University College London

RNA dysregulation in models of frontotemporal dementia and amyotrophic lateral sclerosis

Doctoral Thesis

Author:
Jack Humphrey

Supervisors:
Prof Adrian Isaacs
Dr Vincent Plagnol
Dr Pietro Fratta

UCL Institute of Neurology

UCL Genetics Institute

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Jack Humphrey

Abstract

Amyotrophic Lateral Sclerosis (ALS) and Frontotemporal dementia (FTD) are two rare but devastating neurodegenerative diseases that share pathological features and genetic factors. A central question in both diseases is the role of the RNA-binding proteins transactive response DNA-binding protein 43kDa (TDP-43) and fused in sarcoma (FUS). These proteins play a vital role in RNA regulation in all cells but in diseased neurons they alter their cellular localisation to form potentially pathogenic aggregates. This process can be linked to rare genetic mutations in the *TARDBP* and *FUS* genes, although most cases of ALS and FTD have no known genetic cause.

My work uses the revolutionary technology of RNA sequencing to measure and compare gene expression and RNA splicing in different cellular and animal models of sporadic and genetic disease. Here I present the results of four studies that investigate the biology of TDP-43 and FUS, assessing both their normal cellular roles and the impact of rare disease-causing mutations.

In these projects I analyse RNA sequencing data to discover novel gene expression and RNA splicing phenomena. This includes the repression of cryptic splicing by TDP-43 but not FUS, the progressive downregulation of mitochondrial and ribosomal transcripts in a mouse model of FUS ALS, a gain of splicing function by TDP-43 mutations affecting constitutive exon splicing, and widespread changes in intron retention caused by FUS knockout or aggressive FUS mutations. I also discover a novel mechanism for how FUS might regulate its own translation.

This work expands on what is currently known about the roles in RNA regulation for TDP-43 and FUS and provides new avenues for understanding both the causes and progression of ALS and FTD.

Impact Statement

This thesis comprises four studies on TDP-43 and FUS, two proteins linked to the devastating neurodegenerative diseases Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. I analysed RNA sequencing data to develop insights in to the functions of these two proteins in different models of disease. This work has uncovered new roles for TDP-43 and FUS in specific types of RNA regulation. Additionally, I have discovered new mechanisms to explain how disease-associated mutations in the two proteins affect these roles. Conclusions from my work are applicable to the RNA biology field as a whole as well as efforts to understand and treat these diseases. Insights into RNA-binding proteins and RNA regulation from this work can be transferred to many non-neurological diseases including muscular diseases, retinal diseases and certain cancers (Scotti and Swanson, 2015). Several of my chapters focus on developing new methods or adapting existing methods to analyse RNA splicing, a key mechanism of RNA regulation. RNA-sequencing is becoming a standard experiment across all fields of biology and the use of sequencing data in analysis of RNA splicing is an area of extreme interest and method development.

Three peer-reviewed papers have been published from this thesis. As of November 2018, "Quantitative analysis of cryptic splicing associated with TDP-43 depletion" (Humphrey et al., 2017) has been cited 11 times, "Humanized mutant FUS drives progressive motor neuron degeneration without aggregation in FUS Delta14 knockin mice" (Devoy et al., 2017) has been cited 11 times and "Mice with endogenous TDP-43 mutations exhibit gain of splicing function and characteristics of amyotrophic lateral sclerosis" (Fratta et al., 2018) has been cited 9 times and was the subject of a "News and Views" article in *The EMBO Journal* (Rouaux et al., 2018). These outcomes illustrate the importance of this work to the neurodegenerative disease and RNA biology fields.

Three new mouse models of ALS/FTD have been generated for projects carried out in this thesis: the FUS Δ14 mouse, the TDP-43 RRM2mut mouse, and the TDP-43 LCDmut mouse. All three mice are available for labs around the world to use in further experiments. RNA-seq data created from these mice have been made publicly available and can be downloaded from the Sequence Read Archive[1]. Several of my chapters rely on the re-analysis of published RNA sequencing data. I am pleased to see that sequencing data analysed in this thesis (chapter 5) has been reused in a further paper on TDP-43 (Sivakumar et al., 2018).

I have tried to make all software written during this thesis publicly available where possible. This can be downloaded from the GitHub platform. This includes the RNA-seq analysis pipeline used in all chapters[2], software developed to find and classify cryptic splicing (chapter 3)[3], and software written to classify splicing events found in the TDP-43 LCDmut and RRM2mut mice (chapter 5)[4].

[1] `https://www.ncbi.nlm.nih.gov/sra`
[2] `https://github.com/plagnollab/RNASeq_pipeline`
[3] `https://github.com/jackhump/CryptEx`
[4] `https://github.com/jackhump/Two_TDP-43_Mutant_Mice`

Contents

List of Figures

List of Tables

List of abbreviations

ALS	Amyotrophic lateral sclerosis
CLIP	UV crosslinking and immunoprecipitation
eCLIP	Enhanced CLIP
ES cell	Embryonic stem cell
ENU	N-ethyl-N-nitrosourea
FDR	False discovery rate
FTD	Frontotemporal dementia
FUS	RNA-binding protein Fused in Sarcoma
GO	Gene ontology
GWAS	Genome-wide association study
hnRNP	Heterogeneous nuclear ribonucleoprotein
iCLIP	Individual nucleotide resolution CLIP
KO	Knockout
LCD	Low complexity domain
LCDmut	TDP-43 Low complexity domain mutation
LINE	Long interspersed nuclear element
MUT	Mutant
NLS	Nuclear localisation signal
NMD	Nonsense-mediated decay
PSI	Percentage spliced in
PTC	Premature termination codon
RBP	RNA-binding protein
RNA	Ribonucleic acid
RRM	RNA-recognition motif
RRM2mut	TDP-43 RNA recognition motif 2 mutation
RNA-seq	RNA sequencing
RT-PCR	Reverse transcription polymerase chain reaction
SINE	Short interspersed nuclear element
snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleoprotein
SRA	Sequence read archive
TDP-43	Transactive response DNA binding protein, 43 kDa (protein)
TARDBP	Transactive response DNA binding protein (gene)
TSS	transcription start site
UTR	untranslated region
WT	Wildtype

Acknowledgements

My interest in the genetics of neurological disease was kindled by my time during my undergraduate degree with James Cox in the lab of John Wood. While there I realised the importance of computational methods and the insight they can bring to biology. When I began my PhD in September 2014 I knew nothing about RNA sequencing or computer programming, nor the great body of work on the molecular basis of ALS and FTD. That I've managed to complete a thesis on these topics is entirely down to the kindness and patience of so many brilliant scientists in both the Genetics Institute and the Institute of Neurology.

I want to thank Warren Emmett and Kitty Lo for being my closest mentors and friends in the Plagnol lab. Warren in particular taught me everything I know about splicing, and quite a lot about life too. In my office at the Darwin Building I have to thank Lucy Van Dorp, Dave Curtis, Chris Steele, Cian Murphy, Seth Jarvis, Claire Tkacz, Mike Scott, Leilei Cui, Flo Camus, Niko Pontikos and Francois Balloux for making computational biology an exciting and fun discipline to work in.

From my colleagues in the Fratta lab I must specifically thank Agnieszka Ule, Nicol Birsa and Prasanth Sivakumar for providing me with both experimental data and with friendship. In the Institute of Neurology I worked with some wonderful people including Anny Devoy, Justin Tosh, Laura Pulford (RIP), Lizzy Fisher, Lauren Gittings, Katie Wilson, Teresa Niccoli, Nejc Haberman, Igor Ruiz De Los Mozos, Nobby Chakrabarti and Charlotte Capitanchik as well as the rest of the Fratta, Isaacs, Fisher, Ule, and Luscombe labs. None of this work would have been possible without the fantastic administrative support of Rosie Baverstock-West, Agata Blaszczyk, Susmita Datta, James Michaels, Tristan Clark and David Gregory.

I am truly grateful for the support given to me by my three supervisors: Adrian Isaacs, Vincent Plagnol and Pietro Fratta. My PhD has been deeply rewarding and I feel I've really benefited from their diverse experiences and approaches, as well as occasional differences of opinion. It's truly been a dream team of PhD supervision and I look forward to collaborating with you all in the future.

Finally, I owe a deep debt of gratitude to my friends and family for the love and support they've shown me throughout these 4 years.

For Aliss Pollock

Without you the world would be a far duller place.

1 Introduction

1.1 Amyotrophic Lateral Sclerosis and Frontotemporal Dementia comprise a spectrum of disease

Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disorder primarily affecting the motor neurons of the cerebral cortex and the spinal cord. It affects between 2 and 16 people per 100,000 (Logroscino et al., 2010). Patients gradually lose voluntary motor control of their limbs and the muscles involved in speech and swallowing. Death usually occurs within 2-3 years after the first sign of symptoms, usually from infection caused by the inability to swallow. Frontotemporal Dementia (FTD) is a progressive neurodegenerative disorder primarily affecting the frontal and temporal lobes of the cerebral cortex. It affects 15-22 people per 100,000 and is the second most common dementia after Alzheimer's disease (Onyike and Diehl-Schmid, 2013). Depending on the subtype of FTD, patients exhibit worsening behavioural inhibition, language production or comprehension. Both disorders peak in incidence at around 60 years of age, are invariably fatal and have no cure. These two disorders are now recognised to be two ends of a spectrum of disease called ALS/FTD. This is in part due to a sharing of symptoms in some cases, as FTD patients can exhibit motor deficits and ALS patients can exhibit cognitive decline, but also due to a striking concordance in pathology and genetics.

Both disorders have recognisable brain pathology upon autopsy, with the affected brain regions showing aggregated protein inclusions in the nucleus and cytoplasm of neurons and glia. In FTD around 35% of patients have inclusions that contain Tau, a microtubule-associated protein also linked to Parkinson's and Alzheimer's disease (Rademakers et al., 2004). The rest of FTD patients present inclusions containing one of two proteins: TAR DNA-binding protein 43kDa (TDP-43) (Neumann et al., 2006; Arai et al., 2006) or fused in sarcoma (FUS) (Neumann et al., 2009). In ALS almost all patients present with TDP-43 positive inclusions (Neumann et al., 2006; Arai et al., 2006) and a small number display FUS inclusions (Vance et al., 2009; Kwiatkowski et al., 2009), firmly cementing the link between the two disorders and a suggesting key role for TDP-43 and FUS in ALS/FTD.

The progress in understanding the pathology of ALS/FTD has been mirrored by the progress in locating causative genes. This was initially done by linkage studies, where blocks of shared genetic variation were identified in the affected members of large families. *SOD1* was the first gene linked to ALS in series of families over 20 years ago (Rosen et al., 1993). Patients with SOD1 mutations have SOD1 containing inclusions and not TDP-43. Mutations in *MAPT*, the gene coding for the Tau protein were then found in familial FTD cases (Hutton et al., 1998), linking the protein pathology with alterations to the gene itself. This theme continued in the discovery a series of rare mutations in *TARDBP*, the gene that codes for

TDP-43 in familial cohorts of ALS, FTD and combined ALS/FTD (Sreedharan et al., 2008; Kabashi et al., 2008; Benajiba et al., 2009; Borroni et al., 2009). Mutations in FUS were then found in ALS patients (Vance et al., 2009; Kwiatkowski et al., 2009), completing the link between pathology and genetics. Shortly after this, a long-standing mystery was solved. Multiple ALS and FTD families had been linked to a region on chromosome 9, which was revealed to be a large expansion in the intron of the *C9orf72* gene (Renton et al., 2011; DeJesus-Hernandez et al., 2011). In individuals of Caucasian ancestry the expansion is found in 5-10% of sporadic ALS and FTD cases, 40% of ALS and 25% of FTD cases with a family history (Majounie et al., 2012), more than all the other known genes put together, making *C9orf72* the single largest genetic contribution to ALS and a major contributor to FTD. Patients with *C9orf72* expansions have TDP-43 pathology, as well as RNA foci containing the expanded *C9orf72* RNA and dipeptide repeat proteins which are translated from the *C9orf72* repeat sequence in both sense and antisense direction (DeJesus-Hernandez et al., 2011; Renton et al., 2011).

The emergence of next-generation sequencing technologies has moved the gene hunting field from conducting linkage in family pedigrees to large-scale studies comparing the allele frequencies between groups of affected and unaffected people, at first in exomes (the total protein coding portion of the genome) and now to whole genomes. These studies have found extremely rare mutations that individually account for a very small number of cases but provide vital information on which cellular pathways are involved in disease (Taylor et al., 2016; Pottier et al., 2016). Broadly, the proteins these genes code for can be grouped by their functions into three categories. *OPTN* (Maruyama et al., 2010) , *UBQLN2* (Deng et al., 2011), *SQSTM1* (Fecto et al., 2011), *CHMP2B* (Skibinski et al., 2005) and *TBK1* (Cirulli et al., 2015; Freischmidt et al., 2015) have all been linked to protein degradation. *MAPT*, *DCTN1* (Puls et al., 2003) , *CHCHD10* (Bannwarth et al., 2014), *TUBA4A* (Smith et al., 2014) and *KIF5A* (Nicolas et al., 2018) have been linked to microtubule transport and stability. The third group of genes encode RNA-binding proteins, and this is the function of *TARDBP* and *FUS*. Other RNA-binding proteins associated with ALS/FTD through genetics are *ATXN2* (Elden et al., 2010), *TAF15* (Ticozzi et al., 2011), *hnRNPA1* and *hnRNPA2B1* (Kim et al., 2013), *MATR3* (Johnson et al., 2014), *SFPQ* (Thomas-Jinu et al., 2017), and *TIA1* (Mackenzie et al., 2017). The proteins these genes code for have been linked to splicing, transcription, translation and transport of mRNA. The converging evidence from the pathology and genetics together create the RNA hypothesis of ALS/FTD, where impaired RNA regulation due to mutations or mislocalisation of RNA binding proteins is progressively toxic to neurons. Understanding and refining this hypothesis is the focus of my PhD, although the other genetic results suggest it is only one facet of a complex set of disease pathways.

1.2 RNA splicing is a key step for RNA regulation

The discovery of a large discrepancy between the length of a gene and the length of its mRNA ushered in new paradigm for biology: RNA splicing (Berget et al., 1977; Chow

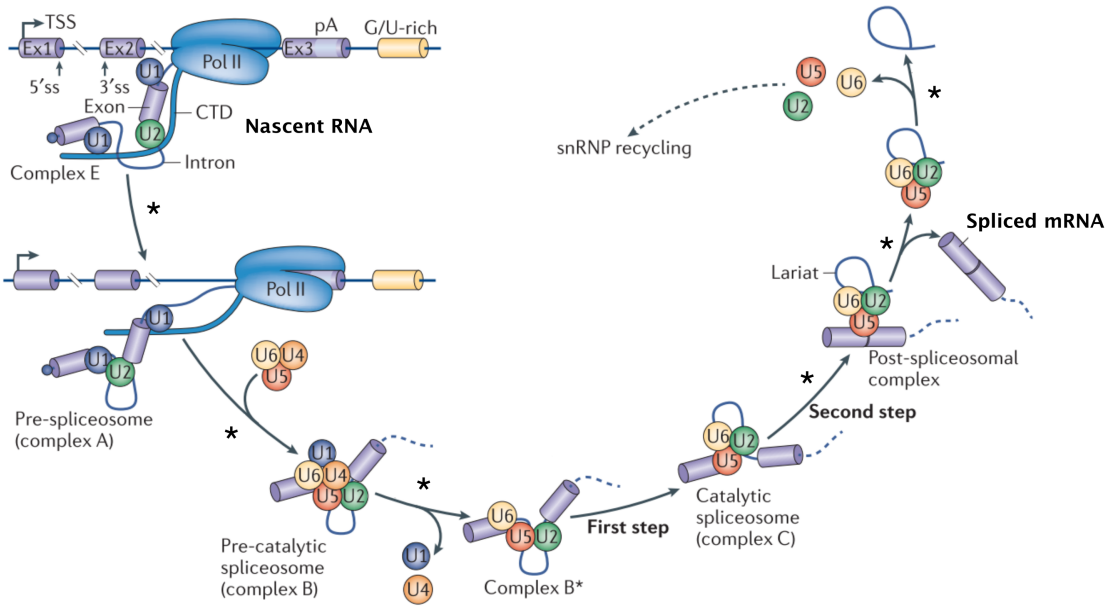


Figure 1.1: Splicing of a U2-dependent intron. Splicing proceeds through the recruitment of snRNP particles on the nascent RNA and the formation of different complexes, which transition through different arrangements. * denotes the transition is dependent on the action of ATP and RNA helicases to proceed. Figure taken from (Matera and Wang, 2014). Names of specific helicases have been removed for simplicity.

et al., 1977). This is in essence a process of distinguishing which sections of a nascent RNA molecule are to be kept (the exons), and which are to be removed (the introns) to create a final transcript. The majority of protein-coding genes are made up of multiple exons and so require splicing for the creation of their mature mRNAs. Beyond this, 95% of multi-exon genes are alternatively spliced (Pan et al., 2008; Wang et al., 2008). Alternative splicing is the process where a particular set of exons in a gene are selected to be spliced together. For protein-coding genes this is a mechanism for creating functional diversity in the proteome, where a single gene can encode multiple proteins with different functions. This has consequences for understanding gene regulation and cell biology.

Splicing depends on the reliable recognition of exons and subsequent removal of introns by the spliceosome complex. The spliceosome is a dynamic molecular machine made up of small nuclear ribonucleoprotein (snRNP) complexes containing small nuclear RNAs (snRNA) and their partnering proteins (Matera and Wang, 2014). Splicing proceeds by a set of interactions between the nascent RNA and the snRNP complexes (Fig. 1.1). Eukaryotes have two types of introns which are spliced by two different spliceosomes made up of mostly different sets of snRNPs: major or U2-dependent introns, and minor or U12-dependent introns (Tarn and Steitz, 1996). Major introns make up 99.5% of human introns. I will describe the process of U2-dependent intron splicing.

snRNAs form base-pairing interactions with consensus RNA sequences called splice sites. These demarcate the boundaries between introns and exons. The 5´ splice site is recognised by the U1 snRNP and the 3´ splice site and the upstream branch point sequence is recognised by the U2 snRNP. U1 snRNP binding precedes the U2 snRNP and the two complexes first

pair together across an exon in a process known as exon definition. This is due to the long length of introns in higher eukaryotes that immediate pairing across the intron is unfavoured. This complex transitions to an intron-defined arrangement across the adjacent intron which brings the two splice sites and the branch point into close contact. At this point the U4/5/6 tri-snRNP is recruited and through a set of ATP-dependent rearrangements the spliceosome catalyses two transesterification reactions to remove the intron and join the two exons together. The intron is circularised at the branchpoint to form a lariat structure which is then degraded.

During splicing, other sequences in the nascent RNA can recruit RNA-binding proteins to modulate splicing by promoting or repressing the splicing of particular exons. This is the mechanism behind alternative splicing and is an interplay between the *cis*-acting RNA sequence and *trans*-acting proteins. Whether an exon is included in a mature mRNA transcript depends on the local environment around the splicing reaction. Sequence motifs can be classified based on which factors they recruit and by what the result of binding is, whether binding enhances or silences the splicing of the exon. Therefore a motif in an exon that promotes its inclusion is an exonic splicing enhancer element. Whether a sequence is used to silence or enhance the splicing of an exon depends on its location within the exon-intron structure. For example, the neuro-oncological ventral antigen (NOVA) proteins have a binding preference for YCAY sequences, where Y indicates a pyrimidine (Buckanovich et al., 1996; Jensen et al., 2000). NOVA binding directly immediately downstream or distally upstream of an exon acts to promote its inclusion, whereas binding directly on top of an exon or adjacent to the upstream exon promotes exon skipping (Ule et al., 2006). Multiple splicing factors can bind the same sequence elements and function as a network depending on their levels of expression (Wang et al., 2013). These splicing networks can be created by protein-protein interactions between factors or through protein-RNA interactions whereby a splicing factor can control its own splicing and the splicing of other factors. Therefore different tissues or development time points can promote the splicing of particular sets of genes through changing the expression levels of different RNA-binding proteins.

The inclusion of a particular exon can have a wide range of consequences for an RNA molecule and its eventual protein, if it is to be translated. The inclusion of an alternative or cassette exon may alter stability, add a new functional domain, change localisation or change protein-protein interactions. One way that splicing can alter RNA stability is through the nonsense-mediated decay (NMD) pathway. Exons can destabilise their host transcript by introducing premature termination codons (PTCs), stop codons that appear upstream of the final stop codon. These can make a transcript sensitive to degradation by NMD (McGlincy and Smith, 2008). This occurs during the pioneer round of translation, where the ribosome detects a PTC appearing before the final exon-exon junction, which is demarcated by the exon-junction complex. This triggers degradation of the transcript. If the PTC is within 50 nucleotides of the final exon-exon junction, the transcript can escape NMD. Intriguingly, splicing factors themselves often have NMD-sensitive isoforms (Ni et al., 2007). The splicing of NMD-sensitive isoforms in a splicing factor transcript is often triggered by the binding of that same splicing factor protein (Jangi and Sharp, 2014). This is a mechanism by which

splicing factors regulate their own translation: a phenomenon known as autoregulation (Rosenfeld et al., 2002).

Splicing generally occurs as the nascent RNA is being transcribed by RNA polymerase II (Beyer and Osheim, 1988; Ameur et al., 2011). This allows for interaction behind the transcription and splicing machinery. Two reciprocal models for this have been proposed: a recruitment model, where RNA motifs recruit splicing factors which themselves recruit transcriptional modifiers, and a kinetic model, where the modulation of elongation speed alters this recruitment by changing the availability of RNA motifs for recruitment (Kornblihtt et al., 2004). Alternate exons have weaker splice sites, whose sequence deviates from the high affinity consensus sequence (Stamm et al., 2000). Pausing or slowing transcription speed can allow these weaker splice sites to be recognised and compete with stronger constitutive splice sites to promote alternate exon inclusion. Alternatively, more time can allow for low-affinity silencing elements to be used to promote exon skipping. A number of exons sensitive to transcription speed are found in genes for splicing factors (Ip et al., 2011), suggesting deep connections between transcription and splicing.

Of all the cells in the human body, neurons arguably make the largest demands upon the transcription and splicing machinery. Neuron-specific genes tend to be much longer than in other tissues (Sibley et al., 2015) and an individual neuronal gene can be processed by alternate splicing to create 1000s of mRNAs and subsequent protein isoforms (Treutlein et al., 2014). The distinct compartments of a neuron's architecture requires exquisite fine-tuning of protein function to suit its location, for example on either side of a synapse. There is also the matter of transport. Motor neurons can have axons over 1 metre in length, along which an mRNA would have to travel to reach ribosomes close to a synapse for local translation. Small defects in splicing could have catastrophic consequences for neurons and motor neurons in particular.

One example of neuronal vulnerability to splicing dyregulation is the phenomenon of cryptic exons. Due to the reduced evolutionary conservation of intronic sequence, pairs of 3´ and 5´ splice sites can emerge randomly to create new exons, with long neuronal introns being most vulnerable. These cryptic exons (also known as pseudoexons) can arise due to mutations that create new splice sites or remove the existing binding sites for splicing repressors. These type of mutations have been implicated in a number of genetic diseases (Eng et al., 2004; Buratti et al., 2007a; Vorechovsky, 2006; Meili et al., 2009). Inclusion of a cryptic exon, untested by evolution, can destabilise its host transcript or radically alter the eventual protein structure.

Another source of cryptic exons are transposable elements. One such example are Alu elements, the predominant transposable element in primates which are often found within introns in the antisense direction (Deininger and Prescott, 2011). The consensus Alu sequence consists of two arms joined by an adenine-rich linker ending with a poly-adenine tail. When transcribed in the antisense direction these uridine-rich sequences can act as cryptic polypyrimidine tracts and only a few mutations are required to convert them into viable exons in a process termed exonisation (Sorek et al., 2002). *De novo* mutations that lead to Alu exonisation have been found in a range of diseases (Vorechovsky, 2010) suggesting a

need for regulation of potentially damaging Alu exons. Alu exonisation is repressed by the RNA binding protein hnRNP C, which competes with the spliceosome component protein U2AF65, the partner of the U2 snRNA, for binding cryptic 3´ splice sites (Zarnack et al., 2013). Due to the potentially negative effects of incorporation of new exons, recognition of cryptic exons needs to be repressed. It is unknown how many other RNA-binding proteins play a role in repressing the recognition of cryptic splice sites.

Splicing is a key biological mechanism for increasing protein function and regulating gene expression. Studying RNA processing has been made substantially easier and more powerful by the emergence of new technology to survey the entire transcriptome at once: RNA sequencing.

1.3 RNA-sequencing is a revolutionary technology to quantify gene expression and splicing

RNA sequencing, henceforth written as RNA-seq, is the application of modern high throughput sequencing to directly determine the sequences and quantity of RNA molecules in a group of cells (Wang et al., 2009). Unlike the older microarray technology which relies on choosing a set of RNA probes to measure, RNA-seq is hypothesis-free. It is also highly sensitive, allowing for the detection of lowly expressed transcripts. Instead of measuring the intensity of a microarray probe, the abundance of a particular RNA molecule is calculated simply by counting the number of sequencing reads that contain its sequence. As sequencing technology has improved and reduced in cost, more complicated aspects of RNA regulation are now observable. Alternative splicing can be measured by the number of sequencing reads split across multiple exons: the splice junction. Complicated isoforms can be reconstructed from splice junctions where sequencing is sufficiently deep.

The real power of an RNA-seq experiment is that it is an open platform. Once the data is generated it can be downloaded and used in light of whatever the most up-to-date references, annotations or novel hypotheses happen to be. As it is now a requirement for publication that all raw sequence data is made available, this allows for large scale re-analysis and meta-analysis in light of new discoveries and ideas. Throughout my work I capitalise on this by re-analysing public datasets and combining their results to find new and interesting biology. This ease of replication is a triumph for modern science and will hopefully lead to more robust findings.

1.4 RNA-binding proteins implicated in ALS/FTD

The RNA hypothesis of ALS and FTD has emerged from evidence from pathology and genetics that centres around RNA-binding proteins, chief among them TDP-43 and FUS. My work attempts to better understand the biological roles of these two proteins and link them to animal models of disease.

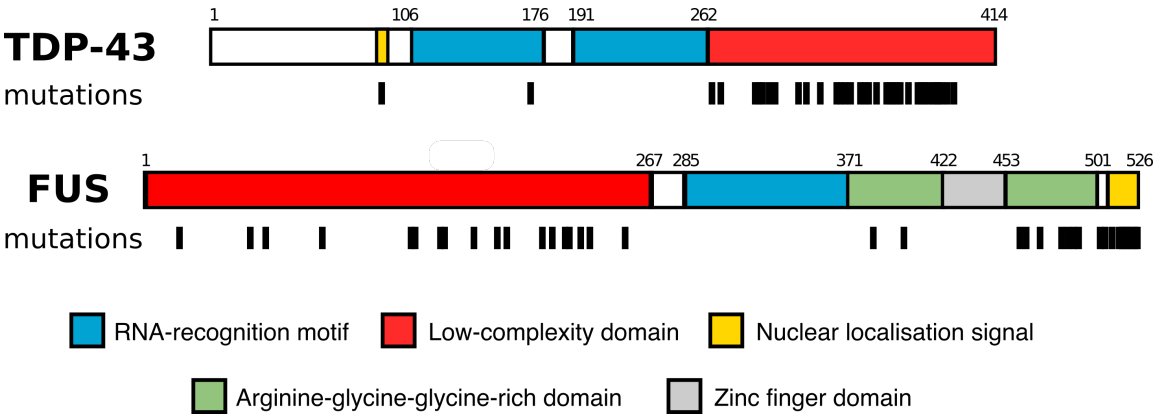


Figure 1.2: Protein domains of TDP-43 and FUS. Structures of the two proteins, coloured by functional domain. Positions of each mutation are represented by black bars. Figure adapted from (Kapeli et al., 2017).

TDP-43

TAR DNA-binding protein 43 kDa (TDP-43) is a ubiquitously expressed RNA and DNA-binding protein encoded by the *TARDBP* gene. It contains two RNA-recognition motifs, a nuclear localisation signal and a long glycine-rich low-complexity domain (Fig 1.2). Loss of TDP-43 from the nucleus accompanied by TDP-43 positive inclusions in the cytoplasm of cortical and spinal cord neurons is the hallmark pathology found in nearly all sporadic cases of ALS as well as the majority of non-Tau cases of FTD (Neumann et al., 2006; Arai et al., 2006) as well as cases of Alzheimer's disease (LaClair et al., 2016) and Huntington's Disease (Doi et al., 2008). Cytosolic TDP-43 has also been found in mouse models of traumatic brain injury, suggesting mislocalisation is a general neuronal stress response (Moisse et al., 2009). In addition, missense mutations in *TARDBP* have been shown to cause familial ALS (Sreedharan et al., 2008). Mutations cluster in the low-complexity domain (Kapeli et al. 2017; Fig. 1.2). These findings point to a key role of TDP-43 in the development of ALS/FTD. There is still much debate on whether the mislocalisation of TDP-43 plays a role in neurodegeneration through a loss of normal nuclear function or a gain of cytoplasmic function. A further question is the influence of rare mutations in the TDP-43 low-complexity domain and how they contribute to pathology.

Roles in RNA regulation

TDP-43 is a predominantly nuclear protein but can shuttle to the cytoplasm (Ayala et al., 2008). This has implicated it both in nuclear roles in transcription and splicing, but also in RNA transport and translation. In splicing, TDP-43 binds UG-rich RNA motifs (Buratti et al., 2001; Buratti and Baralle, 2001b), validated by X-ray crystallography (Lukavsky et al., 2013). TDP-43 binding can either repress or enhance cassette exon splicing depending on the motif position, as was found for individual genes (Mercado et al., 2005; Bose et al., 2008; Shiga et al., 2012). This was expanded genome-wide by RNA-protein interaction experiments to find RNA targets of TDP-43 and correlate TDP-43 binding with cassette exon splicing (Polymenidou et al., 2011; Tollervey et al., 2011; Kapeli et al., 2016). For

cassette exons, TDP-43 binds on top of or close to exons which it represses and further away from the exons it enhances (Tollervey et al., 2011), similarly to NOVA proteins. However, the majority of TDP-43 binding was found to be deep within introns and on 3´ untranslated regions (3´ UTRs). Long intron genes were strongly downregulated under TDP-43 depletion (Polymenidou et al., 2011), suggesting an important role for TDP-43 in their splicing and stabilisation. TDP-43 knockdown was found to cause the emergence of a set of cryptic exons from poorly conserved introns (Ling et al., 2015), suggesting that intronic binding of TDP-43 acts to repress cryptic splice sites. Unlike hnRNP C this has not been linked to a particular class of retrotransposon element, despite TDP-43 being observed to bind a range of retrotransposons including Alu elements (Li et al., 2012; Zarnack et al., 2013; Kelley et al., 2014). Attempts to correlate TDP-43's effect on the splicing of a gene with a change in the level of the subsequent protein have found few examples (De Conti et al., 2015; Štalekar et al., 2015).

In the cytoplasm TDP-43 binds a set of genes at the 3´ UTR (Colombrita et al., 2012) and can form cytoplasmic RNA granules which in neurons are transported along axons (Alami et al., 2013; Fallini et al., 2012). This gives TDP-43 a putative role in local translation. TDP-43 has been shown to either repress or enhance the translation of a small number of target transcripts (Majumder et al., 2012, 2016; Neelagandan et al., 2018). In addition, many of the proteins identified to interact with TDP-43 are involved in translation (Freibaum et al., 2010). TDP-43 controls its own expression by autoregulation. This is achieved by TDP-43 binding the 3´ UTR of the *TARDBP* transcript (Ayala et al., 2011; Koyama et al., 2016). Under conditions of cellular stress, translation is temporarily paused when actively translated mRNAs and RNA-binding proteins assemble into processing bodies and stress granules (Anderson and Kedersha, 2008). This assembly relies on protein-protein interactions between low-complexity domains of RNA-binding proteins. TDP-43 has been observed to interact with stress granules (Colombrita et al., 2009) and ALS-linked mutations in the TDP-43 low-complexity domain can increase the formation of stress granules in response to a cellular stressor (Liu-Yesucevitz et al., 2010). These granules can then rapidly disassemble once stress is stopped.

Roles in disease

Under normal physiological conditions in both neurons and glia, TDP-43 shuttles continually between the nucleus and cytoplasm (Ayala et al., 2008). However, in post-mortem brain tissue in ALS/FTD, TDP-43 is observed to leave the nucleus, often completely, and form protein aggregates in the cytoplasm (Neumann et al., 2006). Efforts in understanding TDP-43 in disease have looked at both loss and gain- of function, as well as the relevance of TDP-43 mutations to pathogenesis. It is not yet clear how the localisation of TDP-43 changes in the transition between normal physiology to the extreme state of end-stage disease.

The TDP-43 aggregates found in post-mortem disease brains contain fragments of the TDP-43 C-terminal as well as full length TDP-43 that are phosphorylated and ubiquitinated (Neumann et al., 2006; Arai et al., 2006; Bosque et al., 2013; Hasegawa et al., 2008). These

aggregates are toxic to neurons (Zhang et al., 2009) and also contain a number of other RNA-binding proteins (Dammer et al., 2012). TDP-43 can spontaneously aggregate with itself, and mutations in the low complexity domain increase the propensity to do so (Johnson et al., 2009). TDP-43 aggregates are types of amyloid (Fang et al., 2014), specific structures that are toxic to neurons and found in multiple neurodegenerative diseases. The relationship between the dynamic role of TDP-43 in RNA granules, including stress granules, and the formation of toxic TDP-43 aggregates is not yet fully understood.

Global loss of TDP-43 is lethal in the mouse embryo (Kraemer et al., 2010) and postnatal deletion leads to rapid death (Chiang et al., 2010). Conditional knockout of TDP-43 in mouse postnatal motor neurons causes a gradual degeneration of affected neurons and atrophy of muscle (Iguchi et al., 2013).

Overexpression of wildtype human TDP-43 in mice caused motor neuron degeneration in the spinal cord and severe motor impairment in proportion to the dose of the transgene (Wils et al., 2010; Shan et al., 2010). However toxicity caused by TDP-43 overexpression occurs in mice without the observation of TDP-43 aggregates (Wegorzewska et al., 2009; Barmada et al., 2010). Conversely, wildtype human TDP-43 was not found to be toxic when expressed at a physiological level (Arnold et al., 2013) and only the ALS-causing mutant forms were found to cause motor neuron degeneration. Again, this occurred without observed TDP-43 loss in the nucleus nor cytoplasmic aggregation. This works suggests that TDP-43 toxicity to neurons does not require TDP-43 aggregates to form. Clearly neurons are very sensitive to the expression level of TDP-43 protein, and this is compounded by autoregulation. Therefore, any changes in TDP-43 protein levels through knockdown or over-expression will interfere with this feedback loop, making it hard to gauge the true expression change of a particular targeting strategy.

FUS

Fused in sarcoma is a ubiquitously expressed RNA-binding protein encoded by the *FUS* gene. The FUS protein consists of a long low-complexity domain, an RNA-recognition motif, two arginine-glycine-glycine domains, a zinc finger domain and a N-terminal nuclear localisation signal (Fig. 1.2). FUS is a member of the FET family of RNA binding proteins, sharing high sequence homology with EWSR1 and TAF15 (Kovar, 2011). Mutations in *EWSR1* and *TAF15* have both been found in a small number of ALS patients (Neumann et al., 2011; Couthouis et al., 2011; Ticozzi et al., 2011; Couthouis et al., 2012). Over 40 mutations in FUS have been found to cause ALS, accounting for around 5% of familial cases and 1% of sporadic cases (Vance et al., 2009; Kwiatkowski et al., 2009). Mutations cluster in the low complexity domain and the nuclear localisation signal. FUS-ALS is distinguished from sporadic ALS by its aggressive early onset and the presence of FUS protein in cytoplasmic aggregates and not TDP-43. In FTD, only a small number of cases have found to have FUS mutations (Van Langenhove et al., 2010; Broustal et al., 2010). However, FUS-positive inclusions are seen in around 10% of FTD cases in the absence of any FUS mutation (Neumann et al., 2009). Unlike in ALS, FUS aggregates seen in FTD also

contain EWSR1 and TAF15, as well as the nuclear import protein Transportin (Neumann et al., 2011, 2012). Additionally, FUS has also been detected in aggregates from cell and mouse models of Huntington's disease (Doi et al., 2008; Kino et al., 2016).

Roles in RNA regulation

Although predominantly localised to the nucleus, FUS appears have a role in every step of RNA processing in both the nucleus and cytoplasm. As a splicing factor it binds to GGU motifs within introns and 3´ UTR sequences to either enhance or repress exon inclusion and polyadenylation (Rogelj et al., 2012; Lagier-Tourenne et al., 2012). As with TDP-43, FUS preferentially binds within introns and 3´ UTRs (Lagier-Tourenne et al., 2012; Rogelj et al., 2012; Ishigaki et al., 2012). However the overlap between FUS and TDP-43 RNA targets is small (Lagier-Tourenne et al., 2012; Rogelj et al., 2012; Colombrita et al., 2012; Honda et al., 2014). A small number of genes were found to have FUS binding antisense to the promoter (Ishigaki et al., 2012). In genes with long (>100kb) introns, FUS binding has a sawtooth pattern which appears to decline over the length of the intron (Rogelj et al., 2012; Lagier-Tourenne et al., 2012). This pattern may be due to the interaction between FUS and RNA polymerase II (Schwartz et al., 2012) and not reflect a particular binding specificity of FUS as such a pattern was also observed for TDP-43 and U2AF65 (Rogelj et al., 2012). FUS depletion leads to downregulation of long intron genes (Lagier-Tourenne et al., 2012), suggesting FUS plays a role in stabilising the splicing of particularly long introns. FUS also has a role in polyadenylation through its interaction with RNA polymerase II (Schwartz et al., 2012) as it can stall transcription at 3´ UTRs to encourage premature polyadenylation (Masuda et al., 2015). FUS has been observed to modulate 3´ end processing of RNA, as observed in the GluA1 AMPA receptor subunit (Udagawa et al., 2015). FUS knockdown has also been shown to alter levels of intron retention, particularly in splicing factors (van Blitterswijk et al., 2013; Nakaya et al., 2013).

In transcription, FUS interacts with RNA polymerase II, which could modulate transcription elongation speed (Schwartz et al., 2012). FUS interacts with both the major spliceosome via the U1 snRNP (Sun et al., 2015; Yu et al., 2015) and the minor spliceosome through binding to the U11 snRNP (Reber et al., 2016), both of which define the 5´ splice site at major or minor introns. Within splicing factor networks FUS also interacts with TDP-43, MATRN3, hnRNP A1, PTBP1 and other SR splicing factors (Lagier-Tourenne et al., 2012; Yamaguchi and Takanashi, 2016; Yang et al., 1998; Meissner et al., 2003; Kamelgarn et al., 2016). Some of these interactions are RNA-dependent (Kamelgarn et al., 2016). The role of FUS in splicing is therefore highly complex and nuanced.

Beyond mRNA, FUS facilitates the creation of both microRNA (Morlando et al., 2012) and circular RNA (Errichelli et al., 2017), two RNA species with complex regulatory functions. In the cytoplasm FUS has also been observed in RNA transport granules (Kanai et al., 2004; Fujii and Takumi, 2005). FUS is also present in cytoplasmic SMN complexes which create the spliceosomal snRNP complexes (Yamazaki et al., 2012; Groen et al., 2013). As with TDP-43, FUS aggregation has been linked to stress granule formation. FUS itself is recruited

into stress granules (Andersson et al., 2008; Yasuda et al., 2013). Rare mutations in FUS increase cytoplasmic FUS and increase stress granule recruitment (Dormann et al., 2010; Bosco et al., 2010). This requires RNA-binding activity of FUS to occur (Daigle et al., 2013). FUS aggregates found in ALS/FTD patients also contain stress granule proteins (Dormann et al., 2010) suggesting a link between stress granule formation and disease. Like TDP-43, FUS can control its own translation by autoregulation. This is thought to occur by FUS binding nearby one its exons, creating an NMD-sensitive exon skipping isoform (Zhou et al., 2013).

Roles in disease

Although causative ALS mutations have been found throughout the FUS protein, mutations associated with the lowest age of onset and shortest disease course are clustered in the proline-tyrosine nuclear localisation signal (NLS). Nuclear import of FUS is controlled by the nuclear import receptor protein transportin binding to the NLS (Dormann et al., 2010). The most aggressive FUS mutations either mutate the key proline residue in the terminal PY motif (P252L; Chiò et al. 2009) or remove the NLS entirely through a frameshift (G466VfsX14; DeJesus-Hernandez et al. 2010) or a stop codon (R495X; Bosco et al. 2010). Patients with these NLS ablating mutations tend to die in their early 20s whereas patients with NLS mutations further from the PY sequence have disease onsets resembling sporadic ALS (Shang and Huang, 2016). The fact that NLS removal causes such an aggressive form of the disease suggests that mislocalisation of FUS in the cytoplasm is the key pathogenic event. However, whether this is due to a loss of nuclear FUS or additional toxicity resulting from increased cytoplasmic FUS is still unclear. Like TDP-43, FUS protein can spontaneously aggregate *in vitro* through its low complexity domain (Murray et al., 2017). NLS mutations do not alter the propensity of FUS to self-aggregate (Sun et al., 2011). Instead NLS mutations mislocalise FUS to the cytoplasm, which may encourage aggregation. The interaction between transportin and the FUS NLS has recently been recognised to promote the dissolution of FUS aggregates (Guo et al., 2018; Yoshizawa et al., 2018). Reduced nuclear FUS would also impair autoregulation, increasing translation of FUS protein. This would then increase cytoplasmic FUS. In the mouse, complete knockout of endogenous FUS is lethal in an inbred C57BL/6 J background (Hicks et al., 2000; Kuroda et al., 2000) but survive until adulthood on a mixed background with no apparent motor deficits at 90 weeks of age (Kino et al., 2015). Overexpression studies have found a dosage dependence for the expression of human FUS to cause disease symptoms (Verbeeck et al., 2012; Mitchell et al., 2013; Shiihashi et al., 2016), with neurodegeneration and death only seen when human mutant FUS was highly expressed. Overexpression of NLS mutant FUS does not appear to alter the splicing of FUS RNA targets (Shiihashi et al., 2016), despite mutant FUS interacting with the endogenous wildtype FUS (Qiu et al., 2014). Other studies suggest a gain of function mediated by mutant FUS being responsible for neurodegeneration. Post-natal knockout of endogenous FUS in motor neurons did not lead to motor dysfunction (Sharma et al., 2016). A direct comparison between FUS knockout and mutation demonstrated embryonic lethality in both conditions, but with motor neuron loss only seen in the mutant

mice (Scekic-Zahirovic et al., 2016). Until recently, there have been no mouse models of FUS ALS that study mutant FUS at a physiological expression level.

1.5 Aims of my PhD

Understanding the role of TDP-43 and FUS in disease requires studying their roles in neuronal cells. The aim of my PhD is to capitalise on the RNA-seq revolution and bring together disparate datasets to assess the nature of RNA regulation by TDP-43 and FUS. I then apply that insight to novel animal models of disease generated by the UCL Institute of Neurology.

1.6 Overview of chapters

Cryptic splicing occurs in published TDP-43 but not FUS depletion data

TDP-43 has been observed to repress non-conserved intronic sequence from recognition by the spliceosome (Ling et al., 2015). I combined multiple public datasets on TDP-43 or FUS depletion and developed a method to discover and quantify cryptic splicing. I expanded the number of cryptic exons found to be repressed by TDP-43. Conversely I did not find evidence of cryptic exon splicing in FUS knockdown in human and mouse cells.

FUS mutant mice show progressive changes in mitochondrial and ribosomal transcripts

Most attempts to model FUS ALS in mice either knock out endogenous FUS or over-express human mutant forms of FUS. Any study of RNA processing and/or motor neuron toxicity in these models is therefore confounded. Dr Anny Devoy developed a new mouse model where an aggressive ALS mutation was knocked in to endogenous mouse FUS. This allows for the study of mutant FUS when expressed at a physiological level. I analysed RNA-seq data collected from two tissues and time points and observed a progressive change in mitochondrial and ribosomal transcripts restricted to the spinal cord.

Loss and gain of TDP-43 splicing function in two mutant mouse lines

ALS mutations in TDP-43 cluster in the C-terminal low-complexity region of the protein. I studied RNA-seq datasets from two mutant mouse lines generated by random mutagenesis. One line had a mutation in the RNA-recognition motif which reduced the RNA-binding ability of TDP-43. This was accompanied by cryptic exons, a sign of a loss of splicing function. The second line had a mutation in the low-complexity domain. I observed widespread skipping of normally constitutive exons, the inverse phenomenon to cryptic exon repression.

ALS-causative FUS mutations impair FUS autoregulation through intron retention

The most aggressive forms of ALS arise from mutations in the nuclear localisation signal of FUS. I combined 3 embryonic mouse datasets where FUS was either knocked out or the nuclear localisation signal was removed. This allowed me to jointly model the effects of the two different conditions, increasing detection power and confidence. I observed substantial overlap in both gene expression and splicing, suggesting that the FUS mutations primarily reduce the nuclear function of FUS. I identified a novel mechanism by which FUS could regulate its own translation.

2 Methods

In this chapter I discuss computational methods that used throughout my PhD. My project has consisted of processing large amounts of RNA-seq data and has required a stable workflow that can be run in parallel across multiple computers. I describe the creation of an RNA-seq library and then each step of the RNA-seq analysis pipeline in detail. I then discuss tools for analysing differential gene expression and splicing between conditions. These provide lists of genes and splicing events but do not by themselves provide biological insight. It is therefore important to integrate these results with other sources of data. I describe methods for using RNA-protein interaction data, motif searching, sequence conservation and gene ontology.

2.1 Library preparation and sequencing

RNA-seq involves selecting a pool of RNA molecules and converting them into a sequencing library of complementary DNA (cDNA) fragments. A sequencing machine then samples fragments from this library and convert the sequence of each fragment into digital information. It is important to fully understand the steps taken in the preparation of an RNA-seq library when analysing the resulting data. Which RNA species are selected, how they are converted and how they are sequenced are all important decisions that will affect the biological questions that can be asked of the data and how the downstream analysis can be carried out.

Considerations for the design of RNA-seq experiments

RNA sequencing libraries are created from either total cellular RNA (total RNA) or from polyadenylated RNA only. Total RNA libraries must be depleted of ribosomal RNA species, which consist of 80-90% of all RNA in the cell (Wilhelm and Landry, 2009). Ribosomal depletion is performed using biotinylated probes that hybridise to ribosomal transcripts. The most popular kit for this is the Ribo-Zero method (Illumina). In contrast, polyA+ libraries are created by RNA extraction with poly(Thymine) oligomers which hybridize to polyadenylated RNA. This is sometimes referred to as mRNA-seq, but as multiple non-coding RNAs are polyadenylated this is in fact a misnomer.

Choosing between the two library preparation methods depends on both the biological question of interest but the quality of RNA to be sequenced. polyA+ RNA-seq has a coverage bias towards the 3´ end of transcripts and this is exacerbated when RNA is highly fragmented. Total RNA in contrast has a much more even coverage throughout the body of each transcript. For this reason, total RNA libraries are preferred when working with

post-mortem tissue samples due to the high rates of RNA degradation from freeze-thaw cycles in frozen tissue and the effect of formalin on fixed tissue.

When quantifying gene expression, total RNA and polyA+ sequencing libraries are highly concordant (Cui et al., 2010; Zhao et al., 2018). The effect of 3´ bias can be reduced by focusing analysis on the 3´ end of transcripts where coverage will be highest. When investigating splicing, total RNA libraries will contain higher proportions of intronic reads, originating from unspliced nascent RNA (Ameur et al., 2011). This can confound analysis. However, total RNA has the benefit of giving information on the expression of a wide range of non-coding RNA transcripts. If the user is particularly interested in small RNAs such as microRNAs and small nuceolar RNAs, size fractionation is carried out before sequencing to enrich the library for these species.

For the sequencing itself, there are 4 key considerations: whether single- or paired-end sequencing will be done, whether the library will be stranded,the number of reads sequenced per sample and the length of the reads. Single- and paired-end sequencing refers to whether one end of the cDNA fragment or both ends will be sequenced. Stranded libraries retain the direction of transcription for each RNA fragment. This allows for the discovery of antisense transcripts, such as promoter antisense and 3´ antisense transcripts, which are ubiquitious in eukaryotes but poorly understood (Lavorgna et al., 2004) and improves quantification of transcripts from genes that overlap from opposing ends.

These four parameters affect the types of analyses that can be performed on the sequencing data. They also substantially influence the cost of the experiment. The original 25 base pair unstranded single-end sequencing libraries produced when the technology was in its infancy (Mortazavi et al., 2008) are far cheaper than 300bp stranded paired-end, the current maximum with Illumina technology.

The most obvious measurable outcome is alignment rate: the proportion of sequencing reads that uniquely align to the genome of interest. Longer reads increase the unique alignment rate, as do paired-end libraries due to the added constraint on alignment given by both reads in the pair aligning close together. For detecting lowly expressed transcripts, sequencing depth is crucial. For performing the current state-of-the-art splicing analysis, the most important metric is the number of read that span splice junctions. This is less important for the previous generation of software which use exon coverage instead. With longer reads, the chances of any given read overlapping an intron increase. Therefore long ($\geq$ 100bp) reads are highly recommended. A study comparing RNA-seq samples at different simulated read lengths recommended 50bp single-end acceptable for differential gene expression and at least 100bp paired-end for splicing (Chhangawala et al., 2015).

Sequencing library preparation

The description below is of a polyA+ paired-end and stranded RNA-seq library preparation for Illumina sequencing. Total RNA is extracted from cells or tissues using Trizol (Thermo Fisher) or a similar phenol/chloroform reagent (Chomczynski and Sacchi, 1987).

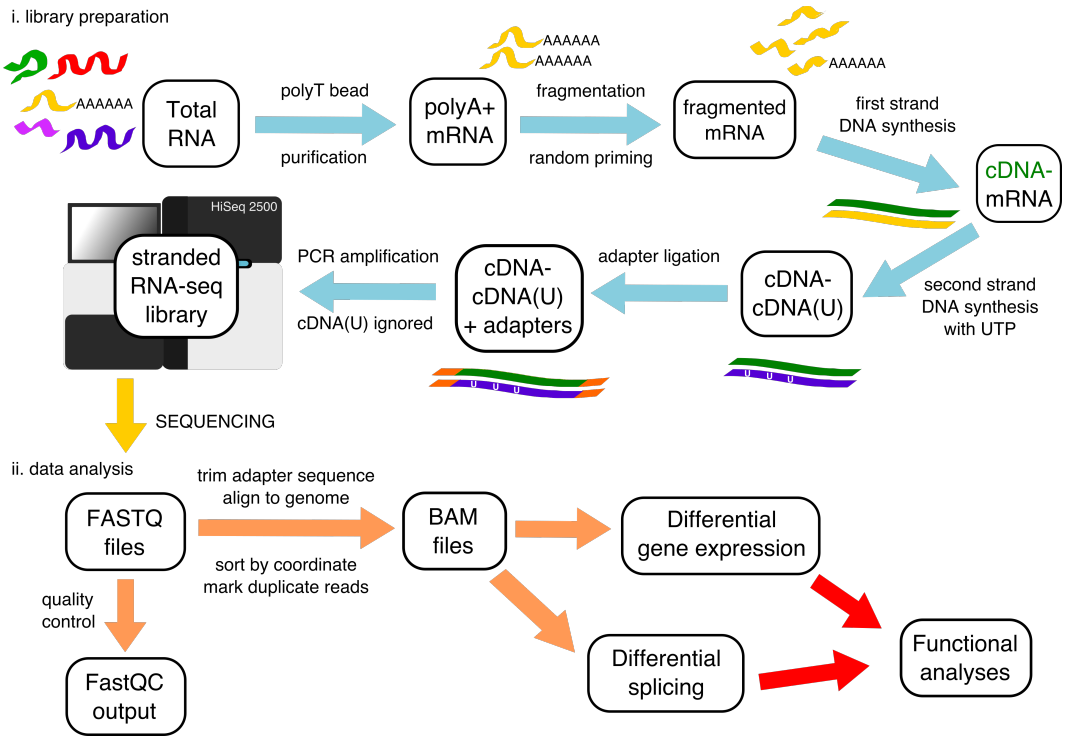


Figure 2.1: Pipeline for stranded RNA-seq preparation and analysis
i) Library preparation from total RNA extraction and strand-specific amplification
ii) Bioinformatic analysis workflow to align reads and test for differential gene expression and exon usage

To enrich for mRNA species the RNA is mixed with magnetic beads that are bound to poly(Thymine) oligomers complementary to the poly(Adenine) tail at the 3´ end of mRNA. The RNA is then fragmented. As paired-end sequencing extracts information from each end of the fragment it is important to consider the fragment size in light of the subsequent length of the reads, as with short fragments and long reads the read pairs will overlap leading to redundant information. Pseudo-random barcodes are ligated to each fragment to allow for reverse transcription. This pseudo-randomness does cause bias towards particular sequences (Van Gurp et al., 2013). The first round of reverse-transcription is carried out forming double stranded DNA-RNA hybrids. These complexes are melted and a second reverse-transcription reaction is carried out in the presence of uracil instead of thymine, creating double stranded DNA. The uracilated DNA strand is now in the same orientation as the original mRNA fragment. Polymerase chain reaction amplification is then carried out where the uracil containing DNA and the RNA are ignored by the polymerase. The resulting library is strand-specific as all the DNA fragments are antisense to the original RNA fragments. Further adapters are added for sequencing in the flowcell and to give the fragments from each sample an identifier as multiple samples are usually pooled together. The standard Illumina paired-end sequencing reaction ligates the amplified fragments to wells in the flowcell which then form clusters of identical amplified fragments. Sequencing then occurs by sequential addition of fluorescent nucleotides along the length of a fragment in both directions, the colour of which is recorded by a high resolution sensor. This is "sequencing by synthesis", the Solexa/Illumina method (Bentley et al., 2008). The identity of each nucleotide is determined and given a quality score based on the confidence of the

measurement. The reads are converted from the basecall (BCL) format to a FASTQ file by the sequencing centre using *bcl2fastq* (Illumina). The FASTQ file specification (Cock et al., 2009) encodes the unique read ID, the read sequence and a quality score. In paired-end sequencing, reads of each pair share a read ID . Forward and reverse reads are written into separate files. The FASTQ file can be thought of as a digital encoding of the sequences from both ends of each cDNA fragment.

2.2 The Plagnol lab RNA-seq pipeline

A pipeline connects together multiple software tools to convert raw data into summaries and statistics. Our RNA-seq pipeline converts raw sequencing reads into counts of genes, exons and splice junctions for use in downstream analyses. This was collaboratively developed by the group of Vincent Plagnol but adapted and occasionally broken by myself. The pipeline itself is written in the Bash programming language `http://www.gnu.org/software/bash/`. Individual modules for differential expression and splicing are written in the R statistical programming language (Gentleman and Ihaka, 1996). Code for the pipeline is freely available to all from (`github.com/plagnollab/RNASeq_pipeline/`).

The pipeline has been optimised for the UCL Computer Science cluster. The cluster is made up of hundreds of individual computers or nodes, controlled from a head computer using the Sun Grid Engine system (Univa) for organising and distributing computational work. An individual set of instructions to run on a single node is called a job. This allows for individual steps of the pipeline to be run in parallel or series by distributing steps across multiple jobs. Jobs can then be linked so that the next step commences once all jobs have completed. This makes running the pipeline from start to finish as fast and efficient as possible. However, I have encountered numerous problems while doing my research due to instability and breakages of individual nodes in the cluster. This can be mitigated by the use of cluster-aware pipeline frameworks such as SnakeMake or NextFlow (Köster and Rahmann, 2012; Di Tommaso et al., 2017). These frameworks can restart or reassign jobs when a particular node fails, and increase hardware requirements for nodes when steps repeatedly fail. I hope that the next version of the pipeline will be written using one of these frameworks.

Quality control and read alignment

The first step in any analysis is quality control of the FASTQ files. Popular tools like *FastQC* (`bioinformatics.babraham.ac.uk/projects/fastqc/`) analyse a set of FASTQ files and produce visualisations of multiple diagnostic tests. It can be useful to observe the range of quality scores and how they alter throughout the length of a read to diagnose faults during the sequencing reaction. Another important diagnostic is the presence of adapter sequences within reads. This can occur when the fragmentation step is too aggressive or when the original RNA sample is heavily degraded, often the case in human post-mortem tissue samples. With short fragments, the sequential addition of nucleotides runs into the

sequencing adapter sequence, making these reads difficult to align. These universal adapter sequences can be removed by software such as *Trim Galore!* (`bioinformatics.babraham.ac.uk/projects/trim_galore/`). This can also remove low quality sequence from the ends of reads, which often occurs towards the end of an Illumina sequencing run.

Following trimming, the FASTQ files must then be aligned to the genome of the species of interest. There have been great advancements in speed and accuracy in alignment algorithms for DNA and RNA. The key difference between DNA and RNA alignment is the need for read splitting for RNA. As most RNAs are spliced, any cDNA fragment that originates from the boundary between two exons will need to be split and both pieces separately aligned to the genome. The interval between two pieces is then recorded by the aligner as a splice junction, the demarcation of where an intron was excised. The current state-of-the-art algorithm in both speed and accuracy at resolving splice junctions is *STAR* (Dobin et al., 2013). *STAR* derives its speed from loading the entire genome into memory and then aligning millions of reads per hour using a seed-and-extend algorithm, where small pieces of each read are aligned and incrementally extended to find the best possible split alignment. The alignment information for each read is recorded in a BAM file (Li et al., 2009). This format encodes the original read sequence and quality score along information on the alignment, including where the alignment is unique and whether the read was split or clipped by the aligner. Reads from each pair are initially recorded next to each other and the BAM file is ordered by the read name. However for most downstream analyses the BAM file must be re-ordered by genomic position. The pipeline does this using the *Novosort* algorithm (`http://www.novocraft.com/products/novosort/`).

If 3´ bias is present in a polyA+ RNA-seq library, this can be observed by computing read coverage across all genes with a diagnostic package such as *QoRTS* or RNASeQC (Hartley and Mullikin, 2015b; Deluca et al., 2012). As most differential expression software assumes even coverage throughout a gene body, heavily biased samples can skew the estimates of gene or transcript expression.

For downstream expression analysis, the aligned reads are then quantified for a set of features. These features can either can be whole genes or individual exons. There are multiple annotations for the human and mouse genome for known genes and exons are and our pipeline uses the Ensembl transcript annotation as our reference (Cunningham et al., 2015). Uniquely aligning reads that overlap each gene or individual exon are counted using *HTSeq* (Anders et al., 2015).

Differential gene expression

The most common application of RNA-seq is for an experiment comparing the abundances of different RNAs between conditions. This could be for example between the knockdown of a particular gene and a control, or between a group of disease patients and a group of healthy controls. These experiments should be made up of multiple biological replicates, where RNA libraries have been prepared from different organisms or cell culture samples under the same conditions. This is so a fair assessment can be made of the biological variation

in RNA abundance within each condition. This is in comparison to technical replicates, where the same library is sequenced multiple times, which can only explain the variance in the sequencer itself. As the concordance between technical RNA-seq replicates is very high (Mortazavi et al., 2008), these are generally shunned in favour of biological replicates. RNA-seq is still an expensive experiment and the high cost limits a lab to sequence only a small number of biological replicates to typically less than 5 samples per condition. There is also an ethical consideration when working with model organisms on how many samples should be appropriate. Small sample sizes limit statistical power to detect small variations in RNA abundances and the confidence in the truth of one's results. There are multiple algorithms to test for differential expression but we settled on using the *DESeq2* package for its statistical robustness and speed (Love et al., 2014). It is desiged to compensate for the small sample sizes used in a typical RNA-seq experiment.

DESeq2 makes use of the fact that each sequencing library measures the abundance of tens of thousands of transcripts. The number of reads generated by each library can be highly variable and so this is also accounted for. *DESeq2* normalises the read counts for each gene in each sample by the size of each sample's library, the *size factor*. It then assumes that the normalised read counts fit a negative binomial probability distribution. It estimates the variance or dispersion in read count for each gene across all samples. To compensate for small sample sizes, which give a very high estimate of dispersion, it compares the dispersion between all genes and shrinks each estimate to a local average based on genes expressed at a similar level. Using these shrunken dispersion estimates, the software then fits two generalised linear models: a null model where condition has no effect and an alternative model where the change in condition explains the change in gene expression. The two models are compared with a Wald test on which model fits the data better, computing a P-value. The P-values generated for each gene are adjusted to correct for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The output of a differential expression analysis gives both an adjusted P value for each gene and a fitted log_2 of the fold change in expression between the two conditions. This gives an estimate of the effect size. Fold-changes on the log_2 scale have the useful aesthetic property of being symmetrical around positive and negative powers of two. Therefore a doubling of expression will have a log_2 fold change of 1, and a halving will have a log_2 fold change of -1.

2.3 Differential splicing

Differential splicing is an analysis that looks for changes in splicing between conditions. How this is done depends on the frame of reference used. I will discuss examples of different approaches.

Generally, methods for differential splicing have followed the increasing sophistication and reducing cost of sequencing technology. Increases in read length and and greater depth of sequencing due to plummeting sequencing costs have driven the development of new software. These have made use of the increasing availability of splice junction reads within samples.

I can group software packages into one of four categories based on what they consider to be the fundamental unit of differential splicing. All these modern packages work best with high depth paired-end data. The approaches based on junction reads also require long reads to maximise the information available.

Exon quantifiers

These packages focus on the change in usage of particular exons between conditions. Therefore the fundamental unit of analysis is the exon. They test for differences in the usage of a particular exon as a ratio of read counts for each exon against all exons. This adjusts for differences in baseline gene expression between conditions. This approach was first used in the *DEXSeq* package (Anders et al., 2012). *DEXSeq* requires a flattened list of "union exons" which is a set of non-overlapping intervals formed from collapsing exons from overlapping transcripts together. RNA-seq seq reads can then be counted at each union exon. This method does not require splice junctions so lower depth and shorter read libraries can be used. *DEXSeq* normalises the per-exon counts to a library *size factor* and compares the normalised counts from each exon with the counts from all the exons in a gene. It then fits a generalised linear model to compare the ratio of exon usage between biological conditions and estimates a fold-change, which is shrunken with a Bayesian shrinkage procedure. The output of a *DEXSeq* analysis is each union exon in a gene is given a log_2 fold change between conditions and an adjusted P value from the test. Although this can be used to demonstrate whether differential exon usage occurs due to a condition, it is inherently difficult to extract meaningful information about the underlying biology. A significant union exon may not correspond to real exon. As they consider the whole gene they are sensitive to poorly annotated genes and extreme biases in coverage. Therefore, hits from *DEXSeq* need to be closely analysed in the context of the gene they reside in. This approach is also dependent on transcript annotation so cannot pick up novel events. *DEXSeq* has been adapted to look at differential intron usage (Li et al., 2015). This work directly inspired my work on *Cryptex*, a software pipeline for using splice junction reads to find novel exons which I discuss in chapter 3. The intepretability problem has been reduced by a software package called *JunctionSeq* (Hartley and Mullikin, 2015a), which also quantifies splice junctions and can pick up novel junctions, but uses a *DEXSeq* framework for differential usage.

Transcript quantifiers

Here, the fundamental unit is the transcript. Genes can have many different transcripts which may be highly similar in exon content. These tools estimate the abundance of different transcripts and compare abundances between conditions. Transcript assembly algorithms like *RSEM* (Li and Dewey, 2011), *Cufflinks* (Trapnell et al., 2010) and *Stringtie* (Pertea et al., 2015) assemble aligned reads into transcripts using annotation but also find novel transcripts. Pseudoaligners like *Kallisto* (Bray et al., 2016) and *Salmon* (Patro et al., 2017) can work straight from the FASTQ files without prior alignment. All these algorithms require long reads for enough splicing information and very dependent on the list of tran-

scripts they are given. Some genes can generate hundreds of very similar transcript isoforms or overlap with other genes and so these methods will struggle to accurately assign reads correctly. (Zhang et al., 2017) compared isoform estimation between pseudoalignment tools to alignment-dependent tools (*RSEM*, *Cufflinks*) and found the alignment free methods outpeformed in both speed and accuracy.

Local splicing event quantifiers

A compromise between the two previously mentioned, these methods take the splicing event as fundamental unit and compare changes in inclusion of a splicing event between conditions. This relies on the presence of reads that span splice junctions and so require high depth and long reads. The power of the splice junction rests in its ability to exactly demarcate where an intron has been excised from a transcript. The coordinates and counts of each splice junction can then be used to construct and quantify different types of splicing events. Using this information in analysis provides a much more accurate picture of splicing beyond exon quantifiers which only assess changes in read coverage across exons. The trade-off is that as splice junction reads are the minority of all reads there is reduced power to detect differential splicing events compared to exon quantification methods.

Typically, splicing events are discovered and classified by constructing a graph database from splice junction reads. This models splicing events as a set of overlapping nodes and edges. The topology of the graph then determines the type of splicing event. These events may be cassette exons, alternate splice sites, intron retention, mutually exclusive exons or alternate starts and ends. This classification greatly aids downstream interpretation as different categories of event may be regulated by different mechanisms. Each event is then tested for differences in inclusion/exclusion across conditions. This approach has been implemented in the *SGSeq*, *SUPPA2*, *rMATs*, *JUM*, *MAJIQ* and *Whippet* packages (Goldstein et al., 2016; Trincado et al., 2018; Vaquero-Garcia et al., 2016; Wang and Rio, 2018; Shen et al., 2014; Sterne-Weiler et al., 2018). SGSeq has the unique ability to find novel events not present in annotation. It discovers and classifies events and then provides counts reads supporting inclusion and exclusion of each event. These counts are then fed into the *DEXSeq* framework and a general linear model fitted to test the effect of condition. *Whippet* combines features of the pseudoaligners with the local event quantifiers. It first builds an index of splicing events from transcript annotation and then aligns reads to the index straight from the FASTQ file. This approach cannot detect novel splicing junctions but can discover novel arrangements of known exons. âĂIJAll these packages, while providing a clear output of different splicing events, cannot detect other mRNA processing events that only be distinguished by read coverage and not junctions, such as the alternate polyadenylation between two tandem 3´ UTRs. Each of the local splicing quantifiers analyse splicing with different biological and statistical assumptions so it is difficult to directly compare the software on a set of simulated events. Attempts at comparisons have been driven by software authors themselves to demonstrate superiority of one tool over another but have shown that the tools largely agree with each other on the majority of events (Trincado et al., 2018; Vaquero-Garcia et al., 2018).

Annotation-free quantifiers

The fourth and final class of algorithm don't depend on any transcript annotation at all. The *derFinder* package (Collado-Torres et al., 2017) assesses changes in read coverage between samples, which is particularly useful in looking for novel non-coding RNA species or in the aforementioned tandem 3´ UTR problem. The *LeafCutter* package (Li et al., 2018) looks for differentially used splice junctions independently of annotation and has shown that across a set of tissues around 30% of the high confidence splice junctions recovered are novel to any annotation database.

Deciding on which tools for use for each project

One of the analytical priorities for my work on TDP-43 and FUS has been to quantify novel splicing changes. These are splicing changes that are not recorded in annotation databases. In chapter 3 I developed my own tool to look at cryptic splicing without requiring long reads, as I reanalysed older RNA-seq data. The idea behind *CryptEx* was that with shorter reads there are very few splice junctions in an individual sample but by first combining all samples together, novel exons can be discovered and distinguished from random intronic read coverage. Once those exons are then determined, the total read coverage over that position can be used to test for differential usage using *DEXSeq*. Since that project was published there has been a movement in the field towards tools that can accurately quantify novel splicing but the requirement for read length is still present. This is now less of a problem as read lengths of 100bp and greater are now routine.

For the work on TDP-43 mutant mice described in chapter 5, I used different splicing methods to match the quality of the data, which was generated over 5 years with different generations of library preparation and sequencing technology. For the oldest data, generated from mouse embryonic fibroblasts and head samples with low read depth and short length, I used the *DEXSeq* package to estimate differential exon usage and *CryptEx* to find novel exons (not included in chapter). With the newer data from mouse embryonic spinal cord, sequenced with longer paired-end reads and at higher depth, I wanted to use a local splicing event quantifier to take advantage of the high proportion of splice junction spanning reads. I chose the *SGSeq* package as it could quantify splicing events from BAM files and used a graph database approach to find novel splicing events as well as classify annotated splicing events. This provided a much richer dataset to work with than the previous *DEXSeq* method. For my work on combining FUS datasets from multiple groups I again used *SGSeq* for splicing analysis as it uses *DEXSeq* for statistical testing of splicing event usage between conditions. The general linear model framework provided by *DEXSeq* allowed me to add dataset-specific covariates and analyse all samples jointly.

2.4 Functional analyses

Downstream of differential expression and splicing analysis, I have used other sources of data to find and understand potential causes and mechanisms for the changes I detect.

RNA-protein interaction data

RNA-binding proteins bind to RNA sequence or structural motifs in sets of target transcripts. This binding influences the expression and splicing of these transcripts. These binding preferences and targets can be observed experimentally using UV crosslinking and immunoprecipitation, or CLIP. This uses a specific wavelength of ultraviolet light to form crosslinks, covalent bonds between amino acids and nucleotides in close proximity (Ule et al., 2003). These RNA-protein complexes are then purified using an antibody to the protein of interest. The overhanging RNA and protein is then digested, leaving only a small polypeptide bound to a short fragment of RNA. The RNA is then reverse-transcribed and sequenced. With the advent of next-generation sequencing, this was taken transcriptome-wide with the development of high-throughput CLIP (HITS-CLIP) (Licatalosi et al., 2008). However, CLIP does not provide precise information on which nucleotides in the fragment are bound, as the method depends on the reverse transcriptase reading past the crosslink nucleotide. As the reverse transcriptase frequently stalls at the crosslinking position, this can be exploited. By capturing these truncated cDNAs, the position of the UV crosslink can be determined from the nucleotide directly upstream of where the cDNA aligns to the genome. This technique, individual nucleotide CLIP or iCLIP, allows for the precise binding sites of a protein to be discovered (König et al., 2010; Huppertz et al., 2014). Once the libraries are sequenced and the reads aligned to the genome, unique cDNAs are found by removing duplicate reads and then counted. Binding regions are determined with a shuffling procedure which clusters reads within a set distance and compares the counts of each cluster to a shuffled control within the same genomic feature (Wang et al., 2010). This produces lists of clusters and their component peaks discovered at a false discovery rate of 5%. The processing of iCLIP data used in my work was done by the iCOUNT server, maintained by Tomaž Curk and colleagues (Curk, 2016).

An alternative to iCLIP, enhanced CLIP (eCLIP), reduces the number of PCR duplicate reads in the library (Van Nostrand et al., 2016). eCLIP was used by the ENCODE consortium to profile multiple RNA-binding proteins. Data for enriched peaks and clusters is processed and freely available at `https://www.encodeproject.org/`. Unlike iCLIP, eCLIP does not include a verification step where the correct size of RNA-protein complex is checked by eye. Therefore one cannot be sure that the cDNA libraries are not from RNAs that interact with other co-purified proteins (Chakrabarti et al., 2018).

I can integrate eCLIP and iCLIP data with my RNA-seq results to correlate binding of RNA targets with changes in gene expression and splicing. RNA maps are a powerful visulisation tool first used to profile the protein NOVA (Ule et al., 2006). They plot the distribution of CLIP tags or reads across a set of genomic features. Typically these features are cassette

exons from RNA-seq that are either included or skipped more when the protein of interest is perturbed. By comparing a set of altered events to a control set of no changes, it is possible to see the scale of enrichment for binding within a particular region. This has been used to show positional specificity around splice sites for multiple RNA-binding proteins (Ule et al., 2006; Wang et al., 2010; König et al., 2010). As well as constructing RNA maps, I have used iCLIP and eCLIP data to look at CLIP binding on different types of genomic features, such as introns and 3´ untranslated regions.

CLIP-based methods do not provide an absolute measure of binding affinity between a protein and its target RNAs. The number of unique overlapping cDNA reads at a position is a combination of the binding affinity at that position and the underlying expression of the target. It remains challenging to distinguish high affinity from low affinity binding without the use of indirect methods, such as using RNA-seq to estimate the underlying expression of each target. Direct comparisons of cDNA counts between CLIP profiles of the same protein under different conditions are therefore difficult to make. In addition, data can only be generated from sites that are both expressed in the cell type tested and can be aligned to the genome. The latter precludes the use of CLIP methods on understanding binding to repetitive regions within specific genes (Chakrabarti et al., 2018). There are also inherent technical biases in UV crosslinking and RNase digestion that affect the cDNA profile. Despite these caveats, CLIP-based methods, in combination with RNA-seq, are powerful tools for improving our understanding of RNA biology.

Functional annotation with ontologies

High throughput analyses can produce large lists of genes. Understanding how these genes fall into specific functional sets and/or pathways provides important biological insight.

The Gene Ontology (GO) initiative seeks to annotate every gene with function information (Ashburner et al., 2000; Carbon et al., 2017). GO terms are split into three domains: molecular function, cellular component and biological process. Terms have defined relationships with each other in a hierarchical structure; more specific terms are the children of broader terms. The Kyoto Encyclopedia of Genes and Genomes (KEGG) annotates genes to pathways of interacting molecules as well as to sets of genes linked to particular human diseases (Ogata et al., 1999). Both databases share experimental information across orthologous genes between species. Annotations may be from direct experimental evidence or inferred through computational methods such as protein sequence and structure comparison, which make up the majority of annotations as of 2010 (du Plessis et al., 2011).

Whether a list of genes from an analysis is enriched for a particular GO or KEGG set is tested with a hypergeometric test, comparing the observed proportion of genes in the list that overlap the set with the expected number under the null. It is important that multiple testing correction is employed due to the vast number of sets tested in a single analysis. This can be performed using web-based interfaces such as *PantherDB* (Thomas et al., 2003) and gProfileR (Reimand et al., 2016). *gProfileR* is particularly useful as it simultaneously tests all three categories of GO term along with the KEGG sets. *GOseq* is a method for

gene ontology enrichment testing from differential expression results (Young et al., 2010). As longer genes will have more read coverage there is more statistical power to detect a change in expression of those genes. *GOseq* weights gene sets by the length of genes within them when computing enrichment tests. I used *GOseq* in chapter 4 but due to the enhanced ease of use and speed I chose to use *gProfiler* in chapter 6.

These methods do not incorporate cell or tissue information into the statistics. When neuronal GO or KEGG terms are enriched when studying neuronal samples it is not surprising. Ontology databases are highly biased towards already well-studied genes and diseases like cancer (Haynes et al., 2018), which reflects trends in biomedical research and funding. Results from GO enrichment tests are unstable throughout time due to constantly increasing databases (Tomczak et al., 2018). There are dangers with using gene annotation data for biological interpretation as it is relatively simple to construct a story that makes "biological sense" from a random set of genes and annotations (Pavlidis et al., 2012). For these reasons, ontology results should be interpreted with a degree of caution.

2.5 Conclusions

- Sequencing libraries should be carefully thought out in light of the biological question
- Pipelines allow efficient processing of large volumes of data but can be unstable
- Differential splicing analysis depends on both the quality of sequencing library and the frame of reference chosen
- Combining RNA-seq results with CLIP-based methods can give useful information
- Gene annotations can provide insight but should be used with care

3 Cryptic splicing occurs in published TDP-43 but not FUS depletion data

This chapter has been published as (Humphrey et al., 2017). See appendices for full reproduction of the manuscript.

3.1 Overview

Ling and colleagues observed the inclusion of cryptic exons when TDP-43 was depleted in HeLa or mouse embryonic stem cells (Ling et al., 2015). These cryptic exons were shown to originate from poorly evolutionarily conserved sequence and shared no positions between the two species. These findings raise the possibility that impaired exon recognition contributes to TDP-43's role in ALS aetiology.

I aimed to replicate and expand the findings of Ling and colleagues. I first developed a quantitative genome-wide analysis pipeline to detect and classify cryptic splicing. I then applied this pipeline to seven datasets (four human and three murine models) to systematically quantify cryptic splicing alterations associated with depletion of TDP-43. In addition I investigated FUS in order to determine whether modulation of cryptic splicing was a common feature of RNA-binding proteins implicated in ALS. Furthermore I analysed hnRNP C, as it had been previously shown to repress cryptic exons originating from Alu repeat elements. Lastly, I used independent protein-RNA interaction datasets, conservation data, repeat element annotation and splice site scoring to investigate the potential mechanisms linking TDP-43 depletion with the cryptic exon phenomenon.

3.2 Methods

Data preparation

Table 3.1 lists all the public data used in this study. All RNA-seq data was downloaded in the FASTQ format and processed using our RNA-seq pipeline (see methods chapter). Processed iCLIP peak data was downloaded from the iCOUNT server. Both the human and mouse TDP-43 iCLIP data have been previously published (Tollervey et al., 2011; Rogelj et al., 2012).

Processed eCLIP data (previously described by (Van Nostrand et al., 2016)) was downloaded from the ENCODE project. The narrowPeaks bed format was used with the first nucleotide of the cluster defined as the peak. Peak coordinates from iCLIP and eCLIP were converted to the hg38 and mm10 builds using the *LiftOver* tool from UCSC.

Table 3.1: List of accessions

Assay	Accession Code	Downloaded	Target	Cell/Tissue	Pubmed ID
RNA-seq	PRJNA282887	SRA[1]	TDP-43	Mouse ES	26250685
RNA-seq	PRJNA282692	SRA	TDP-43	Human Hela	26250685
RNA-seq	PRJNA127211	SRA	TDP-43	Mouse ES	20660762
RNA-seq	PRJNA141971	SRA	TDP-43	Mouse adult brain	21358643
RNA-seq	ENCSR129RWD	ENCODE[2]	control	K562 mRNA	-
RNA-seq	ENCSR134JRE	ENCODE	TDP-43	K562 mRNA	-
RNA-seq	ENCSR372DZW	ENCODE	control	K562 total RNA	-
RNA-seq	ENCSR455TNF	ENCODE	TDP-43	K562 total RNA	-
RNA-seq	PRJNA174534	SRA	FUS	Mouse adult brain	23023293
RNA-seq	ENCSR084SCN	ENCODE	control	K562 mRNA	-
RNA-seq	ENCSR325OOM	ENCODE	FUS	K562 mRNA	-
RNA-seq	PRJEB3048	SRA	hnRNP C	HeLa	23374342
iCLIP	20100222_LUjt3	iCOUNT[3]	TDP-43	Mouse embryonic brain 1	22934129
iCLIP	20091102_LUjt5	iCOUNT	TDP-43	Mouse embryonic brain 2	22934129
iCLIP	20100222_LUjt3	iCOUNT	TDP-43	Human neural stem cells	21358640
iCLIP	20101125_LUjt8	iCOUNT	TDP-43	Human SH-SY5Y 1	21358640
iCLIP	20091102_LUjt5	icount.biolab.si	TDP-43	Human SH-SY5Y 2	21358640
eCLIP	multiple	ENCODE	Multiple RBPs	HepG2 and K562	-

Table 3.2: All RNA-sequencing data used in this study ES: embryonic stem cell. K562: human leukaemia cell line. siRNA: small interfering RNA. shRNA: short hairpin RNA. ASO: antisense oligonucleotide. PE: paired end sequencing. SE: single end sequencing. For single end sequencing, depth is measured in millions of mapped reads whereas paired end sequencing depth is measured in millions of mapped fragments.

	Species	Cell	Protein	Depletion method	Library	Read type	Depth	Citation
1	Human	HeLa	TDP-43	siRNA	mRNA	100bp PE	97-116M	Ling, 2015
2	Mouse	ES	TDP-43	deletion	mRNA	100bp PE	70-75M	Ling, 2015
3	Human	K562	TDP-43	shRNA	RNA	100bp PE	55-62M	ENCODE
4	Human	K562	TDP-43	shRNA	mRNA	100bp PE	25-29M	ENCODE
5	Mouse	Brain	TDP-43	ASO	mRNA	75bp SE	35-60M	Polymenidou, 2011
6	Mouse	ES	TDP-43	deletion	mRNA	40bp SE	2-11M	Chiang, 2010
7	Human	K562	FUS	shRNA	mRNA	100bp SE	12-21M	ENCODE
8	Mouse	Brain	FUS	ASO	mRNA	72bp SE	20-60M	Lagier-Tourenne, 2012
9	Human	HeLa	hnRNP C	siRNA	mRNA	72bp SE	26-28M	Zarnack, 2013

Cryptic splicing definition

I define cryptic splicing as the inclusion of sequence into mRNA transcripts that is not annotated by an existing database of known exons. Splicing aware alignment software such as *STAR* (Dobin et al., 2013) cut short reads that originate from a spliced transcript and align the pieces separately, marking the distance between them as a splice junction. Splice junctions can be used to reaffirm known splicing patterns or infer novel splicing. I therefore base my computation of cryptic splicing as a relative increase in splice junctions that join known splice sites to unannotated positions within introns, with this increase correlating with the depletion of a particular RNA binding protein. Different repositories have differing levels of proof for annotating exons but I define an annotated exon as one listed in the

[1]Sequence read archive: `www.ncbi.nlm.nih.gov/sra`

[2]Encyclopedia of DNA elements: `www.encodeproject.com`

[3]iCOUNT iCLIP web server: `www.icount.biolab.si`

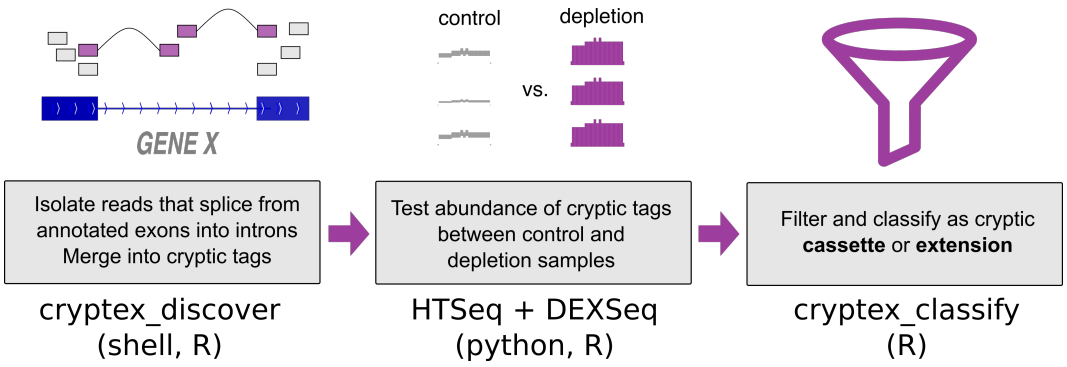


Figure 3.1: Schematic of the *Cryptex* pipeline

Ensembl list of transcripts (release 82) (Cunningham et al., 2015) which contains many more alternatively spliced exons than the RefSeq database (Pruitt et al., 2014) but is still more conservative than the GENCODE database (Harrow et al., 2012).

Cryptic splicing discovery with the *CryptEx* pipeline

RNA-seq is a constantly improving technology. As I wanted to analyse all TDP-43 depletion data created from 2010 onwards, the use of modern RNA-seq analysis software, with its requirement for long reads and high depth paired-end stranded sequencing, was sadly out of the question. The cryptic splicing discovery pipeline (*CryptEx*) I developed is designed (see Fig. 3.1) to be used on any RNA-seq library, whether single or paired end, stranded or unstranded, total RNA or polyA-selected RNA. This flexibility inevitably results in a large number of false positive hits which have to be aggressively filtered downstream. I am indebted to Yafang Li and colleagues whose paper on intron retention (Li et al., 2015) was the inspiration for my repurposing of existing bioinformatic tools to mine published data for an overlooked RNA phenotype.

I wrote the code for *CryptEx* in the Bash shell scripting language, making use of the *SAMTools* (version 1.2) (Li et al., 2009) and *BEDTools* (version 2.25.0) (Quinlan and Hall, 2010) libraries of commands for manipulating read alignments and lists of genomic features respectively. Reads were counting using HTSeq (Anders et al., 2015). The statistical testing for differential cryptic exon usage was carried out using the *DEXSeq* package (Anders et al., 2012) in the R statistical computing environment. I subsequently wrote all code for downstream processing and filtering of cryptic hits in the R language (version 3.1.1) due to the wealth of existing software packages for genomic analysis, data processing and visualisation. I made extensive use of the *Biostrings, data.table, DEXSeq, dplyr, GenomicRanges, ggplot2, gridExtra, optparse, plyr, stringr,* and *tidyr* packages. The code for reproducing the results of this chapter is available in a GitHub repository `www.github.com/jackhump/CryptEx`.

In order to discover all possible splice junctions that travel into the intron, spliced reads were extracted from each aligned bam file using *SAMTools*, discarding any secondary alignments. To extract only the spliced reads that overlap an annotated exon the lists of spliced reads were intersected in *BEDTools* with a flattened list of exons, created using the `dexseq_prepare_annotation.py` Python script included with the *DEXSeq* package (Anders et al.,

2012). An inverse intersection was then performed with the same exon list to retain only the spliced reads which do not bridge two annotated exons. The intronic mapping sections of each read were split off from the rest and retained, thus keeping a set of aligned reads that splice to, but are not part of, an annotated exon. The results of this for each sample were merged together irrespective of condition. Split reads that were within 500bp of each other were merged into larger intervals, hereby referred to as tags. This ideally captures both the upstream and downstream splice junction to a central cryptic cassette exon. To keep only the tags that are splicing within the gene body another intersection was performed with a list of introns. This was generated from the same flattened exon file by an R script written by Devon Ryan (`seqanswers.com/forums/showthread.php?t=42420`). The tags were then incorporated into the flattened list of exons, the GFF file. Each tag was given a unique identification number including a reference to the upstream annotated exon, allowing for comparisons of different datasets. The reads that overlap annotated exons and tags were counted using *HTSeq* (Anders et al., 2015) on the default settings, ignoring reads marked as PCR duplicates. The read counts were used to calculate differential usage of each exon with respect to condition with *DEXSeq*.

All the cryptic tags with an adjusted P-value (false discovery rate) $< 5\%$ and a $|\log_2(\text{fold change})| > 0.6$ were extracted from the *DEXSeq* results table. The splice junctions from the alignment of each sample were used to work out the coordinates of the canonical junction that spans the intron within which the cryptic tag is or isn't spliced in control samples. Using splice junctions from the depletion condition samples, the upstream and downstream junctions that connect the adjacent annotated exons to the cryptic tag were re-discovered and quantified. Any cryptic example that did not have at least one upstream or downstream junction per sample or had fewer than ten canonical splice junctions was removed. These junctions were used to calculate per-condition mean Percent Spliced In (PSI) values which are a ratio of cryptic splicing over the sum of cryptic and canonical splicing (Katz et al., 2010). As a number of cryptic splicing events are present at a low level in control samples, ΔPSI values were created for both upstream and downstream splicing for each tag. This is the difference in PSI between the depletion samples and the control samples. Any cryptic tag that had either an upstream or downstream $\Delta\text{PSI} < 5\%$ was removed.

iCLIP/eCLIP enrichment

I downloaded lists of iCLIP and eCLIP peaks (see Table 3.1) and used the UCSC LiftOver tool to convert the coordinates into the human build 38 and mouse build 10. The coordinates of each cryptic tag were flanked by 100 base pairs on either side to capture binding around the putative splice sites. In order to compare the overlap between cryptic exons and RNA-protein binding peaks, two sets of null exons were created for comparison, which maintain the same length as their corresponding cryptic exon but sample either the intronic sequence outside of the flanked exon or that of the adjacent introns within the same gene if available. Overlaps between exons and iCLIP and eCLIP peaks were calculated using BedTools. The proportions of overlap between the cryptic exons and the two sets of null exons were compared using a proportion test with the null hypothesis that the proportion

of exons overlapping an iCLIP or eCLIP peak would be the same.

Motif enrichment analysis

FASTA sequence was generated for the cryptic exons flanked by 100 nucleotides either side and submitted to the *MEME* web tool (Bailey et al., 2009) under the default settings. The analysis was repeated using the *HOMER* algorithm (Heinz et al., 2010) on RNA mode. Motifs were created using *WebLogo* (Crooks et al., 2004). Frequencies of the 16 possible dinucleotides were compared between flanked cryptic exon sequences with adjacent intron sequences from the same set of genes.

Transposable element enrichment

Lists of transposable elements in human and mouse (hg38 and mm10 respectively) were previously generated by the *RepeatMasker* tool (Smit et al., 2015) and were downloaded from UCSC. Overlap between different transposable elements and the cryptic exons was calculated in each orientation using *BedTools*.

Conservation analysis

PhyloP compares the sequence alignments of multiple species to produce per base conservation scores (Pollard et al., 2010). Average conservation score per cryptic tag was calculated using *bigWigSummary* (UCSC) for both human and mouse data. The lists of splice junctions created by *STAR* when aligning each sample were used to identify the coordinates of the exons adjacent to the cryptic exon. The randomly sampled intronic sequence from the cryptic-containing intron was used as a negative control.

Protein prediction analysis

Any cryptic exon which did not fall within the coding sequence of a transcript was omitted. Splice junctions were used to determine the upstream and downstream exons adjacent to each cryptic exon. These exons were matched to their corresponding annotated exon in the Ensembl transcript file for each species to work out the correct frame of translation. Nucleotide sequences for transcripts either including or excluding the cryptic tag were created. If the cryptic exon had been previously flagged as an extension then the entire continuous intronic sequence was inserted up to the remaining cryptic splice site. These transcripts were then translated *in silico* and defined as premature termination codon (PTC)-containing if the inclusion transcript contained a stop codon and as a frameshift if the sequence of the downstream exon no longer matched. A null distribution of PTC-containing or frame shifted transcripts was created by shuffling the identity of the central exon 1000 times.

Splice junction scoring

The strength of 5´ and 3´ splice sites was calculated for the human cryptic exons using *maxEnt* (Yeo et al., 2004). Higher scores indicate the increased log odds of a given splice site being a true splice site. The 5´ splice site is defined as the last 3 nucleotides of the upstream exon flanked by 6 intronic nucleotides, of which the first two are invariably GU. The 3´ splice site is defined as the last 20 intronic nucleotides of which the final two are invariably AG, flanked by the first 3 nucleotides of the downstream exon. The splice sites of annotated exons were used as a positive control. Randomly generated sequence with invariant AG or GT was used as a negative control. Paired t-tests were carried out to test the direction of change between the cryptic and annotated splice sites for each class of cryptic exon.

3.3 Results

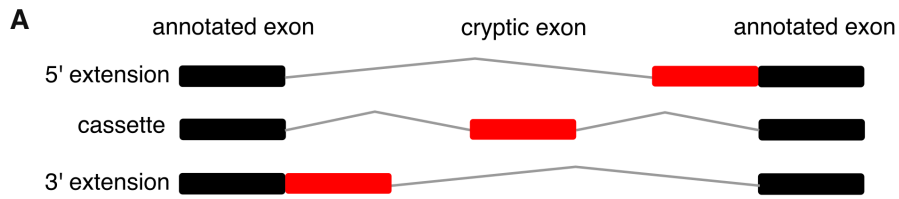
Depletion of TDP-43 but not FUS results in cryptic exons

I downloaded and analyzed 9 publicly available RNA-seq datasets (Table 3.1. This comprised TDP-43 depletion (three human and three murine, datasets 1-6), FUS depletion (1 human, 1 murine, datasets 7-8) and as a positive control a human hnRNP C depletion dataset for which cryptic exons have previously been reported (dataset 9). While these datasets differ in library preparation method, read depth and length, and protein depletion method (Table 3.2), the FUS datasets match the TDP-43 datasets by cell type.

Significant cryptic exons discovered by *Cryptex* were classified into three categories: (i) cassette-like, where novel 3´ and 5´ splice sites are recognised, which forms a completely new exon; (ii) 5´ extension, where a novel 3´ splice site is recognised and an existing exon is extended upstream of its annotated start and (iii) 3´ extension, where a novel 5´ splice site is recognised and an exon is extended downstream of its annotated end (Fig. 3.2A). *Cryptex* does not consider fully retained introns, but other methods have been designed for this purpose (Li et al., 2015; Bai et al., 2015; Braunschweig et al., 2014).

Comparing the two human ENCODE K562 cell line TDP-43 depletion datasets (3-4), the poly-A selected mRNA-Seq dataset yielded far more splicing events than the total RNA dataset, presumably due to polyA+ selection leading to a higher coverage of mature spliced mRNA species (Fig. 3.2B). In total 95 human cryptic exons were discovered and classified, with the majority only detected in the mRNA-seq dataset. 11 cryptic splicing events were shared between datasets 3 and 4 (Fig. 3.2C). Of the 26 human cryptic exons reported by Ling, 12 were detected in at least one of the two datasets 3 and 4.

Both mouse datasets differ in both cell type (adult striatum in dataset 5 vs embryonic stem (ES) cell in dataset 6) and read depth (35-60M in dataset 5 vs 2-10M in dataset 6). 52 cryptic exons were identified in total, with 46 detected in the adult striatum and 15 in ES cells, with 6 exons observed in both. Of the 46 cryptic splicing events identified in



B

	unfiltered output	classified as cryptic exon			
TDP43 depletion		**5' extension**	**cassette**	**3' extension**	
Mouse adult brain	1080	7	18	17	Polymenidou, 2011
Mouse ES cells	44	4	8	3	Chiang, 2010
Mouse ES cells	NA	2	29	15	Ling, 2015
Human K562 mRNA	1530	26	31	27	ENCODE
Human K562 total RNA	359	7	6	9	ENCODE
HeLa cells	NA	4	15	11	Ling, 2015
FUS depletion					
Mouse adult brain	1575	0	0	1	Lagier-Tourenne, 2012
Human K562 mRNA	29	3	0	0	ENCODE
hnRNPC depletion					
HeLa cells	2683	266	227	150	Zarnack, 2013

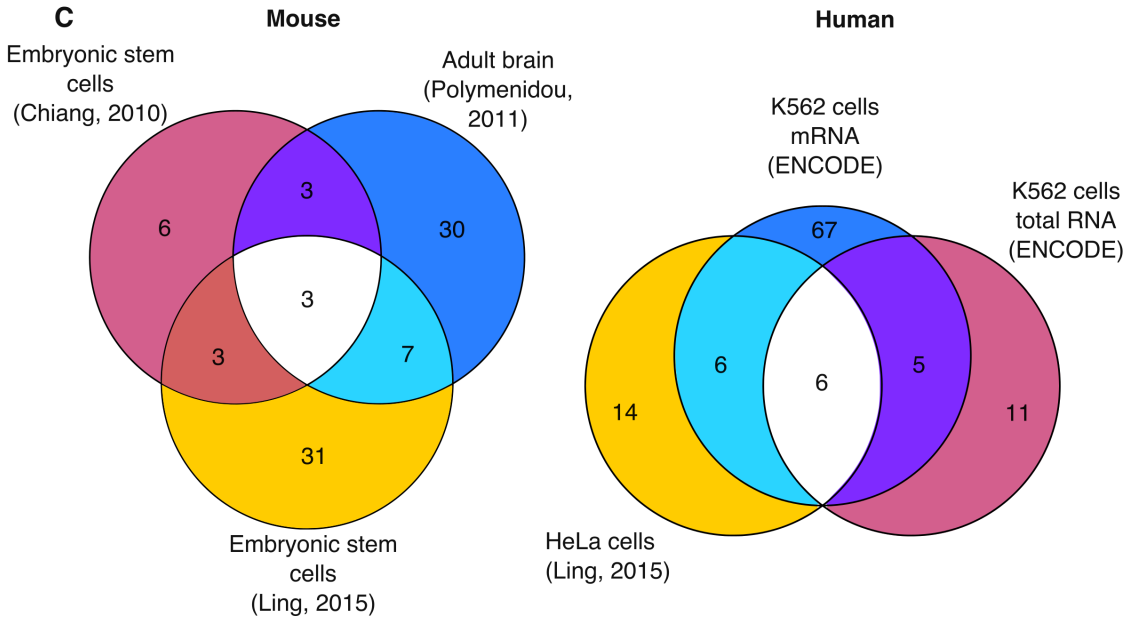


Figure 3.2: Cryptic splicing discovered by the *CryptEx* pipeline (A) Schematic of the three classes of cryptic exon. Black boxes represent annotated exons and red boxes represent a cryptic exon. Grey lines represent the spliced intron. **(B)** Tally of the three classes of cryptic exon discovered by the Cryptex pipeline across the nine datasets. "Unfiltered output" refers to the number of differentially used cryptic splicing events at a false discovery rate (FDR) $<$ 5% before undergoing cryptic exon classification. Counts from Ling et al's data are taken from the paper itself. **(C)** Venn diagrams showing the overlap between the six TDP-43 depletion datasets.

murine samples by Ling et al, 13 were detected in at least one of datasets 5 and 6. Side by side visual inspection suggests that differences in library preparation and read depth are behind the low concordance rates in both human and mouse, as cryptic exons detected in

the higher depth dataset (K562 mRNA and mouse adult brain) can be observed by eye in the lower depth dataset (K562 total RNA and mouse ES cell). These exons currently fail to be detected by the *CryptEx* algorithm.

No cryptic splicing events were shared between human and mouse as previously reported (Ling et al., 2015). Note that to report overlap with Ling and colleagues (datasets 1 and 2), the raw data was unsuitable for the cryptic exon discovery pipeline due to a lack of biological replicate samples. Instead the sequence data was aligned and the splice junctions generated by the aligner were used to classify previously reported cryptic exons.

In contrast, while a large number of novel splicing events were observed in the FUS depletion datasets, our algorithm only classified 3 in mouse and 1 in human as cryptic exons. FUS depletion was not observed to produce any cassette-like cryptic exons in either species. The coordinates of each cryptic exon found in human and mouse are reproduced in the appendices.

Cryptic exons are bound by TDP-43

TDP-43 linked cryptic exons were grouped into unions of all cassette-like exons and extension events discovered in human and mouse, totalling 95 human and 52 murine cryptic exons. I then explored whether TDP-43 binding could explain the observed splicing changes in RNA-Seq data, as observed by Ling and colleagues. I took two complementary approaches: (i) searching for enriched motifs in the RNA sequence including and surrounding the cryptic exons (Fig. 3.3A,D) and (ii) correlating the positions of cryptic exons with TDP-43 protein-RNA interaction data (Fig. 3.3B,C).

TDP-43 can repress or enhance the inclusion of a given exon by either binding within or adjacent to the exonic sequence (Tollervey et al., 2011). Hence, for our motif search, I flanked cryptic exon sequences by 100 nucleotides on either side. UG-rich motifs were found to be enriched in both mouse and human cryptic exons using two different algorithms: *MEME* (Fig. 3.3A) and *HOMER* (appendices). Of the 52 mouse cryptic exons, 29 had a run of UG up to 40 nucleotides in length. Similarly, human cryptic exons were enriched in a UG motif but not in a continuous manner. By comparing the frequencies of 16 possible dinucleotides between the flanked cryptic exon sequence and the sequence of the adjacent intron either up or downstream of the cryptic-containing intron it was possible to resolve the enrichment of UG dinucleotides (Fig. 3.3D). UG and GU were enriched in flanked cryptic exon sequence in both human (fold change GU = 1.53; UG = 1.48; $P < 10^{-50}$; proportion test) and mouse (fold change GU = 2.14; UG = 1.85; $P < 10^{-50}$; proportion test). If TDP-43 binding was uniform throughout an intron or gene then one would expect to see similar proportions of overlap between cryptic exons, surround intronic sequence and adjacent exons (Fig. 3.3B). However, both species show an enrichment in TDP-43 binding peaks specific to the cryptic exons in every iCLIP dataset used, with as much as 25% of human cryptic exons and 50% of mouse cryptic exons overlapping at least one iCLIP peak each (both species: $P < 10^{-16}$, proportion test) (Fig. 3.3C).

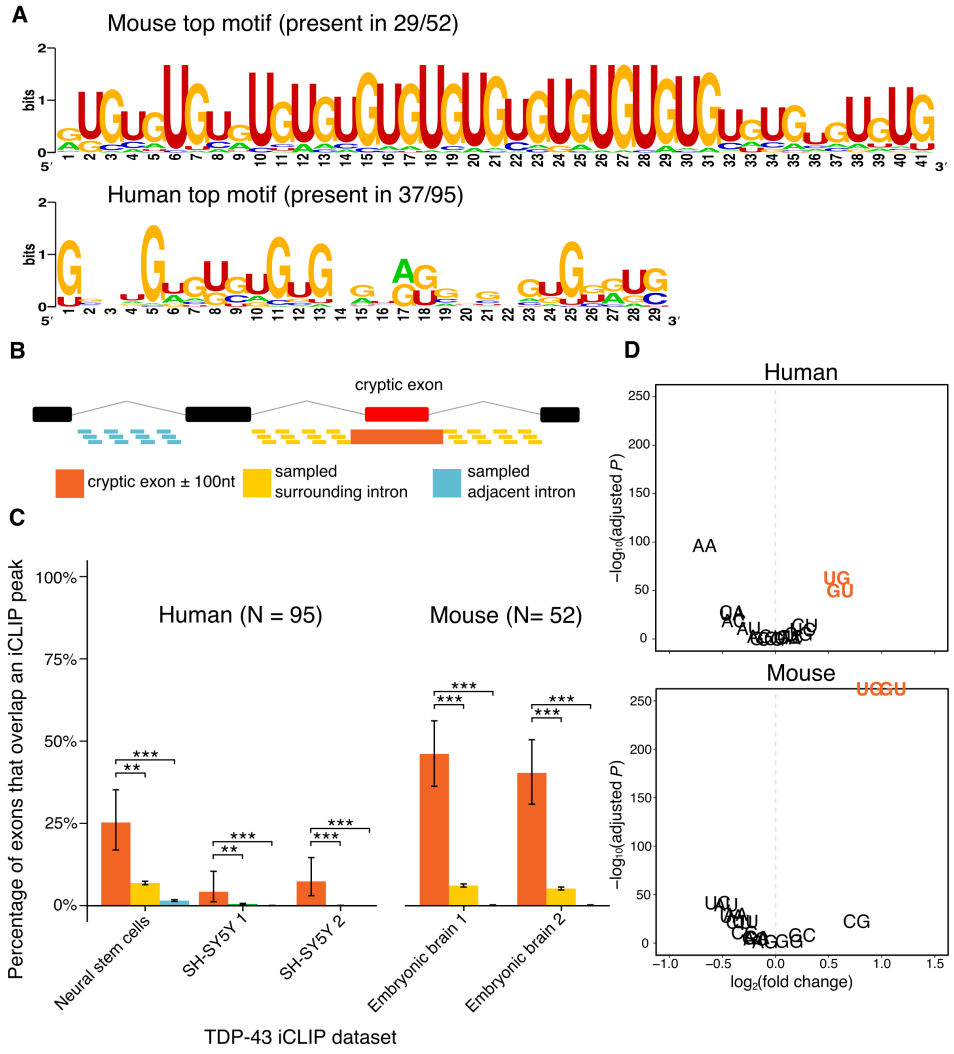


Figure 3.3: Evidence of TDP-43 binding cryptic exons. (A) Results of *MEME* motif search. Only the motif with the greatest enrichment compared to background sequence is presented. **(B)** Schematic of iCLIP peak enrichment test. For illustration a cassette-like cryptic exon (green box) is shown between two annotated exons (black boxes) separated by intronic sequence (black lines). The proportion of the group of cryptic exons flanked either side by 100 nucleotides (orange) that overlap at least one iCLIP peak is compared to the proportion of overlaps in a group of length matched sequences from either the surrounding intron (yellow) or an adjacent intron (blue) each randomly sampled 100 times per gene. **(C)** iCLIP peak overlap enrichment for the 95 human cryptic exons found in either K562 cell TDP-43 depletion dataset and the 52 cryptic exons found in either mouse TDP-43 depletion dataset. D) Dinucleotide enrichment in the flanked cryptic exons compared to adjacent introns. Error bars denote 95% confidence intervals of the binomial distribution. $^{*}P < 0.05$, $^{**}P < 0.001$, $^{***}P < 10^{-16}$. All P-values adjusted for multiple testing by Bonferroni method.

Cryptic exon recognition is unrelated to the binding of transposable elements by TDP-43

TDP-43 has been demonstrated to bind antisense Alu elements, which are the source of cryptic exons repressed by hnRNP C (Zarnack et al., 2013; Kelley et al., 2014). I therefore

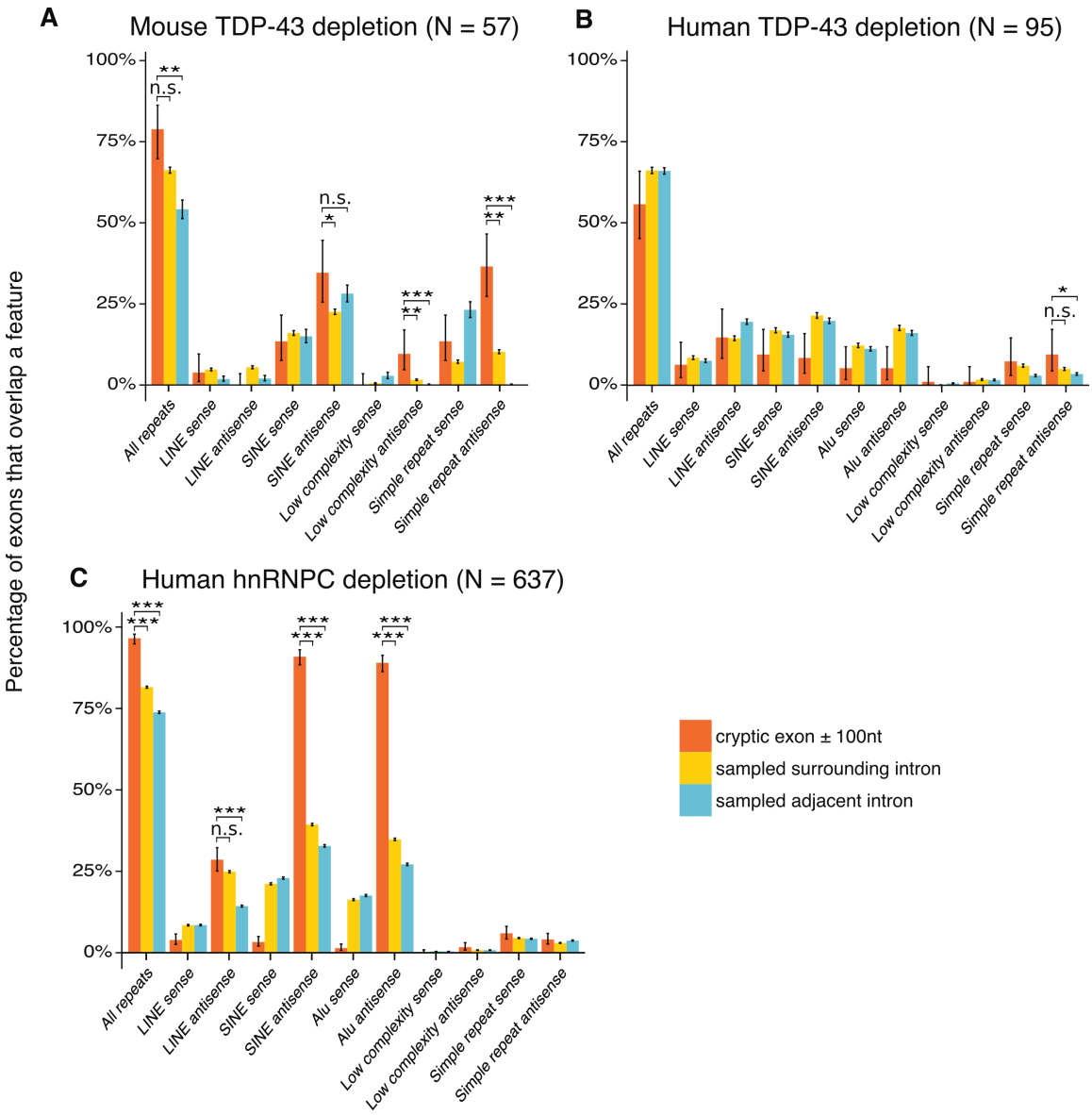


Figure 3.4: Cryptic exons in transposable elements. Overlap between different families of repetitive element with lists of exons, separated by orientation. **(A)** Mouse TDP-43 depletion. **(B)** Human TDP-43 depletion. **(C)** Human hnRNP C depletion. The proportion of the cryptic exons that contain a particular element are shown in orange. Length matched random samples from the surrounding intron (yellow) and adjacent introns (blue) are used as controls. LINE: Long Interspersed Nuclear Element; SINE: Short Interspersed Nuclear Element. $^{*}: P < 0.05$, $^{**}: P < 0.001$, $^{***}: P < 10^{-16}$. All P-values corrected for multiple testing with Bonferroni method.

investigated whether TDP-43 induced cryptic exons preferentially overlap specific families of transposable elements and/or class of repetitive sequences. Transposable and repeat elements annotations were obtained using the *RepeatMasker* software, and these features were split by family and orientation. Although Alu elements are a subfamily within the primate SINE element family, I included them separately given the prior hnRNP C result. Control regions were obtained as before. Cryptic exons show only a modest enrichment patterns in the simple repeat" family, owing to the aforementioned UG motifs (Fig. 3.4A/B). This contrasts with hnRNP C depletion, which shows a striking enrichment of antisense SINE elements of which all are of the Alu type ($P < 10^{-16}$, proportion test; Fig. 3.4C), a

result consistent with previous analyses of dataset 9 (Zarnack et al., 2013).

Cryptic exons are poorly conserved and generate premature stop codons

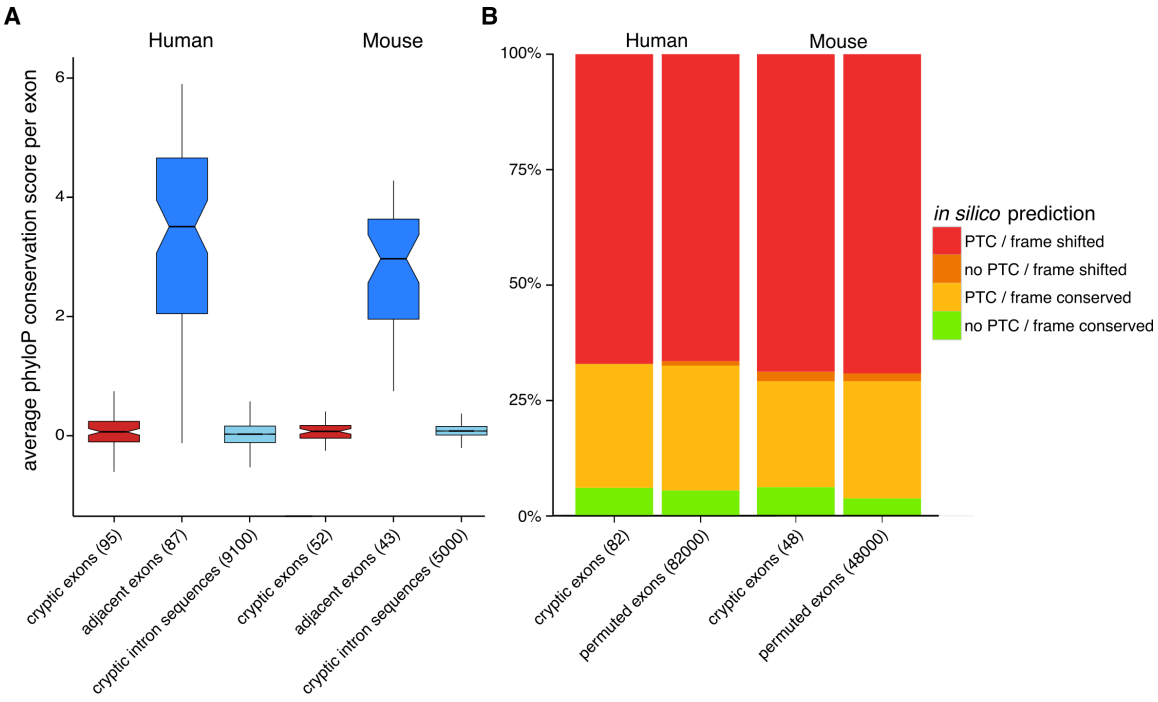


Figure 3.5: Conservation and premature termination codon analysis. (A) Per exon average PhyloP conservation scores in cryptic exons, adjacent exons within the same set of genes (when available) and randomly sampled sequences from the cryptic containing intron. Box plots show first quartile, median and third quartile with the notches representing the 95% confidence interval of the median. Whiskers represent the minimum and maximum values that fall within 1.5 times the interquartile range. **(B)** The functional impact of cryptic exon inclusion on the host transcripts in human and mouse. Colours indicate the category of prediction and box size indicates the proportion of the total group of exons in each category. Categories from top to bottom: at least one premature termination codon (PTC) introduced and frame shifted (red); no PTCs introduced but frame shifted (orange); PTCs introduced but frame conserved (yellow); no PTCs introduced and frame conserved (green). For each species there is a corresponding set of null exons where the central exon has been permuted 1000 times.

I then quantified the extent of evolutionary conservation of cryptic exons using the multiple species alignment conservation scores generated by *PhyloP*. I calculated mean conservation scores per exon for the cryptic exons and compared them to scores from both the annotated exons and randomly sampled intronic sequences from the same genes. No differences were observed between cryptic exons and matched intronic sequences (Fig. 3.5A), and a much lower conservation level than adjacent annotated exons.

I also investigated the consequences of inclusion of cryptic exons on translation of the transcript. Potential outcomes for each gene are: (i) a functional transcript, (ii) a premature termination codon (PTC) or (iii) a frameshift variant. I compared these estimates to random simulations where the identity of the included exon has been permuted 1000 times. The results were consistent with the null expectation of neutral evolution, with around 66% of cryptic exons leading to a frameshift due to length mismatches and less than 10% of cryptic exons predicted to create functional transcripts (Fig. 3.5B).

Cryptic exon containing genes are downregulated

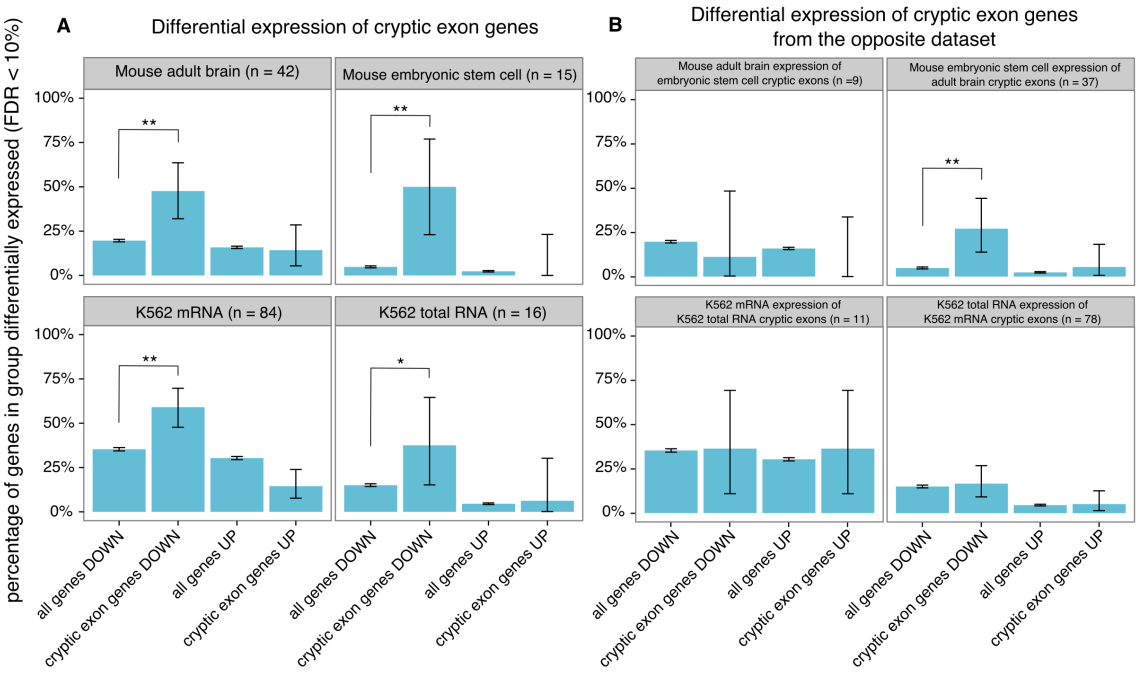


Figure 3.6: Differential expression of cryptic exon genes. (A) Significantly differentially expressed genes between TDP-43 depletion and control samples (false discovery rate = 10%) in datasets 3-6. Comparisons were made between all genes tested with an expression level at or greater than the lowest expressed cryptic exon ("all genes") and the genes where cryptic exons were discovered ("cryptic exon genes") with each group divided by direction of change. **(B)** As above but for cryptic exon genes from the other dataset of the same species. Error bars show 95% confidence intervals for each proportion. $^{*}: P < 0.05$, $^{**}: P < 0.001$. All P-values adjusted by Bonferroni correction.

I then investigated, in datasets 3-6, whether genes containing cryptic exons showed a specific pattern of altered expression. I calculated the proportion of the cryptic exon containing genes in each dataset that were differentially expressed at a FDR of 10%. This was then compared with the proportion of differential expression of all genes with an expression level at or greater than the lowest expressed cryptic exon found in that dataset. I then plotted the number of differentially expressed genes in each dataset as a proportion of the total, separated by direction. In all four TDP-43 depletion datasets, the cryptic exon containing genes as a group are more likely to be significantly downregulated compared to the genome-wide proportion ($P < 0.001$, hypergeometric test; Fig. 3.6A).

Furthermore, I performed the same analysis for each dataset with the cryptic exon containing genes that were only found in the other dataset of the same species (Fig. 3.6B). Surprisingly, in the mouse ES cell dataset 6 there was an enrichment of downregulated genes that contain cryptic exons only detectable in the mouse adult brain dataset 5 ($P < 0.001$, hypergeometric test). Visual inspection of these 10 introns in the mouse ES cell data suggests that 7 of them may harbour cryptic exons in the ES cell data that are currently undetectable by the *CryptEx* algorithm.

Human cryptic exons are driven by the recognition of strong splice sites that are normally repressed

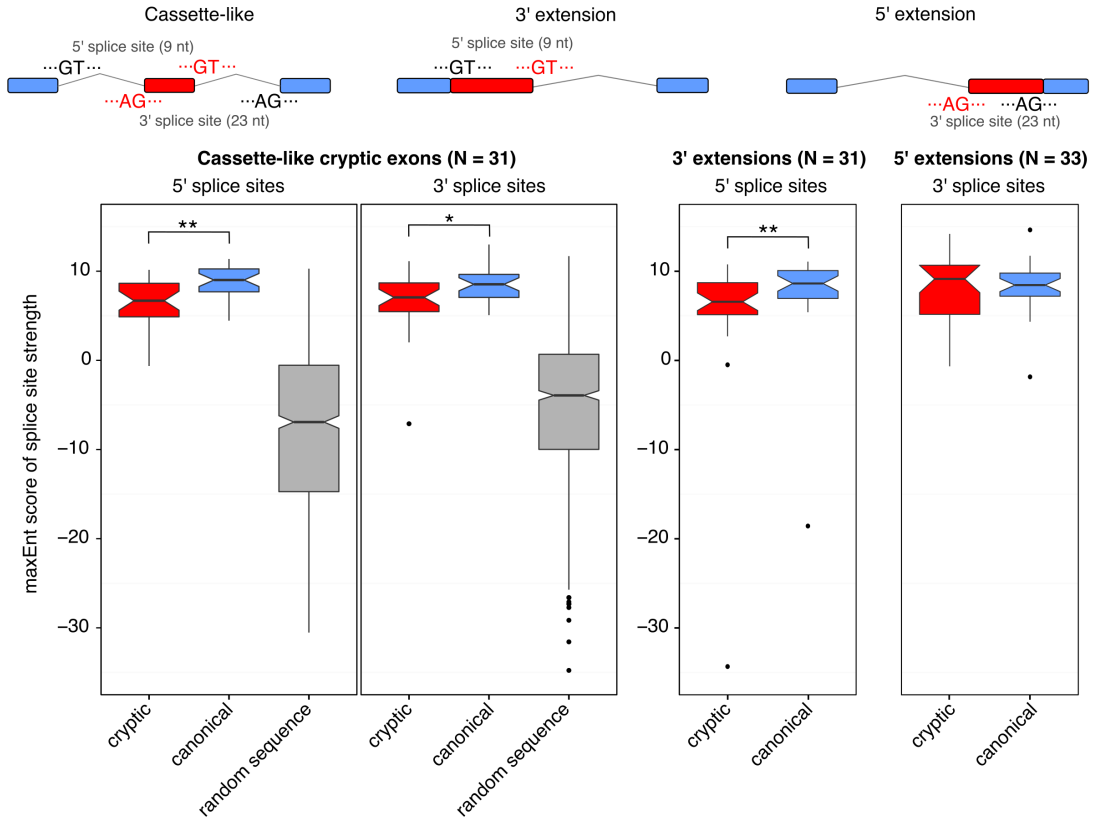


Figure 3.7: Scoring cryptic splice sites against canonical splice sites. Cassette-like cryptic exons have 5´ and 3´ splice sites that are recognised by the spliceosome under TDP-43 depletion. 5´ and 3´ splice site scores are plotted separately. Cryptic splice sites are shown in red, canonical splice sites are shown in blue. Random sequences with AG or GT consensus are plotted in grey. Cryptic extensions are the result of a cryptic splice site competing with the canonical splice site. 5´ extensions result from cryptic 3´ splice sites whereas 3´ extensions result from cryptic 5´ splice sites. Box plots show first quartile, median and third quartile with the notches representing the 95% confidence interval of the median. Whiskers represent the minimum and maximum values that fall within 1.5 times the interquartile range. Outliers are plotted as black dots. Cryptic splice sites are compared to canonical splice sites with paired t-tests. * : $P < 0.05$, ** : $P < 0.001$.

Whereas cassette-like cryptic exons appear as separate exons distinct from their surrounding exons, extension events must rely on a switch from a canonical splice site to a newly accessible splice site. I hypothesised that these extension events result from competition between two splice sites upon TDP-43 depletion. This would require the sequence of and around the cryptic splice site to be similarly recognisable to the spliceosome. Using the *MaxEnt* statistical model to score splice sites by comparing their DNA sequences with constitutive observed canonical sequences, I scored the 5´ and 3´ splice sites of our cryptic exons and compared them with the scores of the surrounding canonical splice sites. The model compares splice sites from annotated exons with decoy splice sites that retain the consensus AG/GT at the 3´ or 5´ splice site respectively. Therefore I also scored randomly generated

sequences which retained the consensus AG/GT positions. Although the canonical splice sites were on average stronger than their corresponding cryptic splice site ($P < 0.05$, paired t-test), the majority had scores far greater than those from random sequence (Fig. 3.7), suggesting that they are able function as genuine, albeit weaker, splice sites when TDP-43 is depleted.

TDP-43 cryptic exons are bound by other RNA binding proteins

Proteomic studies have demonstrated that TDP-43 interacts with a number of RNA-binding proteins (RBPs), including multiple members of the heterologously expressed ribonucleoprotein (hnRNP) family and other splicing factors (Blokhuis et al., 2016; Ling et al., 2010; Freibaum et al., 2010). The splicing of specific annotated exons has been shown to depend on the interaction of TDP-43 with multiple splicing factors (Mohagheghi et al., 2016). I hypothesised that some cryptic exons may be included indirectly through a loss of interactions with different RBPs. Van Nostrand and colleagues have performed eCLIP, a higher throughput modification of the iCLIP protocol, on 73 different RBPs including TDP-43 and FUS (Van Nostrand et al., 2016). The experiments were carried out in 2 human cell lines (K562 and HepG2) with 29 of the RBPs being tested in both cell lines. I performed the same overlap analysis between our human cryptic exons and each set of eCLIP peaks, using the same two sets of control sequences as before. Each eCLIP experiment was performed in duplicate. This gives each RBP four possible enrichment results using a proportion test. For each RBP, the highest P-value from the four tests was reported and corrected for multiple testing. Only proteins with a resulting $P < 0.05$ are reported. Unsurprisingly TDP-43 had the highest number of overlapping exons ($p < 10^{-22}$; proportion test), followed by U2AF65, TIA1, SRSF7, U2AF35, PPIG, SRSF1 and IGF2BP1 (Fig. 3.8A). Hierarchical clustering was performed on the RBPs. The three largest clusters consist of TDP-43 alone, the U2 snRNP binding proteins U2AF35 and U2AF65, and a third cluster containing the other proteins (Fig. 3.8B).

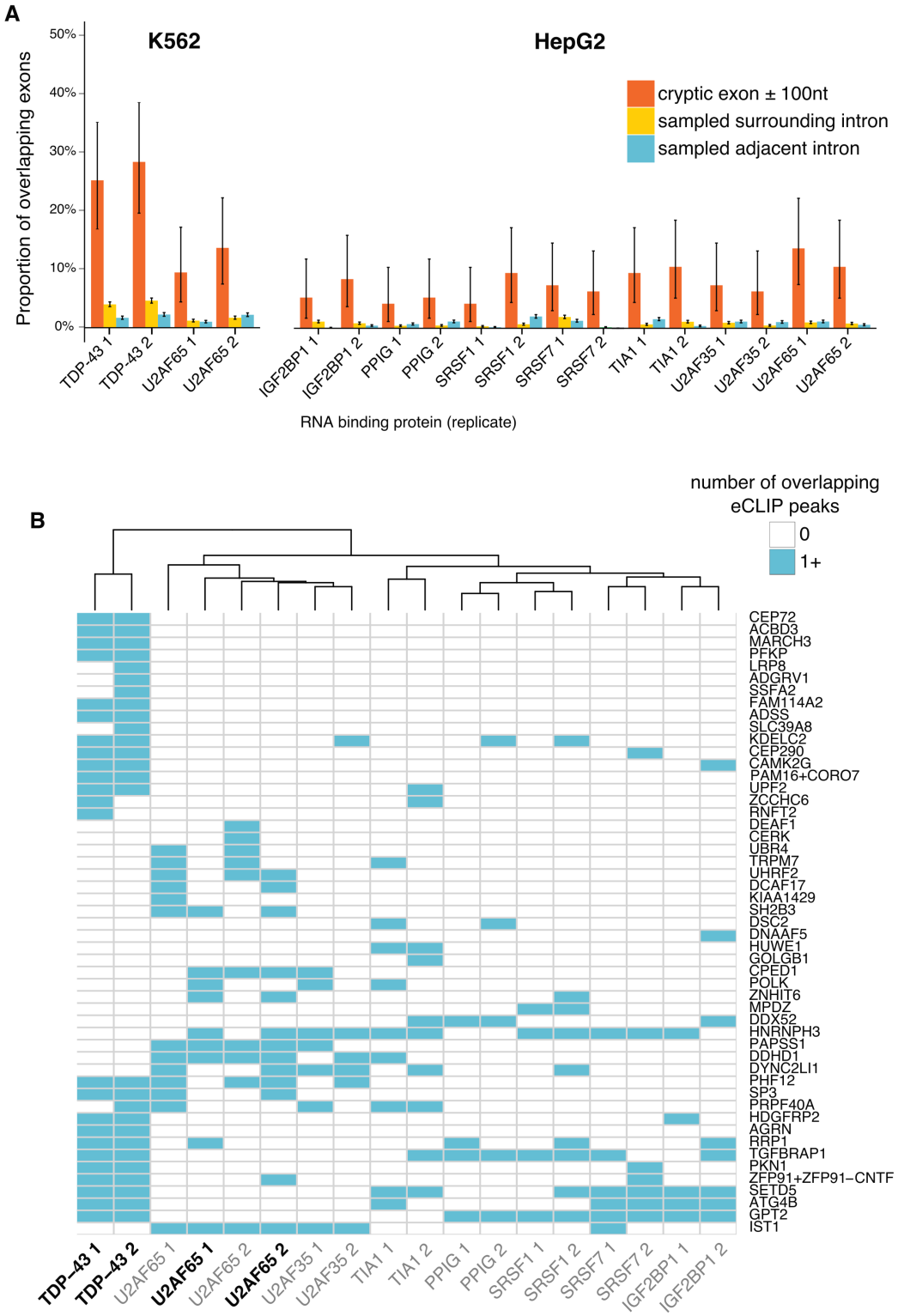


Figure 3.8: Mining of ENCODE eCLIP data in K562 and HepG2 cells. (A) RNA-binding proteins (RBPs) with significant ($P < 0.05$) proportion of overlapping cryptic exons (red) compared to sampled intronic sequence from the same (green) and adjacent introns (blue). **(B)** Comparison of eCLIP datasets from different RNA binding proteins (columns) showing the overlap with the cryptic exons (rows). RBPs in bold typeface are from K562 cells whereas those in regular typeface are from HepG2 cells.

3.4 Discussion

I designed an analytical strategy to identify cryptic splicing that takes advantage of biological replicates in RNA sequencing data. I have applied this tool to a set of human and murine TDP-43 depletion datasets, as well as datasets that deplete hnRNP C or FUS. The results are consistent with the previous findings that depletion of TDP-43 or hnRNP C leads to the inclusion of novel cryptic exons in both human and mouse. Although FUS undoubtedly plays an important role in splicing and mRNA stability and shares a number of targets with TDP-43 (Lagier-Tourenne et al., 2012), the low number of cryptic exons observed due to FUS depletion suggests that it does not play a major role in cryptic splicing and is a key point of differentiation with TDP-43. This can explained by the action of TDP-43 as a local splicing repressor (Tollervey et al., 2011), a function that is not shared by FUS (Rogelj et al., 2012).

Further examination of TDP-43 linked exons suggests they tend to possess the necessary UG-rich sequence elements to be bound by TDP-43 and using iCLIP data I observed that a subset of the cryptic exons are shown to be bound by TDP-43 *in vivo*. I went on to investigate the origins of these TDP-43 bound cryptic exons, as has been done for the targets of hnRNP C. I observed that unlike hnRNP C linked cryptic exons, which invariably originate from antisense Alu elements, TDP-43 linked cryptic exons do not originate from any single family of transposable element. Furthermore their sequences show very low species conservation, akin to random intronic sequence, but remarkably they contain splice sites very close in strength to those of their adjacent annotated exons. Differential expression analysis suggests that the bulk of cryptic exon containing genes are significantly downregulated upon TDP-43 depletion. I hypothesize that this is due to nonsense mediated decay of the inclusion transcript, which I predict would occur in over 90% of cryptic exon containing transcripts. Comparing my human cryptic exons with a proteomic study of TDP-43 depletion in human SH-SY5Y cells (Štalekar et al., 2015) showed that protein levels were changed for 3 of the 95 human cryptic exon containing genes. Two, *HUWE1* and *GOLGB1* had protein levels that were 8% and 31% of the control cells respectively whereas the third, *HNRNPH3* was found to be 7-fold increased under TDP-43 depletion. Interestingly, the cryptic exon discovered in hnRNP H3 falls upstream of the start codon whereas those found in *HUWE1* and *GOLGB1* are predicted to trigger NMD by inclusion of intronic RNA into the coding sequence (see Supplementary Table 1). Another gene with cryptic splicing seen in both K562 datasets and in the initial Ling HeLa data, *AGRN*, has been shown to be decreased at the protein level in the cerebrospinal fluid of ALS patients compared to healthy controls and other neurological diseases (Collins et al., 2015). Correct splicing of *AGRN* has shown to be crucial for the formation of the neuromuscular junction (Ruggiu et al., 2009).

My understanding of the role of TDP-43 in cryptic splicing is that of a safeguard against the inclusion of potentially damaging intronic sequence into transcripts. However, the relationship between the UG-rich sequences and the strong 5´ and 3´ splice sites and their changes over evolutionary time are unknown as I observed no conserved cryptic exons between human and mouse. In each species, the cryptic exon sequences represent newly forming exons

emerging from neutral evolution following divergence. These sequences are under blanket splicing repression by TDP-43 and other factors and so are not under selection.

Using publicly available ENCODE eCLIP data, I identified a number of RNA binding proteins that also bind subsets of human cryptic exons under normal conditions, that is, in the presence of TDP-43. It is unsurprising that the splicing factors U2AF35 and U2AF65 are enriched as they preferably bind pyrimidine-rich 3´ splice site sequences which all cryptic exons appear to possess. That only 10-15% of cryptic exons show U2AF35/65 binding may be due to competition from TDP-43 in a manner similar to that seen between hnRNP C and U2AF65 (Zarnack et al., 2013). TIA1 is an exciting finding due to its role in the formation of stress granules, which are key regulators of RNA stability (Gilks et al., 2004). In addition, IGF2BP1 has been reported to be bound to TDP-43 in HEK293T and HeLa cell extracts (Ling et al., 2010; Freibaum et al., 2010), whereas SRSF7 was reported as binding to TDP-43 in mouse N2A cells (Blokhuis et al., 2016). None of the observed proteins have been reported to change their protein level in response to TDP-43 depletion (Štalekar et al., 2015).

As the majority of cryptic exons are predicted to lead to nonsense-mediated decay of the inclusion transcript it seems peculiar that we can observe these transcripts at all. I hypothesise that cryptic splicing may be much more widespread than can be observed by RNA sequencing as I predict that the majority of cryptic exons are substrates for nonsense-mediated decay (Losson and Lacroute, 1979; McGlincy and Smith, 2008). There may be more cryptic exons that are degraded more efficiently by nonsense-mediated decay than others and so may be uncovered if nonsense-mediated decay was inhibited. Over half of all cryptic exon genes are significantly downregulated in each dataset (Fig 3.6B), suggesting that cryptic exon inclusion may be a key mechanism in the widespread changes in RNA expression that occur upon TDP-43 depletion. This has been explored in hnRNP C, where a large increase in the number and observed inclusion of Alu-derived cryptic exons was seen when both hnRNP C and the nonsense mediated decay-associated protein UPF1 were co-depleted compared to just hnRNP C or UPF1 alone (Attig et al., 2016).

Two genes, *ATG4B* and *GPSM2*, have previously been demonstrated to have cryptic exon inclusion RNA transcripts in ALS patient brain samples, suggesting a role for cryptic splicing in disease (Ling et al., 2015). My analysis also identified a cryptic exon in *ATG4B* in human cells, but not *GPSM2*; however I did not analyse human brain data. By expanding the list of cryptic exons, it will be interesting to explore whether these are also dysregulated in ALS patient brains. However, such analysis may prove challenging owing to the likely small concentrations of RNA originating from diseased cells in brain homogenate and the likelihood of degradation by NMD. Alternate strategies may involve mass spectrometry screens for the subset of cryptic exon containing genes that escape the NMD process, for example because the cryptic exon is in frame. Such proteins may represent useful biomarkers for ALS.

A complementary bioinformatic method was used increase the number of cryptic exons seen in the original Ling data (Tan et al., 2016). This study extends the cryptic splicing phenomenon to RBM17, another RNA-binding protein.

Defining cryptic exons by their absence in existing annotation is not a suitable long term strategy. For example, the *SORT1* gene contains an exon normally repressed by TDP-43 and not constitutively included in any tissue (Prudencio et al., 2012). But due to the studies that have documented its existence the "cryptic" exon in *SORT1* is annotated in all transcript databases. There are certainly many examples of annotated exons that are extremely rarely included and may be repressed by splicing factors like TDP-43. Future studies may have to take a more measured approach based on the frequency of inclusion in multiple tissues and conditions to assess whether an exon is truly cryptic or not.

Epilogue: Cryptic exons are a widespread phenomenon seen in many RNA-binding proteins

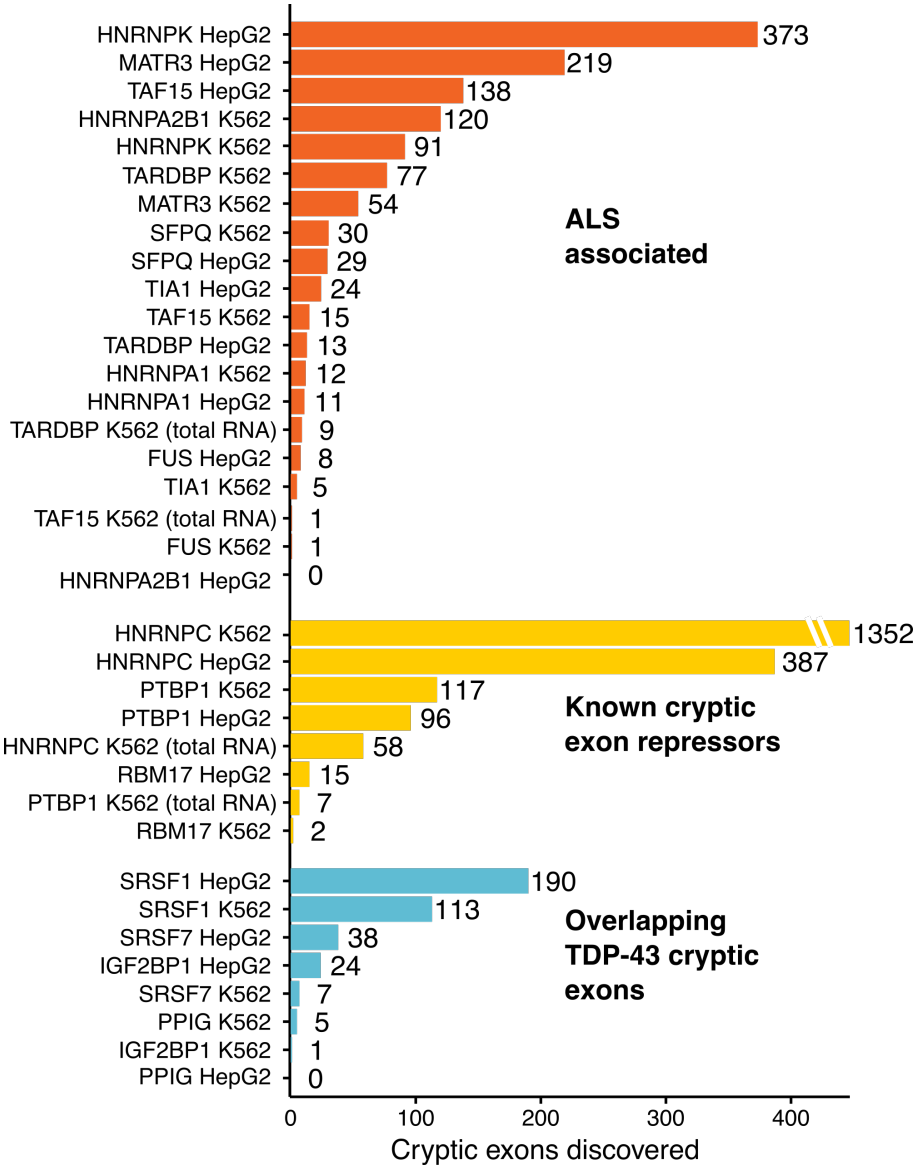


Figure 3.9: Cryptic exons found in different ENCODE shRNA knockdowns

Following completion and publication of this work, I downloaded a selection of shRNA knockdowns of RNA-binding proteins from the ENCODE portal. This included every available protein linked to ALS, including TIA1 and SFPQ, recently shown to harbour rare

disease-causing mutations (Mackenzie et al., 2017; Thomas-Jinu et al., 2017). In addition I downloaded knockdowns of hnRNP K, a splicing factor which interacts with TDP-43 (Freibaum et al., 2010) and plays a role in regulating TDP-43 aggregation (Moujalled et al., 2015). All data was aligned and run through the *Cryptex* pipeline for cryptic exon discovery. Cryptic exon abundances were compared between the knockdowns and their scrambled shRNA controls.

hnRNP K knockdown in HepG2 and K562 cells lead to the inclusion of 373 and 91 cryptic exons respectively (Fig 3.9), the largest effect seen in this set of proteins. MATR3 knockdown led to 219 and 54 cryptic exons. MATR3 has now been linked to cryptic splicing through binding within LINE elements (Attig et al., 2018). Perhaps most surprising is the finding of 138 and 15 cryptic exons being included in the knockdown of TAF15 in HepG2 cells and K562 cells respectively. TAF15 is closely related to FUS, in which I only observe 8 and 1 cryptic exons. This suggests diverging splicing functions between TAF15 and FUS, despite the two proteins binding similar motifs and overlapping RNA targets (Kapeli et al., 2016). As a comparison I downloaded data from proteins linked to cryptic splicing in other papers such as PTBP1 and RBM17 (Ling et al., 2016; Tan et al., 2016), as well as the previously discussed HNRNPC. Knockdown of either hnRNP C or PTBP1 leads to a robust number of cryptic exons in both cell types. Comparatively few events are seen in RBM17, despite 1475 genes being previously observed to harbour cryptic splicing changes in an *Rbm17* knockout mouse (Tan et al., 2016). In addition I included the proteins seen to be enriched in eCLIP binding to TDP-43 cryptic exons. Surprisingly nearly all proteins assessed show evidence of cryptic splicing upon their knockdown, with SRSF1 knockdown having the most. Together this shows that repression of cryptic splice sites is probably a more general function of RNA-binding proteins than previously realised. This makes the seeming lack of cryptic splicing seen in FUS knockdown all the more intriguing.

3.5 Summary

In this project I replicate and confirm the presence of cryptic exons after TDP-43 depletion and show they have a negative impact on the genes they reside in, leading to decreased expression levels. I have extended the scale and understanding of cryptic exons and their relation to TDP-43. In addition, I have demonstrated a key difference between FUS and other ALS-associated RNA-binding proteins. Further work is warranted to determine the relevance of cryptic exons to ALS and FTD pathogenesis.

4 FUS mutant mice show progressive changes in mitochondrial and ribosomal transcripts

Work presented in this chapter has been published as part of (Devoy et al., 2017). See appendices for full reproduction of the published manuscript.

4.1 Overview

This chapter describes work carried out in collaboration with Dr Anny Devoy of the UCL Institute of Neurology. Dr Devoy created a "humanised" mouse model of ALS resulting from a mutation in the FUS RNA-binding protein. I analysed RNA-seq taken from two tissues and time points and demonstrated a specific transcriptomic signature that correlates with a progressive neurodegenerative phenotype seen in aged mutant mice. This involves the downregulation of mitochondrial and ribosomal transcripts.

4.2 Contributions

- Transgenic mice were created by Dr Anny Devoy
- RNA sequencing libraries were prepared by Dr Anny Devoy
- RT-PCR validation was performed by Dr Anny Devoy
- Fig. 4.1 was created by Dr Anny Devoy

All bioinformatic analysis and interpretation was designed and performed by myself in consultation with Dr Anny Devoy and my supervisors.

4.3 Background

All previous studies alter FUS expression to levels that wildly differ to the normal biological situation. Overexpression or knockout are clearly toxic but these do not help to differentiate the effects of the mutations of normal FUS function. Until recently there have been no studies where the effects of a human FUS mutation are seen on the mouse at a physiological level of expression. The FUS Δ14 mutation was found in an early onset ALS patient who died at the age of 22 following a very rapid disease course (DeJesus-Hernandez et al., 2010). The mutation alters the 3´ splice site of exon 14 of FUS, causing it to be skipped. Comparison of the Δ14 mutation with other, late-onset ALS causing mutations have shown an increased propensity by FUS Δ14 to accumulate in the cytoplasm (Verbeeck et al., 2012).

Analysis of mice carrying a single copy of the FUS Δ14 mutation enables a reconstruction of progressive neurodegenerative disease. By analysing RNA-seq data collected across the lifespan of the mice, I can observe specific RNA dysregulation caused by the mutant protein.

The FUS Δ14 mouse is a humanised model of ALS

The FUS Δ14 mouse was created by directed mutagenesis of the mouse Fus exon 14 splice site as well as humanisation of exon 15 with 4 separate mutations.The result of splicing exons 13 and 15 together is a frameshift which at the protein level removes the C-terminal nuclear localisation signal (Fig. 4.1A) but leaves a novel peptide sequence which can used to create specific antibodies. Reverse-transcription PCR to detect FUS mRNA showed the Δ14 FUS mRNA to be expressed at a similar level to wildtype FUS (Fig. 4.1B).

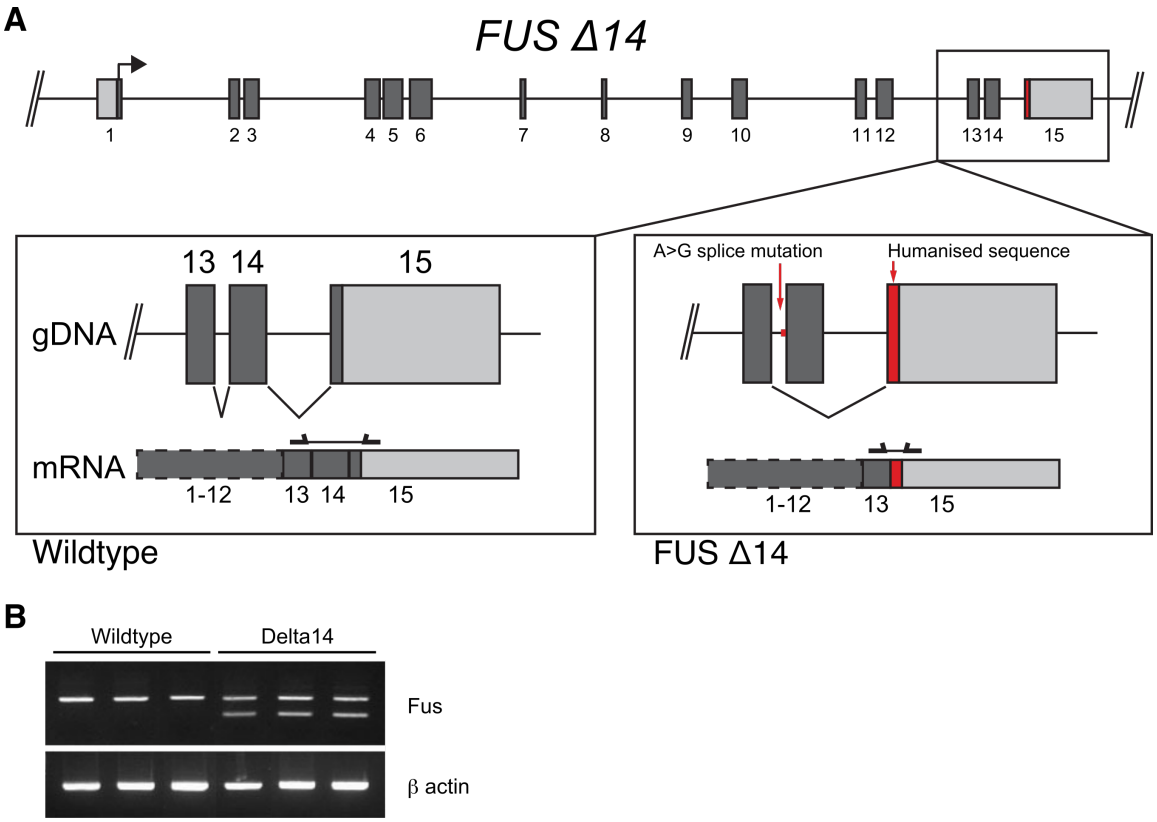


Figure 4.1: The FUS Δ14 model. (A) The FUS locus with a closeup of the terminal three exons in the wildtype mouse (left) and the FUS Δ14 mouse (right). **(B)** RT-PCR of FUS mRNA from spinal cord of wildtype and mutant mice.

4.4 Methods

Data preparation

All RNA-seq data used is listed in table 4.1. All samples were aligned to the mm10 mouse reference genome using the previously discussed analysis pipeline.

Table 4.1: All RNA-sequencing data used in this study Library info describes the type of library prepared, the length of each read and whether the sequencing was single or paired end. PE: paired end sequencing. Depth is defined the number of uniquely mapped read pairs, in millions.

Species	Cell type	Time point	Library info	Depth	Number
Mouse	Spinal Cord	3 months	stranded mRNA 75bp PE	33-43M	4 vs 4
Mouse	Spinal Cord	12 months	stranded mRNA 75bp PE	35-46M	4 vs 4
Mouse	Motor/Frontal Cortex	3 months	stranded mRNA 75bp PE	35-52M	4 vs 4
Mouse	Motor/Frontal Cortex	12 months	stranded mRNA 75bp PE	35-48M	4 vs 4

Differential gene expression

Differential expression was carried out with *DESeq2* (Love et al., 2014) comparing wildtype with mutant mice. All P-values were adjusted at a false discovery rate of 10%. To assess the variance between each sample the raw counts for each gene were normalised to account for library size and then normalised again by the regularized log normalisation method (Love et al., 2014). The counts were further transformed into Z-scores, which express the number of standard deviations from the mean of all counts for that gene. Heatmaps were created for all significantly differentially expressed genes using the *pheatmap* R package (Kolde, 2012).

Gene Ontology analysis

Gene ontology (GO) analysis is a method to extract inference from gene expression experiments by annotating each gene and protein within a unified vocabulary of molecular and cellular function. This can be used to identify dyregulated pathways or broader trends (Ashburner et al., 2000). For each differential expression results set the multiple correction threshold was dropped to $P < 0.005$ for each gene to increase the number of input genes. The resulting list of hits was then annotated for GO terms and a hypergeometric test was applied to test the enrichment of particular categories against a background set of genes. It is important to account for selection bias for long genes in RNA-seq data as longer genes have more power to be differentially expressed than shorter genes (Young et al., 2010). The R package *GOseq* implements a GO term annotation and enrichment test which takes account of this length bias (Young et al., 2010). The resulting P-values were then Bonferroni corrected for multiple testing. In each category the number of genes that were up- or downregulated in the $\Delta 14$ mice relative to wildtype littermates were expressed as a percentage.

Differential splicing

Differential splicing was analysed for the 12 month spinal cord samples using SGSeq (Goldstein et al., 2016) to find novel and annotated splicing events. Counts of inclusion and exclusion were used to fit a model with DEXSeq (Anders et al., 2012) to test the effect of condition on splicing event inclusion. Splicing events were reported at FDR < 0.05.

4.5 Results

Gene expression changes are tissue- and time point-specific

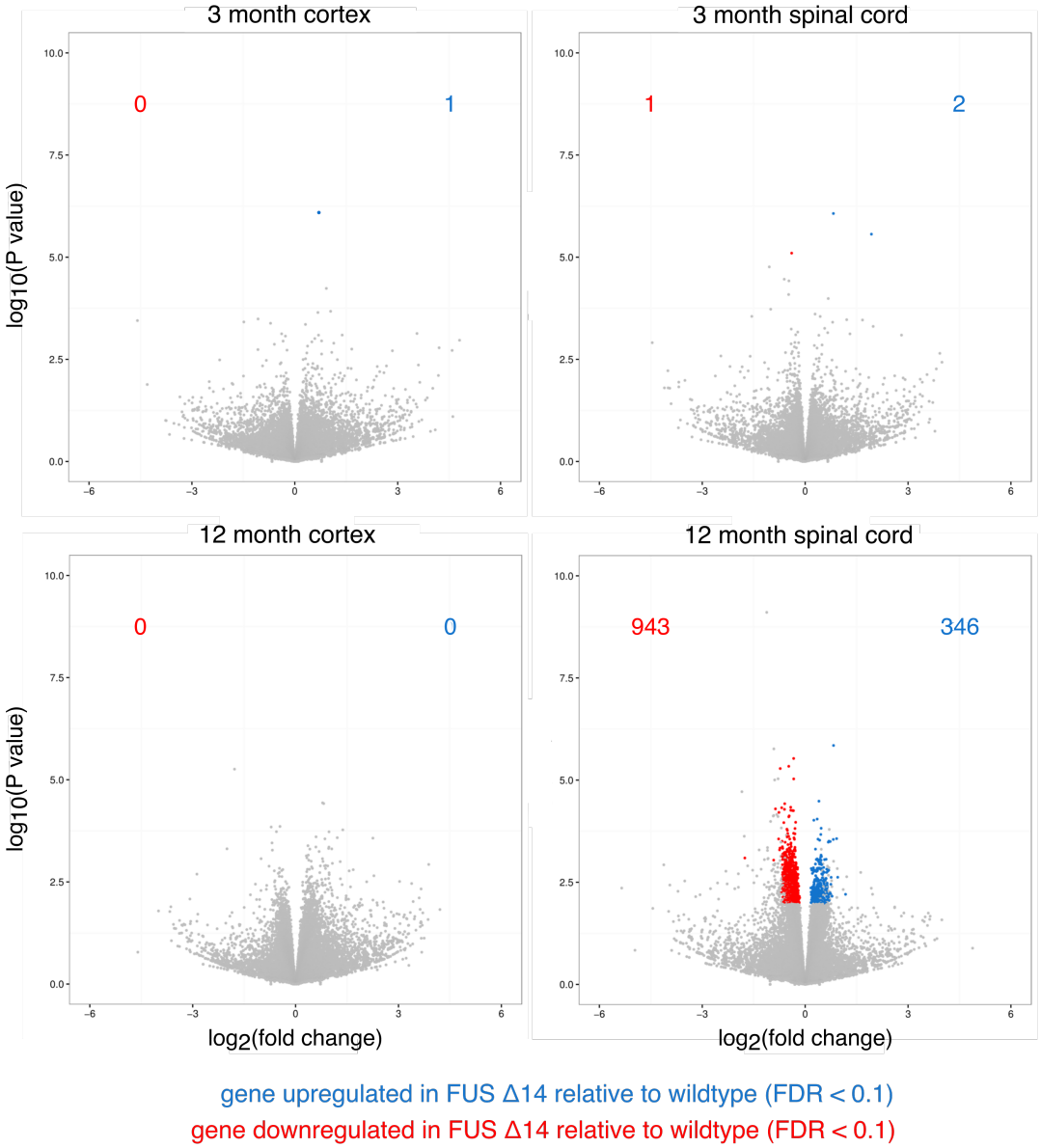


Figure 4.2: Differential gene expression analysis on the Δ14 mouse across two tissues and time points. Each gene is represented as a point. The x axis is the log_2 of the ratio between the average expression in the Δ14 mice against that of the wildtype mice. The y axis is the log_10 of the unadjusted P-value for the differential expression test. The numbers of upregulated genes (adjusted $P < 0.1$ with a log_2(fold change) > 0) are in blue and the the downregulated (adjusted $P < 0.1$ with log_2(fold change) < 0) are in red.

To examine progressive changes in RNA regulation, total RNA was extracted from both spinal cord and forebrain from mice across two time-points: 3 months and 12 months. Four male mice of each genotype (Δ14 and wildtype littermate) were used at each timepoint. At 3 months of age, the forebrain and spinal cord samples had 1 and 3 genes differentially expressed respectively. At 12 months of age there were no differentially expressed genes found

in the forebrain but 1,289 found in the spinal cord, with genes predominantly decreased in expression (Fig. 4.2). To examine the variance between each sample in the 12 month spinal cord dataset, the read counts for each gene were normalised and converted to Z-score values. Plotted as a heatmap it is clear that the size of each change is small and there is considerable inter-condition variability (Fig. 4.3).

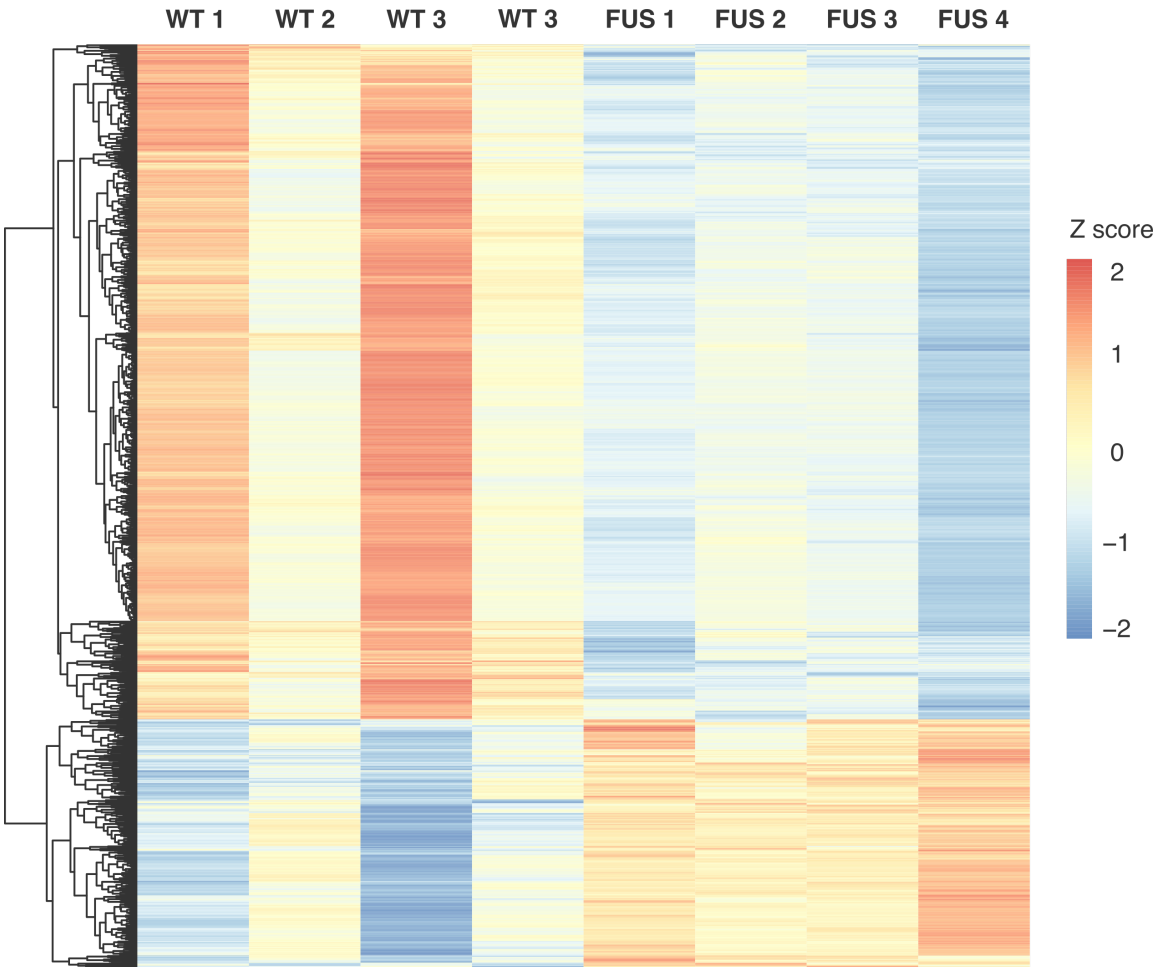


Figure 4.3: Z-score heatmap of the 1289 differentially expressed genes in 12 month spinal cord dataset WT: wildtype littermate; FUS: FUS Δ14 heterozygote.

Gene ontology analysis indicates changes in mitochondrial and ribosomal pathways

Genes from the 12 month spinal cord dataset differentially expressed at a relaxed threshold of $P < 0.005$ were sent for gene ontology enrichment analysis. This compares the number of genes that are members of a particular ontology category with the expected distribution. This analysis identified a strong enrichment in genes belonging to mitochondria, ribosome and proteasome cateogories. Strikingly, almost all the genes in these categories were down-regulated (Fig. 4.4).

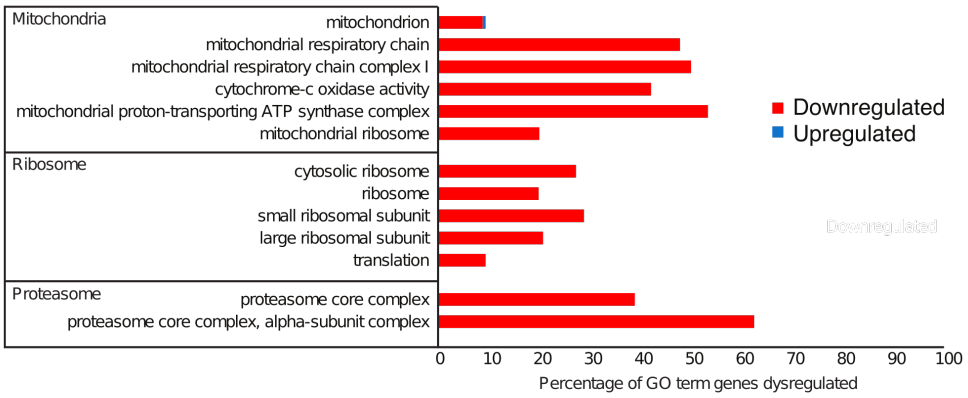


Figure 4.4: Gene ontology categories significantly enriched in the 12 month FUS Δ14 spinal cord samples Each significant gene ontology category (adjusted P < 0.05) is expressed as a proportion of upregulated (red) or downregulated (blue) genes.

Splicing events are limited to FUS itself and few other genes

Due to the lack of gene expression changes seen in other comparisons, differential splicing was assessed on 12 month spinal cord samples only. 11 splicing events were found at FDR $<$ 0.05, 4 of which were found in *Fus* itself. The top 2 splicing events in *Fus* overlap exon 14, the skipping of which in the Δ14 mice is classified as both an alternative last exon and an alternative start site. The latter is a mis-annotation by *SGSeq* due to the poor mappability of the central sequence of exon 14, leading to a gap in reads that span both ends of the exon. The other 2 splicing events in *Fus* are in the middle of the transcript where two introns, 6 and 7, are retained more in the wildtype mice and less retained in the Δ14 mice. Two overlapping complex events are seen in the *Mbp* gene which encodes for Myelin basic protein, a major component of the myelin sheath. *Mbp* is alternatively spliced to create different isoforms (de Ferra et al., 1985). A complex event involving 4 cassette exons is subtly altered in Δ14. *Ewsr1* is a closely-related RNA-binding protein that interacts with FUS at the protein and RNA level (Kapeli et al., 2016; Lagier-Tourenne et al., 2012). A novel intron retention event in *Ewsr1* is retained less in FUS Δ14 mice. This may in fact be a differential polyadenylation event that has previously been observed (Rogelj et al., 2012) and mis-classified by the splicing software. The other splicing events found affect 4 other genes of unknown significance.

4.6 Discussion

My gene expression analysis demonstrates a progressive and tissue-specific change in RNA regulation, with 1,289 genes differentially expressed in the spinal cord of 12 month mutant mice. This finding is complemented by other contributions to the project. Behavioural experiments have shown the Δ14 mice to have a progressive loss of motor function that is observable from 12 months of age. In addition, a reduction in number of motor neurons in the lumbar spinal cord has been observed from 12 months onwards, accompanied by an increase

Table 4.2: Splicing events found in 12 month spinal cord samples All splicing events found comparing FUS Δ14 mice to wildtype at FDR < 0.05. log_2FC: log_2 fold change.

Gene	type	log_2FC	FDR	coordinates (mm10)
Fus	alternate last exon	−0.132	$1e^{-118}$	chr7:127981291-127981458
Fus	alternate start site	−0.413	$1e^{-116}$	chr7:127981522-127981791
Fus	retained intron	−0.217	$1e^{-13}$	chr7:127974434-127975888
Fus	retained intron	−0.191	$1e^{-6}$	chr7:127972770-127974400
Mbp	complex	−0.014	0.007	chr18:82575633-82584116
Mbp	complex	0.022	0.007	chr18:82575633-82584112
Ewsr1	retained intron	−0.066	0.014	chr11:5079485-5078952
Cyhr1	alternate 3´ splice site	0.127	0.028	chr15:76659955-76659439
Gm32856	alternate 3´ splice site	−0.317	0.038	chr8:129281397-129282476
Cd47	cassette exon	0.148	0.038	chr16:49906812-49910869
A330023F24Rik	retained intron	1.420	0.044	chr1:195021564-195021688

in cytoplasmic mislocalisation specifically of the mutant FUS. The broad downregulation of mitochondrial, ribosomal and proteasomal genes specifically in the spinal cord of late-stage mice is an interesting finding that deserves further investigation. A recent study of a mouse model where mouse FUS was entirely replaced by either wildtype or ALS mutant human FUS showed similar downregulation of ribosomal (but not mitochondrial) genes (López-Erauskin et al., 2018). An enrichment of mitochondrial GO categories was previously seen in a FUS knockdown experiment conducted in human embryonic kidney cells (Schwartz et al., 2012), suggesting that mitochondrial changes may be due to a loss of normal FUS function. Mitochondrial defects have been observed in the brains of FTD-FUS patients accompanied by FUS translocating to mitochondria (Deng et al., 2015). FUS overexpression has also been observed to cause mitochondrial defects at the neuromuscular junction (So et al., 2018). This points to an important role for FUS in mitochondria that may be perturbed by the Δ14 mutation. As mutant FUS has been shown to impair axonal transport (Guo et al., 2017), this impairment could explain the changes seen in both ribosomal and proteasomal transcripts.

FUS knockout and knockdown experiments have demonstrated that large numbers of splicing events are sensitive to FUS protein levels (Rogelj et al., 2012; Lagier-Tourenne et al., 2012; Ishigaki et al., 2012; Honda et al., 2014; Scekic-Zahirovic et al., 2016). The small number of splicing events found in the 12 month spinal cord samples suggest that the Δ14 mutation has little affect on splicing with the exception of the *Fus* transcript itself. The two *Fus* intron retention events seen to be less retained in Δ14 compared to wildtype as well as the events seen in *Mbp* and *Ewsr1* are deserving of further study. Defects in myelination have been seen in another FUS NLS mutation (Scekic-Zahirovic et al., 2017) and this may explain the changes seen in the myelin component *Mbp*. The fellow FET family member *Ewsr1* is also mis-spliced in FUS Δ14. As it is also an intron retention event which changes in the same direction as those seen in *Fus*, this could imply a connection between FUS and *Ewsr1* at the RNA level.

5 Loss and gain of TDP-43 splicing function in two mutant mouse lines

Work presented in this chapter has been published as part of (Fratta et al., 2018). See appendices for full reproduction of the published manuscript.

5.1 Overview

TDP-43 is a ubiquitously expressed RNA-binding protein with multiple roles in RNA processing including mRNA splicing. TDP-43 mislocalisation and aggregation is a common ALS pathology observed in the majority of patient brains. Additionally, rare TDP-43 mutations are causative for ALS. How TDP-43 mutations lead to disease and how TDP mislocalisation occurs in the absence of mutations is unknown. Whether loss of TDP-43 from the nucleus or a gain of TDP-43 in the cytoplasm is the pathological consequence is also under debate. TDP-43 mutations cluster in the low-complexity domain of the protein which has been implicated in aggregation and interaction with other proteins.

We generated two mutant mouse lines to study different aspects of the role of TDP-43 in mRNA splicing. By comparing the two we discovered that mutations in the low-complexity domain of the protein lead to a gain of splicing function. This is radically distinct from mutations that affect RNA-binding which act as a relatively simple loss of splicing function.

5.2 Contributions

- All RT-PCRs and Western blots presented were performed, quantified and plotted by Prasanth Sivakumar
- All RNA-seq and iCLIP sequencing libraries were created by Prasanth Sivakumar, DrAgnieszka Ule and Dr Pietro Fratta
- Mice were handled by Dr Thomas Ricketts and Dr Cristian Bodo
- Preliminary analysis of splicing was performed by Dr Warren Emmett and Kitty Lo
- I processed all RNA-seq data and performed all the splicing analyses bar Fig. 5.3, performed by Dr Kitty Lo, and figures 5.4 and 5.12, which were created by Prasanth Sivakumar.

5.3 Methods

RNA sequencing

Details on all RNA-seq datasets are presented in table 5.1, including a published dataset of TDP-43 knockdown in mouse adult striatum (Polymenidou et al., 2011). All RNA-seq libraries were polyA+ enriched.

Table 5.1: List of accessions

Tissue	Genotype	N	Read length	Range uniquely mapped reads
Embyonic fibroblasts	RRM2mut	3	50nt x 2	4-13M
	LCDmut	3	50nt x 2	10-13M
	TDP-43 shRNA	3	50nt x 2	7-12M
Embryonic head	RRM2mut	3	40nt x 2	26-48M
	LCDmut	3	40nt x 2	27-34M
Adult spinal cord	RRM2mut	4	75nt x 2	41-53M
	LCDmut	4	75nt x 2	45-58M
Embryonic Brain	RRM2mut	4	100nt x 2	31-36M
Adult striatum (**?**)	TDP-43 ASO	4	75bp x 1	35-60M

Data processing

All data pre-processing, quality control and alignment were done with the standard RNA-seq pipeline (see chapter 2)

Differential splicing

Three different generations and qualities of sequencing data were generated over the course of the study. Therefore the methods I used to measure splicing changes were tailored to each dataset. For the low depth and short read embryonic fibroblast and embryonic head samples, I used DEXSeq package (Anders et al., 2012) to estimate changes in differential exon usage of annotated exons only. Due the high depth and long read length of the RRM2mut embryonic brain and LCDmut adult spinal cord samples I used the SGSeq package (Goldstein et al., 2016). Although SGSeq will discover and classify more complex splicing events, I focussed solely on cassette exons for their ease of interpretation.

Annotation of splicing events

As TDP-43 depletion is associated with the splicing of non-annotated cryptic exons (Ling et al., 2015) I wanted to examine both mouse lines for novel splicing events. However, as transcript annotation progresses the number of novel splicing events will diminish over time. Instead I decided to classify splicing events by the levels of inclusion rather than annotation. For each exon, the percentage spliced in (PSI) was computed and the difference in mean PSI between mutants and controls (ΔPSI) was calculated. Exons were classified as extreme inclusion or cryptic exons if they show negligible inclusion in wildtype ($\text{PSI}_{control} < 5\%$) and

an increased ΔPSI ($> 5\%$). Extreme exon skipping events or "skiptic exons" occured where an exon that is apparently constitutive ($\text{PSI}_{control} > 95\%$) is then skipped in the mutants ($\Delta\text{PSI} < -5\%$).

iCLIP analyses

Analysis of high-throughput iCLIP libraries was conducted using the iCount pipeline, mapping reads to the mm10 build of the mouse genome. Only uniquely-mapped sense reads from each dataset were used. All peak calling and false discovery rate correction was carried out as described in (Huppertz et al., 2014; König et al., 2010). Peaks were then clustered together and the resulting clusters were used in all further analysis. RNA maps are a visualisation tool for examining the enrichment of a set of features within a set of RNA sequences at a nucleotide level (Ule et al., 2006). They are very effective at aggregating multiple genomic loci together to demonstrate position specific enrichments of features such as sequence motifs within RNA-protein interaction data (CLIP, iCLIP, eCLIP). I developed software that would overlap one set of genomic coordinates with another and transform the output of this intersection into a large matrix that could be normalised and then visualised. RNA maps were created for groups of cassette exons by quantifying per-nucleotide iCLIP coverage across the entire length of each parent intron that contained the splice sites of each cassette exon. To maximise potential coverage, I pooled together all iCLIP replicates created by the Fratta lab with TDP-43 iCLIP generated previously (Rogelj et al., 2012). Analysis was then restricted to 300nt around the parent intron splice sites and 300nt around the cassette exon splice sites. Per-nucleotide iCLIP coverage was defined as the number of overlaps with at least one iCLIP cluster at an individual nucleotide divided by the total number of exon sequences. The normalised iCLIP coverage distributions are presented with gaussian smoothing for aesthetic appeal. Due to variance in exon lengths, it was simply noted whether the exon overlapped with at least one iCLIP cluster and this is plotted as a proportion of all exons with a separate axis. For the cryptic and skiptic exons, The 20 exons with the greatest total coverage are plotted individually.

Long intron genes

To assess the relationship between intron length and differential expression, I found the longest intron in each gene using annotations from GENCODE mouse release 25 (Harrow et al., 2012) by writing a Python script (2.7.1) that parses the GENCODE GTF file. I converted unadjusted differential gene expression P-values from DESeq2 into Z-scores and give them the sign of the log_2 fold change. Genes were ordered by signed Z-score and binned into groups of 200. The plots present mean intron length and standard error of the mean for each group.

To assess the dependence between iCLIP coverage and intron length, total TDP-43 iCLIP coverage across the entire length of genes was calculated and normalised to give a per-nucleotide coverage proportion. Genes were divided into those contained introns >100kb (see above) and to whether they were upregulated or downregulated in the RRM2mut

compared to wildtype littermates. Coverage distributions were compared using a Mann-Whitney-Wilcoxon non-parametric test in R.

Permutation of splicing results

I permuted the sample order of the 4 wildtype and 4 LCDmut homozygotes 50 times to get all possible permutations and reran the splicing analysis for each comparison. Distributions of *P*-values are presented as quantile-quantile plots to visualise the inflation from the expected distribution under the null hypothesis of there being no difference between the two groups.

Functional analysis of extreme cassette splice events

Cassette exons and their parent introns were extracted from the SGSeq results. A per-nucleotide list of PhyloP conservation scores (Pollard et al., 2010) for the mouse aligned to 59 other vertebrates (mm10.60way.phyloP60way.bw) was downloaded from UCSC. Mean scores were calculated for each exon using bigWigSummary (UCSC). The extreme cassette exons were compared to all exons annotated in the GENCODE mouse release 25.

Cassette exon splicing can destabilise its host transcript with either its inclusion or exclusion leading to a downstream frameshift and the presence of premature termination codons (Lewis et al, 2003). To predict the functional consequences of exon inclusion or skipping on the host transcript a script was written in R that predicted the upstream and downstream exons that flank the extreme cassette exons using both GENCODE annotation and the spliced reads in from the aligned RNA-seq samples. If both flanking exons were predicted to be in the coding sequence then the exon sequences were concatenated with and without the central exon and translated *in silico* in the predicted codon frame of the upstream exon. If skipping or inclusion of the central exon caused a frameshift and/or a premature stop codon this was noted. To assess the correlation between the presence of an extreme cassette exon and changes in expression of its host gene, the proportion of genes that are significantly up- or downregulated at FDR $< 10\%$ was assessed in three sets: extreme cassette exons, non-extreme cassette exons and as a control, genes with no cassette splicing expressed at a level at or greater than the most lowly expressed extreme exon gene. The proportions of up- and downregulated genes were compared between the control genes and the two groups of cassette exon containing genes with a binomial test in R.

Statistical analyses

All differential expression results are significant at a Benjamini-Hochberg false discovery rate of 10%. All differential splicing results presented are significant at a false discovery rate of 1% unless specifically stated.

5.4 Results

A random mutagenesis screen produces two mutant mouse lines with point mutations in *Tardbp*

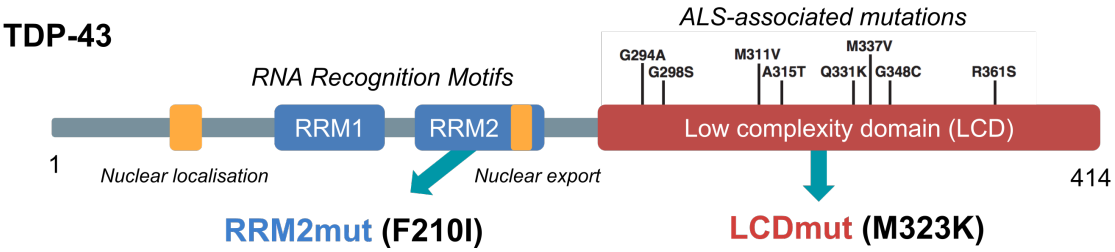


Figure 5.1: The two TARDBP mutations and their location within the TDP-43 protein. RRM2mut affects the second RNA recognition motif whereas LCDmut affects the C-terminal low complexity domain.

N-ethyl-N-nitrosourea (ENU) is a very potent mutagenic compound. Dosing male mice with ENU induces point mutations in sperm cells (De Angelis et al., 2000). Large banks of mutant mouse sperm are maintained at MRC Harwell (Acevedo-Arozena et al., 2008) and the RIKEN in Japan (Gondo et al., 2010). From these resources two mouse lines with mutations in *Tardbp* were chosen, F210I and M323K. The F210I mutation lies within the second RNA recognition motif (RRM2) of the TDP-43 protein whereas M323K lies within the C-terminal low complexity domain (LCD). This region is a hotspot for ALS mutations (Fig. 5.1). The M323K mutation lies within a 20 amino acid alpha-helical region previously found to be important for liquid phase separation and protein aggregation (Conicella et al., 2016). Developing these two mice allowed us to interrogate TDP-43 function and compare a mutation that would be predicted to impair the RNA binding ability of TDP-43 (F210I) with a mutation that potential resembles ALS (M323K). Due to their positions within the protein, the two mutations will be henceforth referred to as RRM2mut and LCDmut.

We derived mice from mutant sperm and backcrossed for 10 generations onto a mixed C57BL/6J - DBA/2J background to remove unwanted background mutations. RRM2mut is embryonic lethal in the homozygous state but not in heterozygosity, whereas homozygous LCDmut mice are viable and live normal lifespans. The two mutant lines were crossed together to create compound heterozygotes. The RRM2mut/LCDmut mice were viable, suggesting the two mutations complement each other. As TDP-43 is known to form oligomers (Afroz et al., 2017) this may be a case of the two mutant proteins balancing each other when forming heterodimers. Close observation of aged LCDmut revealed gradual muscle weakness and a reduction in motor neuron numbers in the spinal cord (data not shown), suggesting that the patient-like M323K mutation indeed causes symptoms of neurodegeneration reminiscent of ALS. Due to the extreme differences in phenotype, particularly the neurodegeneration seen in LCDmut adults, I was curious to explore the effect of the mutations on RNA splicing.

The two mutations have opposing effects on splicing known TDP-43 targets

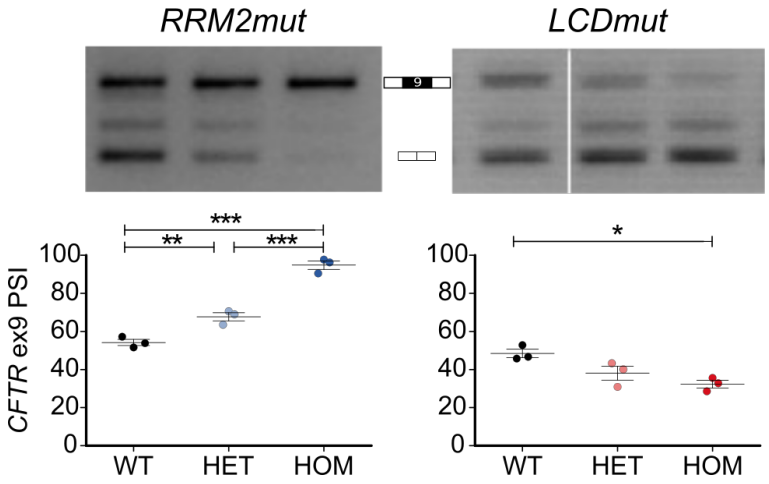


Figure 5.2: Opposite effects on splicing CFTR exon 9 minigene, a known TDP-43 splicing target. RT-PCR traces from primers flanking CFTR exon 9, quantified as PSI ratios for each genotype. *** $P < 0.0001$ (ANOVA); ** $P < 0.01$; *** $P < 0.001$ (Tukey post-hoc test).

Exon 9 of the *CFTR* gene was the first mRNA splicing target of TDP-43 to be described in the literature (Buratti and Baralle, 2001a). Knocking down TDP-43 leads to increased exon 9 inclusion, suggesting that TDP-43 acts to promote exon skipping. We made use of a minigene construct created from *CFTR* exon 9 and its two flanking introns and exons (Buratti et al., 2007b). Reverse-Transcriptase Polymerase Chain Reaction (RT-PCR) was performed to amplify between primers that flank exon 9. Analysis of the gel electrophoresis traces shows two primary bands: a larger band corresponding to exon 9 inclusion and a smaller band corresponding to exon 9 skipping. RRM2mut has a clear dose-dependent increase in exon 9 inclusion compared to skipping and thus resembles a loss of TDP-43 splicing function (Fig. 5.2). Conversely, LCDmut has a dose-dependent increase in exon 9 skipping. This occured despite no differences in protein levels of the two mutant proteins from wildtype TDP-43 being detected by western blotting (Fig. 5.12). This suggests that the pro-skipping action of TDP-43 is increased at the CFTR locus and the LCDmut mutation causes a gain of splicing function.

To look transcriptome-wide at the effects of the two mutations on simple cassette exon splicing the lab generated low-depth RNA-seq data from mouse embryonic fibroblasts. To compare both mutations to a simple loss of TDP-43 a short hairpin RNA (shRNA) was designed to be complementary to *Tardbp* mRNA. Binding of the shRNA should target *Tardbp* mRNA for degradation, reducing TDP-43 protein levels. I performed a cassette exon splicing analysis and converted the fold change and *P*-value from each exon into a signed Z-score. This allows for two way comparisons between the shRNA knockdown of TDP-43 (TDP-KD), LCDmut and RRM2mut. Separating each graph into quadrants makes it clear that all significantly changed exons found in each mutation are changed in the same direction between RRM2mut and TDP-KD (Fig. 5.3). However, the exon changes are opposing when comparing LCDmut and TDP-KD, as well as between LCDmut and

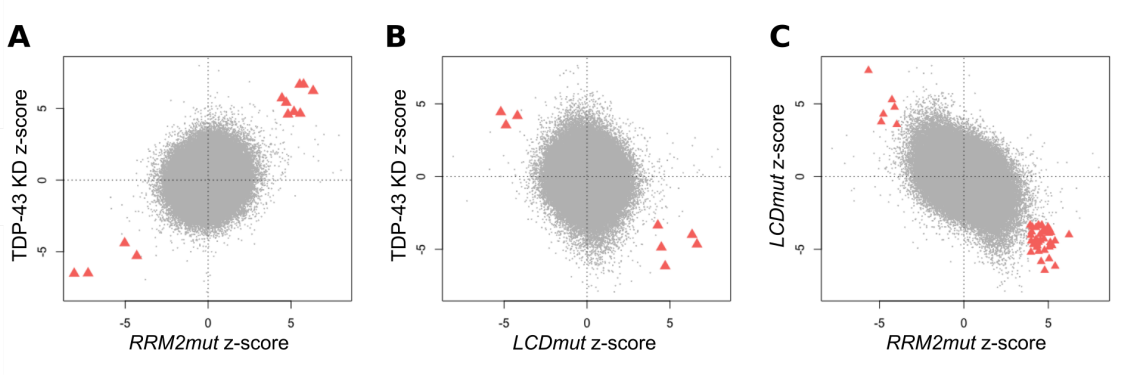


Figure 5.3: Comparing exon usage between the two mutations and a TDP-43 knockdown. Signed Z-scores for all exons found by DEXSeq in the mouse embryonic fibroblasts, comparing a TDP-43 shRNA knockdown to RRM2mut **(A)** and LCDmut **(B)** and comparing the two mutations **(C)**. Exons significant at FDR $< 10\%$ plotted as red triangles, non-significant exons plotted as grey dots.

RRM2mut . This provides evidence at transcriptome scale that RRM2mut is a loss of splicing function whereas LCDmut is behaving oppositely.

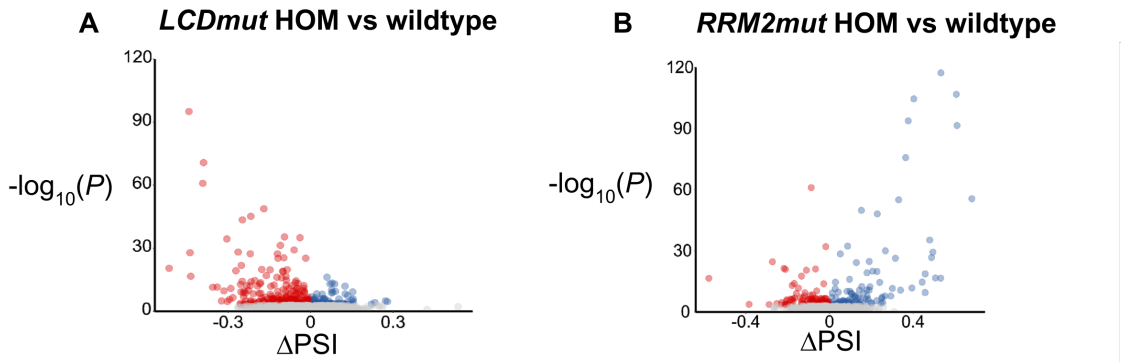


Figure 5.4: Direction of cassette exon splicing in the two mutant lines. All splicing events found by SGSeq plotted by P-value and direction (ΔPSI; see methods) for LCDmut **(A)** and RRM2mut **(B)**.

To deeply interrogate splicing changes at time points that were relevant to the phenotypes of interest (death in RRM2mut and neurodegeneration in LCDmut) high depth RNA-seq data was generated from RRM2mut embryonic brains and LCDmut adult brains (30 days post-natal). Higher depth and longer read data allows for better discrimination of novel splicing events and so I ran a novel splicing analysis using SGSeq (Goldstein et al., 2016). Plotting the direction and P-value of all cassette exons found in the two mutations strongly suggests that the strongest splicing changes are in exon inclusion in RRM2mut and in exon skipping in LCDmut (Fig. 5.4).

The splice-site competition model of cassette exon splicing allow an RNA-binding protein to either repress or enhance cassette exon splicing depending on the position it binds within the intron. When RNA binding proteins bind on top of or close to an exon, the recognition of that exon's splice sites by the spliceosome is blocked and the exon is no longer included in a transcript. Conversely, if an RNA binding protein binds deeper within the intron it

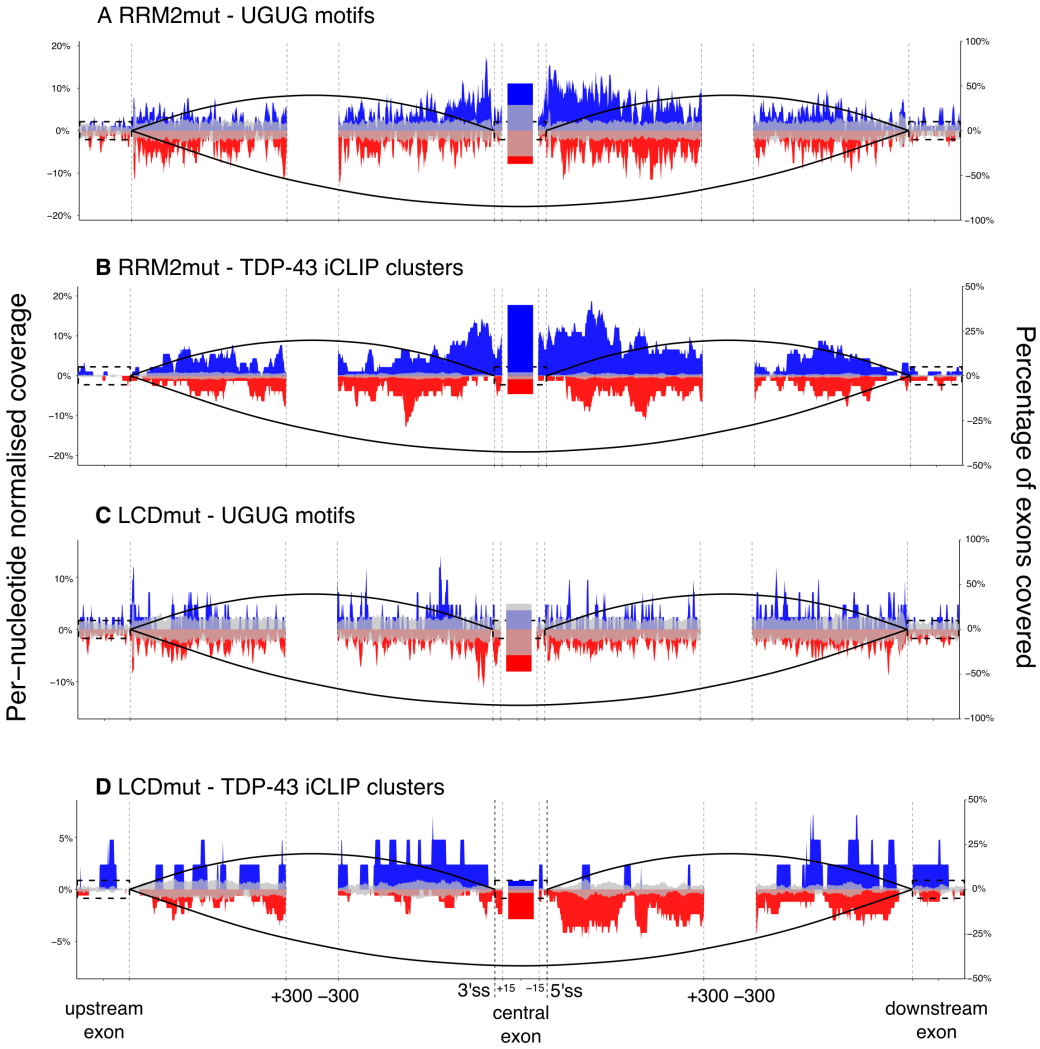


Figure 5.5: RNA maps visualise positional enrichment of iCLIP peaks and sequence motifs within cassette exon splicing loci. RNA maps constructed from differentially included (blue) and skipped (red) cassette exons. **(A)** TGTG motifs in RRM2mut. **(B)** TDP-43 iCLIP in RRM2mut. **(C)** TGTG motifs in LCDmut; **(D)** TDP-43 iCLIP in RRM2mut.

can act to recruit the spliceosome to the exon and enhance exon inclusion. I created RNA maps to look for positional enrichment within and around the the cassette exons, both for UGUG sequence motifs as well as RNA-protein interaction information from iCLIP experiments. For each set of exons I used a random set of 1994 cassette exons which were not significantly changed as a null distritbution. Motif-based maps were calibrated using the invariant AG and GT dinucleotides at the 3´ and 5´ splice sites respectively (see appendices). For RRM2mut, the 91 cassette exons with increased inclusion were enriched in UGUG sequence motifs (Fig 5.5A) and iCLIP peaks (Fig 5.5B). Both sequence features either directly overlap the exon or are within 100 bp either side. The 78 skipped exons in RRM2mut were depleted in both feature types directly on top and close by. LCDmut cassette exons show the inverse distribution to RRM2mut as the 168 skipped exons are enriched for features that directly overlap and closely flank the exons (Fig 5.5C/D), whereas the 42 included exons were depleted of sequence features that were close or overlapping. All exons howed enrichment of TDP-43 features at the distal flanking 5´ and 3´ splice sites,

suggesting long range cooperation around the distal splice sites. This data suggests that whereas the RRM2mut cassette exons are shifted due to a reduction in TDP-43, the LCDmut exons are shifting in a direction suggesting an increase of TDP-43 in the mutant cells.

RRM2mut leads to a loss of splicing function and cryptic exons

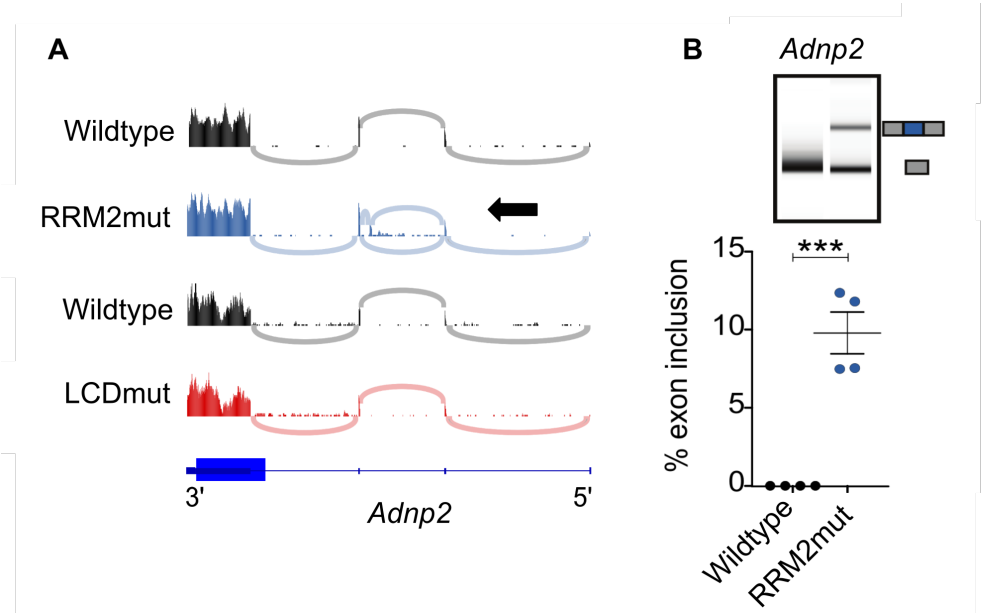


Figure 5.6: RRM2mut leads to cryptic exon splicing. **(A)** Representative RNA-seq traces show cryptic exon in *Adnp2* included in RRM2mut specifically. **(B)** RT-PCR of *Adnp2* selectively amplifies a band corresponding to cryptic exon inclusion in RRM2mut samples. P < 0.001; t-test(two-sided).

A hallmark of TDP-43 depletion is the widespread inclusion of non-conserved cryptic exons (Ling et al., 2015). The bias towards cassette exon inclusion suggested that a number of RRM2mut repressed exons may be cryptic exons. Rather than relying on isoform annotation I filtered all cassette exons found in RRM2mut and selected those that were barely or not at all spliced in wildtypes but were included in RRM2mut samples, resulting in 33 candidate cryptic exons being discovered. A representative example of a cryptic exon in *Adnp2* is shown in Fig. 5.6, changing from 0% inclusion in wildtype mice to 10% inclusion in RRM2mut but not in LCDmut mice.

Another feature of TDP-43 loss of function is a striking downregulation of genes with long introns (>100kb). This phenomenon was first observed in mice where an antisense oligonucleotide strategy was used to lower TDP-43 in the striatum (TDP-ASO); (**?**). Long intron genes are over-represented in neuronal cells (Sibley et al., 2015) and it is thought that TDP-43 binds within these long introns to promote their processing and splicing. I re-analysed RNA-seq data from this study and compared it to the RRM2mut and LCDmut sequencing data. I ran a differential gene expression analysis with DESeq2 and ranked all genes by their direction of expression change between controls and ASO treatment/mutations. I then binned genes into groups of 200 and extracted the lengths of longest intron in each gene from GENCODE annotation. While in LCDmut, gene length is evenly distributed

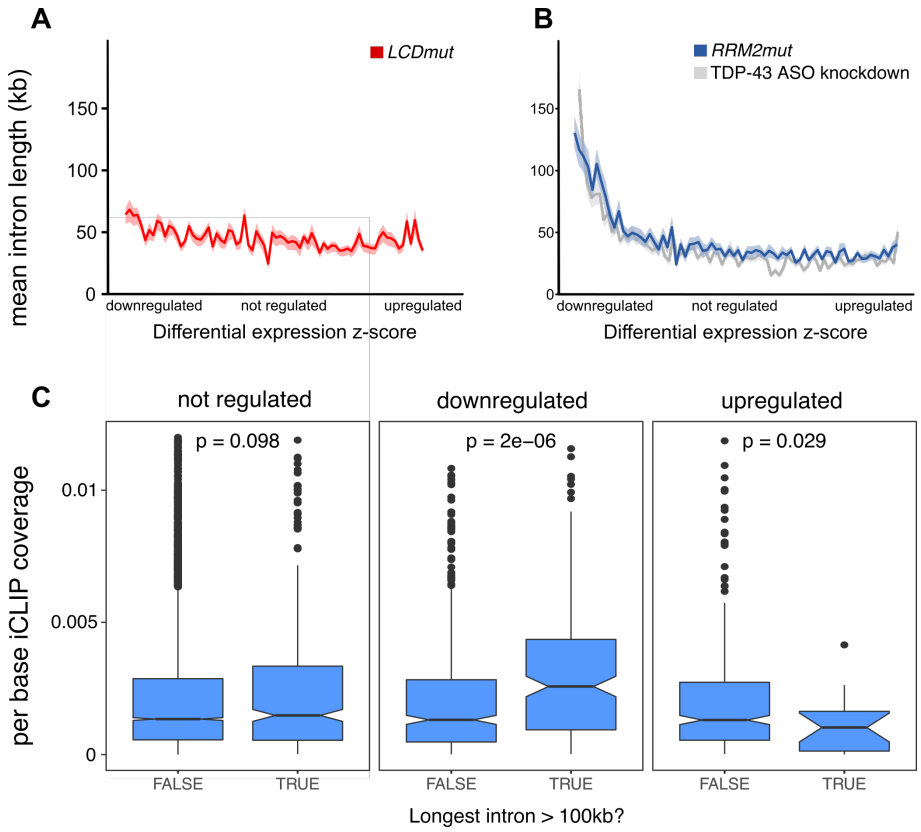


Figure 5.7: RRM2mut gene expression has bias for long gene downregulation, mirroring TDP-43 knockdown data. (A,B) Mean intron length in kilobases for bins of genes ordered by signed Z-score. TDP-43 antisense oligonucleotide (ASO) knockdown data taken from Polymenidou et al. (2011). **(C)** The per-nucleotide overlap of iCLIP interaction peaks for each gene is significantly increased in long intron genes downregulated in RRM2mut mice compared to wildtype. *P*-values are from Mann-Whitney-Wilcoxon test.

between downregulated, unchanged and upregulated genes (Fig. 5.7A), there is a clear bias for the most downregulated genes having long introns in both RRM2mut and TDP-ASO (Fig. 5.7B). An orthogonal approach is to look at TDP-43 protein-RNA interaction data performed on wildtype cells with the iCLIP method (Huppertz et al., 2014). I calculated the proportion of nucleotides in each gene that had an iCLIP peak overlapping, suggesting direct TDP-43 binding. Genes that were downregulated in RRM2mut had no difference in the distribution of iCLIP peak overlaps except for those downregulated genes that also contained at least one intron longer than 100 kilobases (P=2e-6; Fig. 5.7C). Long intron genes were modestly depleted in iCLIP peaks when the genes were upregulated (P=0.029).

LCDmut shows the inverse of cryptic splicing - skiptic splicing

I applied the same cryptic exon filtering strategy to LCDmut and found a small number of potential cryptic exons. However, when applying the inverse critera to find exons that are 95-100% included in wildtype and then skipped in LCDmut I uncovered 48 exons. These exons are constitutively spliced in wildtype samples and yet are skipped in LCDmut, making them the inverse of cryptic exons. I therefore christened them "skiptic" exons - a portmanteau of cryptic and skipping. A selection of skiptic exons were validated by RT-PCR (Fig. 5.8).

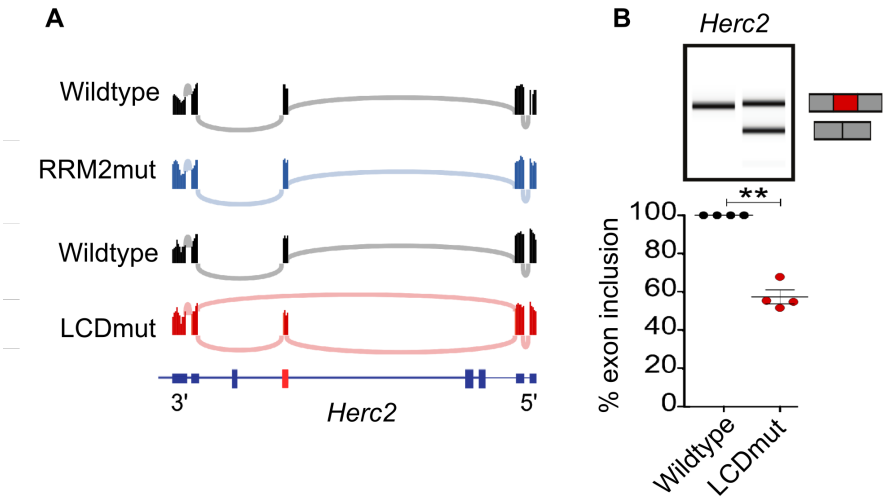


Figure 5.8: LCDmut leads to skiptic exon splicing. **(A)** Representative RNA-seq traces show a constitutive exon in *Herc2* skipped in LCDmut specifically - a skiptic exon. **(B)** RT-PCR of *Herc2* selectively amplifies a band corresponding to exon skipping skipping in LCDmut samples. P $<$ 0.001; t-test(two-sided). **(C)** PhyloP conservation scores for 1000 randomly chosen mouse exons compared to the 48 skiptic exons found in LCDmut.

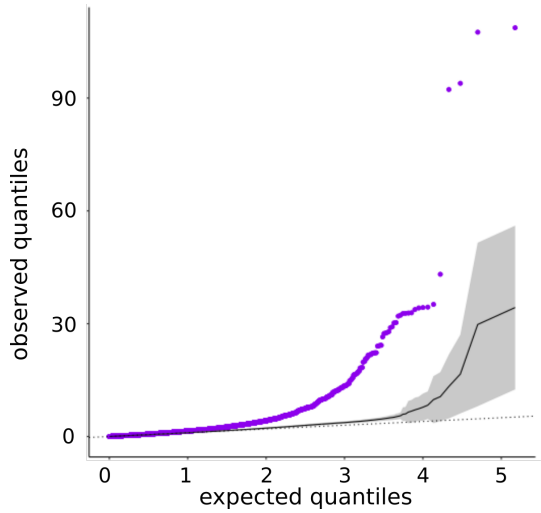


Figure 5.9: Permuting sample order shows clear splicing difference between LCDmut mice and controls. Quantile-quantile plots show the difference between expected and observed distribution of *P*-values generated from multiple tests. True case-control sample labelling of LCDmut and littermate controls (purple) shows clear inflation of low *P*-values when compared to all permutations of sample ordering (grey; plotted as mean ± standard deviation).

To demonstrate that the 48 skiptic exons found in LCDmut were not statistical anomalies due to the small sample size I employed a permutation strategy. The wildtype and LCDmut labels were shuffled 50 times to cover all possible sample orders and the splicing analysis repeated with the permuted labels. A quantile-quantile plot contrasts the observed distribution of *P*-values generated by a large number of statistical tests with the theoretical expected distribution. If there is no difference between genotypes then some low *P*-values are expected by chance due to the small sample size and the large number of tests. However, there is a clear difference in splicing between LCDmut and wildtype mice that drives a large inflation in the number of low *P*-value splicing events far away from the expected

distribution (Fig. 5.9). A table of counts of all exons found at each permutation is presented in the appendices.

Both skiptic and cryptic splicing show evidence of direct TDP-43 binding

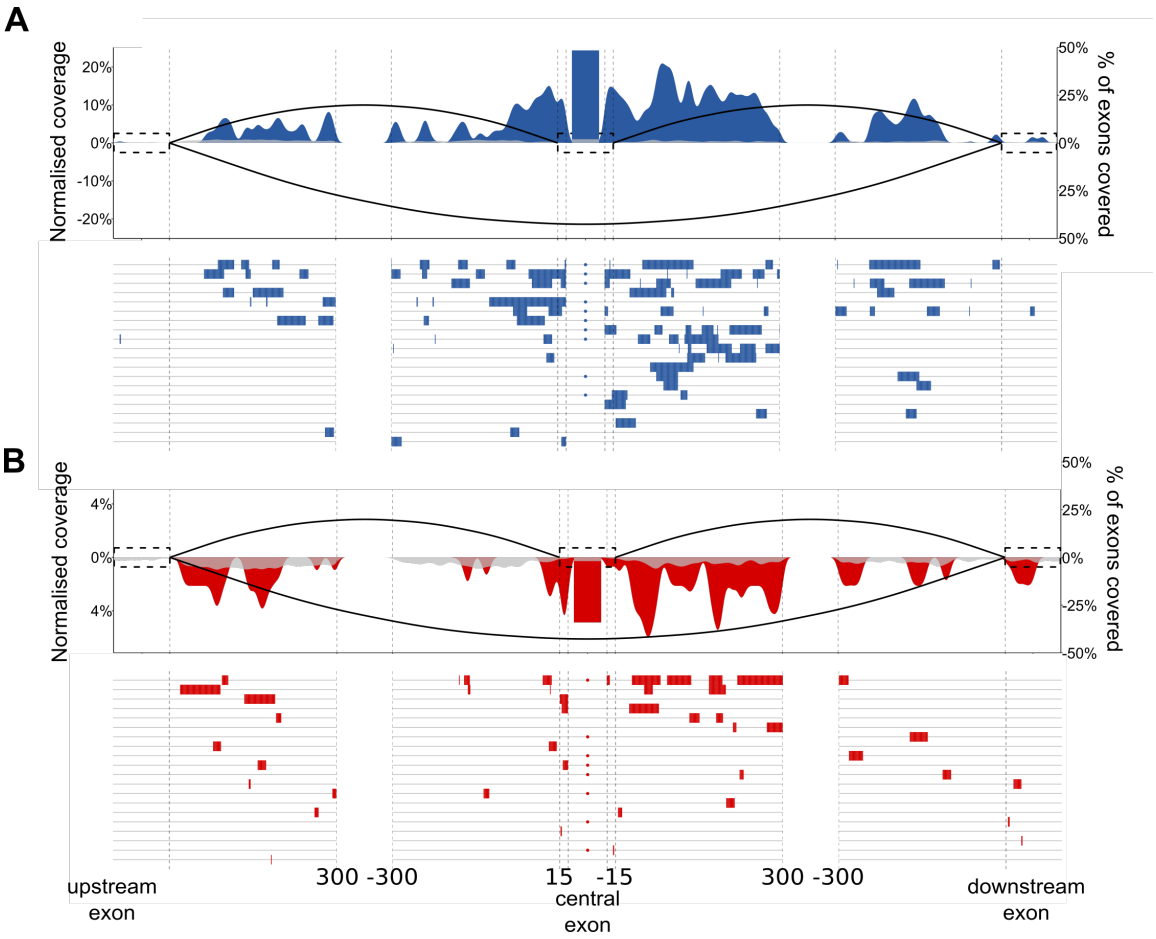


Figure 5.10: RNA maps of skiptic and cryptic exons show direct binding by TDP-43. (A) The 33 cryptic exons found in the RRM2mut embryonic mice. Traces show normalised iCLIP peak coverage within and around each exon. Left y-axis: the proportion of exons with an iCLIP peak at that nucleotide. Right y axis: the proportion of central exons or flanking exons that overlap any iCLIP peaks. Below, individual positions of iCLIP peaks for the top 20 cryptic exons. Circles in the centre denote whether there are any iCLIP peaks overlapping the central exon. **(B)** as before for the 48 skiptic exons found in the LCDmut adult mice.

Cryptic exons associated with TDP-43 depletion have been demonstrated to originate from mRNA that is directly bound by TDP-43 itself (Ling et al., 2015). This suggests that these splicing changes emerge because TDP-43 can no longer act to repress cryptic exon recognition by the splicing machinery and other factors. I remade the RNA maps for iCLIP protein-RNA interaction peaks for the 33 cryptic and 48 skiptic exons found in RRM2mut and LCDmut to test whether they appeared to be directly bound by TDP-43. The RRM2mut cryptic exons show overlapping and closely flanking TDP-43 binding and strikingly so do the LCDmut cryptic exons. Importantly, the iCLIP data used for these maps is mainly drawn from wildtype TDP-43, suggesting that while TDP-43 may be normally binding the skiptic exons it does not do so sufficiently strongly to repress their inclusion.

	total	overlap	%
All exons in GENCODE vM12	744,786	37,276	5%
All constitutive exons found in all samples	239,897	17,828	7.4%
All cassette exons found in all samples	5,656	361	6.4%
Significant cassette exons (FDR < 0.05) between LCDmut and wildtype mice	260	49	18.8%
Skiptic exons (control PSI > = 0.95; dPSI > = -0.05; FDR < 0.05)	47	31	66%

Table 5.2: Proportions of exons with any TDP-43 binding from iCLIP

Skiptic splicing is predicted to be deleterious to gene expression

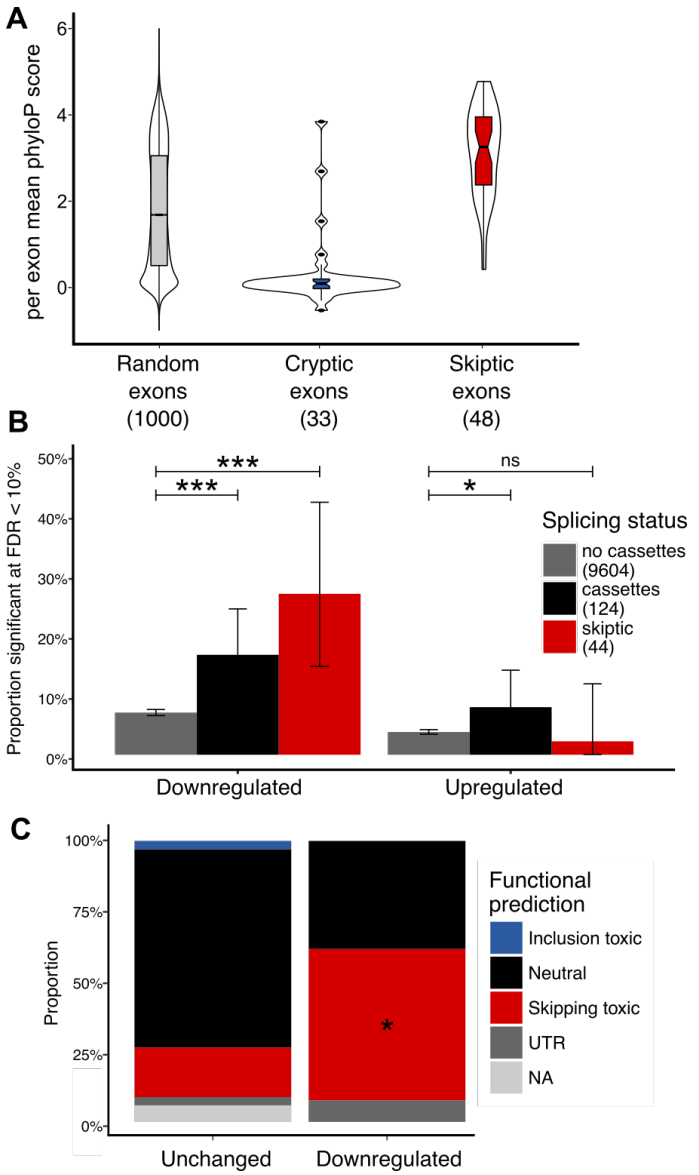


Figure 5.11: Functional analyses of the skiptic exons. **(A)** Per-exon mean PhyloP conservation scores presented as box plots for each set of exon, showing the median and interquartile range. Notches are the 95% confidence interval of the median. **(B)** The relationship between splicing status and differential expression in LCDmut. Significantly downregulated genes (FDR < 0.1) are enriched in skiptic exons compared to genes expressed at a similar level (P=1.19e-6), as well as non-skiptic cassette exons (P=6.16e-5). Upregulated genes are mildly enriched in non-skiptic cassette exons (P=0.029) but not in skiptic exons (P=0.88). All P-values generated from a binomial test. **(C)** Proportions of predicted downstream consequence of exon skipping for either non-regulated or downregulated skiptic exons. P=0.034; chi-squared test.

Cryptic exons originate from very poorly conserved DNA sequence (Fig. 5.11A), suggesting that there is no functional protein coding information encoded in them. Inclusion of these essentially randomised sequences would more likely than not lead to degradation of the host mRNA transcript through the inclusion of premature stop codons or frameshifts and nonsense-mediated decay. Skiptic exons are highly conserved (Fig. 5.11A), with a much higher average conservation level than a randomly chosen set of annotated exons from GENCODE. This is to be expected from their close to 100% inclusion rates within transcripts. Their constitutive splicing status means that their inclusion is near guaranteed and so unlike cassette exons, which are swapped in and out dynamically between tissues and developmental time points, I hypothesis that there is no evolutionary pressure to maintain a length divisible by 3. In the event of these exons being skipped they would likely lead to a shift of reading frame, potentially leading to degradation of the host transcript through nonsense-mediated decay. To test whether skiptic exon splicing correlated with host transcript degradation I combined the differential splicing and differential expression analyses together. I separated genes into sets based on i) whether they were significantly upregulated or downregulated in LCDmut compared to wildtype (FDR $< 10\%$) and ii) whether they contained either a cassette exon or a skiptic exon that changed in inclusion levels between LCDmut and wildtype (FDR $< 1\%$). Genes that contained skiptic exons were more likely to be downregulated than those without any cassette exon splicing (P=1.19e-6; binomial test; Fig. 5.11B). The same trend cannot be seen in the other direction as there as no increased likelihood for skiptic containing genes to be upregulated. Finally I attempted to predict *in silico* whether the skipping of a skiptic exon would lead to nonsense mediated decay. By combinining the central cassette exon with the flanking upstream exons and translating the concatenated sequence it is possible to assess whether skipping the central exon would frameshift the downstream exon, generating premature stop codons. Using this method I discovered that 15 out of 47 of the skiptic exons would lead to the inclusion of a premature stop codon. I compared the distribution of predictions between skiptic exons in downregulated and unchanged genes and found an increased representation of predictions of toxic skipping from 18% to 54% (P=0.034; chi-squared test).

LCDmut impairs TDP-43 autoregulation

Many RNA-binding proteins have been shown to regulate their own expression. This is termed autoregulation (Lareau et al., 2007; Wollerton et al., 2004)]. When levels of the protein are high they will bind to their mRNA and shift production to an untranslated isoform, through nonsense-mediated decay (Losson and Lacroute, 1979; McGlincy and Smith, 2008) or through nuclear retention (Boutz et al., 2015). The 3´ untranslated region (3´ UTR) of *Tardbp* is remarkably complex regulatory hub (Fig. 5.12A). There are 3 experimentally validated polyadenylation and cleavage sites, pA1, 2, and 4. There also 2 introns within the 3´ UTR (6 and 7; (Ayala et al., 2011; Koyama et al., 2016)). 3´ UTR splicing is a mechanism for regulating TDP-43 translation as an intron that follows a stop codon should trigger nonsense-mediated decay (Losson and Lacroute, 1979) . Both spliced isoforms pA_2* and pA_4* are predicted to undergo nonsense-mediated decay due to the stop codon

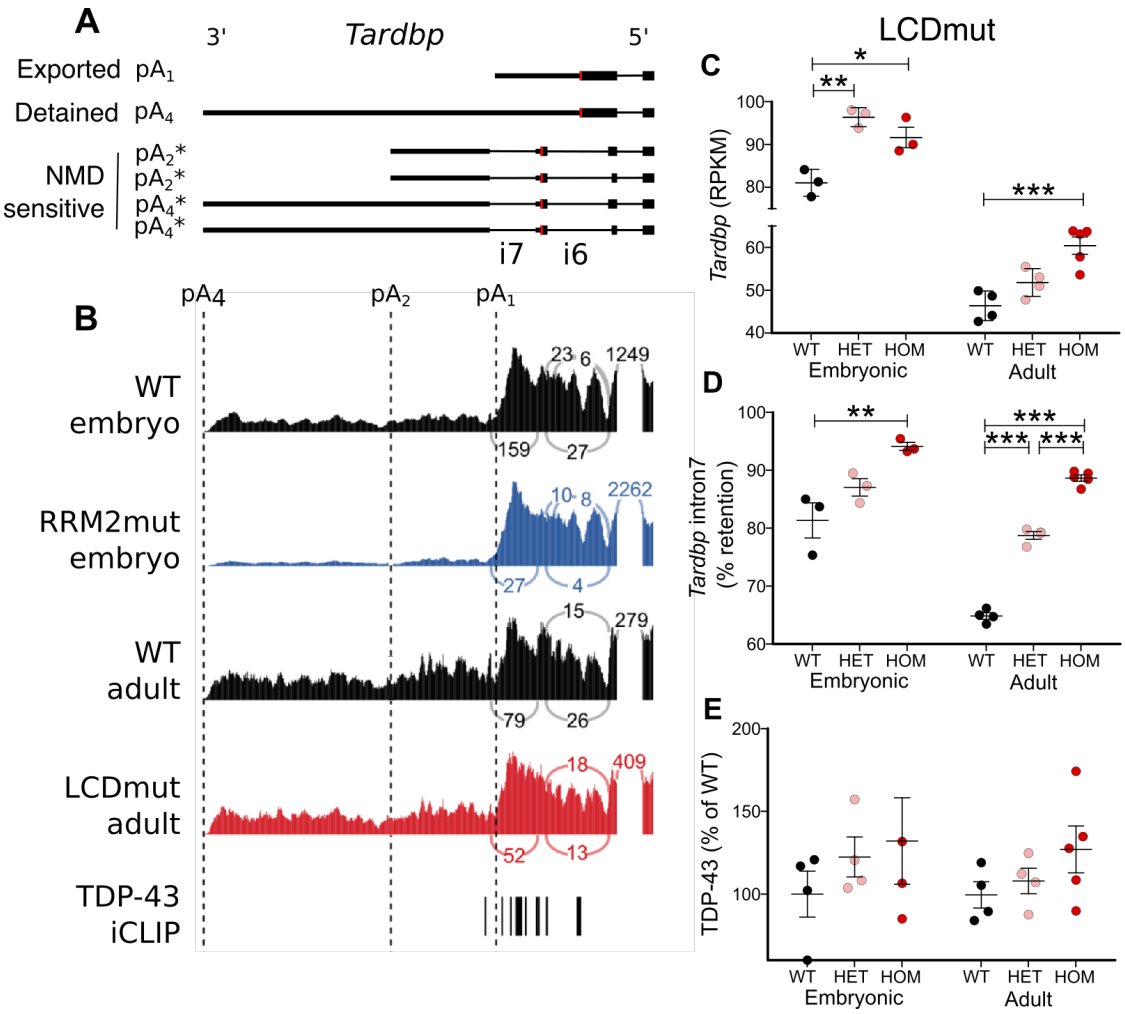


Figure 5.12: LCDmut and TDP-43 autoregulation. **(A)** The 3´ UTR of *Tardbp*. At least six potential 3´ UTR isoforms have been proposed by (Koyama et al., 2016), consisting of 3 different polyadenylation (pA) sites. * indicates that this isoform is predicted to be degraded by nonsense-mediated decay. Stop codons indicated by red bars. Untranslated sequence represented by thinner bars, coding sequence with thicker bars. **(B)** Sashimi plots show representative RNA-seq read coverage and splice junction counts from RRM2mut and LCDmut mice with their respective controls. TDP-43 iCLIP data from (Rogelj et al., 2012) is presented as individual peaks. **(B)** Expression of all *Tardbp* isoforms (RPKM) increases in LCDmut at two time points. Embryonic dataset ANOVA, P=0.0032, n=3; adult dataset ANOVA, P=0.0009, n=4–5; error bars: SD; Bonferroni multiple comparison tests are plotted as P-value: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ **(C)** Retention of 3´ UTR intron 7 increases in LCDmut. Embryonic dataset ANOVA, P=0.0113, n=3; adult dataset ANOVA, $P < 0.0001$, $n = 4$–5; error bars: SEM; Bonferroni multiple comparison tests are plotted as P-value: $^{**}P < 0.01$; $^{***}P < 0.001$. **(D)** Quantification of TDP-43 protein levels in relation to β-actin in Western blots. Results are normalised to the mean of wildtype (100%). Embryonic dataset: ANOVA P=0.480, n=4; adult dataset: ANOVA P=0.491, n=3; error bars: SD.

is further than 50bp from the intron in each (Nagy and Maquat, 1998). Additionally, the long pA4 site is preferentially retained in the nucleus where it is degraded by the exosomal complex (Ayala et al., 2011). Only the pA1 isoform is translated into TARDBP protein (Koyama et al., 2016). TDP-43 binds the 3´ UTR of its own mRNA overlapping the pA_1 site (Polymenidou et al., 2011; Tollervey et al., 2011). This prevents polyadenylation at the pA1 site and encourages the creation of the retained pA4 transcript and the splicing of NMD-sensitive pA_2* and pA_4* isoforms (Koyama et al., 2016). This system allows TDP-43

protein levels to regulate the stabilty of *Tardbp* mRNA.

I assessed the the *Tardbp* 3´ UTR locus in the RNA-seq data to observe whether the two mutations had effects on autoregulation. RRM2mut shifts the balance of UTR isoforms from near equal amounts of pA_1 and pA_4 to predominantly pA_1, presumably due to its reduced RNA binding ability (Fig. 5.12A). In LCDmut, a clear upregulation of Tardbp mRNA expression is seen in both embryonic and adult RNA-seq samples (Fig. 5.12B). When I quantified the level of intron 7 inclusion, a proxy for the proportion of NMD-sensitive pA_2* and pA_4* isoforms, LCDmut increased intron 7 retention in a dose dependent manner, suggesting a loss of NMD-sensitive UTR isoforms (Fig. 5.12C). However, when assessing the total TDP-43 protein levels by Western blotting, no difference was observed between LCDmut and wildtype cells (Fig. 5.12D). Together this suggests a subtle impairment in the autoregulation feedback loop in LCDmut cells.

Skiptic splicing can be observed in human ALS patients

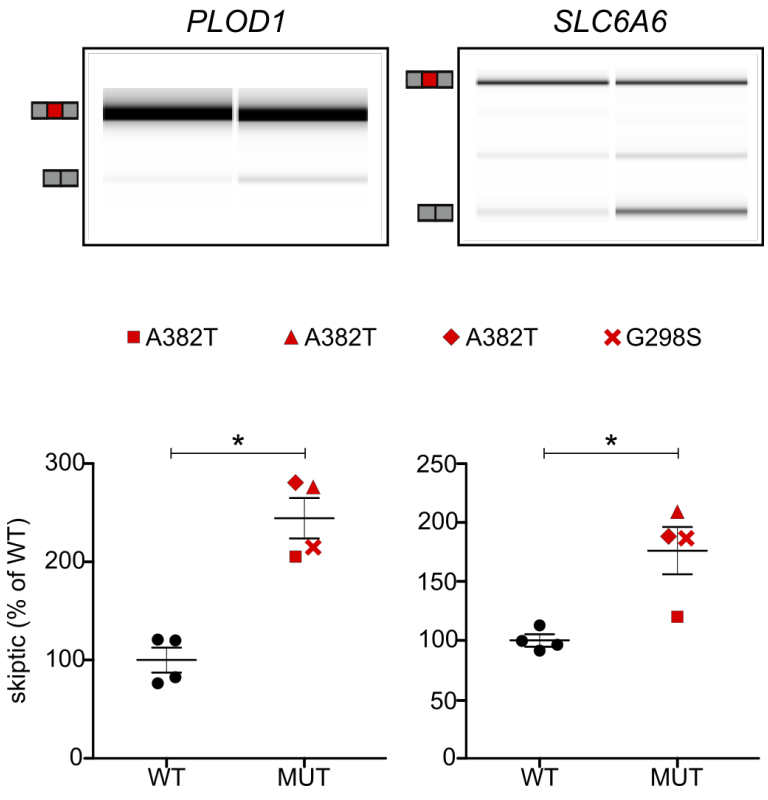


Figure 5.13: Skiptic exon splicing in TARDBP ALS. RT-PCR traces for two skiptic exons in fibroblasts taken from 4 *TARDBP* ALS patients and 4 non-neurological control lines. For clarity, quantification was taken of the skiptic exon band only instead of the ratio. $*P < 0.05$, $n = 4$, error bars: SEM

Due to the high species conservation of the skiptic exons, I reasoned that it might be possible to see evidence of their skipping in human ALS patients with heterozygous mutations in the low complexity region of *TARDBP*. Seven skiptic exons were tested using RT-PCR in

four patients with the A382T and G298S mutations and four control fibroblast lines. Two skiptic exons were found to have a mild but statistically signficant increase in the intensity of the band matching the skipping transcript. This provides the first evidence of a gain of TDP-43 splicing function in patients with ALS-associated *TARDBP* mutations.

5.5 Discussion

In this study two complementary mouse models were developed to study the splicing function of TDP-43. By using point mutations I have been able to observe the modulation of TDP-43-regulated splicing without dramatically altering TDP-43 expression. This has been particularly useful when studying RRM2mut. TDP-43 is very highly expressed during embryonic development and knockout mice die at E6.5 (Ricketts et al., 2014)) whereas RRM2mut mice survive until E18.5, allowing greater scope for studying TDP-43 loss in development. We have conclusively proven RRM2mut to be a loss-of-function mutation through our experiments on known TDP-43 splicing targets and also by looking transcriptome-wide. This is also shown by the observation of widespread cryptic exon inclusion and the downregulation of long intron genes, two clear molecular phenotypes of TDP-43 loss.

LCDmut is an intriguing mutation as I have discovered a completely novel and unexpected gain of splicing function. RNA maps demonstrated that the splicing targets altered in LCDmut are enriched in TDP-43 binding through iCLIP data, even when using iCLIP from wildtype TDP-43. This suggests a shift from TDP-43 binding from passive to active regulation at these loci.

Although in the fibroblast data there are a small number of overlapping exons that are spliced in opposing directions, the majority of exons altered in LCDmut are not changed in RRM2mut. This suggests that regulation of these exons is not simply bi-directional, with LCDmut increasing TDP-43 to cause skipping of an exon at sites where its loss causes inclusion.

One hypothetical mechanism for the gain of splicing function seen in LCDmut is through an impairment of TDP-43 autoregulation which would increase TDP-43 protein levels. The splicing of intron 7 decreases in a dose-dependent manner with LCDmut, suggesting a reduction in the creation of NMD-sensitive 3´ UTR isoforms. It is curious that this is reflected with a clear increase of *Tardbp* mRNA but not at the protein level. This was also seen in mice heterozygous for a TDP-43 null allele (Ricketts et al., 2014). This may be due reduced sensitivity of western blotting when compared to RNA-seq, or that the Western blotting was carried out on total cellular protein rather than fractionated by cellular compartment. A recent paper on a similar TDP-43 LCD mutant mouse (Q331K; White et al. (2018)) observed an increase in *Tardbp* mRNA levels which was accompanied by an increase in *nuclear but not cytoplasmic* TDP-43 protein levels. Our assay (Fig. 5.12D) looked at total TDP-43 protein levels so this may explain why no difference was seen. I cannot rule out the possibility of changes in mRNA or protein localisation, or altered regulation by microRNAs that may also complicate the picture.

The most unexpected finding in LCDmut is the discovery of skipped constitutive exons: skiptic exons. Previously TDP-43 had only been studied in the context of alternate cassette exons. Whereas cryptic exons emerge from normally repressed sections of introns that contain strong splice sites, skiptic exons appear to be an over-correction by TDP-43. Focusing our RNA maps on these skiptic exons shows that both cryptic and skiptic exons are simi-

larly enriched by TDP-43 binding in wildtype cells. Therefore TDP-43 may have a minor role in supporting the inclusion of these exons but with the LCDmut mutation this shifts to promoting their excision from the host transcript. That I see only 44 genes affected by skiptic transcripts may be down to the sheer redundancy and collaboration between multiple RNA-binding proteins. The skiptic exons may represent transcripts that are most sensitive to TDP-43 expression for their correct splicing, but why they do not also show changes in TDP-43 loss is mysterious.

Cryptic exons show a higher enrichment for TDP-43 iCLIP peaks than the skiptic exons. This is potentially an artefact of comparing iCLIP from predominantly embryonic tissue with splicing changes from adult mice but it could point to an alternate mechanism to explain the gain of splicing function. Low-complexity domains are crucial for assembling RNA-binding proteins into structures Gueroussov et al. (2017) and its possible that the LCDmut mutation shifts TDP-43 into assembling more strongly with certain groups of proteins to play a stronger role in splicing regulation at certain loci. Experiments to determine the protein-protein "interactome" of TDP-43 have been published (Freibaum et al., 2010) and it would be intriguing to compare LCDmut to wildtype TDP-43.

Recently a study investigated a similar model of TDP-43 ALS where the ALS patient mutation Q331K was knocked in to mice (White et al., 2018). They reported a similar gain of splicing function in the homozygous mice, although they looked at a small number of individual splicing events and not transcriptome-wide.

These results are complementary to ours and provide parallel evidence that low-complexity mutations in TDP-43 can alter the splicing function of the protein, potentially through altering autoregulation. It would be interesting to see whether the skiptic exons observed in adult LCDmut spinal cord can be detected in the now published Q331K frontal cortex.

My work on the long intron genes and the RNA maps raises interesting questions about the role of TDP-43 in splicing beyond merely binding on top of or closely flanking splice sites. The cassette exons altered in either RRM2mut or LCDmut show enrichment in TDP-43 iCLIP peaks in the distal introns, albeit at a lower intensity. The long-intron genes that are downregulated selectively in RRM2mut also show a general enrichment in iCLIP peaks throughout the intron. Conversely, the upregulated long genes are depleted in iCLIP peaks compared to other upregulated genes. Together this suggests perhaps an additive role in splicing, where the introns that can be bound by as many molecules of TDP-43 will be processed differently, overriding the need for targeted TDP-43 binding to splice sites. This will be interesting to explore with future datasets, with iCLIP at greater sequencing depth.

The two skiptic exons validated in human ALS patients are intriguing high quality transcriptome-wide data is needed to determine whether TDP-43 gain of splicing function is occurring in patient cells. Although the skiptic exons should be conserved between mouse and human, the flanking regulatory sequences that flank the exons may not, so it is not surprising that only two of the seven skiptic exon candidates could be seen to change. It has been previously reported that TDP-43 depletion has a different effect size on the same exon in human than in mouse Mohagheghi et al. (2016). This is supposedly due not to TDP-43 binding, which is

invariant, but the role of other RNA-binding proteins. This combinatorial model of splicing is currently intractable to investigate transcriptome-wide.

While our human validation work was on patients with *TARDBP* mutations, it is interesting to speculate on the relevance of TDP-43 splicing to sporadic disease. Increased total and cytoplasmic *TARDBP* mRNA has been observed in sporadic ALS patient neurons, particularly in neurons with TDP-43 aggregations (Koyama et al., 2016). One can imagine a scenario whether altered TDP-43 autoregulation in sporadic ALS would initially over-compensate and increase *TARDBP* mRNA and cause a splicing gain of function. Eventually when TDP-43 is completely unable to enter the nucleus there would be a shift to a TDP-43 loss of function phenotype, heralded by cryptic exon inclusion and long intron gene downregulation. This idea would be interesting to test in a longitudinal human cell model, for example ALS patient stem cell-derived motor neurons.

LCDmut mice exhibit symptoms of motor neuron degeneration (decreased motor neuron numbers, reduction in grip strength - data not shown), but do so without TDP-43 aggregation in motor neurons. This uncoupling of TDP-43 aggregation and disease symptoms has been observed in other TDP-43 mutant mouse models (Arnold et al., 2013; Gordon et al., 2018).

Further work is needed to untangle the mechanism by which a low-complexity domain mutation leads to a gain of splicing function. As well as pursuing the change in autoregulation and what this might do to TDP-43 translation, I believe it is worth investigating changes in the TDP-43 interactome such a mutation might cause.

6 ALS-causative FUS mutations impair FUS autoregulation through intron retention

6.1 Overview

The most aggressive FUS mutations in ALS are those that completely abolish the nuclear localisation signal (NLS). Patients carrying a single copy of the P525L mutation, where the critical proline of the PY motif is mutated (Chiò et al., 2009), die around 20 years of age with a disease course of less than 2 years (Shang and Huang, 2016). The removal of the NLS mislocalises FUS to the cytoplasm but whether toxicity comes from reducing the nuclear functions of FUS or through a new function in the cytoplasm is still being debated.

The Fratta lab generated RNA sequencing data from embryonic mouse neuronal tissue where FUS was either knocked out or the NLS removed through mutation. I combined the data with two previously published datasets with similar NLS mutations and FUS knockout mice. To assess differential expression and splicing I employed a joint modelling strategy which boosted the power of detection and increased the confidence in my findings. I found that in both expression and splicing, FUS NLS mutations act as a diminished knockout, with little evidence to support a gain of toxic function in the cytoplasm. When examining FUS itself, I observed that loss of nuclear FUS correlated with the reduction in the splicing of a retained intron transcript, which was validated by RT-PCR. The FUS intron retention transcript is insensitive to nonsense-mediated decay. My data suggest that disease-associated FUS mutations impair FUS autoregulation which is normally maintained through a retained intron transcript. This work has important implications for both the RNA biology and neurodegenerative disease fields.

6.2 Contributions

- Mice were handled by Dr Cristian Bodo
- RNA sequencing libraries were prepared by Dr Nicol Birsa.
- RNA was extracted for PCR by Dr Nicol Birsa and Matthew Bentham
- Primers designed and RT-PCRs were performed by David Robaldo and Dr Carmelo Milioto
- Methods for RT-PCR were written by David Robaldo

All bioinformatic analysis and interpretation was designed and performed by myself. The validation experiments were designed by myself and Dr Pietro Fratta.

6.3 Background

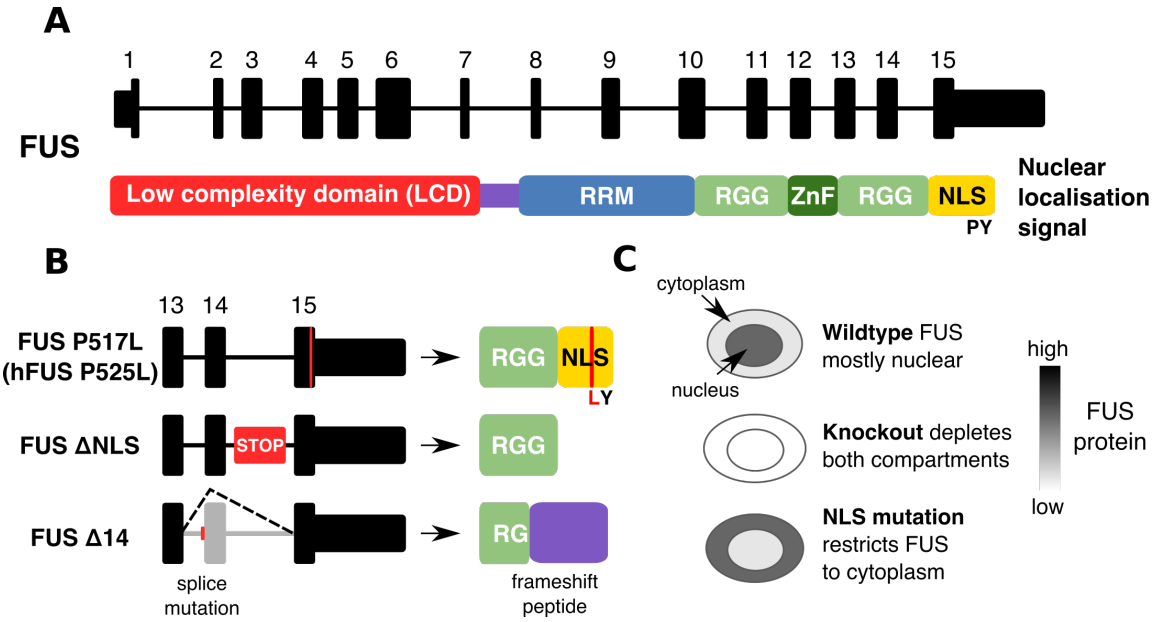


Figure 6.1: The structure of the FUS protein and known ALS mutations. **(A)** The FUS protein is comprised of a low complexity domain (LCD), an RNA recognition motif (RRM) domain, two Arginine-Glycine-Glycine (RGG) domains, a zinc finger domain (Znf), a nuclear export signal (NES) and a nuclear localisation signal (NLS). **(B)** The three FUS NLS mutations used in this study. The Bozzoni group knocked in a point mutation to create the FUS P525L line, a missense mutation equivalent to the human ALS P517L mutation. The Dupuis group created a FUS ΔNLS line where the entire NLS is removed. We have used the FUS Δ14 mouse, where a frameshift mutation leads to the skipping of exon 14 and a frameshifting of the remaining NLS sequence. **(C)** In wildtype cells FUS protein is predominantly nuclear but can shuttle to the cytoplasm. When FUS is knocked out it will be reduced in both compartments but if the NLS is mutated or deleted then FUS will accumulate in the cytoplasm.

For this study, three datasets of FUS knockout and FUS nuclear localisation signal mutation were used. When combined I refer to them collectively as FUS KO and FUS MUT respectively. Table 6.1 describes the 3 datasets. Table 6.2 describes the sequencing libraries generated.

Bozzoni

Capauto and colleagues from the Bozzoni group generated RNA-seq data from purified motor neurons cultured from mouse induced pluripotent stem cells (Capauto et al., 2018). Their knockout uses a gene trap in *Fus* intron 12 first used by Hicks et al (2000). This construct leads to a partial reduction of FUS protein levels. Their NLS mutation is the P525L point mutation (Chiò et al., 2009), which in mice corresponds to P517L. All cells are homozygous for their mutations. The two conditions share a single set of controls, which are motoneurons derived from wildtype stem cells. The read depth and length for all the Bozzoni samples is high and but the sample number is n=3 for each condition, the lowest of the three datasets. Their differential expression analysis found 40 genes in common, with 198 specific to knockout and 419 specific to NLS mutation. Splicing changes were not assessed in their publication.

Dupuis

Scekic-Zahirovic and colleages from the Dupuis group generated RNA-seq data from mouse embryonic brain (embryonic day E18.5) (Scekic-Zahirovic et al., 2016). Their knockout construct uses a gene trap in *Fus* intron 1 (generated in-house) which leads to complete FUS loss. Their NLS mutation introduces a stop codon after exon 14, terminating transcription upstream of the NLS sequence (ΔNLS). This mimics the R495X mutation which removes the entire NLS sequence (Bosco et al., 2010). All mice are homozygous for their genetic modifications. The two conditions have their own separate wildtype littermate controls. The depth and read length of the RNA-seq libraries is lower than the other two datasets but this is compensated for by being polyA+ enriched, which will mean a greater proportion of sequencing reads aligning to coding exons. They found a strong overlap between differentially expressed genes in the two transgenic lines, with 353 shared genes, 433 specific to FUS NLS mutations and 1205 specific to FUS knockout. They also performed a targeted cassette exon splicing analysis (RASL-seq) and again found overlap between knockout and NLS mutation, with 75 shared cassette exons, 98 specific to NLS mutation and 177 specific to knockout.

Fratta

The third dataset was created for this study by the Fratta group from mouse embryonic spinal cord samples (embryonic day E18.5). The knockout construct used to generate mice is the same as the Dupuis dataset and hence should be a robust knockout of FUS. The NLS mutation is the Δ14 mutation discussed in the previous chapter. A splice acceptor site mutation in exon 14 site prevents exon inclusion, leading to a downstream frameshift and removal of the NLS. This mutation corresponds to the human G466VfsX14 mutation (DeJesus-Hernandez et al., 2010). Each condition, Δ14 and knockout have their own wildtype littermate samples as controls. For both conditions, both heterozygous and homozygous mice generated and sequenced. Heterozygous mice were not used in the joint modelling. The RNA-seq data produced has the longest reads at 2 x 150bp as well as high sequencing depth. This provides sufficient resolution to quantify differential splicing.

By combining the newly generated Fratta dataset with that of the two previous groups I have put together the largest analysis to date on FUS and gene expression. Furthermore I have performed a comprehensive analysis on splicing, looking for both annotated and novel splicing events in both FUS knockout and NLS mutation. This joint modelling approach boosts statistical power and allows me to demonstrate that the majority of differentially expressed genes and differential splicing events are shared between FUS knockout and NLS mutation.

6.4 Methods

RNA sequencing

This describes the Fratta dataset. The other two datasets are already published. All mice used were sacrificed at embryonic day 18.5. Mice carrying one or two copies of either the FUS knockout transgene from (Scekic-Zahirovic et al., 2016) or the FUS Δ14 genotype from (Devoy et al., 2017) were compared to their wildtype littermates. Total RNA was extracted from mouse spinal cord. cDNA libraries were created using a TruSeq stranded total RNA RiboZero protocol (Illumina). Libraries were sequenced on an Illumina HiSeq to generate paired end 150bp reads.

Dataset	Tissue	Controls	Age	Knockout (KO)	Mutation (MUT)
Bozzoni (Capauto et al., 2018)	Motor neurons cultured from iPSCs	Shared	-	Gene trap in exon 12	P517L knock-in, corresponding to human P525L
Dupuis (Scekic-Zahirovic et al., 2016)	Whole brain	Separate	E18.5	Gene trap in intron 1	Stop codon after exon 14 (ΔNLS)
Fratta (this study)	Spinal cord	Separate	E18.5	Gene trap in intron 1	Δ exon 14

Table 6.1: The three FUS mouse datasets

Dataset	Replicates per condition	Library type	Mapped reads (millions)	Read type	SRA accession
Bozzoni	3	Total RNA	34-52	2 x 100bp	SRP111475
Dupuis	4-5	mRNA	15-25	1 x 50bp	SRP070906
Fratta	4	Total RNA	52-65	2 x 150bp	-

Table 6.2: RNA-seq statistics of the three datasets

Data processing

All RNA sequencing data was processed with our in-house pipeline (outlined in Methods chapter). RNA-seq data was aligned to the mm10 mouse genome build.

Differential Expression

Each dataset consists of FUS knockout samples, FUS NLS mutation samples and wildtype controls. In the Bozzoni dataset the controls are shared but in the other two datasets the knockout and mutation samples have their own separate controls for use in two-way comparisons. Differential gene expression was tested with DESeq2 (Love et al., 2014). Initially each comparison (wildtype vs knockout or wildtype vs mutation) was performed separately for each dataset, creating six individual analyses. To boost power and create a set of high confidence changes, two joint models were created using either the knockout (KO) or mutation (MUT) samples with their specific controls. The joint model uses all the samples of the same comparison together in a general linear model with a dataset covariate. DESeq2 uses a Bayesian shrinkage strategy when estimating the log_2 fold change. For each gene the log_2 fold change is the linear combination of the three individual datasets. Genes are reported as significantly differentially expressed at a false discovery rate (FDR) threshold of 0.05 (Benjamini and Hochberg, 1995). For plots, gene expression values are raw counts multiplied by each sample's size factor generated by DESeq2. These normalised counts are then normalised to the wildtype samples for each dataset to visualise the relative change in expression.

To assess the level of overlap between the KO and MUT joint models, two different overlap thresholds were employed. The first, a more conservative threshold, depends on a gene being significant at FDR < 0.05 in both datasets. The second, more relaxed threshold, calls a gene as significant if it falls below FDR < 0.05 in one dataset and has an uncorrected P-value < 0.05 in the other.

Differential Splicing

SGSeq was run on all the samples together to discover and classify all potential splicing events using the default parameters for finding novel splicing (Goldstein et al., 2016). Differential splicing for individual comparisons and joint models with a dataset-specific covariate were performed using DEXSeq (Anders et al., 2012). The same overlap threshold strategies were employed as for differential gene expression. SGSeq looks for all potential splicing events in each sample and then counts the reads supporting either the inclusion or exclusion of that splicing variant. Percentage Spliced In (PSI) values (Katz et al., 2010) for each splicing variant were calculated by taking the read counts supporting the inclusion event and dividing by the total reads in that event.

Gene Ontology

Gene Ontology enrichment testing was performed with the GProfileR package (Reimand et al., 2016). GO and KEGG categories were hand-curated to remove redundant terms and restricted to a minimum overlap of 5 genes per set. All *P*-values are reported after Bonferroni correction.

iCLIP and functional analyses

FUS iCLIP data from mouse brain (Rogelj et al., 2012) was reprocessed by the iCOUNT iCLIP analysis pipeline (http://icount.biolab.si/). I downloaded the set of FUS iCLIP clusters that passed enrichment against background at FDR < 0.05. Only iCLIP clusters with a minimum of two supporting reads were used. Untranslated region (UTR) and coding exon (CDS) annotation were taken from GENCODE mouse (comprehensive; mouse v12). Any intron-retention, nonsense mediated decay or "cds end nf" transcripts were removed. UTR coordinates were split into 5´ and 3´ UTR based on whether they overlapped an annotated polyadenylation site or signal (GENCODE mouse v18 poladenylation annotation). 3´ UTRs were extended by 5kb downstream to capture any unannotated sequence. Introns were defined as any gaps in the transcript model between CDS and UTR coordinates. Promoter-antisense coordinates were taken by flanking the 5ÚTR sequence by 5kb upstream and inverting the strand. Overlaps between iCLIP clusters and genomic features were created for each set of differentially expressed genes, split into upregulated (log_2 fold change $>$ 0) or downregulated (log_2 fold change < 0). Overlaps were done in a strand-specific manner, with only iCLIP clusters in the same direction being used.

Whether an iCLIP cluster overlaps a genomic region depends on both the affinity of the chosen protein for RNA sequence of the motif and the abundance of the RNA in the cell. In addition, a longer region would be more likely to overlap an iCLIP cluster by random chance than a shorter region. When comparing sets of genomic regions, whether genes or splicing events, this must be taken into account. See appendices for distributions of lengths and expressions between significant and non-significant genes and splicing events.

To test for enrichment of FUS iCLIP clusters in upregulated and downregulated genes, each set of tested genes was compared to a set of null genes with no evidence of differential expression ($P > 0.05$ in both models). The null set was selected for having both lengths and expression values within the first and third quartile of the test gene set. The expression values were calculated by taking the mean number of reads covering each gene in the Fratta wildtype samples, with each sample read count first normalised by the library size factor for each sample calculated by DESEq2.

The proportion of each set of genes overlapping an iCLIP peak was compared to that of the null set with a χ^2 test of equal proportions.

For the splicing events found in the joint models, three enrichment tests were performed for different genomic features. For these tests the coordinates of the entire encompassing intron were used for each splicing variant. Each test set of splicing events was compared to

a matched set of null splicing events where $P > 0.05$ in both joint models. The null events were chosen to have length and expression levels within the first and third quartiles of that of the test set.

Proportions of overlap between splicing events and the null set were tested using a χ^2 test of equal proportions.

As a positive control in both analyses, the same overlaps were computed with iCLIP clusters from U2AF65, also from (Rogelj et al., 2012).

The coordinates of polyadenylation cleavage sites were downloaded from the PolyA Site Atlas (Gruber et al., 2016). The proportions of splicing events that overlapped a polyadenylation cleavage site were compared to the null.

Per nucleotide PhyloP conservation scores (Pollard et al., 2010) comparing mouse (mm10) with 60 other vertebrates was downloaded from UCSC. The median PhyloP score was calculated for each splicing variant and compared.

RT-PCR

Primers were designed using Primer3 (Koressaar and Remm, 2007) and *in silico* PCR (UCSC). For both human and mouse FUS, the forward primer was designed for exon 6 and the reverse primer designed to span the spliced exon 8/9 junction to preferentially amplify spliced FUS mRNA. An additional third primer was designed to amplify a section of either intron 6 or intron 7.

Cells were obtained from mouse spinal cord and/or cultured mouse embryonic fibroblasts resuspended in Trizol (Thermo Fisher). RNA was extracted using miRNeasy Mini Kit (Qiagen) following the manufacturer's instructions cDNA was obtained from extracted RNA using SuperScript IV Reverse Transcriptase kit (Thermo Fisher). Briefly, a mix was made of RNA template (500ng for mouse brain; 100ng for cultured cells (cycloheximide treatment), 10 mM dNTP, 50 mM oligo d(T)20, 50 mM random hexamer followed by 5 min of incubation at 65°C and 1 min in ice. Mix was then complemented with 5X SuperScript IV Reverse Transcriptase buffer, 100 nM DTT, RNase OUT and SuperScript IV Reverse Transcriptase buffer followed by incubation at 23°C, 55°C and 80°C, 10 min each.

RT-PCR was carried out using 10X AccuPrime Taq DNA polymerase mastermix system (Invitrogen). Each PCR reaction mix contained 5 ng of gDNA, 10 mM of forward and reverse primers. cDNA was amplified with the following conditions: Intron 6 retention: One cycle of 5 min at 95°C, followed by 30 cycles of 30 sec at 95°C, 30 sec at 56°C, and 30 sec at 68°C, and finishing with 5 min incubation at 68°C. Intron 7 retention: One cycle of 5 min at 95°C, followed by 30 cycles of 30 sec at 95°C, 30 sec at 61°C, and 30 sec at 68°C, and finishing with 5 min incubation at 68°C. Srsf7 NMD positive control: One cycle of 5 min at 95°C, followed by 35 cycles of 30 sec at 95°C, 30 sec at 58°C, and 15 sec at 68°C, and finishing with 5 min incubation at 68°C. Amplified products were finally obtained using Agilent 4200 TapeStation System following the manufacturer's instructions. Results were

analysed on TapeStation analysis software (Agilent). Intron retention events are plotted as the percentage of integrated area of band corresponding to intron retention.

Target	Direction	Sequence
mFUS exon 6	F	GTTATGGCAATCAGGACCAGAG
mFUS intron 6	R	TTGGCTCCCAAGTTCTCACA
mFUS intron 7	F	GGAGAAACTGGATGGATGCAC
mFUS exon 8/9	R	CCTGTTCAGAATCATGACGAGA
hFUS exon 6	F	TCCTCCATGAGTAGTGGTGGT
hFUS intron 6	R	GTTCAGGCTCCCAAGTTCTC
hFUS intron 7	F	TTCTCTCGGGTGAGAGAACC
hFUS exon 8/9	R	GTCTGAATTATCCTGTTCGGAGTC
mSRSF7	F	CGACGAAGAAGAAGCAGGTTTC
mSRSF7	R	TCTGGCCTCTTATGCTGATCAC

Table 6.3: List of primers used in RT-PCR. mFUS: mouse *FUS*; hFUS: human *FUS*

Cycloheximide treatment and fractionation

Mouse embryonic fibroblasts were treated with 100ug/ml cycloheximide (Sigma) for 6 hours before RNA was extracted with Trizol (Thermo Fisher) and RT-PCR performed as before. As a positive control, primers targeting the NMD-sensitive exon 4 of *Srsf7* were used from (Edwards et al., 2016).

6.5 Results

Modelling differential expression jointly increases power and demonstrates significant overlap between FUS knockout and FUS NLS mutation

Differential expression compares the abundances of transcripts from each gene between two conditions. As most RNA-sequencing datasets comprise small numbers of samples, the number of genes found to be significantly changed between conditions depends on the degree of difference between conditions, the number of samples per condition, and the read depth covering each gene. When I analyse each dataset individually, comparing knockout and NLS mutation to their respective controls in each dataset, the Dupuis dataset had the largest number of differentially expressed genes in both conditions (FDR < 0.05), an order of magnitude more than Bozzoni or Fratta (Table 6.4). Despite the differences in numbers, in all three datasets the knockout condition produces more differentially expressed genes than NLS mutation, suggesting a larger effect on FUS function.

I then combined the three datasets together in a joint analysis for the knockout and NLS mutation samples and their respective controls. I will refer to these two joint models as KO and MUT respectively. This increases the sample size of each condition from 3-5 to 11-12 which should markedly improve the estimation of per-gene variation. DESeq2 uses a general linear model framework so I added a dataset-specific covariate. This strategy will reward genes where the direction of change is the same between all three datasets and punish genes where the datasets differ in direction. The two comparisons are nominally independent from each other as only the Bozzoni wildtype samples are shared. Modelling all the data together allows the three independent studies of FUS function to contribute to a high-confidence set of expression changes. At FDR < 0.05 the KO joint model contains 2136 significantly changed genes and the MUT model contains 754. When comparing the genes found by the joint analysis to the individual analyses there is only a moderate overlap. Of the 2916 genes found in the Dupuis knockuts, only 1007 of those genes are present in the joint KO analysis. This suggests that a large number of genes called as significant in each dataset cannot be replicated in the other two datasets, despite being the same condition and all being embryonic neuronal tissue.

For each gene the resulting joint log_2 fold change and P-value is an estimated combination of the three datasets. I compared the values found by the individual analyses with the joint models(Fig. 6.2). For the KO models, the Fratta knockout comparison has larger log_2 fold change values for the same genes when compared to Bozzoni and Dupuis knockouts. The fitted log_2 fold change is a compromise between the estimated fold changes of all three datasets. When inspecting the distribution of P-values, the Dupuis KO dataset has a excess of low P-values compared to the other two. Therefore the joint modelling strategy increases the number of genes and harmonises three different datasets together into a set of high confidence differentially expressed genes.

The joint models provide two set of genes where we can be confident of a shared signal between the three datasets. I next looked for evidence of a shared gene expression signal

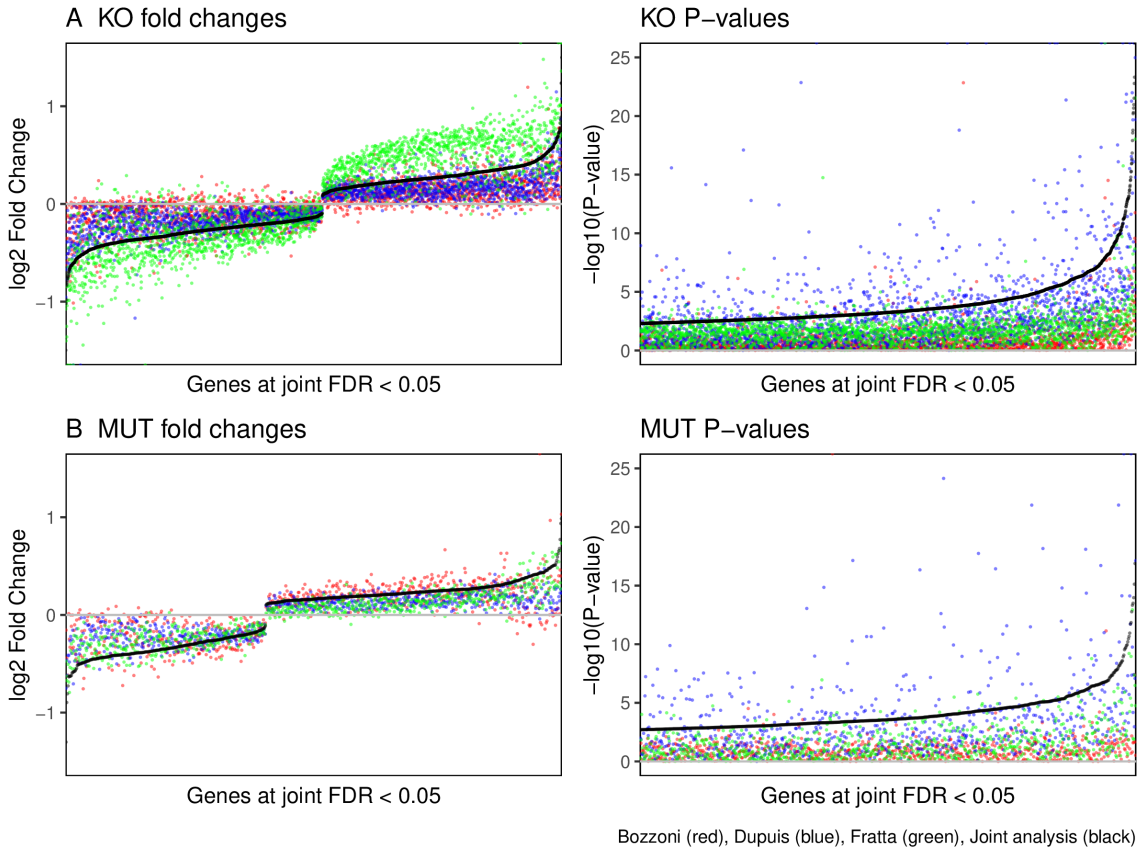


Figure 6.2: Joint differential expression increases power and adjusts effect sizes. **(A)** Plotting the log_2 fold changes (left) and P-values (right) for the 2136 KO genes and 754 MUT genes found at FDR < 0.05 in the two joint analyses. Values found in the individual analyses, Bozzoni (red), Dupuis(blue) and Fratta (green) are compared to the value produced by the joint analysis (black). **(B)** As before but for the MUT analyses.

between the KO model and the MUT model. I first used a conservative threshold for overlap, where a gene must be significant at FDR < 0.05 in both models. This produced an overlap of 425 shared genes between KO and MUT (Fig. 6.3A), with 329 mutation-specific genes and 1711 knockout specific genes. I next used a relaxed overlap criteria where a gene overlaps if it reaches FDR < 0.05 in one model and an uncorrected P < 0.05 in the other. This increased the overlap to 1318 genes, reducing the specific genes to 186 in the MUT model, and 961 in KO. The majority of genes are now shared between the two models.

	Bozzoni MUT	Dupuis MUT	Fratta MUT	Bozzoni KO	Dupuis KO	Fratta KO
Individual hits	19	1552	88	100	2916	151
Overlapping joint model	5	368	57	51	1007	114
Unique to dataset	14	1184	31	49	1909	37
Joint model	754			2136		
Overlap (strict)	329		**425**		1711	
Overlap (relaxed)	186		**1318**		961	

Table 6.4: Results from separate and joint differential expression analysis

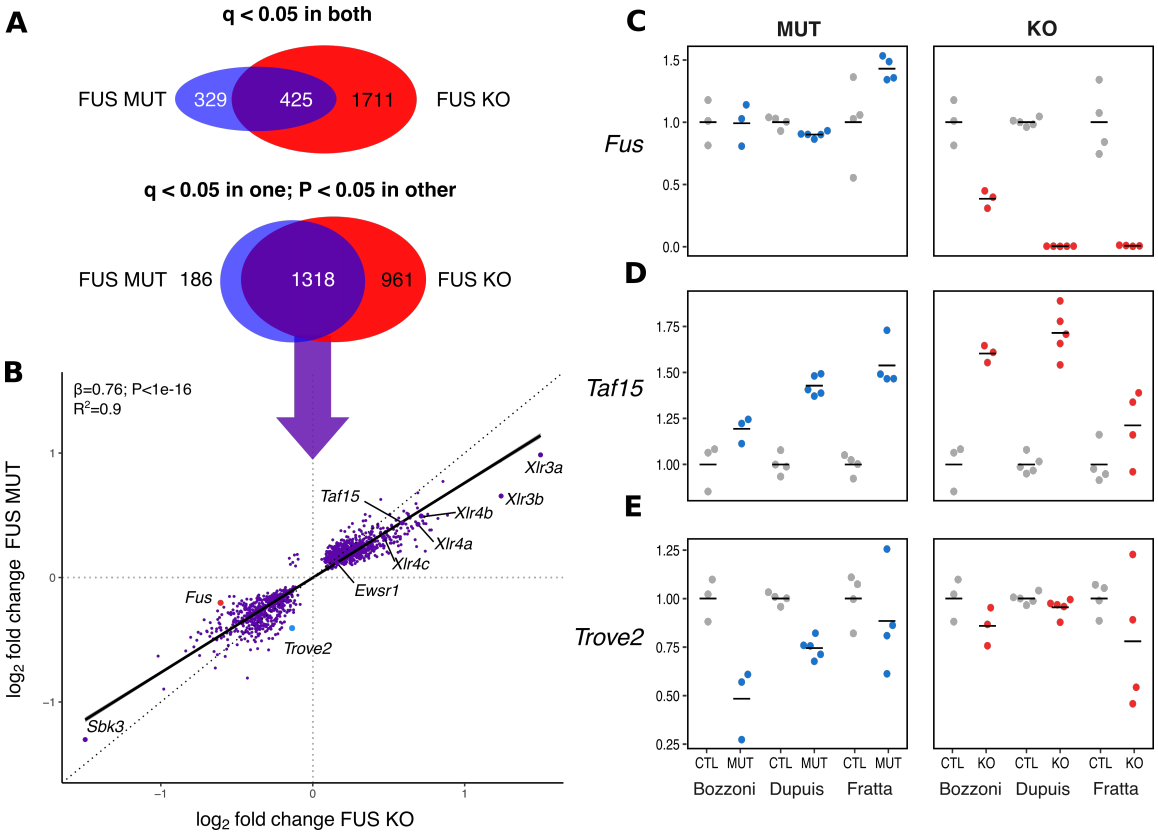


Figure 6.3: Overlapping the joint differential expression models shows that FUS mutations affect gene expression in generally the same direction. (A) Venn diagrams show the overlap between the FUS KO and FUS MUT joint differential expression models with a strict FDR cut-off (upper) and a more relaxed *P*-value cut-off. **(B)** Plotting the fitted log_2 fold-change values for each of the 1318 overlapping genes in KO and MUT shows a bias towards weaker changes in MUT compared to KO. **(C, D,E)**: Normalised read counts in each dataset for *Fus*, *Taf15* and *Trove2*. Samples are plotted relative to the mean expression in controls (CTL).

Comparing the $\log_2$ fold changes found for the 1318 overlapping genes between FUS KO and FUS MUT showed that only 7 genes are altered in opposing directions (Fig. 6.3B) . Fitting a linear model between the fold changes of the two datasets shows that the effect of FUS MUT on gene expression is 76% that of FUS KO. ($\beta = 0.76$; P $<$ 1e-16 F-test; $R^2 = 0.90$). This suggests that while FUS KO and FUS MUT affect the same genes in the same directions, the magnitude of change is greater in FUS KO than FUS MUT. This relationship is not an artefact of the relaxed overlap criteria as fitting the model on just the 425 strictly overlapping genes resulted in $\beta = 0.8; P < 1e-16$. The relative weakness of NLS mutations compared to knockouts can be explained as NLS mutant FUS can still be detected in the nucleus at low levels (Devoy et al., 2017; Scekic-Zahirovic et al., 2016).

Visualising individual genes in all the datasets demonstrates the power of the joint models. *Fus* itself is unchanged in the FUS MUT joint model but strongly downregulated in FUS KO (Fig. 6.3C). The Bozzoni knockout is weaker than the one used by Dupuis and Fratta, with a reduction of *Fus* RNA to only 40% of wildtype, compared to near 100% knockout in the other two datasets. The three NLS mutations are inconsistent for *Fus* expression. In Bozzoni it is unchanged, in Dupuis it is slightly downregulated (presumably due to a loss

of reads covering the final exon and in Fratta it is slightly upregulated.

The other members of the FET family of RNA-binding proteins, *Taf15* and *Ewsr1* have been shown to reciprocally interact with FUS at the protein and RNA level (Kapeli et al., 2016; Lagier-Tourenne et al., 2012). *Taf15* is strongly upregulated in both MUT and KO (Fig. 6.3D), as is *Ewsr1*. *Taf15* upregulation has been repeatedly observed and validated (Kino et al., 2015; Scekic-Zahirovic et al., 2016). The *Trove2* gene encoding the 60 kDa SS-A/Ro ribonucleoprotein, is downregulated in MUT only and unchanged in KO (Fig. 6.3E) and has also been validated (Scekic-Zahirovic et al., 2016).

The X-linked lymphocyte receptor genes, *Xlr3a, Xlr3b, Xlr4a* and *Xlr4b* are all strongly upregulated in both conditions but more so in KO than MUT. These genes form a cluster of paralogous genes on the X chromosome and are paternally imprinted in mice (Raefski and O'Neill, 2005). They appear to be specific to mice. *Xlr4b* overexpression has been shown to alter dendritic spine growth in mouse neurons (Cubelos et al., 2010). Although these changes have been observed in another FUS knockout mouse model (Kino et al., 2015),they are potentially artefacts from an imbalance of sexes between conditions. Embryonic mice are not typically sexed but I was able to sex the samples *in silico* using the read counts aligning to Y chromosome genes (see Appendices). Although there is an imbalance between sexes it is clearly not driving the large upregulation of *Xlr* genes. The upregulation of *Xlr3a* is strongest within the all-male Bozzoni dataset (Fig. 6.4A) and can also be seen in the mixed-sex Dupuis and Fratta datasets.

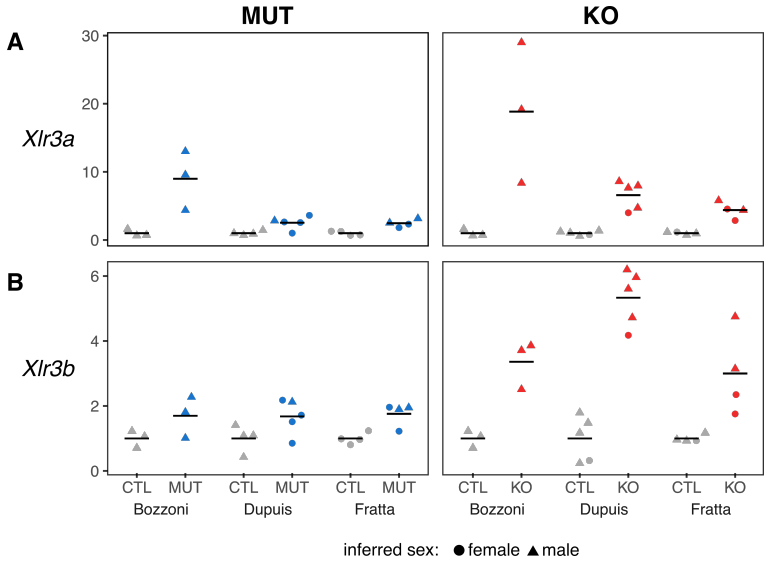


Figure 6.4: Xlr genes are upregulated in both FUS KO and FUS MUT in mice of both sexes. Normalised read counts in each dataset, plotted relative to the mean of each dataset and condition-specific control group for *Xlr3a* **(A)** and *Xlr3b* **(B)**.

Synaptic and RNA-binding genes are a common gene expression response to FUS nuclear depletion

Gene ontology (GO) terms enriched in the overlapping genes were strongly direction specific, with upregulated genes enriched in terms involving RNA binding, splicing and metabolism

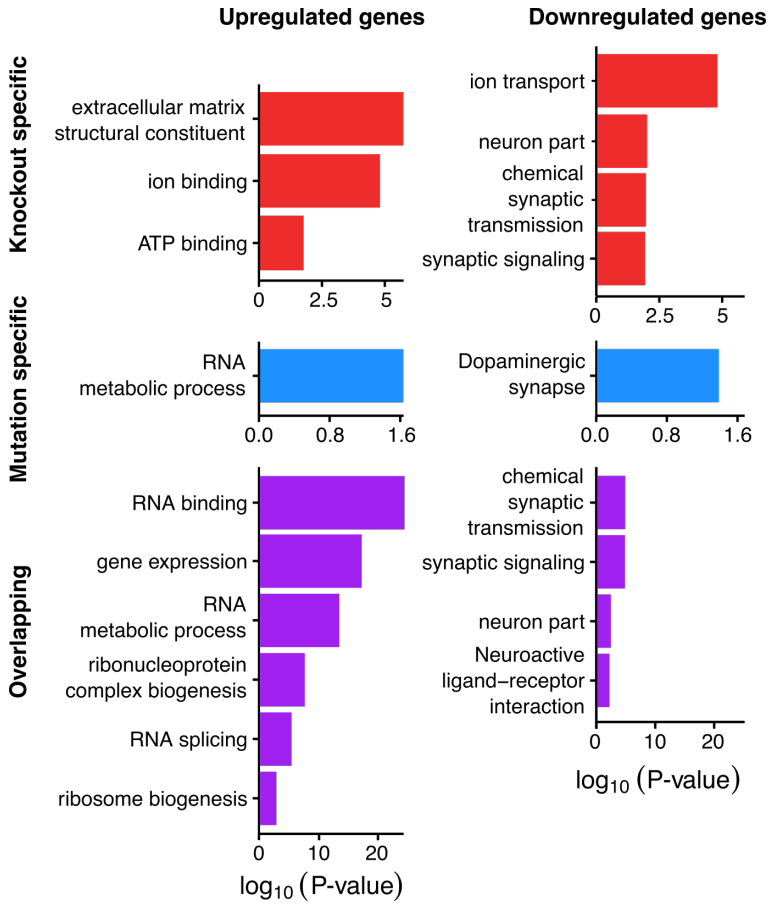


Figure 6.5: Overlapping genes are enriched in neuronal and RNA terms. Enriched Gene Ontology terms in the three groups of genes, split by direction.

whereas downregulated genes were enriched in synaptic and neuronal terms (Fig. 6.5A). Knockout-specific and mutation-specific genes were less clearly enriched in specific functions. Knockout specific genes were involved in extracellular membrane functions, ion channels and amino acid transport whereas the mutant specific genes showed an enrichment in dopaminergic synapses and RNA metabolism.

Downregulated genes are enriched for FUS binding peaks

Individual nucleotide resolution crosslinking and immunopreciation (iCLIP) is an experimental technique for identifying the RNA targets of RNA-binding proteins. Using iCLIP and other RNA-protein interaction techniques, FUS has been shown to preferentially bind within introns and 3´ untranslated regions (3´ UTRs) rather than exons. (Lagier-Tourenne et al., 2012; Rogelj et al., 2012; Ishigaki et al., 2012; Masuda et al., 2015; Kapeli et al., 2016). FUS binding at the 3´ UTR has been shown to influence polyadenylation of certain genes (Masuda et al., 2015) but may also have a role in directing mRNA localisation or competing with microRNAs, which in animals predominantly bind 3´ UTR sequences to trigger degradation (Lee et al., 1993; Carthew and Sontheimer, 2009). A small number of genes have been reported to have FUS binding sites upstream and antisense of the promoter, which correlated with upregulation of the gene upon FUS knockdown (Ishigaki et al., 2012).

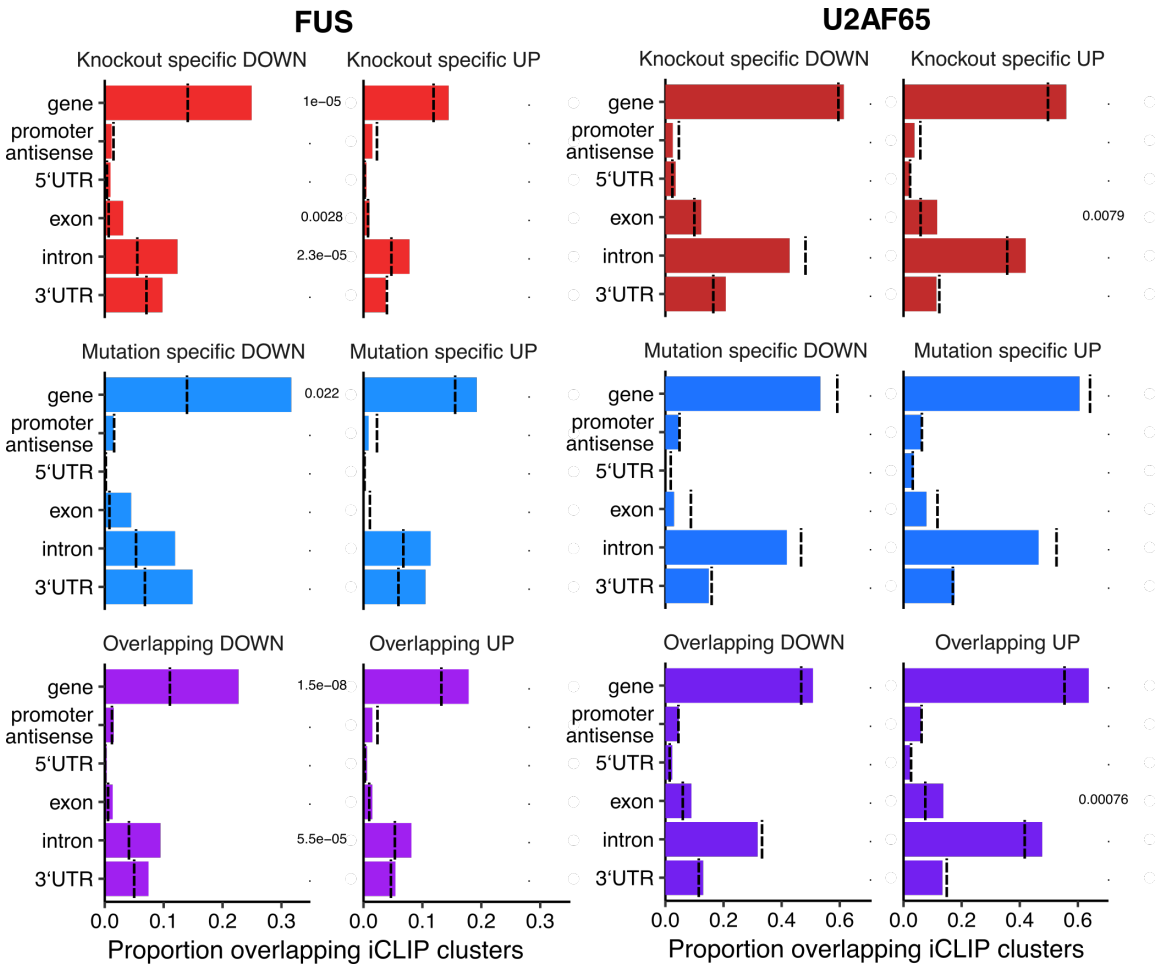


Figure 6.6: Downregulated genes are enriched in FUS iCLIP binding peaks. Proportions of each set of genes that have any iCLIP clusters within the entire gene or in any particular gene feature, split by direction. Sets of gene were tested for enrichment for either FUS or U2AF65 iCLIP clusters, compared to sets of null genes with similar length and expression characteristics. *P*-values are from χ^2 test of equal proportions and adjusted for multiple testing with Bonferroni correction.

I used the coordinates of FUS binding sites enriched relative to a background control in a published FUS iCLIP dataset from embryonic day 18 mouse brain (Rogelj et al., 2012) to investigate a relationship between specific FUS binding and the direction of gene expression in the different gene sets. As a positive control I overlapped the same sets of genes with iCLIP clusters from U2AF65, a protein that interacts with the U2 snRNP at the 3' splice site and should therefore be ubiquitous.

As previously reported, FUS preferentially binds to introns and 3'UTRs over exons (Rogelj et al., 2012; Lagier-Tourenne et al., 2012). Promoter-antisense binding is found in a small number of genes and at a similar proportion. Comparing each set of genes to its length and expression-matched null demonstrates a clear enrichment in FUS iCLIP peaks in downregulated genes in all three categories (Knockout specific DOWN P = 1e-5; Mutation specific DOWN P = 0.02; Overlapping DOWN P = 1.5e-8; Fig. 6.6). Splitting these genes into features reveals an enrichment in introns especially (Knockout specific DOWN P = 2.3-e5; Overlapping DOWN P = 5.5e-5). In contrast, the gene level overlaps with U2AF65 iCLIP clusters do not differ between the significant and null sets with the exception of a mild en-

richment overlapping exons in the upregulated introns (Knockout specific UP P = 0.0079; Overlapping UP P = 0.00076). These data suggest that the downregulated genes are bound by FUS, particularly in introns, and these genes may be vulnerable to nuclear FUS loss.

FUS modulates the inclusion of a set of highly conserved RNA-binding protein introns

I then used the same joint modelling approach to assess evidence of differing effects between FUS knockout and NLS mutation on alternative splicing. For the individual analyses I used SGSeq (Goldstein et al., 2016) to discover and quantify all possible alternative splicing isoforms, both novel and annotated. The read counts supporting each splicing variant were used to test for differential usage between conditions for either NLS mutation or knockout and their respective controls. For the joint analyses, I ran SGSeq on all the samples simultaneously and then fit two separate models for all knockout samples and their controls (KO) and all mutation samples and their controls (MUT), incorporating a dataset-specific covariate. Models were fit using DEXSeq (Anders et al., 2012). Table 6.5 summarises the numbers of splicing events in the KO and MUT models at FDR < 0.05. The joint models increased power as for MUT and KO respectively 93 and 890 events are found to be significantly altered, more than the sums of the individual analyses. There is also a very good concordance between each individual analyses and their joint model, with the exception of the Bozzoni mutant samples (only 7 out of 31).

Comparing the joint FUS knockout and NLS mutation splicing models, the two overlap at both a strict and relaxed significance threshold. With the relaxed overlap criteria only 16 splicing events remain specific to the NLS mutations (Fig. 6.7A) . There are 501 knockout specific splicing events and 405 that overlap between the two conditions. Manual curation of the 16 mutant specific events gave little confidence of a mutation-specific effect on splicing, as only one splicing event appeared convincingly in all 3 datasets. The single event that did appear to be real was an intron retention event in *FUS*, far upstream of the mutated NLS region.

	Bozzoni MUT	Dupuis MUT	Fratta MUT	Bozzoni KO	Dupuis KO	Fratta KO
Individual hits	47	1	79	275	54	316
Overlapping joint model	9	0	33	143	31	206
Unique to dataset	38	1	46	132	23	110
Joint model	93			890		
Overlap (strict)	33		**60**	830		
Overlap (relaxed)	16		**405**	501		

Table 6.5: Results from separate and joint splicing analysis

Comparing the different types of splicing event shows a similar distribution between knockout-specific and overlapping splicing events, both of which are dominated by complex splicing

events. These events are combinations of multiple types, such as a cassette exon accompanied by a retained intron and multiple alternative 3´ and 5´ splice sites, all within the same locus. When using tools like SGSeq that can pick up novel splicing events it can be expected that the majority of events are indeed complex. This has been seen with other splicing tools (Vaquero-Garcia et al., 2016). An example of a complex event is in *Ybx1*, which comprises a cassette exon within a retained intron (Fig. 6.7D). The second largest category of event are retained introns. An example of this class is seen in *Ewsr1*, where a normally retained intron is less retained in both FUS knockout and NLS mutation (Fig. 6.7E). The third largest category are the cassette exons, which can either be skipped or included. An example is an annotated cassette exon in the neuronal gene *Nrxn3*, which is included more in FUS knockout and NLS mutation when compared to wildtype (Fig. 6.7F). RNA-seq traces of all events mentioned above are in the appendices. Alternate 5´ and 3´ splice sites can be found in all three sets of genes, with alternate 5´ sites appearing at twice the rate of alternate 3´ splice sites. This discrepancy could be explained by the interaction of FUS with the U1 snRNP (Yu et al., 2015; Yu and Reed, 2015). Knockdown or NLS mutation of FUS could lead to impaired U1 snRNP binding and hence disrupted 5´ splice site recognition.

Comparing the strict to the relaxed overlap criteria between the KO and MUT joint models reduces the number of events found to be specific to FUS mutation to 16, with 405 overlapping events and 501 found to be specific to FUS knockout (Table 6.7). Plotting the log_2 fold changes of the overlapping splicing events, there is a similar trend towards larger fold changes in the KO than the MUT joint models ($\beta = 0.7$, P $<$ 1e-16; F-test; $R^2 = 0.89$). There are no overlapping splicing events that change in opposing directions between the two joint models.

The three largest categories of splicing events for the knockout-specific and overlapping events were complex, retained intron and cassette exons. I subjected each category to enrichment tests both for gene ontology terms and other genomic features. There was a clear enrichment in RNA-binding and neuronal GO terms in the overlapping splicing events, with terms relating to RNA binding dominating retained introns (Fig 6.8A). Conversely neuronal terms were found in cassette exons. Knockout-specific splicing events were enriched for different sets of genes, including microtubule and nucleolus.

The complex and intron retention events may arise from differences in polyadenylation site usage. FUS has been observed to modulate polyadenylation (Masuda et al., 2015). To investigate this I compared each set of splicing events with a set of annotated polyadenylation sites (Gruber et al., 2016). Only overlapping retained introns were enriched for polyadenylation sites (P=0.004), suggesting some intron retention events are mislabelled 3´ UTRs (Fig 6.8B).

Direct regulation of splicing events by FUS binding to introns has been proposed using RNA-protein interaction experiments (Lagier-Tourenne et al., 2012; Rogelj et al., 2012; Ishigaki et al., 2012). To look for evidence of FUS regulating these splicing and polyadenylation events I used the same set of FUS iCLIP clusters from (Rogelj et al., 2012). Complex events and retained introns were enriched in having iCLIP clusters within the affected introns

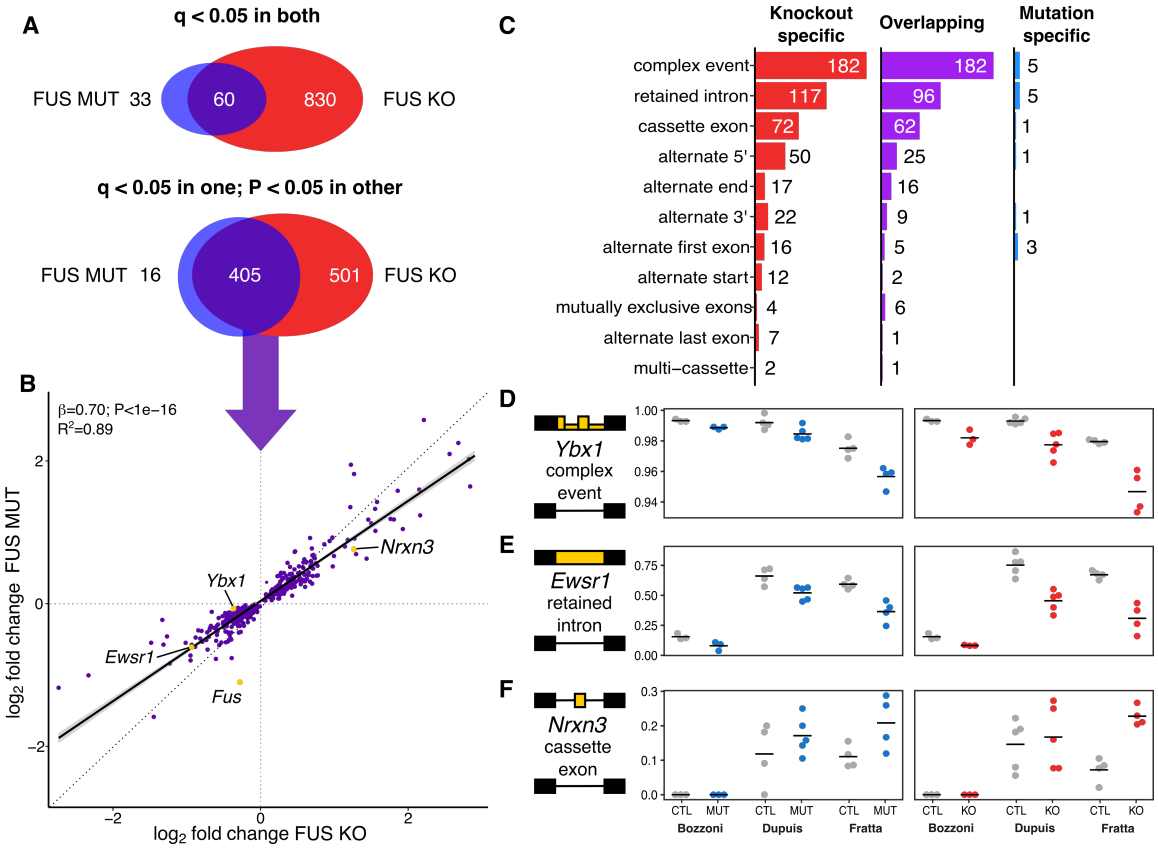


Figure 6.7: Splicing changes strongly overlap between KO and MUT joint models. **(A)** Overlapping KO and MUT splicing events results in a significant overlap. The number of mutation specific events drops to just 16 when a more relaxed overlap threshold is used. **(B)** Splicing events plotted by their log_2 fold change values in the KO and MUT joint models. There is a bias towards larger changes in the KO than MUT ($\beta = 0.7$; .P $<$ 1e-16). Dotted line $y = x$; bold line fitted regression **(C)** Categories of splicing variant found in each set of events. Complex events are defined as splicing events which are made up of multiple categories. **(D-F)** Examples of a cassette exon, retained intron and a complex event in all three datasets.

in both overlapping and knockout-specific contexts. The strongest enrichment was seen in overlapping retained introns (P=5e-32; Fig 6.8A). No enrichment was seen in cassette exons, suggesting that cassette exon splicing changes are not the direct result of a change in FUS binding.

RNA-binding proteins often contain intronic sequences that are very highly conserved (Lareau et al., 2007) and these sequences are often used in the regulation of their translation (Ni et al., 2007). To test whether the splicing events show high sequence conservation, I calculated the median PhyloP score using the 60-way comparison between mouse and other species for each encompassing intron (Pollard et al., 2010). Sets of events were then tested on the proportion of the set with a median phyloP score > 0.5, where a score of 0 is neutral and 1 is highly conserved. Only retained introns were shown to be enriched in sequence conservation, and at a greater externt for overlapping (P=3e-18) than in knockout-specific events (P=0.0002; Fig 6.8D). For cassette exons the two flanking introns are included and so the median conservation score will reflect the conservation of those introns, even when the central cassette exon itself is highly conserved.

Taken together, these results show that nuclear loss of FUS through either knockout or

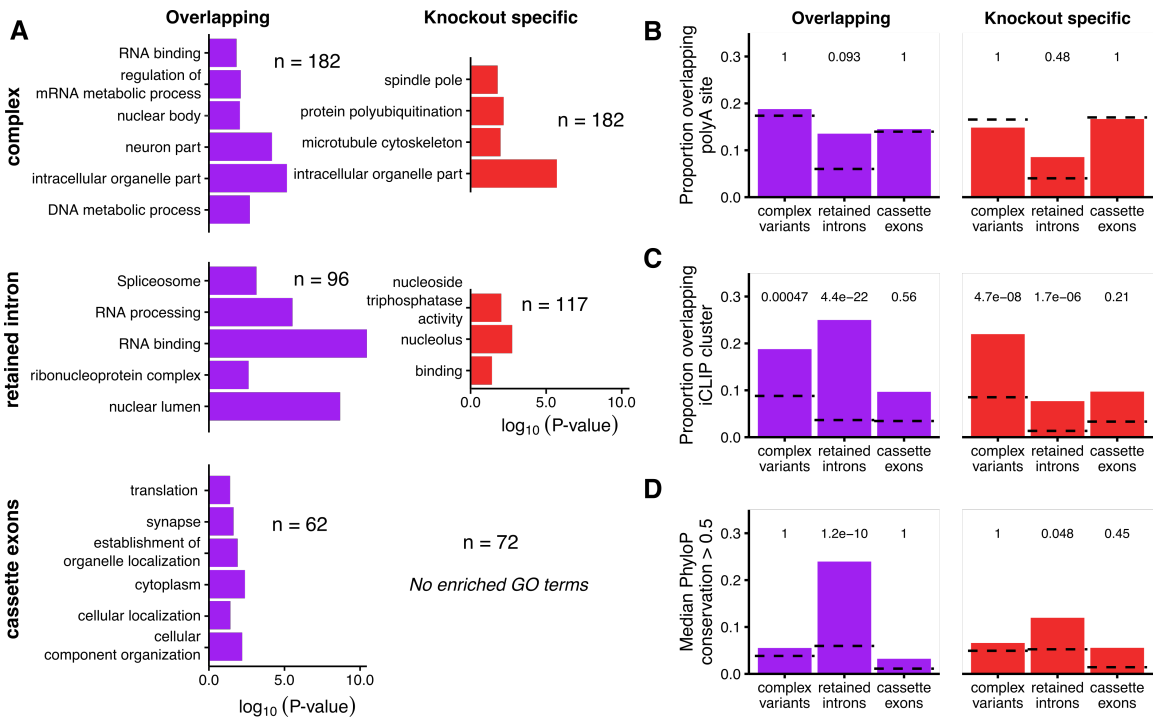


Figure 6.8: Intron retention events are highly conserved and occur in RNA binding proteins. **(A)** Significantly enriched Gene Ontology terms found in genes split by category and splicing variant type. **(B)** The proportion of each type of splicing variant in each category that overlaps a polyadenylation cleavage site. **(C)** The proportion of each type of splicing variant in each cateoory that overlaps a FUS iCLIP peak. **(D)** The proportion of each type of splicing variant in each category that have a median PhyloP conservation score greater than 0.5. Each set of splicing events was compared to a null set of non-significant splicing events with matched length and wildtype expression. P-values adjusted for multiple testing with the Bonferroni method.

NLS mutation leads to a set of splicing changes concentrated in conserved intron retention events affecting other RNA-binding proteins. Cassette exons do not tend to be bound by FUS beyond the null expectation and originate from less conserved introns.

As the splicing events are enriched for similar gene ontology terms as the differentially expressed genes, I reasoned that these may affect the same group of genes. However, only 59 differentially expressed genes are found to contain a splicing event. Of those 59, only 12 have FUS iCLIP binding peaks (17%). Those 12 genes all have either complex or retained intron events and includes the U1 splicing factor *Snrnp70*, the FET protein family members *Ewsr1* and *Taf15*, and *Fus* itself. This analysis shows that the role of FUS on gene expression and splicing mostly affects mutually exclusive sets of genes.

FUS autoregulation is dependent on intron retention

The joint splicing analyses found 5 intron retention events in FUS itself. 3 of these are knockout-specific and probably result from increased intronic reads in the partial Bozzoni knockout samples. However, The 2 remaining introns (introns 6 and 7) were also found in the FUS NLS mutants. These introns overlap with a large number of FUS iCLIP binding peaks and could be the site of regulation of the *Fus* transcript by the FUS protein. Many

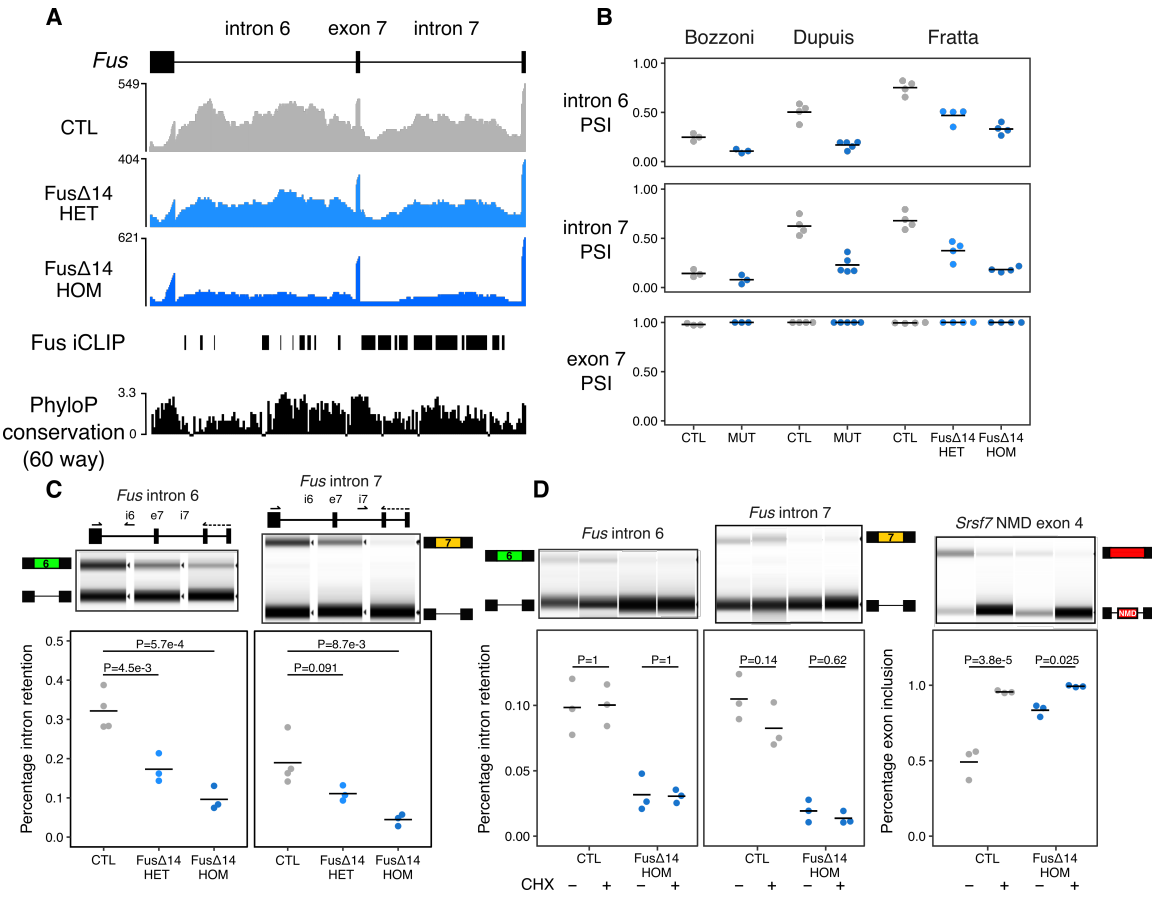


Figure 6.9: Fus intron retention is an NMD-insensitive autoregulation mechanism. **(A)** FUS introns 6 and 7 are highly conserved and have multiple FUS iCLIP binding peaks. Retention of introns 6 and 7 decreases with increasing dose of FUS Δ14. RNA-seq coverage for wildtype, FUS Δ14 heterozygous and FUS Δ14 homozygous samples are accompanied by FUS iCLIP (Rogelj et al., 2012) and PhyloP conservation (60 way) tracks. **(B)** Percentage spliced in (PSI) values of intron 6, intron 7 and exon 7 in the three datasets, including the FUS Δ14 heterozygotes. **(C)** RT-PCR validation of the reduction in intron 6 and 7 inclusion with increasing dose of FUS Δ14 mutation. Left panel - FUS intron 6 (ANOVA genotype P=5.1e-4; t-test CTL vs HET P=4.5e-3; CTL vs HOM P=5.7e-4) Right panel - FUS intron 7 (ANOVA genotype P=8.5e-3; t-test CTL vs HET P=0.091, CTL vs HOM P=8.7e-3) **(D)** Translation blocked with cycloheximide (CHX) to observe whether the intron retention transcript is sensitive to nonsense-mediated decay. Left panel: FUS intron 6 retention is not altered with CHX treatment. ANOVA treatment P=0.96; genotype P=1.9e-5; t-test CTL untreated vs CTL treated P=1; HOM untreated vs HOM treated P=1. Middle panel: FUS intron 7 retention is unchanged by CHX treatment. ANOVA treatment P=0.10; genotype P=3.7e-6; t-test CTL untreated vs CTL treated P=0.14; HOM untreated vs HOM treated P=0.62. Right panel - *Srsf7* exon 4, a known NMD target, is increased by CHX treatment. ANOVA treatment P=5.3e-4; genotype P=0.011; t-test CTL untreated vs CTL treated P=3.8e-5; HOM untreated vs HOM treated P=0.025. All reported P-values are corrected for multiple testing.

RNA-binding proteins have the ability to bind their own RNA to control the level of their own expression, a phenomenon known as autoregulation (Rosenfeld et al., 2002; Jangi and Sharp, 2014). When protein levels are high, protein-RNA binding shifts splicing towards the production of an non-coding isoform. This is commonly through exposing transcripts to nonsense-mediated decay (NMD) (McGlincy and Smith, 2008), by including an exon containing a premature stop codon as in HNRNP L and NOVA (Rossbach et al., 2009; Dredge et al., 2005), skipping a frame-preserving exon as in PTBP1 (Wollerton et al., 2004), or splicing within the 3´ UTR as in TDP-43 (Ayala et al., 2011).

RNA-protein interaction experiments have revealed a large cluster of FUS binding across introns 6 and 7 of the FUS gene (Lagier-Tourenne et al., 2012), both of which show very high sequence conservation between species. This region was previously suggested to be the locus of autoregulation due to both an annotated cassette exon event in exon 7 and an a early polyadenylation transcript being present in transcript annotation databases (Ensembl, Refseq). Skipping of exon 7 is predicted to cause a frameshift which should trigger NMD. Zhou and colleagues first investigated the mechanism of FUS autoregulation by inhibiting NMD and looking at changes in splicing of exon 7 (Zhou et al., 2013). The exon-skipping transcript increased when FUS was overexpressed and decreased when FUS was knocked down, suggesting this splicing event to be the autoregulatory mechanism. However this mechanism does not explain the high sequence conservation of the diffuse FUS binding pattern across introns 6 and 7 (Fig. 6.9A). When examining RNA-sequencing coverage of the FUS gene in the Fratta, Bozzoni and Dupuis datasets, I could not observe any changes in the inclusion of exon 7 in any sample. However I could observe a strong retention of both introns 6 and 7 that decreased in the presence of FUS NLS mutations. This phenomenon can be seen in all three datasets (Fig. 6.9B), despite the baseline level of retention in wildtype cells being highly variable. We generated RNA sequencing data from mice heterozygous for the Δ14 NLS mutation. Comparing wildtypes, heterozygous and homozgous FUS Δ14 samples showed that retention of introns 6 and 7 decreased with dose of the mutation (Fig. 6.9A/B) .

We designed an RT-PCR assay to validate the intron retention changes with a three primer method that could amplify a spliced transcript spanning *Fus* exons 6-7-8-9 with a second band for either exon 6-intron 6 or intron 7-exon8. Intron retention decreased in a mutation dose-dependent manner (intron 6 P=5.7e-4; intron 7 P=8.7e-4; ANOVA; Fig. 6.9C). We failed to detect a band corresponding to the skipping of exon 7 in any sample.

Retaining two introns would be expected to cause NMD through premature stop codons, which are abundant in both intron 6 and 7. This has been proposed as the main degradation pathway for intron retention transcripts (Wong et al., 2013), although the nuclear retention and elimination (NRE) has also been implicated (Yap and Makeyev, 2013). To test whether the intron retention FUS transcript undergoes NMD we reran the RT-PCR experiments after inhibiting translation in mouse fibroblasts with cycloheximide for 6 hours. This should cause NMD to be inhibited as NMD requires transcripts to be bound to the ribosomes. No effect on intron retention by cycloheximide treatment was seen in either wildtype or FUS Δ14 homozygous fibroblasts (Fig. 6.9D). As a positive control I used the inclusion of a known NMD-sensitive exon in *Srsf7* (Edwards et al., 2016), which increased in both genotypes.

These experiments suggest that depleting nuclear FUS protein downregulates the production of an intron retention transcript insensitive to nonsense-mediated decay.

6.6 Discussion

This study is the largest transcriptome-wide assessment of FUS function to date. By combining three separate datasets of FUS knockdown and NLS mutation I have been able to discover a large repertoire of differentially expressed genes and differential splicing events that have not previously been reported. Comparison of the two conditions demonstrates that NLS mutations predominantly act to reduce nuclear FUS, as the majority of expression and splicing changes overlap and change in the same direction. This leaves only a small number of NLS mutation-specific gene expression changes and almost no mutation-specific splicing changes other than in FUS itself. By studying the mutation-specific splicing changes in FUS I have discovered a novel intron retention transcript that could explain the mechanism by which FUS regulates its own translation.

The lack of a mutation specific effect in gene expression and splicing appears to contrast with previous work. Studies on wildtype and mutant FUS have found evidence of a specific toxic effect of NLS mutant FUS (Verbeeck et al., 2012; Mitchell et al., 2013; Shiihashi et al., 2016; Scekic-Zahirovic et al., 2016). Scekic-Zahirovic and colleagues observed motor neuron loss in embryonic mice homozygous for NLS mutations, but not in the embryos homozygous for a FUS knockout allele (Scekic-Zahirovic et al., 2016), one of the datasets used in this study. Their analysis of gene expression saw separation between the two conditions. My subsequent analysis with joint modelling and relaxed overlap thresholds demonstrates that there is in fact little separation. The specific toxicity of NLS mutant FUS to motor neurons may therefore be due to the other molecular roles of FUS not measured in this study.

The joint modelling approach has allowed me to combine three different RNA sequencing datasets to produce a consensus set of RNA phenotypes. Due to the stochastic nature of transcription and splicing combined with the small sample sizes employed, these datasets are inherently variable. By combining repeated observations under the same genetic conditions we can better discover true FUS-associated changes. While joint modelling increases detection power, it rewards conformity between experiments and I cannot discount the possibility that each dataset has its own cell-type or mutation-specific effects confounded with the dataset itself. It is arguable that these dataset-specific changes are more likely than not to be biologically irrelevant such as transgene insertion artefacts. In addition, with increased power I can now detect changes in expression and splicing at very small effect sizes. While this provides more information, the biological relevance of these small changes is harder to investigate. These changes are less likely to be a direct result of FUS RNA or protein interaction and may emerge through multiple downstream effectors or pathways.

Employing a relaxed significance threshold demonstrated that the majority of changes are shared between FUS knockout and NLS mutation. The direction of change was identical for 99% of overlapping genes with a clear bias towards stronger changes in the knockout (Fig. 6.3B). This is unsurprising, as NLS mutations do not completely abolish FUS nuclear import (Scekic-Zahirovic et al., 2016; Devoy et al., 2017). There is a widespread downregulation of neuronal and synaptic genes in FUS nuclear depletion. This could be due to defects in RNA transport as FUS has been found in RNA transport granules (Kanai et al., 2004; Fujii and

Takumi, 2005). Conversely, RNA binding genes were upregulated in both conditions. FUS is known to interact on a protein-protein level with multiple splicing factors (Yang et al., 1998; Meissner et al., 2003; Groen et al., 2013) and particularly members of the U1 snRNP complex (Sun et al., 2015; Yu et al., 2015). CLIP experiments show that FUS binds a number of these factors at the RNA level (Nakaya et al., 2013). Therefore FUS is part of a complex network of RNA and protein interactions with multiple splicing factors. However, why FUS loss acts to upregulate these factors is yet to be discovered. FUS has been shown to act as both a splicing repressor and enhancer depending on its position, particularly at minor/U12 introns (Reber et al., 2016). This phenomenon may explain some of my gene expression changes and is worth investigating further. I managed to replicate the finding of upregulation of the X-linked lymphocyte receptor (*Xlr*) cluster of genes, first seen in adult FUS knockout mice (Kino et al., 2015), suggesting that FUS loss causes chronic dysregulation of these genes. As Xlr gene overexpression can alter dendritic spine growth and are regulated by the Cux1/2 transcription factors (Cubelos et al., 2010), the mechanism of how FUS loss leads to their upregulation is intriguing. No FUS iCLIP clusters were observed overlapping any Xlr gene, discounting direct post-transcriptional regulation. As these genes appear to be mouse-specific they are not directly relevant to disease but could provide a useful model for the effect of FUS on transcription.

Overlapping the joint splicing models demonstrated that almost all splicing changes attributed to NLS mutations could also be observed in FUS knockouts. I did not observe a convincing set of mutation-specific splicing changes which is unsurprising as splicing is a nuclear function. The overlap between knockdown and expression has been seen in cells where overexpression of NLS mutant FUS cannot rescue splicing changes in FUS knockdown cells (Sun et al., 2015). Due to the multiplicity of FUS interactions between multiple splicing factors, disturbance of any or all of these factors could be confounding the splicing changes. For example, the splicing profiles of FUS NLS mutations overlap those from knockdowns of known interactor SMN (Mirra et al., 2017). I observed approximately double the number of 5´ splice site changes than 3´ splice site changes. This could be due to the shorter length of the consensus 5´ splice site sequence compared to the 3´ making cryptic splice sites more likely by chance, but it may hint at a consequence of the interaction between FUS and the U1 snRNP. Although the interaction between FUS and the U1 snRNP has been well studied (Nakaya et al., 2013; Yu et al., 2015; Yu and Reed, 2015), my analysis has uncovered a large array of splicing events to study the role of FUS in splicing.

The dominance of retained introns and complex events (where multiple splice sites are altered simultaneously) points to a more nuanced picture of splicing changes than has previously been investigated in RNA-binding protein biology. I concede however that the large number of complex events seen in the joint splicing models is probably a result of the juction-centric method I used, as well as analysing all the data jointly. An isoform-centric approach where reads are assigned to specific transcripts (Trapnell et al., 2010; Bray et al., 2016) perhaps would more sensitively tease out the different changes that are happening in complex introns.

Nevertheless, complex events as well as intron retention events are enriched in FUS iCLIP

binding and are more likely to affect splicing factors than the cassette exons I observed. This suggests a more subtle role for FUS in splicing than simply altering alternate cassette exons which have been the focus on previous studies on FUS and splicing (Rogelj et al., 2012; Lagier-Tourenne et al., 2012; Ishigaki et al., 2012; Honda et al., 2014; Scekic-Zahirovic et al., 2016). Interestingly, widespread intron retention among splicing factor genes has been seen in human motor neurons derived from induced pluripotent stem cells containing mutations in the ALS gene *VCP* (Luisier et al., 2018). This study observed splicing changes over several developmental time points and found a developmental delay in splicing, particularly in intron retention. These retained introns were enriched in RNA-binding gene ontology terms, similarly to what I have observed. I overlapped their set of 143 genes that contain differentially retained introns with the set of 219 genes I found to have either retained introns or complex events shared between FUS knockdown and FUS mutation - the "Overlapping" set. 12 genes were common between the two sets, a small but significant overlap (P = 1e-5, Fisher exact test). Significantly, this included FUS itself. This suggests a wider role for intron retention changes in FUS and other splicing factor intron retention in ALS beyond the study of FUS mutations. FUS can bind the C-terminus of RNA polymerase II (Schwartz et al., 2012) and a study on the role of FUS in alternate polyadenylation suggested that FUS binding to pre-mRNA can stall RNA polymerase II (Masuda et al., 2015). Transcriptional speed is known to affect splicing as pausing transcription can allow more time for splicing factors to bind weaker affinity sequences (Kornblihtt et al., 2004). Splicing factor genes themselves are particular vulnerable to changes in transcription speed (Ip et al., 2011). This suggests that the splicing factor intron retention events seen in FUS knockout and mutation, including *Fus* itself, may arise from FUS interacting with transcription rather than binding and repressing particular splice sites.

The high sequence conservation seen in retained introns suggests a regulatory role for these splicing events, despite the unexpectedly low overlap between splicing changes and differentially expressed genes. One example of this is the U1 splicing factor *Snrnp70*, whose highly conserved intron retention transcript has previously been shown to increase in FUS knockdown, which should lead to increased degradation of *Snrnp70* mRNA through nonsense-mediated decay (Nakaya et al., 2013). Two overlapping splicing events are found in *Snrnp70* that suggest the NMD-sensitive conserved intron section is more retained in FUS knockout. I observed upregulation of *Snrnp70* in the FUS knockouts, although this gene-level metric may confound the change in isoforms suggested by the splicing analysis. The other FET family members *Taf15* and *Ewsr1* are both upregulated and both contain conserved intron retention events that are more frequently skipped in FUS depletion. For *Taf15*, its upregulation when FUS is depleted is probably a redundancy mechanism due to the shared RNA motifs and target genes of the two proteins (Ibrahim et al., 2013; Kapeli et al., 2016). The splicing changes I identify in both *Taf15* and *Ewsr1* suggest a mechanism for how their expression is increased when nuclear FUS is depleted. A previous study did not find the mRNA stability of *Taf15* to be changed in FUS knockdown (Colombrita et al., 2012) but this requires re-evaluation in light of the new data.

By studying our RNA-seq data in all three NLS mutant datasets I observed a novel intron

retention isoform whose retention inversely correlates with the dose of NLS mutation. As the change can also be observed in mice heterozygous for the FUS knockout allele it is unlikely to be due to the mutant NLS itself (data not shown). Although cassette exon skipping of exon 7 was previously suggested to be the splicing event of FUS autoregulation (Zhou et al., 2013), neither our RNA-seq or RT-PCR observed exon 7 skipping in the presence of NLS mutation. This could simply be due to exon 7 skipping being present but at much lower level than intron retention changes. Zhou and colleagues could not have detected intron retention changes as the length of the retention transcript exceeds the maximum amplicon length for PCR. Our three-primer approach robustly demonstrated the changes in retention occur in both introns 6 and 7 but they cannot conclusively prove simultaneous retention of both introns. This would require a long-read sequencing technique. The NMD inhibition experiments suggest that the FUS intron retention transcript is insensitive to nonsense-mediated decay. Although, cycloheximide inhibits translation in order to inhibit NMD and off-target target effects cannot be excluded. To elucidate this as the mechanism of FUS regulation, further experiments are needed to validate this in human cells. One possible mechanism is that the transcript is instead detained in the nucleus. This is a regulatory pathway proposed for intron retention transcripts to prevent translation (Boutz et al., 2015). A analogous mechanism for autoregulation has been proposed for TDP-43, where the binding of TDP-43 protein to *Tardbp* mRNA shifts polyadenylation to create a long 3´ UTR transcript which is detained in the nucleus and degraded by the exosome. The long transcript can also be spliced to form NMD-sensitive isoforms which can be exported to the cytoplasm and then degraded (Ayala et al., 2011; Koyama et al., 2016). It is unclear why the system would require two separate degradation pathways. In the light of this, the discovery of the FUS intron retention transcript could be a complementary mechanism to the one set out by Zhou and colleagues. This would then answer the outstanding question of the high evolutionary conservation of both introns and the length of FUS binding throughout.

This study combines multiple RNA sequencing experiments to better understand and catalogue changes in gene expression and RNA splicing in response to FUS nuclear depletion. My analysis suggests that these changes are shared due to the extreme overlap and concordance of direction of gene expression and splicing. It does not conclusively link any of these alterations to motor neuron toxicity and degeneration, which is only seen in NLS mutant embryos and not in FUS knockouts (Scekic-Zahirovic et al., 2016). It is unclear whether toxicity can be attributed to the 186 mutation-specific genes due to the sharing of gene ontology categories with the overlapping set (RNA processing and synaptic genes).

The discovery of a novel mechanism for FUS autoregulation is valuable for understanding the role of FUS in disease. Autoregulation acts to maintain protein homeostasis but FUS mutations can interfere with this. When FUS nuclear import is abolished by mutations in the NLS, there is no way for FUS protein to regulate its own expression, leading to ever-more cytoplasmic FUS. Increased FUS protein has been seen in ALS patients with a mutation in the FUS 3'UTR, which alters a microRNA binding site (Modigliani et al., 2014), highlighting the relevance of protein homeostasis to disease. High concentrations of cytoplasmic FUS protein increase the likelihood for aggregation which NLS mutant FUS is

more prone to do (Bosco et al., 2010). This increased propensity may be due to the recent finding that NLS acts to solubilise FUS aggregates through binding to Transportin (Guo et al., 2018; Yoshizawa et al., 2018; Hofweber et al., 2018). In FUS ALS, patients only have a single copy of NLS mutant FUS. Selectively removing the mutant copy, at either a DNA or RNA level, should all but prevent cytoplasmic mislocalisation while maintaining nuclear FUS levels due to the compensatory nature of autoregulation. Indeed, mice heterozygous for a FUS knockout allele have FUS mRNA and protein levels at 75% of wildtypes and exhibit no motor neuron degeneration (**?**). This could feasibly be increased to 100% by modulating FUS autoregulation by promoting the splicing of introns 6 and 7. This work expands our understanding of the complex role of FUS in gene expression and splicing and uncovers a new model for FUS autoregulation. I hope this will be useful to the disease field in designing targeted therapies for FUS ALS.

7 Conclusions

In this thesis I have analysed multiple RNA-seq datasets to uncover new insights into TDP-43 and FUS. I have explored the physiological roles of the two proteins in RNA regulation through loss-of-function experiments. Furthermore, by using mutant mice as models of disease, I have uncovered mechanisms by which disease-associated mutations can impart a gain of function. This comparison between loss and gain of functionality will be crucial for understanding the onset and course of ALS/FTD.

In chapter 3 I developed software to discover and classify the full extent of cryptic splicing repression by TDP-43 and applied it to multiple mouse and human datasets. I demonstrated that this repression is due to a combination of TDP-43 binding motifs and strong splice site sequences co-occurring within poorly conserved introns. This mechanism was not seen in FUS knockdown, suggesting a divergence in function for the two proteins. Later work has shown that cryptic exon repression is seen in a range of RNA-binding proteins, including the ALS-associated MATRIN3, SFPQ and TIA1.

In chapter 4 I analysed data from a novel FUS mutant mouse which modelled aggressively early onset FUS ALS. The mice are heterozygous for a mutation which removes the FUS nuclear localisation signal, leading to cytoplasmic mislocalisation of the mutant FUS protein without depleting wildtype FUS from the nucleus. By looking across two tissues and time-points I identified a specific transcriptomic signature seen only in late-adult mouse spinal cord. This was dominated by changes in mitochondrial and ribosomal genes. Splicing changes at this tissue and time point were restricted to FUS itself and a small number of other genes including the myelin protein *Mbp* and the FUS-interacting RNA-binding protein *Ewsr1*.

In chapter 5 I compared two TDP-43 mutant mice lines. One exhibited a loss of splicing function due to the mutation affecting an RNA-recognition motif. As expected, this led to the inclusion of cryptic exons. The other had a mutation within the TDP-43 low-complexity domain, a hotspot for ALS mutations. I demonstrated that this mutation imparts a gain of splicing function characterised by a set of aberrant constitutive exon skipping events. Some of these "skiptic" exons could be identified in TDP-43 mutant human patients. The gain of splicing function could arise from a change in TDP-43 autoregulation or from a shift in how TDP-43 interacts with other splicing factors.

In chapter 6 I compared three independent studies where FUS was either knocked out or mutated to remove its nuclear localisation signal in mouse neuronal tissues. Unlike chapter 3, these mutant mice were homozygous for their mutations. My joint analyses combining either the knockout or mutant FUS samples together increased the detectable number of RNA regulation changes compared to analysing the datasets separately. Overlapping the two joint models demonstrated significant similarity between the two FUS conditions in

both gene expression and splicing, suggesting that FUS NLS mutations deplete FUS from the nucleus. There were few RNA regulation changes found specifically in the mutations. Looking in detail at the FUS locus itself, I identified a novel mechanism by which FUS could regulate its own translation through intron retention. This autoregulation mechanism may be complementary to one previously reported exon skipping (Zhou et al., 2013).

All together, my results suggest specific consequences for RNA regulation occuring with both loss of TDP-43 or FUS from the nucleus and from cytoplasmic mislocalisation.

7.1 Issues arising

Loss and gain of splicing function in TDP-43 and FUS

Both loss and gains of function are relevant to ALS/FTD because of the nuclear depletion of TDP-43 and FUS in the end-stage of the disease is accompanied by cytoplasmic aggregates. The formation of aggregates may be encouraged by rare mutations. My work has attempted to understand the impact on splicing that could occur in human ALS/FTD patients.

For TDP-43, knocking down or knocking out the protein in human or mouse cells uncovered a physiological role for TDP-43 in repressing non-conserved cryptic exons. This was also seen in the RRM2mut mouse line where one of the RNA-recognition motifs was mutated. However, a mutation in the low-complexity domain, where most ALS mutations are found, led to the inverse phenomenon, the skipping of constitutive exons. These two phenomena appear to be mutually exclusive, suggesting that the LCDmut does not cause a loss of splicing function but instead an aberrant gain of function. Although the mechanism for this has not been identified, it points to a role for TDP-43 in maintaining constitutive exon splicing. This was reinforced by wild-type TDP-43 iCLIP binding peaks observed on and around the "skiptic" exons.

Since I began studying TDP-43 and cryptic exons, many more RNA-binding proteins have been found to have a role in cryptic exon repression including PTBP1/2, RBM17, hNRNP L and MATR3 (Ling et al., 2016; Tan et al., 2016; McClory et al., 2018; Attig et al., 2018). My own assessment of ENCODE knockdown data on a large number of RNA-binding proteins suggests that cryptic exon repression is a near-universal phenomenon. This included the majority of ALS/FTD associated proteins tested, with the exception of FUS. Skiptic splicing is the inverse of cryptic splicing. The two phenomena represent opposite ends of aberrant splicing caused by extreme changes in protein homeostasis. The skiptic exons emerge due to rare mutations in TDP-43 and the cryptic exons through artificial (or biological) TDP-43 depletion. Due to the tight regulation of the levels of RNA-binding proteins, these aberrant splicing events shouldn't occur within the boundaries of normal physiology. Therefore, there is an outstanding question as to when in the course of disease these events first begin, whether for use as a biomarker or as a therapeutic target. It would also be interesting to look for more examples of skiptic exons in other RNA-binding proteins.

The discovery of cryptic and skiptic splicing leads to multiple questions on the nature of

cooperation between splicing factors. Most introns are bound by multiple splicing factors and yet a number of them are specifically mis-spliced as cryptic exons when TDP-43 is depleted. This suggests that in certain introns TDP-43 cannot be substituted for when it is depleted. This is a lack of biological redundancy. Where these non-redundant sequences appear would depend on which splicing factors are expressed at the time. This is shown by a cell-type specificity for particular TDP-43 cryptic exons, as seen by myself and others (Jeong et al., 2017). It would also suggest that each set of cryptic exons in a particular cell type are unique to each RNA-binding protein. This could be tested with the ENCODE corsortium data.

Another key mechanism to be understood in RNA biology is that of autoregulation. Both TDP-43 and FUS proteins bind their own transcripts and modulate their own translation. However both proteins appear to do this by a complex set of mechanisms involving both NMD and nuclear detention. The TDP-43 autoregulatory locus is the 3´ UTR, the site of TDP-43 binding (Polymenidou et al., 2011). This binding alters 3´ UTR length to create a nuclear detained isoform, accompanied by NMD-sensitive spliced UTR isoforms (Ayala et al., 2011; Koyama et al., 2016). Both the RRM2mut and LCDmut mutations appear to alter TDP-43 autoregulation and this may account for the gain in splicing function seen in LCDmut. FUS binds introns 6 and 7 (Lagier-Tourenne et al., 2012), which was previously identified to cause skipping of the central exon to create an NMD-sensitive transcript (Zhou et al., 2013) My study of the FUS introns 6 and 7 has suggested that retention of both introns creates an NMD-insensitive transcript that is potentially detained in the nucleus. I would predict that these mechanisms play a role in ALS/FTD where either TDP-43 or FUS are mislocalised from the nucleus to the cytoplasm in affected neurons and glia. Reduction in autoregulation would lead to a positive feedback mechanism where enhanced translation would then increase cytoplasmic TDP-43 or FUS.

Beyond autoregulation there is the intriguing phenomena of *cross-regulation*, where depleting one splicing factor impacts the expression or splicing of many other splicing factors. This has been previously seen in a study of multiple hNRNP proteins (Huelga et al., 2012). This study demonstrated asymmetric relationships between pairs of proteins, where depletion of one protein would upregulate another but not vice versa. With FUS nuclear depletion I observed upregulation of the FET family members *Taf15* and *Ewsr1* accompanied by changes in intron retention seen in both transcripts. This suggests a feedback mechanism where reduction in nuclear FUS leads to a direct increase in FET family levels, potentially to compensate for reduced nuclear FUS due to the overlapping motifs and targets of the three proteins, although TAF15 knockdown fails to alter *FUS* gene expression (Kapeli et al., 2016). Many other splicing factor genes were upregulated and alternatively spliced in FUS nuclear depletion, suggesting a generalised response of the splicing machinery and associated factors to loss of one particular protein. This phenomena must depend on nuclear FUS levels, as this was only seen in the homozygous NLS mutant mice and not in the FUS Δ14 adult mice studied in chapter 4. There I found no large changes in splicing factor gene expression and splicing except for *Fus* and *Ewsr1*. It has been previously observed that TDP-43 binds the same region of the *FUS* transcript that FUS binds itself, linking the two

proteins (Lagier-Tourenne et al., 2012). It would be interesting to look at changes in the putative FUS autoregulatory transcript in my collection of TDP-43 depletion and mutation mice.

The TDP-43 LCDmut mouse was created to study the effect of mutation on the low-complexity domain. It is striking that low-complexity domains are found in almost all RNA-binding proteins associated with ALS/FTD. In TDP-43 most disease-associated mutations are found there. In disease, the low complexity domains have been implicated in protein aggregation and stress granule formation but it still unclear what the physiological role of these domains are. One hypothesis is that these domains allow the assembly of multiple RNA-binding proteins on a transcript (Gueroussov et al., 2017). This would allow splicing factors to be brought in contact with a transcript that lacks known sequence motifs and would answer why many altered splicing events seen in my work have no nearby CLIP binding peaks.

RNA-seq: the value of replication and annotation-free analysis

RNA-seq allows phenomena previously seen in individual genes to be demonstrated genome-wide. In the example of TDP-43 and splicing, initial studies on individual cassette exons have given way to hundreds of splicing events being found across different tissues and developmental time points. The creation of RNA maps allow general principles of a protein's effect on splicing to be demonstrated and clearly visualised. We can build up complete pictures of what an RNA-binding protein can do.

With each RNA-seq study, new phenomena and mechanisms can be identified. Crucially, as each published dataset is available for re-analysis, these phenomena can be replicated in multiple species and tissues. By combining multiple published datasets in chapter 3 I have strengthened the case for TDP-43 and cryptic exons. The novel mechanism of FUS autoregulation found in chapter 6 is much stronger for having been observed in 3 independent studies. Replication of RNA phenomena is crucial as there is so much variability between experiments caused by RNA extraction, library preparation, and analysis software. With so many findings arising from individual RNA-seq experiments in a particular species or cell type there is a real need for these to be replicated.

Annotation initiatives have been successful in cataloguing genes and transcripts in many species. They are very useful for any scientist interested in RNA biology by providing the genomic coordinates for a set of experimentally-validated splicing isoforms. However, there is great value in looking outside what is given to you by annotation. In under-studied cell types like neurons there is a wealth of unannotated splicing in normal cells. When perturbing splicing one could be expect there to be unusual or aberrant splicing that would not normally be observed.

Both cryptic exons and skiptic exons are examples of novel splicing events that are not detected when relying on annotated transcripts alone. I anticipate more developments in the RNA-seq field on tools that discover and quantify novel splicing. Beyond this, there is a

pressing need to verify novel RNA isoforms at the protein level, if they are indeed translated. This will require new computational tools for shotgun mass spectrometry.

7.2 Future directions

Translating work into human patients

It is regrettable that throughout my PhD I have not had the opportunity to explore more human RNA-seq samples, particularly those from human ALS/FTD patients. This is partly due to the inherent variablity between human subjects but also because of the technical hurdles in analysis. Unlike laboratory strains of mice, humans are highly heterogenous in both their genetics and the environmental exposures. Even patients with identical mutations in a disease-associated gene could expected to have many distinct modifying genetic variants as well as completely different life experiences. Therefore any case-control study on transcriptomics in humans requires a much larger sample sizes to account for the increased variability in measurements of gene expression and splicing. Alternatively, this genetic variability can be exploited by a study design that correlates genetic variation with other features. Genome-wide association studies (GWAS) correlate common or rare genetic variation with clinical phenotypes. The latest GWAS on ALS combines nearly 100,000 samples (Nicolas et al., 2018). Gene expression and splicing phenotypes from RNA-seq can be correlated with genetic variation to produce lists of genetic variants associated with specific changes in RNA regulation. These genetic variants can then also be associated with clinical phenotypes, as found in a large Alzheimer's disease cohort (Raj et al., 2018).

There are also technical considerations in sequencing post-mortem brains. Both diseased and control donor brains will often come from human patients who were kept on artificial respiration before death. This is known to affect RNA quality in post-mortem tissue (Durrenberger et al., 2010). Most post-mortem tissue RNA-seq samples are prepared using total RNA libraries to avoid bias towards the 3´ end that comes with polyA+ purification of degraded RNA. Total RNA libraries will dilute the amount of reads that align to protein-coding genes, requiring higher coverage and greater cost per sample. There are now large consortia established to collect and sequence both RNA and DNA from large numbers of patient tissues. As part of my post-doctoral studies I am looking forward to working with these samples.

An alternative route for studying ALS/FTD in humans is to grow patient neurons from induced pluripotent stem cells (iPSCs). These can be created from skin samples from patients which are reverted a pluripotent state and then differentiated into disease-specific cell types such as motor neurons. What is particularly powerful is that patient cell lines can be edited to remove the causative mutation, creating an isogenic control line. This effectively controls for human genetic variability, and so any phenotypes can be directly associated to the effect of the initial mutation. Studies comparing iPSC-derived motor neurons are providing powerful insights into changes in RNA regulation within motor neurons, as seen

for microRNAs in FUS (De Santis et al., 2017) and intron retention in VCP (Luisier et al., 2018).

Diving deeper into RNA regulation

All the work in my PhD was done with bulk-cell short-read RNA-seq. There are limitations to this approach that are both biological and technical. Bulk RNA sequencing samples RNA from all cells in a population. This limits the observations that can be made to changes in expression and splicing that can be seen across a wide range of cells. It also demands low variability between cell type mixtures between conditions to avoid this confounding results. A rapidly developing solution to this hurdle is single-cell RNA sequencing, where individual cells are first separated from a tissue and sequencing libraries prepared using microfluidic technology. Studies are now emerging of 10s of 1000s of individual cells being profiled. This has allowed the entire mouse nervous system to be classified into different cell types (Zeisel et al., 2018). The technology for this is still in its infancy and there is a problem with sequencing throughput to get enough reads per sample to robustly measure splicing. It would be very exciting if this technology could be applied to ALS/FTD patient brains.

Bulk sequencing of cells also removes the distinction of RNA localisation within different cellular compartments. For example, both FUS and TDP-43 bind transcripts in the cytoplasm and have some role in axonal transport (Fallini et al., 2012; Fujii and Takumi, 2005). Separating RNA into nuclear or cytoplasmic fractions would give information about nuclear retention of particular transcript, as I hypothesise for the FUS intron retention transcript identified in chapter 6. This can be done biochemically by fractionation, or physically by the use of growing neuronal cells in compartments to physically isolate RNA (Taliaferro et al., 2016). Alternatively, sequencing the nascently transcribed RNA would allow observation of unstable transcripts. The nagging question from chapter 3 is that if I predict most cryptic exons to be degraded by NMD, why do I see any at all? If we were to sequence nascently transcribed RNA in TDP-43 depletion and controls we could identify unstable transcripts that rapidly degrade upon export to the cytoplasm. This may demonstrate cryptic splicing repression by TDP-43 to be much more widespread and why certain cryptic exon transcripts appear to evade NMD.

The other hurdle with the current state of RNA research is the length of sequencing reads. The current gold standard for Illumina sequencing is paired end 300bp reads. This means that splicing analysis requires the reconstruction of complex isoforms from individual short fragments. Long-read sequencing techniques such as Pacific Biosciences and Oxford Nanopore promise to sequence entire transcripts in a single read. This would greatly simplify analysis but are still developing. Both techniques suffer from a high error rate which will complicate read alignment. Throughput is also an issue. It is not yet known what sequencing depth is sufficient to analyse alternate splicing with long reads. Oxford Nanopore technology can be used to sequence native RNA rather than cDNA, which allows the observation of RNA modifications. A compromise between long and short reads are synthetic long reads, where short Illumina reads are generated from an individual transcript suspended in

an oil drop. This has allowed the reconstruction of individual transcripts and the curious finding that some splicing events are co-regulated within a transcript (Tilgner et al., 2015).

As with the emergence of RNA-seq 10 years ago, technical breakthroughs in RNA biology will uncover new and exciting types of RNA regulation. I hope these developments will enhance our understanding of the molecular mechanisms of ALS/FTD.

Bibliography

Acevedo-Arozena, A., Wells, S., Potter, P., Kelly, M., Cox, R. and Brown, S. Enu mutagenesis, a way forward to understand gene function. *Annu Rev Genomics Hum Genet*, 9: 49–69, 02 2008.

Afroz, T., Hock, E.M., Ernst, P., Foglieni, C., Jambeau, M., Gilhespy, L.A., Laferriere, F., Maniecka, Z., Plückthun, A., Mittl, P. et al. Functional and dynamic polymerization of the ALS-linked protein TDP-43 antagonizes its pathologic aggregation. *Nature Communications*, 8(1):1–14, 2017. doi: 10.1038/s41467-017-00062-0.

Alami, N.H., Smith, R.B., Carrasco, M.A., Williams, L.A., Winborn, C.S., Han, S.S.W., Kiskinis, E., Winborn, B., Freibaum, B.D., Kanagaraj, A. et al. Axonal transport of TDP-43 mRNA granules is impaired by ALS-causing mutations. *Neuron*, 81(3):536–543, 2013. doi: 10.1016/j.neuron.2013.12.018.

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L. and Feuk, L. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural and Molecular Biology*, 18 (12):1435–1440, 2011. doi: 10.1038/nsmb.2143.

Anders, S., Reyes, A. and Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Research*, pages 2008–2017, 2012. doi: 10.1101/gr.133744.111.Freely.

Anders, S., Pyl, P.T. and Huber, W. HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.

Anderson, P. and Kedersha, N. Stress granules: the Tao of RNA triage. *Trends in Biochemical Sciences*, 33(3):141–150, 2008. doi: 10.1016/j.tibs.2007.12.003.

Andersson, M.K., Ståhlberg, A., Arvidsson, Y., Olofsson, A., Semb, H., Stenman, G., Nilsson, O. and Åman, P. The multifunctional FUS, EWS and TAF15 proto-oncoproteins show cell type-specific expression patterns and involvement in cell spreading and stress response. *BMC Cell Biology*, 9:1–17, 2008. doi: 10.1186/1471-2121-9-37.

Arai, T., Hasegawa, M., Akiyama, H., Ikeda, K., Nonaka, T., Mori, H., Mann, D., Tsuchiya, K., Yoshida, M., Hashizume, Y. et al. TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochemical and Biophysical Research Communications*, 351(3):602–611, 2006. doi: 10.1016/j.bbrc.2006.10.093.

Arnold, E.S., Ling, S.C., Huelga, S.C., Lagier-Tourenne, C., Polymenidou, M., Ditsworth, D., Kordasiewicz, H.B., McAlonis-Downes, M., Platoshyn, O., Parone, P.A. et al. ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neu-

ron disease without aggregation or loss of nuclear TDP-43. *Proceedings of the National Academy of Sciences*, 110(8):E736–45, 2013. doi: 10.1073/pnas.1222809110.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. Gene Ontology: Tool for The Unification of Biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Attig, J., De Los Mozos, I.R., Haberman, N., Wang, Z., Emmett, W., Zarnack, K., Koenig, J. and Ule, J. Splicing repression allows the gradual emergence of new alu-exons in primate evolution. *eLife*, 5(0):1–27, 2016. doi: 10.7554/eLife.19545.

Attig, J., Agostini, F., Gooding, C., Chakrabarti, A.M., Singh, A., Haberman, N., Zagalak, J.A., Emmett, W., Smith, C.W., Luscombe, N.M. et al. Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. *Cell*, 174(5):1067–1081.e17, 2018. doi: 10.1016/j.cell.2018.07.001.

Ayala, Y.M., Zago, P., D'Ambrogio, A., Xu, Y.F., Petrucelli, L., Buratti, E. and Baralle, F.E. Structural determinants of the cellular localization and shuttling of TDP-43. *Journal of Cell Science*, 121(22):3778–3785, 2008. doi: 10.1242/jcs.038950.

Ayala, Y.M., Conti, L.D., Dhir, A., Romano, M., Ambrogio, A.D., Tollervey, J., Ule, J., Baralle, M., Baralle, F.E., De Conti, L. et al. TDP-43 regulates its mRNA levels through a negative feedback loop. *The EMBO Journal*, 30(2):277–288, 2011. doi: 10.1038/emboj.2010.310.

Bai, Y., Ji, S. and Wang, Y. IRcall and IRclassifier: two methods for flexible detection of intron retention events from RNA-Seq data. *BMC Genomics*, 16 Suppl 2:S9, 2015.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue):W202–8, 2009.

Bannwarth, S., Ait-El-Mkadem, S., Chaussenot, A., Genin, E.C., Lacas-Gervais, S., Fragaki, K., Berg-Alonso, L., Kageyama, Y., Serre, V., Moore, D.G. et al. A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement. *Brain*, 137(8):2329–2345, 2014. doi: 10.1093/brain/awu138.

Barmada, S.J., Skibinski, G., Korb, E., Rao, E.J., Wu, J.Y. and Finkbeiner, S. Cytoplasmic Mislocalization of TDP-43 Is Toxic to Neurons and Enhanced by a Mutation Associated with Familial Amyotrophic Lateral Sclerosis. *Journal of Neuroscience*, 30(2):639–649, 2010. doi: 10.1523/JNEUROSCI.4988-09.2010.

Benajiba, L., Ber, I.L., Camuzat, A., Lacoste, M., Thomas-Anterion, C., Couratier, P., Legallic, S., Salachas, F., Hannequin, D., Decousus, M. et al. TARDBP mutations in motoneuron disease with frontotemporal lobar degeneration. *Annals of Neurology*, 65(4):470–474, 2009. doi: 10.1002/ana.21612.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995. doi: 10.2307/2346101.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008. doi: 10.1038/nature07517.

Berget, S.M., Moore, C. and Sharp, P.A. A spliced sequence at the 5'-terminus of adenovirus late mRNA. *Proceedings of the National Academy of Sciences USA*, 74(8):3171–3175, 1977. doi: 10.1073/PNAS.74.8.3171.

Beyer, A.L. and Osheim, Y.N. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes & development*, 2(6):754–765, 1988. doi: 10.1101/gad.2.6.754.

Blokhuis, A.M., Koppers, M., Groen, E.J.N., van den Heuvel, D.M.A., Dini Modigliani, S., Anink, J.J., Fumoto, K., van Diggelen, F., Snelting, A., Sodaar, P. et al. Comparative interactomics analysis of different ALS-associated proteins identifies converging molecular pathways. *Acta Neuropathologica*, 132(2):175–196, 2016.

Borroni, B., Bonvicini, C., Alberici, A., Buratti, E., Agosti, C., Archetti, S., Papetti, A., Stuani, C., Di Luca, M., Gennarelli, M. et al. Mutation within TARDBP leads to frontotemporal dementia without motor neuron disease. *Human Mutation*, 30(11), 2009. doi: 10.1002/humu.21100.

Bosco, D.A., Lemay, N., Ko, H.K., Zhou, H., Burke, C., Kwiatkowski, T.J., Sapp, P., Mckenna-Yasek, D., Brown, R.H. and Hayward, L.J. Mutant FUS proteins that cause amyotrophic lateral sclerosis incorporate into stress granules. *Human Molecular Genetics*, 19(21):4160–4175, 2010. doi: 10.1093/hmg/ddq335.

Bose, J.K., Wang, I.F., Hung, L., Tarn, W.Y. and Shen, C.K.J. TDP-43 overexpression enhances exon 7 inclusion during the survival of motor neuron pre-mRNA splicing. *Journal of Biological Chemistry*, 283(43):28852–28859, 2008.

Bosque, P.J., Boyer, P.J. and Mishra, P. A 43-kDa TDP-43 Species Is Present in Aggregates Associated with Frontotemporal Lobar Degeneration. *PLoS ONE*, 8(5), 2013. doi: 10.1371/journal.pone.0062301.

Boutz, P.L., Bhutkar, A. and Sharp, P.A. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes and Development*, 29(1):63–80, 2015. doi: 10.1101/gad.247361.114.

Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M. and Blencowe, B.J. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*, 24(11):1774–1786, 2014. doi: 10.1101/gr.177790.114.

Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016. doi: 10.1038/nbt.3519.

Broustal, O., Camuzat, A., Guillot-Noël, L., Guy, N., Millecamps, S., Deffond, D., Lacomblez, L., Golfier, V., Hannequin, D., Salachas, F. et al. FUS mutations in fron-

totemporal lobar degeneration with amyotrophic lateral sclerosis. *Journal of Alzheimer's Disease*, 22(3):765–769, 2010. doi: 10.3233/JAD-2010-100837.

Buckanovich, R.J., Yang, Y.L. and Darnell, R.B. The onconeural antigen nova-1 is a neuron-specific rna-binding protein, the activity of which is inhibited by paraneoplastic antibodies. *The Journal of Neuroscience*, 16(3):1114–1122, 1996.

Buratti, E. and Baralle, F.E. Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of CFTR exon 9. *Journal of Biological Chemistry*, 276(39):36337–36343, 2001a.

Buratti, E., Chivers, M., Kralovicova, J., Romano, M., Baralle, M., Krainer, A.R. and Vorechovsky, I. Aberrant 5' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Research*, 35(13):4250–4263, 2007a.

Buratti, E., Stuani, C., De Prato, G. and Baralle, F.E. SR protein-mediated inhibition of CFTR exon 9 inclusion: molecular characterization of the intronic splicing silencer. *Nucleic Acids Research*, 35(13):4359–4368, 2007b.

Buratti, E. and Baralle, F.E. Characterization and Functional Implications of the RNA Binding Properties of Nuclear Factor TDP-43, a Novel Splicing Regulator of CFTR Exon 9. *Journal of Biological Chemistry*, 276(39):36337–36343, 2001b. doi: 10.1074/jbc.M104236200.

Buratti, E., Dörk, T., Zuccato, E., Pagani, F., Romano, M. and Baralle, F.E. Nuclear factor TDP-43 and SR proteins promote in vitro and in vivo CFTR exon 9 skipping. *The EMBO Journal*, 20(7):1774–1784, 2001. doi: 10.1093/emboj/20.7.1774.

Capauto, D., Colantoni, A., Lu, L., Santini, T., Peruzzi, G., Biscarini, S., Morlando, M., Shneider, N.A., Caffarelli, E., Laneve, P. et al. A Regulatory Circuitry Between Gria2, miR-409, and miR-495 Is Affected by ALS FUS Mutation in ESC-Derived Motor Neurons. *Molecular Neurobiology*, pages 1–17, 2018. doi: 10.1007/s12035-018-0884-4.

Carbon, S., Dietze, H., Lewis, S.E., Mungall, C.J., Munoz-Torres, M.C., Basu, S., Chisholm, R.L., Dodson, R.J., Fey, P., Thomas, P.D. et al. Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Research*, 45(D1): D331–D338, 2017. doi: 10.1093/nar/gkw1108.

Carthew, R.W. and Sontheimer, E.J. Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4):642–655, 2009. doi: 10.1016/j.cell.2009.01.035.

Chakrabarti, A.M., Haberman, N., Praznik, A., Luscombe, N.M. and Ule, J. Data Science Issues in Studying ProteinâĂŞRNA Interactions with CLIP Technologies. *Annual Review of Biomedical Data Science*, 1(1):235–261, 2018. doi: 10.1146/annurev-biodatasci-080917-013525.

Chhangawala, S., Rudy, G., Mason, C.E. and Rosenfeld, J.A. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology*, 16(1):1–10, 2015. doi: 10.1186/s13059-015-0697-y.

Chiang, P.M., Ling, J., Jeong, Y.H., Price, D.L., Aja, S.M. and Wong, P.C. Deletion of TDP-43 down-regulates Tbc1d1, a gene linked to obesity, and alters body fat metabolism. *Proceedings of the National Academy of Sciences*, 107(37):16320–16324, 2010. doi: 10.1073/pnas.1002176107.

Chiò, A., Restagno, G., Brunetti, M., Ossola, I., Calvo, A., Mora, G., Sabatelli, M., Monsurrò, M.R., Battistini, S., Mandrioli, J. et al. Two Italian kindreds with familial amyotrophic lateral sclerosis due to FUS mutation. *Neurobiology of Aging*, 30(8):1272–1275, 2009. doi: 10.1016/j.neurobiolaging.2009.05.001.

Chomczynski, P. and Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry*, 162(1):156–159, 1987. doi: 10.1016/0003-2697(87)90021-2.

Chow, L.T., Gelinas, R.E., Broker, T.R. and Roberts, R.J. An Amazing Sequence Arrangement at the 5' Ends of Adenovirus 2 Messenger RNA. *Cell*, 12:1–8, 1977. doi: 10.1074/jbc.270.36.21411.

Cirulli, E.T., Lasseigne, B.N., Petrovski, S., Sapp, P.C., Dion, P.A., Leblond, C.S., Couthouis, J., Lu, Y.F., Wang, Q., Krueger, B.J. et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science*, 347(6229):1436–1441, 2015. doi: 10.1126/science.aaa3650.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L. and Rice, P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 2009. doi: 10.1093/nar/gkp1137.

Collado-Torres, L., Nellore, A., Frazee, A.C., Wilks, C., Love, M.I., Langmead, B., Irizarry, R.A., Leek, J.T. and Jaffe, A.E. Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Research*, 45(2):e9, 2017. doi: 10.1093/nar/gkw852.

Collins, M.A., An, J., Hood, B.L., Conrads, T.P. and Bowser, R.P. Label-Free LC-MS/MS proteomic analysis of cerebrospinal fluid identifies Protein/Pathway alterations and candidate biomarkers for amyotrophic lateral sclerosis. *Journal of Proteome Research*, 14 (11):4486–4501, 2015.

Colombrita, C., Zennaro, E., Fallini, C., Weber, M., Sommacal, A., Buratti, E., Silani, V. and Ratti, A. TDP-43 is recruited to stress granules in conditions of oxidative insult. *Journal of Neurochemistry*, 111(4):1051–1061, 2009. doi: 10.1111/j.1471-4159.2009.06383.x.

Colombrita, C., Onesto, E., Megiorni, F., Pizzuti, A., Baralle, F.E., Buratti, E., Silani, V. and Ratti, A. TDP-43 and FUS RNA-binding proteins bind distinct sets of cytoplasmic messenger RNAs and differently regulate their post-transcriptional fate in motoneuron-like cells. *Journal of Biological Chemistry*, 287(19):15635–15647, 2012. doi: 10.1074/jbc.M111.333450.

Conicella, A.E., Zerze, G.H., Mittal, J. and Fawzi, N.L. ALS Mutations Disrupt Phase

Separation Mediated by α-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. *Structure*, 24(9):1537–1549, 2016. doi: 10.1016/j.str.2016.07.007.

Couthouis, J., Hart, M.P., Shorter, J., Dejesus-hernandez, M., Erion, R., Oristano, R., Liu, A.X., Ramos, D., Jethava, N., Hosangadi, D. et al. A yeast functional screen predicts new candidate ALS disease genes. *Proceedings of the National Academy of Sciences*, 108 (52):20881–90, 2011. doi: 10.1073/pnas.1109434108.

Couthouis, J., Hart, M.P., Erion, R., King, O.D., Diaz, Z., Nakaya, T., Ibrahim, F., Kim, H.J., Mojsilovic-petrovic, J., Panossian, S. et al. Evaluating the role of the FUS/TLS-related gene EWSR1 in amyotrophic lateral sclerosis. *Human Molecular Genetics*, 21(13): 2899–2911, 2012. doi: 10.1093/hmg/dds116.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, 2004.

Cubelos, B., Sebastián-Serrano, A., Beccari, L., Calcagnotto, M.E., Cisneros, E., Kim, S., Dopazo, A., Alvarez-Dolado, M., Redondo, J.M., Bovolenta, P. et al. Cux1 and Cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. *Neuron*, 66(4):523–535, 2010. doi: 10.1016/j.neuron.2010.04.038.

Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., Geng, J., Zhang, B., Yu, X., Yang, J. et al. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics*, 96(5):259–265, 2010. doi: 10.1016/j.ygeno.2010.07.010.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. et al. Ensembl 2015. *Nucleic Acids Research*, 43(D1):D662–D669, 2015. doi: 10.1093/nar/gku1010.

Curk, T. icount: protein-rna interaction iclip data analysis. *(in preparation)*, 0:0–0, 2016.

Daigle, G.G., Lanson, N.A., Smith, R.B., Casci, I., Maltare, A., Monaghan, J., Nichols, C.D., Kryndushkin, D., Shewmaker, F. and Pandey, U.B. RNA-binding ability of FUS regulates neurodegeneration, cytoplasmic mislocalization and incorporation into stress granules associated with FUS carrying ALS-linked mutations. *Human Molecular Genetics*, 22(6):1193–1205, 2013. doi: 10.1093/hmg/dds526.

Dammer, E.B., Fallini, C., Gozal, Y.M., Duong, D.M., Rossoll, W., Xu, P., Lah, J.J., Levey, A.I., Peng, J., Bassell, G.J. et al. Coaggregation of RNA-binding proteins in a model of TDP-43 proteinopathy with selective RGG motif methylation and a role for RRM1 ubiquitination. *PLoS ONE*, 7(6), 2012. doi: 10.1371/journal.pone.0038658.

De Angelis, M.H., Flaswinkel, H., Fuchs, H., Rathkolb, B., Soewarto, D., Marschall, S., Heffner, S., Pargent, W., Wuensch, K., Jung, M. et al. Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nature Genetics*, 25(4):444–447, 2000. doi: 10.1038/78146.

De Conti, L., Akinyi, M.V., Mendoza-Maldonado, R., Romano, M., Baralle, M. and Buratti, E. TDP-43 affects splicing profiles and isoform production of genes involved in the

apoptotic and mitotic cellular pathways. *Nucleic Acids Research*, 43(18):8990–9005, 2015. doi: 10.1093/nar/gkv814.

de Ferra, F., Engh, H., Hudson, L., Kamholz, J., Puckett, C., Molineaux, S. and Lazzarini, R.A. Alternative splicing accounts for the four forms of myelin basic protein. *Cell*, 43(3 PART 2):721–727, 1985. doi: 10.1016/0092-8674(85)90245-4.

De Santis, R., Santini, L., Colantoni, A., Peruzzi, G., de Turris, V., Alfano, V., Bozzoni, I. and Rosa, A. FUS mutant human motoneurons display altered transcriptome and microRNA pathways with implications for ALS pathogenesis. *Stem Cell Reports*, 9(5): 1450–1462, 2017. doi: 10.1016/j.stemcr.2017.09.004.

Deininger, P. and Prescott, D. Alu elements: know the SINEs. *Genome Biology*, 12(12): 236, 2011.

DeJesus-Hernandez, M., Kocerha, J., Finch, N., Crook, R., Baker, M., Desaro, P., Johnston, A., Rutherford, N., Wojtas, A., Kennelly, K. et al. De novo truncating FUS gene mutation as a cause of sporadic amyotrophic lateral sclerosis. *Human Mutation*, 31(5):1377–1389, 2010. doi: 10.1002/humu.21241.

DeJesus-Hernandez, M., Mackenzie, I.R., Boeve, B.F., Boxer, A.L., Baker, M., Rutherford, N.J., Nicholson, A.M., Finch, N.A., Gilmer, H.F., Adamson, J. et al. Expanded GGGGCC hexanucleotide repeat in non-coding region of C9ORF72 causes chromosome 9p-linked frontotemporal dementia and amyotrophic lateral sclerosis. *Neuron*, 72(2):245–25609, 2011. doi: 10.1016/j.neuron.2011.09.011.

Deluca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.D., Williams, C., Reich, M., Winckler, W. and Getz, G. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, 2012. doi: 10.1093/bioinformatics/bts196.

Deng, H.X., Chen, W., Hong, S.T., Boycott, K.M., Gorrie, G.H., Siddique, N., Yang, Y., Fecto, F., Shi, Y., Zhai, H. et al. Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. *Nature*, 477(7363):211–215, 2011. doi: 10.1038/nature10353.

Deng, J., Yang, M., Chen, Y., Chen, X., Liu, J., Sun, S., Cheng, H., Li, Y., Bigio, E.H., Mesulam, M. et al. FUS Interacts with HSP60 to Promote Mitochondrial Damage. *PLoS Genetics*, 11(9):1–30, 2015. doi: 10.1371/journal.pgen.1005357.

Devoy, A., Kalmar, B., Stewart, M., Park, H., Burke, B., Noy, S.J., Redhead, Y., Humphrey, J., Lo, K., Jaeger, J. et al. Humanized mutant FUS drives progressive motor neuron degeneration without aggregation in 'FUSDelta14' knockin mice. *Brain*, 140(11):2797–2805, 2017. doi: 10.1093/brain/awx248.

Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E. and Notredame, C. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4): 316–319, 2017. doi: 10.1038/nbt.3820.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

Doi, H., Okamura, K., Bauer, P.O., Furukawa, Y., Shimizu, H., Kurosawa, M., Machida, Y., Miyazaki, H., Mitsui, K., Kuroiwa, Y. et al. RNA-binding protein TLS is a major nuclear aggregate-interacting protein in Huntingtin exon 1 with expanded polyglutamine-expressing cells. *Journal of Biological Chemistry*, 283(10):6489–6500, 2008. doi: 10.1074/jbc.M705306200.

Dormann, D., Rodde, R., Edbauer, D., Bentmann, E., Fischer, I., Hruscha, A., Than, M.E., MacKenzie, I.R., Capell, A., Schmid, B. et al. ALS-associated fused in sarcoma (FUS) mutations disrupt transportin-mediated nuclear import. *The EMBO Journal*, 29(16):2841–2857, 2010. doi: 10.1038/emboj.2010.143.

Dredge, B.K., Stefani, G., Engelhard, C.C. and Darnell, R.B. Nova autoregulation reveals dual functions in neuronal splicing. *The EMBO Journal*, 24(8):1608–1620, 2005. doi: 10.1038/sj.emboj.7600630.

du Plessis, L., Škunca, N. and Dessimoz, C. The what, where, how and why of gene ontology-a primer for bioinformaticians. *Briefings in Bioinformatics*, 12(6):723–735, 2011. doi: 10.1093/bib/bbr002.

Durrenberger, P.F., Fernando, S., Kashefi, S.N., Ferrer, I., Hauw, J.J., Seilhean, D., Smith, C., Walker, R., Al-Sarraj, S., Troakes, C. et al. Effects of antemortem and post-mortem variables on human brain mRNA quality: A brainNet Europe study. *Journal of Neuropathology and Experimental Neurology*, 69(1):70–81, 2010. doi: 10.1097/NEN.0b013e3181c7e32f.

Edwards, C.R., Ritchie, W., Wong, J.J.L., Schmitz, U., Middleton, R., An, X., Mohandas, N., Rasko, J.E.J. and Blobel, G.A. A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. *Blood*, 127(17):e24–e34, 2016. doi: 10.1182/blood-2016-01-692764.

Elden, A.C., Kim, H.J., Hart, M.P., Chen-Plotkin, A.S., Johnson, B.S., Fang, X., Armakola, M., Geser, F., Greene, R., Lu, M.M. et al. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature*, 466(7310):1069–1075, 2010. doi: 10.1038/nature09320.

Eng, L., Coutinho, G., Nahas, S., Yeo, G., Tanouye, R., Babaei, M., Dörk, T., Burge, C. and Gatti, R.A. Nonclassical splicing mutations in the coding and noncoding regions of the ATM gene: maximum entropy estimates of splice junction strengths. *Human Mutation*, 23(1):67–76, 2004.

Errichelli, L., Dini Modigliani, S., Laneve, P., Colantoni, A., Legnini, I., Capauto, D., Rosa, A., De Santis, R., Scarfò, R., Peruzzi, G. et al. FUS affects circular RNA expression in murine embryonic stem cell-derived motor neurons. *Nature Communications*, 8:14741, 2017. doi: 10.1038/ncomms14741.

Fallini, C., Bassell, G.J. and Rossoll, W. The ALS disease protein TDP-43 is actively transported in motor neuron axons and regulates axon outgrowth. *Human Molecular Genetics*, 21(16):3703–3718, 2012. doi: 10.1093/hmg/dds205.

Fang, Y.S., Tsai, K.J., Chang, Y.J., Kao, P., Woods, R., Kuo, P.H., Wu, C.C., Liao, J.Y., Chou, S.C., Lin, V. et al. Full-length TDP-43 forms toxic amyloid oligomers that are present in frontotemporal lobar dementia-TDP patients. *Nature Communications*, 5:1–13, 2014. doi: 10.1038/ncomms5824.

Fecto, F., Yan, J., Vemula, S.P., Liu, E., Yang, Y., Chen, W., Zheng, J.G., Shi, Y., Siddique, N., Arrat, H. et al. SQSTM1 mutations in familial and sporadic amyotrophic lateral sclerosis. *Archives of Neurology*, 68(11):1440–1446, 2011. doi: 10.1001/archneurol.2011.250.

Fratta, P., Sivakumar, P., Humphrey, J., Lo, K., Ricketts, T., Oliveira, H., Brito-Armas, J.M., Kalmar, B., Ule, A., Yu, Y. et al. Mice with endogenous TDP-43 mutations exhibit gain of splicing function and characteristics of amyotrophic lateral sclerosis. *The EMBO Journal*, page e98684, 2018. doi: 10.15252/embj.201798684.

Freibaum, B.D., Chitta, R.K., High, A.A. and Taylor, J.P. Global analysis of TDP-43 interacting proteins reveals strong association with RNA splicing and translation machinery. *Journal of Proteome Research*, 9(2):1104–1120, 2010.

Freischmidt, A., Wieland, T., Richter, B., Ruf, W., Schaeffer, V., Müller, K., Marroquin, N., Nordin, F., Hübers, A., Weydt, P. et al. Haploinsufficiency of TBK1 causes familial ALS and fronto-temporal dementia. *Nature Neuroscience*, 18(5):631–636, 2015. doi: 10.1038/nn.4000.

Fujii, R. and Takumi, T. TLS facilitates transport of mRNA encoding an actin-stabilizing protein to dendritic spines. *Journal of Cell Science*, 118(24):5755–5765, 2005. doi: 10.1242/jcs.02692.

Gentleman, R. and Ihaka, R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

Gilks, N., Kedersha, N., Ayodele, M., Shen, L., Stoecklin, G., Dember, L.M. and Anderson, P. Stress granule assembly is mediated by prion-like aggregation of tia-1. *Molecular biology of the cell*, 15(12):5383–5398, 2004.

Goldstein, L.D., Cao, Y., Pau, G., Lawrence, M., Wu, T.D., Seshagiri, S. and Gentleman, R. Prediction and quantification of splice events from RNA-seq data. *PLoS ONE*, 11(5):1–18, 2016. doi: 10.1371/journal.pone.0156132.

Gondo, Y., Fukumura, R., Murata, T. and Makino, S. ENU-Based Gene-Driven Mutagenesis in the Mouse: A Next-Generation Gene-Targeting System. *Experimental Animals*, 59(5):537–548, 2010. doi: 10.1538/expanim.59.537.

Gordon, D., Dafinca, R., Scaber, J., Alegre-Abarrategui, J., Farrimond, L., Scott, C., Biggs, D., Kent, L., Oliver, P.L., Davies, B. et al. Single-copy expression of an amyotrophic lateral sclerosis-linked TDP-43 mutation (M337V) in BAC transgenic mice leads to altered

stress granule dynamics and progressive motor dysfunction. *Neurobiology of Disease*, 121 (October 2018):148–162, 2018. doi: 10.1016/j.nbd.2018.09.024.

Groen, E.J.N., Fumoto, K., Blokhuis, A.M., Engelen-Lee, J.Y., Zhou, Y., van den Heuvel, D.M.A., Koppers, M., van Diggelen, F., van Heest, J., Demmers, J.A.A. et al. ALS-associated mutations in FUS disrupt the axonal distribution and function of SMN. *Human Molecular Genetics*, 22(18):3690–3704, 2013. doi: 10.1093/hmg/ddt222.

Gruber, A.J., Schmidt, R., Gruber, A.R., Martin, G., Ghosh, S., Belmadani, M., Keller, W. and Zavolan, M. 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):1145–1159, 2016. doi: 10.1101/gr.202432.115.

Gueroussov, S., Weatheritt, R.J., O'Hanlon, D., Lin, Z.Y., Narula, A., Gingras, A.C. and Blencowe, B.J. Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing. *Cell*, 170(2):324–339.e23, 2017. doi: 10.1016/j.cell.2017.06.037.

Guo, L., Kim, H.J., Wang, H., Monaghan, J., Freyermuth, F., Sung, J.C., O'Donovan, K., Fare, C.M., Diaz, Z., Singh, N. et al. Nuclear-Import Receptors Reverse Aberrant Phase Transitions of RNA-Binding Proteins with Prion-like Domains. *Cell*, 173(3):677–692.e20, 2018. doi: 10.1016/j.cell.2018.03.002.

Guo, W., Naujock, M., Fumagalli, L., Vandoorne, T., Baatsen, P., Boon, R., Ordovás, L., Patel, A., Welters, M., Vanwelden, T. et al. HDAC6 inhibition reverses axonal transport defects in motor neurons derived from FUS-ALS patients. *Nature Communications*, 8(1):1–14, 2017. doi: 10.1038/s41467-017-00911-y.

Harrow, J., Frankish, A., Gonzalez, J.M. and Frazer, K.A. GENCODE : The reference human genome annotation for The ENCODE Project. *Genome Research*, 22:1760–1774, 2012. doi: 10.1101/gr.135350.111.

Hartley, S.W. and Mullikin, J.C. Detection and Visualization of Differential Exon and Splice Junction Usage in RNA-Seq Data with JunctionSeq. *Nucleic Acids Research*, 44:e127, 2015a.

Hartley, S.W. and Mullikin, J.C. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. *BMC bioinformatics*, 16(1):224, 2015b. doi: 10.1186/s12859-015-0670-5.

Hasegawa, M., Arai, T., Nonaka, T., Kametani, F., Yoshida, M., Hashizume, Y., Beach, T.G., Buratti, E., Baralle, F., Morita, M. et al. Phosphorylated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Annals of Neurology*, 64(1):60–70, 2008. doi: 10.1002/ana.21425.

Haynes, W.A., Tomczak, A. and Khatri, P. Gene annotation bias impedes biomedical research. *Scientific Reports*, 8(1):1–7, 2018. doi: 10.1038/s41598-018-19333-x.

Heinz, S., Sven, H., Christopher, B., Nathanael, S., Eric, B., Lin, Y.C., Peter, L., Cheng,

J.X., Cornelis, M., Harinder, S. et al. Simple combinations of Lineage-Determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4):576–589, 2010.

Hicks, G.G., Singh, N., Nashabi, a., Mai, S., Bozek, G., Klewes, L., Arapovic, D., White, E.K., Koury, M.J., Oltz, E.M. et al. Fus deficiency in mice results in defective B-lymphocyte development and activation, high levels of chromosomal instability and perinatal death. *Nature Genetics*, 24(2):175–9, 2000. doi: 10.1038/72842.

Hofweber, M., Hutten, S., Bourgeois, B., Spreitzer, E., Niedner-Boblenz, A., Schifferer, M., Ruepp, M.D., Simons, M., Niessing, D., Madl, T. et al. Phase Separation of FUS Is Suppressed by Its Nuclear Import Receptor and Arginine Methylation. *Cell*, 173(3): 706–719.e13, 2018. doi: 10.1016/j.cell.2018.03.004.

Honda, D., Ishigaki, S., Iguchi, Y., Fujioka, Y., Udagawa, T., Masuda, A., Ohno, K., Katsuno, M. and Sobue, G. The ALS/FTLD-related RNA-binding proteins TDP-43 and FUS have common downstream RNA targets in cortical neurons. *FEBS Open Bio*, 4: 1–10, 2014. doi: 10.1016/j.fob.2013.11.001.

Huelga, S., Vu, A., Arnold, J., Liang, T., Liu, P., Yan, B., Donohue, J., Shiue, L., Hoon, S., Brenner, S. et al. Integrative Genome-wide Analysis Reveals Cooperative Regulation of Alternative Splicing by hnRNP Proteins. *Cell Reports*, 1(2):167–178, 2012. doi: 10.1016/j.celrep.2012.02.001.

Humphrey, J., Emmett, W., Fratta, P., Isaacs, A.M. and Plagnol, V. Quantitative analysis of cryptic splicing associated with TDP-43 depletion. *BMC Medical Genomics*, 10(1): 1–17, 2017. doi: 10.1186/s12920-017-0274-1.

Huppertz, I., Ina, H., Jan, A., Andrea, D., Easton, L.E., Sibley, C.R., Yoichiro, S., Mojca, T., Julian, K. and Jernej, U. iCLIP: Protein–RNA interactions at nucleotide resolution. *Methods*, 65(3):274–287, 2014.

Hutton, M., Lendon, C.L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., Pickering-Brown, S., Chakraverty, S., Isaacs, A., Grover, A. et al. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*, 393(6686):702–5, 1998. doi: 10.1038/31508.

Ibrahim, F., Maragkakis, M., Alexiou, P., Maronski, M.A., Dichter, M.A. and Mourelatos, Z. Identification of In Vivo, Conserved, TAF15 RNA Binding Sites Reveals the Impact of TAF15 on the Neuronal Transcriptome. *Cell Reports*, 3:301–308, 2013. doi: 10.1016/j.celrep.2013.01.021.

Iguchi, Y., Katsuno, M., Niwa, J.I., Takagi, S., Ishigaki, S., Ikenaka, K., Kawai, K., Watanabe, H., Yamanaka, K., Takahashi, R. et al. Loss of TDP-43 causes age-dependent progressive motor neuron degeneration. *Brain*, 136(5):1371–1382, 2013. doi: 10.1093/brain/awt029.

Ip, J.Y., Schmidt, D., Pan, Q., Ramani, A.K., Fraser, A.G., Odom, D.T. and Blencowe,

B.J. Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Research*, 21(3):390–401, 2011. doi: 10.1101/gr.111070.110.

Ishigaki, S., Masuda, A., Fujioka, Y., Iguchi, Y., Katsuno, M., Shibata, A., Urano, F., Sobue, G. and Ohno, K. Position-dependent FUS-RNA interactions regulate alternative splicing events and transcriptions. *Scientific Reports*, 2:1–8, 2012. doi: 10.1038/srep00529.

Jangi, M. and Sharp, P.A. Building robust transcriptomes with master splicing factors. *Cell*, 159(3):487–498, 2014. doi: 10.1016/j.cell.2014.09.054.

Jensen, K.B., Musunuru, K., Lewis, H.A., Burley, S.K. and Darnell, R.B. The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. *Proceedings of the National Academy of Sciences*, 97(11):5740–5745, 2000. doi: 10.1073/pnas.090553997.

Jeong, Y.H., Ling, J.P., Lin, S.Z., Donde, A.N., Braunstein, K.E., Majounie, E., Traynor, B.J., LaClair, K.D., Lloyd, T.E. and Wong, P.C. Tdp-43 cryptic exons are highly variable between cell types. *Molecular Neurodegeneration*, 12(1):13, 2017. doi: 10.1186/s13024-016-0144-x.

Johnson, B.S., Snead, D., Lee, J.J., McCaffery, J.M., Shorter, J. and Gitler, A.D. TDP-43 is intrinsically aggregation-prone, and amyotrophic lateral sclerosis-linked mutations accelerate aggregation and increase toxicity. *Journal of Biological Chemistry*, 284(30):20329–20339, 2009. doi: 10.1074/jbc.M109.010264.

Johnson, J.O., Pioro, E.P., Boehringer, A., Chia, R., Feit, H., Renton, A.E., Pliner, H.A., Abramzon, Y., Marangi, G., Winborn, B.J. et al. Mutations in the Matrin 3 gene cause familial amyotrophic lateral sclerosis. *Nature Neuroscience*, 17(5):664–666, 2014. doi: 10.1038/nn.3688.

Kabashi, E., Valdmanis, P.N., Dion, P., Spiegelman, D., McConkey, B.J., Velde, C.V., Bouchard, J.P., Lacomblez, L., Pochigaeva, K., Salachas, F. et al. TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nature Genetics*, 40(5):572–574, 2008. doi: 10.1038/ng.132.

Kamelgarn, M., Chen, J., Kuang, L., Arenas, A., Zhai, J., Zhu, H. and Gal, J. Proteomic analysis of FUS interacting proteins provides insights into FUS function and its role in ALS. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1862(10):2004–2014, 2016. doi: 10.1016/j.bbadis.2016.07.015.

Kanai, Y., Dohmae, N. and Hirokawa, N. Kinesin transports RNA: Isolation and characterization of an RNA-transporting granule. *Neuron*, 43(4):513–525, 2004. doi: 10.1016/j.neuron.2004.07.022.

Kapeli, K., Pratt, G.A., Vu, A.Q., Hutt, K.R., Martinez, F.J., Sundararaman, B., Batra, R., Freese, P., Lambert, N.J., Huelga, S.C. et al. Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. *Nature Communications*, 7:12143, 2016. doi: 10.1038/ncomms12143.

Kapeli, K., Martinez, F.J. and Yeo, G.W. Genetic mutations in RNA-binding proteins

and their roles in ALS. *Human Genetics*, 136(9):1193–1214, 2017. doi: 10.1007/s00439-017-1830-7.

Katz, Y., Wang, E.T., Airoldi, E.M. and Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, 2010.

Kelley, D.R., Hendrickson, D.G., Tenen, D. and Rinn, J.L. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome biology*, 15(12):537, 2014. doi: 10.1186/s13059-014-0537-5.

Kim, H.J., Kim, N.C., Wang, Y.D., Scarborough, E.A., Moore, J., Diaz, Z., MacLea, K.S., Freibaum, B., Li, S., Molliex, A. et al. Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature*, 495(7442):467–473, 2013. doi: 10.1038/nature11922.

Kino, Y., Washizu, C., Kurosawa, M., Yamada, M., Miyazaki, H., Akagi, T., Hashikawa, T., Doi, H., Takumi, T., Hicks, G.G. et al. FUS/TLS deficiency causes behavioral and pathological abnormalities distinct from amyotrophic lateral sclerosis. *Acta neuropathologica communications*, 3:24, 2015. doi: 10.1186/s40478-015-0202-6.

Kino, Y., Washizu, C., Kurosawa, M., Yamada, M., Doi, H., Takumi, T., Adachi, H., Katsuno, M., Sobue, G., Hicks, G.G. et al. FUS/TLS acts as an aggregation-dependent modifier of polyglutamine disease model mice. *Scientific Reports*, 6(October):1–14, 2016. doi: 10.1038/srep35236.

Kolde, R. Pheatmap: pretty heatmaps. *R package version 61*, 2012.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M. and Ule, J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17(7):909–915, 2010. doi: 10.1038/nsmb.1838.

Koressaar, T. and Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics*, 23(10):1289–1291, 2007. doi: 10.1093/bioinformatics/btm091.

Kornblihtt, A.R., De La Mata, M., Fededa, J.P., Muñoz, M.J. and Nogués, G. Multiple links between transcription and splicing. *RNA*, 10(10):1489–1498, 2004. doi: 10.1261/rna.7100104.

Köster, J. and Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012. doi: 10.1093/bioinformatics/bts480.

Kovar, H. Dr. Jekyll and Mr. Hyde: The two faces of the FUS/EWS/TAF15 protein family. *Sarcoma*, 2011(Table 1), 2011. doi: 10.1155/2011/837474.

Koyama, A., Sugai, A., Kato, T., Ishihara, T., Shiga, A., Toyoshima, Y., Koyama, M., Konno, T., Hirokawa, S., Yokoseki, A. et al. Increased cytoplasmic *TARDBP* mRNA in affected spinal motor neurons in ALS caused by abnormal autoregulation of TDP-43. *Nucleic Acids Research*, page gkw499, 2016. doi: 10.1093/nar/gkw499.

Kraemer, B.C., Schuck, T., Wheeler, J.M., Robinson, L.C., Trojanowski, J.Q., Lee, V.M.Y.

and Schellenberg, G.D. Loss of Murine TDP-43 disrupts motor function and plays an essential role in embryogenesis. *Acta Neuropathologica*, 119(4):409–419, 2010. doi: 10.1007/s00401-010-0659-0.

Kuroda, M., Sok, J., Webb, L., Baechtold2, H., Urano, F., Yin, Y., Chung, P., Rooij, D.G., Akhmedov, A., Ashley, T. et al. Male sterility and enhanced radiation sensitivity in TLS-/- mice. *The EMBO Journal*, 19(3):453–462, 2000. doi: 10.1093/emboj/19.3.453.

Kwiatkowski, T.J., Bosco, D.A., Leclerc, A.L., Tamrazian, E., Vanderburg, C.R., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E.J., Munsat, T. et al. Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science*, 323(5918):1205–1208, 2009. doi: 10.1126/science.1166066.

LaClair, K.D., Donde, A., Ling, J.P., Jeong, Y.H., Chhabra, R., Martin, L.J. and Wong, P.C. Depletion of TDP-43 decreases fibril and plaque β-amyloid and exacerbates neurodegeneration in an Alzheimer's mouse model. *Acta Neuropathologica*, 132(6):859–873, 2016. doi: 10.1007/s00401-016-1637-y.

Lagier-Tourenne, C., Polymenidou, M., Hutt, K.R., Vu, A.Q., Baughn, M., Huelga, S.C., Clutario, K.M., Ling, S.C., Liang, T.Y., Mazur, C. et al. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nature Neuroscience*, 15(11):1488–1497, 2012. doi: 10.1038/nn.3230.

Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C. and Brenner, S.E. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, 446(7138):926–929, 2007. doi: 10.1038/nature05676.

Lavorgna, G., Dahary, D., Lehner, B., Sorek, R., Sanderson, C.M. and Casari, G. In search of antisense. *Trends in Biochemical Sciences*, 29(2):88–94, 2004. doi: 10.1016/j.tibs.2003.12.002.

Lee, R.C., Feinbaum, R.L. and Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75: 843-85:843–854, 1993. doi: 10.1016/0092-8674(93)90529-Y.

Li, B. and Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 2011. doi: 10.1186/1471-2105-12-323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

Li, W., Jin, Y., Prazak, L., Hammell, M. and Dubnau, J. Transposable Elements in TDP-43-Mediated Neurodegenerative Disorders. *PLoS ONE*, 7(9):1–10, 2012. doi: 10.1371/journal.pone.0044099.

Li, Y., Rao, X., Mattox, W.M., Amos, C.I. and Liu, B. RNA-Seq Analysis of Differential Splice Junction Usage and Intron Retentions by DEXSeq. *PLoS ONE*, 10(9)(e0136653), 2015. doi: http://doi.org/10.1371/journal.pone.0136653.

Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K. and Pritchard, J.K. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1):151–158, 2018. doi: 10.1038/s41588-017-0004-9.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X. et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, 2008. doi: 10.1038/nature07488.

Ling, J.P., Pletnikova, O., Troncoso, J.C. and Wong, P.C. TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science*, 349(6248):650–655, 2015.

Ling, J.P., Chhabra, R., Merran, J.D., Schaughency, P.M., Wheelan, S.J., Corden, J.L. and Wong, P.C. PTBP1 and PTBP2 repress nonconserved cryptic exons. *Cell Reports*, 17(1): 104–113, 2016. doi: 10.1016/j.celrep.2016.08.071.

Ling, S.C., Albuquerque, C.P., Han, J.S., Lagier-Tourenne, C., Tokunaga, S., Zhou, H. and Cleveland, D.W. ALS-associated mutations in TDP-43 increase its stability and promote TDP-43 complexes with FUS/TLS. *Proceedings of the National Academy of Sciences*, 107 (30):13318–13323, 2010.

Liu-Yesucevitz, L., Bilgutay, A., Zhang, Y.J., Vanderwyde, T., Citro, A., Mehta, T., Zaarur, N., McKee, A., Bowser, R., Sherman, M. et al. Tar DNA binding protein-43 (TDP-43) associates with stress granules: Analysis of cultured cells and pathological brain tissue. *PLoS ONE*, 5(10), 2010. doi: 10.1371/journal.pone.0013250.

Logroscino, G., Traynor, B.J., Hardiman, O., Chiò, A., Mitchell, D., Swingler, R.J., Millul, A., Benn, E. and Beghi, E. Incidence of amyotrophic lateral sclerosis in Europe. *Journal of neurology, neurosurgery, and psychiatry*, 81(4):385–90, 2010. doi: 10.1136/jnnp.2009.183525.

López-Erauskin, J., Tadokoro, T., Baughn, M.W., Myers, B., McAlonis-Downes, M., Chillon-Marinas, C., Asiaban, J.N., Artates, J., Bui, A.T., Vetto, A.P. et al. ALS/FTD-Linked mutation in FUS suppresses intra-axonal protein synthesis and drives disease without nuclear loss-of-function of FUS. *Neuron*, 0(0):1–15, 2018. doi: 10.1016/j.neuron.2018.09.044.

Losson, R. and Lacroute, F. Interference of nonsense mutations with eukaryotic messenger. *Proc. Natl. Acad. Sci. USA*, 76(10):5134–5137, 1979. doi: 10.1073/PNAS.76.10.5134.

Love, M.I., Huber, W. and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. doi: 10.1186/s13059-014-0550-8.

Luisier, R., Tyzack, G.E., Hall, C.E., Mitchell, J.S., Devine, H., Taha, D.M., Malik, B., Meyer, I., Greensmith, L., Newcombe, J. et al. Intron retention and nuclear loss of SFPQ are molecular hallmarks of ALS. *Nature Communications*, 9(1):2010, 2018. doi: 10.1038/s41467-018-04373-8.

Lukavsky, P.J., Daujotyte, D., Tollervey, J.R., Ule, J., Stuani, C., Buratti, E., Baralle, F.E., Damberger, F.F. and Allain, F.H.T. Molecular basis of UG-rich RNA recognition by the

human splicing factor TDP-43. *Nature Structural & Molecular Biology*, 20(12):1443–1449, 2013. doi: 10.1038/nsmb.2698.

Mackenzie, I.R., Nicholson, A.M., Sarkar, M., Messing, J., Purice, M.D., Pottier, C., Annu, K., Baker, M., Perkerson, R.B., Kurti, A. et al. TIA1 Mutations in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Promote Phase Separation and Alter Stress Granule Dynamics. *Neuron*, 95(4):808–816.e9, 2017. doi: 10.1016/j.neuron.2017.07.025.

Majounie, E., Renton, A.E., Mok, K., Dopper, E.G.P., Waite, A., Rollinson, S., Chiò, A., Restagno, G., Nicolaou, N., Simon-Sanchez, J. et al. Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: A cross-sectional study. *The Lancet Neurology*, 11(4):323–330, 2012. doi: 10.1016/S1474-4422(12)70043-1.

Majumder, P., Chen, Y.T., Bose, J.K., Wu, C.C., Cheng, W.C., Cheng, S.J., Fang, Y.H., Chen, Y.L., Tsai, K.J., Lien, C.C. et al. TDP-43 regulates the mammalian spinogenesis through translational repression of Rac1. *Acta Neuropathologica*, 124(2):231–245, 2012. doi: 10.1007/s00401-012-1006-4.

Majumder, P., Chu, J.F., Chatterjee, B., Swamy, K.B. and Shen, C.K.J. Co-regulation of mRNA translation by TDP-43 and Fragile X Syndrome protein FMRP. *Acta Neuropathologica*, 132(5):721–738, 2016. doi: 10.1007/s00401-016-1603-8.

Maruyama, H., Morino, H., Ito, H., Izumi, Y., Kato, H., Watanabe, Y., Kinoshita, Y., Kamada, M., Nodera, H., Suzuki, H. et al. Mutations of optineurin in amyotrophic lateral sclerosis. *Nature*, 465(7295):223–226, 2010. doi: 10.1038/nature08971.

Masuda, A., Takeda, J.i., Okuno, T., Okamoto, T., Ohkawara, B., Ito, M., Ishigaki, S., Sobue, G. and Ohno, K. Position-specific binding of FUS to nascent RNA regulates mRNA length. *Genes and Development*, pages 1045–1057, 2015. doi: 10.1101/gad.255737.114.itation.

Matera, A.G. and Wang, Z. A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology*, 15(2):108–21, 2014. doi: 10.1038/nrm3742.

McClory, S.P., Lynch, K.W. and Ling, J.P. HnRNP L represses cryptic exons. *RNA*, 24(6):761–768, 2018. doi: 10.1261/rna.065508.117.

McGlincy, N.J. and Smith, C.W.J. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem. Sci.*, 33(8):385–393, 2008.

Meili, D., Kralovicova, J., Zagalak, J., Bonafe, L., Fiori, L., Blau, N., Thony, B. and Vorechovsky, I. Disease-causing mutations improving the branch site and polypyrimidine tract: Pseudoexon activation of LINE-2 and antisense alu lacking the poly(t)-tail. *Human Mutation*, 30(5):823–831, 2009.

Meissner, M., Lopato, S., Gotzmann, J., Sauermann, G. and Barta, A. Proto-oncoprotein TLS/FUS is associated to the nuclear matrix and complexed with splicing factors PTB,

SRm160, and SR proteins. *Experimental Cell Research*, 283(2):184–195, 2003. doi: 10.1016/S0014-4827(02)00046-0.

Mercado, P.A., Ayala, Y.M., Romano, M., Buratti, E. and Baralle, F.E. Depletion of TDP 43 overrides the need for exonic and intronic splicing enhancers in the human apoA-II gene. *Nucleic Acids Research*, 33(18):6000–6010, 2005.

Mirra, A., Rossi, S., Scaricamazza, S., Di Salvio, M., Salvatori, I., Valle, C., Rusmini, P., Poletti, A., Cestra, G., Carrì, M.T. et al. Functional interaction between FUS and SMN underlies SMA-like splicing changes in wild-type hFUS mice. *Scientific Reports*, 7(1): 2033, 2017. doi: 10.1038/s41598-017-02195-0.

Mitchell, J.C., McGoldrick, P., Vance, C., Hortobagyi, T., Sreedharan, J., Rogelj, B., Tudor, E.L., Smith, B.N., Klasen, C., Miller, C.C.J. et al. Overexpression of human wild-type FUS causes progressive motor neuron degeneration in an age- and dose-dependent fashion. *Acta Neuropathologica*, 125(2):273–288, 2013. doi: 10.1007/s00401-012-1043-z.

Modigliani, S.D., Morlando, M., Errichelli, L., Sabatelli, M. and Bozzoni, I. An ALS-associated mutation in the FUS 3'-UTR disrupts a microRNAâĂŞFUS regulatory circuitry. *Nature Communications*, 5(4335), 2014. doi: 10.1038/ncomms5335.

Mohagheghi, F., Prudencio, M., Stuani, C., Cook, C., Jansen-West, K., Dickson, D.W., Petrucelli, L. and Buratti, E. TDP-43 functions within a network of hnRNP proteins to inhibit the production of a truncated human SORT1 receptor. *Human molecular genetics*, 25(3):534–545, 2016. doi: 10.1093/hmg/ddv491.

Moisse, K., Volkening, K., Leystra-Lantz, C., Welch, I., Hill, T. and Strong, M.J. Divergent patterns of cytosolic TDP-43 and neuronal progranulin expression following axotomy: Implications for TDP-43 in the physiological response to neuronal injury. *Brain Research*, 1249:202–211, 2009. doi: 10.1016/j.brainres.2008.10.021.

Morlando, M., Dini Modigliani, S., Torrelli, G., Rosa, A., Di Carlo, V., Caffarelli, E. and Bozzoni, I. FUS stimulates microRNA biogenesis by facilitating co-transcriptional Drosha recruitment. *The EMBO Journal*, 31(24):4502–4510, 2012. doi: 10.1038/emboj.2012.319.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008. doi: 10.1038/nmeth.1226.

Moujalled, D., James, J.L., Yang, S., Zhang, K., Duncan, C., Moujalled, D.M., Parker, S.J., Caragounis, A., Lidgerwood, G., Turner, B.J. et al. Phosphorylation of hnRNP K by cyclin-dependent kinase 2 controls cytosolic accumulation of TDP-43. *Human Molecular Genetics*, 24(6):1655–1669, 2015. doi: 10.1093/hmg/ddu578.

Murray, D.T., Kato, M., Lin, Y., Thurber, K.R., Hung, I., McKnight, S.L. and Tycko, R. Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains. *Cell*, 171(3):615–627.e16, 2017. doi: 10.1016/j.cell.2017.08.048.

Nagy, E. and Maquat, L.E. A rule for termination-codon position within intron-containing

genes: When nonsense affects RNA abundance. *Trends in Biochemical Sciences*, 23(6): 198–199, 1998. doi: 10.1016/S0968-0004(98)01208-0.

Nakaya, T., Alexiou, P., Maragkakis, M. and Chang, A. FUS regulates genes coding for RNA-binding proteins in neurons by binding to their highly conserved introns. *RNA*, pages 498–509, 2013. doi: 10.1261/rna.037804.112.4.

Neelagandan, N., Gonnella, G., Dang, S., Janiesch, P.C., Miller, K., Katrin, K., Marques, R.F., Indenbirken, D., Alawi, M., Grundhoff, A. et al. TDP-43 enhances translation of specific mRNAs linked to neurodegenerative disease. *Nucleic Acids Research*, 0(0):1–21, 2018. doi: 10.1093/nar/gky972.

Neumann, M., Sampathu, D.M., Kwong, L.K., Truax, A.C., Micsenyi, M.C., Chou, T.T., Bruce, J., Schuck, T., Grossman, M., Clark, C.M. et al. Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. *Science*, 314(5796): 130–133, 2006. doi: 10.1126/science.1134108.

Neumann, M., Rademakers, R., Roeber, S., Baker, M., Kretzschmar, H.A. and MacKenzie, I.R.A. A new subtype of frontotemporal lobar degeneration with FUS pathology. *Brain*, 132(11):2922–2931, 2009. doi: 10.1093/brain/awp214.

Neumann, M., Bentmann, E., Dormann, D., Jawaid, A., Dejesus-Hernandez, M., Ansorge, O., Roeber, S., Kretzschmar, H.A., Munoz, D.G., Kusaka, H. et al. FET proteins TAF15 and EWS are selective markers that distinguish FTLD with FUS pathology from amyotrophic lateral sclerosis with FUS mutations. *Brain*, 134(9):2595–2609, 2011. doi: 10.1093/brain/awr201.

Neumann, M., Valori, C.F., Ansorge, O., Kretzschmar, H.A., Munoz, D.G., Kusaka, H., Yokota, O., Ishihara, K., Ang, L.C., Bilbao, J.M. et al. Transportin 1 accumulates specifically with FET proteins but no other transportin cargos in FTLD-FUS and is absent in FUS inclusions in ALS with FUS mutations. *Acta Neuropathologica*, 124(5): 705–716, 2012. doi: 10.1007/s00401-012-1020-6.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E. and Ares, M. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes and Development*, 21(6):708–718, 2007. doi: 10.1101/gad.1525507.

Nicolas, A., Kenna, K.P., Renton, A.E., Ticozzi, N., Faghri, F., Chia, R., Dominov, J.A., Kenna, B.J., Nalls, M.A., Keagle, P. et al. Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron*, 97(6):1268–1283.e6, 2018. doi: 10.1016/j.neuron.2018.02.027.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34, 1999. doi: 10.1093/nar/27.1.29.

Onyike, C.U. and Diehl-Schmid, J. The epidemiology of frontotemporal dementia. *International Review of Psychiatry*, 25(2):130–7, 2013. doi: 10.3109/09540261.2013.776523.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J. and Blencowe, B.J. Deep surveying of alternative

splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, 2008. doi: 10.1038/ng.259.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, 2017. doi: 10.1038/nmeth.4197.

Pavlidis, P., Jensen, J.D., Stephan, W. and Stamatakis, A. A critical assessment of storytelling: Gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, 29(10):3237–3248, 2012. doi: 10.1093/molbev/mss136.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3), 2015. doi: 10.1038/nbt.3122.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

Polymenidou, M., Lagier-tourenne, C., Hutt, K.R., Stephanie, C., Moran, J., Liang, T.Y., Ling, S.c., Sun, E., Wancewicz, E., Mazur, C. et al. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nature Neuroscience*, 14:459–468, 2011. doi: 10.1038/nn.2779.Long.

Pottier, C., Ravenscroft, T.A., Sanchez-Contreras, M. and Rademakers, R. Genetics of FTLD: overview and what else we can expect from genetic studies. *Journal of Neurochemistry*, 138:32–53, 2016. doi: 10.1111/jnc.13622.

Prudencio, M., Jansen-West, K.R., Lee, W.C., Gendron, T.F., Zhang, Y.J., Xu, Y.F., Gass, J., Stuani, C., Stetler, C., Rademakers, R. et al. Misregulation of human sortilin splicing leads to the generation of a nonfunctional progranulin receptor. *Proceedings of the National Academy of Sciences*, 109(52):21510–5, 2012. doi: 10.1073/pnas.1211577110.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M. et al. RefSeq: An update on mammalian reference sequences. *Nucleic Acids Research*, 42(1):756–763, 2014. doi: 10.1093/nar/gkt1114.

Puls, I., Jonnakuty, C., LaMonte, B.H., Holzbaur, E.L., Tokito, M., Mann, E., Floeter, M.K., Bidus, K., Drayna, D., Oh, S.J. et al. Mutant dynactin in motor neuron disease. *Nature Genetics*, 33(4):455–456, 2003. doi: 10.1038/ng1123.

Qiu, H., Lee, S., Shang, Y., Wang, W.Y., Au, K.F., Kamiya, S., Barmada, S.J., Finkbeiner, S., Lui, H., Carlton, C.E. et al. ALS-associated mutation FUS-R521C causes DNA damage and RNA splicing defects. *Journal of Clinical Investigation*, 124(3):981–999, 2014. doi: 10.1172/JCI72723.

Quinlan, A.R. and Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

Rademakers, R., Cruts, M. and Van Broeckhoven, C. The role of tau (MAPT) in fron-

totemporal dementia and related tauopathies. *Human Mutation*, 24(4):277–295, 2004. doi: 10.1002/humu.20086.

Raefski, A.S. and O'Neill, M.J. Identification of a cluster of X-linked imprinted genes in mice. *Nature Genetics*, 37(6):620–624, 2005. doi: 10.1038/ng1567.

Raj, T., Li, Y.I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.C., Ng, B., Gupta, I., Haroutunian, V. et al. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nature Genetics*, 2018. doi: 10.1038/s41588-018-0238-1.

Reber, S., Stettler, J., Filosa, G., Colombo, M., Jutzi, D., Lenzken, S.C., Schweingruber, C., Bruggmann, R., Bachi, A., Barabino, S.M. et al. Minor intron splicing is regulated by FUS and affected by ALS-associated FUS mutants. *The EMBO Journal*, 35(14):1504–1521, 2016. doi: 10.15252/embj.201593791.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H. and Vilo, J. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, 44(W1):W83–W89, 2016. doi: 10.1093/nar/gkw199.

Renton, A.E., Majounie, E., Waite, A., Sim??n-S??nchez, J., Rollinson, S., Gibbs, J.R., Schymick, J.C., Laaksovirta, H., van Swieten, J.C., Myllykangas, L. et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*, 72(2):257–268, 2011. doi: 10.1016/j.neuron.2011.09.010.

Ricketts, T., McGoldrick, P., Fratta, P., De Oliveira, H.M., Kent, R., Phatak, V., Brandner, S., Blanco, G., Greensmith, L., Acevedo-Arozena, A. et al. A nonsense mutation in mouse Tardbp affects TDP43 alternative splicing activity and causes limb-clasping and body tone defects. *PLoS ONE*, 9(1), 2014. doi: 10.1371/journal.pone.0085962.

Rogelj, B., Easton, L.E., Bogu, G.K., Stanton, L.W., Rot, G., Curk, T., Zupan, B., Sugimoto, Y., Modic, M., Haberman, N. et al. Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Scientific Reports*, 2:1–10, 2012. doi: 10.1038/srep00603.

Rosen, D.R., Siddique, T., Patterson, D., Figlewicz, D.A., Sapp, P., Hentati, A., Donaldson, D., Goto, J., O'Regan, J.P. and Deng, H.X. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature*, 362(6415):59–62, 1993. doi: 10.1038/362059a0.

Rosenfeld, N., Elowitz, M.B. and Alon, U. Negative autoregulation speeds the response times of transcription networks. *Journal of Molecular Biology*, 323(5):785–793, 2002. doi: 10.1016/S0022-2836(02)00994-4.

Rossbach, O., Hung, L.H., Schreiner, S., Grishina, I., Heiner, M., Hui, J. and Bindereif, A. Auto- and Cross-Regulation of the hnRNP L Proteins by Alternative Splicing. *Molecular and Cellular Biology*, 29(6):1442–1451, 2009. doi: 10.1128/MCB.01689-08.

Rouaux, C., Gonzalez De Aguilar, J. and Dupuis, L. Unmasking the skiptic task of TDP-43. *The EMBO Journal*, 37(11):e99645, 2018. doi: 10.15252/embj.201899645.

Ruggiu, M., Herbst, R., Kim, N., Jevsek, M., Fak, J.J., Mann, M.A., Fischbach, G., Burden, S.J. and Darnell, R.B. Rescuing Z agrin splicing in nova null mice restores synapse formation and unmasks a physiologic defect in motor neuron firing. *Proceedings of the National Academy of Sciences*, 106(9):3513–3518, 2009.

Scekic-Zahirovic, J., Sendscheid, O., Oussini, H.E., Jambeau, M. and Ying, S. Toxic gain of function from mutant FUS protein is crucial to trigger cell autonomous motor neuron loss. *The EMBO Journal*, pages 1–21, 2016.

Scekic-Zahirovic, J., Oussini, H.E., Mersmann, S., Drenner, K., Wagner, M., Sun, Y., Allmeroth, K., Dieterlé, S., Sinniger, J., Dirrig-Grosch, S. et al. Motor neuron intrinsic and extrinsic mechanisms contribute to the pathogenesis of FUS-associated amyotrophic lateral sclerosis. *Acta Neuropathologica*, 133(6):887–906, 2017. doi: 10.1007/s00401-017-1687-9.

Schwartz, J.C., Ebmeier, C.C., Podell, E.R., Heimiller, J., Taatjes, D.J. and Cech, T.R. FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. *Genes and Development*, 26(24):2690–2695, 2012. doi: 10.1101/gad.204602.112.

Scotti, M.M. and Swanson, M.S. RNA mis-splicing in disease. *Nature Reviews Genetics*, 17 (1):19–32, 2015. doi: 10.1038/nrg.2015.3.

Shan, X., Chiang, P.M., Price, D.L. and Wong, P.C. Altered distributions of Gemini of coiled bodies and mitochondria in motor neurons of TDP-43 transgenic mice. *Proceedings of the National Academy of Sciences*, 107(37):16325–16330, 2010. doi: 10.1073/pnas.1003459107.

Shang, Y. and Huang, E.J. Mechanisms of FUS mutations in familial amyotrophic lateral sclerosis. *Brain Research*, 1647:65–78, 2016. doi: 10.1016/j.brainres.2016.03.036.

Sharma, A., Lyashchenko, A.K., Lu, L., Nasrabady, S.E., Elmaleh, M., Mendelsohn, M., Nemes, A., Tapia, J.C., Mentis, G.Z. and Shneider, N.A. ALS-associated mutant FUS induces selective motor neuron degeneration through toxic gain of function. *Nature Communications*, 7:10465, 2016. doi: 10.1038/ncomms10465.

Shen, S., Park, J.W., Lu, Z.x., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q. and Xing, Y. 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, 2014. doi: 10.1073/pnas.1419161111.

Shiga, A., Ishihara, T., Miyashita, A., Kuwabara, M., Kato, T., Watanabe, N., Yamahira, A., Kondo, C., Yokoseki, A., Takahashi, M. et al. Alteration of POLDIP3 splicing associated with loss of function of TDP-43 in tissues affected with ALS. *PLoS One*, 7(8): e43120, 2012.

Shiihashi, G., Ito, D., Yagi, T., Nihei, Y., Ebine, T. and Suzuki, N. Mislocated FUS is sufficient for gain-of-toxic-function amyotrophic lateral sclerosis phenotypes in mice. *Brain*, 139(9):2380–2394, 2016. doi: 10.1093/brain/aww161.

Sibley, C.R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni,

D., Ryten, M., Weale, M.E., Hardy, J. et al. Recursive splicing in long vertebrate genes. *Nature*, 521(7552):371–5, 2015. doi: 10.1038/nature14466.

Sivakumar, P., De Giorgio, F., Ule, A.M., Neeves, J., Nair, R.R., Bentham, M., Birsa, N., Humphrey, J., Plagnol, V., Acevedo-Arozena, A. et al. TDP-43 mutations increase HNRNP A1-7B through gain of splicing function. *Brain*, pages 1–4, 2018. doi: 10.1093/brain/awy260.

Skibinski, G., Parkinson, N.J., Brown, J.M., Chakrabarti, L., Lloyd, S.L., Hummerich, H., Nielsen, J.E., Hodges, J.R., Spillantini, M.G., Thusgaard, T. et al. Mutations in the endosomal ESCRTIII-complex subunit CHMP2B in frontotemporal dementia. *Nature Genetics*, 37(8):806–808, 2005. doi: 10.1038/ng1609.

Smit, A., Hubley, R. and Green, P. RepeatMasker open-4.0. `http://www.repeatmasker.org`, 2015. Accessed: 2016-2-1.

Smith, B., Ticozzi, N., Fallini, C., Gkazi, A., Topp, S., Kenna, K., Scotter, E., Kost, J., Keagle, P., Miller, J. et al. Exome-wide Rare Variant Analysis Identifies TUBA4A Mutations Associated with Familial ALS. *Neuron*, 84(2):324–331, oct 2014. doi: 10.1016/j.neuron.2014.09.027.

So, E., Mitchell, J.C., Memmi, C., Chennell, G., Vizcay-Barrena, G., Allison, L., Shaw, C.E. and Vance, C. Mitochondrial abnormalities and disruption of the neuromuscular junction precede the clinical phenotype and motor neuron loss in hFUSWTtransgenic mice. *Human Molecular Genetics*, 27(3):463–474, 2018. doi: 10.1093/hmg/ddx415.

Sorek, R., Ast, G. and Graur, D. Alu-containing exons are alternatively spliced. *Genome Research*, 12(7):1060–1067, 2002.

Sreedharan, J., Blair, I.P., Tripathi, V.B., Hu, X., Vance, C., Rogelj, B., Ackerley, S., Durnall, J.C., Williams, K.L., Buratti, E. et al. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science*, 319(5870):1668–1672, 2008.

Štalekar, M., Yin, X., Rebolj, K., Darovic, S., Troakes, C., Mayr, M., Shaw, C.E. and Rogelj, B. Proteomic analyses reveal that loss of TDP-43 affects RNA processing and intracellular transport. *Neuroscience*, 293:157–170, 2015. doi: 10.1016/j.neuroscience.2015.02.046.

Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O. and Zhang, M.Q. An Alternative-Exon Database and Its Statistical Analysis. *DNA and Cell Biology*, 19(12):739–756, 2000. doi: 10.1089/104454900750058107.

Sterne-Weiler, T., Weatheritt, R.J., Best, A.J., Ha, K.C. and Blencowe, B.J. Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Molecular Cell*, pages 1–14, 2018. doi: 10.1016/J.MOLCEL.2018.08.018.

Sun, S., Ling, S.C., Qiu, J., Albuquerque, C.P., Zhou, Y., Tokunaga, S., Li, H., Qiu, H., Bui, A., Yeo, G.W. et al. ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. *Nature Communications*, 6:6171, 2015. doi: 10.1038/ncomms7171.

Sun, Z., Diaz, Z., Fang, X., Hart, M.P., Chesi, A., Shorter, J. and Gitler, A.D. Molecular determinants and genetic modifiers of aggregation and toxicity for the als disease protein fus/tls. *PLoS Biology*, 9(4), 2011. doi: 10.1371/journal.pbio.1000614.

Taliaferro, J.M., Vidaki, M., Oliveira, R., Olson, S., Zhan, L., Saxena, T., Wang, E.T., Graveley, B.R., Gertler, F.B., Swanson, M.S. et al. Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Molecular Cell*, 61(6):1–13, 2016. doi: 10.1016/j.molcel.2016.01.020.

Tan, Q., Yalamanchili, H.K., Park, J., De Maio, A., Lu, H.C., Wan, Y.W., White, J.J., Bondar, V.V., Sayegh, L.S., Liu, X. et al. Extensive cryptic splicing upon loss of RBM17 and TDP43 in neurodegeneration models. *Human Molecular Genetics*, 25(23):5083–5093, 2016. doi: 10.1093/hmg/ddw337.

Tarn, W.Y. and Steitz, J.A. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell*, 84(5):801–811, 1996. doi: 10.1016/S0092-8674(00)81057-0.

Taylor, J.P., Brown, R.H. and Cleveland, D.W. Decoding ALS: from genes to mechanism. *Nature*, 539(7628):197–206, 2016. doi: 10.1038/nature20413.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. and Narechania, A. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, 13(9):2129–2141, 2003. doi: 10.1101/gr.772403.

Thomas-Jinu, S., Gordon, P.M., Fielding, T., Taylor, R., Smith, B.N., Snowden, V., Blanc, E., Vance, C., Topp, S., Wong, C.H. et al. Non-nuclear Pool of Splicing Factor SFPQ Regulates Axonal Transcripts Required for Normal Motor Development. *Neuron*, 94(4):931, 2017. doi: 10.1016/j.neuron.2017.04.036.

Ticozzi, N., Vance, C., LeClerc, A.L., Keagle, P., Glass, J.D., McKenna-Yasek, D., Sapp, P.C., Silani, V., Bosco, D.A., Shaw, C.E. et al. Mutational analysis reveals the FUS homolog TAF15 as a candidate gene for familial amyotrophic lateral sclerosis. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 156(3):285–290, 2011. doi: 10.1002/ajmg.b.31158.

Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., Harel, I., Bustamante, C.D., Rasmussen, M. and Snyder, M.P. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nature Biotechnology*, 33(7):736–742, 2015. doi: 10.1038/nbt.3242.

Tollervey, J.R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., Konig, J., Hortobagyi, T., Nishimura, A.L., Zupunski, V. et al. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat Neurosci*, 14(4):452–458, apr 2011.

Tomczak, A., Mortensen, J.M., Winnenburg, R., Liu, C., Alessi, D.T., Swamy, V., Vallania, F., Lofgren, S., Haynes, W., Shah, N.H. et al. Interpretation of biological experiments

changes with evolution of the Gene Ontology and its annotations. *Scientific Reports*, 8 (1):1–10, 2018. doi: 10.1038/s41598-018-23395-2.

Trapnell, C., Williams, B.a., , G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010. doi: 10.1038/nbt.1621.

Treutlein, B., Gokce, O., Quake, S.R. and Südhof, T.C. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proceedings of the National Academy of Sciences*, 111(13):E1291–9, 2014. doi: 10.1073/pnas.1403244111.

Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J. and Eyras, E. SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19(1):1–11, 2018. doi: 10.1186/s13059-018-1417-1.

Udagawa, T., Fujioka, Y., Tanaka, M., Honda, D., Yokoi, S., Riku, Y., Ibi, D., Nagai, T., Yamada, K., Watanabe, H. et al. FUS regulates AMPA receptor function and FTLD/ALS-associated behaviour via GluA1 mRNA stabilization. *Nature Communications*, 6(0):7098, 2015. doi: 10.1038/ncomms8098.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A. and Darnell, R.B. CLIP Identifes Nova-Regulated RNA Networks in the Brain. *Science*, 302(0):1212–1215, 2003. doi: 10.1126/science.1090095.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J. and Darnell, R.B. An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119):580–586, 2006. doi: 10.1038/nature05304.

van Blitterswijk, M., Wang, E.T., Friedman, B.A., Keagle, P.J., Lowe, P., Leclerc, A.L., van den Berg, L.H., Housman, D.E., Veldink, J.H. and Landers, J.E. Characterization of FUS Mutations in Amyotrophic Lateral Sclerosis Using RNA-Seq. *PLoS ONE*, 8(4):1–8, 2013. doi: 10.1371/journal.pone.0060788.

Van Gurp, T.P., McIntyre, L.M. and Verhoeven, K.J.F. Consistent errors in first strand cDNA due to random hexamer mispriming. *PLoS ONE*, 8(12):2–5, 2013. doi: 10.1371/journal.pone.0085583.

Van Langenhove, T., Van Der Zee, J., Sleegers, K., Engelborghs, S., Vandenberghe, R., Gijselinck, I., Van Den Broeck, M., Mattheijssens, M., Peeters, K., De Deyn, P.P. et al. Genetic contribution of FUS to frontotemporal lobar degeneration. *Neurology*, 74(5): 366–371, 2010. doi: 10.1212/WNL.0b013e3181ccc732.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K. et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514, 2016.

Vance, C., Rogelj, B., Hortobágyi, T., De Vos, K.J., Nishimura, A.L., Sreedharan, J., Hu,

X., Smith, B., Ruddy, D., Wright, P. et al. Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science*, 323(5918):1208–1211, 2009.

Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., González-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W. and Barash, Y. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, 5:1–30, 2016. doi: 10.7554/eLife.11752.

Vaquero-Garcia, J., Norton, S. and Barash, Y. LeafCutter vs. MAJIQ and comparing software in the fast-moving field of genomics. *bioRxiv*, page 463927, 2018. doi: 10.1101/463927.

Verbeeck, C., Deng, Q., DeJesus-Hernandez, M., Taylor, G., Ceballos-Diaz, C., Kocerha, J., Golde, T., Das, P., Rademakers, R., Dickson, D.W. et al. Expression of Fused in sarcoma mutations in mice recapitulates the neuropathology of FUS proteinopathies and provides insight into disease pathogenesis. *Molecular Neurodegeneration*, 7(1):1, 2012. doi: 10.1186/1750-1326-7-53.

Vorechovsky, I. Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Research*, 34(16):4630–4641, 2006.

Vorechovsky, I. Transposable elements in disease-associated cryptic exons. *Human Genetics*, 127(2):135–154, 2010.

Štalekar, M., Yin, X., Rebolj, K., Darovic, S., Troakes, C., Mayr, M., Shaw, C.E. and Rogelj, B. Proteomic analyses reveal that loss of TDP-43 affects RNA processing and intracellular transport. *Neuroscience*, 293:157–170, 2015.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008. doi: 10.1038/nature07509.

Wang, Q. and Rio, D.C. JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. *Proceedings of the National Academy of Sciences*, page 201806018, 2018. doi: 10.1073/pnas.1806018115.

Wang, Y., Xiao, X., Zhang, J., Choudhury, R., Robertson, A., Li, K., Ma, M., Burge, C.B. and Wang, Z. A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nature Structural and Molecular Biology*, 20(1):36–45, 2013. doi: 10.1038/nsmb.2459.

Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N.M., Rot, G., Zupan, B., Curk, T. and Ule, J. iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biology*, 8(10), 2010. doi: 10.1371/journal.pbio.1000530.

Wang, Z., Gerstein, M. and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 2009.

Wegorzewska, I., Bell, S., Cairns, N.J., Miller, T.M. and Baloh, R.H. TDP-43 mutant

transgenic mice develop features of ALS and frontotemporal lobar degeneration. *Proceedings of the National Academy of Sciences*, 106(44):18809–14, 2009. doi: 10.1073/pnas.0908767106.

White, M.A., Kim, E., Duffy, A., Adalbert, R., Phillips, B.U., Peters, O.M., Stephenson, J., Yang, S., Massenzio, F., Lin, Z. et al. TDP-43 gains function due to perturbed autoregulation in a Tardbp knock-in mouse model of ALS-FTD. *Nature Neuroscience*, 2018. doi: 10.1038/s41593-018-0113-5.

Wilhelm, B.T. and Landry, J.R. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48(3):249–257, 2009. doi: 10.1016/j.ymeth.2009.03.016.

Wils, H., Kleinberger, G., Janssens, J., Pereson, S., Joris, G., Cuijt, I., Smits, V., Ceuterick-de Groote, C., Van Broeckhoven, C. and Kumar-Singh, S. TDP-43 transgenic mice develop spastic paralysis and neuronal inclusions characteristic of ALS and frontotemporal lobar degeneration. *Proceedings of the National Academy of Sciences*, 107(8):3858–3863, feb 2010. doi: 10.1073/pnas.0912417107.

Wollerton, M.C., Gooding, C., Wagner, E.J., Garcia-Blanco, M.A. and Smith, C.W. Autoregulation of Polypyrimidine Tract Binding Protein by Alternative Splicing Leading to Nonsense-Mediated Decay. *Molecular Cell*, 13(1):91–100, 2004. doi: 10.1016/S1097-2765(03)00502-1.

Wong, J.J.L., Ritchie, W., Ebner, O.A., Selbach, M., Wong, J.W.H., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K. et al. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell*, 154(3):583–595, 2013. doi: 10.1016/j.cell.2013.06.052.

Yamaguchi, A. and Takanashi, K. FUS interacts with nuclear matrix-associated protein SAFB1 as well as Matrin3 to regulate splicing and ligand-mediated transcription. *Scientific Reports*, 6(0):35195, 2016. doi: 10.1038/srep35195.

Yamazaki, T., Chen, S., Yu, Y., Yan, B., Haertlein, T.C., Carrasco, M.A., Tapia, J.C., Zhai, B., Das, R., Lalancette-Hebert, M. et al. FUS-SMN Protein Interactions Link the Motor Neuron Diseases ALS and SMA. *Cell Reports*, 2(4):799–806, 2012. doi: 10.1016/j.celrep.2012.08.025.

Yang, L., Embree, L.J., Tsai, S. and Hickstein, D.D. Oncoprotein TLS interacts with serine-arginine proteins involved in RNA splicing. *Journal of Biological Chemistry*, 273(43):27761–27764, 1998. doi: 10.1074/jbc.273.43.27761.

Yap, K. and Makeyev, E.V. Regulation of gene expression in mammalian nervous system through alternative pre-mRNA splicing coupled with RNA quality control mechanisms. *Molecular and Cellular Neuroscience*, 56:420–428, 2013. doi: 10.1016/j.mcn.2013.01.003.

Yasuda, K., Zhang, H., Loiselle, D., Haystead, T., Macara, I.G. and Mili, S. The RNA-binding protein Fus directs translation of localized mrnas in APC-RNP granules. *Journal of Cell Biology*, 203(5):737–746, 2013. doi: 10.1083/jcb.201306058.

Yeo, G., Gene, Y. and Burge, C.B. Maximum entropy modeling of short sequence motifs

with applications to RNA splicing signals. *Journal of Computational Biology*, 11(2-3): 377–394, 2004.

Yoshizawa, T., Ali, R., Jiou, J., Fung, H.Y.J., Burke, K.A., Kim, S.J., Lin, Y., Peeples, W.B., Saltzberg, D., Soniat, M. et al. Nuclear Import Receptor Inhibits Phase Separation of FUS through Binding to Multiple Sites. *Cell*, 173(3):693–705.e22, 2018. doi: 10.1016/j.cell.2018.03.003.

Young, M.D., Wakefield, M.J., Smyth, G.K. and Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2):R14, 2010. doi: 10.1186/gb-2010-11-2-r14.

Yu, Y. and Reed, R. FUS functions in coupling transcription to splicing by mediating an interaction between RNAP II and U1 snRNP. *Proceedings of the National Academy of Sciences*, 112(28), 2015. doi: 10.1073/pnas.1506282112.

Yu, Y., Chi, B., Xia, W., Gangopadhyay, J., Yamazaki, T., Winkelbauer-Hurt, M.E., Yin, S., Eliasse, Y., Adams, E., Shaw, C.E. et al. U1 snRNP is mislocalized in ALS patient fibroblasts bearing NLS mutations in FUS and is required for motor neuron outgrowth in zebrafish. *Nucleic Acids Research*, 43(6):3208–3218, 2015. doi: 10.1093/nar/gkv157.

Zarnack, K., Koenig, J., Tajnik, M., Martincorena, I., Eustermann, S., Stevant, I., Reyes, A., Anders, S., Luscombe, N.M. and Ule, J. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, 152(3): 453–466, 2013. doi: 10.1016/j.cell.2012.12.023.

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L.E., La Manno, G. et al. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014.e22, 2018. doi: 10.1016/j.cell.2018.06.021.

Zhang, C., Zhang, B., Lin, L.L. and Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, 18(1):1–11, 2017. doi: 10.1186/s12864-017-4002-1.

Zhang, Y.J., Xu, Y.F., Cook, C., Gendron, T.F., Roettges, P., Link, C.D., Lin, W.L., Tong, J., Castanedes-Casey, M., Ash, P. et al. Aberrant cleavage of TDP-43 enhances aggregation and cellular toxicity. *Proceedings of the National Academy of Sciences*, 106 (18):7607–7612, may 2009. doi: 10.1073/pnas.0900688106.

Zhao, S., Zhang, Y., Gamini, R., Zhang, B. and Von Schack, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: PolyA+ selection versus rRNA depletion. *Scientific Reports*, 8(1):1–12, 2018. doi: 10.1038/s41598-018-23226-4.

Zhou, Y., Liu, S., Liu, G., Öztürk, A. and Hicks, G.G. ALS-associated FUS mutations result in compromised FUS alternative splicing and autoregulation. *PLoS Genetics*, 9 (10), 2013. doi: 10.1371/journal.pgen.1003895.

8 Appendices

8.1 Appendices to chapter 3

Human cryptic exons flanked by 100nt

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1		1e-12	-2.972e+01	12.63%	0.44%	166.5bp (201.8bp)
2 *		1e-11	-2.706e+01	11.58%	0.41%	74.6bp (106.1bp)
3 *		1e-11	-2.685e+01	6.32%	0.01%	89.4bp (0.0bp)
4 *		1e-11	-2.682e+01	23.16%	3.45%	105.3bp (216.9bp)
5 *		1e-11	-2.659e+01	7.37%	0.04%	81.5bp (66.3bp)
6 *		1e-10	-2.470e+01	20.00%	2.71%	75.4bp (216.6bp)
7 *		1e-10	-2.440e+01	24.21%	4.35%	94.6bp (182.4bp)
8 *		1e-10	-2.354e+01	14.74%	1.29%	73.1bp (104.5bp)
9 *		1e-10	-2.341e+01	11.58%	0.62%	94.0bp (108.6bp)
10 *		1e-10	-2.340e+01	10.53%	0.45%	73.9bp (85.8bp)

Mouse cryptic exons flanked by 100nt

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1 *		1e-11	-2.658e+01	19.23%	0.54%	24.4bp (82.6bp)
2 *		1e-10	-2.528e+01	21.15%	0.92%	65.3bp (66.0bp)
3 *		1e-9	-2.276e+01	15.38%	0.36%	58.9bp (40.8bp)
4 *		1e-9	-2.245e+01	17.31%	0.60%	81.4bp (73.2bp)
5 *		1e-9	-2.156e+01	11.54%	0.09%	78.6bp (9.1bp)
6 *		1e-9	-2.103e+01	13.46%	0.24%	72.1bp (74.4bp)
7 *		1e-8	-2.067e+01	9.62%	0.03%	102.8bp (24.8bp)
8 *		1e-8	-2.002e+01	15.38%	0.53%	75.7bp (119.5bp)
9 *		1e-8	-1.890e+01	9.62%	0.05%	47.7bp (61.9bp)
10 *		1e-7	-1.832e+01	15.38%	0.70%	124.3bp (143.7bp)

Figure 8.1: Cryptic exon motifs found by *HOMER*

All mouse cryptic exons discovered by *CryptEx*

gene	Ensembl ID	exon ID	chr	exon start	exon end	+/-	5'ΔPSI	3'ΔPSI	classification	observed in	EScell log_2FC	brain log_2FC	*phyloP*	prediction
Abr	ENSMUSG00000017631	E026i1	chr11	76468524	76468676	-	0.09	0.00	5' extension	EScell;brain	-1.75	-0.59	0.77	PTC/frame shifted
Pnpla6	ENSMUSG00000004565	E016i1	chr8	3524947	3525140	+	0.29	0.00	5' extension	brain	-1.56	-0.47	0.08	PTC/frame shifted
Lrrc49	ENSMUSG00000047766	E017i3	chr9	60637648	60637695	-	0.06	0.00	5' extension	brain	-1.19	-0.25	0.17	PTC/frame shifted
Ccdc149	ENSMUSG00000045790	E009i3	chr5	52434250	52434424	-	0.16	0.00	5' extension	brain	.	-0.28	0.06	PTC/frame conserved
Ak5	ENSMUSG00000039058	E014i6	chr3	152643869	152643915	-	0.16	0.00	5' extension	brain	.	-0.53	0.17	PTC/frame shifted
Kctd10	ENSMUSG00000001098	E017i2	chr5	114376677	114376866	-	0.07	0.01	5' extension	brain	.	.	0.02	PTC/frame shifted
Tmem14a	ENSMUSG00000025933	E005i1	chr1	21229121	21229173	+	0.37	0.00	5' extension	brain	-1.02	-0.60	-0.08	PTC/frame conserved
Apod	ENSMUSG00000022548	E006i1	chr16	31298051	31298471	-	0.04	0.08	3' extension	brain	.	0.74	0.09	PTC/frame shifted
Erbb2ip	ENSMUSG00000021709	E026i1	chr13	103862527	103862588	-	0.00	0.25	3' extension	brain	.	0.62	-0.05	PTC/frame shifted
Trim8	ENSMUSG00000025034	E001i2	chr19	46506869	46506928	+	0.00	0.10	3' extension	Ling;brain	.	.	-0.23	PTC/frame shifted
Usp15	ENSMUSG00000099088	E015i3	chr10	123150961	123151125	-	0.00	0.32	3' extension	Ling;brain	.	.	0.15	PTC/frame conserved
Fam135a	ENSMUSG00000026153	E008i1	chr1	24022506	24022670	-	0.00	0.25	3' extension	brain	.	.	0.00	PTC/frame shifted
Rasal1	ENSMUSG00000029602	E007i1	chr5	120653508	120653879	+	0.03	0.16	3' extension	brain	.	.	-0.23	PTC/frame conserved
Brms1l	ENSMUSG00000012076	E008i1	chr12	55864624	55865091	+	0.04	0.23	3' extension	brain	-0.81	.	0.14	PTC/frame conserved
Cacna1b	ENSMUSG00000004113	E050i1	chr2	24691959	24692366	-	0.00	0.39	3' extension	brain	.	-0.58	0.21	PTC/frame shifted
Nt5c3b	ENSMUSG00000017176	E010i1	chr11	100433637	100433684	-	0.00	0.07	3' extension	brain	-1.15	-0.50	0.12	PTC/frame conserved
Spcs2	ENSMUSG00000035227	E004i1	chr7	99856412	99856514	-	0.01	0.29	3' extension	Ling;EScell;brain	.	0.34	0.40	PTC/frame shifted
Ccdc88a	ENSMUSG00000032740	E004i8	chr11	29421354	29421401	+	0.00	0.27	3' extension	brain	-1.63	.	-0.04	PTC/frame conserved
Adap1	ENSMUSG00000056413	E014i3	chr5	139302108	139302166	-	0.00	0.36	3' extension	brain	.	-0.47	0.08	PTC/frame shifted
Ap3b2	ENSMUSG00000062444	E031i2	chr7	81490178	81490277	-	0.04	0.10	3' extension	brain	0.50	-0.28	0.18	PTC/frame shifted
Fam73a	ENSMUSG00000054942	E015i2	chr3	152292556	152292657	-	0.02	0.09	3' extension	Ling;brain	-1.39	-0.63	-0.04	PTC/frame shifted
Ptprn2	ENSMUSG00000056553	E019i11	chr12	117154249	117154298	+	0.00	0.13	3' extension	brain	.	-1.01	0.24	PTC/frame shifted
Dlg2	ENSMUSG00000052572	E004i17	chr7	91543136	91543191	+	0.00	0.23	3' extension	brain	.	-0.71	1.67	PTC/frame shifted
Cd200	ENSMUSG00000022661	E003i1	chr16	45389970	45390184	-	0.04	0.06	3' extension	brain	.	.	-0.13	Not in CDS
Dock4	ENSMUSG00000035954	E049i1	chr12	40835769	40835912	+	0.09	0.20	Cassette	brain	.	-0.68	0.53	PTC/frame shifted
C4b	ENSMUSG00000073418	E032i1	chr17	34739389	34739500	-	0.06	0.30	Cassette	brain	.	.	0.00	PTC/frame shifted
Camk1g	ENSMUSG00000016179	E017i2	chr1	193368866	193368952	-	0.70	0.47	Cassette	brain	.	.	-0.06	Not in CDS
Reep3	ENSMUSG00000019873	E001i1	chr10	67014087	67014164	-	0.67	0.39	Cassette	brain	.	0.44	-0.01	PTC/frame shifted
Drosha	ENSMUSG00000022191	E009i1	chr15	12833221	12833389	+	0.10	0.15	Cassette	brain	.	.	-0.06	PTC/frame conserved
Synj2bp	ENSMUSG00000021139	E036i2	chr12	81509671	81510249	-	0.41	0.07	Cassette	brain	.	.	0.11	PTC/frame shifted
Adnp2	ENSMUSG00000053950	E002i1	chr18	80138152	80138304	-	0.57	0.50	Cassette	Ling;EScell;brain	0.33	0.25	0.03	PTC/frame shifted
Adipor2	ENSMUSG00000030168	E010i2	chr6	119369859	119369918	-	0.30	0.44	Cassette	Ling;brain	-1.06	.	0.07	PTC/frame shifted
Cdh22	ENSMUSG00000053166	E013i2	chr2	165183238	165183371	-	0.40	0.67	Cassette	brain	.	-0.35	0.13	Not in CDS
Ermn	ENSMUSG00000026830	E001i1	chr2	58049485	58049561	-	0.16	0.16	Cassette	brain	.	.	0.06	PTC/frame shifted
Ift81	ENSMUSG00000029469	E003i1	chr5	122556905	122556991	-	0.19	0.08	Cassette	Ling;brain	.	.	0.08	benign/frame shifted
Vps13d	ENSMUSG00000020220	E043i1	chr4	145099351	145099463	-	0.60	0.20	Cassette	brain	.	-0.81	0.23	PTC/frame shifted
Ube2d1	ENSMUSG00000019927	E010i1	chr10	71263584	71263717	-	0.15	0.36	Cassette	EScell;brain	.	0.35	0.04	PTC/frame shifted
Hgsnat	ENSMUSG00000037260	E001i1	chr8	25945947	25945999	-	0.28	0.15	Cassette	Ling;brain	-0.59	.	0.20	PTC/frame conserved
Elmod1	ENSMUSG00000041986	E003i2	chr9	53923941	53923995	-	0.27	0.19	Cassette	brain	.	.	0.97	benign/frame conserved
Mkx	ENSMUSG00000061013	E007i7	chr18	6986138	6986235	-	0.24	0.50	Cassette	brain	.	.	-0.08	PTC/frame shifted
Hace1	ENSMUSG00000038822	E013i1	chr10	45639722	45639803	+	0.06	0.20	Cassette	Ling;brain	.	-0.38	0.10	benign/frame conserved
Fxyd2	ENSMUSG00000059412	E003i1	chr9	45408987	45409173	+	0.42	0.18	Cassette	brain	-1.89	.	-0.25	PTC/frame shifted
Thoc7	ENSMUSG00000053453	E007i1	chr14	13957061	13957631	-	0.21	-0.10	5' extension	EScell	-0.74	.	0.04	PTC/frame shifted
Nme6	ENSMUSG00000032478	E020i1	chr9	109837872	109838225	+	0.40	0.00	5' extension	Ling;EScell	-0.84	.	0.21	PTC/frame conserved
Wtip	ENSMUSG00000036459	E001i1	chr7	34111074	34111099	-	0.13	0.00	5' extension	EScell	-0.46	.	0.26	PTC/frame shifted
Gsta4	ENSMUSG00000032348	E005i2	chr9	78207196	78207382	+	0.02	0.07	3' extension	Ling;EScell	-0.45	.	0.11	PTC/frame shifted
Slc7a6	ENSMUSG00000031904	E001i1	chr8	106169186	106169256	+	0.00	0.12	3' extension	EScell	.	-0.23	-0.25	Not in CDS
Flnb	ENSMUSG00000025278	E040i1	chr14	7935368	7935630	+	0.09	0.43	Cassette	Ling;EScell;brain	-1.34	.	-0.12	PTC/frame shifted
B4galnt3	ENSMUSG00000041372	E002i1	chr6	120204263	120204465	-	0.07	0.10	Cassette	EScell	.	.	-0.15	PTC/frame shifted
Psma2	ENSMUSG00000015671	E005i1	chr13	14620930	14621038	+	0.06	0.13	Cassette	EScell	.	.	0.19	PTC/frame conserved
Abr	ENSMUSG00000017631	E024i1	chr11	76464580	76464758	-	0.18	0.07	Cassette	EScell;brain	-1.75	-0.59	-0.25	PTC/frame shifted
Wasf3	ENSMUSG00000029636	E009i1	chr5	146469766	146469912	+	0.10	0.22	Cassette	EScell	.	.	-0.02	benign/frame conserved

3. All human cryptic exons discovered by *CryptEx*

gene	Ensembl ID	exon ID	chr	exon start	exon end	+/-	5'ΔPSI	3'ΔPSI	classification	observed in	mRNA FC	total FC	*phyloP*	prediction	*maxEnt* 5'	*maxEnt* 3'
PHF12	ENSG00000109118	E028i1	chr17	28914903	28915475	-	0.14	0.00	5' extension	mRNA;total	0.23	.	0.448493	PTC/frame conserved	.	9.93
ZFPM2	ENSG00000169946	E016i1	chr8	105419822	105419906	+	0.35	0.00	5' extension	mRNA	.	.	-0.131815	PTC/frame conserved	.	11.28
ACBD3	ENSG00000182827	E004i1	chr1	226156701	226156780	-	0.12	0.00	5' extension	mRNA	-0.31	.	0.22671	PTC/frame shifted	.	9.63
RPS6KA3	ENSG00000177189	E031i2	chrX	20241592	20241686	-	0.18	0.01	5' extension	mRNA	-0.42	.	2.38738	PTC/frame shifted	.	8.09
CPED1	ENSG00000106034	E034i7	chr7	121194471	121194593	+	0.07	0.01	5' extension	mRNA	.	.	0.398074	PTC/frame shifted	.	9.56
HUWE1	ENSG00000086758	E033i1	chrX	53551990	53552074	-	0.09	0.05	5' extension	mRNA	-0.56	.	0.114241	PTC/frame shifted	.	6.87
DYNC2LI1	ENSG00000138036	E006i1	chr2	43774402	43774936	+	0.08	0.04	5' extension	mRNA	-0.63	.	-0.228969	PTC/frame shifted	.	2.09
ZMYND8	ENSG00000101040	E019i2	chr20	47254235	47254424	-	0.07	0.03	5' extension	mRNA	-0.36	.	0.101961	PTC/frame shifted	.	3.77
HEPH	ENSG00000089472	E030i1	chrX	66257344	66257516	+	0.12	0.00	5' extension	mRNA	-0.58	-0.34	0.245606	PTC/frame shifted	.	14.19
BTNL9	ENSG00000165810	E031i1	chr5	181059149	181059215	+	0.09	0.00	5' extension	mRNA	.	.	-0.101275	PTC/frame shifted	.	-0.66
PAPSS1	ENSG00000138801	E008i1	chr4	107654948	107655036	-	0.08	0.00	5' extension	mRNA;total	-0.87	.	0.746813	PTC/frame conserved	.	10.84
CLASP1	ENSG00000074054	E049i1	chr2	121471187	121471274	-	0.16	0.00	5' extension	mRNA	-0.19	.	-0.0460715	PTC/frame shifted	.	4.56
IST1	ENSG00000224470	E053i1	chr16	71922108	71922437	+	0.12	0.01	5' extension	mRNA	.	.	0.177172	PTC/frame shifted	.	13.96
UBR4	ENSG00000272084	E084i1	chr1	19145598	19145708	-	0.05	0.00	5' extension	mRNA	.	.	0.710974	PTC/frame shifted	.	7.28
ANKRD27	ENSG00000105186	E019i1	chr19	32630070	32630151	-	0.15	0.01	5' extension	mRNA	-0.37	-0.24	-0.103821	PTC/frame shifted	.	12.52
CCDC77	ENSG00000120647	E020i1	chr12	422421	422891	+	0.07	0.01	5' extension	mRNA	-0.84	.	-0.0231545	PTC/frame shifted	.	3.79
LRPPRC	ENSG00000138095	E024i1	chr2	43942675	43942815	-	0.06	0.00	5' extension	mRNA	-1.01	.	0.14133	PTC/frame shifted	.	3.85
ADGRV1	ENSG00000164199	E085i1	chr5	90803514	90803596	+	0.07	0.00	5' extension	mRNA	-0.47	0.32	-0.185377	PTC/frame shifted	.	9.14
ASNSP1	ENSG00000248498	E006i1	chr8	46595259	46595315	-	0.13	0.00	5' extension	mRNA	.	.	1.14798	Not in CDS	.	8.29
SMAD2	ENSG00000175387	E010i1	chr18	47868051	47868103	-	0.08	0.02	5' extension	mRNA	-0.39	.	0.289826	PTC/frame shifted	.	9.95
ZNHIT6	ENSG00000117174	E003i1	chr1	85662210	85662278	-	0.08	0.04	5' extension	mRNA	-0.80	.	0.0230322	PTC/frame conserved	.	5.75
NLRP1	ENSG00000091592	E045i1	chr17	5585499	5585635	-	0.15	0.00	5' extension	mRNA	.	.	-0.1356	Not in CDS	.	5.48
CEP290	ENSG00000198707	E024i1	chr12	88086709	88086789	-	0.13	0.00	5' extension	mRNA	-1.15	.	0.157693	PTC/frame shifted	.	7.83
METTL8	ENSG00000123600	E037i1	chr2	171433562	171433649	-	0.50	0.00	5' extension	mRNA	.	.	0.223589	Not in CDS	.	11.06
MPDZ	ENSG00000107186	E015i1	chr9	13110935	13111024	-	0.35	0.00	5' extension	mRNA	-0.57	.	0.11092	PTC/frame shifted	.	10.52
ERBB2IP	ENSG00000112851	E013i2	chr5	65997127	65997228	+	0.06	0.04	5' extension	mRNA	-0.39	.	0.211071	PTC/frame shifted	.	9.62
PRPF40A	ENSG00000196504	E001i1	chr2	152657053	152657141	-	0.04	0.05	3' extension	mRNA	-0.77	.	-0.0653744	PTC/frame shifted	-0.49	.
UPF2	ENSG00000151461	E002i1	chr10	11926921	11927006	-	0.00	0.16	3' extension	Ling;mRNA	.	.	0.00227976	PTC/frame shifted	9.85	.
RRP1	ENSG00000160214	E008i1	chr21	43791522	43792578	+	0.02	0.08	3' extension	mRNA	0.82	.	-0.184698	PTC/frame shifted	6.51	.
LRP8	ENSG00000157193	E033i1	chr1	53274090	53274694	-	0.00	0.27	3' extension	mRNA;total	.	.	1.77583	PTC/frame conserved	2.70	.
PKP4	ENSG00000144283	E020i1	chr2	158588958	158588987	+	0.00	0.17	3' extension	mRNA	-0.31	.	0.0933565	PTC/frame shifted	3.91	.
DDX52	ENSG00000278053	E022i1	chr17	37629848	37629963	-	0.01	0.08	3' extension	mRNA;total	-0.39	-0.30	0.981172	PTC/frame conserved	7.96	.
SETD5	ENSG00000168137	E065i1	chr3	9468576	9468615	+	0.00	0.13	3' extension	Ling;mRNA;total	-0.24	.	2.57544	PTC/frame conserved	7.36	.
KDELC2	ENSG00000178202	E018i2	chr11	108497751	108498085	-	0.00	0.19	3' extension	Ling;mRNA	-0.78	.	-0.0860609	PTC/frame conserved	8.95	.
GPT2	ENSG00000166123	E009i1	chr16	46900812	46900883	+	0.00	0.16	3' extension	mRNA	-0.98	-0.72	1.01736	PTC/frame shifted	3.30	.
GSE1	ENSG00000131149	E007i1	chr16	85650416	85650709	+	0.04	0.06	3' extension	mRNA	0.29	.	-0.525771	PTC/frame shifted	5.99	.
MBOAT2	ENSG00000143797	E029i2	chr2	8986052	8986136	-	0.00	0.25	3' extension	mRNA	-0.66	.	-0.0138646	PTC/frame shifted	8.34	.
DNAAF5	ENSG00000164818	E024i1	chr7	783225	783309	+	0.00	0.05	3' extension	mRNA	0.49	-0.37	-0.238231	PTC/frame shifted	9.25	.
CLASP1	ENSG00000074054	E012i1	chr2	121387587	121387655	-	0.04	0.07	3' extension	mRNA	-0.19	.	-0.0066914	PTC/frame conserved	2.72	.
ZFP91	ENSG00000255073	E012i1	chr11	58616974	58617182	+	0.04	0.05	3' extension	mRNA	.	.	-0.213922	PTC/frame shifted	6.97	.
FAM178B	ENSG00000273634	E017i1	chr2	96923867	96924228	-	0.00	0.30	3' extension	mRNA	.	.	-0.351309	PTC/frame shifted	3.14	.
ANAPC1	ENSG00000153107	E020i1	chr2	111799529	111799699	-	0.00	0.12	3' extension	mRNA	-0.90	.	0.0276639	PTC/frame conserved	5.28	.
CERK	ENSG00000100422	E005i1	chr22	46691869	46692022	-	0.03	0.14	3' extension	mRNA	.	.	-0.579144	PTC/frame shifted	9.22	.
ARHGAP11B	ENSG00000187951	E033i1	chr15	30755981	30756067	+	0.00	0.11	3' extension	mRNA	.	.	-0.13533	Not in CDS	10.77	.
GOLGA8A	ENSG00000175265	E037i1	chr15	34436658	34436837	-	0.00	0.10	3' extension	mRNA	-0.58	.	-0.147737	Not in CDS	8.72	.
TGFBRAP1	ENSG00000135966	E014i2	chr2	105307479	105307570	-	0.00	0.19	3' extension	mRNA	.	.	0.807447	PTC/frame shifted	4.75	.
CAMK2G	ENSG00000148660	E039i1	chr10	73873200	73873283	-	0.00	0.21	3' extension	mRNA	.	.	0.336877	PTC/frame shifted	6.32	.

3. All human cryptic exons discovered by *CryptEx* (continued)

gene	Ensembl ID	exon ID	chr	exon start	exon end	+/-	5'ΔPSI	3'ΔPSI	classification	observed in	mRNA FC	total FC	*phyloP*	prediction	*maxEnt* 5'	*maxEnt* 3'
DCAF17	ENSG00000115827	E008i1	chr2	171441954	171442195	+	0.04	0.07	3' extension	mRNA	-0.50	.	0.207598	PTC/frame conserved	5.13	.
RP1-179N16.6	ENSG00000246982	E002i2	chr6	36160683	36160784	-	0.02	0.13	3' extension	mRNA	.	.	0.161445	Not in CDS	9.16	.
RNFT2	ENSG00000135119	E016i2	chr12	116790033	116790087	+	0.00	0.15	3' extension	Ling;mRNA	.	.	0.118938	PTC/frame shifted	7.03	.
SH2B3	ENSG00000111252	E005i1	chr12	111446092	111446216	+	0.00	0.11	3' extension	mRNA	.	-0.52	-0.401417	PTC/frame shifted	5.85	.
SENP7	ENSG00000138468	E020i1	chr3	101349160	101349188	-	0.00	0.05	3' extension	mRNA	-1.11	.	0.116452	PTC/frame shifted	6.06	.
GOLGB1	ENSG00000173230	E013i1	chr3	121682916	121682980	-	0.03	0.08	3' extension	mRNA	-0.40	0.26	0.918212	PTC/frame conserved	9.35	.
AGRN	ENSG00000242590	E015i1	chr1	1044440	1045080	+	0.46	0.63	Cassette	Ling;mRNA;total	.	.	-0.913282	PTC/frame shifted	5.78	4.71
CEP72	ENSG00000112877	E012i1	chr5	648049	648223	+	0.69	0.70	Cassette	Ling;mRNA;total	.	-0.30	-0.102364	PTC/frame shifted	9.81	-7.11
RAP1GAP	ENSG00000076864	E010i1	chr1	21598718	21598820	-	0.27	0.34	Cassette	mRNA	.	-0.39	-0.293165	PTC/frame shifted	6.51	7.06
PFKP	ENSG00000067057	E010i1	chr10	3099556	3099819	+	0.29	0.19	Cassette	Ling;mRNA;total	.	.	-0.0943177	PTC/frame shifted	4.88	2.02
PKN1	ENSG00000123143	E016i2	chr19	14450087	14450214	+	0.22	0.10	Cassette	Ling;mRNA	.	-0.49	0.0079453	PTC/frame shifted	4.88	6.94
FAM114A2	ENSG00000055147	E040i1	chr5	154037367	154037457	-	0.27	0.23	Cassette	Ling;mRNA;total	.	.	0.205094	Not in CDS	8.11	11.04
HDGFRP2	ENSG00000167674	E015i1	chr19	4492012	4492152	+	0.15	0.10	Cassette	Ling;mRNA	0.90	.	-0.606886	benign/frame conserved	6.14	7.67
PRPF38A	ENSG00000134748	E003i1	chr1	52405054	52405321	+	0.05	0.07	Cassette	mRNA	.	.	0.106102	PTC/frame conserved	6.99	6.54
ATG4B	ENSG00000168397	E047i1	chr2	241668874	241669149	+	0.10	0.24	Cassette	Ling;mRNA	0.47	.	-0.307633	PTC/frame conserved	10.13	10.45
ADSS	ENSG00000035687	E020i1	chr1	244433371	244433496	-	0.11	0.10	Cassette	mRNA	.	.	0.185587	PTC/frame shifted	7.83	7.65
NSUN2	ENSG00000037474	E024i1	chr5	6619315	6619467	-	0.07	0.07	Cassette	mRNA	-0.39	-0.39	-0.517291	PTC/frame shifted	9.14	11.12
TAOK3	ENSG00000135090	E007i1	chr12	118151926	118151997	-	0.05	0.06	Cassette	mRNA	.	.	0.0363395	PTC/frame conserved	8.56	8.68
CORO7	ENSG00000103426	E081i1	chr16	4367297	4367729	-	0.25	0.10	Cassette	mRNA	.	.	-0.121763	PTC/frame shifted	3.34	5.19
HBG2	ENSG00000196565	E016i1	chr11	5258008	5258776	-	0.58	0.34	Cassette	mRNA;total	.	.	-0.0145071	Not in CDS	5.91	10.53
SSFA2	ENSG00000138434	E024i1	chr2	181911125	181911233	+	0.23	0.22	Cassette	Ling;mRNA;total	-0.32	.	0.0525638	PTC/frame conserved	6.69	6.70
MAR3	ENSG00000173926	E009i1	chr5	126917510	126917676	-	0.14	0.05	Cassette	mRNA	0.51	.	0.178889	PTC/frame shifted	3.82	7.74
HERC1	ENSG00000103657	E014i1	chr15	63629529	63629697	-	0.05	0.09	Cassette	mRNA	-0.74	-0.31	0.164758	PTC/frame conserved	3.53	4.58
SP3	ENSG00000172845	E011i1	chr2	173957233	173957376	-	0.08	0.20	Cassette	mRNA	-0.29	.	0.236352	PTC/frame conserved	3.78	10.06
DDHD1	ENSG00000100523	E023i1	chr14	53096116	53096218	-	0.21	0.42	Cassette	mRNA	.	.	3.39222	benign/frame conserved	8.04	5.98
RP11-488L18.4	ENSG00000227671	E001i4	chr1	247197079	247197247	-	0.10	0.07	Cassette	mRNA	.	.	0.00769496	Not in CDS	5.17	8.01
UHRF2	ENSG00000147854	E017i2	chr9	6465743	6465846	+	0.16	0.34	Cassette	mRNA	-0.58	.	0.670822	PTC/frame shifted	5.91	8.18
REV3L	ENSG00000009413	E028i1	chr6	111364943	111365042	-	0.19	0.10	Cassette	mRNA	-0.42	.	1.78129	Not in CDS	10.08	2.92
SLC12A2	ENSG00000064651	E012i1	chr5	128143124	128143182	+	0.08	0.20	Cassette	mRNA	-0.56	.	0.00065344	benign/frame conserved	10.15	8.69
MED23	ENSG00000112282	E025i1	chr6	131599465	131599512	-	0.12	0.12	Cassette	mRNA	-0.75	.	0.108663	PTC/frame shifted	3.14	10.13
POLK	ENSG00000122008	E019i1	chr5	75580537	75580602	+	0.14	0.16	Cassette	mRNA	-0.74	.	0.330324	benign/frame conserved	8.55	2.26
TRPM7	ENSG00000092439	E027i1	chr15	50588190	50588250	-	0.08	0.07	Cassette	mRNA	-0.35	.	0.391366	benign/frame conserved	8.27	5.73
Tretained intronM37	ENSG00000108395	E029i1	chr17	59076653	59076734	-	0.11	0.13	Cassette	mRNA	-0.33	.	0.0130702	PTC/frame conserved	8.72	3.56
C4orf32	ENSG00000174749	E001i1	chr4	112180673	112180858	+	0.20	0.19	Cassette	mRNA	-1.16	-0.62	0.271923	PTC/frame shifted	4.80	5.88
ZCCHC6	ENSG00000083223	E004i1	chr9	86301739	86301828	-	0.06	0.06	Cassette	mRNA	.	0.25	0.139962	PTC/frame shifted	-0.63	6.03
ANKRD26	ENSG00000107890	E045i1	chr10	27098179	27098326	-	0.06	0.11	Cassette	mRNA	-1.17	.	0.103968	PTC/frame conserved	9.06	8.66
KIAA1429	ENSG00000164944	E039i3	chr8	94548144	94548207	-	0.08	0.08	Cassette	mRNA	-0.23	.	0.0650052	PTC/frame shifted	9.72	9.69
ZFYVE27	ENSG00000155256	E020i1	chr10	97754726	97754917	+	0.09	-0.01	5' extension	total	0.94	.	-0.0173503	PTC/frame shifted	.	9.37
MYO16	ENSG00000041515	E036i1	chr13	109106632	109106716	+	0.17	0.00	5' extension	total	0.27	0.47	0.452463	PTC/frame shifted	.	4.85
INPP5A	ENSG00000068383	E009i1	chr10	132705192	132705232	+	0.12	0.00	5' extension	total	-0.39	-0.37	-0.138522	PTC/frame shifted	.	10.69
SCP2	ENSG00000226147	E024i1	chr1	53002787	53002870	+	0.09	0.00	5' extension	total	.	.	0.254126	PTC/frame shifted	.	3.40
DEAF1	ENSG00000177030	E006i2	chr11	659434	659796	-	0.06	-0.01	5' extension	total	0.43	.	-0.0376266	PTC/frame shifted	.	10.92
PPP1R14B	ENSG00000173457	E012i1	chr11	64249088	64249175	-	0.00	0.15	3' extension	total	.	.	-0.0648612	Not in CDS	-34.33	.
SLC19A1	ENSG00000173638	E028i1	chr21	45541454	45541659	-	0.00	0.06	3' extension	total	0.46	-0.81	-0.685436	Not in CDS	6.03	.
HNRNPH3	ENSG00000096746	E009i1	chr10	68333071	68333102	+	0.00	0.08	3' extension	total	-0.35	.	0.783294	Not in CDS	6.58	.
SLC39A8	ENSG00000138821	E016i2	chr4	102324249	102324998	-	0.02	0.06	3' extension	total	.	.	-0.220484	PTC/frame conserved	8.55	.
DSC2	ENSG00000134755	E017i1	chr18	31101220	31101268	-	0.00	0.46	3' extension	total	.	.	-0.509009	PTC/frame conserved	7.96	.
ANXA3	ENSG00000138772	E023i1	chr4	78600147	78600348	+	0.00	0.10	3' extension	total	-1.23	-0.36	-0.0017593	PTC/frame shifted	9.14	.

8.2 Appendices to chapter 4

category	P-value	DEG	Down	Up	term
GO:0044429	3.90E−20	77	76	1	mitochondrial part
GO:0070469	3.13E−19	24	24	0	respiratory chain
GO:0005743	4.95E−19	55	54	1	mitochondrial inner membrane
GO:0005739	1.52E−18	134	126	8	mitochondrion
GO:0005746	2.21E−18	22	22	0	mitochondrial respiratory chain
GO:0044455	2.46E−17	32	32	0	mitochondrial membrane part
GO:0005740	6.67E−17	63	62	1	mitochondrial envelope
GO:0031966	2.00E−16	60	59	1	mitochondrial membrane
GO:0005747	4.11E−14	16	16	0	mitochondrial respiratory chain complex I
GO:0030964	4.11E−14	16	16	0	NADH dehydrogenase complex
GO:0045271	4.11E−14	16	16	0	respiratory chain complex I
GO:1990204	1.79E−12	20	20	0	oxidoreductase complex
GO:0045259	9.35E−09	9	9	0	proton-transporting ATP synthase complex
GO:0009055	1.65E−08	14	14	0	electron carrier activity
GO:0015078	2.43E−08	16	16	0	hydrogen ion transmembrane transporter activity
GO:0005753	6.01E−08	8	8	0	mitochondrial proton-transporting ATP synthase complex
GO:0016469	2.15E−07	11	11	0	proton-transporting two-sector ATPase complex
GO:0055114	3.90E−07	62	57	5	oxidation-reduction process
GO:0004129	6.04E−07	8	8	0	cytochrome-c oxidase activity
GO:0015002	6.04E−07	8	8	0	heme-copper terminal oxidase activity
GO:0016676	6.04E−07	8	8	0	oxidoreductase activity, acting on a heme group..
GO:0016675	9.69E−07	8	8	0	oxidoreductase activity, acting on a heme group...
GO:0045263	1.19E−06	6	6	0	proton-transporting ATP synthase complex
GO:0006119	1.22E−06	9	9	0	oxidative phosphorylation
GO:0042773	2.31E−06	7	7	0	ATP synthesis coupled electron transport
GO:0033177	3.78E−06	7	7	0	proton-transporting two-sector ATPase complex
GO:0008137	5.95E−06	7	7	0	NADH dehydrogenase (ubiquinone) activity
GO:0050136	5.95E−06	7	7	0	NADH dehydrogenase (quinone) activity
GO:0000276	7.68E−06	5	5	0	mitochondrial proton-transporting ATP synthase complex
GO:0003954	9.07E−06	7	7	0	NADH dehydrogenase activity
GO:0016655	1.30E−05	8	8	0	oxidoreductase activity, acting on NAD(P)H, etc
GO:0022900	1.34E−05	9	9	0	electron transport chain
GO:0042775	1.47E−05	6	6	0	mitochondrial ATP synthesis coupled electron transport
GO:0070069	3.21E−05	5	5	0	cytochrome complex
GO:0005761	4.78E−05	10	10	0	mitochondrial ribosome
GO:0005759	9.93E−05	18	18	0	mitochondrial matrix

GO:0044391	1.96E−13	26	26	0	ribosomal subunit
GO:0005840	3.10E−13	32	32	0	ribosome
GO:0003735	5.33E−12	24	24	0	structural constituent of ribosome
GO:0022626	2.35E−10	18	18	0	cytosolic ribosome
GO:0030529	6.74E−10	53	51	2	ribonucleoprotein complex
GO:0015935	3.21E−09	15	15	0	small ribosomal subunit
GO:0022627	6.33E−09	12	12	0	cytosolic small ribosomal subunit
GO:0015934	1.39E−05	11	11	0	large ribosomal subunit
GO:0006412	1.61E−05	35	35	0	translation
GO:0000313	4.78E−05	10	10	0	organellar ribosome
GO:0005839	5.95E−06	7	7	0	proteasome core complex
GO:0019773	7.68E−06	5	5	0	proteasome core complex, alpha-subunit complex
GO:0004298	1.34E−05	7	7	0	threonine-type endopeptidase activity
GO:0070003	1.34E−05	7	7	0	threonine-type peptidase activity
GO:0070062	1.96E−09	147	117	30	extracellular vesicular exosome
GO:0043230	2.15E−09	147	117	30	extracellular organelle
GO:0065010	2.15E−09	147	117	30	extracellular membrane-bounded organelle
GO:0031982	5.04E−08	174	134	40	vesicle
GO:0031988	8.20E−08	163	127	36	membrane-bounded vesicle
GO:0005198	2.08E−06	35	28	7	structural molecule activity
GO:0003723	0.000106434	83	74	9	RNA binding
GO:0045116	0.000111545	4	4	0	protein neddylation
GO:0019866	6.81E−19	57	55	2	organelle inner membrane
GO:0031967	2.50E−14	76	71	5	organelle envelope
GO:0031975	3.14E−14	76	71	5	envelope
GO:0031090	6.92E−08	97	88	9	organelle membrane
GO:0044444	3.05E−15	324	275	49	cytoplasmic part
GO:0043227	6.35E−12	467	365	102	membrane-bounded organelle
GO:0005737	3.03E−11	424	340	84	cytoplasm
GO:0044446	2.58E−10	244	207	37	intracellular organelle part
GO:0044422	8.66E−10	249	211	38	organelle part
GO:0043226	9.72E−10	489	381	108	organelle
GO:0043231	3.43E−09	414	332	82	intracellular membrane-bounded organelle
GO:0043229	7.98E−09	449	360	89	intracellular organelle
GO:0044421	9.02E−08	173	135	38	extracellular region part
GO:0032991	1.19E−07	217	176	41	macromolecular complex
GO:0005576	1.26E−07	191	147	44	extracellular region
GO:0044424	2.32E−07	494	392	102	intracellular part
GO:0044445	1.08E−06	21	21	0	cytosolic part
GO:0005622	2.31E−06	497	394	103	intracellular

Table 8.1: All gene ontology terms found in the 12 month FUS Δ14 spinal cords. DEG: differentially expressed genes at FDR 10%. Down and Up refer to the direction of differential gene expression of the genes in each ontology category.

8.3 Appendices to chapter 5

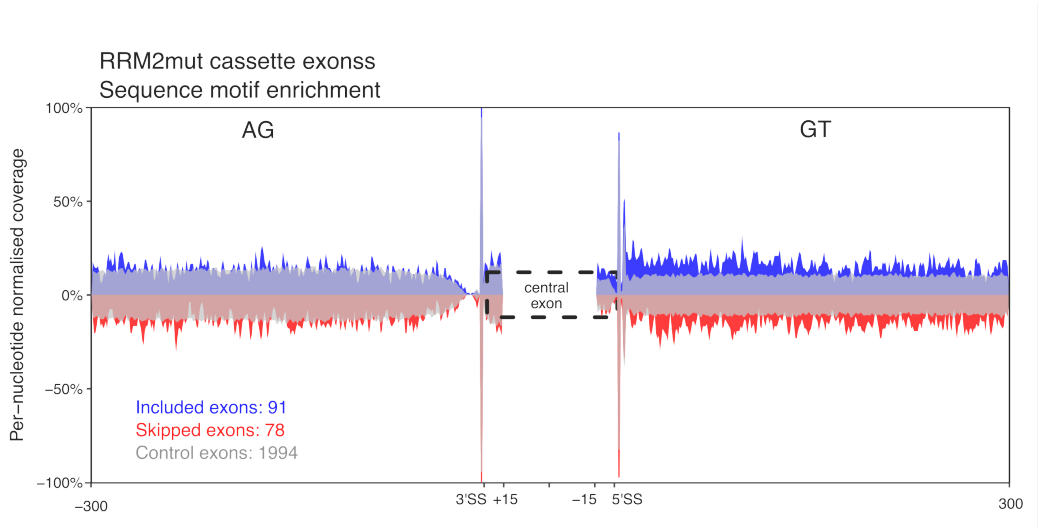


Figure 8.2: RNA maps of AG and GT dinucleotides are invariant at the 5' and 3' splice sites RNAmaps constructed from differentially included (blue) and skipped (red) cassette exons from RRM2mut.

Table 8.2: Results of permuting sample order and repeating splicing analysis. Exons refers to the number of differentially spliced exons found at FDR 5%, cryptics are those that satisfy the cryptic exon critera and skiptics are those that satisfy the skiptic exon criteria (see chapter). Groups in bold are the correct sample ordering. Note that definitions of skiptic and cryptic depend on which condition is the reference, which accounts for the different numbers between the two correct orderings.

groups	exons	cryptic	skiptic
M323K_WT_1+M323K_WT_2+M323K_WT_3+M323K_WT_4	920	2	47
M323K_HOM_1+M323K_HOM_2+M323K_HOM_3+M323K_HOM_4	920	4	9
M323K_HOM_2+M323K_HOM_3+M323K_HOM_4+M323K_WT_3	38	0	0
M323K_HOM_1+M323K_WT_1+M323K_WT_2+M323K_WT_4	38	0	0
M323K_HOM_1+M323K_HOM_2+M323K_HOM_4+M323K_WT_4	25	0	1
M323K_HOM_3+M323K_WT_1+M323K_WT_2+M323K_WT_3	25	0	0
M323K_HOM_1+M323K_HOM_3+M323K_HOM_4+M323K_WT_4	25	0	2
M323K_HOM_2+M323K_WT_1+M323K_WT_2+M323K_WT_3	25	1	0
M323K_HOM_1+M323K_HOM_2+M323K_HOM_3+M323K_WT_1	24	0	0
M323K_HOM_4+M323K_WT_2+M323K_WT_3+M323K_WT_4	24	0	0
M323K_HOM_2+M323K_WT_1+M323K_WT_2+M323K_WT_4	24	1	0
M323K_HOM_1+M323K_HOM_3+M323K_HOM_4+M323K_WT_3	24	0	0
M323K_HOM_1+M323K_WT_1+M323K_WT_2+M323K_WT_3	19	0	0
M323K_HOM_2+M323K_HOM_3+M323K_HOM_4+M323K_WT_4	19	1	0
M323K_HOM_1+M323K_HOM_2+M323K_HOM_4+M323K_WT_2	17	0	0
M323K_HOM_1+M323K_HOM_2+M323K_HOM_3+M323K_WT_3	15	0	0
M323K_HOM_4+M323K_WT_1+M323K_WT_2+M323K_WT_4	15	0	0
M323K_HOM_2+M323K_HOM_3+M323K_HOM_4+M323K_WT_1	12	0	0
M323K_HOM_1+M323K_HOM_2+M323K_HOM_3+M323K_WT_4	7	0	0
M323K_HOM_1+M323K_HOM_3+M323K_HOM_4+M323K_WT_2	7	0	0
M323K_HOM_4+M323K_WT_1+M323K_WT_2+M323K_WT_3	7	0	0
M323K_HOM_3+M323K_WT_2+M323K_WT_3+M323K_WT_4	7	0	0
M323K_HOM_1+M323K_HOM_3+M323K_HOM_4+M323K_WT_1	5	0	0
M323K_HOM_2+M323K_WT_2+M323K_WT_3+M323K_WT_4	5	0	0
M323K_HOM_3+M323K_WT_1+M323K_WT_2+M323K_WT_4	2	0	0
M323K_HOM_1+M323K_HOM_2+M323K_HOM_3+M323K_WT_2	1	0	0
M323K_HOM_4+M323K_WT_1+M323K_WT_3+M323K_WT_4	1	0	0
M323K_HOM_3+M323K_HOM_4+M323K_WT_3+M323K_WT_4	51	0	0
M323K_HOM_1+M323K_HOM_4+M323K_WT_1+M323K_WT_4	11	0	0
M323K_HOM_2+M323K_HOM_3+M323K_WT_2+M323K_WT_3	11	0	0
M323K_HOM_1+M323K_HOM_2+M323K_WT_1+M323K_WT_3	11	1	0
M323K_HOM_3+M323K_HOM_4+M323K_WT_2+M323K_WT_4	11	0	0
M323K_HOM_2+M323K_HOM_4+M323K_WT_3+M323K_WT_4	11	0	0
M323K_HOM_2+M323K_HOM_3+M323K_WT_1+M323K_WT_3	9	0	0
M323K_HOM_1+M323K_HOM_4+M323K_WT_2+M323K_WT_4	9	0	0
M323K_HOM_1+M323K_HOM_2+M323K_WT_3+M323K_WT_4	9	0	0
M323K_HOM_3+M323K_HOM_4+M323K_WT_1+M323K_WT_2	9	0	0
M323K_HOM_2+M323K_HOM_3+M323K_WT_1+M323K_WT_4	9	0	0
M323K_HOM_2+M323K_HOM_3+M323K_WT_3+M323K_WT_4	7	0	0
M323K_HOM_1+M323K_HOM_4+M323K_WT_1+M323K_WT_2	7	0	0
M323K_HOM_3+M323K_HOM_4+M323K_WT_2+M323K_WT_3	6	0	0
M323K_HOM_1+M323K_HOM_2+M323K_WT_1+M323K_WT_4	6	0	0
M323K_HOM_2+M323K_HOM_3+M323K_WT_1+M323K_WT_2	4	0	0
M323K_HOM_1+M323K_HOM_4+M323K_WT_3+M323K_WT_4	4	0	0
M323K_HOM_1+M323K_HOM_2+M323K_WT_2+M323K_WT_4	4	0	1
M323K_HOM_2+M323K_HOM_3+M323K_WT_2+M323K_WT_4	3	0	0
M323K_HOM_3+M323K_HOM_4+M323K_WT_1+M323K_WT_4	21	0	1
M323K_HOM_2+M323K_HOM_4+M323K_WT_2+M323K_WT_4	18	0	0
M323K_HOM_1+M323K_HOM_3+M323K_WT_3+M323K_WT_4	16	0	0
M323K_HOM_2+M323K_HOM_4+M323K_WT_1+M323K_WT_2	16	0	0
M323K_HOM_1+M323K_HOM_3+M323K_WT_2+M323K_WT_4	14	0	1
M323K_HOM_2+M323K_HOM_4+M323K_WT_1+M323K_WT_4	2	0	0

Table 8.3: Information on human fibroblast lines used. B, bulbar; UL, upper limb; LL, lower limb

Fibroblast line	Mutation	Diagnosis	Age at onset	Site of onset	Gender	Age at biopsy
TARDBP 1	G298S	ALS	62	LL	M	64
TARDBP 2	A382T	ALS	59	UL	F	62
TARDBP 3	A382T	ALS	25	LL	F	31
TARDBP 4	A382T	ALS	67	B	M	69
CTRL 1	.	Healthy	.	.	F	67
CTRL 2	.	Healthy	.	.	M	64
CTRL 3	.	Healthy	.	.	M	67
CTRL 4	.	Healthy	.	.	F	69

Table 8.4: List of skiptic exons found in LCDmut adult brain

chr	start	end	Gene	PSI_{WT}	PSI_{LCDmut}	ΔPSI	fold change	q
chr2	144502918	144503031	*Dzank1*	0.9795	0.8160	−0.164	8.9942	1.57E−45
chr7	56157784	56163742	*Herc2*	0.9983	0.7828	−0.216	128.6068	4.71E−42
chr5	29597184	29600641	*Ube3c*	0.9894	0.7449	−0.244	24.1694	1.62E−40
chr10	78284350	78287847	*Agpat3*	0.9964	0.9011	−0.095	27.1469	1.73E−32
chr5	36947237	36952351	*Ppp2r2c*	0.9993	0.9359	−0.063	94.4229	1.71E−26
chr1	118304691	118304770	*Tsn*	0.9978	0.9012	−0.096	44.3227	7.36E−23
chr6	113492124	113492258	*Creld1*	1.0000	0.8747	−0.125	>150	1.05E−22
chr12	50365712	50383400	*Prkd1*	0.9804	0.4517	−0.529	28.0231	4.99E−18
chr4	19618357	19621928	*Wwp1*	0.9984	0.9111	−0.087	56.7242	2.13E−17
chr4	58817550	58820128	*AI314180*	0.9981	0.9037	−0.094	51.0528	1.49E−16
chr19	38265097	38283970	*Lgi1*	0.9847	0.8165	−0.168	11.9947	1.09E−14
chr11	101317623	101317727	*Psme3*	0.9632	0.8737	−0.090	3.4328	2.61E−14
chr7	56184707	56185835	*Herc2*	0.9881	0.8836	−0.104	9.8083	1.25E−13
chr5	108642696	108642779	*Tmem175*	0.9514	0.7273	−0.224	5.6106	6.73E−12
chr11	45852192	45884332	*Clint1*	0.9962	0.8842	−0.112	30.5766	1.85E−10
chr10	7710647	7712501	*Lats1*	0.9959	0.9221	−0.074	19.0151	5.18E−9
chr11	97242047	97244428	*Npepps*	0.9954	0.9425	−0.053	12.6145	5.99E−9
chr11	100440040	100440086	*Nt5c3b*	0.9709	0.8464	−0.125	5.2820	1.21E−7
chr17	22428565	22446797	*Zfp946*	0.9530	0.6858	−0.267	6.6817	5.79E−7
chr2	130713397	130713503	*4930402H24Rik*	0.9908	0.9386	−0.052	6.6376	7.23E−7
chr13	13638376	13641067	*Lyst*	0.9909	0.8287	−0.162	18.8410	8.97E−7
chr6	30439865	30444427	*Klhdc10*	0.9721	0.8957	−0.076	3.7346	1.97E−6
chr19	40340146	40344356	*Sorbs1*	0.9877	0.8305	−0.157	13.7569	2.29E−6
chr5	8143132	8145553	*Adam22*	0.9881	0.8977	−0.090	8.5959	4.96E−6
chr14	18630431	18888266	*Ube2e2*	0.9542	0.8915	−0.063	2.3700	8.43E−6
chr7	126153199	26153279	*Xpo6*	0.9657	0.8754	−0.090	3.6328	2.37E−5
chr8	105194475	105194591	*Cbfb*	0.9736	0.8752	−0.098	4.7209	4.06E−5
chr6	8216470	8224555	*Mios*	0.9674	0.9026	−0.065	2.9852	0.000146
chr5	9131308	9136066	*Dmtf1*	0.9683	0.8230	−0.145	5.5848	0.00015
chr16	17845135	17856752	*Dgcr2*	0.9879	0.9334	−0.054	5.4902	0.000182
chr16	94408716	94410833	*Ttc3*	0.9925	0.8888	−0.104	14.8993	0.000244
chr4	59592318	59594396	*Hsdl2*	0.9801	0.9262	−0.054	3.7039	0.000247
chr2	69741693	69746028	*Ppig*	0.9725	0.9024	−0.070	3.5489	0.00025
chr7	141526432	141526493	*Chid1*	0.9876	0.9272	−0.060	5.8899	0.000367
chr11	79242156	79244463	*Wsb1*	0.9588	0.8991	−0.060	2.4516	0.000532
chr1	119706155	119706313	*Ptpn4*	0.9793	0.8764	−0.103	5.9574	0.000597
chr10	83737933	83758134	*1500009L16Rik*	0.9938	0.8995	−0.094	16.2830	0.000837
chr8	75051657	75052150	*Tom1*	0.9947	0.9301	−0.065	13.2732	0.000883
chr11	58848702	58856061	*Gm12258*	1.0000	0.8671	−0.133	>150	0.000917
chr9	55858525	55859748	*Scaper*	1.0000	0.8226	−0.177	>150	0.001059
chr17	34905456	34905910	*Ehmt2*	0.9718	0.8790	−0.093	4.2930	0.001613
chr14	56957103	56959715	*Zmym2*	0.9958	0.9412	−0.055	14.0286	0.001684
chr5	110396045	110396074	*Fbrsl1*	0.9800	0.7715	−0.208	11.4248	0.002646
chr5	3610288	3615075	*Pex1*	1.0000	0.8614	−0.139	>150	0.004794
chr9	54477528	54501351	*Dmxl2*	0.9831	0.9073	−0.076	5.4917	0.005274
chr17	75540191	75544748	*Fam98a*	0.9778	0.8999	−0.078	4.5140	0.0057
chr11	96773222	96776968	*Snx11*	0.9873	0.9256	−0.062	5.8572	0.008094

8.4 Appendices to chapter 6

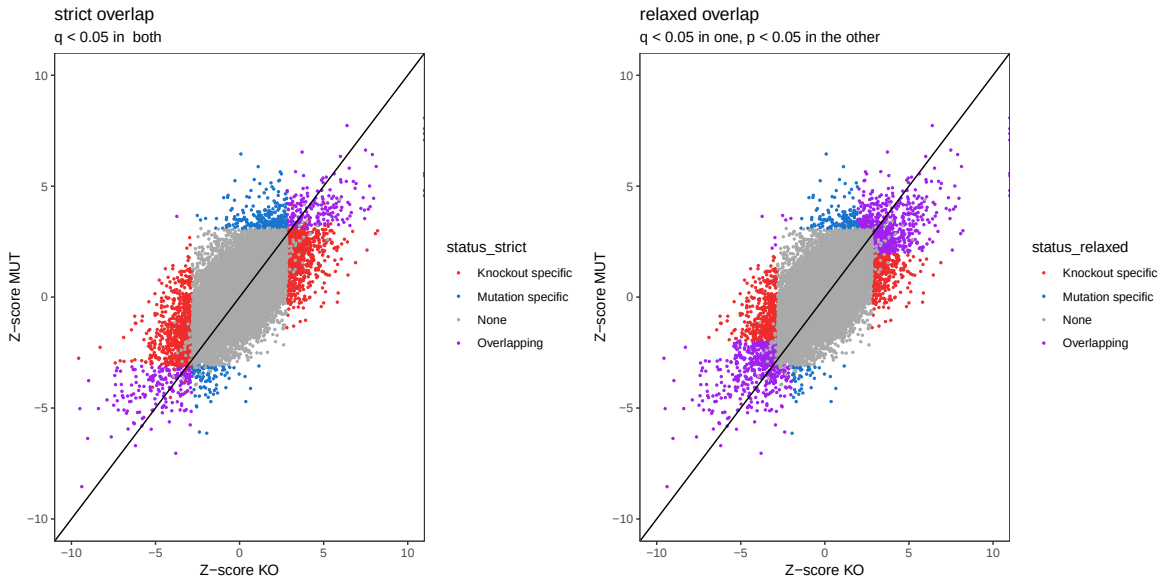


Figure 8.3: Strict and relaxed overlaps of differential expressed genes between FUS KO and FUS MUT joint models. Each gene plotted as a signed Z-score comparisons constructed from the raw p-value and the sign of the estimated log_2 fold change.

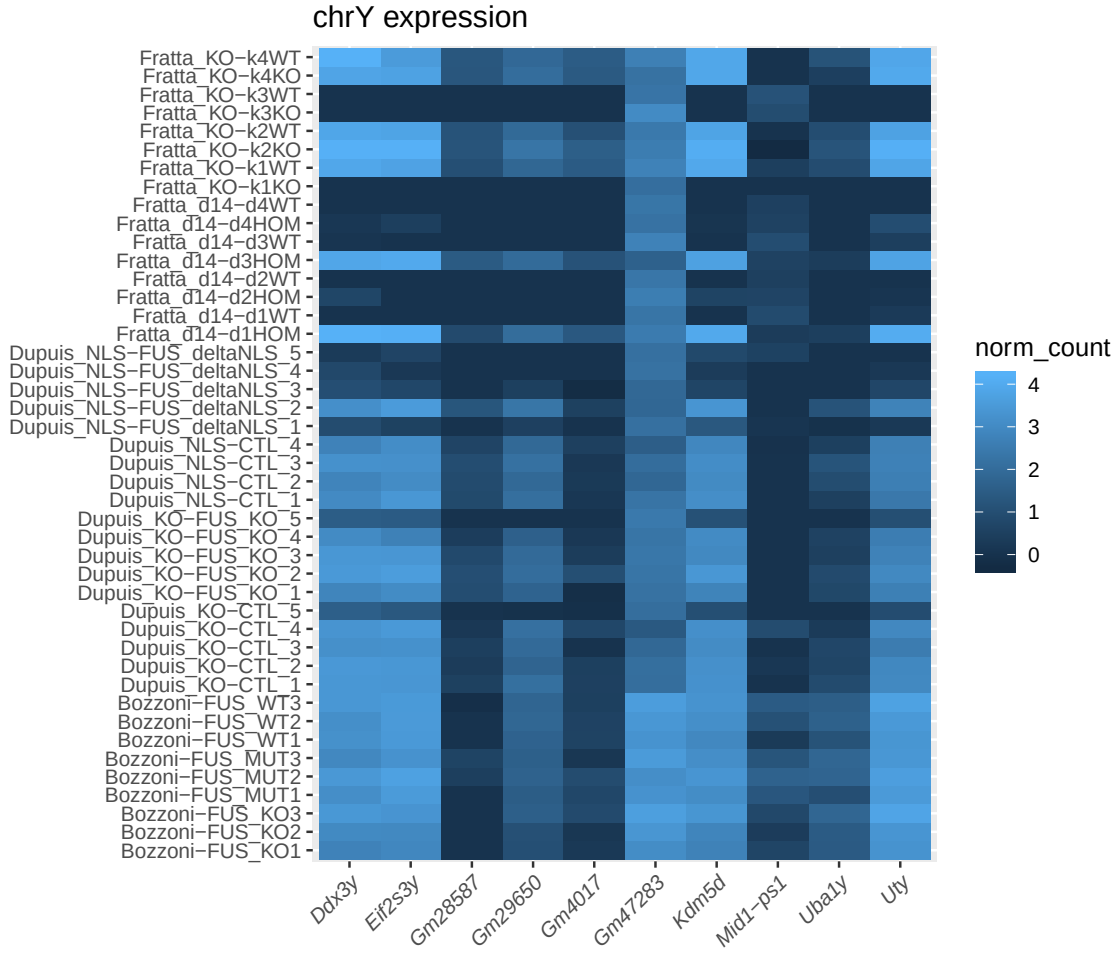


Figure 8.4: chrY expression for each sample allows for sex to be imputed. The top 10 most highly expressed mouse Y chromosome genes are plotted for each sample with a library-size normalised read count. One gene, Gm47283 shows expression in all samples due to it having a paralogue on the X chromosome. The other 9 genes clearly separate samples in to two groups. All Bozzoni samples appear to be male, whereas Dupuis and Fratta samples are mixed.

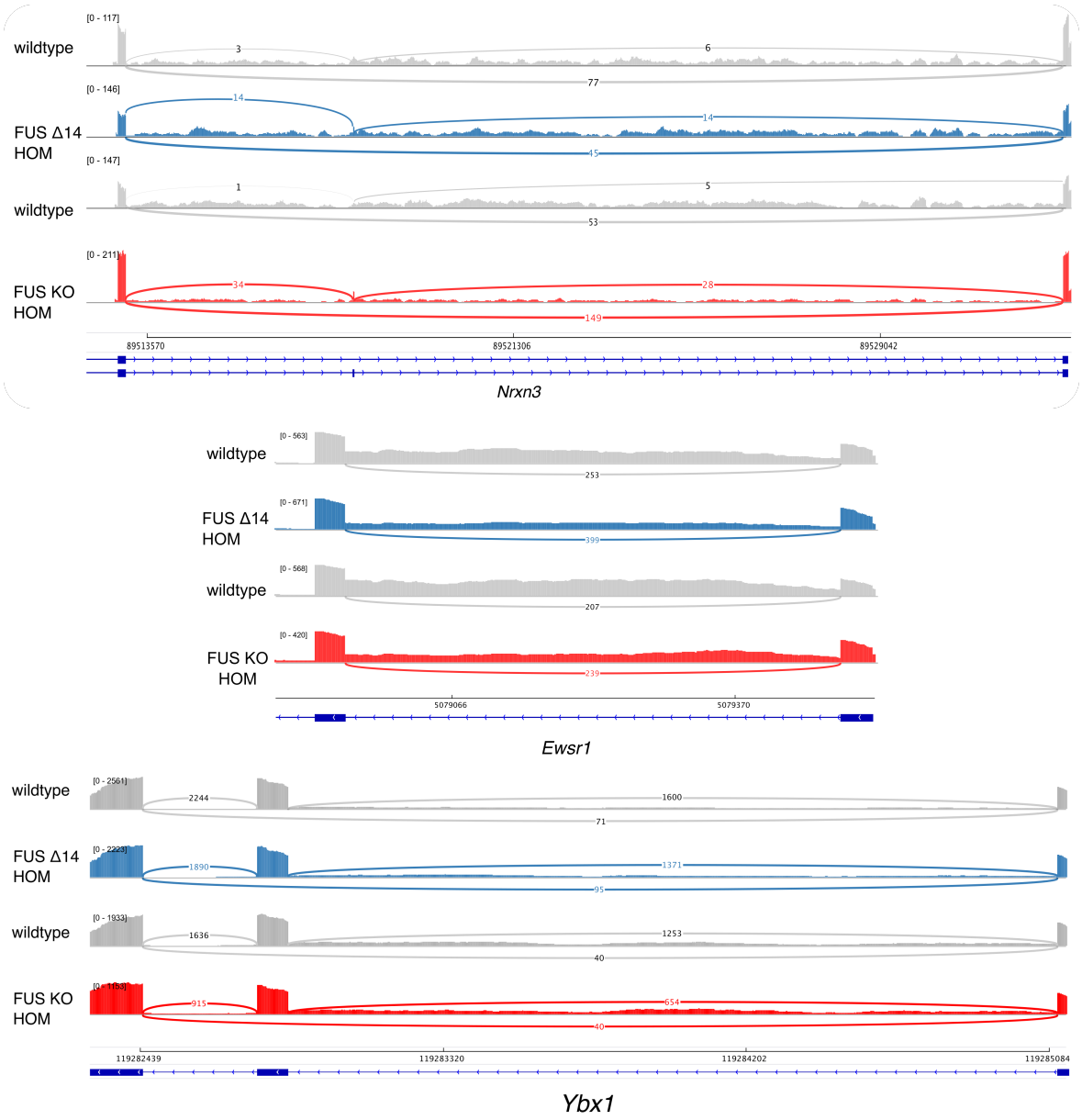


Figure 8.5: RNA-seq traces from three splicing events found in both FUS KO and FUS MUT Sashimi plots from IGV show read coverage and splice junction counts for a cassette exon in *Nrxn3*, a retained intron in *Ewsr1* and a complex splicing event in *Ybx1*. A homozygous FUSΔ14 sample is shown in blue with a littermate control in grey. A homozygous FUS knockout sample is shown in red with its respective littermate control above in grey.

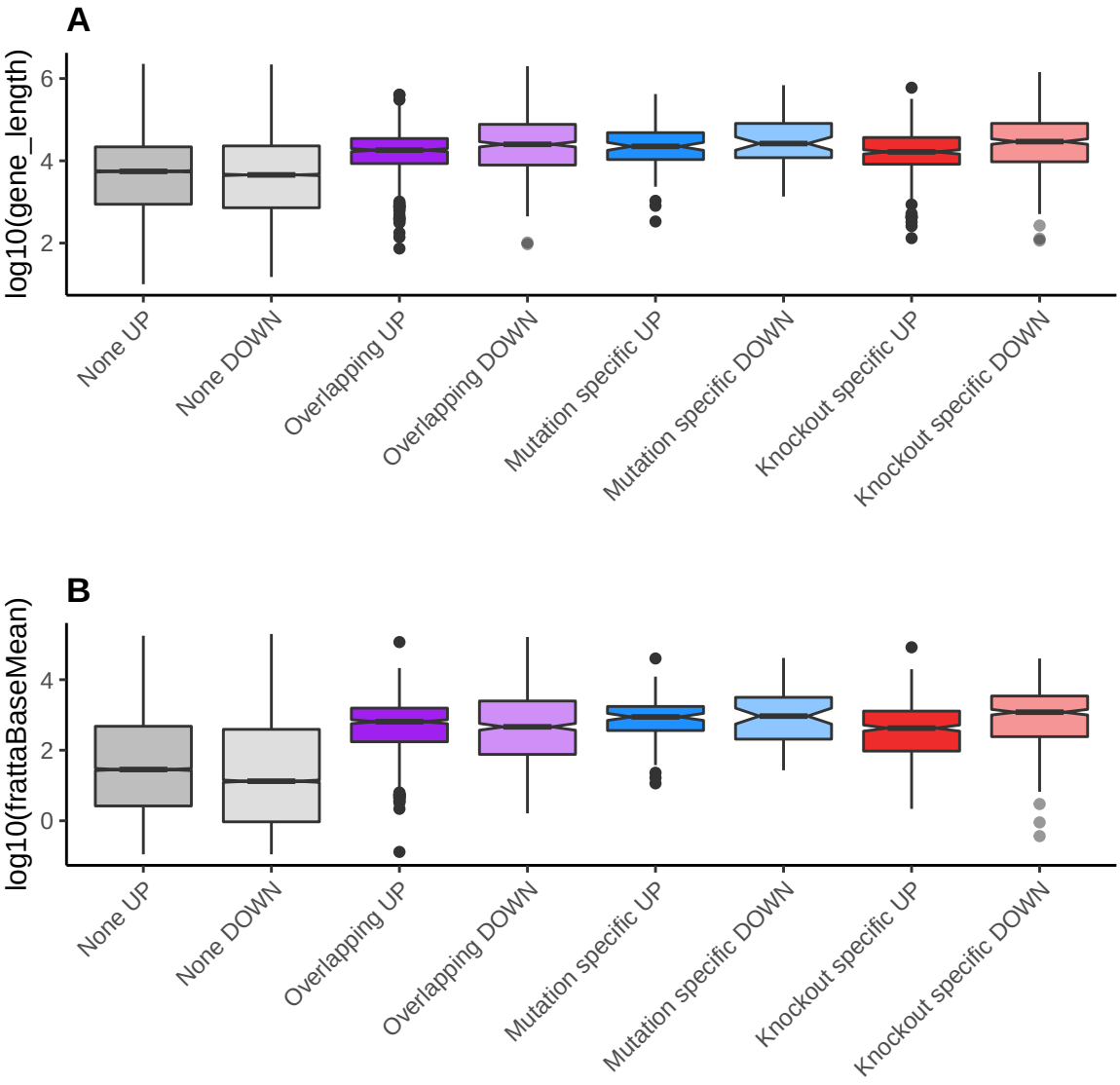


Figure 8.6: Characteristics of differentially expressed genes (A) Sets of non-significant (None) and significant genes grouped by direction of expression. Boxplots show distribution of the log_10(length) in base pairs of the major transcript according to Ensembl. **(B)** Boxplots show distribution of the average gene expression in normalized read counts from the wildtype Fratta samples. Notches denote the 95% confidence interval around the median.

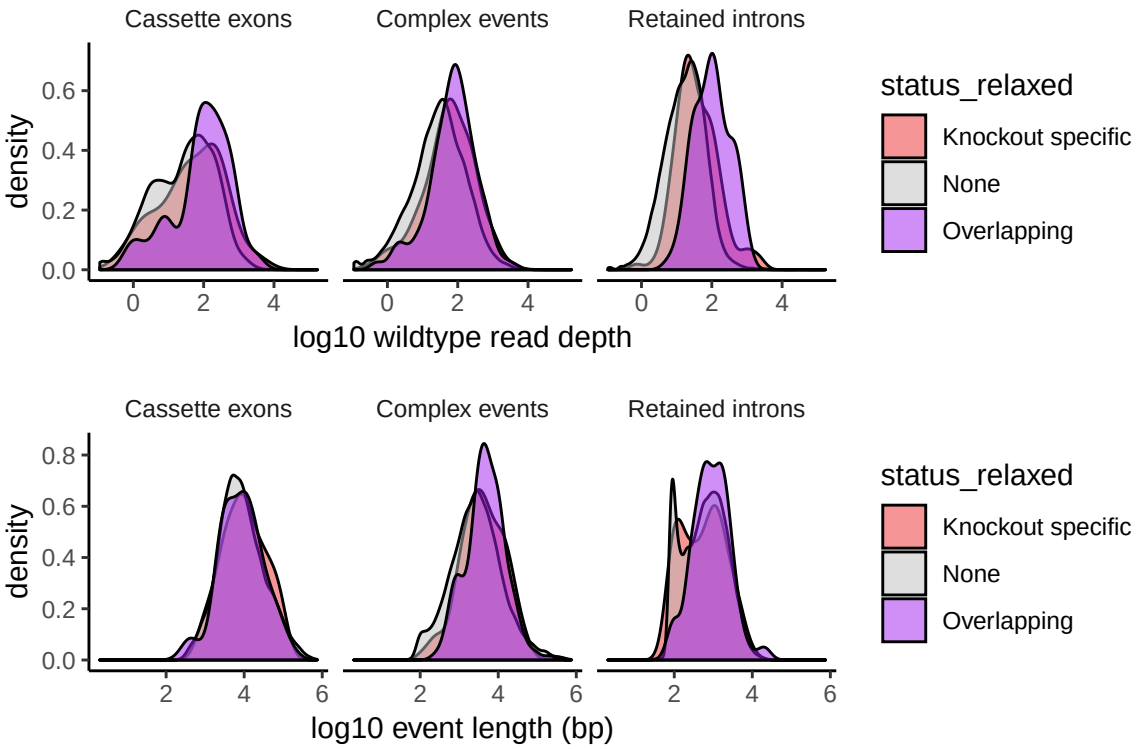


Figure 8.7: Characteristics of differential splicing events Splicing events separated by category, with density plots showing the normalized distribution of read depths (upper panels) or event length (lower panels). Separate distributions are plotted for Overlapping events (purple), Knockout-specific events (red) and non-significant events (grey).

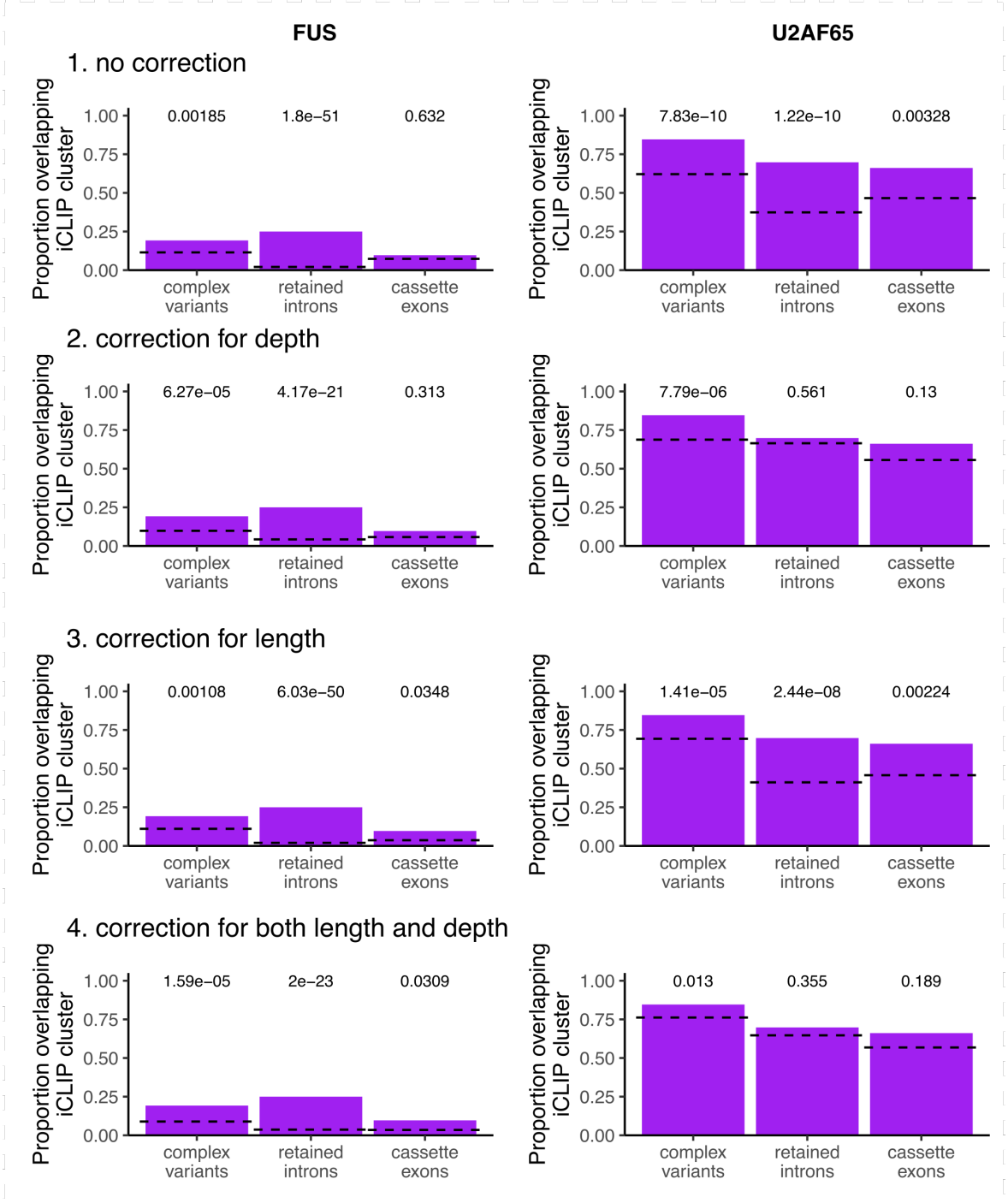


Figure 8.8: Selecting an appropriate background or null set of splicing events Four analyses of FUS and U2AF65 iCLIP clusters overlapping sets of significant and null sets of splicing events. 1. Comparison with all non-significant events shows strong enrichment of U2AF65 clusters in the significant splicing events. 2. Null events have matching depth distribution. 3. Null events have matching length distribution. 4. Null events match significant events for both length and depth.

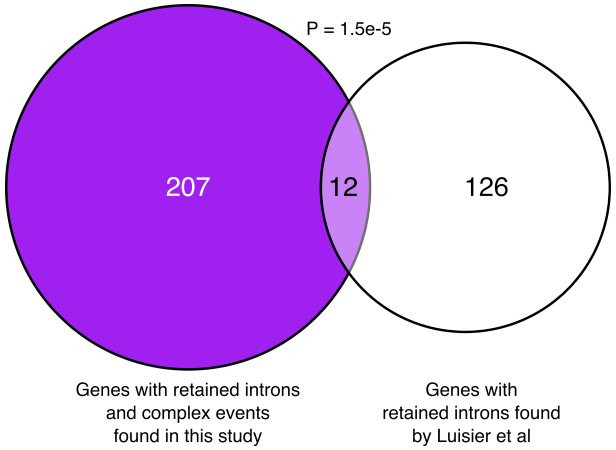


Figure 8.9: FUS-regulated splicing events overlap with those seen in ALS mutant motor neurons Genes with retained introns found in Luisier *et al* on a study of differentiating human motor neurons comparing the effect of the ALS-associated mutation in VCP, overlapped with genes found to have either retained introns or complex events in both the FUS knockout and FUS mutation joint splicing models. P = 1.5e-5, Fisher's exact test. Background set used: 11,251 genes present in SGSeq splicing analysis.

Table 8.5: Splicing events that overlap with Luisier et al

gene name	coords (mm10)	event length/bp	variant type	nearest FUS iCLIP cluster/kb	median phyloP
Atp13a3	chr16:30357274-30361420	4146	complex	0.54	0.234
Ccdc88a	chr11:29494093-29499335	5242	complex	66	0.439
Cdc16	chr8:13767587-13768561	974	retained intron	158	0.244
Fbxl5	chr5:43759825-43760707	882	retained intron	646	0.002
Fus	chr7:127972770-127974400	1630	retained intron	0	1.206
Hnrnpdl	chr5:100036195-100036481	286	retained intron	0.31	0.061
Mfn1	chr3:32562893-32563012	119	retained intron	0	0.255
Ncor1	chr11:62401267-62403799	2532	complex	2.80	0.878
Papola	chr12:105829277-105834710	5433	complex	7.32	0.234
Rbm6	chr9:107833507-107838835	5328	retained intron	13.8	0.150
Srsf5	chr12:80947865-80948129	264	retained intron	0	1.712
Tcerg1	chr18:42550108-42551110	1002	retained intron	25.5	0.428