3′ UTR motifs required for mRNA localization to myelin sheaths in vivo. *PLoS Biology*, 19(1):e3001053.

Yesbolatova, A., Saito, Y., Kitamoto, N. et al (2020). The auxin-inducible degron 2 technology provides sharp degradation control in yeast, mammalian cells, and mice. *Nature Communications*, 11(1):5701.

Yoon, J.H., De, S., Srikantan, S. et al (2014). PAR-CLIP analysis uncovers AUF1 impact on target RNA fate and genome integrity. *Nature Communications*, 5(1):5248.

Zandhuis, N.D., Nicolet, B.P. and Wolkers, M.C. (2021). RNA-Binding Protein Expression Alters Upon Differentiation of Human B Cells and T Cells. *Frontiers in Immunology*, 12:717324.

Zetoune, A.B., Fontanière, S., Magnin, D. et al (2008). Comparison of nonsense-mediated mRNA decay efficiency in various murine tissues. *BMC Genetics*, 9(1):83.

Zhang, H.L., Eom, T., Oleynikov, Y. et al (2001). Neurotrophin-induced transport of a beta-

actin mRNP complex increases beta-actin levels and stimulates growth cone motility. *Neuron*, 31(2):261–275.

Zhang, J. and Maquat, L.E. (1997). Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells. *The EMBO journal*, 16(4):826–833.

Zhang, W., Wagner, B.J., Ehrenman, K. et al (1993). Purification, characterization, and cDNA cloning of an AU-rich element RNA-binding protein, AUF1. *Molecular and Cellular Biology*, 13(12):7652–7665.

Zhang, Y., Cai, J.Z., Xiao, L. et al (2020). RNA-binding protein HuR regulates translation of vitamin D receptor modulating rapid epithelial restitution after wounding. *American Journal of Physiology - Cell Physiology*, 319(1):C208–C217.

Zhang, Z., Hong, Y., Xiang, D. et al (2015). MicroRNA-302/367 Cluster Governs hESC Self-Renewal by Dually Regulating Cell Cycle and Apoptosis Pathways. *Stem Cell Reports*, 4(4):645–657.

Zhao, X., Yang, Y., Sun, B.F. et al (2014). FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Research*, 24(12):1403–1419.

Zheng, S., Gray, E.E., Chawla, G. et al (2012). Psd-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. *Nature neuroscience*, 15(3):381–S1.