University College London

Analysis of degeneration and dysregulation in neurodegenerative diseases

Doctoral Thesis

Author

Seth Jarvis

Supervisors

Dr Vincent Plagnol

Dr Pietro Fratta

Professor Adrian Isaacs

Dr Maria Secrier

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.

Seth Jarvis

Abstract

Neurodegenerative diseases are a group of diseases characterised by progressive loss of structure and function of neurons. Conditions under this umbrella tend to have two main effects on patients, degeneration of motor function, and degeneration of cognitive ability. Two diseases which are closely linked are Amyotrophic Lateral Sclerosis (ALS) which has primarily motor effects, and frontotemporal dementia (FTD) which has primarily cognitive effects. Despite their different presentations, both have been linked to dysregulation in RNA binding proteins (RBPs). All of my work has sought to add to the existing body of knowledge around how neurodegenerative diseases act and progress.

The main body of my work has sought to use RNA sequencing data to analyse data relating to neurodegenerative diseases. 3 of the studies are relating to the effects on RNA expression and splicing of mutations in *C9orf72*, *TAU*, *FUS*, and *TARDBP*. These chapters aim to further elucidate how these genes function, either through analysing RNA expression in novel mouse models, or by comparing RNA expression data from samples in the human brain biobank to relevant controls. In them I find several promising candidates for further investigation with regards to the changes which result from mutations in my genes of interest.

One other chapter uses RNA-sequencing data, and aims to compare data from a total RNA-seq kit, and a kit produced by Lexogen which aims to be able to provide similar information at a lower read depth as total RNA-seq data. While this was more a technical chapter, the samples used were from FUS mutant mice and some of the results of analysis has been published elsewhere. The final chapter involved creating a tool to analyse degeneration of neuromuscular junctions using data from a tool created by a colleague.

Overall, my PhD thesis aims to move the field of research into neurodegenerative diseases forward through a combination of improving our knowledge of diseases, improving our knowledge of the tools we are using, and creating tools for use by future researchers.

Impact Statement

The work of my project is a collaboration between the UCL institute of Neurology, Dementia Research Institute and the UCL Genetics Institute. It has used sequenced data from both mouse models and human brain samples.

4 Published papers have involved my work - "FUS ALS-causative mutations impair FUS autoregulation and splicing factor networks through intron retention" (Humphrey et al., 2020), "FUS-ALS mutants alter FMRP phase separation equilibrium and impair protein translation" (Birsa et al., 2021), "NMJ-Analyser identifies subtle early changes in mouse models of neuromuscular disease" (Mejia Maza et al., 2021), and "A comparison of low read depth QuantSeq 3' sequencing to Total RNA-Seq in FUS mutant mice" (Jarvis et al., 2020). Of these, I am first author on 1 (Jarvis et al., 2020) and co first-author on another (Mejia Maza et al., 2021). There have been a total of 52 citations across the papers.

I analysed RNA-sequencing data from samples. One chapter I have worked on evaluated the differences between two methods of RNA analysis. The comparison was between total RNA-seq, broadly considered the gold standard, and low read depth QuantSeq. My work suggested possible roles for the relatively inexpensive QuantSeq and ways to validate the results. This work has potential to be valuable to many labs doing RNA expression analysis regardless of the focus of their research.

The tool I created with my colleague Alan Mejia Maza to automatically classify neuromuscular junctions as degenerating or healthy may also have a wide use among other groups. Once it has been further refined it may be a key tool in the toolbox of those studying neurodegeneration as it can substantially reduce the time spent manually classifying neuromuscular junctions which is a slow part of existing neurodegeneration research. It will also serve to reduce the level of subjectivity in results making them more easily comparable between labs (this is reliant on ensuring that any data that this model and any future model is trained on is robust).

My analysis of RNA-seq data in neurodegeneration, either through models or human brain samples has revealed possible mechanisms through which some of the mutations of interest may act. This helps our understanding of both ALS and FTD, and may help develop treatments for both conditions.

Indices

Index of Contents

Index of Figures

Index of Tables

Acknowledgements

While writing my thesis, there have been many places from which I received support.

I would like to thank my supervisors: Vincent Plagnol, Pietro Fratta, Adrian Isaacs, and Maria Secrier. I would like to give special thanks to Pietro who has been a consistent supportive presence throughout and has provided insight and understanding throughout, and Maria who has read more drafts of my thesis than I care to admit.

Jack Humphrey showed me some of the ropes of RNA sequencing analysis when I started. Jack Humphrey, Kitty Lo, Vincent Plagnol, and Warren Emmet wrote the RNA-sequencing alignment pipeline which I used for my research.

Nicol Birsa and Agnieszka Ule extracted and organised the sequencing of the samples which I used for my comparison of methods, analysis of post-mortem brain tissue, and analysis of F210I mouse mutants. Carmelo Milioto bred, extracted, and sequenced the mice used for my analysis of C9orf72 mutant mice. Alan Mejia Maza was responsible for creating the confocal microscopy images, the algorithm for extracting features from said images, and provided the data used for my machine learning. They also all provided me with various parts of biological context which I was missing.

MegaCrit, the creators of the game Slay the Spire should be equal parts thanked for the endless hours of sanity provided by play, and cursed because of how much it disrupted my work at some points. Even when not playing, I was often thinking about playing.

Personally, I would like to provide my mother Sarah Jarvis and step father Tim Eason for providing support and a place to live throughout my PhD. I also need to thank James Armour for providing much needed support during a very hard time.

Finally, my wonderful, supportive husband Joshua Green, and our dog Lexi. You both mean the world to me and I couldn’t have done this without either of you.

1. Introduction

1.1. The Spectrum of ALS and FTD

Amyotrophic lateral Sclerosis (ALS) is a progressive neurodegenerative disorder. It primarily causes degeneration of motor neurons and affects between 1 and 5 people per 100,000 (Logroscino et al., 2010; Chiò et al., 2013). It results in gradual loss of control in muscles and limbs, eventually leading to inability to speak or swallow. Death most commonly results from infection due to the inability to swallow and tends to be within 3 years of onset.

Frontotemporal dementia (FTD) is another progressive neurodegenerative disorder. It causes degeneration of the frontal and temporal lobes leading to worsening of either behavioural inhibition, or language production and comprehension. It affects 15-22 people per 100,000(Onyike and Diehl-Schmid, 2013) and is the second most common form of early onset dementia after Alzheimer's disease(Ratnavalli et al., 2002). FTD is not directly life threatening. It is, however, linked with an increase in risk of death(Loi et al., 2022). It increases the risk of infections often through dysphagia, as well as the risk of falls. It may make patients unable to care for themselves, and more likely to engage in impulsivity, particularly in the behavioural variant(Rascovsky et al., 2011).

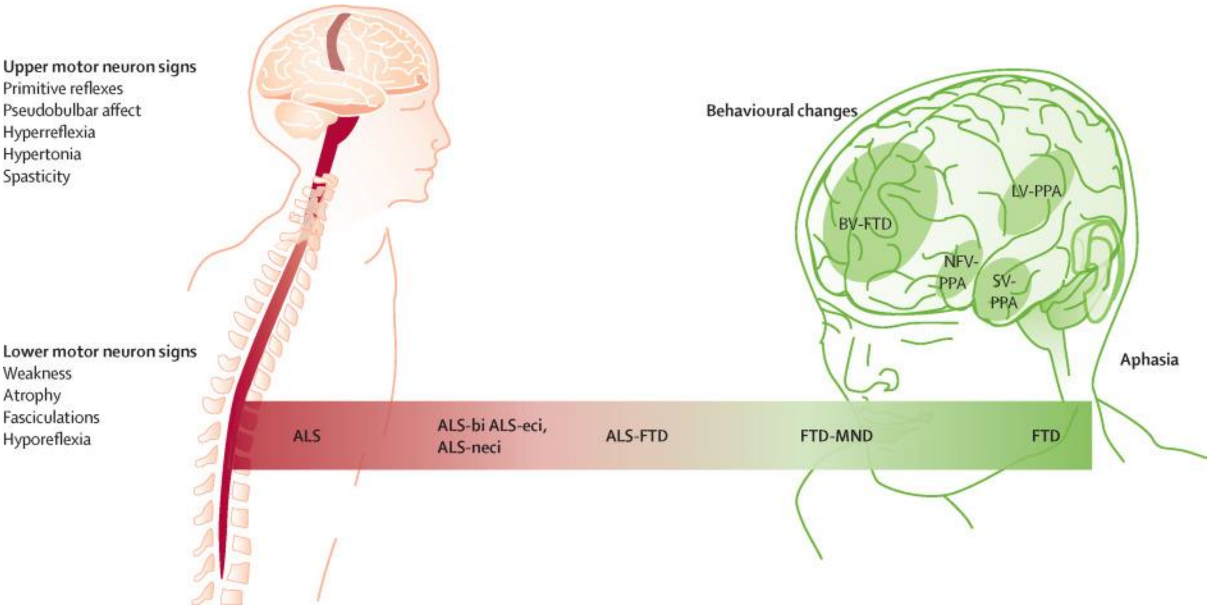

Figure 1-1.Amyotrophic lateral sclerosis and frontotemporal dementia—extremes on the phenotypic spectrum of a single disease from (van Es et al., 2017)

Both disorders peak in incidence around age 60, and are incurable. ALS and FTD are part of a spectrum of disease ALS/FTD. There is an overlap of both the changes observed in the brain and some of the phenotypes. Some ALS patients exhibit cognitive decline, and some FTD patients may start to exhibit motor deficits. The majority of cases in both diseases are sporadic, meaning they have no significant family history. About 10% of ALS cases (Kurland and Mulder, 1955; Taylor et al., 2016) and 1/3 FTD cases (Woollacott and Rohrer, 2016) are familial. Familial cases, while less common, are of particular interest to study as they can provide more obvious insights into the mechanisms of both conditions.

Both disorders have recognisable brain pathology post mortem characterised by aggregated protein inclusions in the regions of the brain most affected. Inclusions of TAU (encoded by the gene *MAPT* on chromosome 17) are found in around 45% of FTD patients (Rademakers et al., 2004). The majority of the remaining FTD patients have inclusions made up of either TAR DNA-binding protein 43kDA (TDP-43) (Neumann et al., 2006) encoded by the gene *TARDBP*, or fused in sarcoma (FUS) (Neumann et al., 2009) encoded by the gene FUS. In ALS, some patients have FUS protein inclusions (Kwiatkowski et al., 2009; Vance et al., 2009), and the 97% have TDP-43 inclusions (Neumann et al., 2006; Scotter et al., 2015). The overlap in these protein inclusions further demonstrates the link between the two diseases. While inclusions of these proteins are relatively common, mutations are rare.

Another gene which is important in both FTD and ALS is *C9orf72*. *C9orf72* was identified as a gene of interest as many familial cases of FTD seemed to have changes in a particular region of chromosome 9. Specifically, it was found that an intron of C9orf72 had a large expansion within it (Renton et al., 2011). Most of the data available is from Caucasians, where about 7% of sporadic and 39.3% of familial ALS patients have mutations in *C9orf72* present. Similarly, mutations in *C9orf72* are present in 6% of sporadic, and 24.8% of familial cases of FTD in Caucasians of north American or European descent. While the evidence from non-Caucasian patients were relatively sparse, some sporadic black (4.1%) and Hispanic (8.3%) individuals had *C9orf72* mutations but no Native American, Asian, or Pacific Islander patients had the mutation. The initial event was therefore dated estimated to be approximately 1500 years ago (Majounie et al., 2012).

Aside from *FUS*, *TARDBP*, and *C9orf72*, there are approximately 20 other genes which have been linked to ALS (Nguyen et al., 2018). The first identified genetic cause of ALS was *SOD1* (Rosen et al., 1993), it produces the enzyme Cu-ZN superoxidase dismutase. It is associated with somewhere in the range of 10-20% of familial ALS, and 0.5-2% of sporadic ALS (National Institute of Health; Chiò et al., 2008). Other genes of interest include *UBQLN2*, *TBK1*, *SQSTM1*, *OPTN*, *NEFH*, and *SETX*.

MAPT, and *C9orf72 are some* of the genes most strongly linked to FTD. While TDP-43 pathology is common, mutations in *TARDBP* occur less often than mutations in GRN, a gene on the same chromosome as *MAPT* (Kumar-Singh, 2011). Familial *FTD*, *GRN*, *MAPT* and

C9orf72 mutations are found in 60% of cases. Mutations which occur in less than 5% of familial cases include *VCP*, *FUS*, *CHMP2B*, *ITM2B*, *TBK1*, and *TBP* (Olszewska et al., 2016).

Genome Wide Association Studies (GWAS) have been performed comparing patients with the conditions to those without. GWAS studies in ALS have found several new loci to be linked to increased likelihood of development of disease and to different disease progression rate. Whilst some loci are directly linked to genes involved in dominant forms of disease such as C9orf72 and SOD1 (Nicolas et al., 2018; van Rheenen et al., 2016, 2021), other loci have highlighted variants linked to disease progression as in UNC13A (Van Es et al., 2009). A novel method used machine learning to integrate functional genomics into GWAS summaries found a rare mutation within the gene KANK1, identifying it as a novel ALS gene. This has subsequently been validated, with reproduction of mutations in human neurons leading to neurotoxicity and TDP-43 mislocalisation(S. Zhang et al., 2022). GWAS studies of FTD have found novel loci linked to disease, including some associated with C9orf72. They have also found a link immune enrichment suggesting immune dysfunction is linked to development of FTD(Broce et al., 2018; Ferrari et al., 2014; Reus et al., 2021).

1.2. Role of RNA in ALS/FTD

Both TDP-43 and FUS are RNA-binding proteins. Given the regularity with which inclusions of these two proteins appear in patients, it is believed that changes in RNA are, if not necessarily causing disease, then at least substantially increasing the likelihood of disease. In this section I aim to give a brief background of several genes which I evaluate the effects of within my thesis, how they work when functioning normally, and some of what we know about how they cause disease.

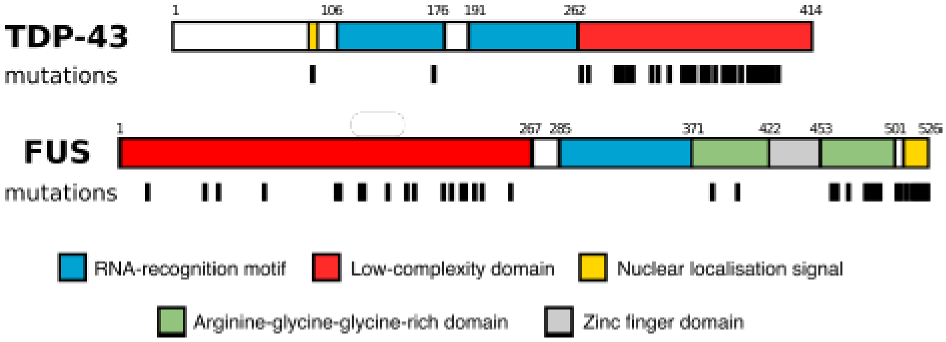

Figure 1-2. Protein domains of TDP-43 and FUS. Structures of the two proteins, coloured by functional domain. Positions of each mutation are represented by black bars. Figure courtesy of Jack Humphrey, and adapted from (Kapeli et al., 2017).

1.2.1 TDP-43

As stated previously, TDP-43 is of interest because it has been linked to most cases of ALS and many of FTD. Cytoplasmic inclusions of TDP-43 are found in both diseases. These inclusions occur regardless in both familial and sporadic cases, and do not seem to require mutations to occur. Cases where there are obvious mutations are useful, as they both allow us to model disease, and understand some of the more common mechanisms by which disease occurs. While pathology of TDP-43 is common, actual mutations seem to be rare. When they do occur, they have been found to cluster around the low-complexity domain (Kapeli et al., 2017), and lead to decreased presence in the nucleus and increased presence in the cytoplasm. There is debate over whether the role TDP-43 plays in neurodegeneration is a result of reduced presence in the nucleus and therefore reduced nuclear function, or a toxic gain of function in the cytoplasm.

TDP-43 is a predominantly nuclear protein but also shuttles to the cytoplasm during normal function (Ayala et al., 2008). TDP-43 plays a role in transcription, splicing, RNA transport, and translation. With regards to splicing, it binds within introns to change behaviour, with the position of binding changing the form which the regulation takes. While this effect was initially discovered in single genes, it has since been validated genome wide through RNA-protein interaction experiments which found some mechanisms through which TDP-43 performs its regulation (Polymenidou et al., 2011; Kapeli et al., 2016; Tollervey et al., 2011). These genome wide studies found that TDP-43 was most heavily associated with 3' untranslated regions and binding in the middle of long introns. Long intron genes are significantly downregulated when TDP-43 is depleted suggesting that this binding has a stabilising role.

In the cytoplasm, TDP-43 binds a set of genes at the 3' untranslated region (Colombrita et al., 2012), and can form RNA granules which are subsequently transported along neurons (Fallini et al., 2012; Alami et al., 2014). It plays a role in translation, both of a small number of target genes and its own gene. It also interacts with proteins that have themselves been linked to translation (Freibaum et al., 2010). Some of the genes it targets are themselves genes linked to translation. TDP-43 can bind to the 3' UTR of *TARDBP*, acting to regulate its own mRNA and thus the production of the protein (Ayala et al., 2011). This has made both knockdown and overexpression models difficult to produce.

Full knock out of *TARDBP* is embryonically lethal, as is being homozygous for certain mutations (Kraemer et al., 2010). Conditional knockout causes gradual degradation of neurons and atrophy of muscles (Iguchi et al., 2013).

The low-complexity domain of TDP-43 has a high quantity of glycine, glutamine, and asparagine. This region is strongly linked to disease: firstly, pathogenic mutations are clustered in this region; further this low complexity domain has important properties that link it to disease pathogenesis. It has been shown to be able to undergo liquid-liquid phase separation to create droplets of cytosolic and nuclear TDP-43. These droplets are likely to play important physiological roles, but when excessive phase separation occurs, it is thought to seed and lead to protein aggregation, the core feature of ALS pathology. It is suggested that it may disrupt nucleocytoplasmic transport, induce clearance of nuclear TDP-43, and cell death. These aggregates appear to be the most common mechanism of TDP-43 toxicity (Baloh, 2011; Gasset-Rosa et al., 2019). This low-complexity domain is also capable of forming many secondary structures, and can have a prion-like effect, causing deterioration in other proteins and making those proteins also capable of inducing misfolding in other proteins(King et al., 2012).

While aggregates can form spontaneously, they are far more likely with mutations in the nuclear localisation signal (Johnson et al., 2009). Some humanised mouse mutations in TDP-43 also caused neurodegeneration when expressed at a standard physiological level but did not cause protein aggregates (Wegorzewska et al., 2009; Barmada et al., 2010). This suggests that while the protein aggregates are toxic, and linked to disease, they are not required for aberrant TDP-43 to cause neurodegeneration. A recent review challenges this notion, and believes that these protein aggregates are key to cause neurodegeneration in patients with ALS (Hergesheimer et al., 2019). Aggregation in the cytoplasm is undeniably a cause of neurodegeneration as it leads to degeneration of cells it occurs in relatively rapidly. Some hypothesise that targeting it may be targeting too late in the process due to the relatively rapid timeframe over which it occurs(Suk & Rousseaux, 2020).

The gain of cytoplasmic function is closely linked to the loss of nuclear function – one does not occur in patients without the other. Both have been isolated and investigated independently. The effects of loss of function have generally been evaluated in models which deplete TDP-43. These models have found that splicing – and particularly cryptic splicing - is impaired both across the whole brain, and in pathways specific to certain cell types(Jeong et al., 2017; Ling et al., 2015; Wu et al., 2019). While stronger evidence is present for the pathogenic effect of cytoplasmic gain of function than loss of function, both play roles in pathogenesis, and neither occur independently in patients.

1.2.2 FUS

FUS is a member of the FET family of RNA-binding proteins. Other proteins in the FET family have been linked to frontotemporal lobe degeneration (the pathological counterpart to FTD), but have not been strongly linked to ALS (MacKenzie and Neumann, 2012). There have been 40 mutations in *FUS* linked to ALS, and it is linked to 5% familial, and 1% sporadic cases (Kwiatkowski et al., 2009). Similarly to TDP-43, *FUS* has a large low-complexity domain. Disease causing mutations are most common in this domain and in the nuclear localisation signal (Shang and Huang, 2016). Familial disease-causing mutations lead to rapid onset, and cause FUS protein aggregates. FUS has also occasionally been linked to FTD pathology, both through mutant *FUS* (Broustal et al., 2010; Van Langenhove et al., 2010), and through protein aggregates in the absence of mutation (Neumann et al., 2009).

FUS acts to regulate splicing and polyadenylation by binding to GUU motifs in introns and 3'UTRs (Ishigaki et al., 2012; Lagier-Tourenne et al., 2012; Rogelj et al., 2012). While the method of action is similar to TDP-43, it is different enough that they share few targets. FUS is presumed to act to stabilise long intron genes, as depletion causes reduced expression in mouse models (Lagier-Tourenne et al., 2012). By binding to RNA polymerase II, FUS also regulates polyadenylation and may have an effect on transcription elongation speed (Schwartz et al., 2012).

There is strong evidence that mislocalisation of FUS is the primary cause of FUS related ALS. The mutations in *FUS* which cause the most severe disease tend to be either in a key proline residue in the NLS, or entirely eliminate the NLS through a frameshift or unexpected stop codon (Chiò et al., 2009; Bosco et al., 2010; DeJesus-Hernandez et al., 2010). Mutations in the NLS are linked to earlier disease onset with a more rapid disease course. While this does suggest that mislocalisation is the primary disease cause, it is still not clear whether it is as a result of loss of function in the nucleus or gain of cytoplasmic function.

Another mechanism of pathogenesis is increased localisation of FUS to the cytoplasm. This is partly linked to the most severe disease course being linked to loss of the nuclear localisation signal (Shang and Huang, 2016), and partly due to the evidence around FUS aggregation. FUS self-aggregates naturally in the cytoplasm (Murray et al., 2017). While mutations in the NLS do not themselves cause FUS to self-aggregate (Sun et al., 2011), they are linked to FUS aggregates: Partly due to increasing localisation to the cytoplasm where FUS is more likely to self-aggregate, and partly because FUS aggregates are normally dispersed through binding of transportin to the NLS (Wang et al., 2018; Yoshizawa et al., 2018). If the NLS is mutated/lost this cannot occur.

FUS aggregates form within the cytoplasm due to the large, low-complexity domain. As is the case with TDP-43, this region is both more likely to bind to other FUS proteins, and has a prion like effect, causing other proteins to misfold (King et al., 2012).

One study seemed to suggest a "two hit hypothesis" for FUS pathology (Vance et al., 2013). When cells with higher levels of cytoplasmic FUS experience oxidative stress, the recruitment of FUS to stress granules may lead to too much FUS being recruited, leading to the formation of FUS aggregates. When there is a large amount of FUS present, it is more likely that it will aggregate as their low-complexity domains are more likely to interact and bind. These aggregates lead to toxicity and neurodegeneration.

Complete knockout of *FUS* is embryonically lethal in an inbred homozygous strain (Barouch et al., 2000; Hicks et al., 2000), but specimens survive until adulthood with no motor deficits at 90 weeks old in the heterozygous strain (Kino et al., 2015). While both *FUS* knockout and *FUS* mutations are lethal when homozygous, only mutant *FUS* causes the degeneration and motor neuron loss associated with disease, confirming the fact that *FUS* gain of function is necessary for neurodegeneration, whilst *FUS* LOF may contribute to this process but is not sufficient (Scekic-Zahirovic et al., 2016).

1.2.3 C9orf72

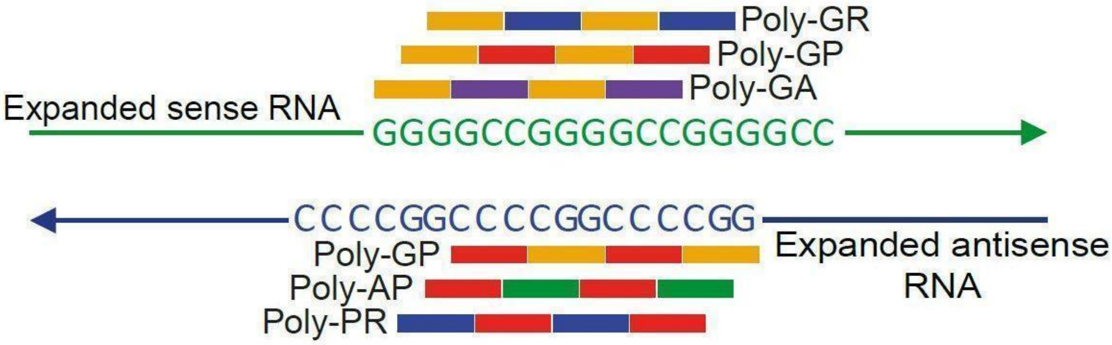

Figure 1-3. Format of C9orf72 and depiction of RAN translation of C9orf72 GGGGCC repeats. Courtesy of the Isaacs lab

Mutations in *C9orf72 (*Chromosome 9, open reading frame 72) highlighted the overlap in pathology between FTD and ALS. As stated previously, it is one of the most common mutations among Caucasian people, accounting for 40% of sporadic cases of familial ALS, and 25% of FTD patients. It was first identified as the cause of disease through several genome wide association studies, which were eventually traced back to a locus on chromosome 9 termed 9p21 (Pearson et al., 2011; Mok et al., 2012). Through deep sequencing, a GGGGCC repeat in the first intron of *C9orf72* was identified as the most common cause (Renton et al., 2011).

Not a huge amount is known about the protein C9orf72. It is produced by the gene of the same name. It is strongly related to the Differentially Expressed in Normal and Neoplasia (DENN) family of proteins (Levine et al., 2013). It is suspected to play a role in membrane traffic in conjunction with Rab-GTPase switches (Levine et al., 2013). Both through implication as these mechanisms are related to autophagy, and direct complexes which have been observed (M. Yang et al., 2016), C9orf72 has been linked to autophagy. When disrupted it can have both positive and negative effects on initiation of autophagy acting to upregulate initiation through loss of ability to regulate MTORC1 signalling, but also disrupt vesicular transport. Disruption is also linked to alterations of autophagosome formation and lysosome function (Beckers et al., 2021). Mutations which cause the repeat expansion seen in disease also seem to induce a DNA damage response and cause dysfunction in nucleolins (Farg et al., 2017) suggesting a role in DNA formation/stabilisation.

C9orf72 also may have cell specific functions. It has a high level of expression in myeloid cells, macrophages, and dendritic cells compared with other immune cell populations(Atanasio et al., 2016; O'Rourke et al., 2016). Induced pluripotent derived microglia with mutant C9 have been shown to have disrupted phagocytic activity and heightened inflammatory response (Lorenzi et al 2022). This reinforces some links which have been found in to play post-mortem mice and patient brains which have suggested that immune cells such as macrophages, microglia, and astrocytes play a role in neurodegeneration (Rostalski et al., 2019; O'Rourke et al., 2016).   ALS Additionally, several novel transcription sites have been found within myeloid cells and certain CNS tissues cell types suggesting tissue-specific function(Rizzu et al., 2016).

In normal function, the first intron of *C9orf72* contains some copy of the GGGGCC repeat (Renton et al., 2011; Fong et al., 2012). Issues occur when there is a substantial expansion of this repeat. A healthy number of copies is generally considered between 6 and 30 (Renton et al., 2011; Van Mossevelde et al., 2017), with pathogenic levels that have been found as low as 30 and as high as >1000 (Almeida et al., 2013).

There are three proposed mechanisms for how expansion of this repeat causes disease (Mizielinska et al., 2013; Gendron et al., 2014; Zhang et al., 2018):

- Reduced expression of *C9orf72*
- Presence of large numbers of RNA foci (aggregates of RNA produced by both sense and antisense G_4C_2 repeats)
- Production of dipeptide repeat (DPR) proteins through repeat-associated non-ATG (RAN) translation.

Reduced expression of *C9orf72* is the most straightforward change. If not enough of the protein C9orf72 is produced, then it will be unable to undergo its normal physiological duties. This may cause neurodegeneration through loss of stability/degeneration of DNA or incorrect trafficking at the membrane.

The Isaacs lab termed large aggregates of repeat RNA RNA-foci. RNA foci had been described and well-characterised in other neurodegenerative diseases such as Myotonic dystrophy, where they were found to sequester an RBP, named Muscleblind, and alter its normal function. The quest for identifying the equivalent for Muscleblind in C9orf72 did not yield clear cut results. These RNA-foci disrupt nucleocytoplasmic transport (Zhang et al., 2015) and other RBPs such as SF2, SC35, and hnRNP-H have been shown to co-localise with RNA-foci. Of these, only hnRNP-H has been shown to directly bind to RNA foci (Lee et al., 2013). It has been directly shown that high levels of certain proteins produced from the DPRs do directly impair RNA metabolism by delaying breakdown of stress granules, and binding ribosomal proteins and others linked to RNA metabolism such as STAU2(Hartmann et al., 2018; Sun et al., 2020; Y. J. Zhang et al., 2018b).

Finally, through RAN translation, the repeats lead to the production of substantial numbers of DPR proteins. These DPR proteins have been shown to cause neurodegeneration in drosophila (Mizielinska et al., 2014) as well as being toxic to cultured cells (Kwon et al., 2014). The two most toxic DPRs appear to be poly-GR and poly-PR. These proteins are more hydrophilic and less prone to aggregation than other DPRs. They are capable of relocating themselves to the nucleus and seem to disrupt ribosomal RNA production(Kwon et al., 2014). They have also been shown to affect phase separation of proteins with low complexity sequence domains, disrupting the function of membrane-less organelles(Lee et al., 2016).

1.2.4 MAPT

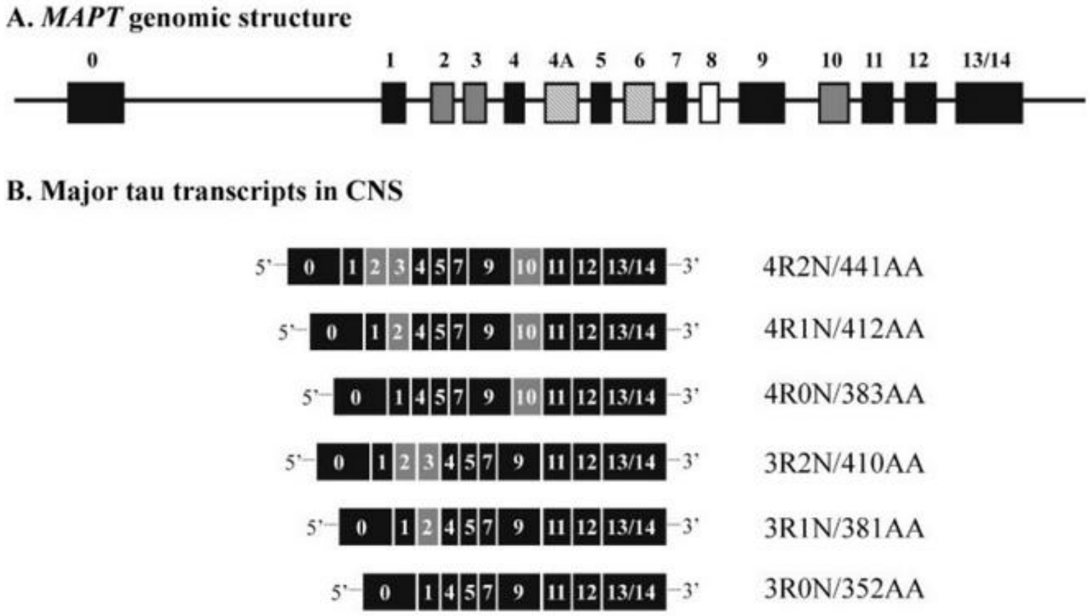

Figure 1-4. A: Schematic presentation of MAPT genomic structure with the 15 exons shown as boxes. B: The six major transcripts resulting from alternative splicing of exons 2, 3, and 10 observed in human brain. A: Filled boxes represent constitutively spliced exons. Alternatively spliced exons are in gray. Hatched exons 4a and 6 are not expressed in the major brain isoforms. Exon 8, shown in white, is not expressed in human MAPT transcripts. Introns and exons are not drawn to scale. B: For each transcript the acronym of the encoded isoform and the number of amino acids (AA) is indicated. Figure from (Rademakers et al., 2004)

Through alternative splicing, the gene Microtubule-Associated Protein Tau (*MAPT)* encodes a group of 6 highly soluble protein isoforms called tau proteins. Tau inclusions are much more common in FTD, and Alzheimer's disease than they are in ALS, but increased levels of phosphorylation of Tau proteins has been found in some ALS patients (Stevens et al., 2019).

Tau proteins are cytoskeletal proteins associated with microtubules. Their primary role seems to be microtubule stabilisation in axons (hence *MAPT*'s name). They perform this through interactions with tubulin, stabilising microtubule formation primarily in the distal region of axons (Cleveland et al., 1977). In addition to their structural roles, tau proteins impair ribosomal function through binding, reducing protein synthesis (Meier et al., 2016; Papanikolopoulou et al., 2019), and play roles in long term memory and certain types of learning (Papanikolopoulou et al., 2019).

The main pathological mechanism of tau is hyperphosphorylation of tau proteins. This results in the accumulation of tau and formation of paired helical filaments (Alonso et al., 2001). These fragments bind into aggregates, these aggregates are one of the primary biomarkers of Alzheimer's disease, initially characterised in 1985 (Brion et al., 1985). The other biomarker is extracellular plaques predominantly made of fibrillar amyloid β (Masters et al., 1985). These aggregates are also found in 40% of FTD patients. It appears that the predominant pathogenic mechanisms of tau are loss of normal tau function, and disruption caused by these aggregates.

1.3. RNA-sequencing

Francis Crick claimed that central dogma of molecular biology is generally stated to be that "DNA makes RNA, and RNA makes protein" (Crick, 1970). This broadly describes the flow of information, from relatively permeant DNA, to RNA and proteins which are generally the less permeant but usually more capable of specific functions. The amount of RNA produced by any particular gene is broadly considered to be a measure of the amount which that gene's level of expression.

RNA Sequencing (RNA-seq) is a form of high throughput sequencing which allows for hypothesis free examination of the transcriptome (the array of all RNA transcripts in a cell) (Wang et al., 2009). This means that knowledge of which genes may be involved in changes of interest are not required and novel changes can be more easily discovered without prior hypothesis. This means it is an incredibly useful method and huge amounts of RNA-seq data are being produced in response to people recognising this utility.

The two most common methods for sequencing RNA sequencing-by-synthesis, and nanopore sequencing. Both methods tend to require the RNA to be separated from the tissue and fragmented. Particular types of RNA such as mRNA can then be extracted if desired. The length of the functional part of these fragments is called the read length, and is determined by the tool being used. In theory, longer reads are better, they can more easily provide more information on the read, including knowledge about the isoform, and splicing. Long-read sequencing is a developing technology however, it is more error prone, and may require different tools to short read sequencing (Heather & Chain 2016; Amarasinghe et al., 2020).

Nanopore works by feeding fragments of RNA through tiny holes (nanopores) in a surface. As each base in the RNA passes through the nanopores, it causes a slight change in ions in the surrounding area, this change is measured and used to sequence samples (Deamer et al., 2016). It has advantages in that it is highly portable, is very good at long read sequencing, and does not require conversion to cDNA - which may introduce artifacts - unlike sequencing-by-synthesis. It will likely become an ever more dominant market force but still has several challenges to contend with. The biggest challenge is that it has a higher error rate than Illumina sequencing (Senol Cali et al., 2018). These errors are randomly distributed and can be more easily fixed during analysis. Since it is a less commonly used technology, some of these tools are less well developed than tools for other methods. Nanopore is also often less well documented, and is more subject to change than similar Illumina data. Given that Nanopore is a relatively young technology, the tools for analysis are still being developed and there are fewer people familiar with pipelines for analysis.

All of the sequencing I have analysed has been performed on some variation of an Illumina dye-based sequencer – a form of sequencing-by-synthesis. This means that most of the comments I make in the remainder of this section are primarily referring to it, although they may be applicable to nanopore based methods as well.

The most common read lengths in Illumina sequencing currently range between 50 and 150bp. In sequencing, RNA is extracted, and then may be filtered to remove types of RNA which may not be desired such as ribosomal RNA, or RNA without poly-A tails. This RNA is then generally fragmented into smaller parts. RNA then undergoes a process called reverse-transcription which converts it to cDNA, a more stable form required for amplification by PCR. The cDNA is then amplified either through bridge or emulsion PCR. This produces large numbers of copies of the transcript (Wang et al., 2009). Generally, only the forward copies of these strands are kept, with the reverse copies being removed. Nucleotides with a particular dye are then integrated into the mixture and used to create complementary copies of the remaining strands. As each new base is added, the strand is excited and the colour of the fluorescence is recorded. This fluorescence is used to identify which base the original strand has at each position (Canard and Sarfati, 1994; Meyer and Kircher, 2010; Clark et al., 2018).

By default, most sequencing methods provide no information on whether the RNA sequenced is from the sense or antisense strand. As genes on the sense and antisense strands may overlap, there have been methods developed (Levin et al., 2010; Mamanova and Turner, 2011) which can attach identifiers to a strand, marking its origin as sense or antisense. These can then be used to improve alignment of the sequencing data, and overall accuracy of analysis.

The data from RNA-seq is hypothesis free. This means that it can be used for exploratory analysis. Even if initial analysis does not provide obvious, useful results, the data is all based on the same basic tool and so, as of right now, all existing RNA-seq data will likely be able to take advantage of newly developed tools for analysis of RNA-seq data. This may be through additional context of discoveries, better optimisation of tools for analysis, or realising an additional role that a subset of the data may play such as splicing and exon usage (Wang et al., 2008; Anders et al., 2012), or the role of non-coding RNAs (Morris and Mattick, 2014).

This is not to say that new methods of sequencing do not hold their advantages. There are somewhere in the region of 100 methods for NGS-sequencing listed on the Illumina website (NGS Library Preparation; Stark et al., 2019) all aiming to provide an advantage over standard RNA-seq. This may be through reduction in cost, increasing ease of use, or through providing specific information on changes in specific types of RNA such as mRNA and ribosomal RNA. Long read sequencing methods are also increasingly being used to answer new questions that have been beyond the reach of short read data.

In addition to analysis of differences between the actual methods of RNA-sequencing data, there are also changes in ways to analyse the data. There are several tools/pipelines for analysis of existing data and comparisons of tools at any particular step is useful for optimisation of existing pipelines or choosing which tools are worth using when developing new ones.

As stated in (Stark et al., 2019), RNA-seq is very much in an “awkward teenage phase”. The glut of new technologies means that there are many potential comparisons to be made between methods. These may be direct comparisons between two methods for sequencing, or between methods for differential expression analysis, or a combination; aiming to evaluate the most effective ways to analyse new datasets. I have performed some comparisons through the course of my PhD, both evaluating methods of sequencing, and trying to optimise analysis of some existing data using new tools which have been developed.

1.3.1 Cell type disambiguation

The human CNS is made up of billions of cells. There are approximately 100bn neurons and it is currently believed, a roughly equal number of glia – a collective term for non-electrical cells, predominantly oligodendrocytes, astrocytes, and microglia(Azevedo et al., 2009; von Bartheld et al., 2016). The specific proportions of cell types within the brain are heterogenous, varying between regions of the brain and between individuals. They may also vary between conditions.

There are 3 main types of neurons present within the brain. Sensory neurons which carry information from sense organs, motor neurons which control muscle activity, and interneurons which connect sensory and motoneurons. Developing neurons are called neuronal projections.

Microglia are immune cells within the brain. In normal function they clear toxic products and dead cells from within the brain (Soulet & Rivest, 2008). They have, also been linked to neurodegenerative disorders such as Alzheimer's Disease. In this case they are harmful to neurons including via mediating synaptic loss and exacerbating tau pathology(Hansen et al., 2018). Astrocytes maintain the environment around neurons, controlling levels of neurotransmitter within synapses, blood flow within the brain, and maintaining general homeostasis (Sofroniew & Vinters, 2010). Finally, Oligodendrocytes predominantly act to produce a myelin sheath around neurons (Bradl & Lassmann, 2010).

One area of interest in my research has been evaluating whether proportions of cell types do vary between conditions within ALS-FTD. There are existing methods for evaluating levels of cell types such as laser capture microdissection (Emmert-Buck et al., 1996) and single-cell RNA-seq (Olsen and Baryawno, 2018). These methods tend to be time or labour intensive, and provide less information on their own than bulk RNA-seq. There are tools which have been developed which use bulk RNA-seq data to generate information on cell types within a set of data from overall tissues such as CibersortX (Newman et al., 2019) and MuSiC (Wang et al., 2019). This process is called cell type deconvolution. I aimed to evaluate the feasibility of using existing methods for cell type deconvolution to evaluate changes in cell types in my tissues. If possible, I also hoped to use this information to improve accuracy of differential expression analysis. Table 1-1 shows several commonly used methods for cell type deconvolution, the broad methods they work, their output, and whether a cell type signature is required or provided.

Deconvolution at its core is a method for calculating values of a source based on an output. It is utilised in various fields, from audio-mixing to evaluating the path which an electrical current takes within the brain. There are two main forms of deconvolution, blind source separation (BSS), and guided blind source separation (gBSS). BSS involves taking raw data, and without knowing the origins, or how the method used to transform them into the final data, calculating the initial input. gBSS knows has some more information on the source data, and uses this in conjunction with the output to separate data (Mohammadi et al., 2017). BSS is a very difficult problem to solve, and requires a substantial amount of tailoring for datasets. gBSS uses known algorithms, combined with a set of data on how the data is mixed to produce results. In the case of RNA based cell type deconvolution, it uses a matrix of data on expression in individual cell types to generate information on those cell types. I deemed that it was not a good use of time to use BSS methods.

When using gBSS, there are two variables with regards to the method: reference dataset, and tool. Several tools had reference datasets for blood or cancer cell deconvolution. There did not seem to be a gold standard reference dataset for deconvolution of brain tissue. When searching for reference datasets, and use of the data, the most common and highest quality reference datasets for deconvolution in RNA-seq data seemed to be for blood samples or cancer tissues. While the tools developed are generalisable, they require a reference dataset. At the time of my analysis, Voineagu lab (Sutton and Voineagu, 2020) were putting substantial time and effort into attempting to develop a gold standard reference dataset. As there was another group working on this problem, who had already encountered several hurdles, developing my own reference dataset was considered beyond the scope of my project. They subsequently decided that developing a reference dataset may not be a worthwhile use of time(Sutton et al., 2022). The existing methods which have integrated brain cell type reference datasets, tend to not be true deconvolution methods, rather giving more relative expression results.

There were some methods (Kelly et al., 2018; Hagenauer et al., 2018; McCoy et al;, 2018, Wang et al., 2020) which are specifically designed with deconvolution of neuronal data in mind. The lack of comparison of relative efficacy meant that when deciding on which method to use, my decision predominantly rested on ease of use, format of output data, and intended use cases. A comparison of existing methods specifically designed for brain cell type deconvolution on known data would be of great utility. In addition, a tool which allows for selection of specific datasets and regions which are already integrated broken down by the factors which (Sutton et al., 2022) would be hugely useful to researchers seeking to perform bulk brain RNA-seq analysis.

Table 1-1. Table of cell type deconvolution methods. From (Sutton et al., 2022)

Algorithm	Class	Signature	Foundation	Output	Citation
DeconRNASeq	Deconvolution	User-specified	Non-negative least squares	Proportions	(Gong & Szustakowski, 2013)
CIBERSORT	Deconvolution	User-specified	Support vector regression	Proportions	(Newman et al., 2015)
Dtangle	Deconvolution	User-specified	Linear mixing model	Proportions	(Hunt et al., 2019)
MuSiC	Deconvolution	User-specified (single-cell only)	Weighted non-negative least squares	Proportions	(X. Wang et al., 2019)
Linseed	Deconvolution	None	Simplex topology	Proportions of unlabelled cell-types	(Zaitsev et al., 2019)
BrainInABlender	Enrichment	In-built (human and mouse brain)	Average scaled expression of marker genes	Enrichment	(Hagenauer et al., 2018)

xCell	Enrichment	In-built (cultured human brain cells)	Gene set enrichment analysis	Enrichment	(Aran et al., 2017)
Coex	Enrichment	None	Weighted gene co-expression network analysis	Enrichment for unlabelled cell-types	(Kelley et al., 2018)

1.4. Aims of my thesis

The goal of my thesis was to bring novel insights into the biology of ALS and FTD through the analysis of RNA-seq data from human-derived samples and mouse models of the disease. Aside from one outlying chapter, this was predominantly through analysing RNA-sequencing data: Either by cataloguing the effects of mutations/pathologies on RNA-expression, or by analysis of the effectiveness of tools for RNA-sequencing analysis.

In the process, I tested and applied various existing RNA-seq data processing and analysis tools, and devised new strategies for statistical analysis, data integration and visualisation. RNA-seq data was used because many of the mutations which have been linked to ALS-FTD are involved in RNA-regulation.

My final thesis is a combination of various analyses. Most were linked to neurodegeneration and RNA-seq analysis. At various points I examine the role effect of mutations in TDP-43, FUS, TAU, and C9orf72 on RNA expression. I have also worked to attempt to extract more information from our RNA-seq data, evaluate methods for analysis, and investigate changes in either makeup or expression in brain cell types. Most of my analysis has been performed on mouse models bred, developed, or created by Nicol Birsa, Agnieszka Ule, or Carmelo Milioto. I have performed some analysis on human brain data which was from the UCL Brain Bank, and trained a machine learning model on data created from a method by Alan Mejia Maza.

1.5. Chapter Abstracts

1.5.1 Comparison of low read depth QuantSeq and RNA-seq – Chapter 3

Transcriptomics is a developing field with multiple new methods of analysis being produced which may hold advantages in price, accuracy, or information; QuantSeq is one such method.

It is a 3’ sequencing method which aims to obtain similar information on differential gene expression with at a lower read depth than standard RNA-seq. It is a method of particular interest to the Fratta lab as it can also be used for information on differential polyadenylation. Changes in polyadenylation have been linked to FUS and therefore we were interested in seeing if QuantSeq was able to provide useful information at much lower read depths. Mouse models of both FUS knockout, and a humanised FUS mutation were sequenced along with littermate controls using both low read depth QuantSeq and total RNA-seq. I then compared the results of the two methods.

1.5.2 Differential Expression in Post Mortem brain tissue in FTD patients – Chapter 4

The Fratta Lab has a large set of RNA-seq data from human brains derived from a mix of healthy, ALS, and FTD patients. We already had a dataset from FTD patients with both C9 and TAU morphology as well as relevant control brains. The goal of my analysis was twofold: first I aimed to examine the changes in expression in our patients when compared to both each other and our control dataset, second, I aimed to see what role differences in cell types might play, and work out how to best integrate this information into differential expression analysis.

1.5.3 Analysis of C9orf72 repeat expansion in mice – Chapter 5

Mutations in C9orf72 are one of the most common causes of both ALS and FTD, specifically, a hexanucleotide expansion in an intron. Carmelo Milioto of the Isaacs lab created mouse models of two proteins created by repeat expansions which have been most strongly linked to disease. Upon receiving the data, my objective was to investigate the changes that these repeats caused over time, and how the two proteins differed from each other.

1.5.4 Analysis of F210I mutant mouse data – Chapter 6

Members of our lab created mice which have a point mutation in one of tardbp’s RNA binding domains. While homozygosity of this mutation is embryonically lethal, heterozygous mice grew to adulthood with no neuropathology. I was provided with RNA-sequencing data from adult heterozygous mice, as well as embryonic data from homozygous mice. My goal was to analyse the heterozygous mouse data which had not been previously analysed and also compare the changes seen to changes in the homozygous mice.

1.5.5 Identification of degenerating neurons using machine learning – Chapter 7

Using confocal microscopy to study neuromuscular junctions (NMJ) is common practice in the study of neurodegeneration. Existing methods are to an extent manual, requiring thresholding and manual judgements of which NMJs are degenerating and which are healthy. My colleague Alan Mejia Maza developed a method which was more automatic, removing the manual thresholding, and which produced more morphological information than other methods. I used the results from his method to create a machine learning model which can automatically classify NMJs into healthy and degenerating based on the output of his method.

2. Methods and RNA-Seq analysis pipeline

2.1. Why we need a pipeline

All but one of my results chapters involves analysis of RNA sequencing data. In order to ensure that this can be done effectively a pipeline was created primarily by other members of the Plagnol lab. I have occasionally personalised some steps or needed to run them manually both due to differing needs between projects, and issues running the pipeline on the UCL Computer Science cluster. An overview of the pipeline can be seen in Figure 2-1.

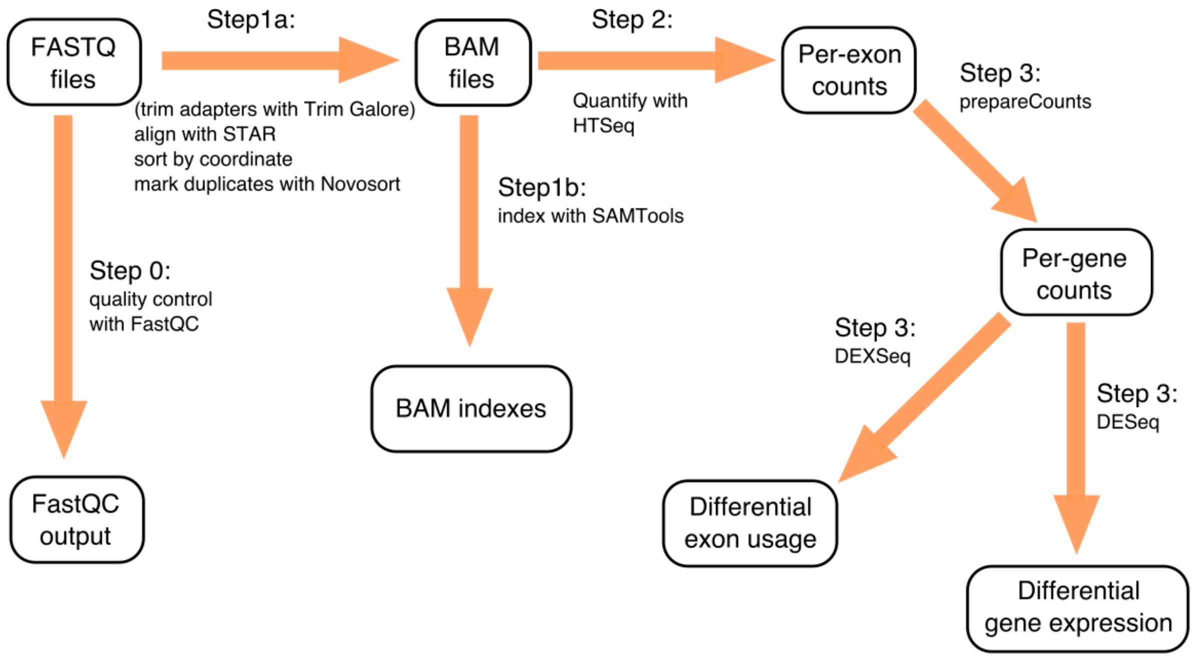

Figure 2-1. Diagram of the ways in which the RNA-seq pipeline processes data. Taken from https://github.com/plagnollab/RNASeq_pipeline/

2.2. Quality control and read alignment

The first two steps in our pipeline are pre-processing steps. They involve running FastQC, then if the data is considered degraded, we recommended that Trim Galore is run. FastQC v0.11.2 (Andrews, 2010) is a tool that produces visual representations of the quality of the reads of each sample. It allows for quick and obvious evaluation of the quality of reads. This can show any issues introduced either during library preparation, or sequencing itself.

Smaller reads may also have adapter sequences which are used during sequencing. High levels of adapter sequences most commonly occur if the original RNA is heavily degraded or there's a flaw in fragmentation. If the proportion of adapter sequences are too high, they may interfere with alignment. We have included another option to run Trim Galore v0.4.5 (Krueger, 2012) which acts to remove both these adapters and low-quality ends of reads, setting the quality cutoff to 20. All other arguments were left as default.

The first main step of the pipeline is properly aligning the reads present in the FASTQ files to the appropriate genome. As well as being aligned to the genome, in RNA sequencing (as opposed to DNA sequencing) information on which sections of the gene have been spliced out, and the location of these splice sites is recorded. This information can be stored as splice junctions and used for further analysis.

As far as we are aware, the current gold standard for RNA-seq alignment which includes splice junctions (ignoring alignment-free comparisons such as Kallisto (Bray et al., 2016)) is Spliced Transcripts Alignment to a Reference henceforth known as STAR, we used v2.4.2a (Dobin et al., 2013), using zcat as the command to read files, and 4 threads, all other arguments were left as their default. It tends to outperform other methods both in terms of accuracy and speed. It tends to be faster on systems with large RAM because it loads the whole genome into memory to allow faster access, and uses the seed-and-extend algorithm which helps to quickly find the best location for splice junctions.

NovoSort V1.03.09 (http://www.novocraft.com) is used to sort the BAM files. The arguments used are: -md (mark duplicates), -xs (secondary alignments do not have duplicates removed), and -f (sorting is forced). The SAMtools V1.2 (Danecek et al., 2021) index function V1.2 is used to index the sorted BAM files using default arguments.

Finally, our pipeline counts the number of reads that are mapped to each exon using HTSeq v0.9.1 (Anders et al., 2015). We use Ensembl transcript annotation as our reference (Cunningham et al., 2015). The functions GenomicArrayOfSets, GFF_Reader, SAM_Reader, BAM_Reader, and pair_SAM_alignments are used where relevant. Arguments are set to default, aside from stranding which is changed depending on the data.

2.3. Differential expression

R version 3.5.1 was used for all differential expression, splicing, and machine learning analysis. It was running within RStudio version 1.3.959.

Differential gene expression is generally the primary rationale for performing an RNA-seq experiment. Experiments tend to be performed by sequencing multiple biological replicates of both the mutant and wildtype. While the cost of sequencing has been rapidly decreasing, it is still often prohibitively expensive to sequence a lot of samples meaning that an algorithm needs to be chosen carefully. We decided to use DESeq2 (Love et al., 2014) because it is faster than other algorithms, and has been designed in such a way that robust statistical inferences can be made from experiments with a small number of replicates. I ran it via the R package EnrichmentBrowser v2.12.1 (Geistlinger et al., 2016). Default Arguments were used aside from setting the argument for which algorithm to use to DESeq2 (aside from the cases where another algorithm was used). DESeq2 V1.26.0 is used to produce normalised reads per counts. This uses a median of ratios to normalise counts described in the paper regarding DESeq (Anders et al., 2010).

The package pheatmap is then used to create heatmaps of expression where relevant. I input the normalised reads produced by DESeq2, generally amongst the most variant genes, and it produces a heatmap. The log of these normalised reads are used to increase readability. All arguments are set to default, and Euclidian distance is used to perform hierarchical clustering.

In most cases, I did not perform differential expression as part of the main pipeline, preferring to run it manually to allow for greater flexibility than using the pipeline. When DESeq2 is run, first the reads are summed then the reads of each gene are normalised for each sample in the library in order to make it easier to compare samples while accounting for library size. It then assumes normalised read counts have a binomial distribution and uses that assumption to fit a negative binomial probability distribution. In order to improve our ability to visualise the data we then shrink the dispersion using the function lfcShrink which improves performance on some datasets and makes the rest easier to visualise (Love, n.d.).

The software then fits two generalised linear models: one with the null hypothesis that there is no difference between expression of a gene across the two conditions, and one for the alternative hypothesis that there is. The results of the two models are then compared using a Wald test to see which fits best. Due to the large number of tests, multiple testing correction needs to be performed. DESeq2 uses the Benjamini-Hochberg (Hochberg, 1995) method to do this.

The final output of DESeq2 has the name and Ensembl ID of each gene, the mean reads across all samples, the fold change of number of reads across the two conditions, as well as the adjusted and unadjusted P-values. P-values are calculated by DESeq2 using a Wald test, and adjusted using Benjamini and Hochberg correction. P-values are a measure of the likelihood that a value would be as extreme or more extreme than the one observed in the analysis purely due to chance. Adjustment is required as when performing a large number of tests, it is likely that a large number of samples would appear to be significant. The adjusted p-value adjusts the p-values to correct for the number of tests which have been performed – in this case the number of genes which have been tested for significance. Unless explicitly stated otherwise, the adjusted p-value of under 0.05 is the significance threshold used.

Gene ontology (GO) analysis is a method used with the goal of reducing the complexity of interpretation of analysis. It does this by producing measures of whether a pathway seems to be significantly enriched/depleted compared with what is expected. The GO project was the first large scale project relating to this. Their goal was to standardise terms used to describe mechanism, and produce a database of genes and the pathways with which they are linked(Consortium, 2004). I performed GO term analysis by finding all genes which were significantly differentially expressed. I then used a combination of topGO v 2.34.0 (Alexa and Rahnenfuhrer, 2020), and pathview v1.22.3 (Luo and Brouwer, 2013) to produce GO term networks based on these genes. To calculate differences, both Fisher's exact test and a Kolmogorov-Smirnov like test are used by topGO. Those networks were then re-created using yEd (yworks, 2019) to allow for easier formatting and text size changing to improve readability. Relative enrichment / depletion was measured against a background gene list of all genes which were expressed within my samples that had not been filtered out due to low expression.

2.4. Differential splicing

Another possible use for RNA-Seq data is the investigation of differential splicing or exon usage. In order to look at differential exon usage, we use the R package DEXSeq v1.5.3 (Anders et al., 2012) which, like DESeq2, fits 2 generalised linear models over a negative binomial distribution and tells us whether it is more likely that the results we see are due to random chance or a real difference between conditions, which causes this change in the reads seen in each exon.

In order to look at differential splicing we use a package called SGSeq 1.4.0 (Goldstein et al., 2016) which normalises the reads of each exon and then finds the type of splicing events that are observed (both annotated and novel), and the number of reads of each event that we see in each sample. We then instruct DEXSeq to treat each splicing event as a gene and each variant of that particular splicing event as an exon. This allows us to identify alternative splicing events.

2.5. Evaluating changes in cell types

Brain tissues are relatively heterogeneous in their makeup. As I am also working with neurodegeneration, I surmised that it might be possible to evaluate whether there are differences in the relative balance or levels of RNA expression of certain cell types in some of our brain data.

In order to investigate the possibility of changes I used the R package BrainInABlender (BIAB) (Hagenauer et al., 2018). This package performs cellular deconvolution on bulk RNA-seq data to provide relative levels of each of a predefined set of brain cells. It takes normalised reads of genes as an input, and further scales them to ensure no one sample exerts too much influence on the balance in other cells. Comparisons are then performed between these normalised, variance stabilised data, and built-in datasets of representative expression of 10 different brain cell types. This comparison is used to work out the relative levels of each cell type in datasets.

Integrating the results of BIAB into our differential expression analysis was also something with which I experimented. I attempted to use formulas which treated each cell type as an individual covariate e.g.

Age + Sex + Cell type 1 + Cell Type 2 + Cell Type 3 + condition

This style of formulas caused errors in DESeq2 which seemed to be due to insufficient differences between our cell types. The format

Age + Sex + Cell type 1:Cell Type 2:Cell type 3:condition + condition

was able to run more consistently. In my final analysis I settled on the approach of only including cell types within the model if there are significant differences in relative cell type balance between cells, and that, when desired, they should be integrated as an interaction term with the condition in DESeq2's formula.

2.6. Machine Learning

Supervised machine learning is a process of providing a dataset to a computer with certain attributes and groups into which they need to be classified. An algorithm then finds the best ways to classify each sample in the training dataset into the groups and tests its accuracy. The most accurate models can then be used to predict the classes of new data (Jiang et al., 2020).

In chapter 7, models were trained on data created by my colleague Alan Mejia Maza's imagej plugin NMJ Analyser (Mejia Maza et al., 2021). The models were created using the R package Caret V 6.0-78 (Kuhn 2008). Two models were created, one which used the raw data, and one which equalised the number of denervated and healthy NMJs (1000 each). In both cases, 80% of the data was used for training models, and the remaining 20% was used to evaluate models. Both models created were random forest models created using 10-fold cross validation.

3. Comparison of low read depth QuantSeq and RNA-seq

3.1. Publication

This comparison of methods has been published in the journal Frontiers in Genetics (Jarvis et al., 2020). Elements of this analysis have also been published in Science Advances (Birsa et al., 2021) and Nucleic Acids Research (Humphrey et al., 2020). The text present here is largely adapted from (Jarvis et al., 2020).

3.2. Introduction

As explained in the introduction, mutations in Fused in Sarcoma (*FUS*) are linked to development of amyotrophic lateral sclerosis (ALS). Changes in the levels of FUS cause significant changes in splicing and expression (Ishigaki et al., 2012; Coady and Manley, 2015; Humphrey et al., 2020). Members of the Fratta lab worked to develop a mouse model which carries mutations in endogenous FUS, this mutation causes skipping of the penultimate exon in FUS – exon 14 – and will henceforth be known as FUS d14 (Devoy et al., 2017). It is considered a humanized mutation, as it is a mutation found in humans which has been linked with onset of ALS (DeJesus-Hernandez et al., 2010). It expresses ALS at endogenous levels when heterozygous, and causes progressive motor neuron loss in midlife, so has a similar effect to that observed in patients. This allows us to see the effects of the mutation without the overexpression present in some other FUS models. Concurrent with our mouse models we produced a set of mice who's FUS had been fully knocked out (FUS KO), these, along with littermate controls for both mutations, were sequenced using total RNA-seq and the changes were compared. I found the expression changes induced by these mutations appeared to mostly be caused by loss of function (Humphrey et al., 2020). It does this by changing levels of intron retention events in RBPs, causing them to be unable to correctly function.

We also sequenced the same samples using a form of 3' sequencing called QuantSeq to compare the performance of these two different approaches in evaluating gene expression. QuantSeq aims to be able to provide useful information about differential expression at lower read depths than other methods, as well as providing data on differential polyadenylation (Moll et al., 2014). After fragmentation of the constituent RNA, QuantSeq extracts only fragments which have a polyA signal attached. These fragments are the only ones sequenced meaning that multiple fragments from the same initial piece of RNA will not be sequenced, reducing redundancy. This allows for a more accurate gauge of expression at lower read depths.

In order to investigate the viability of using QuantSeq as a possible replacement for total RNA-seq, I compared the results of the sequencing of two datasets; Fus KO and Fus d14, each against their own wild-type (WT) littermate controls. I observed differences between the genes that are found to be differentially expressed using the two methods and aimed to identify some possible causes.

3.3. Methods

FUS knockout mice were obtained from the Mouse Knockout Project [FUStm1(KOMP)Vlcg]. FUS d14 mice were created as previously described (Devoy et al., 2017). All procedures for the care and treatment of animals were in accordance with the Animals (Scientific Procedures) Act 1986 Amendment Regulations 2012. The total number of samples was four FUS KO mice and four WT littermate controls, and four FUS d14 mice with four of their own littermate WT controls. All mouse work was performed by Nicol Birsa.

For RNA sequencing experiments FUS d14 or KO heterozygous and homozygous mice were compared to their respective WT littermates. Spinal cords were collected from E17.5 mouse embryos. Tissues were snap frozen, genotyped and total RNA was extracted from the appropriate samples using Qiazol followed by the mini RNAeasy kit (Qiagen). RNA samples used for sequencing all had RIN values of 9.9 or 10. The same samples were used for total and QuantSeq sequencing and preparation and extraction was performed by Nicol Birsa. For total RNA-seq, cDNA libraries were made at the Oxford Genomics facility using a TruSeq stranded total RNA RiboZero protocol (Illumina).

Libraries were sequenced on an Illumina HiSeq to generate paired end 150 bp reads. For QuantSeq libraries the 3_0 mRNA-seq library prep kit REV for Illumina (Lexogen) was used QuantSeq and samples were sequenced by the Lexogen facility (Austria). In QuantSeq the average number of reads was 933,955 in d14 and 1,453,108 in KO. In RNA-seq the average number of reads was 35,678,902 in d14 and 39,159,292 in KO. The full number of reads found in each sample can be found in Table 3-1.

Alignment of the RNA-seq samples was performed by the pipeline as described in chapter 2 using an in-house pipeline. Gregor Rot aligned our QuantSeq samples in the method described in his paper (Rot et al., 2017). This method also uses STAR, but adds in filters to reduce instances of internal priming by filtering alignments containing stretches of six consecutive A or with 70% A coverage within 10 bp of the polyA signal. Differential Expression was performed using the DESeq2 algorithm via EnrichmentBrowser as described in Chapter 2.3.

Paired T-tests were used to test for correlation where appropriate. The full code for differential expression and comparisons between the two samples can be found on GitHub (https://github.com/SethMagnusJarvis/PhDFigureCreation/tree/main/QuantSeq)

Table 3-1. Number of reads sequenced in QuantSeq and RNA-seq

Sample Name	QuantSeq Count	RNA-seq Count
d1WT	582093	35709893
d2WT	1326069	38443056
d3WT	1123340	43643022
d4WT	875931	37918784
d1HOM	736595	36154653
d2HOM	929399	40040056
d3HOM	913675	32262506
d4HOM	984539	21414181
k1WT	1456620	33683581
k2WT	1440645	33593705
k3WT	1429041	39451942
k4WT	1295541	37243756
k1KO	1276747	29531417
k2KO	1718545	35790444
k3KO	1503195	54350352
k4KO	1504529	46666652

3.4. Results

3.4.1 Analysis of overall differences between methods

We performed our methods comparison on two sets of samples derived from embryonic spinal cords of: (a) four Fus KO mice and WT littermate controls, and (b) four Fus d14 mice with their own littermate WT controls. The same RNA samples were then sequenced either using standard total RNA-seq for library preparation or QuantSeq kits produced by Lexogen. QuantSeq selectively amplifies regions of RNA close to a polyA signal, whilst total RNA-seq sequences all pieces of RNA present within the cell regardless of a presence of a polyA signal. The core differences between the two methods are illustrated by Figure 3-1.

Our intention with this data was to evaluate the relative performance of low read depth QuantSeq against RNA-seq to see whether it would be a suitable supplement to, or replacement for RNA-seq in our lab given as it is comparatively inexpensive, and can provide accurate information on differential polyadenylation, a topic in which I am interested as ALS has been shown to have an effect on polyadenylation.

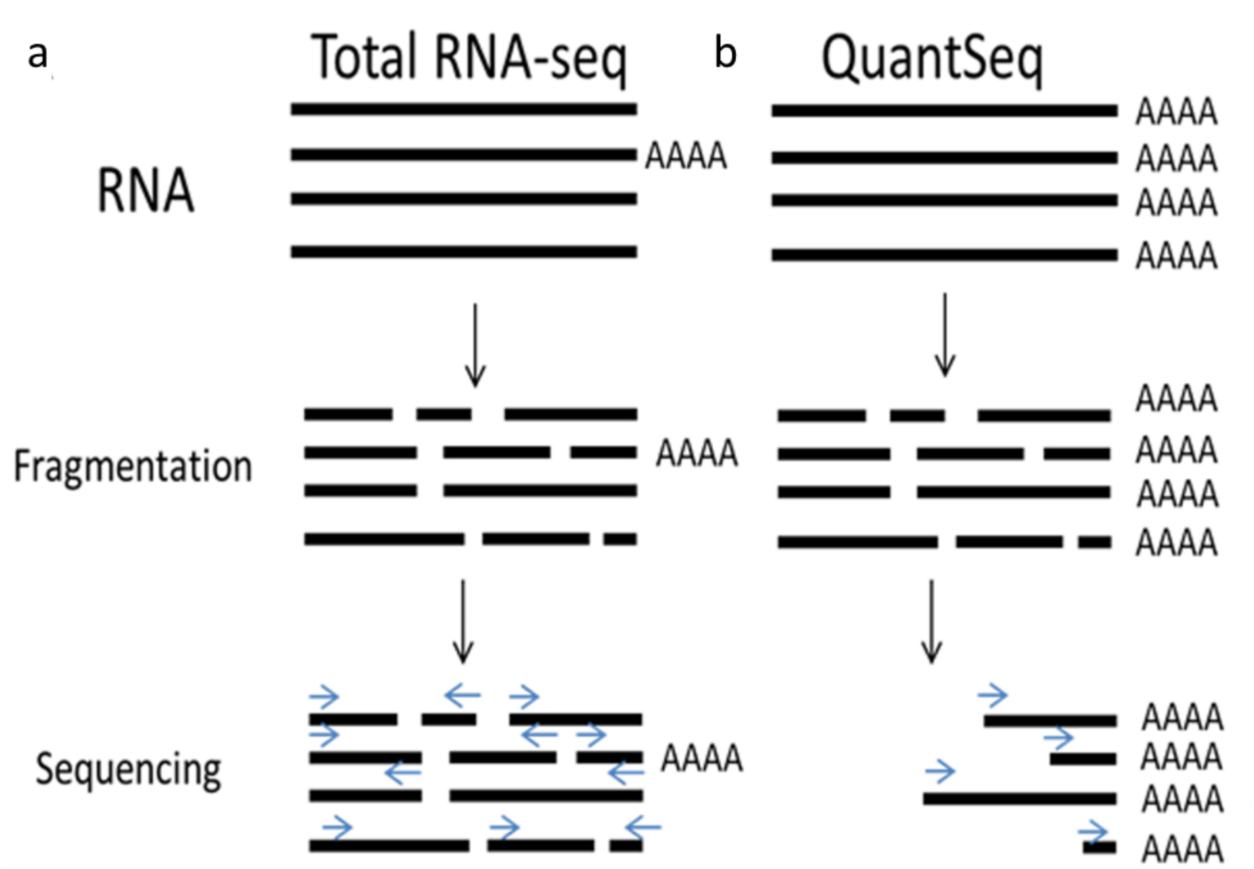

Figure 3-1. Plot comparing QuantSeq and Total RNA-seq. It demonstrates that Total RNA-seq uses all fragments from all reads, while QuantSeq only uses fragments which have polyA tails attached. From Jarvis, S. et al 2020

We made PCA plots of normalised reads by gene in both datasets (Figure 3-2). The way the data segregated seemed to show that the two samples were not directly comparable. This was because the RNA-seq samples clustered more strongly with the other RNA-seq samples, than they did with themselves, or other samples sequenced with the same mutation..

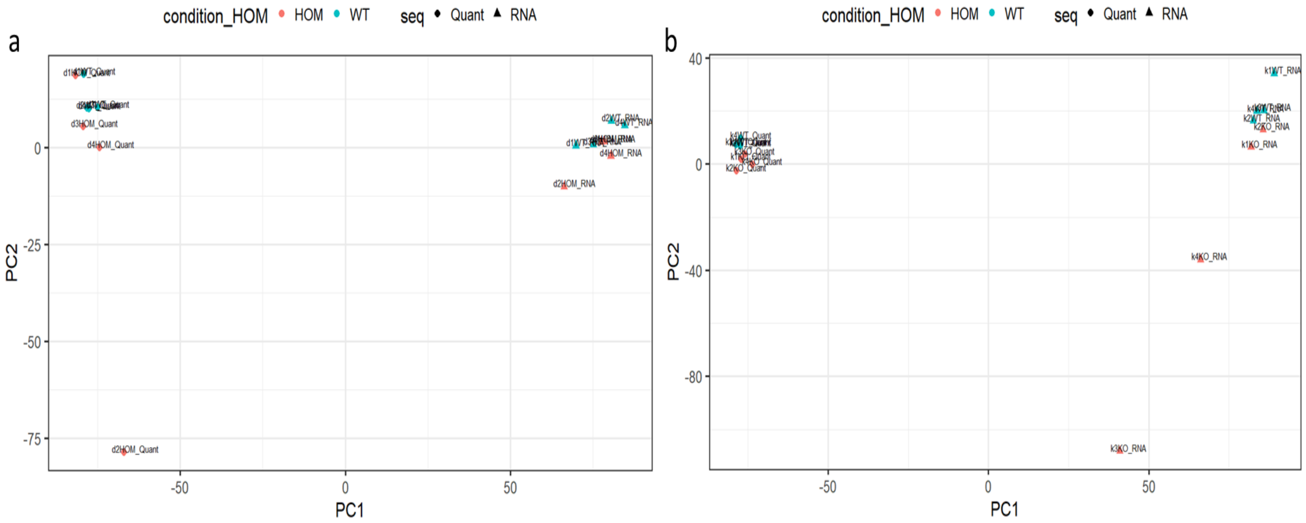

Figure 3-2. PCA of normalised RNA expression by gene in (a) d14 and (b) KO. Both figures show that the samples are more similar to the samples which were sequenced using the same method than they were to the samples taken from the same organism sequenced using different methods.

In order to evaluate whether there was a correlation overall despite the differences, I investigated whether the mean expression of genes correlated between methods in each mutation (Figure 3-3). There appeared to be a positive correlation in the graphs and I found that this was reflected in a paired t-test in both datasets (cor = 0.3515, p < 0.0005 in the d14 datasets, and cor = 0.3586, p < 0.0005 in the KO datasets). In other words, there is a statistically significant positive correlation in the number of reads per gene in QuantSeq and RNA-Seq. Some of the differences observed may be partially due to how small the size of the correlation is, but it also means that differences observed are mainly due to other factors.

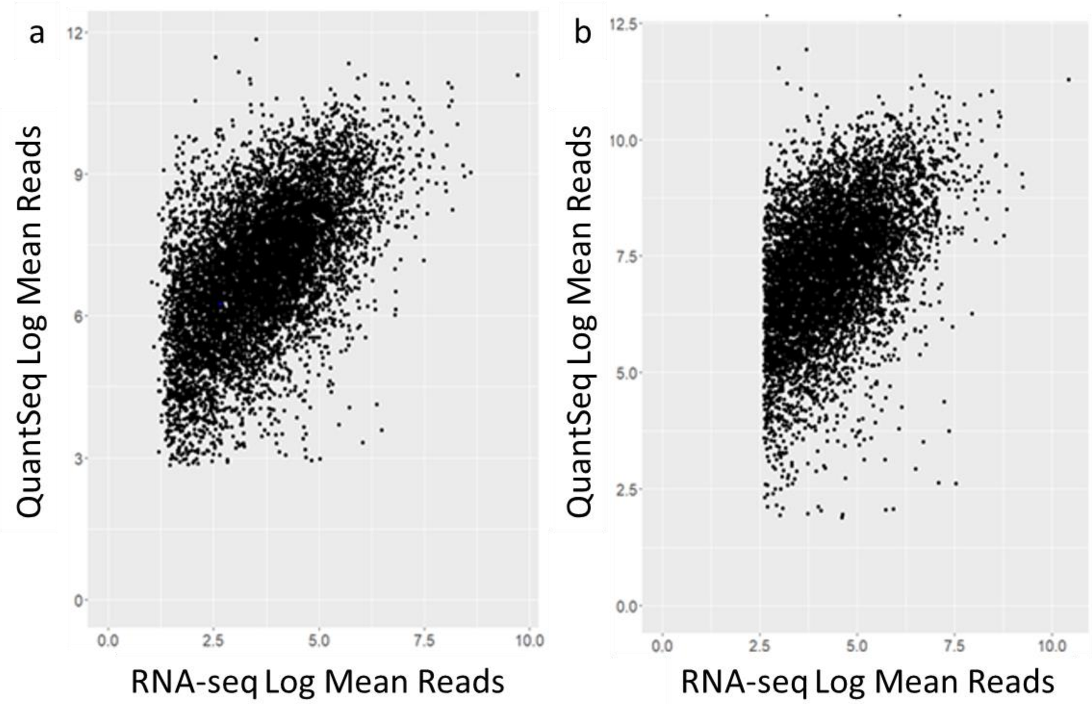

Figure 3-3. Scatter plots of log Mean reads QuantSeq vs log mean reads in RNA-seq in (a) d14 and (b) KO from Jarvis, S. et al 2020

When performing this step, I noticed that several genes found in the RNA-seq dataset were not found at all in the QuantSeq dataset as they did not appear in the list of genes which had a mean number of reads in QuantSeq when I joined the tables. Some of these were due to low expression in QuantSeq resulting in them being filtered out, but some were pseudogenes or genes which had not been experimentally confirmed. QuantSeq is only capable of interacting with RNA which has polyA signals attached, this means that it can only work with RNA which is relatively heavily processed and likely to be translated into a protein. These genes therefore cannot be detected by QuantSeq and were removed from future analysis.

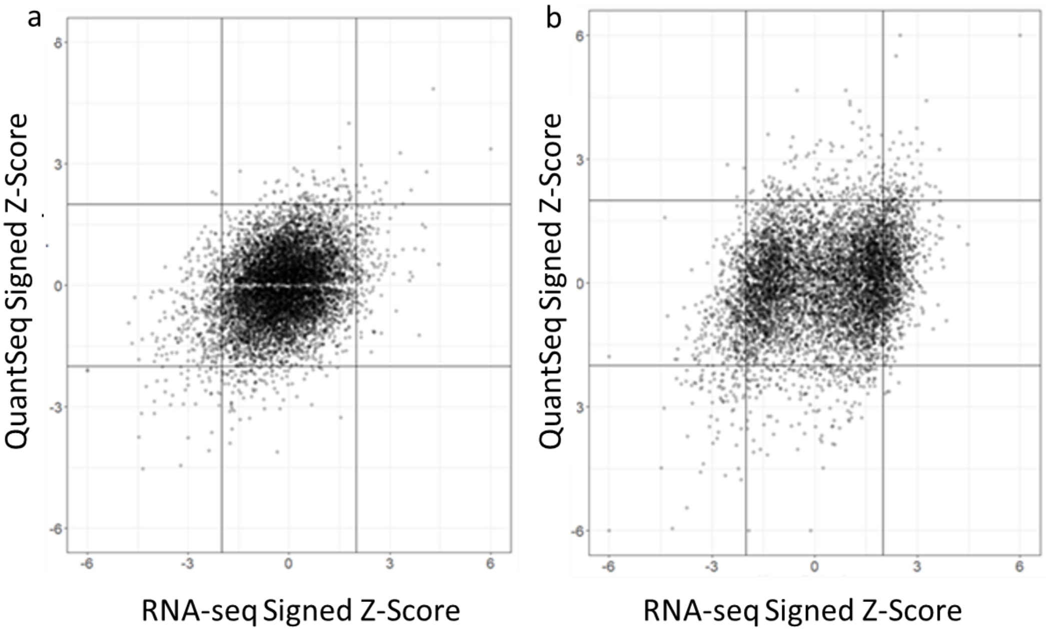

Figure 3-4. Comparison of Z-scores in QuantSeq and RNA-seq in (a) d14 and (b) KO datasets. Z-scores are a signed measure of significance based on the p-value and fold change of the gene. Both panels broadly show that genes which are significant using one method are either also significant with the same direction of fold change in the other method, or are non-significant. from Jarvis, S. et al 2020

After running differential expression analysis using DESeq2 on the sequenced samples and their relevant controls, I calculated the signed z-scores – a measure based on the p-value and the fold change of the gene. To see if the trend in expression resulted in a similar trend in differential expression of genes I created Z-score plots comparing RNA-seq and QuantSeq data (Figure 3-4). Genes with an unsigned z-score greater than 2 (equivalent to an unadjusted p value of 0.023), meant these genes were most likely to be significantly differentially expressed, rarely exceeded the significance thresholds in opposite directions in analysis of QuantSeq and RNA-seq.

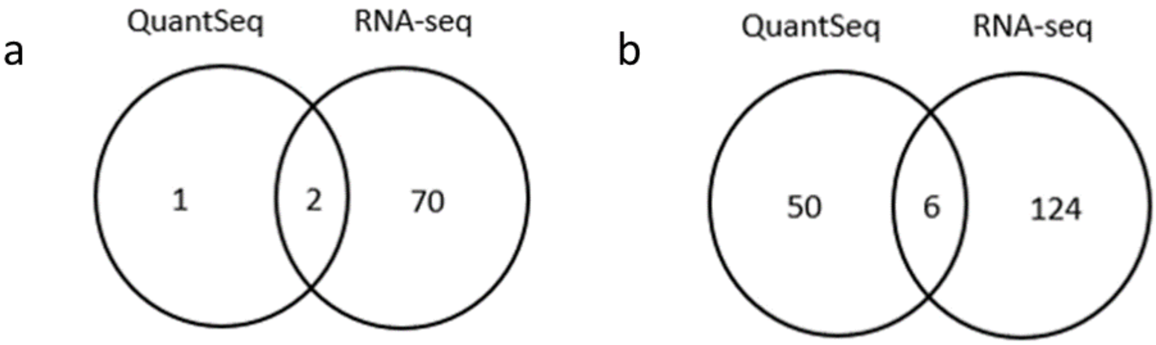

Figure 3-5. Venn diagrams showing overlap of significant genes (padj <0.05) between QuantSeq and RNA-seq in (a) d14 and (b) KO datasets. It shows that in the d14 samples 2/3 genes significant in QuantSeq are also significant in RNA-seq, and in KO that only 10.7% of genes significant in QuantSeq are significant in RNA-seq.

When I directly compared genes which were significantly differentially expressed (adjusted p-value < 0.05) (Figure 3-5) I found that of the 70 genes which were found to be significantly differentially expressed in RNA-seq, only 2 were also found in QuantSeq (*Pspc1* and *Selenop*). QuantSeq did find one gene to be significantly differentially expressed that RNA-seq did not, Mt2 – a gene linked to metal ion binding. Substantially more genes were found to be significantly differentially expressed in our KO QuantSeq dataset. Only 10.71% of the genes which were found to be significant in QuantSeq were also found to be significant in RNA-seq despite both groups having more genes significantly differentially expressed overall. The genes which were found to be significantly differentially expressed in both datasets were *Trim72*, *Fus*, *Bcas1*, *Gjd2*, *Ahi1*, and *Chodl* are linked to, among other things, cell repair and development of the nervous system(Cong et al., 2020; Ferland et al., 2004).

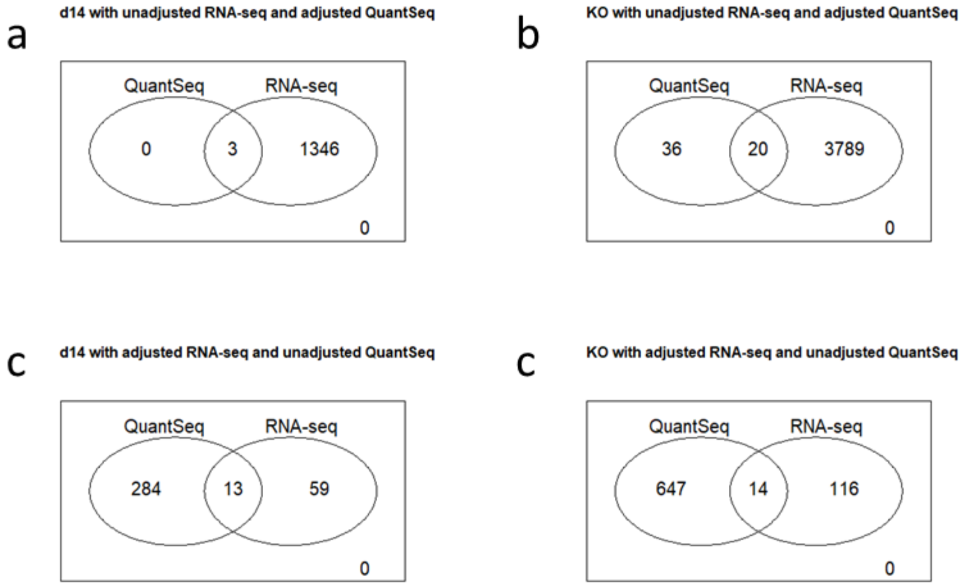
a
d14 with unadjusted RNA-seq and adjusted QuantSeq
QuantSeq
RNA-seq
0
3
1346
0
b
KO with unadjusted RNA-seq and adjusted QuantSeq
QuantSeq
RNA-seq
36
20
3789
0
c
d14 with adjusted RNA-seq and unadjusted QuantSeq
QuantSeq
RNA-seq
284
13
59
0
c
KO with adjusted RNA-seq and unadjusted QuantSeq
QuantSeq
RNA-seq
647
14
116
0

Figure 3-6. Venn diagrams showing the comparison genes found significant using an unadjusted p-value of 0.05 in one dataset and an adjusted p-value of 0.05 in the other dataset, using unadjusted RNA-seq p-values in a-b and unadjusted QuantSeq p-values in c-d

Adjusted p-values are used to reduce the likelihood that a difference in levels of expression is due to random chance. I thought it might be useful to test relaxing the significance threshold of one of the comparisons at a time to evaluate whether the overlap in significance was increased. There is a chance that the stringent adjustment was meaning that genes where there truly was a difference were being obscured as it did not quite reach significance. Since we were evaluating a specific pool of genes, I felt this was not unreasonable. It would also let us see if it was potentially a read-depth issue, causing subtle changes in expression to be missed as they did not reach the extremes required for adjusted significance (Figure 3-6). In the KO data I found that when I relaxed the QuantSeq threshold, the percentage of RNA-seq genes that overlapped rose from 10.71% to 35.71%. There was a bigger increase when I relaxed the RNA-seq threshold, where the overlap with QuantSeq rose from 10 to 36% of the total genes QuantSeq found to be significant.

To continue our investigation into the differences between our two methods of sequencing I ranked each gene by its mean expression relative to the expression of other genes in the same sample (Figure 3-7). I found a slight but significant correlation of the rank in both the KO (r^2 0.2795, p-value <0.0005) and d14 (r^2 0.323, p-value <0.0005) samples across genes which were significant in either sample. This supports the hypothesis that read depth and sampling coverage are a major source of the discrepancies observed.

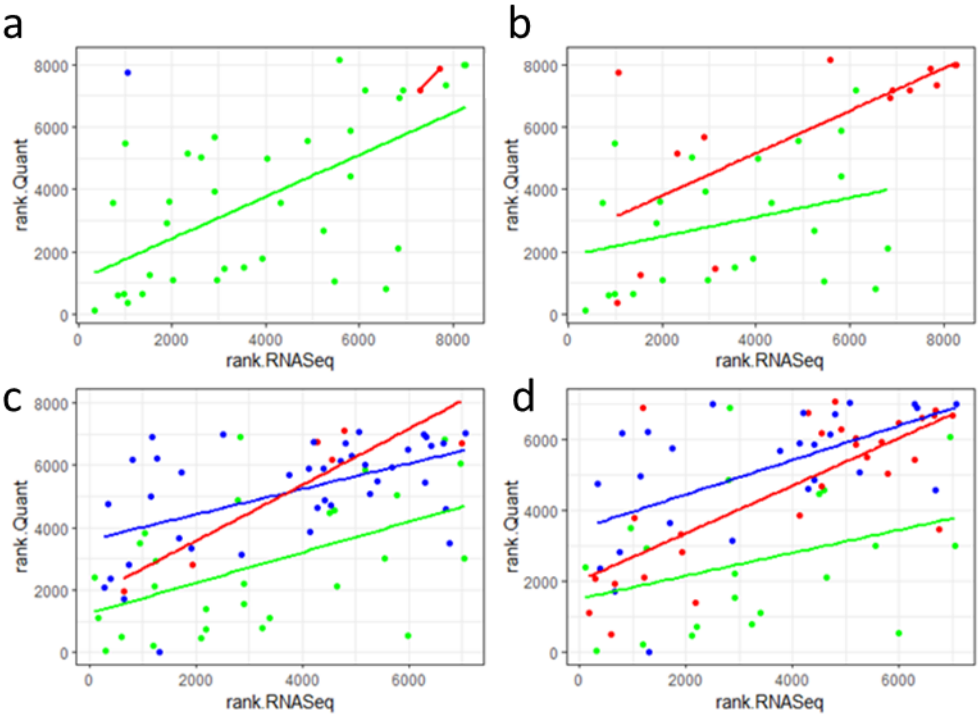

Figure 3-7. Comparison of genes found to be significant in one dataset, ranked by their relative expression in d14 (A-B) and KO (C-D). Genes are coloured green if only significant (padj < 0.05) in RNA-seq, blue if only significant in QuantSeq, and red if significant in both. Panels B & D use a relaxed threshold for red, where if a gene is adjusted significant after multiple testing correction in one dataset, it only needs to have an unadjusted p-value threshold of 0.05 in the other. From Jarvis, S. et al 2020

Ten genes which were significantly differentially expressed in QuantSeq were not present in the matrix of Reads from our QuantSeq data. Of these, eight were found as fusion genes, and two had been filtered due to having very low reads. The majority of genes found to be significantly differentially expressed in RNA-seq were not found in QuantSeq data. Figure 3-8 shows the breakdown of these genes. The majority of these genes had very low expression, although some had no expression at all in our data. Several of the genes which were not found were mostly antisense or long non-coding RNAs, these would not be found in QuantSeq as it requires RNA transcripts to have polyA tails which non-functional genes do not have.

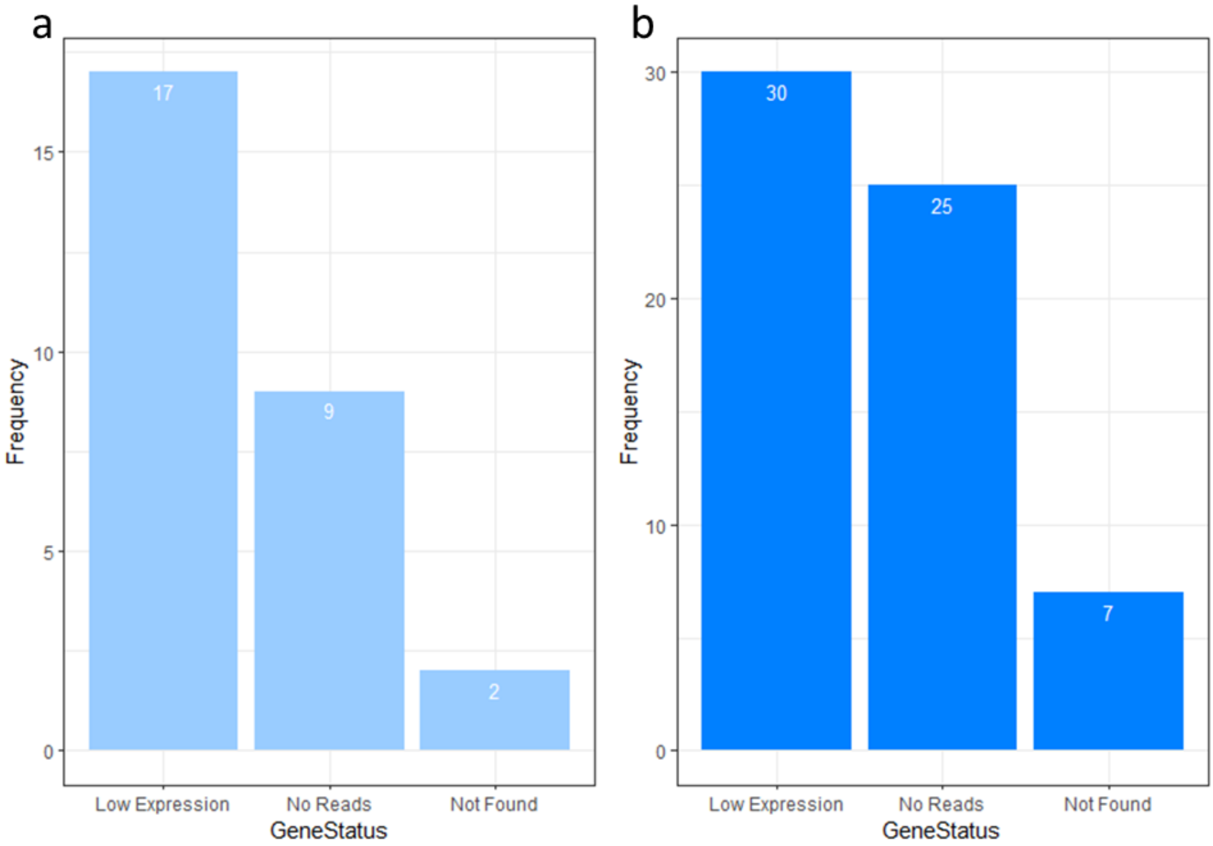

Figure 3-8. Bar plots showing the distribution of genes that were found significant in RNA-seq but were not present in our QuantSeq differential expression results after filtering pseudogenes and genes to be experimentally confirmed in (A) d14 and (B) KO. From Jarvis, S. et al 2020

Table 3-2. Table showing overlap of significant GO terms in each dataset

	GO terms only significant in Total RNA-seq	GO terms only significant in QuantSeq	GO terms significant in both
d14	28	12	5
KO	186	109	98

3.4.2 Investigation of possible causes of differences between methods

We compared the biological process GO terms arising from the different analyses using genes which were found to be significant with an unadjusted p-value <0.05. This allowed us to have a broader base of GO-terms to potentially see any overlap. Table 3-2 shows the results of this comparison. The majority of GO terms in both datasets did not overlap, however I could see, especially in the KO dataset (where more GO-terms were found to be significant) that there was a 25% overlap of the total number of GO terms found to be significant. These included GO terms related to sensory perception, localisation, transport, and RNA-metabolic processes. This tells us that while the two datasets do differ in which genes they find significant, the processes that those genes serve do overlap to an extent. The full results of which GO terms were found to be significant can be found in Table 3-4 and Table 3-5.

Table 3-3. Minimum number of genes required to cover 10 and 50% of total reads

	d14 QuantSeq	d14 RNA-seq	KO QuantSeq	KO RNA-seq
First 10%	15	61	16	64
First 50%	698	1159	646	1224

As stated in the methods, RNA-seq has about 30 times the average number of reads per sample as QuantSeq. I used the sample function in R to take random reads from our RNA-seq dataset, at several different levels between 1.5 million reads and 20 million reads per sample to see how they compared. There was a positive correlation overall, with more reads resulting in more genes found to be differentially expressed. The results in KO can be observed in Figure 3-9. At these lower levels of sampling, QuantSeq does find more genes to be differentially expressed than RNA-seq in KO aside from at 20 million reads per sample and finds as many/more genes differentially expressed below 10 million reads per sample in the d14 dataset. The number of genes found to be significant by RNA-seq is highly correlated with the number of reads (adjusted r^2 0.9049, p-value 0.0083).

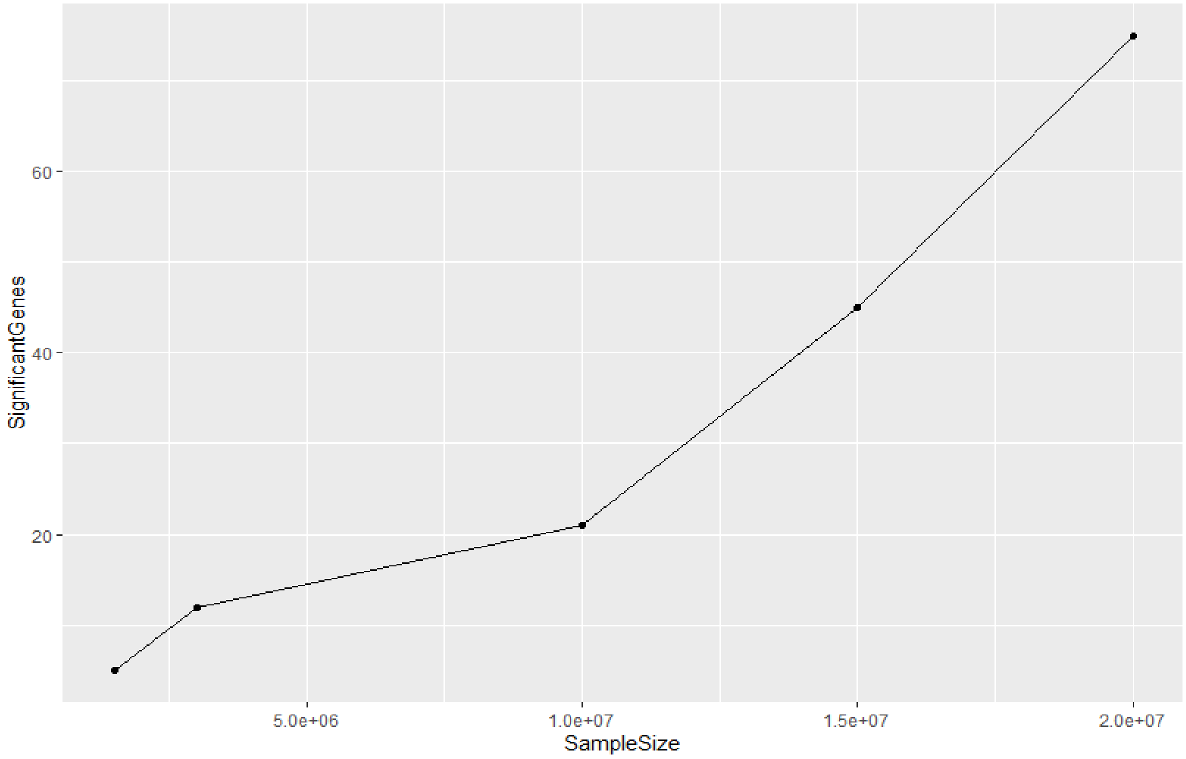

Figure 3-9. Plot showing the number of genes with padj < 0.05 plotted against the number of reads that the KO RNA-seq dataset had been downsampled to. From Jarvis, S. et al 2020

Whilst the number of detected differentially expressed genes may depend mostly on the number of reads as shown above, our investigation progressed to explore reasons for the differences between genes that were observed to be differentially expressed between the two methods. RNA-seq finds about 5x the genes that QuantSeq does represented in the top 10% of reads, and about double the number of genes in the top 50% of reads (Table 3-3). There are about half the number of genes with reads in QuantSeq compared with RNA-seq meaning by the time 50% of reads have been accounted for, the proportion of total genes represented is similar. This suggests that RNA-seq tends to distribute reads more evenly across genes whereas QuantSeq has a large number of reads concentrated in relatively few genes.

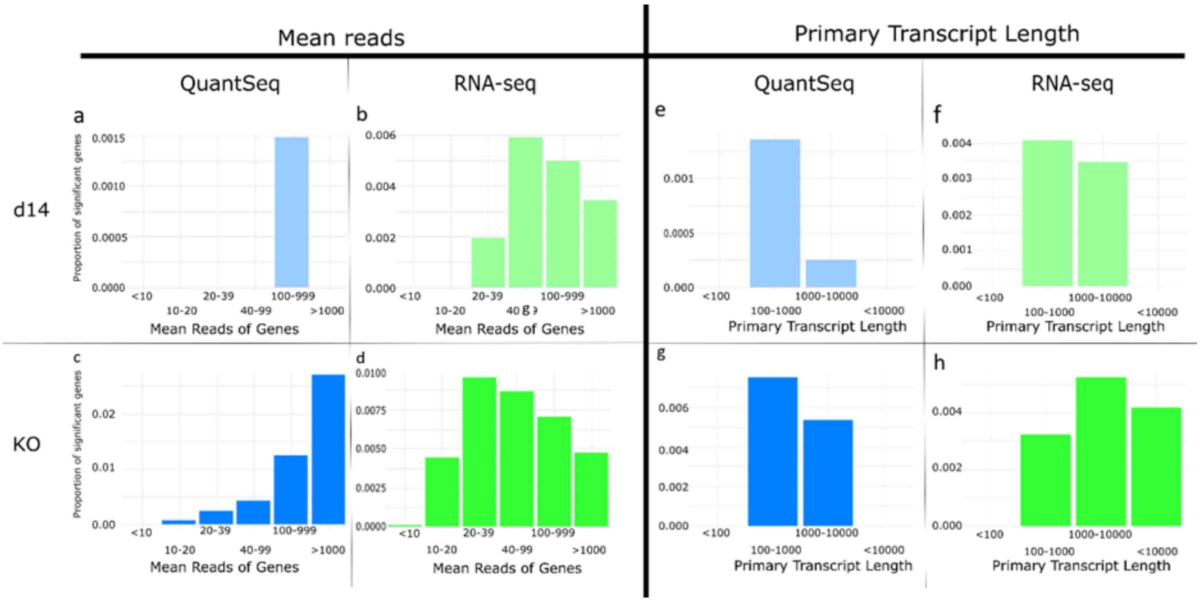

Figure 3-10. Possible sources of difference between QuantSeq and RNA-seq using d14. (A, C, E, G) and KO (B, D, F, H) datasets: (A-D) Bar plots showing the proportion of genes that are significantly differentially expressed (padj < 0.05) separated by the mean number of reads in the gene using (A and C) QuantSeq, and (B and D) RNA-seq sequencing; (E-H) Bar plots showing the proportion of genes that are significantly differentially expressed (padj < 0.05) separated by the length of the Appris Primary 1 transcript in the gene using the (E and G) QuantSeq, and (F and H) RNA-seq sequencing. From Jarvis, S. et al 2020

We wished to see if there were any obvious reasons for the differences I observed between the methods beyond differential read depth. Figure 3-10 A-D shows the correlation between the mean number of reads found by a method, and the proportion of genes which were found to be significant with an unadjusted p-value < 0.05. As can be seen, there is appears to be a positive trend in QuantSeq, more reads in a gene appears to increase its sensitivity to significantly differentially expressed events. In RNA-seq there is a minimum number of reads desired per gene, but past that point it plateaus and may even have a slight negative correlation (this disappears when using adjusted p-values).

Subsequently I wondered whether the length of a gene’s primary transcript had an effect on whether a gene was found to be significant (Figure 3-10 E-H). I found a negative correlation in both datasets in QuantSeq, and a slight positive correlation in RNA-seq between primary transcript length and proportion of genes found to be significant. The correlation persists in QuantSeq when using adjusted significance values but is no longer present in RNA-seq.

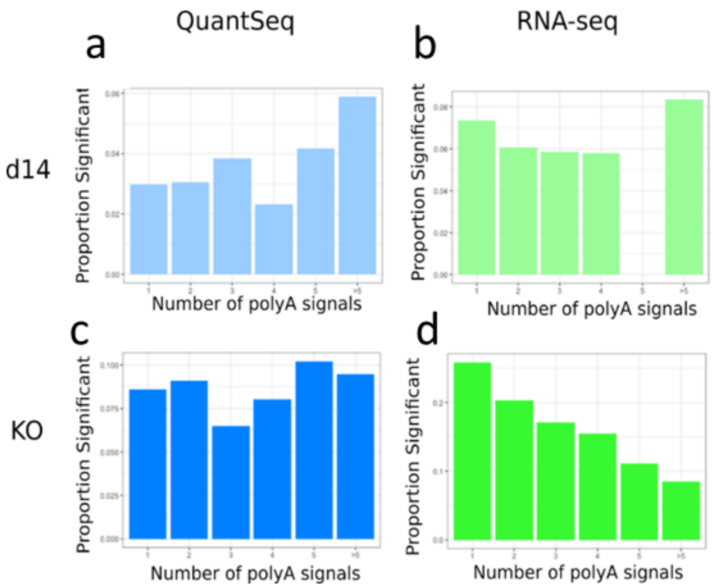

Figure 3-11. Bar plots showing the proportion of genes that are significantly differentially expressed (p-value < 0.05) separated by the number of polyA signals each gene has found within the polyA atlas. From Jarvis, S. et al 2020

As QuantSeq uses the polyA region to prime during sequencing I tested whether either method had a bias for genes with more polyA signals. I found that there was no correlation in any but the RNA-seq d14 dataset (Figure 3-11). This negative correlation is no longer present when using adjusted p-values.

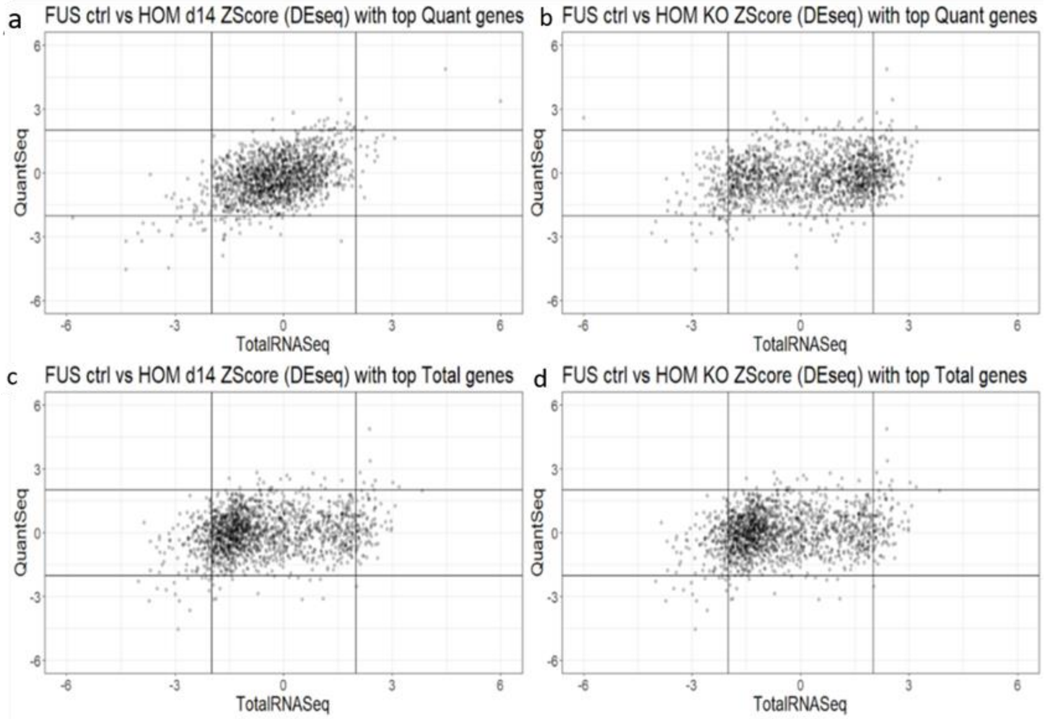

Figure 3-12. Plots comparing Z-Scores of RNA-Seq and QuantSeq in d14 & KO experiments of the top 25% most expressed genes in QuantSeq and RNA-seq. (a) the most expressed genes in QuantSeq using data from the d14 mutation, (b) the most expressed genes in QuantSeq using data from the KO mutation, (c) the most expressed genes in total RNA-seq using data from the d14 mutation, (d) the most expressed genes in total RNA-seq using data from the KO mutation. They do not show clear differences from the prior z-score plots aside from being more sparsely populated. From Jarvis, S. et al 2020

Finally, I wanted to see how the top 25% most expressed genes (Figure 3-12), and genes with a single polyA signal (Figure 3-13) changed our Z-score plots. Neither change made a substantial difference to the results.

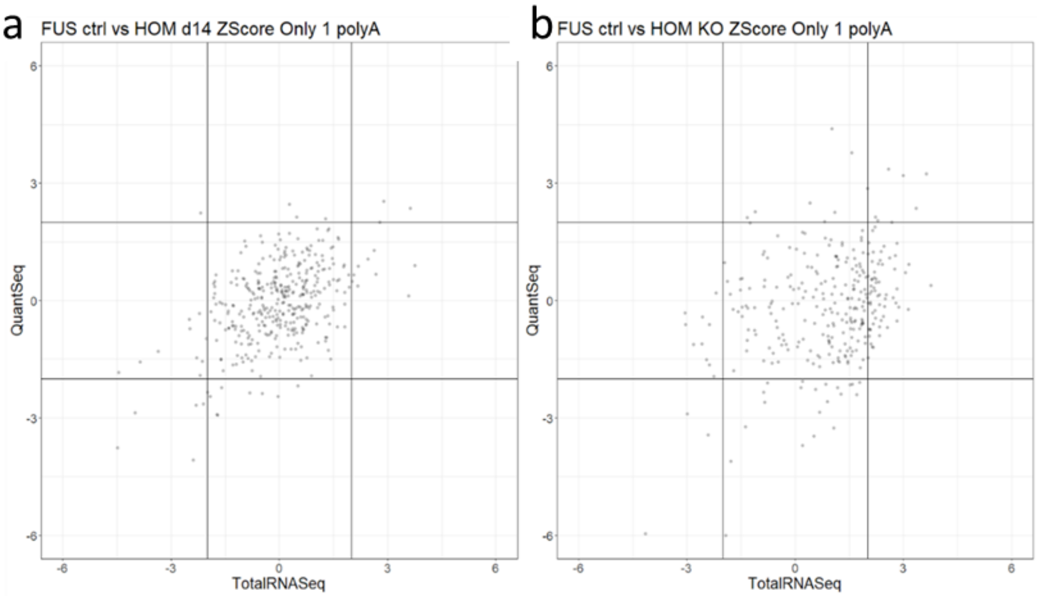

Figure 3-13. Plots comparing Z-Scores of RNA-Seq and QuantSeq in d14 & KO experiments of genes which only have 1 polyA signal in (a) d14, (b) KO. They do not show clear differences from the prior z-score plots aside from being more sparsely populated. From Jarvis, S. et al 2020

3.5. Discussion

Due to the rapid development of various NGS technologies including RNA-seq, various methods are aiming to provide a better service, more information on certain methods, or at a lower cost. In the case of low read depth QuantSeq the hope is to provide both a lower cost service and more information on differential polyadenylation over total RNA-seq. While the read depth I used is not recommended on the Lexogen site, it was sequenced by them so falls within parameters of what they consider acceptable usage.

QuantSeq is able to provide information on differential polyadenylation in ways that RNA-seq is not. As it binds to the poly-A signal, we can evaluate which regions of each gene have polyA signals attached and from that gain some insight into the process. While it cannot provide information on splicing, this additional information means that it has a distinct use case irrespective of how it performs in other use cases. We felt that it was useful to compare how it performs at low read depts to see inexpensive low-read depth analysis was another potential use case.

The largest methodological difference between the two methods is how the libraries are prepared. This leads to substantial changes in the characteristics of sequenced RNA. QuantSeq sequences one fragment per piece of RNA, and exclusively sequences fragments with a polyA signal attached. Total RNA-seq fragments and sequences a random sample of all RNA within the cell. This means that QuantSeq will tend to sequence more mature RNA, and will not sequence non-coding genes. This will lead to some difference in levels of expression at a gene level as the two methods are sampling from a different population of RNA. QuantSeq is also likely to sequence less ribosomal RNA as while ribosomal RNA is sometimes polyadenylated, it does not tend to be abundant (Slomovic et al., 2006). All of my analysis within this chapter does need to be viewed through the lens of these differences.

I found substantial differences in the genes found to be significantly differentially expressed in QuantSeq and total RNA-seq. Total RNA-seq is a very widely used method and so it was concerning that QuantSeq's results were different. The relatively low reads of the genes which weren't significant in QuantSeq but were in RNA-seq did suggest that the issue was with the read depth rather than with the method itself.

Given the appearance of our Z-score plots, and how low the overlap of genes found to be significantly differentially expressed between the two datasets were, I felt it was important to identify any possible sources of differences beyond the difference in read depth. By sequencing both cases and controls using both methods I have done our best to ensure that I correct for biases within each method, that being said, I suspected that the methods may have a bias towards certain types of genes. I did find that at this read depth, QuantSeq was more likely to find a gene to be significant when it had a high number of reads whereas RNA-seq only required a certain baseline of reads before more stopped making a difference. That may change at higher read depth. Unlike in other work (Ma et al., 2019), I did not find that RNA-seq was more likely to find genes to be significant if they had a longer primary transcript, this is likely because they used an mRNA-sequencing kit for library preparation. QuantSeq's likelihood of finding a gene to be differentially expressed did seem slightly negatively correlated with primary transcript length. I didn't manage to find any other contributing factors to the differences observed.

While polyA selection has been linked to changes in splicing, Humphrey et al., 2017 found that the overlap between splicing and expression changes in these datasets was minimal therefore I felt it was unlikely to be responsible for the differences observed.

When I down sampled the RNA-seq data I did find that there were reads in more genes in our RNA-seq data than our QuantSeq data, this is likely to have led to the relative underperformance of RNA-seq at this low level of read depth, in combination with QuantSeq advantage of only having one transcript per read, potentially increasing its sensitivity to changes at this read depth. As Illumina does not recommend this low read depth (Illumina Inc, 2017), the poor performance was expected, but useful as a comparative to QuantSeq.

Colleagues in the Fratta laboratory have validated many of the genes which RNA-seq found to be exclusively differentially expressed in the lab. Some of the genes which QuantSeq found to be exclusively differentially expressed, including 3 of the 5 most differentially expressed genes in our knockout samples (*Mcur1*, *FTSJ1*, and *GPR17*), have been linked to neurodegeneration, and in the case of *GPR17* specifically to ALS elsewhere (Liao et al., 2017; Angelova et al., 2020; Bonfanti et al., 2020). The rest of the genes found to be significant exclusively in QuantSeq sadly have not been tested.

Corley et al., 2019 found a much better correlation between RNA-seq and QuantSeq results when performed at a similar read depth. They also used an mRNA library preparation method. They used Kallisto – a fast, pseudoalignment tool – for analysis. They suggest that the combination of low read depth QuantSeq and Kallisto is a good tool for investigative alignment particularly where polyadenylation is of interest. I think that my research supports this idea of a combination. It could provide useful insights into differential expression in samples at a lower cost and faster than other methods. The additional insight into differential polyadenylation is especially useful. Given the differences between the two methods however, experimental validation of any hits observed using methods such as qPCR and nanostring is especially important.

I have highlighted some ways that these methods differ and do advise caution when using low read depth QuantSeq given how it differs from the current gold standard of RNA-sequencing. At higher read depths, QuantSeq does provide useful insight (Oh et al. 2020), and as stated before, correlates well with the results from RNA-seq analysis. At this low read depth, while some of the hits have been validated, many have not, and several genes which I expected to be differentially expressed were not found to be. Given this, while we aren't sure of the sensitivity, low read depth QuantSeq may be useful for broad inexpensive surveys of many samples, the most promising results of which can then be experimentally validated.

3.5.1 Future work

The primary questions arising from my analysis are:

- Do genes exclusively significantly differentially expressed in QuantSeq show changes in qPCR?
- Do similar comparisons in other datasets show similar differences?

The best method for validating DEGs exclusively found in QuantSeq would likely be an experimental method such as qPCR. If the exclusively DEGs in QuantSeq do seem to show differences, then they may be future targets for analysis of the changes that result from mutations in FUS. It may also suggest a limitation of using total RNA-seq on datasets which result from changes in genes known to be linked to RNA regulation. If the genes don't show changes, then it suggests that low read depth QuantSeq may not be useful even for initial analysis of changes.

Validating whether the differences between the two methods are also seen in other methods would also be important. This validation would ideally be performed by a similar experiment; a group would sequence a dataset consisting of a mutation that is known to cause large numbers of DEGs using both low read depth QuantSeq and RNA-seq. If the differences in DEGs are not seen in this other dataset, then this would substantially expand the utility of low read-depth QuantSeq. It might then be worth investigating why our data had the differences I observed. There are two major possible causes of these differences. One is that our mutation is in a gene that is heavily linked to RNA-regulation. The other is that QuantSeq exclusively examines mRNA, while total RNA-seq examines all RNA. Total RNA-seq looks at all transcripts in the cell while mRNA-seq looks exclusively at coding regions. If there are substantial differences between expression in coding and non-coding regions in our data it may be responsible for some of the differences observed. If the differences persist then the utility of low read depth QuantSeq depends on whether DEGs exclusively found in QuantSeq are found to be changed in lab-based validation.

The implications of this work are potentially interesting. If the DEGs are validated, or the differences in which DEGs are found to be significant do not persist then this suggests that low read depth QuantSeq may be a powerful tool for cheap, widespread initial exploration of differential expression data. If the differences are not seen in other datasets, then low read depth QuantSeq may be a useful replacement for RNA-seq wholesale; it is less expensive, and provides easier access to information on differential polyadenylation.

Given the lack of statistical power found in a large number of results in the QuantSeq data, an additional opportunity may arise. The lower power suggests that the tools which currently exist may not be best suited to low read depth QuantSeq data. One possible avenue for people working with the data would be to further develop tools to extract the maximum possible information from the data.

Table 3-4. Results of GO term analysis of all genes in d14 in both datasets

Description	GO Term	Fold Enrichment Total	FDR Total	Fold Enrichment Quant	FDR Quant
sensory perception of chemical stimulus	GO:0007606	-0.13	1.13E-10	-NA	0.00518
sensory perception of smell	GO:0007608	-0.12	3.36E-10	-NA	0.0115
regulation of biological quality	GO:0065008	+1.34	0.01	+1.69	0.00665
cellular process	GO:0009987	+1.11	0.0182	+1.25	0.0013
developmental process	GO:0032502	+1.25	0.0256	+1.62	0.00141

Table 3-5. Results of GO terms found to be significant in both datasets in KO filtered to terms which have a fold enrichment >= 1.5.

Description	GO Term	Fold Enrichment Total	FDR Total	Fold Enrichment Quant	FDR Quant
cellular nitrogen compound metabolic process	GO:0034641	1.54	1.16E-18	1.51	0.00152
gene expression	GO:0010467	1.52	2.46E-09	1.86	3.58E-05
RNA metabolic process	GO:0016070	1.54	2.39E-07	1.71	0.0115
translation	GO:0006412	2.15	1.01E-06	2.7	0.00541
peptide biosynthetic process	GO:0043043	2.09	1.70E-06	2.54	0.0108
peptide metabolic process	GO:0006518	1.86	7.06E-06	2.16	0.0273
RNA processing	GO:0006396	1.63	1.14E-05	2.04	0.00352
regulation of cellular amide metabolic process	GO:0034248	1.74	0.00103	2.23	0.0433
positive regulation of cell migration	GO:0030335	1.54	0.00601	2	0.0373
regulation of cytoskeleton organization	GO:0051493	1.52	0.0079	2.49	0.00018
divalent inorganic cation homeostasis	GO:0072507	1.5	0.0198	2.23	0.00665
cellular divalent inorganic cation homeostasis	GO:0072503	1.52	0.0221	2.35	0.00413
calcium ion homeostasis	GO:0055074	1.5	0.031	2.07	0.0401
cellular calcium ion homeostasis	GO:0006874	1.5	0.0324	2.15	0.0311

4. Differential Expression in Post Mortem brain tissue in FTD patients

4.1. Publication

This work has yet to be published.

4.2. Introduction

As stated in the introduction, the most common cause of frontotemporal dementia is mutations in the gene *C9orf72*. The 2nd most common cause is mutations in the gene *MAPT* which encodes the protein TAU.

TAU proteins act to maintain stability of microtubules in neuronal cells. Mutations in *MAPT* can lead to TAU proteins being abnormally phosphorylated and accumulating within cells (Alonso et al., 1997). This accumulation then leads to the development of the behavioural variant of frontotemporal dementia (Rademakers et al., 2004). Unlike C9, mutations in MAPT have not been linked to the development of ALS.

FTD is a neurodegenerative disorder, and some pathology has been linked to changes in RNA. In order to investigate changes caused by various pathologies RNA-seq is often performed on bulk brain tissues. As discussed in 1.3.1, brain tissue is predominantly comprised of neurons and glial cells. Different cell types may have different levels of expression both in marker genes – genes that are known to be associated with the cell types, and other genes. Differences in the number or degree of activity in various cell types between samples may result in biological noise. This makes it harder to identify true changes in expression(Shen-Orr & Gaujoux, 2013). Given this possibility for noise, I aimed to use cell type disambiguation techniques in order to determine whether there were substantial differences in the proportions of various cell types in the tissue data being analysed. Cell type disambiguation tends to rely on using levels of expression in various marker genes to find some measure related to levels of various cell types. If my investigation does find substantial changes in cell types between either samples or conditions, I am aiming to determine whether it is worthwhile correcting for these changes in our differential expression analysis as different cell types express some genes at different levels.

There are various cell type disambiguation techniques which have been developed, but analysis of cell types in brains is relatively uncommon. Most of these tools are tissue agnostic, but all require reference data from various cell types. Table 1-1 has a list of commonly used tools and whether they require reference data, and if they do whether it is included. As there is currently no gold standard for brain cell type expression analysis, and developing my own was beyond the scope of this chapter, I restricted the tools I used to ones with an integrated reference dataset for brain cell types. As discussed in 1.3.1, there are some tools which fit this criterion. Of them, I felt that two were best suited to my work, BrainInABlender (Hagenauer et al., 2018) and Xcell (Aran et al., 2017). Both of these tools calculate an enrichment score, rather than estimated proportions of total cell abundance as produced by some other methods such as CibersortX (Newman et al., 2019).

The overall aim of this analysis was to find differences in levels of RNA expression between patients with FTD, and control brains. It also aimed to find if there were clear differences in mechanism between the two different FTD pathologies. Members of the Fratta lab sequenced samples from the UCL brain bank, some were brains from healthy donors, some were from donors who had FTD which stemmed from a mutation in C9orf72, and some from donors with FTD which stemmed from MAPT mutations. In addition to evaluating changes between the types of brain, I also aimed to find whether there were significant changes between brain cell types and determine the utility of including brain cell types as covariates in the differential expression model. If there are substantial changes in cell types between samples it might add biological noise to the differential expression analysis. By accounting for this, truly differentially expressed genes may become clearer.

4.3. Methods

Brains were donated to the UCL brain Bank. Our dataset consisted of 4 healthy brains, 5 brains from patients with FTD with C9 pathology, and 4 brains from patients with FTD with TAU pathology.

Table 4-1. Table of metadata including read depth of samples

Sample Name	Condition	Age	Sex	Reads
CTL_1	CTL	63	M	43323837
CTL_2	CTL	85	M	59456990
CTL_3	CTL	79	F	52218081
CTL_4	CTL	71	M	57055405
C9_1	C9	72	M	46165285
C9_2	C9	68	M	33036357
C9_3	C9	66	F	54544627
C9_4	C9	45	M	64851575
C9_5	C9	74	F	65318616
TAU_1	TAU	75	M	34795333
TAU_2	TAU	68	M	38549539
TAU_3	TAU	74	M	51863825
TAU_4	TAU	72	M	60039216

Alignment and read counting were performed as discussed in chapter 2. Differential expression was performed using DESeq2 (Love et al., 2014). I also normalised reads using DESeq2, and used them to quantify cell type abundance using both BrainInABlender (Hagenauer et al., 2018) and xCell (Aran et al., 2017) with default parameters for both methods.

PCA plots of the BrainInABlender results took used relative levels of all cell types to calculate PCA score.

The formula used when integrating the results of cell type disambiguation into the differential expression formula is:

~ Condition:Astrocyte:Endothelial:Microglia:Mural:Neuron_All:
Neuron_Interneuron:Neuron_Projection:Oligodendrocyte:
Oligodendrocyte_Immature:RBC + Condition

Age and sex are added to the start of the formula in cases where it does not cause the model to fail to converge. What this means is that the model will evaluate how each of the interaction terms interact with one another, and the condition, as well as the effect of the condition alone. Given as changes in relative enrichment in one cell type will be linked to changes in another, this seemed to be a prudent step, and ensured that the model didn't fail to converge.

The full code developed for the differential expression and other analyses can be found here: https://github.com/SethMagnusJarvis/PhDFigureCreation/tree/main/FTDBrain.

4.4. Results

Two causes of frontotemporal dementia are mutations in C9orf72 and Tau. I used sequenced samples from UCL's brain bank consisting of 4 control samples, 5 FTD patients with a C9 mutation and 4 FD patients with a TAU mutation. The mean age was 70 with most people being between 60 and 80. There were 3 women, one control and two FTD C9 mutants. Our objective was both to analyse these samples in order to see the effect which mutations in these genes had on RNA expression, and to evaluate the effectiveness of tools to disambiguate cell types in human brain data and correct for any changes seen in our differential expression results. This was useful both as part of analysis of this dataset, and in preparation of analysis of a much larger cohort of patient brain data which has been sequenced in the US.

4.4.1 Differential expression in TAU and C9 mutants

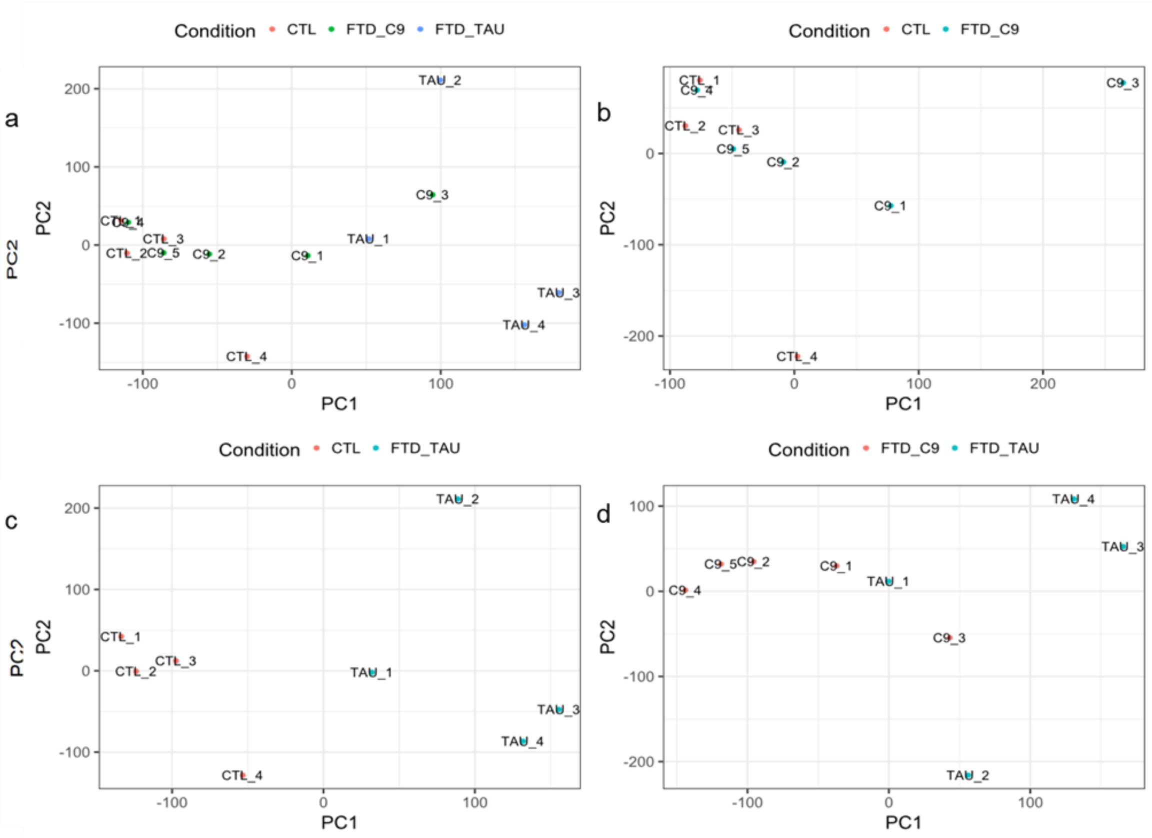

Figure 4-1. PCA of normalised reads per gene in (a) WT vs Mut, (b) WT vs C9, (c) WT vs TAU, and (d) C9 vs TAU. It shows some segregation between the three types of sample, with the most pronounced differences being between TAU and the other samples. Sample C9_3 is more similar to TAU samples than to other C9 samples.

As an initial quality control step, I ran principal component analysis (Figure 4-1) on reads normalised by DESeq2, both examining all samples overall, and each possible comparison (CTL vs C9, CTL vs TAU, C9 vs TAU). I saw the most obvious differences in our TAU mutants, which were distinct from both the FTD C9 patients, and the controls. There was some separation between C9 and control, but it was not as pronounced.

We proceeded to make heatmaps of the genes with most variance in their expression using hierarchical clustering (Figure 4-2). I once again see that TAU mutants cluster together and separately from other samples. There are two C9 samples (C9_1 and C9_3) which seem to be most similar. They seem to be more similar to the main tau cohort than one TAU sample is to the rest of the tau cohort. The C9 mutants show distinct expression patterns compared to the control samples aside from one control sample which clusters in their midst.

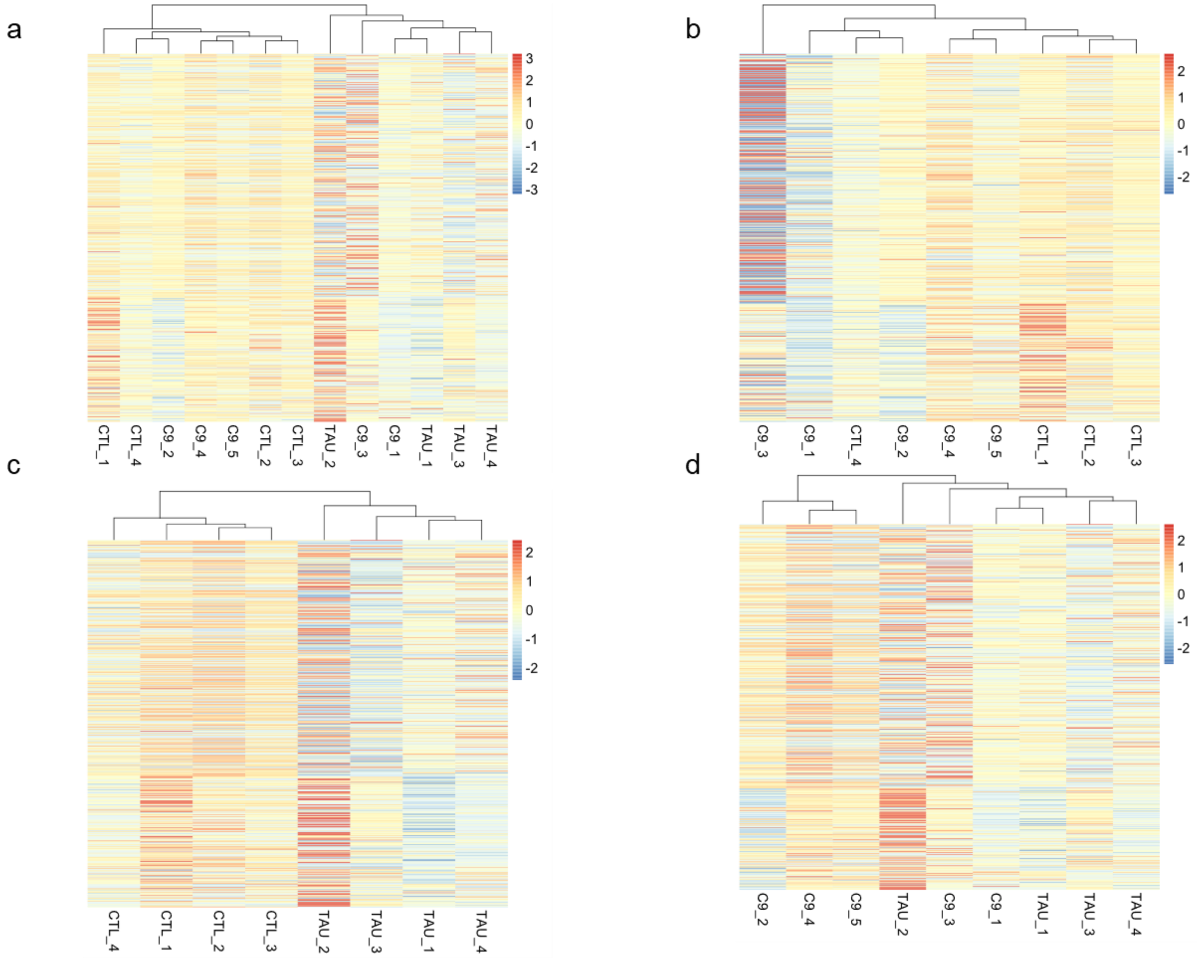

Figure 4-2. Heatmap of normalised expression per gene among the most expressed genes in each dataset. Each line represents the expression of an individual gene using normalised values as produced by DESeq2. Colour coding indicates log of these normalised reads. In (a) All WT vs all mutants, (b) WT vs C9, (c) WT vs TAU, and (d) C9 vs TAU

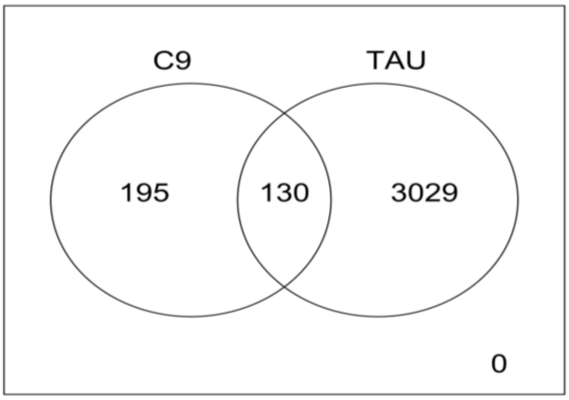

Figure 4-3. Venn diagram comparing genes differentially expressed when samples with C9 and TAU etiology were compared to our control samples. It shows that the TAU samples have 9.7x the number of genes significantly differentially expressed when compared with control and that 40% of genes in which are significantly different in C9 etiology are also significantly different from control in TAU etiology.

Our PCA plots suggested that the TAU mutation had a much stronger transcriptional phenotype compared to C9. I therefore expected to uncover more differentially expressed genes under this condition when compared to the control samples than C9. This was found to be correct. Our Venn diagram showing the overlap of the two samples (Figure 4-3) showed that TAU had far more genes differentially expressed than C9, and that of the genes found to be significantly differentially expressed from control in C9, two thirds were also differentially expressed in TAU. Figure 4-4 shows the volcano plots of the differential expression results in the comparisons and further demonstrate the differences in size of mutation in the comparisons.

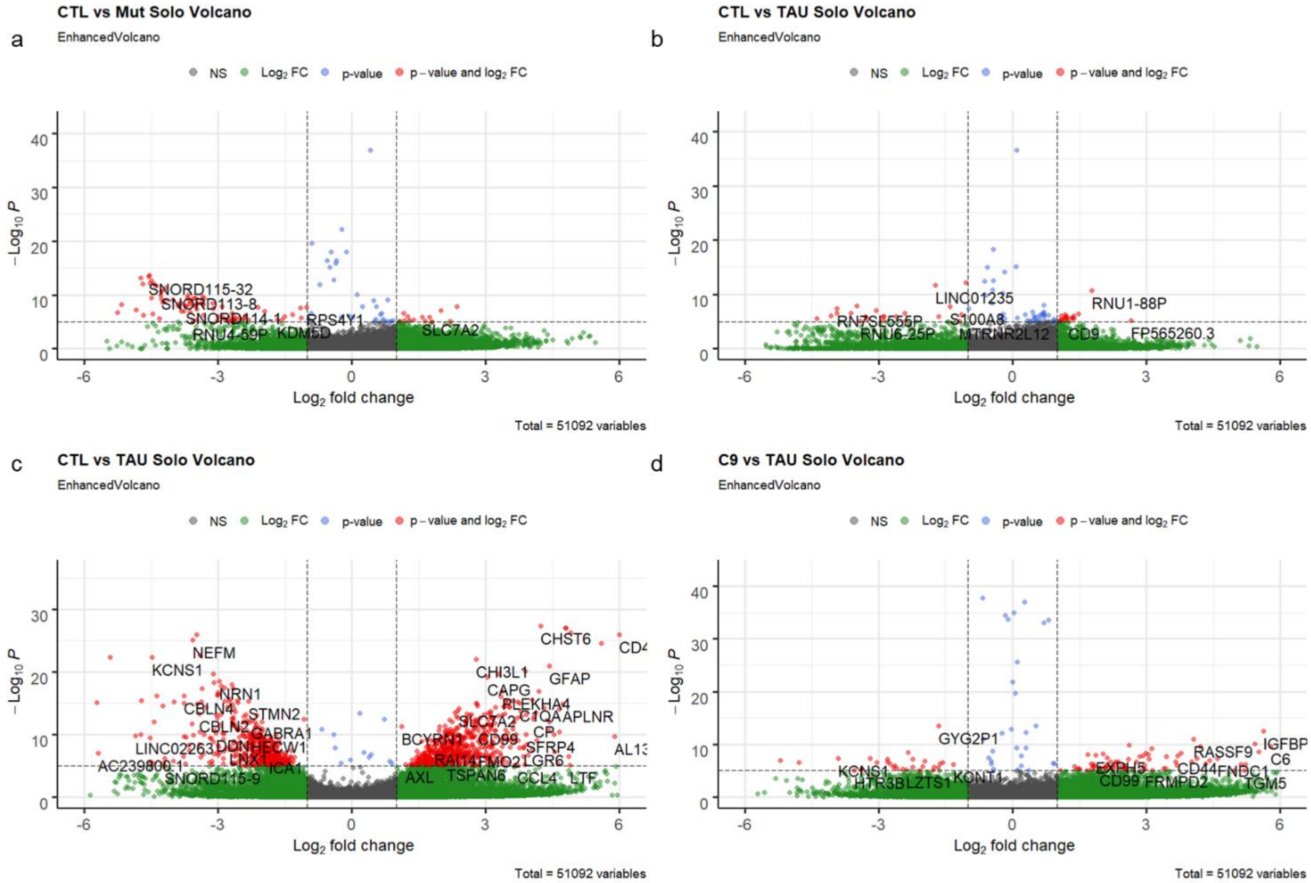

Figure 4-4. Volcano plots of differential expression in all 4 comparisons of samples in (a) All WT vs all mutants, (b) WT vs C9, (c) WT vs TAU, and (d) C9 vs TAU. Samples were colored depending on whether they met certain thresholds of significance met (log2FC >abs(2), p-value <=10^{-6}), they are colored green if they exclusively meet the fold change threshold, blue if they exclusively meet the significance threshold, and red if they meet both thresholds.

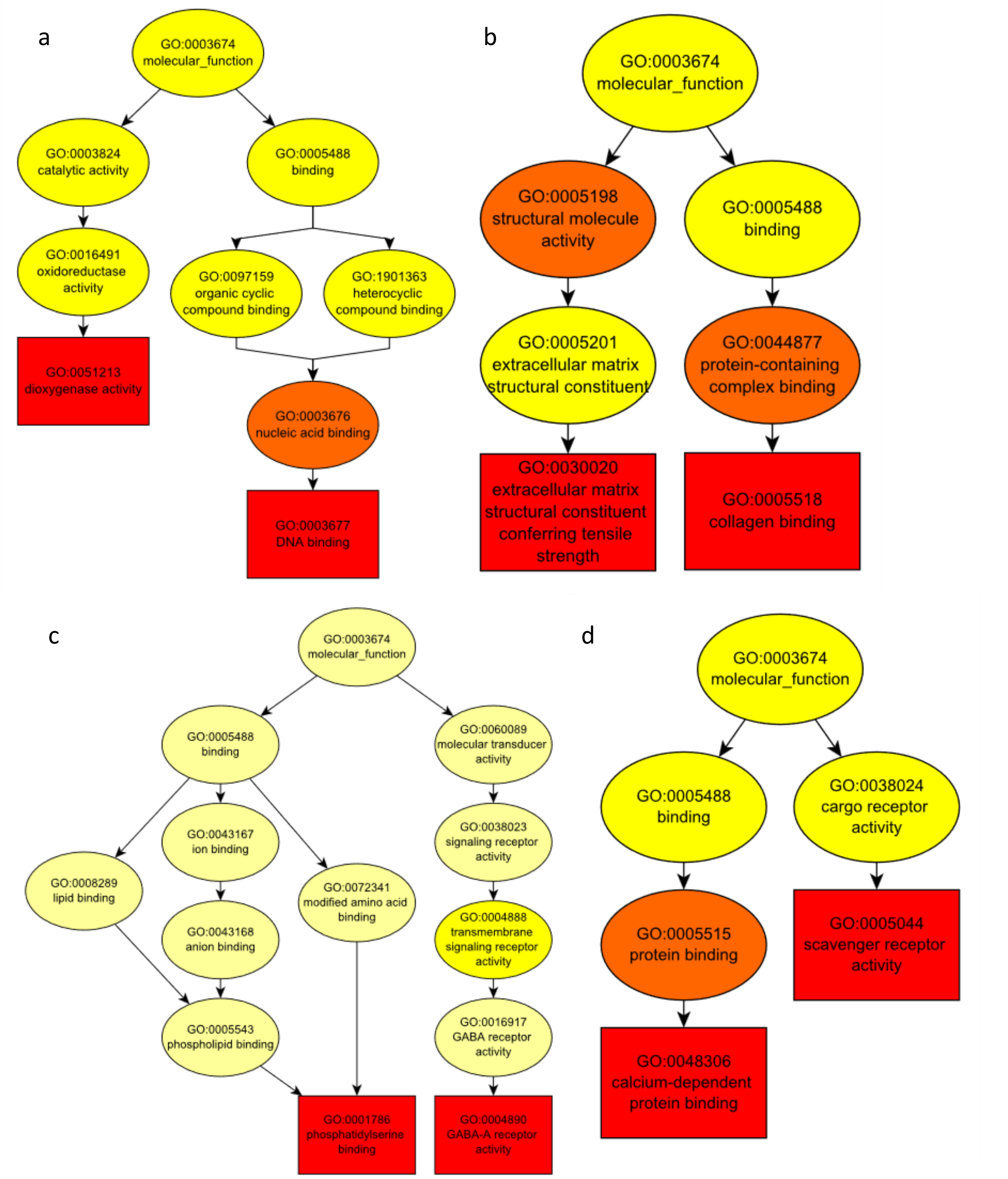
a
GO:0003674 molecular_function
GO:0003824 catalytic activity
GO:0005488 binding
GO:0016491 oxidoreductase activity
GO:0097159 organic cyclic compound binding
GO:1901363 heterocyclic compound binding
GO:0051213 dioxygenase activity
GO:0003676 nucleic acid binding
GO:0003677 DNA binding
b
GO:0003674 molecular_function
GO:0005198 structural molecule activity
GO:0005488 binding
GO:0005201 extracellular matrix structural constituent
GO:0044877 protein-containing complex binding
GO:0030020 extracellular matrix structural constituent conferring tensile strength
GO:0005518 collagen binding
c
GO:0003674 molecular_function
GO:0005488 binding
GO:0060089 molecular transducer activity
GO:0043167 ion binding
GO:0038023 signaling receptor activity
GO:0008289 lipid binding
GO:0072341 modified amino acid binding
GO:0043168 anion binding
GO:0004888 transmembrane signaling receptor activity
GO:0005543 phospholipid binding
GO:0016917 GABA receptor activity
GO:0001786 phosphatidylserine binding
GO:0004890 GABA-A receptor activity
d
GO:0003674 molecular_function
GO:0005488 binding
GO:0038024 cargo receptor activity
GO:0005515 protein binding
GO:0005044 scavenger receptor activity
GO:0048306 calcium-dependent protein binding

Figure 4-5. Most significant molecular function GO terms in significantly differentially expressed genes using Kolmogorov-Smirnov in (a) WT vs Mut, (b) WT vs C9, (c) WT vs TAU, and (d) C9 vs TAU. This is a subgraph induced by the most significant GO terms identified by the fisher exact algorithm for scoring GO terms for enrichment. Rectangles indicate the most significant terms. Node colour represents the relative significance, ranging from dark red (most significant) to bright yellow (least significant). For each node, some basic information is displayed. The first two lines show the GO identifier and a trimmed GO name

In order to see what processes the differentially expressed genes were involved in, I calculated the GO term enrichments for each of our datasets (Figure 4-5). The most significant GO terms were linked to deoxygenase activity and DNA-binding in our combination of control to both FTD groups at once. In CTL vs C9 the most significant GO terms were linked to the extracellular matrix and collagen binding. In CTL vs TAU the most significant GO terms were linked to Phosphatidylserine binding and GABA receptor activation. Finally, in our comparison of C9 Mutants to TAU mutants, our most significant GO terms were linked to calcium dependent protein binding and scavenger receptor activation.

4.4.2 Cell type enrichment

Brains have relatively heterogeneous cell types, and so correcting for differences in cell types can be useful in the analysis. I used the R packages BrainInABlender (BIAB) (Hagenauer et al., 2018) and XCell (Aran et al., 2017) to examine changes in cell types between samples, both to identify differences in cell types, and to improve differential expression accuracy.

The first step was to create a PCA plot of the samples to see whether there were obvious differences in levels of enrichment between the different etiologies. Figure 4-6 shows that the grouping of samples is similar to that observed previously; TAU mutants seem to share the least similarities with the other two samples. Figure 4-7 shows boxplots of each individual cell type, it shows that the only statistically significant differences in relative enrichment of cell types are between the TAU aetiologies and control samples. The outlier observed in both the gene level and cell type enrichment level PCAs in C9 (C9_3) may be due to changes in relative levels in oligodendrocytes in that sample. This is further reinforced in Figure 4-8 – a heatmap of relative enrichment of each cell type by sample. If oligodendrocytes are truly enriched in this sample compared to others, it would mean that genes related to oligodendrocytes are more heavily enriched in that sample which would cause it to be segregated from other samples in PCA analysis. While we do not have full information about the disease courses of patients who were analysed, as stated previously, there does not seem to be a substantial difference between the rate of disease progression of patients with tau aetiology compared to those with C9.

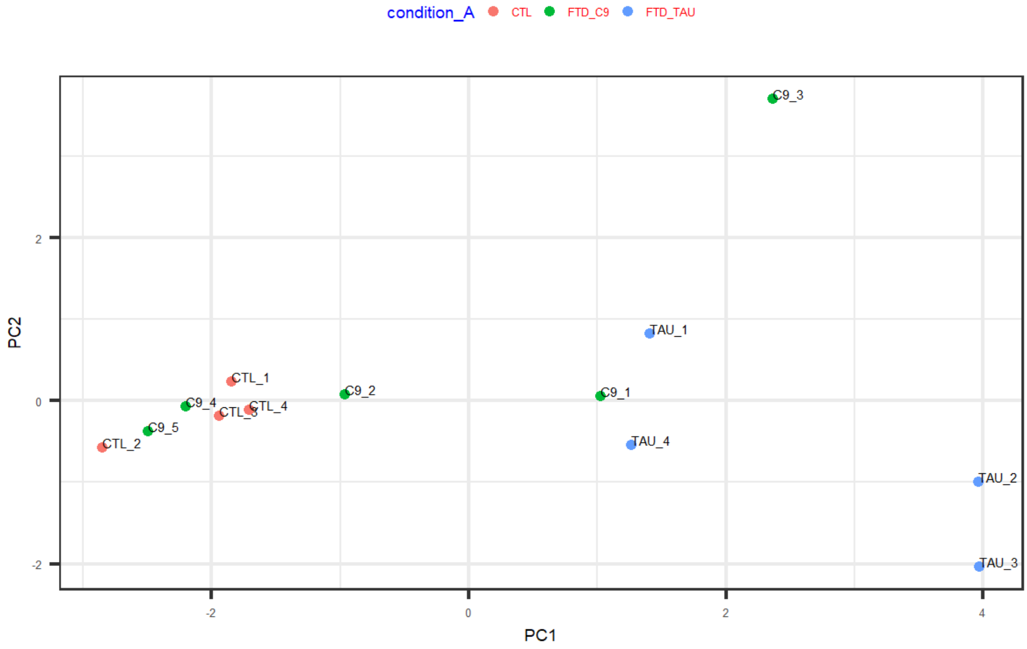

Figure 4-6. PCA plot of relative enrichment of cell types in each sample compared to what is expected based on expression of cellular marker genes from brain in a blender. This shows a similar separation to the PCA plot of normalised expression, with patients with a TAU etiology having the clearest segregation from other samples.

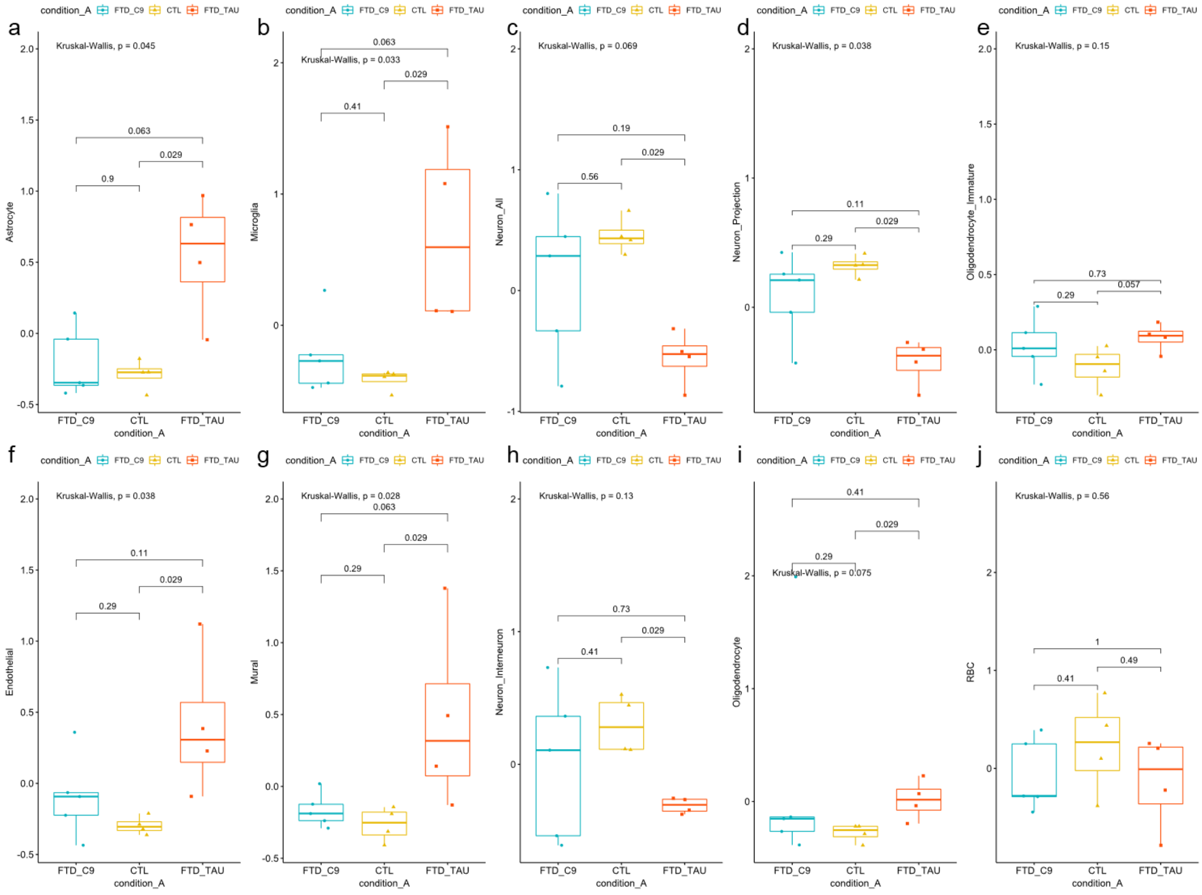

Figure 4-7. Boxplots of relative enrichment of cell types from BIAB in (A) Astrocytes, (B) Endothelial, (C) Microglial, (D) All Neurons, (E) Interneurons, (F) Neuron Projections, (G) Oligodendrocytes, (H) Immature Oligodendrocytes,

and (J) Red blood cells. A Kruskal-Wallis test for significance was performed among the overall samples, both to identify whether there is a significant difference in the levels of enrichment overall, and if there is a significant difference between any two types of samples.

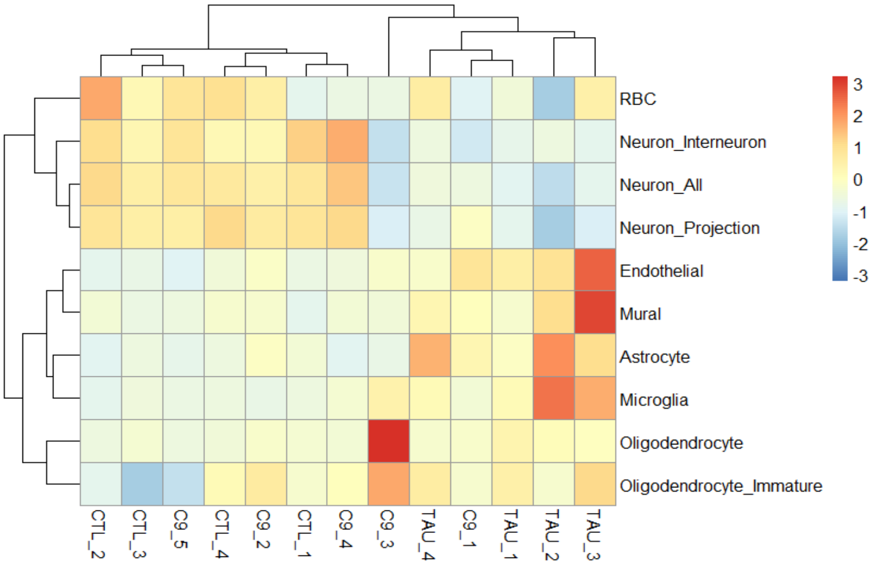

Figure 4-8. Heatmap of relative enrichment of cell types from BrainInABlender across the different samples and conditions. Colour gradient represents the relative level of each cell according to BIAB.

Overall, however, the only statistically significant differences in cell types were observed between the control samples and the TAU mutants (Figure 4-7). The C9 Samples did appear to have levels of cell type specific enrichment and depletion which were more similar to control than TAU. They were, however, different enough from the control samples that their levels of cell types were not found to be significantly different from TAU.

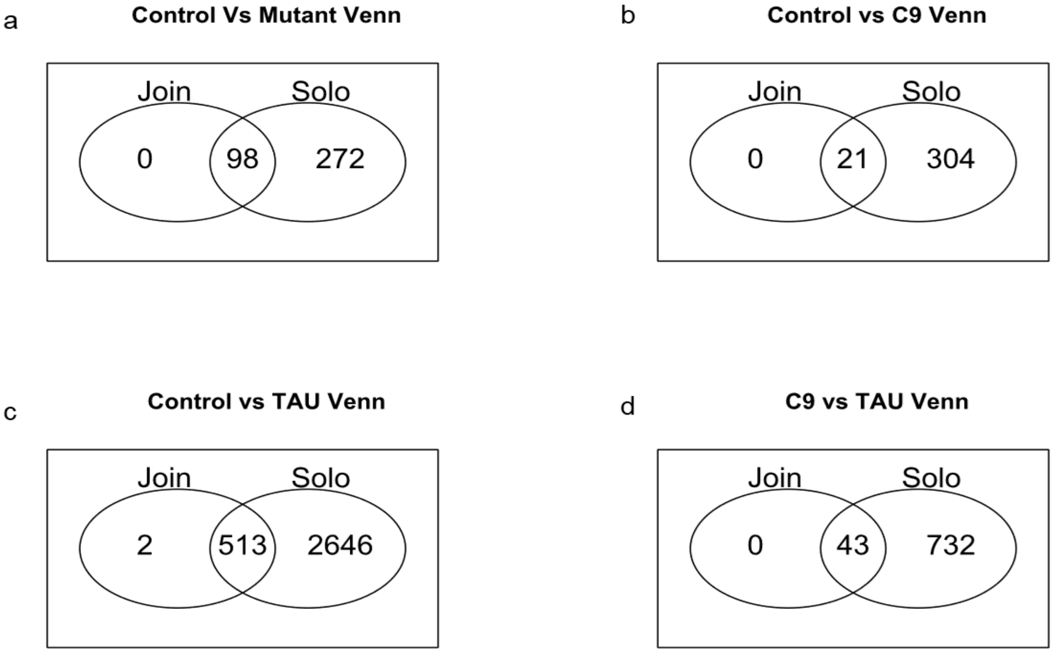

Figure 4-9. Venn diagrams overlap of differential expression results. "Solo" refers to using exclusively age, sex and condition as covariate, and "Join" to the analysis that also includes the interaction between the levels of all cell types from BIAB and the condition, in addition to the age, sex, and condition in (a) WT vs Mut, (b) WT vs C9, (c) WT vs TAU, and (d) C9 vs TAU

Given the variability present in the cell type makeup of brain tissue, I hoped to include the interaction of the types of cells with the condition as a covariate in the differential expression model. When this interaction was included, I found that the number of genes found to be significantly differentially expressed in all comparisons Figure 4-9). It is reassuring that almost all of the genes which are found to be significantly differentially expressed when correcting for cell type are also found to be differentially expressed when not correcting for it. When creating volcano plots (Figure 4-10). I found that including cell types as interaction terms seems to cause DESeq2 to substantially overestimate fold changes in genes.

The most significant GO terms when using these interaction terms can be seen in Figure 4-11. Many GO terms changes between the two samples. The terms appeared to tend towards being less specific when the BIAB results were included than in Figure 4-5. This is possibly due to the number of genes found to be significant decreasing which would make it harder to find more specific effect sizes as there would be fewer genes linked to that pathway. This included the WT vs C9 samples changing from having specific terms related to the extracellular matrix and collagen, to more general protein binding and translation initiation factor activity. Similarly, the WT vs TAU no longer found GABA-A receptor and phosphatidylserine binding to be significant. Instead, it found that the more general term carbohydrate binding was found to be significant along with extracellular matrix binding.

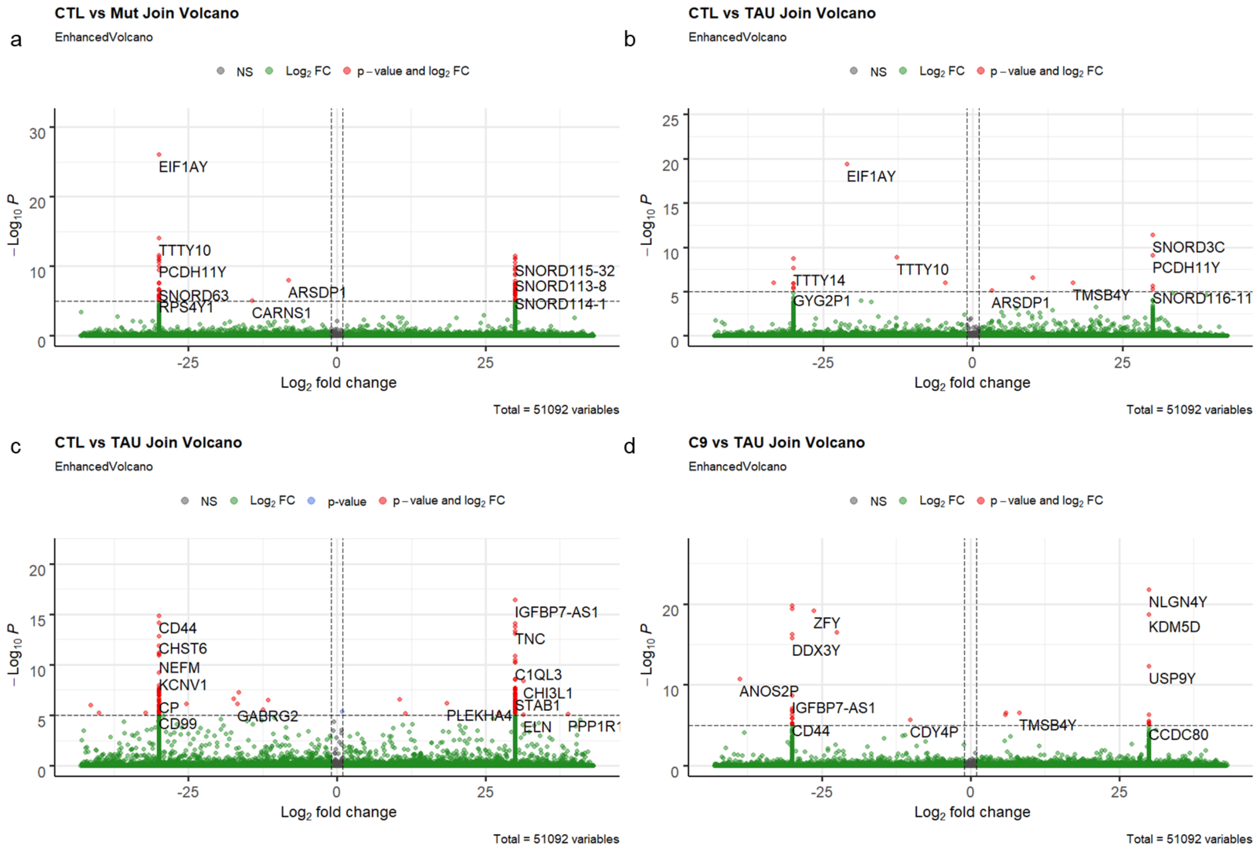

Figure 4-10. Volcano plots of differential expression results, including BIAB cell type enrichment as a covariate in (a) WT vs Mut, (b) WT vs C9, (c) WT vs TAU, and (d) C9 vs TAU. Samples were colored depending on whether they met certain thresholds of significance met (log2FC >abs(2), p-value <=10^{-6}), they are colored green if they exclusively meet the fold change threshold, blue if they exclusively meet the significance threshold, and red if they meet both thresholds. Fold changes appear to have partially converged around certain locations. The reason appeared to be related to how they were integrated into differential expression analysis. Future work should work on improving the integration of brain cell types into differential expression analysis.

Table 4-2. Table showing number of genes found to be significantly differentially expressed in each sample when relative levels of enrichment of each cell type are included in the formula for differential expression.

Names	Significant genes in WT vs Mut	Significant genes in WT vs C9	Significant genes in WT vs TAU	Significant genes in C9 vs TAU
Astrocyte	4097	520	3507	1346
Endothelial	2111	2190	3396	1649
Microglia	5301	1349	4159	1843
Mural	1832	705	3765	1282
Neuron_All	6477	3704	2872	3631
Neuron_Interneuron	5441	3787	3300	3933
Neuron_Projection	6694	3101	4378	3443
Oligodendrocyte	2689	2217	3324	3071
Oligodendrocyte_Immature	1515	2054	3499	1476
Red Blood Cell	733	450	3395	1239

When correcting for each cell type individually the results are more difficult to interpret, as in almost all cases I see a substantial increase in the number of genes which are found to be differentially expressed (Table 4-2). It is possible that many of these genes may not be differentially expressed. The broad variance seen in cell types may cause these changes.

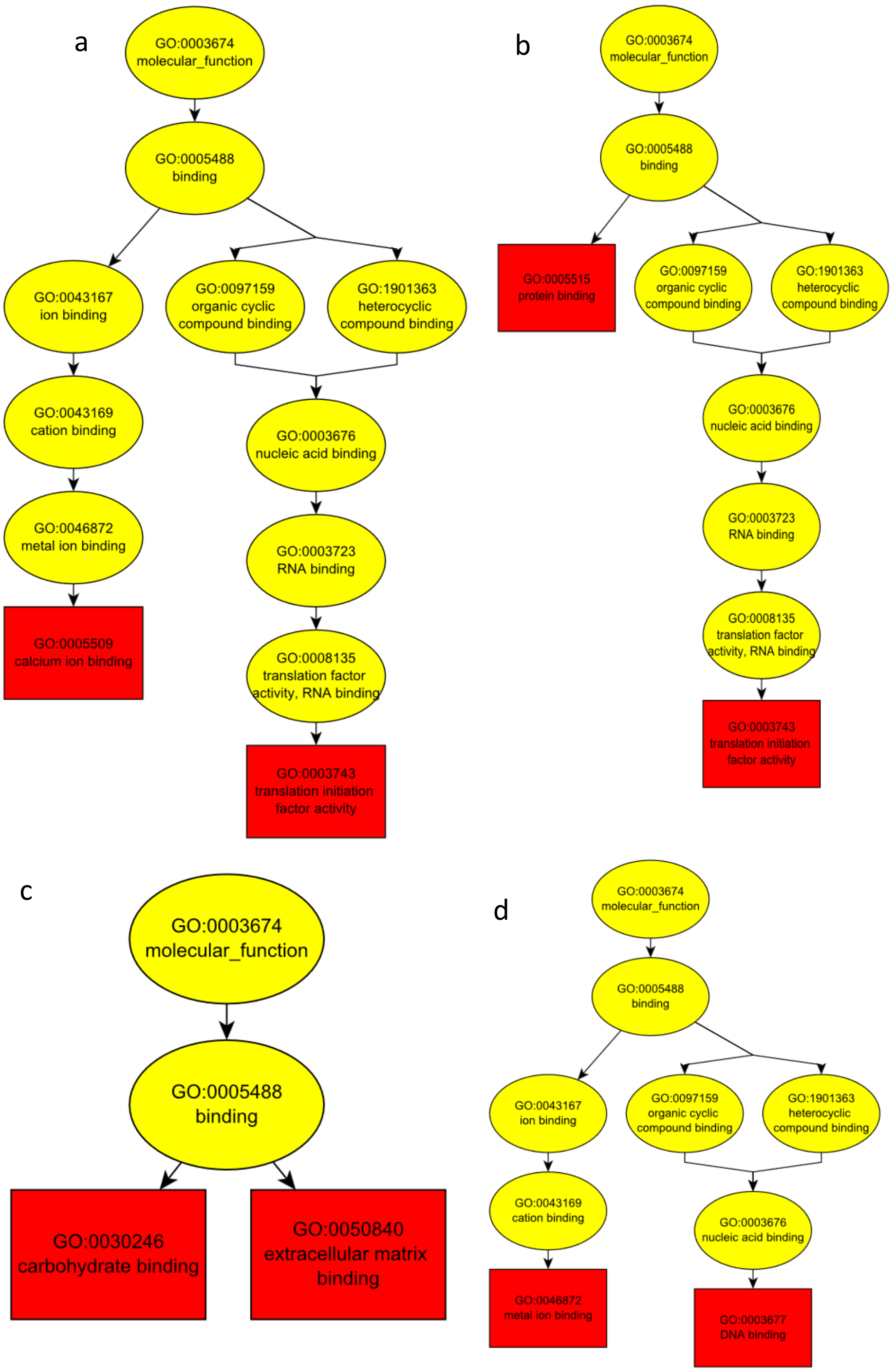
a
GO:0003674 molecular_function
GO:0005488 binding
GO:0043167 ion binding
GO:0097159 organic cyclic compound binding
GO:1901363 heterocyclic compound binding
GO:0043169 cation binding
GO:0003676 nucleic acid binding
GO:0046872 metal ion binding
GO:0003723 RNA binding
GO:0005509 calcium ion binding
GO:0008135 translation factor activity, RNA binding
GO:0003743 translation initiation factor activity
b
GO:0003674 molecular_function
GO:0005488 binding
GO:0005515 protein binding
GO:0097159 organic cyclic compound binding
GO:1901363 heterocyclic compound binding
GO:0003676 nucleic acid binding
GO:0003723 RNA binding
GO:0008135 translation factor activity, RNA binding
GO:0003743 translation initiation factor activity
c
GO:0003674 molecular_function
GO:0005488 binding
GO:0030246 carbohydrate binding
GO:0050840 extracellular matrix binding
d
GO:0003674 molecular_function
GO:0005488 binding
GO:0043167 ion binding
GO:0097159 organic cyclic compound binding
GO:1901363 heterocyclic compound binding
GO:0043169 cation binding
GO:0003676 nucleic acid binding
GO:0046872 metal ion binding
GO:0003677 DNA binding

Figure 4-11. Most significant molecular function GO terms in significantly differentially expressed genes the interaction between the levels of all cell types from BIAB and the condition using Kolmogorov-Smirnov in (a) WT vs Mut, (b) WT vs C9, (c) WT vs TAU, and (d) C9 vs TAU. This is a subgraph induced by the most significant GO terms identified by the fisher exact algorithm for scoring GO terms for enrichment. Rectangles indicate the most

significant terms. Node colour represents the relative significance, ranging from dark red (most significant) to bright yellow (least significant). For each node, some basic information is displayed. The first two lines show the GO identifier and a trimmed GO name

When I used X-cell estimates of the two brain cell types included by default (astrocytes and neurons), only neurons showed significant differences in abundance. The difference was also only significant between CTL and TAU samples. When I broadened this to include all cell types, adipocytes, Fibroblasts, and neurons were the only cell types with significant differences. Adipocytes and fibroblasts had significant differences between TAU and both CTL and C9 samples. Given as neither of these cells should have a significant presence in the brain, I suspect a technical artifact. I would hypothesise that the markers which are used for these cells may overlap with other products of the CNS such as those produced for myelination, and other immune cells. This results in changes in these cells being observed where they may not be truly present.

I re-ran differential expression correcting for the proportions of neurons from XCell. The overlap with differential expression that was not corrected for these cell types was lower than expected (Figure 4-12).

When comparing significantly differentially expressed genes when correcting for the cell type proportions from BIAB against those obtained when correcting for levels of neurons from XCell, I see that while BIAB does find fewer genes to be significantly differentially expressed. Most of the genes found significant when correcting for BIAB cell types are also significant when correcting for XCell's neuron correction results (Figure 4-13).

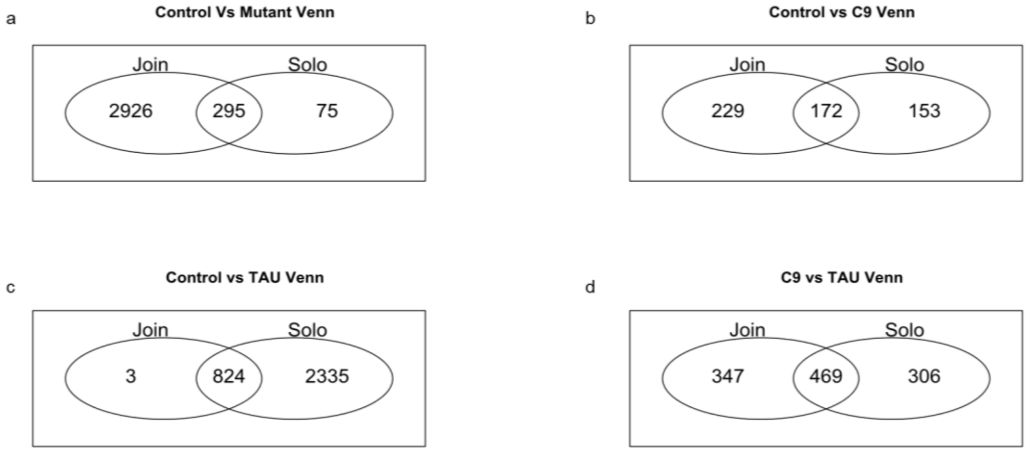

Figure 4-12. Venn diagrams overlap of differential expression results where Solo is using exclusively age, sex and condition as covariate, and join including levels of Neurons suggested by XCell. (a) WT vs Mut, (b) WT vs C9, (c) WT vs TAU, and (d) C9 vs TAU

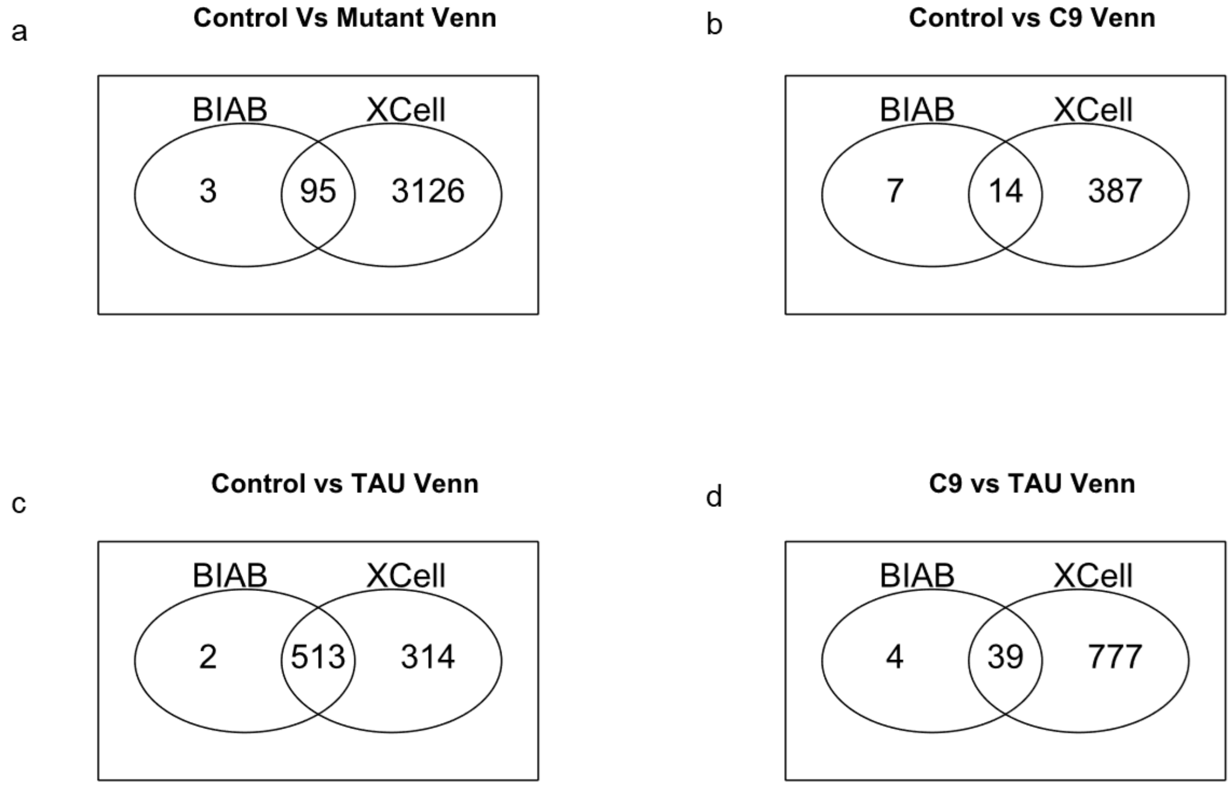

Figure 4-13. Comparisons of significantly differentially expressed genes when correcting for all cell types from Brain in a blender, vs when correcting for Neurons from XCell. (a) WT vs Mut, (b) WT vs C9, (c) WT vs TAU, and (d) C9 vs TAU

4.5. Discussion

I had two main aims when starting this analysis. The first was to uncover changes in gene expression that are related to FTD. The second was to evaluate cell type disambiguation methods. My analysis of differential expression looked at the effects of mutations against control, both in aggregate, and separated by mutation. When investigating cell type disambiguation, I searched for potential tools, and evaluated methods for integrating the results into differential expression analysis.

I believe my work is novel as while some analysis has been performed comparing mutants to wildtype, it is not hugely common. Existing studies such as (Dickson et al., 2019; Prudencio et al., 2015) also only seek to compare FTD patients to control patients, and not to other mutants. Brain cell type disambiguation is also a relatively novel field, particularly compared to more established analysis in cancer. Other studies using a larger portion of the same brain database performed by collaborators with our lab (Humphrey et al., 2022) have since performed brain type disambiguation on larger FTD datasets, but I believe my work was the first to attempt this analysis in FTD. While sample sizes are small, I may have found some useful insights into some of the mechanisms of how these FTD aetiologies cause disease. It will be very important to confirm findings with a larger dataset, either by further RNA-sequencing or quantitative RTPCR of interesting DEGs.

The most obvious result of my differential expression analysis was that mutations in TAU cause more changes in expression than mutations in C9. This was immediately obvious in PCA plots, and further backed up by the differential expression plots where TAU brains as a collective had 10x the number of genes differentially expressed as did C9 mutants. This was unexpected as, while the two pathologies do often have differing clinical presentation, they have similar gross pathologic features, not consistently differing in severity or progression (Van Langenhove et al., 2013).

In the initial analysis of patients with C9 aetiology, in the absence of cell type correction, we did find that there were significant changes in GO terms linked to the extracellular matrix and collagen binding. This finding supports some of the results recently published by the Isaacs lab who found that novel knock-in mice expressing either poly(GR) or poly(PR) dipeptide repeat proteins (DPRs) had increased levels of extracellular matrix proteins identified through quantitative proteomics (Milito et al., 2023). This study also confirmed an increase of extracellular matrix protein gene expression in laser capture micro dissected spinal motor neurons from C9 patients (Highley et al.,2014; Milito et al., 2023). This convergence of different datasets gives confidence in our RNA-seq findings despite small sample size. However, these changes were no longer apparent after correcting for relative cell type enrichment. One possible explanation is that this extracellular matrix signal is driven by reactive gliosis as it has been shown that reactive astrocytes increase extracellular matrix gene expression (Ziff et al., 2022). This would lessen the disease specific relevance for C9 FTD/ALS, but given the emergence of data from several sources this pathway still remains worthy of further follow-up.

Extracellular matrix binding was also found to be one of the most significant GO terms in the comparison of control samples to patients with a TAU aetiology when correcting for cell types. This highlights potential links between the mechanisms of the aetiologies, and given the results of the Isaacs lab (Milito et al., 2023) may be worth further investigation.

Differences the magnitude of genes found to be differentially expressed may be rooted in differences in either proportion or expression in certain cell types in the brain tissue. The only statistically significant changes between datasets in any cell types were between TAU and our control brain dataset. This suggests either changes in the number of certain cell types or changes in expression of certain cell types occur more within TAU samples than within C9. Every cell type aside from immature oligodendrocytes and red blood cells had significant differences between TAU and control. Microglial cells have been linked elsewhere to tau pathology in Alzheimer’s disease (Hansen et al., 2018) so some of the relative enrichments in other changes may be linked to changes in microglia. The lack of significant differences between cell types in C9 samples and TAU samples does suggest that there are some slight changes within C9, but they are small enough that I did not have the statistical power to calculate whether this was significant. All statistically significant differences between samples were between TAU and control.

As discussed in the introduction, I settled on using BrainInABlender as my primary method for cell type disambiguation, while also employing complementary estimates generated by XCell. As enrichment is a relative rather than an absolute measure, it can make comparison between studies difficult. It also stops goodness of fit and error measures from being calculated. This can make biological interpretation difficult without some direct analysis of cell types, and defined hypothesis about what is expected to occur in cell types. However, my goal was to examine differences between samples, so calculating exact cell proportions was not essential in this analysis. The key advantage of these two methods is that they both provide their own reference datasets of cell type markers. There are methods which estimate cell fractions out of the total population, such as Cibersortx (Newman et al., 2019), and dtangle (Hunt et al., 2019), but all of them require reference datasets. While some of them do have reference datasets included, they tend to be for cells related to cancer rather than brain cells.

My analysis was intended as a proof of concept and therefore I felt that using the methods for relative enrichment was acceptable to investigate relative changes between samples, and to test normalisation in our differential expression models.

To the best of my knowledge, relatively little direct work in changes of cell type proportions has been performed in ALS-FTD. Tau pathology has been linked to changes in how glia - and particularly strongly microglia act in Alzheimer's Disease and ALS-FTD(Brelstaff et al., 2021; Kahlson & Colodner, 2015; Perea et al., 2020). Some other comparisons of changes in relative enrichment of cell types in C9 has been performed, which suggests that there is some level of enrichment – however, this was performed in ALS patients, and used a different methodology for selecting markers of enrichment (Humphrey et al., 2022). This enrichment was also not specific to C9, with all ALS patients showing similar changes.

I believe there are two hypotheses for why C9 does not appear to have clear changes in cell type composition compared to wildtype controls:

- There are differences between C9 and Control, however these changes do not reach significance either because they are small changes and our sample size is relatively small, or because the variability within our samples means that we cannot clearly observe them.

- Neurodegeneration may result in overall changes to levels of expression in all cell types. This would result in relatively few changes appearing in the relative levels of expression even if the absolute levels are reduced.

I do not think that there is enough data to support any one of these hypotheses above others. When we integrate cell type information into the differential expression analysis in C9 patients we find that the most significant GO terms appear to be far broader. This included no-longer seeing the more specific terms collagen binding and the extracellular matrix which have been demonstrated elsewhere. This demonstrates that caution needs to be taken when attempting to integrate cell type changes into differential expression analysis. It may lead to disguising true effects.

Prior work has found differences in markers of cell types between FTD patients with C9 aetiology and those without (Dickson et al., 2019). The core difference between this work and my work is they had samples from more patients (24 control, 34 C9FTD patients, 44 non-C9 FTD patients). Due to this, I feel it is more likely that the true changes cannot be distinguished from natural variations between patients due to our relatively small sample size.

In terms of integrating the changes as covariates in differential expression models, while the ideal would be to use each cell type as a separate covariate, this consistently resulted in errors. I believe this is because in most cases, there were either no differences in some cell types, or all of the differences in one or more cell types that were present were already accounted for by the condition. I therefore added the relative enrichment of cell types to the differential expression model as a combined interaction term with relation to the condition. Including this interaction term did cause substantial irregularities in the fold changes calculated by DESeq2. It may be worth investigating why these large changes in fold changes occur, but I consider the most important future work arising to be investigating changes in which genes were deemed to be significantly differentially expressed by DESeq2.

Adding one cell type at a time as a potential confounding factor to individual differential expression models sems to add more noise than it does provide utility. When I included all of the cell estimates from BIAB as a single covariate however, I did find a substantial reduction in the number of genes found to be significantly differentially expressed. It would not be practical to validate whether or not all of the changes are truly not differentially expressed, however, selecting several genes at varied levels of expression as a representative sample would be a worthwhile method of validation.

Overall, several useful insights were gained from this research. First, there is substantial overlap in the mechanisms of TAU and C9 pathology. This is not surprising as the end result of the mutations is similar. I also found that TAU seems to be a mutation with stronger transcriptional consequences in bulk RNA-seq data, although it is not currently clear if this is due to changes in cell type expression.

These larger changes are also reflected in changes of relative enrichment of cell types. The cell type plots only show significant differences in enrichment between TAU mutations and the control. This is an interesting finding and is worth further investigation to see if this is an artefact of the large number of differentially expressed genes or a true difference. I also found that adding the enrichment of all cell types from BrainInABlender to differential expression models may help to reduce biological noise and make finding true differences easier. Until the genes which are only found to be differentially expressed when accounting for cell type in bulk samples are validated, future research should focus on some of the genes which are found to be significantly differentially expressed both when accounting and when not accounting for this cell type information.

4.5.1 4.4.1 Future work

There are three obvious questions that require further validation that have come out of my research into our FTD brain dataset:

- Why does TAU seem to produce many more changes in differentially expressed genes?
- Are there real differences in cell types between datasets?
- How accurate are the changes in genes found to be differentially expressed when including the cell type enrichment?

Previous research into the individual mutations has shown that C9 and TAU pathology in FTD does act via some different pathways(Balendra & Isaacs, 2018; Sha et al., 2012; Young et al., 2021), with C9 bearing some similarities to sporadic FTD cases(Conlon et al., 2018). The differences do not translate into changes in the severity or rate of progression of disease (Van Langenhove et al., 2013). This meant that TAU having an order of magnitude more DEGs was surprising. The first point to investigate would be why this is.

A possible reason for the differences in results may have been the changes in relative enrichment of cell types. C9 pathology did not have any cell types that seemed to be significantly different from control, while TAU had several significant differences in cell types. When including the relative enrichment of brain cell types from BrainInABlender (BIAB), the number of DEGs found in TAU seemed to be much more similar to the numbers found in C9 pathology. This suggests that the differences in the number of DEGs may be linked to the differences in cell types. One initial method of validation may be single cell RNA-seq. It would be ideal to sequence several of the cell types which BIAB suggests show differences in expression. This would allow us to see both whether there are any overall changes in expression, and would allow for clarification of true DEGs. If single cell RNA-seq in human brains is not feasible, then investigating whether there are similar changes observed in mouse models would provide some insight into the truth of our results.

We have performed some initial validation through our C9 mouse dataset analysed in chapter 5. This analysed C9 mice which had a repeat expansion inserted, causing mice to express one of two proteins linked to pathogenesis. In our mutant mice we did not find any significant changes to the proportion of brain cell types between our samples and their wildtype controls. While more validation is required, this finding may be particularly interesting when combined with validation of a similar TAU mouse mutant dataset.

Depending on the results I see other questions would arise. If there are substantial changes in cell types in TAU, it would be ideal to see which of the genes are truly differentially expressed in TAU and how they overlap with C9. It would also be useful to investigate which of the cell types are truly changed and how accurate BIAB is on our dataset. If there are no substantial changes in expression in cell types then further investigation into both the results of the cell type enrichment and differences in the number of differentially expressed genes between C9 and TAU would be necessary.

In a broader context, further development of tools to analyse cell types in brains would be useful. The relative levels of enrichment and depletion of cell types require validation. The reasons for the anomalies observed in fold changes observed when integrating levels of cell types needs to be investigated. Depending on the results of the validation, investigation into the specific genes which are found to be differentially expressed when cell types are integrated may also produce useful insights into pathways which are changed within patients.

The creation of a tool which could easily use multiple types of reference data, and multiple algorithms for deconvolution, along with potentially integration of differential expression analysis accounting for changes in cell types would be useful. The tools which I found which specialised in brain cell type deconvolution tended to produce relative rather than absolute brain cell type deconvolution results, and do now allow for selection of the specific types of tissues from which the RNA-seq data was produced. Given the findings of (Sutton et al., 2022), a tool which had multiple types of neuronal data from several different sources which could be easily selected would make analysis of cell type deconvolution in neuronal data more accurate and accessible. Adding a method to then easily integrate these results into differential expression analysis has the potential to substantially improve analysis of bulk RNA-seq data, and help produce reliable results which would better improve our understanding of changes which occur in the datasets of interest.

5. Analysis of C9orf72 repeat expansion in mice

5.1. Publication

This work has yet to be published.

5.2. Introduction

A GGGGCC (G4C2) hexanucleotide repeat expansion in *C9orf72* gene is the most common genetic cause of FTD and ALS (Majounie et al., 2012). There are currently three non-mutually exclusive proposed mechanisms for how this expansion causes disease: presence of sense and antisense RNA foci, reduced expression of *C9orf72* gene, and presence of dipeptide repeat (DPR) proteins (Zang et al., 2018; Gendron et al., 2014).

Repeat associated non-ATG (RAN) translation of *C9orf72* occurs in all sense and antisense reading frames, leading to six dipeptide repeat proteins: glycine–alanine (poly-GA), glycine–proline (poly-GP) and glycine–arginine (poly-GR) from sense repeat transcripts, and proline–glycine (poly-PG), proline–arginine (poly-PR) and proline–alanine (poly-PA) from antisense repeats. Two repeats in particular, poly-GR and poly-PR have both been linked to pathogenesis (Kwon et al., 2014; Mizielinska et al., 2014).

In order to generate physiological models of C9FTD/ALS and understand the effect of DPR at endogenous levels, the Isaacs lab have successfully generated novel DPR knock-in mouse models in which patient-length DPRs are driven by the endogenous mouse *C9orf72* promoter. They have inserted 400 codon optimised repeats of GR and PR directly after the ATG start codon of the mouse *C9orf72,* generating *C9orf72* knock-in mouse models that express physiological levels of DPRs. Codon optimisation was performed to minimise repetition and guanine-cytosine content as much as possible. The respective mutants are henceforth known as (GR)400, and (PR)400. Both mice develop phenotypes consistent with adult-onset neurodegeneration (Milito et al., 2023). To establish a detailed transcriptional characterisation, we have sequenced both wildtype and mutant samples at 6 and 12 months old.

The goal of this chapter was to evaluate the effect of poly-GR and poly-PR DPRs on RNA-expression when expressed in mice. I sought to compare the mutant mice to their relevant controls and to one another. I also hoped to evaluate whether these DPRs caused changes in relative proportions of cell types within the brain.

5.3. Methods

DPR-specific mouse models were produced by inserting transgene vectors with 400 PR/GR repeats immediately after and in frame of the ATG start codon in mouse *C9orf72* using CRISPR/Cas9 technology. Spinal and brain extraction was performed. The paper establishing these mice is under review as of time of writing. Meso Scale Discovery (MSD) – a form of plate-based assay which uses antibodies to detect molecules of interest (Burguillos, 2013) - was performed to ensure that models expressed their respective DPRs in the spinal cord. Whole-brain RNA extraction was performed and at 6 and 12 months using miRNeasy Micro Kit (Qiagen), this was subsequently sequenced using Illumina NovaSeq. The breeding of mice, culling and extraction were performed by Carmelo Milioto. Further information regarding these mice can be found in (Milito et al., 2023). Read depth was performed at an average of 70,000,000 reads; full per sample information can be found in Table 5-1. The number of samples were chosen as they are a relatively standard number, and have identified significant and relevant differences in other RNA-seq experiments. No power calculations were carried out prior to these experiments.

Alignment, read counting, differential expression analysis, and GO term analysis were performed as discussed in chapter 2. The full code for the specifics of how I performed differential expression, GO term analysis, and figure creation can be found on my GitHub (https://github.com/SethMagnusJarvis/PhDFigureCreation/tree/main/AdrianC9).

Differential expression was performed using DESeq2 as described in section 2.3. Cell type analysis was performed using the R package BrainInABlender. PCA plots of the BrainInABlender results used relative levels of all cell types to calculate PCA score.

In order to search for sequences, I used the command line utility grep which can search raw text files for specific strings, and return the number of occurrences of a particular string within a file.

5.4. Results

5.4.1 Sequencing of samples

We extracted and sequenced the brains of 30 mice, 15 at 6 months, and 15 at 12 months, each group consisted of 5 (GR)400 mutant mice, 5 (PR)400 mutant mice and 5 littermate controls. Our intention was to see what effect these mutations would have on gene expression of these mice.

5.4.2 Removal of outliers

Our initial attempt at analysis did not produce the results I was expecting. There were very few genes which were differentially expressed between mutant mice and comparable wildtype mice. This was true both in the 6-month-old samples, and the 12-month-old samples which had begun to develop motor deficits. We had already recognised that two mice (NM8078_797806, NM8082_800555) had been contaminated, and I had not included them in my analysis.

Table 5-1. Table showing the presence of each mutant sequence in samples. In this table, each sample name corresponds to a different mouse. I used grep to search for a 40 BP long representative sequence of each sample. Read depth is included for those samples with

Sample	Mutant Only Sequence	PR Only Sequence	GR Only Sequence	Condition	Age	Read Depth
(PR)400_8070	11	38	0	(PR)400	6M	60063578
(PR)400_8071	12	29	0	(PR)400	6M	56546275
WT_8072	0	0	0	WT	6M	58594057
(PR)400_8073	18	23	0	(PR)400	6M	62770382
WT_8074	0	0	0	WT	6M	53215370
(GR)400_8075	17	0	82	(GR)400	6M	73325379
(GR)400_8076	17	0	107	(GR)400	6M	64708113
(GR)400_8077	13	0	66	(GR)400	6M	48870851
WT_8078	0	3	0	WT	6M	
(GR)400_8079	35	0	117	(GR)400	6M	66664757
WT_8080	0	0	0	WT	6M	74336161
(GR)400_8081	19	0	62	(GR)400	6M	58270089
(PR)400_8082	6	7	52	(PR)400	6M	
(PR)400_8083	9	15	0	(PR)400	6M	65808400
8084	1	15	0	NA	6M	
WT_8085	0	0	0	WT	12M	81464543
WT_8086	0	0	0	WT	12M	82417595
(PR)400_8087	5	12	0	(PR)400	12M	51246118
(PR)400_8088	18	47	0	(PR)400	12M	77818437
WT_8089	0	0	0	WT	12M	67205224
(PR)400_8090	14	36	0	(PR)400	12M	78825057
WT_8091	0	0	0	WT	12M	73817810
(PR)400_8092	15	38	0	(PR)400	12M	80418384
8093	0	0	7	NA	12M	
(GR)400_8094	34	0	170	(GR)400	12M	77805035
(GR)400_8095	34	0	139	(GR)400	12M	85426812
(GR)400_8096	32	0	147	(GR)400	12M	76775846
(GR)400_8097	34	0	148	(GR)400	12M	76304988
(PR)400_8098	15	35	0	(PR)400	12M	89782534
(GR)400_8099	17	0	117	(GR)400	12M	82504083

Carmelo Milioto performed PCR, searching for contamination with mutant sequences of the samples. For two of the wildtype samples the presence of mutant sequences was identified. Specifically, the samples NM8084_800973, and NM8093_793422. I then confirmed this issue in our RNA by searching for the mutant sequences in our fastq files (Table 5-1). Fortunately, the samples were of different ages so their removal would not have too much effect on overall dataset integrity. I also found one 6-month-old (PR)400 sample to also contain sequences from the other mutation - NM8082_800555. It was not obviously different in the PCR, but I removed all 3 of these samples from future analysis.

After removing these samples, I generated PCA plots of normalised reads combining all samples at once (Figure 5-1 a), as well as one for our 6 months samples and one for our 12-month ones (Figure 5-1 b and c respectively) . The biggest difference between the samples as a whole is the age. In the PCA plot of the 6-month samples, there is no clear visual separation by genotype. There is no phenotype at this age, so the lack of changes within the RNA was not unexpected, as substantial changes would be accompanied by some motor deficit. The separation of mutants from their wildtype samples at 12 months old appears to be more defined. In Figure 5-2 heatmaps looking at the expression of most variant genes found similar results; a strong grouping in the 12-month samples and a mild grouping in the 6-month samples, more so in the PR than GR mutants.

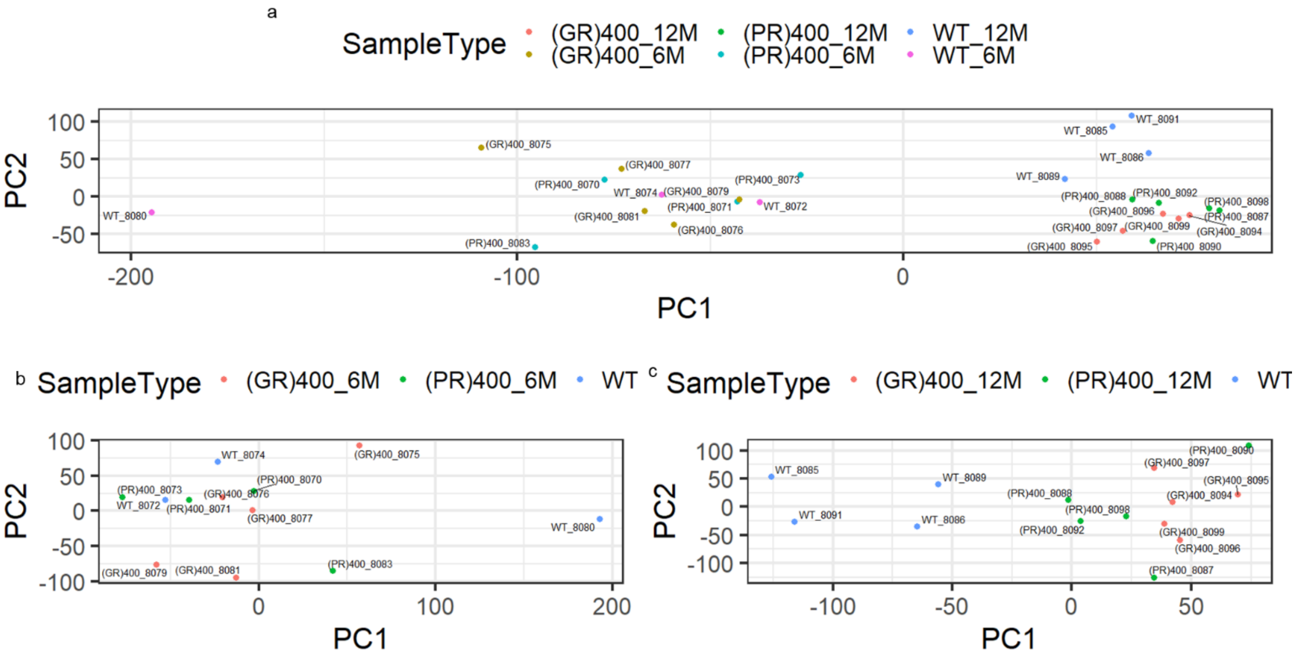

Figure 5-1. PCA plots of normalised reads. (a) all samples, (b) 6-month-old samples, and (c) 12-month-old samples. Age appears to be the largest factor in how samples group. When samples are just performed at 6 months there is not a clear separation between genotypes and wildtypes. At 12 months, there is a clear segregation between mutant and wildtype samples. While there is some mixing between the two mutant lines, they also appear to cluster separately.

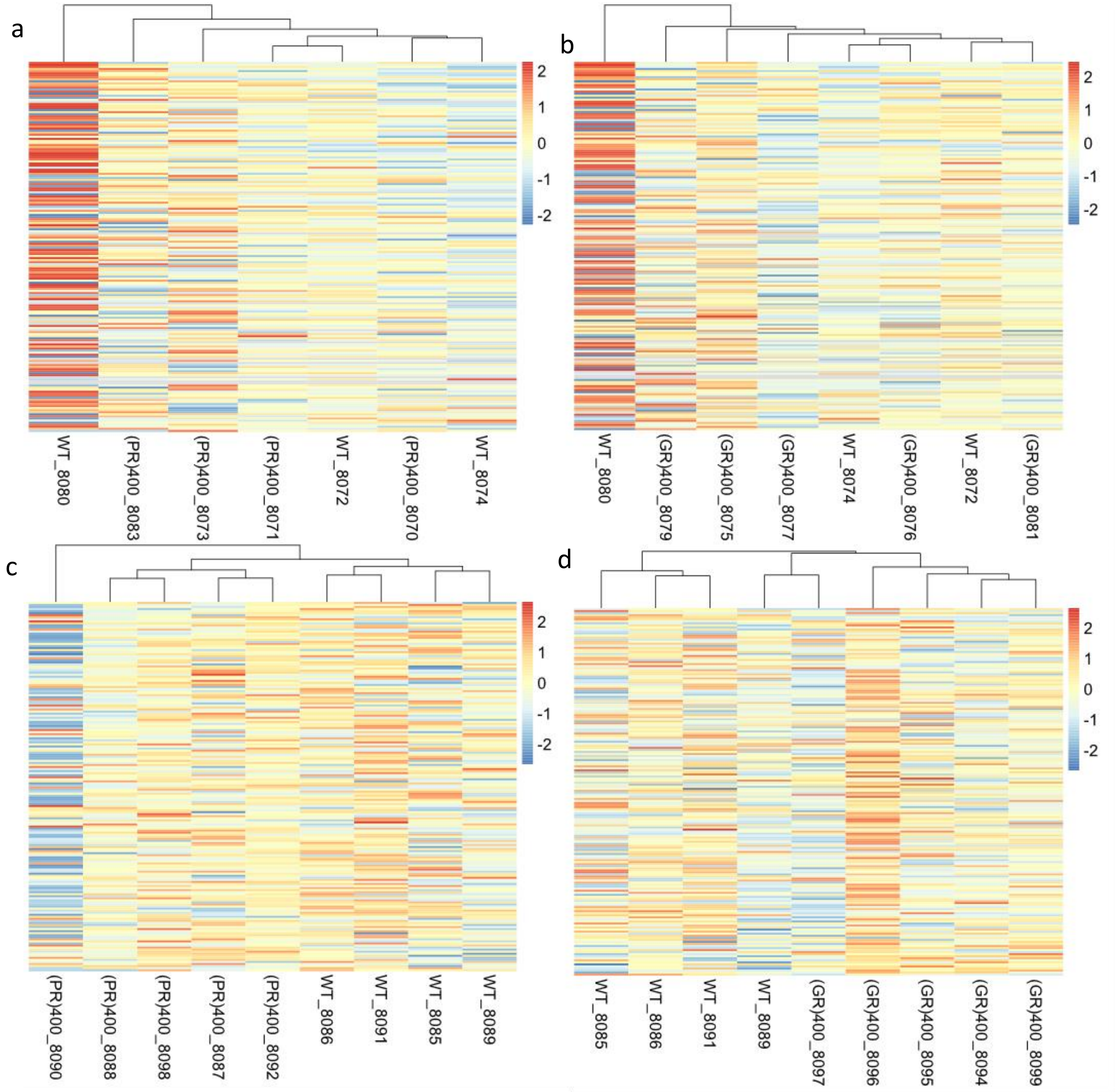

Figure 5-2. Heatmaps of normalised expression of genes in the 99.9th percentile for variance of expression across samples within each plot. Each line represents the expression of an individual gene using normalised values as produced by DESeq2. Colour coding indicates log of these normalised reads. (a) 6-Month-old (PR)400 samples and their relevant wildtype, (b) 6-Month-old (GR)400 samples and their relevant wildtype, (c) 12-Month-old (PR)400 samples and their relevant wildtype, (and (d) 12-Month-old (GR)400 samples and their relevant wildtype.

5.4.3 Differential expression

We separated the samples both by age and mutation status and performed differential expression using DESeq2 (via the EnrichmentBrowser package).

The number of genes found to be significant in both of our 6-month datasets was relatively minimal, only resulting in 2 or so differentially expressed genes per dataset. *C9orf72* was differentially expressed in PR and *Egf* was differentially expressed in both. The PR and GR 12-

month-old datasets had 398 and 1549 differentially expressed (padj < 0.05) genes respectively. The phenotype is much stronger at this age so that is not unexpected. Volcano plots can be seen Figure 5-3.

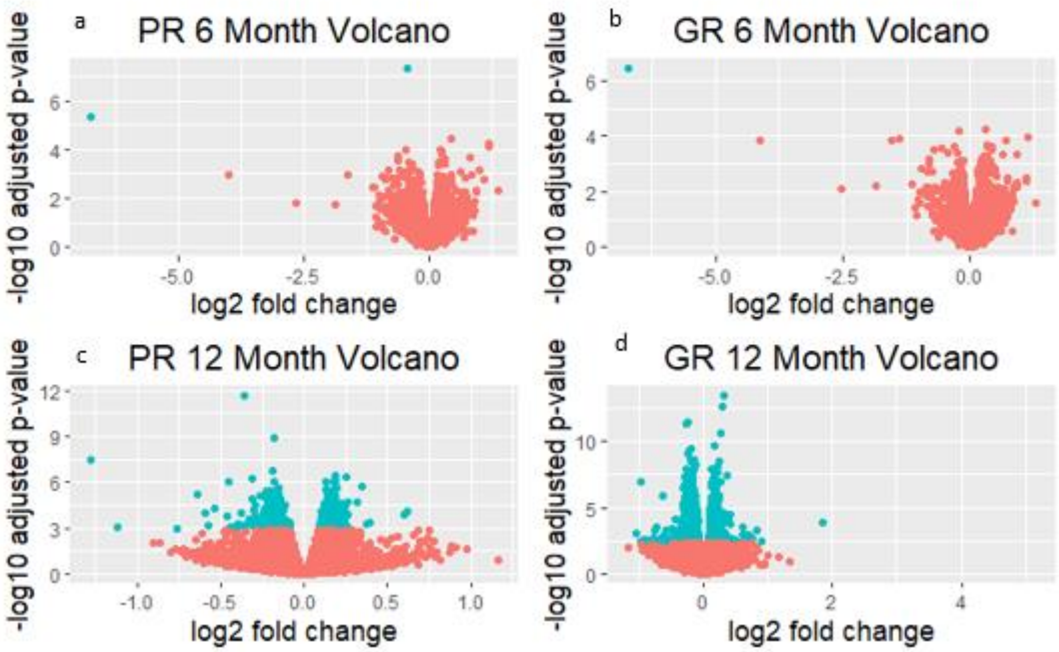

Figure 5-3. Volcano plots of differential expression. (a) 6-Month-old (PR)400 samples and their relevant wildtype, (a) 6-Month-old (PR)400 samples and their relevant wildtype, (b) 6-Month-old (GR)400 samples and their relevant wildtype, (c) 12-Month-old (PR)400 samples and their relevant wildtype, (a) 6-Month-old (PR)400 samples and their relevant wildtype, and (d) 12-Month-old (GR)400 samples and their relevant wildtype. Blue genes are considered to be significantly differentially expressed (padj < 0.05), red genes do not meet the significance threshold

Then I grouped our samples by mutation and by age. There are no significantly differentially expressed genes when grouping all GR samples together and 4 when grouping all PR samples together one of which was *C9orf72*. When grouping by age, 2 genes were differentially expressed when combining 6-month samples (Egf and Ngf), and 1,781 were found when merging the 12-month samples.

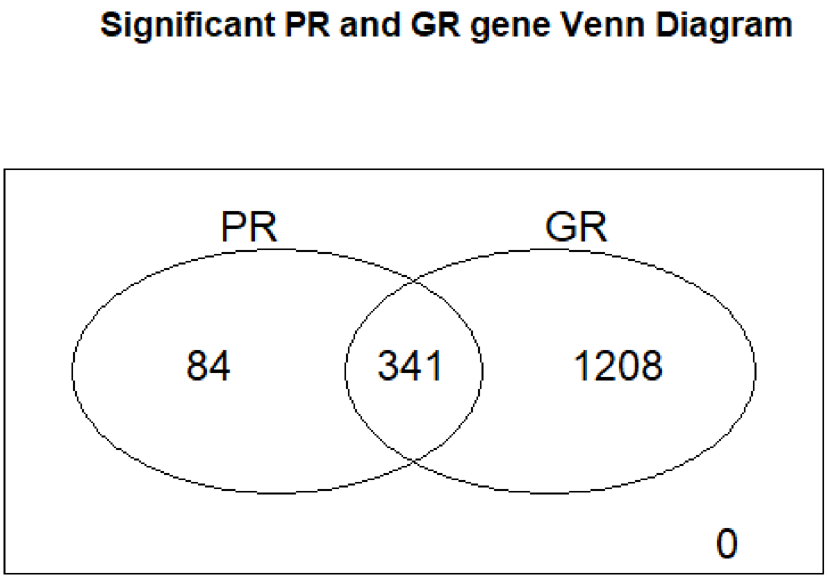

Figure 5-4. Venn Diagram of overlap of differentially expressed genes (padj < 0.05) in our 12-month-old (PR)400 and (GR)400 samples

PR has significantly fewer genes differentially expressed in its 12-month-old samples than does GR (t-test p-value < 0.0001). The majority of genes which are significantly differentially expressed in PR are also significantly differentially expressed in GR. (Figure 5-4).

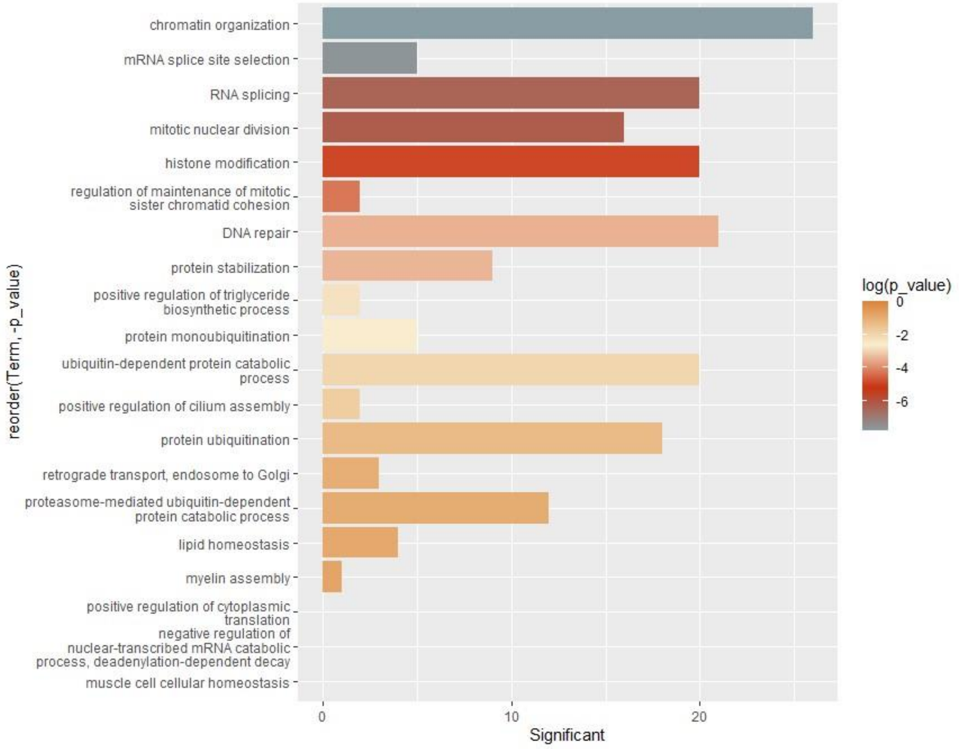

Figure 5-5 Biological GO term enrichment plot in 12-month-old (PR)400 brain samples. Following RNA-seq differential expression analysis, GO term enrichment was performed using the package topGO. The top 20 most significantly GO terms are shown, with the number of significantly differentially expressed genes associated with the GO term on the X axis, and coloured by p-value significance.

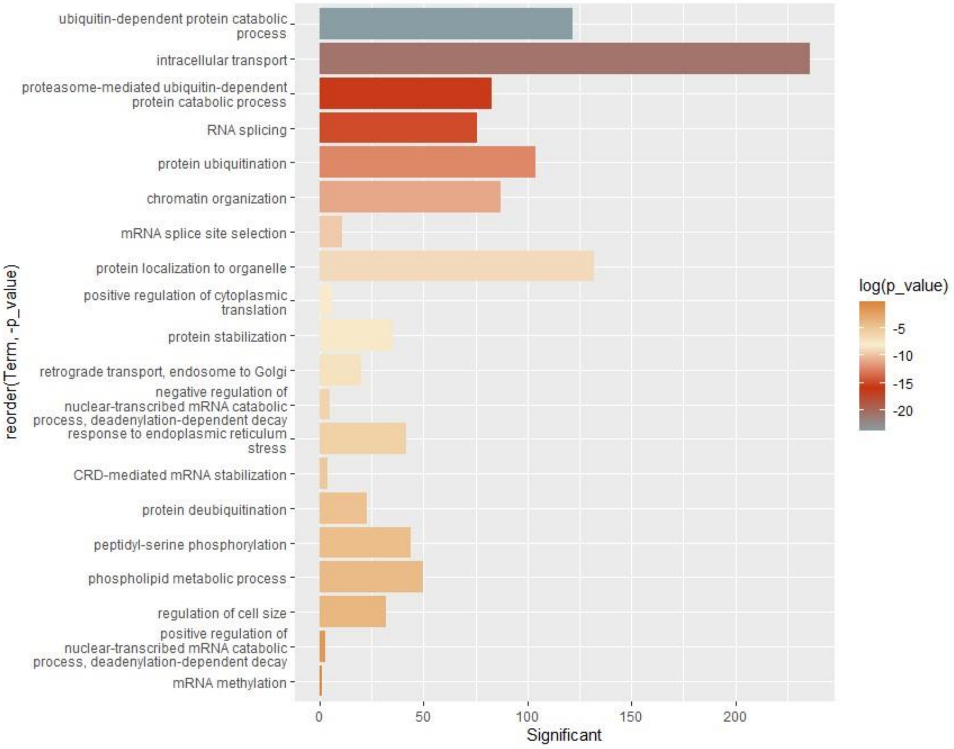

Figure 5-6. Biological GO term enrichment plot in 12-month-old (GR)400 brain samples. Following RNA-seq differential expression analysis, GO term enrichment was performed using the package topGO. The top 20 most significantly GO terms are shown, with the number of significantly differentially expressed genes associated with the GO term on the X axis, and coloured by p-value significance.

In our 12-month-old (PR)400 samples (Figure 5-5) our most significant GO terms included RNA splicing and chromatin organisation, histone modification, and DNA repair. In the 12-month-old (GR)400 samples (Figure 5-6), there are a large number of significant GO terms including three terms related to ubiquitin dependant protein degradation, as well as two terms relating to RNA splicing, and chromatin organisation.

When performing significance testing, we were surprised to find that *C9orf72* was not found to be significantly decreased in either of our GR datasets so wished to plot the expression against wildtype as we expect a 50% reduction due to the knock-in approach removing one healthy C9orf72 allele, which was confirmed by RT-PCR and immunoblot (Milioto et al., 2023). As can be seen in Figure 5-7 there appears to be a difference in overall expression from control in our PR samples at both timepoints (although less so at 12 months) and while there is some in the GR, it is less pronounced, and does not clear the t-test significance threshold of p-value <=0.05.

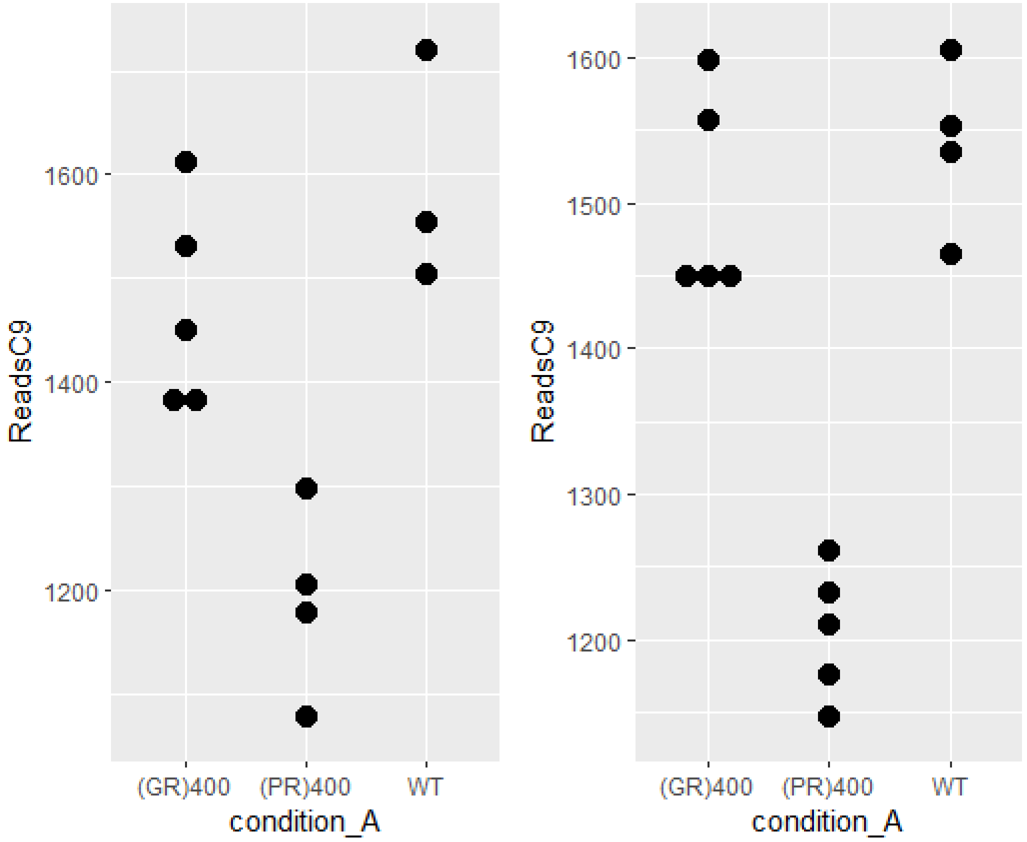

Figure 5-7. Dot plots of normalised expression of C9orf72. (a) 6-month-old, and (b) 12-month-old samples. Data was gathered by selecting all reads mapped to ENSMUSG00000028300 which is the Ensembl ID of C9orf72.

As explained in the previous chapter, brain cell types are one of the most heterogeneous in the body. Given that the samples I was analysing are from brains, I wanted to investigate whether there were changes in specific cell types caused by (GR)400 or (PR)400 knock-in. I ran the R-package BrainInABlender on normalised expression and examined the resultant cell type balance. I created PCAs (Figure 5-8) and heatmaps (Figure 5-9) of the balance of each cell type across samples. Neither type of figure showed a strong correlation with sample type. This lack of difference was further validated when I created boxplots both with and between group p-values and Kruskal-Wallis p-values to compare across all samples, there was no significant difference in any cell type at any age (Figure 5-10). When I separately analysed the data by age this lack of difference continued.

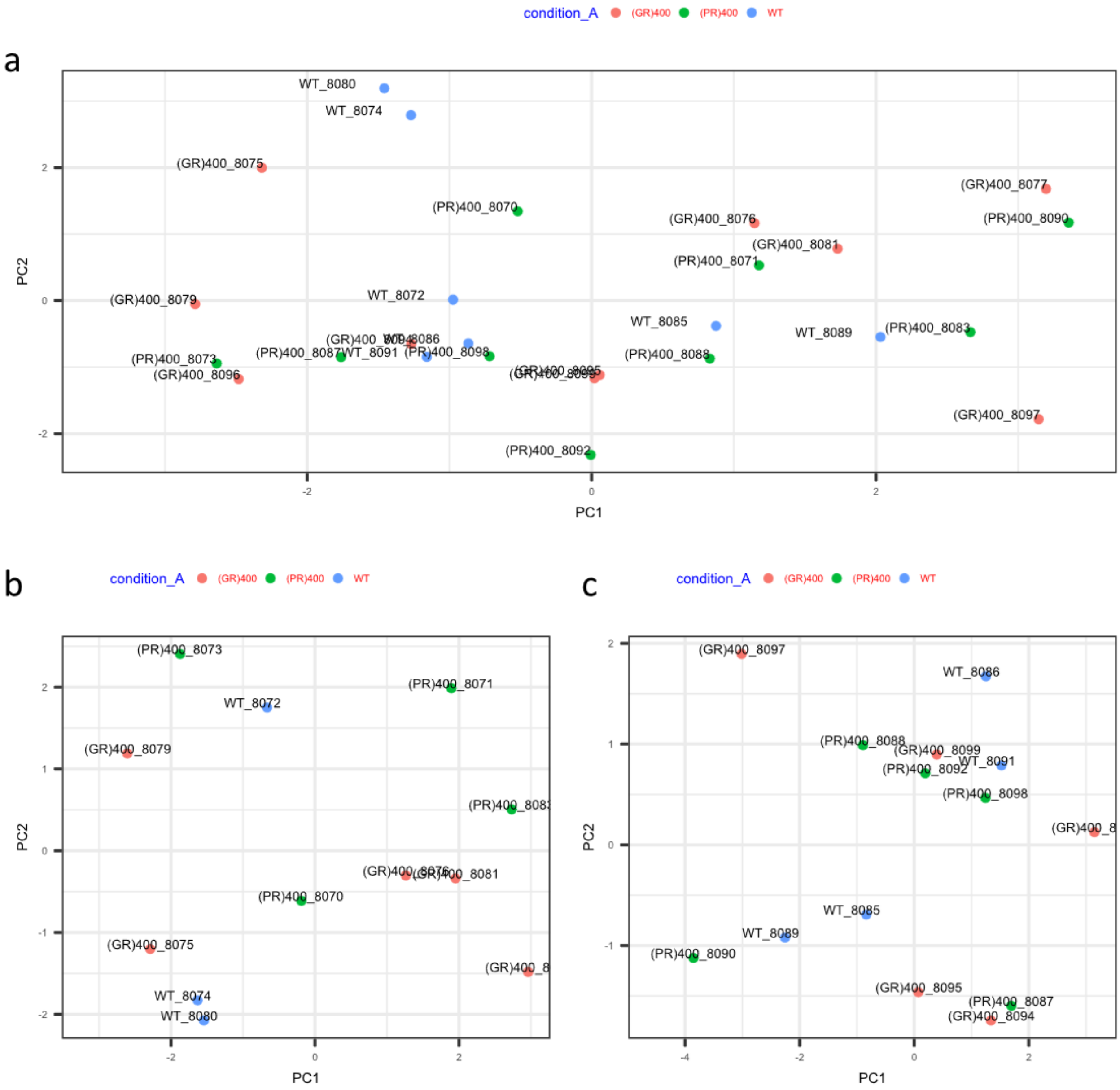

Figure 5-8. PCA plot of relative enrichment of cell types in each sample compared to what is expected based on expression of cellular marker genes from brain in a blender. Plots are in (a) all samples, (b) 6 Month samples, and (c) 12 Month samples. There does not appear to be clear segregation of samples by mutation

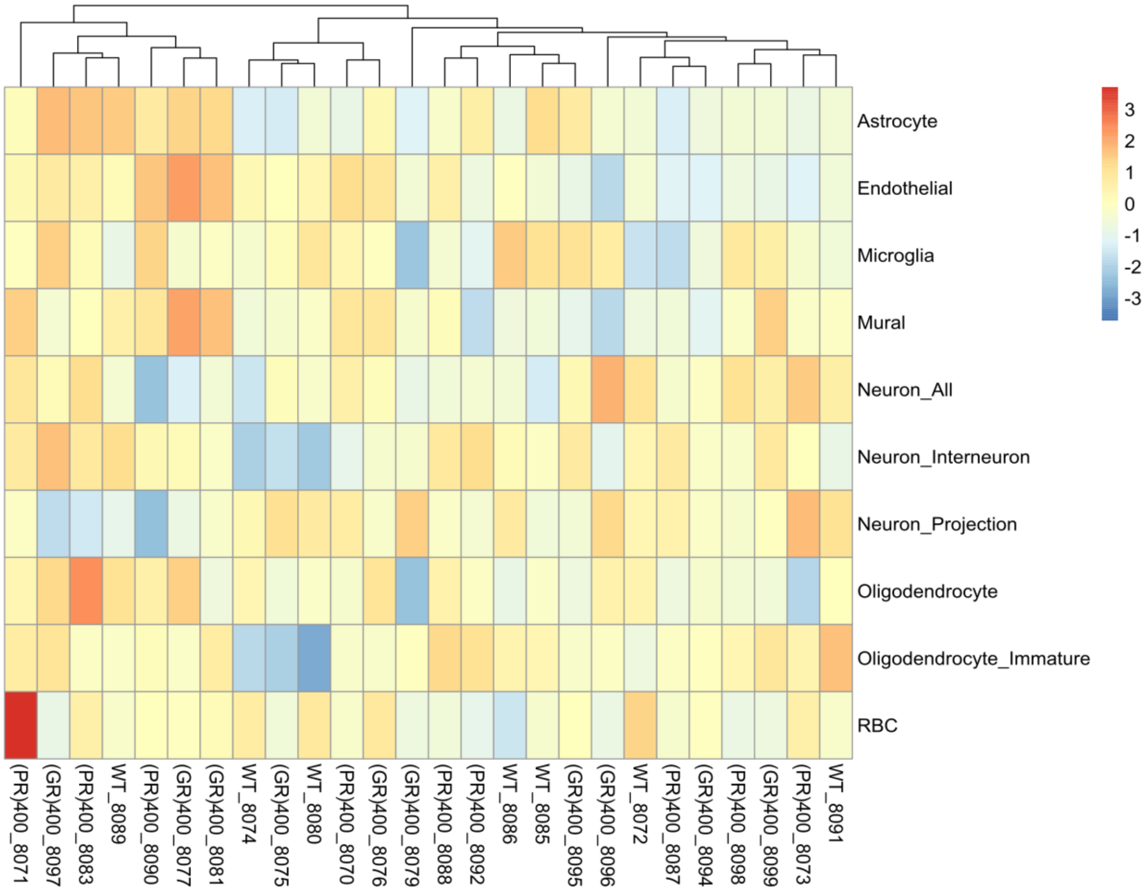

Figure 5-9. Heatmap of relative balance of cell types in all samples coloured by relative enrichment/depletion compared to other cells, normalised on each row. This chart shows all samples which were involved in analysis at all timepoints.

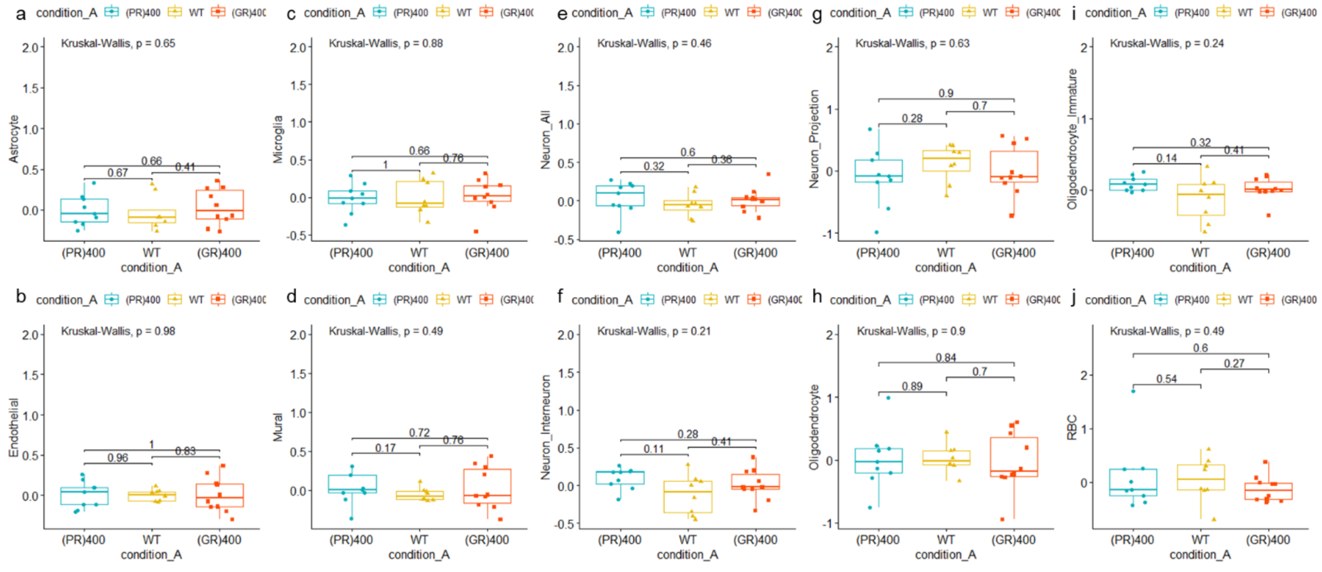

Figure 5-10. Dot plot of relative levels cell type in each sample in each cell, separated by mutation. (A) Astrocytes, (B) Endothelial, (C) Microglial, (D) Mural, (E) All Neurons, (F) Interneurons, (G) Neuron Projections, (H) Oligodendrocytes, (I) Immature Oligodendrocytes, and (J) Red blood cells. Significant differences comparing each group to each other group individually are tested using a t test, and across all samples using a Kruskal-Wallis test.

5.5. Discussion

As stated previously, there are several hypotheses as to why mutations in C9 cause disease. The mouse models created aim to test for changes caused by the most pathological DPR proteins in exclusion of the other possible causes of disease. While this does not resemble changes seen in patients, it is an important step in further developing our understanding of how C9 causes disease.

The main findings of my analysis were that the minimal phenotype in 6-month-old samples reflect an almost total lack of differentially expressed genes at this age, and that the GR seems to be more biologically active, and causes more genes to be differentially expressed. Most of the genes that were found to be expressed in PR expression mutants when compared with wildtype were also differentially expressed in GR expressing mutants when compared with wildtype. This showed a substantial overlap between the two mutations. It was unexpected that *C9orf72* was not found to be significantly differentially expressed at the gene or exon level at any age in GR, one copy of the gene was removed by the inserted sequence, so further investigation into why it was not reduced in GR samples will need to be performed. No change in cell types were observed. This is consistent with no overt gliosis in this model at 6 or 12 months of age. Neuronal loss was observed only in the spinal cord but not the brain, so no change in neurons is also expected. The neuronal loss in spinal cord was only identified after RNA-seq of the brain was performed. In the future it will be interesting to perform RNA-seq on spinal cord from these mice.

Poly(PR) related GO term enrichments were most strongly linked to RNA splicing, chromatin organisation, histone modification, and DNA repair. These results are very interesting, and show some notable overlap with existing studies on poly(PR). For instance, in mice, overexpressing 50 PR repeats showed heterochromatin abnormalities and abnormal histone methylation (Zang et al., 2019). This was thought to be at least in part driven by the specific effects of poly(PR) on heterochromatin protein 1 alpha phase separation. It is intriguing that histone modification was not enriched in the GR dataset, indicating a specific effect of poly(PR) which is consistent with this earlier study. Further work is now needed to investigate effects of poly(PR) on histone modification in our model and how well they correspond to the reported changes in the 50 PR repeat mouse model. It is not surprising that the changes are less marked in our model where poly(PR) expression was driven by the endogenous mouse C9orf72 promoter as compared to high levels of overexpression using AAV driven expression in the PR 50 mice. It was also noted that the majority of differentially expressed genes in the PR 50 mice were downregulated. While we didn't observe this trend, we noted that the genes with the highest fold changes and lowest p-values appeared to have negative fold changes. DNA repair has consistently been identified as a potential player in C9FTD/ALS, while most of the data focused on poly(GR) (Gao et al., 2017), one study shows an effect of poly(PR) on DNA damage (Farg et al., 2017).

Poly(GR) GO term enrichment identified GO terms relating to ubiquitin dependant protein degradation, and RNA splicing as two of the most significantly altered. Both of these pathways have been implicated in neurodegenerative diseases. It is possible that they contribute to neuronal dysfunction in these mice. Of relevance the finding that splicing is dysregulated in C9FTD/ALS patient brains (Prudencio et al., 2015). It would be very interesting to perform differential splicing analysis on our GR 400 dataset to see how well it matches human post-mortem data. We expect our changes to be much earlier in the pathogenic cascade but changes that persist in human post-mortem material would be interesting to follow up. TDP induced splicing changes are now well characterised and induce the inclusion of cryptic exons when TDP function is reduced. However, TDP mislocalisation was not observed in the GR(400) mice (Milioto et al., 2023), therefore we would expect splicing changes to be independent of TDP-43, likely mediated either by direct binding of poly(GR) to RNA, or downstream compensatory effects. The Isaacs lab have recently investigated transcripts which bind to poly(PR) using CLIP and their data suggests it would be interesting to do the same with poly(GR).

Our results in cell typing helped to validate the results from our chapter looking at the brains of FTD patients as well. There did not seem to be any substantial changes to expression in certain cell types/composition of cell types arising from these *C9orf72* mutations. Any of the reasons previously hypothesised in section 4.5 may be responsible for this. An additional hypothesis in this case may be that the DPRs we have produced do not result in cell type specific neurodegeneration.

5.5.1 Future work

The biggest unanswered question arising from my work in the C9 dataset is why does C9 not seem to show significant changes in GR mutants? In theory the one copy of gene should be knocked out but in both the examination of just the gene, and of each exon, I found that there was no significant change either overall or in any single exon. This suggests that the presence of the mutant sequence rather than changes in the levels of expression of C9orf72 are responsible for a lot of the changes observed both in terms of gene expression and pathology.

Validation of whether there are changes through qPCR would raise interesting questions for future research regardless of which result is found. If there truly are few or no changes in expression of C9orf72 in our adult mouse samples why is this? There must be a regulation mechanism at play which can function with this particular GR mutation. If there are changes that our RNA-seq data just is not showing, it would be important to investigate why this is.

Once this investigation has been performed, this model has the potential to provide interesting information on the mechanisms by which DPRs cause disease in the absence of RNA foci. Further interrogation of specific genes which are affected by each DPR may provide more insight into mechanisms of how the DPRs cause disease and highlight potential targets for further research or treatment.

The changes of GO terms linked to RNA splicing suggest that it may be fruitful to perform splicing analysis on these samples at a future data to evaluate which, if any specific changes which are occurring as a result. The read depth of the existing data may be sufficient for splicing analysis.

The creation of additional models may also be a fruitful endeavour. This model aimed to study DPR proteins without the creation of RNA foci. A future model which attempts to evaluate the effect of RNA foci in the absence of DPRs would also be useful if difficult to do. A model which produces RNA-foci and C9orf72 at endogenous levels would also be hugely informative as to the role some of the DPRs play while C9 is at endogenous levels. This would also not be something seen in patients, but would further our understanding of some of the specific mechanisms of disease.

The lack of differences in cell types observed when analysing this dataset does seem to continue our findings from the FTD Brain chapter (Chapter 4). Further investigating the specific mechanisms of changes in brain cell types would be a good strategy for future research. Further developing tools for analysis and integration of brain cell type data into bulk RNA-seq analysis has the potential for broad reaching utility in neurodegeneration.

6. Analysis of F210I mutant mouse data

6.1. Publication

This work has yet to be published.

6.2. Introduction

Transactive response DNA-binding protein (TDP-43) is an RNA-binding protein encoded by the highly conserved gene *TARDBP*. In normal physiological conditions, TDP-43 is localised almost exclusively in the nucleus, where it plays a role in various forms of RNA metabolism such as splicing. A ubiquitinated and hyperphosphorylated version is found in the majority of patients with ALS, and many with FTD irrespective of mutations in the gene (Ling et al., 2013). I therefore hypothesise that this pathology is convergent, with several possible causes leading to mislocalisation of TDP-43 (Suk et al., 2020). This underscores the importance of understanding the role of TDP-43 in pathology to overall understanding of both ALS and FTD.

A large amount of research has been done into the pathogenesis of TDP-43, and despite this, the underlying mechanisms of TDP-43 pathology are unclear. Presence of inclusions in the cytoplasm suggests that there is a toxic gain of function effects, but the nuclear depletion also suggests loss of function. Disambiguating these effects is important for development of therapeutic strategies.

Gain of function effects are typically modelled through overexpression of the affected gene and loss of function through knockout/knockdown models. TDP-43 can regulate its own expression through binding to a site in its 3' UTR (Ling et al., 2013; Sephton et al., 2011). This makes both knockdown and overexpression models technically challenging while homozygous knockout models are embryonically lethal.

Our lab has been working with mice who have a point mutation localised in one of *TARDBP'*s 2 RNA binding domains (RRM2) and gives rise to a phenylalanine to isoleucine change in the final protein. This leads to reduced RNA binding activity and a partial loss of splicing activity (Fratta et al., 2018). These mice continue to express TDP-43 at physiological levels allowing investigation of loss of function without the confounding factor of changes in expression. This mutation has not been encountered in humans, but seeks to use its unique factors to determine some of the effects of this specific loss of function. This mutation is of interest pathologically as mutations which are causative for ALS have been found to cluster around the RNA binding motifs of *TARDBP* (Ederle & Dormann, 2017; Ratti & Buratti, 2016).

As in previous loss of function models, homozygous versions of this mutation are embryonically lethal (Kraemer et al., 2010; Sephton et al., 2010; Wu et al., 2010). HOM embryos also show marked developmental delay compared with both HET mutants and wildtype littermates, but have normal organogenesis and no neuropathology. HET mutants are viable, with no overt motor defects of neuropathology even in aged animals observed. A list of some of the most relevant TDP-43/91Tardbp mouse models can be found in Table 6-1.

Table 6-1. Table of targeted knock in and ENU mutagen Tardbp models. Adapted from (De Giorgio et al., 2019)

Strain name	Transgenic/Gene targeted knock-in/ENU	Genetic Background	Protein	Inclusions/ aggregates	Final-stage disease (terminal MN loss)	Behavioural analysis	Reference
hTDP-43 A315T	Gene-targeted knock-in Endogenous Promoter: mTardbp	129S2/129P2/ Ola×C57BL/6Ntac	Human cDNA gene targeted into mouse Tardbp locus; alteration of 3'UTR aligning importance for autoregulation	Inclusions: Present Ubiquitin: Present TDP-43: Co-localises with Ub p62: Absent	Early signs at ~12-20 weeks. MN loss: 10% at 65 weeks	Mild motor coordination impairment	(Stribl et al., 2014)
Tardbp Q331K (JAX 031345)	Gene-targeted knock-in Promoter: mTardbp	C57BL/6J	mTDP-43; TDP-43 GOF. 45% increase in nuclear TDP-43 in mutant, impaired autoregulation	Absent	Early signs at ~20 weeks; no final MN loss	Mild to no motor coordination impairment at 20 and 24 weeks. Cognitive and memory impairment	(M. A. White et al., 2018)
Tardbp Q331K	Gene-targeted knock-in Promoter: mTardbp	C57BL/6J	mTDP-43; TDP-43 gain-of-splicing function	Not Stated	Not Stated	Not Stated	(Fratta et al., 2018)
Tardbp Q101X (JAX019899)	ENU point mutant Promoter: mTardbp	C57BL/6J×C3H/ HeH	mTDP-43; no difference in protein level between WT/mutant TDP-43	Absent	Early signs at 32-61 weeks; no final MN loss	Mild to No motor coordination impairment	(Ricketts et al., 2014)
Tardbp F210I (BRC#GD000108)	ENU point mutant Promoter: mTardbp	On C57BL/6J embryonic day 18.5; viable HOM on C57BL/6J-DBA/2J	mTDP-43; TDP-43 LOF (shift towards exon inclusion), cryptic exon. Reduced RNA binding	Absent	Absent	Absent	(Fratta et al., 2018)
Tardbp M323K (BRC#GD000110)	ENU point mutant Promoter: mTardbp	On C57BL/6J embryonic lethal; viable HOM on C57BL/6J-DBA/2J	mTDP-43; TDP-43 GOF(increased exon exclusion), skiptic exon. No nuclear depletion. Increased Tardbp intron 7 retention	Inclusions: Present Ubiquitin: Present TDP-43: Absent	Early signs at ~52 weeks; MN loss: 28% at 104 weeks	Motor coordination impairment	(Fratta et al., 2018)

The aim of this chapter is to evaluate the effects of a heterozygous RRM2 mutation in mice on RNA. Both levels of expression and differential splicing are evaluated. The mutants will be compared to wildtype mice. The changes observed will also be compared to the changes observed in homozygous mice when compared with their wildtype mice to evaluate whether there is a difference in the mechanisms of the two mutations beyond any changes in the size of the effect of mutations.

6.3. Methods

The *Tardbp RRM2mut* and *LCDmut* lines were derived from the RIKEN BioResource Centre and are available as BRC no. GD000108 and GD000110 from RIKEN BioResource Centre, respectively. RNA was extracted from MEFs, embryonic E18.5 head or adult spinal cord. Total RNA extracted using an Rneasy fibrous tissue extraction kit (Qiagen). Poly-A tailed RNA was purified using an oligo dT pulldown. Libraries were sequenced using an Illumina HiSeq 2000. The process was the same as that found in (Joyce et al., 2016; E. T. Wang et al., 2012).

5 adult wildtype, and 7 adult heterozygous mice were culled and sequenced. 4 embryonic wildtype mice and 4 embryonic homozygous mice were also culled and sequenced. The breeding and extraction were performed by Nicol Birsa and Agnieszka Ule. The read depth of the HET samples can be found in Table 6-2. The read depth of the HOM samples was between 31-36M.

Table 6-2. Read Depth of HET samples and their WT controls.

Sample	Reads
Het1	70765869
Het2	70218750
Het3	75901401
Het4	75418950
Het5	63184456
Het6	67389214
Het7	69631905
WT1	61407539
WT2	57542722
WT3	67807082
WT4	73074830
WT5	58903412

Alignment and read counting were performed as discussed in chapter 2. The full code for how I performed differential expression and other analysis can be found on my GitHub (https://github.com/SethMagnusJarvis/PhDFigureCreation/tree/main/F210I). Differential expression was performed using DESeq2 as described in section 2.3. Differential splicing analysis was performed as discussed in section 2.4

6.4. Results

6.4.1 Sequencing of samples

Our primary data for this dataset was comprised of 7 Adult mice, heterozygous for the RRM2 mutation, and 5 littermate controls. After sequencing the data was aligned and expression of genes was counted using the method described in chapter 2. I then performed differential expression and splicing analysis, as well as a comparison between this data and a previously analysed HOM dataset.

6.4.2 There are minimal gene expression changes in HET mutants

When comparing the levels of RNA expression in the mutants and their littermate controls, I found that the effect of the HET mutation was relatively small. As can be seen in the PCA of normalised gene expression across all genes (Figure 6-1), I did find some separation between some of the WT and HET samples. While there is some separation, it is not complete, two samples in particular (WT3 and WT4) are almost indistinguishable from our HET mutants. This lack of separation between the groups is even more pronounced when comparing expressions of the most variant genes via heatmaps in Figure 6-2. The Hierarchical clustering shows that 4 of the HET mutants group relatively closely together, one of the wildtype samples however, groups with them, and the other samples are less clearly separated. Given the relatively minor phenotype, this is not totally unexpected.

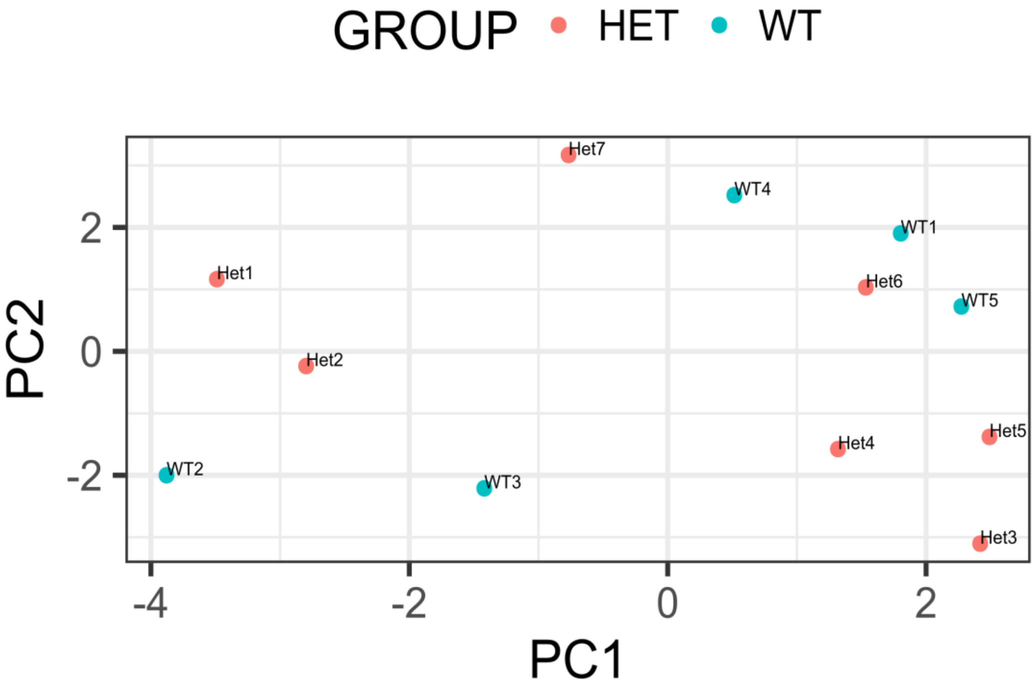

Figure 6-1. A PCA plot of the wildtype and heterozygous samples using normalised expression. It shows no clear separation between the wildtype and mutant samples.

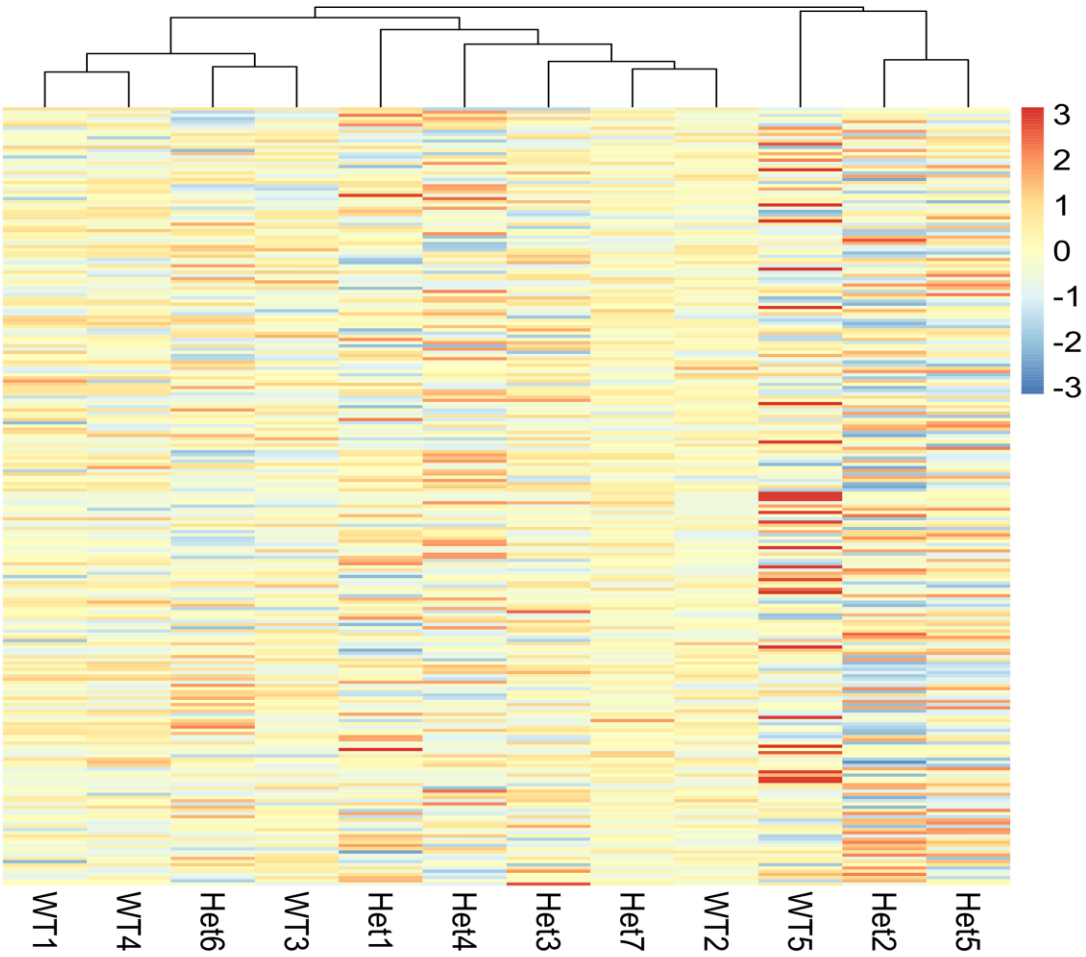

Figure 6-2. Heatmap created using the R package pheatmap using normalised levels of expression to show how samples cluster based on the top 1% most expressed genes. Each line represents the expression of an individual gene using normalised values as produced by DESeq2. Colour coding indicates log of these normalised reads.

Table 6-3. List of the differentially expressed HET mutant genes and a description of their function.

Ensembl ID	Gene Name	Function Description	Fold Change	Adjusted p-value
ENSMUSG00000041459	Tardbp	TAR DNA binding protein	0.2864056	0.001601195
ENSMUSG00000046865	Fbl	fibrillarin	0.3738060	0.008419617
ENSMUSG00000040666	Sh3bgr	SH3-binding domain glutamic acid-rich protein	-0.4510003	0.012066752
ENSMUSG00000067818	Myl9	Calcium ion binding/myosin heavy chain binding	0.5554708	0.006680943
ENSMUSG00000024395	Lims2	LIM and senescent cell antigen like domains 2	0.3522494	0.007746885
ENSMUSG00000033364	Usp37	ubiquitin specific peptidase 37	-0.1926423	0.021157982

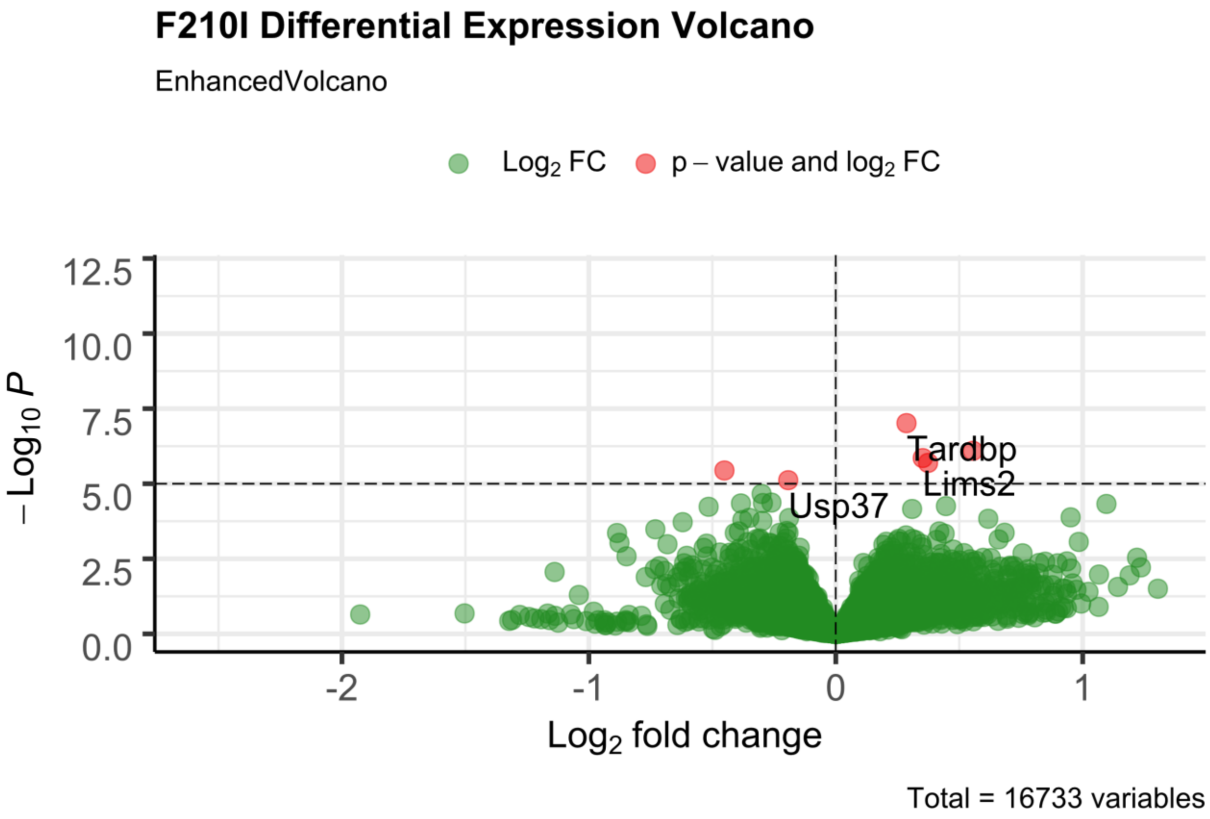

Figure 6-3. volcano plot of differential expression comparing HET mutant mice to wildtype controls. Red genes are considered to be significantly differentially expressed (padj < 0.05); green ones do not meet the significance threshold. It shows that all of the genes which are significantly differentially expressed have relatively small fold changes.

In order to evaluate differential expression, I used DESeq2. As can be seen in the volcano plot (Figure 6-3) there were relatively few genes found to be significantly differentially expressed. Table 6-3 shows the 6 genes which were significantly differentially expressed. TARDBP had a slightly increased expression. Overcompensation of the self-regulatory mechanism may cause this. This increase does not mean an increase in functional TARDBP however. The fold change of TARDBP is one of the smallest observed.

Figure 6-4. shows that while there were relatively few genes found to be significantly differentially expressed, two GO terms in particular were significantly enriched. Specifically, these were GO terms linked to RNA-polymerisation and binding to the 3’-UTR binding of mRNA.

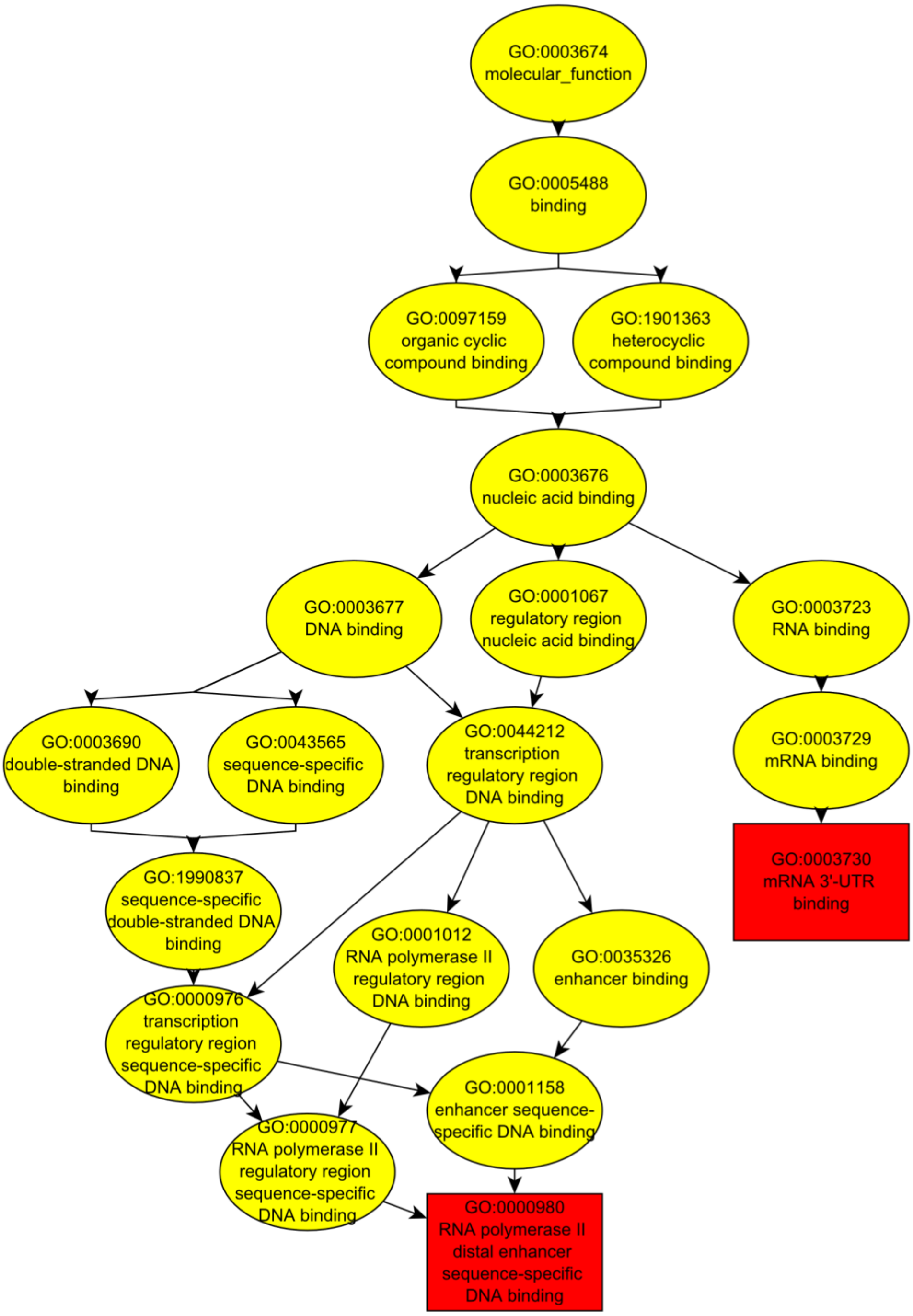

Figure 6-4. GO term maps of molecular function in KS filtering any genes with padj < 0.05. It shows that binding is by far and away the most strongly effected top level GO term, with 3' UTR binding, and a sequence linked to RNA polymerase II being seeming to be most strongly correlated. This is a subgraph induced by the most significant GO terms identified by the fisher exact algorithm for scoring GO terms for enrichment. Rectangles indicate the most significant terms. Rectangle colour represents the relative significance, ranging from dark red (most significant) to bright yellow (least significant). For each node, some basic information is displayed. The first two lines show the GO identifier and a trimmed GO name

6.4.3 Splicing changes

I used the method of calculating differential splicing described in the main methods chapter. It consisted of a combination of SGSeq and DEXSeq. As can be seen in our PCA plot of all of splicing events (Figure 6-5), there does not seem to be a clear separation between the samples. It is not uncommon for PCA not to distinguish genotypes. Given the small number of subtle changes in this dataset, it is not surprising that PCA and hierarchical clustering did not distinguish between hET mutants and WT.

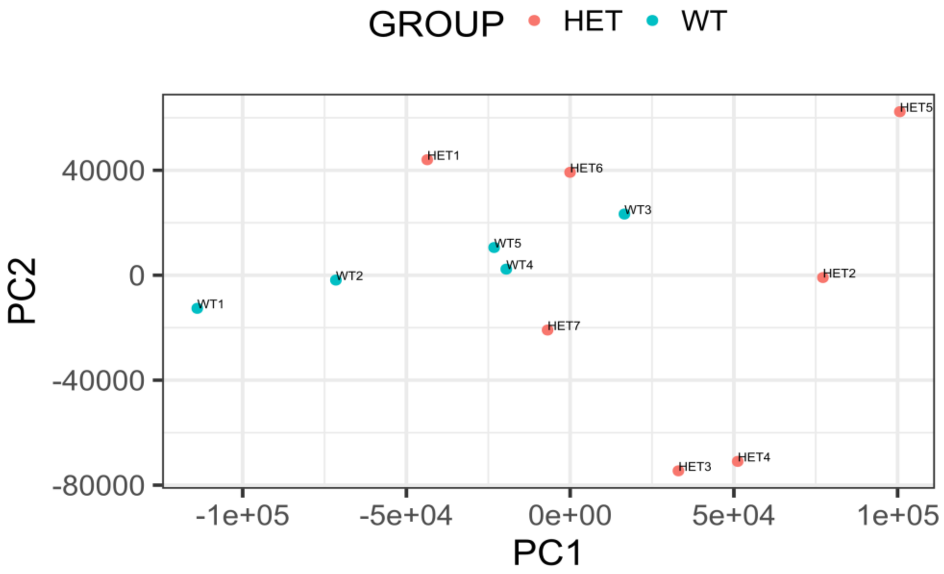

Figure 6-5. PCA of expression of splicing changes in HET data. It uses the normalised level of expression of each splicing event observed within our data.

Figure 6-6 shows the full breakdown of the events which were significantly differentially spliced. Half of these events are cassette exons – broadly regarded as the most common type of splicing event in mammals (Cui et al., 2017; Sammeth et al., 2008), they are a term for the inclusion or loss of one or more exons between other exons in the mature mRNA.

The Volcano plot of splicing (Figure 6-7) showed that the most significant splicing events were in the genes *Sh3bgr, Dnajc5 and Sort1. Sh3bgr* encodes a protein linked to cell migration and angiogenesis (Li et al., 2016). *Dnajc5* encodes for a protein linked to synapses and neurodegeneration (Cadieux-Dion et al., 2013). It has also been strongly to lysosomal storage disorders – a condition which results in the build-up of toxic products due to disruption in the lysosome (Henderson et al., 2016; Nosková et al., 2011). The encoded protein (CSPα) has been found to be neuroprotective in Huntington's and cystic fibrosis (Burgoyne et al., 2015)). *Sort1* is responsible for making a receptor protein found primarily in the CNS (Andersen et al., 2014).

When comparing our list of genes containing differential splicing events to our list of differentially expressed genes, only one gene was found to be in both, *Sh3bgr*. The full list of significantly differentially spliced events can be found in Table 6-4.

F210IHET05

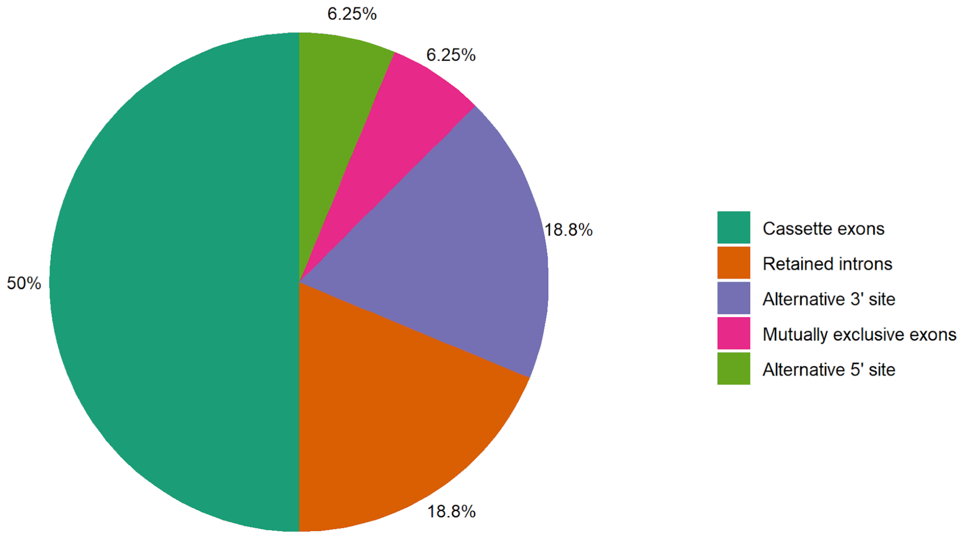

Figure 6-6. Pie Chart showing the proportion each type of differentially spliced event at adjusted significance level <= 0.05. Cassette exons – including or skipping exons – are by far the most common type of event, with alternative 3' and intron retention being the two next most common event types.

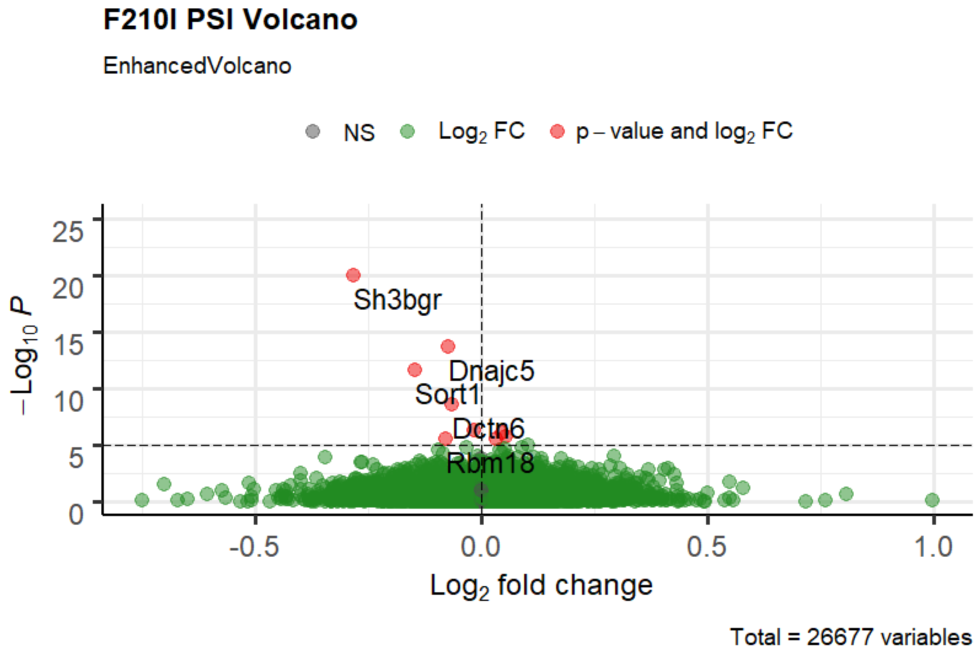

Figure 6-7. Volcano plots of Percentage spliced in (PSI) against log p-value. Red genes are considered to be significantly differentially expressed (padj < 0.05); green ones do not meet the significance threshold. All fold changes are relatively small, but the most significant splicing events tend to have a larger fold change.

All of the genes which have been significantly differentially spliced aside from GSTT3 and RBM18 have either been linked to mutations in *TARDBP* or linked to ALS more broadly(Dash et al., 2022; C. Yang et al., 2014; H.-S. Yang et al., 2020; Ziff et al., 2022). While RBM18 has not been directly linked to ALS or changes in *TARDBP*, like *TARDBP* it contains an RNA binding motif(*RBM18 RNA Binding Motif Protein 18 [Homo Sapiens (Human)] – Gene – NCBI*, n.d.).

Of the differentially spliced events, half were highly expressed (Exon Base Mean> 100). Figure 6-8 shows a slight trend, with splicing events with the highest percentage spliced in index (PSI) – a measure of the ratio of reads including each particular splicing event – almost all had higher mean reads than the median. It is likely that genes which have a higher expression are more actively spliced, although it is possible that there is an issue with power; genes with a lower number of reads are likely to have less splicing information so it is harder to make an accurate judgement.

Table 6-4. List of significantly differentially spliced events in HET mutants.

Gene Name	Exon Base Mean	padj	FDR	Variant Type
Sh3bgr	41.600603	9.223501e-62	9.391094e-62	Skipped Exon: Skipped
Sh3bgr	256.243679	3.321541e-61	3.381894e-61	Skipped Exon: Included
Dnajc5	73.216236	6.068522e-10	6.160618e-10	Skipped Exon: Included
Dnajc5	324.701741	6.805627e-10	6.929286e-10	Skipped Exon: Skipped
Sort1	130.242105	3.193490e-08	3.251517e-08	Skipped Exon: Included + alternative 3’ splice site: Proximal
Sort1	270.838162	3.417108e-08	3.479198e-08	Skipped Exon: Skipped + alternative 3’ splice site: Distal
Dctn6	96.376952	2.879804e-05	2.932130e-05	Skipped Exon: Included + Mutually Exclusive Exons
Dctn6	194.121630	1.388575e-04	1.413806e-04	Skipped Exon: Skipped + alternative 3’ splice site: Distal
Ndufa12	16.072170	3.564028e-03	3.628787e-03	Skipped Exon: Skipped + Two Consecutive Exons Skipped: Included

Ndufa12	729.286697	4.007797e-03	4.080619e-03	Skipped Exon: Included + Two Consecutive Exons Skipped: Skipped
Phldb1	188.673438	6.675040e-03	6.778156e-03	Retained Intron: Retained + alternative 3’ splice site: Proximal
Phldb1	3.633528	9.150234e-03	9.316496e-03	Retained Intron: Excluded+alternative 3’ splice site: Distal
Eif4h	240.776375	9.150234e-03	9.316496e-03	Skipped Exon: Skipped + Retained Intron: Excluded + alternative 5’ splice site: Distal + alternative 3’ splice site: Distal
Eif4h	487.715709	9.700046e-03	9.876298e-03	Skipped Exon: Included + Retained Intron: Retained + alternative 5’ splice site: Proximal + alternative 3’ splice site: Proximal
Gstt3	46.560625	1.180594e-02	1.202046e-02	Skipped Exon: Skipped
Gstt3	7.229119	1.252736e-02	1.275499e-02	Skipped Exon: Included
Rbm18	6.365913	1.252736e-02	1.275499e-02	
Laptm4a	103.626496	3.550280e-02	3.614790e-02	Retained Intron: Retained
Laptm4a	11.105789	3.550280e-02	3.614790e-02	Retained Intron: Excluded

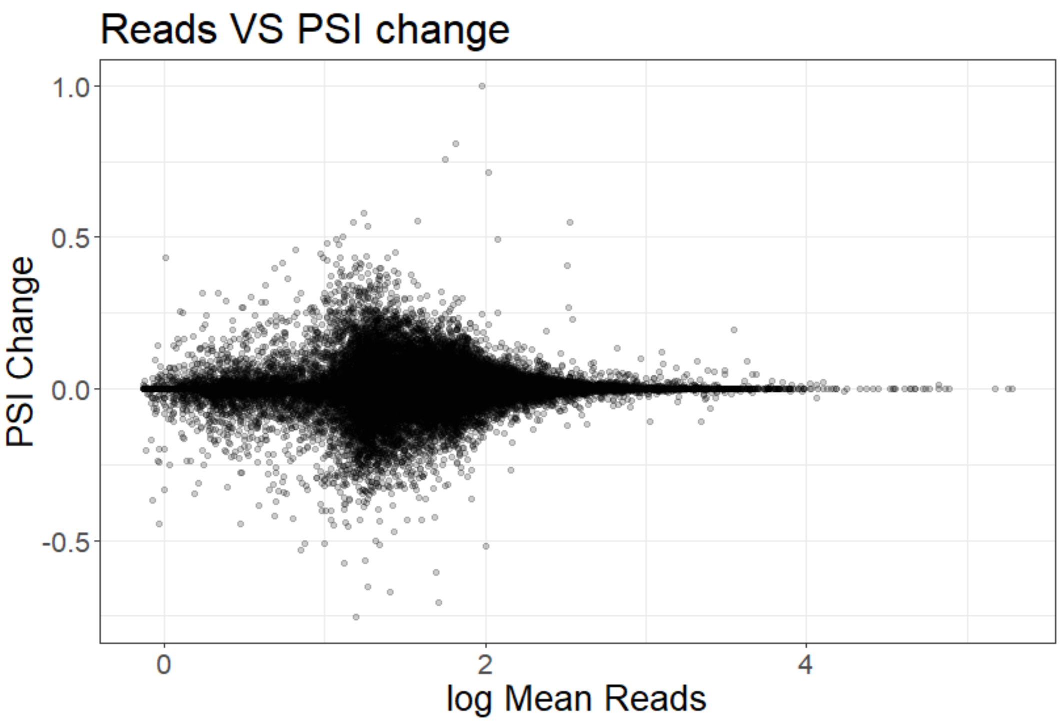

Figure 6-8. Scatter plot of splicing events plotting mean expression vs PSI change. This shows that the genes with the highest PSI had higher mean reads than the median.

6.4.4 Comparison of HET and HOM mutants

The Fratta lab had previously produced homozygous mutant mice. This mutation is embryonically lethal, therefore the sequencing had to be performed on embryos unlike our HET mutants which were performed on adult mice. An additional aim of my research was to see how much the results of our HET mutation differed from our HOM mutants. Data from embryonically culled heterozygous mice had not been collected, so adult HET mouse data was compared with embryonic HOM mouse data.

Initial PCA of our HET and HOM datasets Figure 6-9 showed the characteristic which had the strongest effect was the age at which the samples were taken, this meant that direct comparison of HET vs HOM would contain too much biological noise. I therefore decided to make sure that any analysis done was between the results of differences of mutants and their controls rather than directly between our groups.

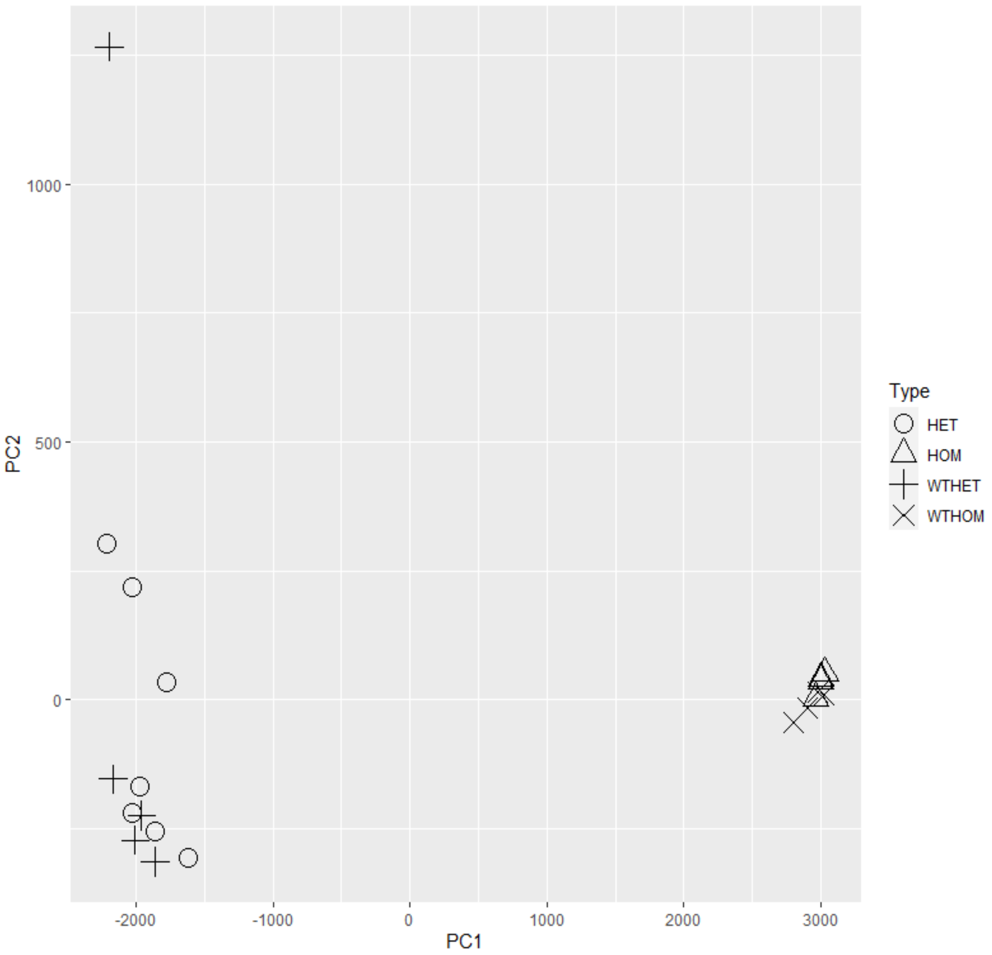

Figure 6-9. PCA of normalised gene expression of both HET and HOM. It shows that samples cluster more closely with other samples of the same age taken at the same time than samples with similar mutations. This suggested a strong batch effect making direct comparisons impractical.

We ran DESeq2 on both datasets and found that of the 6 genes which are differentially expressed in our HET dataset, 3 are found to also be differentially expressed in the HOM as well (*Tardbp*, *Fbl*, and *Sh3bgr*). This can be seen in Figure 6-10. Similarly, Figure 6-11 shows that the 5 of the 13 differential splicing events found within HET are also found to be significant within HOM.

I created figures to give a clearer picture of expression in the genes found to be significantly differentially expressed in HET (Figure 6-12). As well as further reinforcing the batch effect observed in Figure 6-9, it did clearly show differences between the mutants and their relative controls. It also showed that in genes with significant differential expression in HOM, HOM tended to show larger changes than HET. In our z-score plots (Figure 6-13), the trend is broadly positive with most significant z-scores having the same direction of fold change (positive or negative) in both samples. As expected given the relative strength of the phenotypes, our HOM dataset had a lot more significantly differentially expressed genes than our HET.

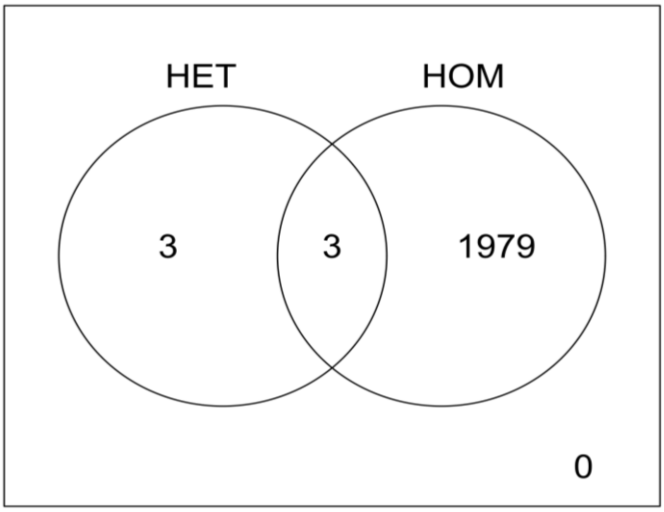

Figure 6-10. Venn diagram of genes found significantly differentially expressed (padj < 0.05) in both HET and HOM. It shows that the HOM mutants had many more genes significantly differentially expressed than did HET mutants, and that only half of the genes significant in HET were significant in HOM suggesting a difference in mechanism over time. The genes which were found to be significantly differentially expressed in both samples are Tardbp, Fbl, and Sh3bgr

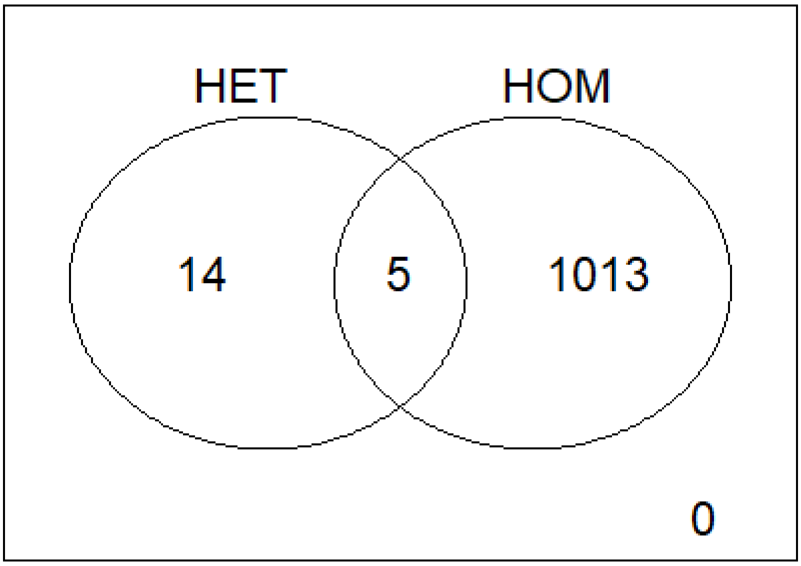

Figure 6-11. Venn diagram of splicing events found significantly differentially expressed (padj < 0.05) in both HET and HOM. It shows that HOM mutants have many more differential splicing events, and that less than half of the splicing events significant in HET were significant in HOM suggesting a difference in mechanism over time. The events which are significantly differentially spliced in both are the two events in Dnajc5, two events in Sort1, and the skipped exon skipped in Eif4h

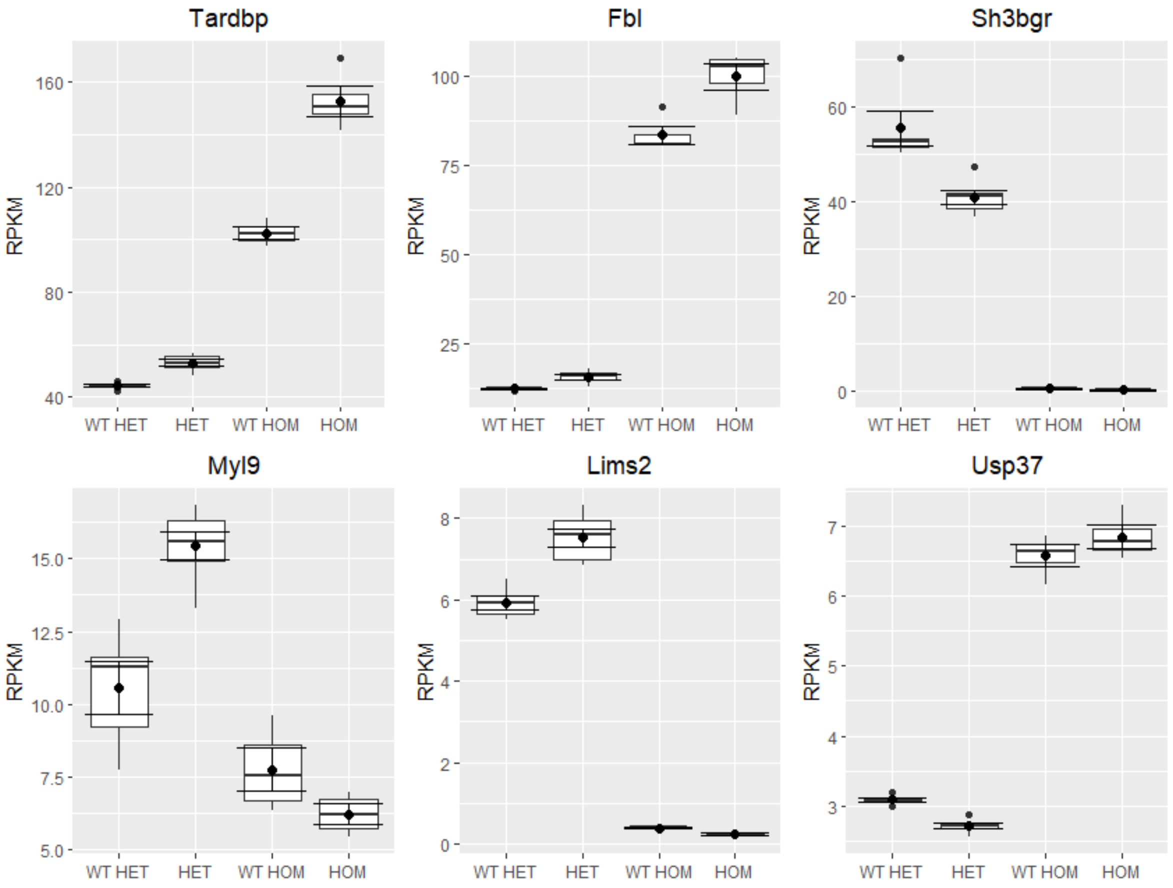

Figure 6-12. Boxplots of RPKM of all genes significantly differentially in our HET data, with genes significantly differentially expressed in our HOM data on the top row. The function of each gene can be found in Table 6-3

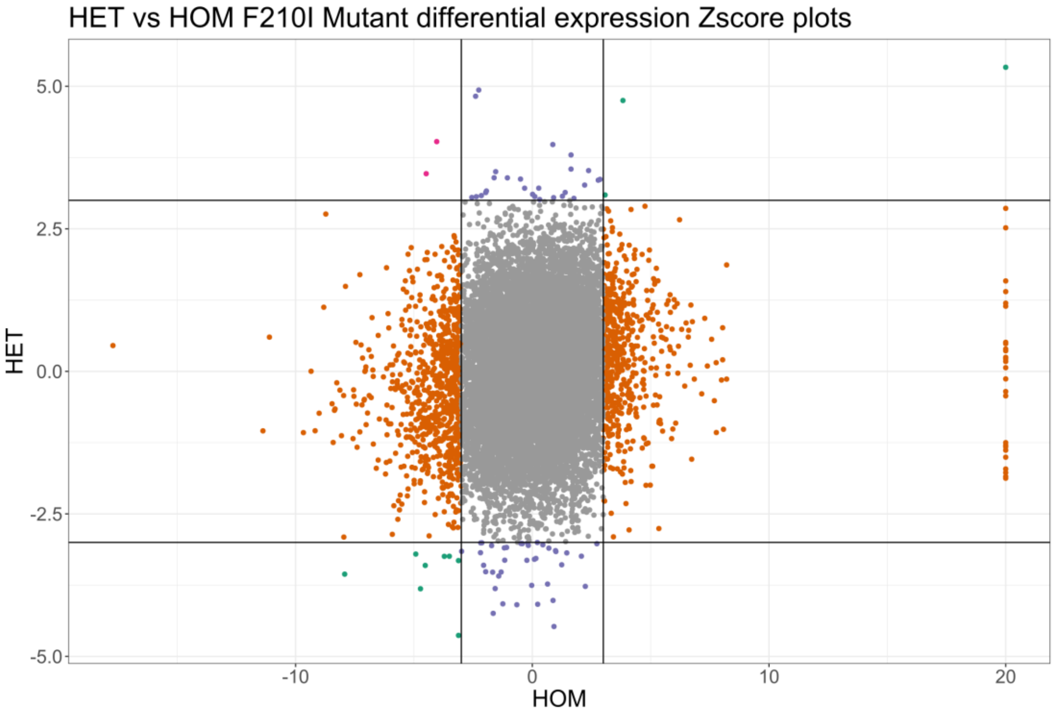

Figure 6-13. Comparison of differential expression Z-scores in HET and HOM coloured based on significance in each set. Orange points are genes which are significant in the HOM dataset and not the HET one, purple are

significant in HET and not HOM, Red are significant in both but in the opposite directions and green samples are significant in the same direction as both. An upper bound of 20 has been imposed on the x-axis due to some genes being so highly significant as to distort the rest of the graph. There are 2 samples with a Z score greater than 3 that have oppositely signed Z scores in the two samples, and 11 which exceed this threshold are signed have the same direction of change

Z-score plot of splicing (Figure 6-14) shows a marginally stronger effect, there are no differential splicing events signed in opposite directions. I finally ran a comparison of both fold change and PSI in the differential splicing events. I plotted a pair of bar plots (Figure 6-15) which evaluated whether a splicing event had a larger absolute fold change/PSI in HET or HOM, and marked it as positive or negative depending on if the larger value was positive or negative. In these graphs the HOM samples tended to have larger fold changes and the HET samples had larger PSI changes. The changes which are higher in HET than HOM present very minor changes, so their biological significance is unclear even if they are truly differentially spliced when compared to comparable WT samples. It does suggest that there may be some splicing changes which are too subtle to be picked up by DEXSeq.

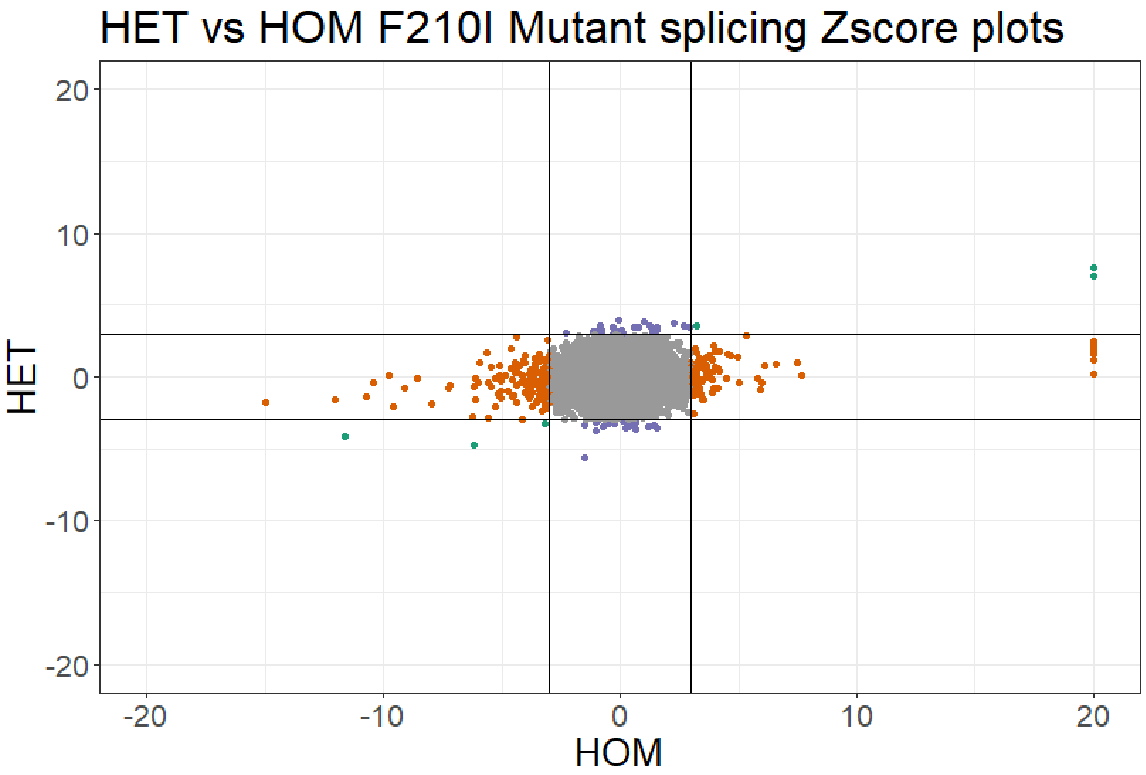

Figure 6-14. Comparison of differential splicing Z-scores in HET and HOM colored based on significance in each set. Orange points are genes which are significant in the HOM dataset and not the HET one, purple are significant in HET and not HOM, and green samples are significant in the same direction in both. An upper bound of 20 has been imposed on the x-axis due to some genes being so highly significant as to distort the rest of the graph. There are no genes with an inverse correlation and 6 where both exceed the threshold and are signed in the same direction.

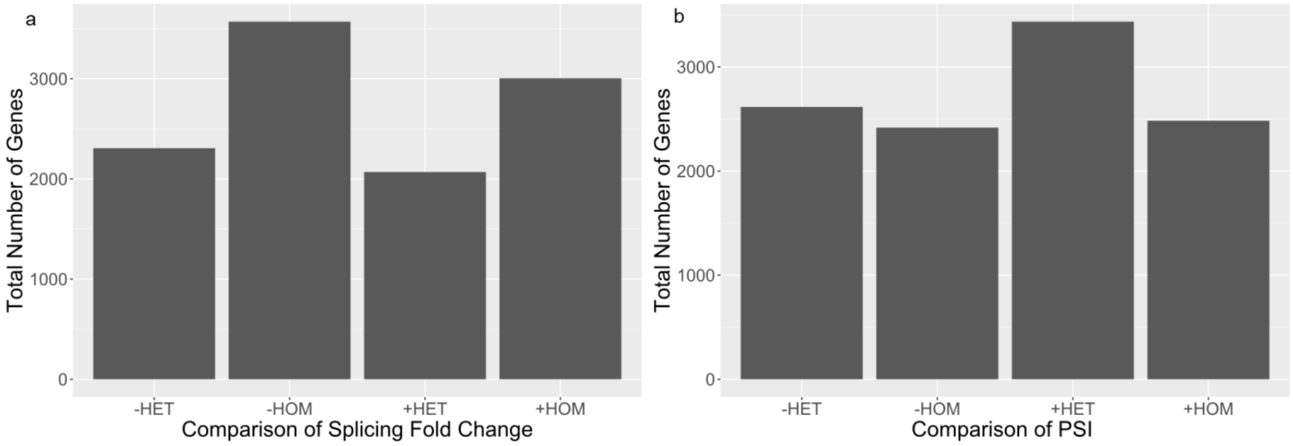

Figure 6-15. Comparison of ratio of splicing. (a) Fold Change, and (b) PSI in HET v HOM: The + or – indicates whether the sign is positively or negative in differential expression/splicing, HET or HOM says which of them had a larger absolute value

6.5. Discussion

TDP43 is a crucial protein for neurodegeneration. Studying its partial loss of function in vivo has been incredibly challenging. When deleting one allele, the self-regulation nullifies the effect, making these models less useful for providing insight into partial loss of function. Our lab had previously described an interesting TDP model where the endogenous TDP had decreased RNA-binding. Whilst the homozygous mice from this model had previously been described as causing E18 lethality, we now decided to investigate the effects of loss of RNA-binding in one allele, and the effects in adulthood. My primary analysis in this section was related to the HET data, as the HET mutation, provides a model of partial loss of function, potentially in a way that is more in line with mutations seen in patients.

Generally, both expression and splicing analysis found very few events, particularly when comparing with the large number of events found within the homozygous embryos (Fratta et al., 2018). Interestingly TARDBP was itself one of the targets found to be upregulated. This was reassuring as it underlines the fact that loss of RNA-binding pushes the system towards overregulation. Although TDP changes are linked to consequences in gene expression, these have been often linked to secondary changes, but the overlap of our HET and HOM analysis does indeed show that half of the DEGs are also present in HOMs supporting them as being strongly linked to TDP loss of function.

Our analysis of splicing also found few genes to be significantly differentially spliced. A number of these were also present in homozygote embryos, showing that our model does indeed have loss of TDP splicing function. When comparing Z score splicing between heterozygotes and homozygotes, there were no significantly discordant events, and all splicing events significant both datasets were concordantly either up or down regulate. I found relatively few differential expression results.

Generally, I found more events to be significantly differentially spliced than I did genes differentially expressed. I think I would have likely found even more if sequencing had been performed at a higher read depth. I believe this for two reasons, the genes which were found to be significantly differentially spliced had more reads than average, and when comparing with the HOM splicing data, the fold change was larger in more genes in the HOM, but HET had more genes with a higher percentage spliced in. Together those two findings suggest that there is substantial splicing activity going on as a result of the HET mutation, but at this read depth, they are subtle enough that sufficient statistical power to identify which ones are occurring for certain is not attained.

The slight enrichment of TARDBP in our HET samples is likely a result of the self-regulation mechanism of TARDBP. TARDBP auto-regulates through negative feedback(Ayala et al., 2011a). As we have disrupted its ability to bind to RNA, it is less able to reduce production leading to a slight increase in the levels of RNA found. This was one of the smaller fold changes within our heterozygous mutations so the levels of TDP-43 are unlikely to be substantially higher than endogenous levels – this is reinforced by the relatively small number of changes found in our HET models. Amongst the other significantly differentially expressed genes, Sh3bgr has been previously linked to mutations in TARDBP – and in particular has been shown to undergo cryptic splicing (Jeong et al., 2017). The others, however, do not appear to have been strongly linked previously to neurodegeneration.

The fact that half of the genes which are significantly differentially expressed in HET are not significantly differentially expressed in HOM could be due to the comparison of adult spinal cord in HET to embryonic brain data in HOM. I could not find any GO terms in common between the genes which were exclusively differentially expressed in HET so they don't seem to share a common mechanism, although given the very small number of DEGs, GO analysis has a limited value in this case. Given the large number of GO terms found in our HOM samples, and HOM not being the predominant focus of this analysis, we did not include tables or figures of their GO term analysis.

Unlike our expression analysis, our splicing analysis did identify very well-known TDP targets such as *Dnajc5*, *Sortilin*, and *Eif4h* (Dash et al., 2022; Polymenidou et al., 2011; C. Yang et al., 2014; H.-S. Yang et al., 2020; Ziff et al., 2022). The splicing change was less strong than in HOMs, confirming a dose dependency of this RNA-loss of function. Overall, the splicing results support the fact that our HET mice have a clear loss of function, and that a minor TDP loss of function may not be associated with overt genotypes. As discussed for the expression analysis, increasing sample size and depth may have allowed me to identify more subtle changes, but it is also clear that the changes which are present are minor. In support of this, our HET dataset had a larger N size (5 v 7 in HET, 4 V 4 in HOM), and a higher mean number of reads (67 million vs 34 million).

RRM2 has two orders of magnitude lower affinity for binding RNA than does RRM1(Kuo et al., 2009). It has also been shown that certain mutations which do almost completely remove the ability of TDP-43 to self-regulate when inserted into RRM1 have little clear effect when inserted into RRM2(Ayala et al., 2011a). That being said, mutations in the RRM2 domain have been strongly linked to development of ALS(Ederle & Dormann, 2017; Maurel et al., 2017; Ratti & Buratti, 2016). My research aimed to evaluate the effect of mutations to the RRM2 domain of TARDBP at endogenous levels, and whether they are capable of causing disease. These mice are a loss of function model, and, while these specific mutations have not been observed in patients, mutations linked to disease have been found to cluster around the RNA-binding motif (Ederle & Dormann, 2017; Ratti & Buratti, 2016). The heterozygous removal of the RRM2 binding motif does not appear to cause substantial neurodegeneration, or motor pathology in our mice.

I felt there might be three potential reasons why our mutation does not cause pathology:

- Mutations in the RRM2 signal of TDP alone are not sufficient to cause degeneration
- Mutations in RRM2 signal of TDP are sufficient to cause degeneration but our mutation either isn't strong enough or has the wrong mechanism
- Cryptic splicing events (Humphrey et al., 2017) observed in human samples are required for pathology and are not present in our mice

Understanding the reasons for this lack of pathology will be important to understand the mechanisms of the disease and are a good target for future investigation. Furthermore, it may be that a mild loss of function as that observed in our mice needs to be complemented with other toxic stimuli. It could be of interest therefore, to see whether our heterozygote loss of function could aggravate other ALS disease models. Even very well-known ALS causing mutations in TDP do not cause a phenotype even after 2 years in mice, and are only able to induce neurodegeneration at a mild level when in homozygosity (Fratta et al., 2018; M. A. White, et al., 2018; Ebstien et al., 2019)

As stated previously, it is likely that, due to the read depth/sample size of our samples, we are not able to pick up some splicing events which are truly different between samples, but have a relatively low magnitude. Re-sequencing samples at a higher read depth, or performing an experiment with more mouse models may be able to help elucidate the changes that are occurring and might give some insight into why this model does not cause disease. The effort may be better spent on a new model which does cause disease – either a different HET mutant or one which does cause the cryptic splicing events seen in human patients. Alternatively, these results may reveal that while changes in TDP are integral to

ALS/FTD, they do not work alone, and investigating other changes associated with this may help to produce a more optimised model.

6.5.1 Future work

The first point to be validated that has resulted from this work is whether the DEGs only found in the HET mutants are (i), truly significantly differentially expressed in HET, and (ii), truly not significantly differentially expressed in HOM. Initial analysis would likely take the form of qPCR. Investigation into changes in gene expression of either the same samples, or similar mouse models would provide insight into whether these changes are true.

While mutations found to occur in RRM2 mutant mice may not be directly linked to the effects that the RRM2 mutation has in humans, it may give guidance as to the sort of changes to look for. A useful starting point would be to test whether the DEGs found significant in our mouse models are also different in patients with similar mutations.

7. Identification of degenerating neurons using machine learning

7.1. Publication

Elements of this work have been published in Scientific Reports (Mejia Maza et al., 2021).

7.2. Introduction

Neuromuscular junctions (NMJs) are the site at which neurons transmit signals to muscles to provide instruction on contractions. They are made of the terminal of a motor neuron, receptors on a muscle fibre, and a Schwann cell which sheaths it. Under normal circumstances, they are not static objects, undergoing continuous innervation and denervation (Slater 2017a). While some irreversible degeneration is part of the natural aging process, it is also one of the earliest signs of some common human neuromuscular disorders (Dupuis et al., 2009; Willadt 2018).

In order to examine neurodegeneration in mice, both manual and automatic methods are used. The manual method entails a visual assessment of innervation status, but classification criteria can differ between labs or studies. One of the most common automatic methods is the ImageJ plugin NMJ-morph (Jones et al., 2016). It is widely used because it accurately and efficiently measures morphological features of NMJs, and can function on several species (Bohem et al., 2020). It is not ideal however as it requires maximum intensity projections, which means that it will not function on the majority of NMJs as they have complex 3D shapes through their interaction with muscle fibres. It also requires manual thresholding of fluorescence, which reduces comparability between studies as it leaves things to the discretion of the person performing the analysis.

Machine learning is a tool which has gained increased prominence. This is predominantly due to a combination of increased computing power, increasing amounts of available training data, and optimisations to existing models. Some work has already been done to train machine learning models to classify cells based on images from microscopy(Kan, 2017; Zinchuk & Grossenbacher-Zinchuk, 2020). This has included training models to recognise various phenotypes as well as cells in an apoptotic or necrotic state (Li et al., 2021; Sommer & Gerlich, 2013; Verduijn et al., 2021). Given the existing literature which uses machine learning tools for similar purposes, and the results of our initial analysis, we felt that it would be fruitful to use machine learning to classify neuromuscular junctions.

My Colleague Alan Mejia Maza has developed a method called NMJ-Analyser (Mejia Maza et al., 2021). NMJ-Analyser enables quantitative assessment of the native 3D conformation of NMJs using an automatic thresholding method, to accurately capture their morphological features. Using the morphological features produced by NMJ-Analyser I worked with Alan Mejia Maza to train machine learning models which could automatically classify neuromuscular junctions into degenerating or healthy.

7.3. Methods

Three ALS mouse strains and a CMT2D mouse model with corresponding age and sex matched littermates were sued to study pathology. The full details of each model can be found in Table 7-1. Lumbrical and flexor digitorum brevis (FDB) muscles located in the hindlimb paws of mice were dissected and stained, as described in (Sleigh, Burgess, et al., 2014; Tarpey et al., 2018).

I mages of NMJs were obtained using a Zeiss LSM 710 confocal microscope (Zeiss, Germany). All strains were imaged at 512 × 512 resolution, except for the FUSΔ14/+ strain, which was at 1024 × 1024. Z-stack images were decomposed into individual pre- and post-synaptic planes or posterior analysis (.TIFF, .PNG and .JPEG).

Images were initially manually assessed and classified using Volocity 3D Image Analysis Software (version 6.5, Perkin Elmer), a family of software products for 3D image acquisition and analysis. When a nerve terminal and motor endplate overlapped with at least 50% coverage the NMJ was considered ‘fully innervated’. When it did not it was considered denervated – distinctions between full and partial denervation were not used for the purposes of training my machine learning models.

Figure 7-1 shows how NMJ-analyser acts to process stacks of images into morphological information. First, pre-processing of the Z-stack images of NMJs is performed. This is to select clearly stained, well defined structures among other parameters. Each NMJ stack is then aligned to create a 3D image of the NMJ. Thresholding is then performed for clarity. The final step in automatic analysis is measurements of the morphology, comparing intensity of each component as well as distance between them at various points.

Table 7-1. Summary of mutant mice. MGI Mouse Genome Informatics number. From (Mejia Maza et al., 2021)

Strain/MGI	Mouse model	Background	Protein expression	Age (months)	Numbers/sex	Disease state	ref
SOD1 $^{G93A/+}$ MGI: 2448770	Transgenic	C57BL/6N-SJL	~ 20-fold overexpression	1	5 WT, 5 mutants (male)	ALS, pre-symptomatic	(Bilsland et al., 2010; Gurney et al., 1994)
				1.5	5 WT, 5 mutants (male)	ALS, pre-symptomatic	
				3.5	4 WT, 9 mutants (male)	ALS, late symptomatic	
Gars $^{C201R/+}$ MGI: 3760297	ENU mutagenesis	C57BL/6J	Physiological expression	1	6 WT, 6 mutants (male)	CMT2D, early symptomatic	(Achilli et al., 2009; Sleigh, Grice, et al., 2014)
				3	3 mutants (male)	CMT2D, symptomatic	
FUS $^{\Delta 14/+}$ MGI: 6100933	Knock-in, partial humanization	C57BL/6J	Physiological expression	3	4 WT, 4 mutants (male)	ALS, pre-symptomatic	(Devoy et al., 2017)
				12	4 WT, 4 mutants (male)	ALS, symptomatic	
TDP43 $^{M323K/M323K}$ MGI: 6355456	ENU mutagenesis	C57BL/6J-DBA/2J	Physiological expression	12	5 WT, 5 mutants (female)	ALS, pre-symptomatic	(Fratta et al., 2018)

The full code for creating the machine learning models as well as the machine learning models produced for this section can be found on GitHub: https://github.com/SethMagnusJarvis/NMJMachineLearning

The morphological data I used to create the machine learning models was produced through analysis with NMJ-Analyser[6], and I used the package caret (Kuhn 2008) to train the machine learning models. 4648 samples were used for training, 1161 were used for testing, these samples were selected randomly. Modelling was performed using random forests with 10-fold cross validation. The seed was set to 1337, and for the equalised models, equalisation was performed by grouping by type and using the sample_n function to select 1000 of each of healthy and degenerating NMJs. The full list of variables used to train the models, and their meaning can be found in Table 7-2.

Table 7-2. Overview of NMJ morphological features. From (Mejia Maza et al., 2021)

NMJ component	**Features**	
Nerve terminal/motor endplate		
Integrity	Cluster_dist	Average distance between fragments or clusters
	Cluster_size	Average size of clusters
	Cluster_numbers	Number of cluster or fragments
	Fragmentation	Fragmentation index (1–1/Cluster numbers)
Shape	Non-compactness	Measures how non compact an NMJ component is
	Shape factor	Numerical description of the 3D shape of an NMJ component and its relation of becoming a more irregular shape
	Rugosity, internal	Number of internal faces, internal irregularity
	Rugosity, external	Number of external faces, external irregularity
Size	Length (µm)	Distance between the major-axis endpoints
	Surface/volume ratio	Measures the amount of surface per unit of volume of an NMJ component
	Surface	Measures the 3D geometrical uppermost layer of an NMJ component. It is also known as surface area of the tridimensional surface
	Volume (µm 3)	Measures the amount of space occupied by an NMJ component
Pre- and post-synaptic		
Interaction	Coverage	Volume of nerve terminal staining as percentage of endplate volume
	Intersection	Area of intersecting area between NMJ components, overlapping
	Average dist	Distance between NMJ components
	Mass-distance	Distance between centre of mass
	Hausdorff distance	Maximum distance between the nearest point between NMJ components

7.4. Results

7.4.1 Statement of purpose

My colleague Alan Mejia Maza has developed a method called NMJ-Analyser which takes as input a stack of confocal microscope images of dissected immunostained neuromuscular junctions. The method then outputs a set of quantitative information based on various morphological features. Our objective was to see whether the output of NMJ-Analyser was sufficient to train a machine learning model to determine whether NMJs were healthy or degenerating. An overview of the workflow can be seen in Figure 7-1.

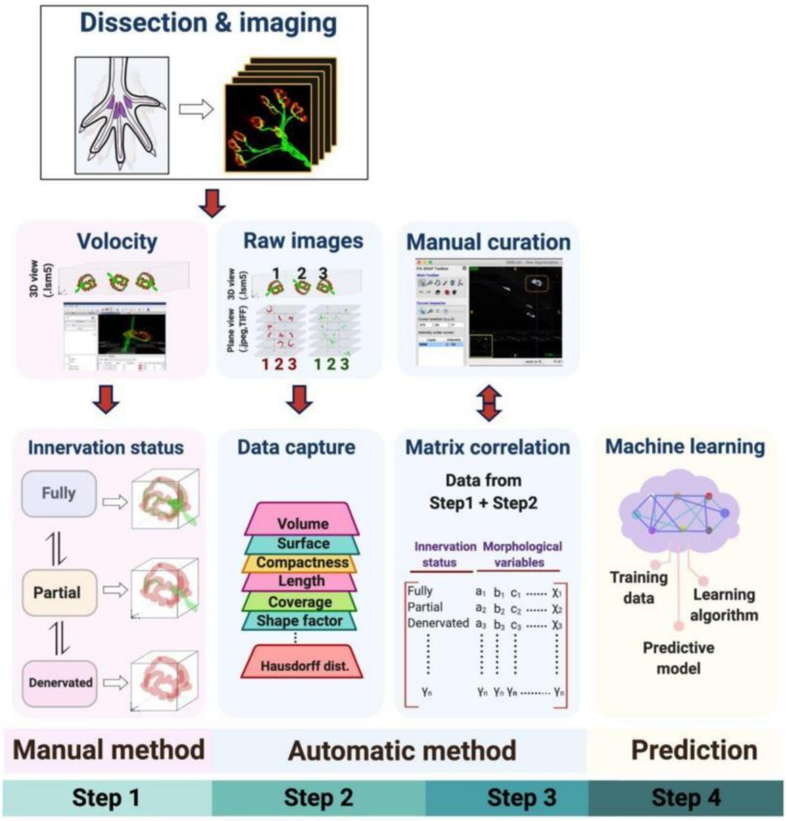

Figure 7-1. Overview NMJ-Analyser's objectives and method. Figure produced for Mejia Maza et al

7.4.2 Overview of NMJ-analyser (Performed by Alan Mejia Maza)

As can be seen in Figure 7-1, processing of the data before it is input into my machine learning model for training involves 3 steps:

1. Dissection, staining digitalization and manual assessment of NMJ status
2. Application of NMJ-Analyser to the stacked raw images and extraction of features
3. Matching manual and automatic assessment

Alan Mejia Maza dissected mouse muscles and stained them. He then digitised the NMJ structures which appear in the staining. Manual classification of the NMJs was then performed, annotating whether they were healthy, partially innervated, or fully denervated.

These image stacks were then passed through the NMJ-analyser python script developed by Alan Mejia Maza. This analyses the image stacks and generates twelve biological relevant parameters for each pre and post synaptic structure, and five for the interaction between the two NMJ components. Several of the parameters are generated for both red and green stained areas resulting in 58 variables in total.

The final step was to create a matrix containing the qualitative measures about innervation and the quantitative results from NMJ-analyser. I used this combined final matrix to train my machine learning models.

7.4.3 Exploratory analysis

In our initial PCA comparing the results from NMJ-analyser (Figure 7-2) we found that there was some visual separation between healthy and degenerating NMJs which made us think that it might be possible to train a machine learning algorithm to classify samples instead of the person collecting the data, which would decrease the level of subjectivity in results. This would increase the accuracy of results, making studies more reliable and increasing the ease of comparison between studies by removing subjectivity in interpretation. It would also save time for researchers as most of the process is automated.

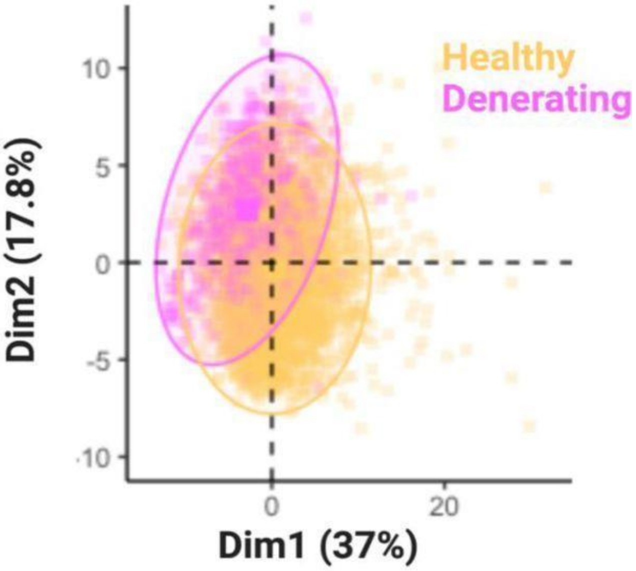

Figure 7-2. PCA plot of NMJ-Analyser's results on our NMJ training dataset. Each dot represents a neuromuscular junction. Figure produced for Mejia Maza et al

7.4.4 Training model

There are 6 commonly used types of machine learning algorithms for classification:

- Logistic regression which uses a sigmoid function to return probability of a label
- Decision trees which build hieratical branches of trees to subset data
- Random forests which are a collection of decision trees that generate a consensus
- Support vector machines which aim to insert a hyperplane between objects within an n-dimensional space to classify them
- k-nearest neighbour which places the data within an n-dimensional space and uses distance between objects to classify them
- I Bayes which calculates conditional probability based on prior knowledge

Logistic regression was not considered suitable for this analysis as it works best on two groups and we were initially intending to try to separate into partially and fully degenerating during classification. Decision trees are prone to overfitting and are considered a less optimal version of random forest classification. Naive bayes was not considered suitable as it is optimised for completely uncorrelated datapoints which our data isn't. This meant that Random Forest, Support Vector Machines, and k-nearest neighbours were considered to be the most suitable models to our classification.

Of those three models I decided to train classifiers using both random forest (RF) and support vector machine (SVM) models. This was because I had existing familiarity with those types of models, and there were well built tools which I could use to train them. Initial attempts at model creation were rather poor – models only had about a 70% accuracy with less than 50% sensitivity in prediction. I thought that this was likely due to the small number of degenerating NMJs so asked Alan Mejia Maza to collect more degenerating motor-neurons. The initial dataset had 362 degenerating NMJs, and 3779 healthy NMJs. The final dataset used to train the models after Alan Mejia Maza had collected more data consisted of 1179 degenerating, and 4630 healthy NMJs.

When training both SVM and RF models, I found that SVMs took up to 10x – between 5 and 10 minutes rather than under one - longer to train making prototyping marginally more difficult, and were about 10% less accurate in our initial exploration of our dataset. When creating the Random Forest model trained on our final dataset, I used 80% of the data as a training dataset, then used the remaining 20% as a test dataset. As can be seen in Figure 7-3, the overall classification accuracy is very high.

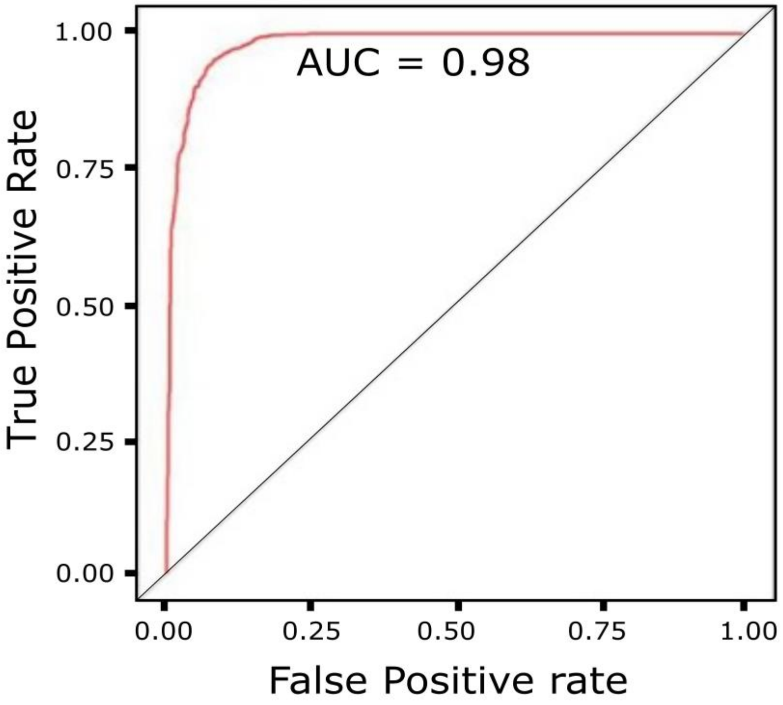

Figure 7-3. AUC curve of our RF model

We also created a model which had an equal number of randomly selected healthy and degenerating NMJs. This model had approximately 5% lower overall accuracy, but was better at identifying degenerating motor neurons so was primarily intended for people who were expecting their data to contain more degenerating motor neurons.

Table 7-3. Results of the machine learning model on the test data

10-fold RF	
Accuracy	0.9552
95% CI	(0.9417, 0.9664)
P-value (Acc > NIR)	<2e-16
Mcnearman’s test (p-value	0.6774
Sensitivity	0.8809
Specificity	0.9741
Prevalence	0.2024
Balanced accuracy	0.9275

7.4.5 K-fold cross-validation

In order to decrease the likelihood that our model had been overfit, I created a model with 10-fold cross-validation i.e., one which had been optimized by creating 10 models, each from different random subsets of the data, and combining their results to increase generalisability. In Table 7-3 I can see that the sensitivity (true positives) is lower than the specificity (true negatives). A likely cause of this bias is the relative scarcity of degenerating samples in our dataset. While I have trained on about 6000 NMJs, only 1/5th of them are degenerating. Table 7-4 shows a breakdown of the top 10 morphological features ranked by their permutation importance (the effect that removing the sample from the model has on prediction accuracy), and both the relative permutation and Gini importance – a measure of the proportion of splits a characteristic is involved in - of each feature. As I can see, the two most important factors for permutation were the shape factors of the endplate and the nerve, while the factor with the highest Gini was “intersection”, a measure of the interaction between the nerve terminal and end plate.

Table 7-4. Breakdown of morphological features on machine learning model accuracy

Morphological Features	Type	Permutation Importance	Gini Importance
Shape factor of endplate	Shape	100.00	47.43
Shape factor of nerve	Shape	58.39	18.63
Intersection	Interaction	35.08	100.00
Integral rugosity of nerves	Shape	34.13	31.29
Integral rugosity of endplates	Shape	31.47	17.74
External rugosity of nerves (green)	Shape	27.40	22.89
Coverage	Interaction	24.81	40.13
External rugosity of nerves (red)	Shape	24.09	24.73
Nerve non-compactness	Shape	23.35	43.20
Surface/volume ratios of nerves	Size	21.12	24.26

7.5. Discussion

As discussed in the paper introducing the method (Mejia Maza et al., 2021), NMJ-Analyser shows higher sensitivity to degeneration than both manual NMJ counting methods and NMJ-Morph. The ability to provide information on the 3D structure independently of shape is important, as NMJ topology may change during disease. This helps to distinguish it from NMJ-Morph (Slater 2017a; Dupuis 2009). Unlike NMJ-Morph it requires no judgement from the user, removing the need for manual thresholding or manual classification of cells.

A comparison between NMJ-Analyser and NMJ-Morph was performed. It found that NMJ-Analyser has a higher sensitivity than NMJ-Morph, and is capable of quantifying NMJs in their native 3D structure which is important due to changes in NMJ topology which have been observed in disease(Marques et al., 2000; Moloney et al., 2014; Slater, 2017; Willadt et al., 2018). This comparison has not been included in this chapter as it was predominantly performed by Alan Mejia Maza and I did not consider it relevant to my work which was training the machine learning models. It can be found in the published paper(Mejia Maza et al., 2021).

The machine learning models I trained are specific and accurate at classification, given the relative rarity of degenerating NMJs. The main model's higher accuracy overall does, as stated, come at the cost of slightly worse performance at identifying degenerating NMJs. I felt that this was a worthwhile trade-off. If any users disagree, I created a model trained on an equal number of degenerating and healthy samples. This reduces the overall accuracy as it identifies more NMJs as degenerating than there are in reality. It was created for cases where there is a higher-than-normal proportion of degenerating NMJs, or where accuracy in identification of NMJs is the most important variable as opposed to overall levels.

In the future ideally there would be a more comprehensive integration of NMJ-Analyser and my machine learning model. Currently, the user needs to run NMJ-analyser separately then input the results matrix for classification in R, ideally, I would integrate this in a single pipeline or script. It may also be valuable to train a new machine learning model on results from other labs and with a larger cohort of degenerating NMJs, further improving generalisability of the model and possibly removing the need for the equalised model.

7.5.1 Future Work

While my analysis was performed only on data collected by Alan Mejia Maza and some of his colleagues, there has been some usage of my tool by external collaborators. Initial use by external labs seems to suggest that the model may be overfitted. This would likely be improved by including training and testing data created by other labs. Further validation from more diverse sources will be useful, particularly if users are willing to manually classify some NMJs and test the accuracy.

With regards to further development there are gains to be made in both usability, and the models themselves. On the usability front, it would be ideal if we could fully automate this system: have a user input their stacks of confocal images, analyse them, and automatically return both the morphological information and whether each NMJ was degenerating or healthy. Barring that, creating a webapp where a user could input their results from NMJ-analyser and receive the resulting counts and confidence intervals would be useful. As it is, I have automated as much as possible and given guidance as to how the analysis can be performed, but it is still more complex than would be ideal.

There are 4 machine learning changes it would be useful to test in future work. These changes are:

- Creating a model with data from more labs to potentially increase generalisability
- Gathering enough data to allow the model to be able to distinguish between healthy, partially degenerating and fully degenerated NMJs
- Producing models for each species as structure of NMJs will differ
- Using a computer vision method to directly predict the status of the NMJs

All of these tasks bar possibly the creation of a machine vision model would require the collection of more data.

Currently, the model is trained on a relatively narrow dataset from a few labs and exclusively on mouse data. If we could gather more data from a range of types of NMJs, ideally from different labs, it would substantially improve the generalisability of the work.

The initial scope of the model was aiming to classify between healthy, partially and fully degenerating NMJs. Sadly, this resulted in performance that wasn't much better than chance, thus I combined partially and fully degenerating NMJs into a single category. If we could gather more data, particularly in partially and fully degenerated NMJs, we would be able to create a more versatile model that could discriminate between the types of NMJs. Similarly, while the processing performed by NMJ-analyser will help ensure some standardisation, it would be ideal if a model could be trained using data from multiple labs to ensure consistency. This model is currently trained exclusively on mouse data. It is currently unclear if being able to classify in multiple species would require a single model trained on multiple species, or one model produced for each species, but in any case, further investigation into development of a method for analysing the NMJs of multiple species would be a worthwhile endeavour.

My initial intention with this project was to develop a machine learning model which could interpret either compiled or raw image stacks from the confocal microscope. This would have required substantially more work from Alan Mejia Maza and would have been doubling up effort to an extent so was put on hold indefinitely. It would be interesting to see how a vision model would compare to the model based on the results of NMJ-analyser. While this might result in less information being output– only returning a classification of the NMJ and confidence–interval – it would be a prime target for automation.

8. Conclusions

My PhD has primarily aimed to elucidate possible mechanisms through which ALS and FTD cause neurodegeneration by using RNA-sequencing methods. This has both been through evaluation and development of new methods and pipelines, as well as analysis of data from both disease models and patient brains. My role in projects was that of leading the research, analysing data provided by colleagues, and to help determine the best path forward and contribute to the larger picture.

In Chapter 3, I compared low read depth QuantSeq, a cost-effective technique to assess gene expression, to the more established total RNA-seq protocol. As a method, its utility in differential polyadenylation analysis means it has its own niche. I wanted to evaluate how generally useful it was as a lower cost potential replacement for total RNA-seq analysis when splicing is not of interest. Other comparisons at both high and low read depths have been performed, but they have been direct comparisons to mRNA methods rather than total-RNA (Corley et al., 2019; Ma et al., 2019).

The primary reason this data was sequenced was as part of a larger study on the mechanisms by which FUS mutations cause disease (Humphrey et al., 2020). The paper found that our humanised mouse mutation d14 limits the role FUS is able to play in autoregulation and splicing, and the method through which it acts – by leading to a dysregulation in intron retention events.

The section of this research which I led was the comparison of the results of RNA-seq and QuantSeq observed in the results of this thesis. This work has subsequently been published as Jarvis et al., 2020. The contributions to the field are two-fold: (1) it demonstrates the utility of using low read depth QuantSeq for initial analysis of changes on many samples, and (2) it questions the ability of QuantSeq to accurately determine DEGs.

I found that there were large differences in both the number of differentially expressed genes (DEGs) and which genes were found to be differentially expressed. Importantly, our datasets comprised both an “extreme” condition, where full knock-out of an RNA binding protein was present vs normal littermates, and an intermediate condition, where a partial loss-of-function mutation of the same RBP (FUS d14 mutation) was compared to its own littermate controls. This was particularly useful as it allowed us to see that discordance increased between the two sequencing approaches when the case when analysing the more subtle mutation, the FUS d14 mutation; as the magnitude of the changes were smaller, QuantSeq did not have the power to reach statistical significance in most cases.

As discussed in the future work section, this may present an opportunity for tool development. This does highlight a potential opportunity for new tools/new statistical methods to be developed to be able to more effectively analyse this kind of low read depth data, and this is a field of expanding interest since the rise of single cell sequencing, where shallow 3’ end sequencing is the norm. Overall, I found that low read depth QuantSeq may be a useful tool for initial exploration of datasets that are likely to have large changes.

I believe that low read depth QuantSeq is likely to be a useful measure in the short to medium term. The cost of RNA-sequencing is still a real factor in planning experiments. Until acquiring samples becomes a more limiting factor than this cost, low read depth QuantSeq may find a use.

One hurdle in the immediate future is the how well it can work with long read sequencing. The field seems increasingly to be moving towards long read sequencing as it can provide a lot more information when analysed properly. In long-read sequencing, I believe that QuantSeq will lose its primary advantage over other methods. Transcripts are generally sequenced in their entirety in long-read sequencing. This means that QuantSeq’s advantage at low read depths will be reduced as one read per transcript will be the default. This means that long read sequencing will only find one read per transcript, and is more able to detect splicing events(Buck et al., 2017; Liu et al., 2017; Oikonomopoulos et al., 2016).

Currently, the two main methods for long read sequencing are nanopore, and PacBio SMRT sequencing. The largest disadvantage of both is their relatively high error rate compared to short read sequencing methods (Kono and Arakawa, 2019; Amarasinghe et al., 2020). For many RNA-related applications the error rate is not as limiting as for genomics investigations. Further, both methods are gradually becoming less error prone. If QuantSeq is able to be utilised for another niche that sets it apart from other RNA pulldown methods in long read sequencing it may be useful, otherwise it is unlikely to see increased adoption. Given, total RNA-seq will be inexpensive enough, and the tools for analysis will be developed enough that all of the information and more which QuantSeq provides will be cheaply available.

In Chapter 4 I analysed bulk RNA sequencing data from *post mortem* human brains. Although this approach is a direct sampling of disease, RNA sourced from brains has known limitations due to relatively low quality, in part due to degeneration occurring post mortem. This makes analysis challenging. Samples came from a mix of healthy brains, as well as from FTD patients with either TAU or C9orf72 mutations. This research was intended to both compare human brains with specific FTD etiologies to control brains, as well as to compare them to one another. To my knowledge, prior to my work a comparison had not sought to compare two subtypes of FTD directly to one another.

I found that TAU pathology resulted in more DEGs than did C9. There also seem to be substantially more differences in cell types between CTL and TAU than between CTL and C9 samples. It is not clear why these differences in cell types are observed, but adding cell types as a covariate when examining differential expression was found to reduce the number of DEGs which may indicate a reduction in biological noise in some cases.

The goals of my analysis of FTD brains were to provide insight into differences between the two pathologies, and, act as a useful pilot study for combining methods of brain cell type deconvolution with RNA-seq data. The results of the two aims are inextricably linked.

The clearest result initially was that TAU pathology seemed to cause far more DEGs than did C9orf72 pathology. Since the conditions do not have markedly different disease progression rates and cause similar underlying conditions, this was unexpected (Van Langenhove et al., 2013).

When evaluating GO terms, I feel the two most interesting findings were the appearance of terms relating to the extracellular matrix and collagen. When not correcting for cell types, these were the most significant GO terms when comparing our C9 patients to controls. This is related to findings in mouse models expressing DPRs (Milito et al., 2023), as they also found changes in the extracellular matrix. When correcting for cell types, TAU patients also found GO terms related to the extracellular matrix to be the most significant highlighting a link between the mechanisms of the two aetiologies.

When I examined the relative enrichment of brain cell types, I found obvious changes in TAU when compared to control samples. At this time, it is not clear how much alterations in the cell composition or changes in expression within cell types drive the differences observed between the two pathologies.

When I added the levels of cell types in TAU vs control in our input to DESeq2, I found that the number of DEGs in TAU decreased to similar levels of those seen in C9 pathologies. This did seem to cause abnormalities in the fold changes with very high peaks at specific fold change levels. This suggested that the method I used for integrating them was not ideal. I do not recommend including relative cell type enrichment when there are minimal changes between two conditions. Given the irregularities observed in fold changes, and the substantial reduction in DEGs across the board when used, I think that, as it stands, the method I used may add biological noise when only small differences between cell types are present. Using simulated data may be a possible useful avenue for investigating whether the issue is specifically with my data, or with the method for integration.

The analysis I have performed has highlighted that there are some clear differences in the effect of C9 and TAU both on RNA metabolism, and on specific cell types within the brain. It has also provided insights into a method to integrate brain cell type data into differential expression analysis, as well as drawbacks with the method. I hope that it also further adds to the evidence of the potential utility of this method in order to reduce biological noise and find true DEGs. I think that future studies should continue improving tools for analysis of cell type specific differences in bulk RNA-seq brain data. I also think that they should investigate cell type specific changes within these two mutants as a possible additional mechanism for neurodegeneration.

In Chapter 5 I analysed bulk RNA-sequencing data from mice models with two different products of the C9orf72 repeat expansion mutations (PR and GR), and relevant littermate controls at both 6 months and 12 months old. Mouse models allow us to study early disease phases and to overcome the technical limitations found when using *post mortem* human material.

These mice in particular aim to mimic the effect of two DPRs which have been most strongly linked to toxicity in C9orf72ALS-FTD (Kwon et al., 2014; Mizielinska et al., 2014). As stated in my introduction, there are three predominant hypotheses of how the repeat expansion observed within C9orf72 causes disease (Mizielinska et al., 2013; Gendron et al., 2014; Zhang et al., 2018). These are:

- Reduced expression of *C9orf72*
- Presence of large numbers of RNA foci (aggregates of RNA produced by both sense and antisense G_4C_2 repeats)
- Production of dipeptide repeat (DPR) proteins through repeat-associated non-ATG (RAN) translation.

Our study predominantly aimed to test the pathogenic effects of DPR proteins in the absence of RNA foci. It also aimed to observe differences in the effects of PR and GR DPRs which may be able to provide further information on the mechanism of disease. Of note, our models all have loss of a normal allele, therefore recapitulating the partial loss of function observed in patients.

Phenotypes had not started to develop at 6 months, and this was echoed in the RNA-sequencing data, with minimal changes to expression in any genes. In the 12-month-old samples, I found large numbers of DEGs in both PR and GR, with GR having more changes overall. As was seen in the C9 mutants in the FTD brain data, there did not seem to be any significant changes in cell types between any of the datasets. I therefore did not feel it was useful to include relative cell type enrichment in differential expression analysis.

The primary goal of my analysis was to investigate the effects of expansion mutations on RNA-expression. Changes I observed in the number of DEGs are consistent with disease progression observed in the mice (Milito et al., 2023). As stated previously, very few genes are significantly differentially expressed at 6 months in either mutation. A large number are found to be significant at 12 months, at which point, symptoms have started to develop.

Some of the specific findings we have correlate with existing literature. Our PR mutant mice appeared to show the strongest changes in GO terms relating to histone modification and DNA repair (Zang et al., 2019). A study on the effect of overexpression of PR in mice found that there were changes observed in terms related to histone modification, my findings show that these changes occur even when these DPRs are expressed at closer to endogenous levels than existing overexpression models. Meanwhile, when evaluating human cells expressing poly(PR) DPRs, markers of DNA damage responses have been found (Farg et al., 2017). My work supports the causative nature of the DPRs in this DNA damage, and with further study may highlight some more specific ways in which this occurs.

Our GR mutant mice showed changes relating to ubiquitin dependant protein degradation and RNA splicing. Splicing is known to be dysregulated in C9FTD/ALS patient brains (Prudencio et al., 2015). Disruptions in splicing have also been well characterised in TDP43 models. As these changes in GO terms related to splicing occur in the absence of TDP mislocalisation (Milioto et al., 2023), it appears to show that GR mutations cause disruption to splicing independent of known pathways by which TDP causes disease.

One surprising finding was that there did not seem to be a change in the level of expression of C9 itself in the GR mutant mice. This is a change which needs further validation as the position of the knock-in insertion should substantially reduce expression of C9 when compared with WT controls. This is particularly because reduction of C9 at the RNA and protein level was shown via RT-PCR and western blotting (Milioto et al., 2023).

Overall, my research suggests that PR and GR DPR proteins do cause disruption to levels of RNA expression in the absence of RNA foci. It also demonstrates some of the mechanisms by which each may act. They are not, however, able to cause the full symptoms of disease in the mice up to 12 months of age. This suggests other potential disease mechanisms may also be involved such as RNA foci.

In chapter 6 I analysed data from mice with a mutation in another crucial gene for ALS/FTD: Tardbp (or TDP-43). The RRM2 mutation of TDP-43 has been previously studied, and disruptions in the RNA binding motifs of TDP-43 induce TDP-43 loss of function, which has been linked to the development of ALS (Ederle & Dormann, 2017; Ratti & Buratti, 2016). We have previously shown this model to be a bona-fide loss of function model (Fratta et al., 2018).

Our models were created to evaluate the effects caused by mislocalisation without full knockout of the gene. Other models such as those in Table 6-1 and (Ayala et al., 2011b; Eréndira Avendaño-Vázquez et al., 2012) have shown that it is difficult to make a model to study *TARDBP* LOF mutations in vivo due to the strong autoregulation of the gene. This results from TDP-43 binding to its own UTR.

Our F210I model has a mutation in the RNA binding capacity which allowed for the study of loss of function without using mice who had fully knocked out TDP-43. While I was involved in the analysis of our homozygous model, I predominantly analysed the mice heterozygous for the mutation, and differences between the heterozygous and homozygous mechanisms of action.

There were very few significant DEGs in our HET data. One of the genes which were found to be significant was Tardbp. Of the others, Sh3bgr had been most strongly linked to neurodegeneration in previous studies. This includes being shown to undergo cryptic splicing in a previous study where TDP-43 was conditionally deleted (Jeong et al., 2017).

When comparing DEGs in HET to those in HOM, half were not found to be significant in the HOM. This may be a difference due to the ages of the mice, the area from which the RNA was taken (spinal cord vs brain), or due something within our analysis i.e., there is a true difference but it does not pass the significance threshold or is in some other way not observed. If there are true differences, they are likely due to differences in levels of TDP-43 or a result of the autoregulation mechanisms which result in equilibrium being restored. It is not obvious whether the differences are true differences or a function of differences in the data. Validation should be performed before a mechanism by which they differ can be discussed.

While HOM versions of our model do seem to almost entirely remove the autoregulation ability of TDP-43, one functional version of the gene seems to be enough to restore equilibrium. As this occurs at endogenous levels, it means that mislocalisation and loss of nuclear function are able to cause pathogenesis. More endogenous loss of function models are needed to potentially isolate which specific effects play the largest role, but our findings show the substantial effect which mislocalisation has, and the ability of autoregulation to restore equilibrium when a single copy is present.

If the differences between the HET and HOM mechanisms are true, further examining them may allow for a broader understanding of how the self-regulation acts, and how this loss of self-regulation causes mutations in TDP-43 to become pathogenic.

In chapter 7 my topic was not transcription, but another very relevant mechanism for ALS biology: the degeneration of neuromuscular junctions. I trained a machine learning model to classify neuromuscular junctions into healthy and degenerating. A common process in the field of neurodegeneration consists of using confocal imaging to evaluate relative levels of denervation of neuromuscular junctions. My colleague Alan Mejia Maza developed a method of analysing confocal image stacks to produce information on morphology of NMJs. I used this information to develop two machine learning models; one made up of the full dataset I had been provided with, and one with a balanced number of degenerating and healthy NMJs. Both of these models have at least a 90% overall accuracy on their respective datasets, but the balanced model is more likely to classify NMJs as degenerating which may be useful in unusually weighted datasets.

Currently analysis of confocal imaging requires significant input from researchers. I hoped to automate this to the maximum possible extent. My initial intention was to train a machine learning model to directly classify NMJs based on confocal image stacks to directly classify NMJs. However, since Alan was generating data which was both independently useful, and could be used to train a machine learning model, collaboration with him made sense.

Through this collaboration, and the development of this machine learning model, I have created a tool which has the potential to be useful to future researchers (Mejia Maza et al., 2021). I hope that this tool will help to both standardise the analysis and classification of NMJs, and reduce the effort and time which classification of NMJs takes.

In conclusion, my main contributions to the field have been:

- the development of a tool to automate classification of neuromuscular junctions,
- a comparison of two methods of differential expression analysis,
- analysis of the effects of mutations in *C9orf72*, *TARDBP*, *SOD1,* and FUS in both human and mouse data
- investigation into integration of cell type proportions into differential expression data, and highlighting future research

Validation of changes which I have observed should be a focus of future work. Additionally, further development of tools may be a fruitful endeavour. The two areas which I would focus on would be building a gold standard RNA library for different cell types from many different backgrounds and preparations, and improving the statistical methods used to be more able to detect small changes. Alternatively, given the results of (Sutton et al., 2022), the creation of a tool which has multiple reference libraries which a user could select depending on the source of their data would substantially improve the accessibility and use of cell type disambiguation within brain sample analysis. Furthermore, easy integration of this tool with differential expression analysis would improve the accuracy of bulk RNA-seq analysis in samples from the brain.

Improvement of statistical tools for analysing differential expression would have hugely broad reaching effects. Since the cost of sequencing remains high, sequencing of large number of samples may be seen as prohibitively expensive, resulting in the statistical power of studies being relatively low. This can be seen in the low sample size of my work. Currently, there appears to be a trade-off in tools for differential expression analysis tools between precision and true-positive rates (T. Wang et al., 2019). Past a certain point however, a fundamental issue with statistics is reached. Depending on the number of samples, the effect size, and the read depth, it is not possible to determine whether an effect is due to chance based on the data observed. Therefore, finding ways to optimise for small study sizes is a fruitful endeavour. As a large number of statistical methods are used for analysis, I do not know what the best course of action for exploration will be to take.

Taking a broader perspective, work to aid in the understanding of the mechanisms of disease such as mine and my colleagues' is already proving useful for development of potential treatments, with drugs currently being trialled. Understanding of the mechanisms of action of *SOD1* has led to the development of the drug Tofersen. This is a drug which prevents the SOD1 protein from being produced, which removes its ability to aggregate (Miller et al., 2022). Similarly, understanding of the pathway through which mutations in *C9orf72* act have led to people testing drugs such as TPN-101 – initially used as a treatment for HIV - which inhibit the enzyme LINE-1 reverse transcriptase. This aims to reduce damage to nerve cells, and appears to suggest neurodegeneration in Drosophila (Krug et al., 2017; Fort-Aznar et al., 2020).

The research presented here is not immediately useful for therapeutic development.
However, the resulting increased understanding of the mechanisms of action of the disease can contribute towards stopping the progression of, and potentially reversing, the damage done by both ALS and FTD. This will be assisted by the development of more tools to more effectively analyse existing and future data.

9. References

Alami, N. H., Smith, R. B., Carrasco, M. A., Williams, L. A., Winborn, C. S., Han, S., Kiskinis, E., Winborn, B., Freibaum, B. D., Kanagaraj, A., Clare, A. J., Badders, N. M., Bilican, B., Chaum, E., Chandran, S., Shaw, C. E., Eggan, K. C., Maniatis, T., & Taylor, J. P. (2014). Axonal transport of TDP-43 mRNA granules is impaired by ALS-causing mutations. Neuron, 81(3), 536–543. https://doi.org/10.1016/j.neuron.2013.12.018

Achilli, F., Bros-Facer, V., Williams, H. P., Banks, G. T., AlQatari, M., Chia, R., Tucci, V., Groves, M., Nickols, C. D., Seburn, K. L., Kendall, R., Cader, M. Z., Talbot, K., Van Minnen, J., Burgess, R. W., Brandner, S., Martin, J. E., Koltzenburg, M., Greensmith, L., … Fisher, E. M. C. (2009). An ENU-induced mutation in mouse glycyl-tRNA synthetase (GARS) causes peripheral sensory and motor phenotypes creating a model of Charcot-Marie-Tooth type 2D peripheral neuropathy. Disease Models & Mechanisms, 2(7–8), 359–373. https://doi.org/10.1242/DMM.002527

Alexa, A., and Rahnenfuhrer, J. (2020). topGO: Enrichment Analysis for Gene Ontology, R package version 2.42.0. Available at: http://www.bioconductor.org/packages/release/bioc/html/topGO.html [Accessed March 17, 2021].

Almeida, S., Gascon, E., Tran, H., Chou, H. J., Gendron, T. F., DeGroot, S., et al. (2013). Modeling key pathological features of frontotemporal dementia with C9ORF72 repeat expansion in iPSC-derived human neurons. Acta Neuropathol. 126, 385–399. doi:10.1007/s00401-013-1149-y.

Alonso, A. D. C., Grundke-Iqbal, I., Barra, H. S., and Iqbal, K. (1997). Abnormal phosphorylation of tau and the mechanism of Alzheimer neurofibrillary degeneration: Sequestration of microtubule-associated proteins 1 and 2 and the disassembly of microtubules by the abnormal tau. Proc. Natl. Acad. Sci. U. S. A. 94, 298–303. doi:10.1073/pnas.94.1.298.

Alonso, A. D. C., Zaidi, T., Novak, M., Grundke-Iqbal, I., and Iqbal, K. (2001). Hyperphosphorylation induces self-assembly of τ into tangles of paired helical filaments/straight filaments. Proc. Natl. Acad. Sci. U. S. A. 98, 6923–6928. doi:10.1073/pnas.121119298.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21, 1–16. doi:10.1186/s13059-020-1935-5.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology, 11(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169. doi:10.1093/bioinformatics/btu638.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. 22, 2008–2017. doi:10.1101/gr.133744.111.

Andersen, J. L., Schrøder, T. J., Christensen, S., Strandbygård, D., Pallesen, L. T., García-Alai, M. M., Lindberg, S., Langgård, M., Eskildsen, J. C., David, L., Tagmose, L., Simonsen, K. B., Maltas, P. J., Rønn, L. C., de Jong, I. E., Malik, I. J., Egebjerg, J., Karlsson, J. J., Uppalanchi, S., Sakumudi, D. R., … Thirup, S. (2014). Identification of the first small-molecule ligand of the neuronal receptor sortilin and structure determination of the receptor-ligand complex. Acta crystallographica. Section D, Biological crystallography, 70(Pt 2), 451–460. https://doi.org/10.1107/S1399004713030149

Andrews, S. (2010). FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Available at http://www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/. doi:citeulike-article-id:11583827.

Angelova, M. T., Dimitrova, D. G., Da Silva, B.,Marchand, V., Jacquier, C., Achour, C., et al. (2020). tRNA 2-O-methylation by a duo of TRM7/FTSJ1 proteins modulates small RNA silencing in Drosophila. Nucleic Acids Res. 48, 2050–2072. doi: 10.1093/nar/gkaa002

Aran, D., Hu, Z., & Butte, A. J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome biology, 18(1), 220. https://doi.org/10.1186/s13059-017-1349-1

Aran, D., Hu, Z., & Butte, A. J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biology, 18(1). https://doi.org/10.1186/S13059-017-1349-1

Atanasio, A., Decman, V., White, D., Ramos, M., Ikiz, B., Lee, H. C., Siao, C. J., Brydges, S., Larosa, E., Bai, Y., Fury, W., Burfeind, P., Zamfirova, R., Warshaw, G., Orengo, J., Oyejide, A., Fralish, M., Auerbach, W., Poueymirou, W., … Lai, K. M. V. (2016). C9orf72 ablation causes immune dysregulation characterized by leukocyte expansion, autoantibody production and glomerulonephropathy in mice. Scientific Reports 2016 6:1, 6(1), 1–14. https://doi.org/10.1038/srep23204

Ayala, Y. M., De Conti, L., Avendaño-Vázquez, S. E., Dhir, A., Romano, M., D'Ambrogio, A., et al. (2011). TDP-43 regulates its mRNA levels through a negative feedback loop. EMBO J. 30, 277–288. doi:10.1038/emboj.2010.310.

Ayala, Y. M., De Conti, L., Avendaño-Vázquez, S. E., Dhir, A., Romano, M., D'Ambrogio, A., Tollervey, J., Ule, J., Baralle, M., Buratti, E., & Baralle, F. E. (2011a). TDP-43 regulates its mRNA levels through a negative feedback loop. The EMBO Journal, 30(2), 277. https://doi.org/10.1038/EMBOJ.2010.310

Ayala, Y. M., De Conti, L., Avendaño-Vázquez, S. E., Dhir, A., Romano, M., D'Ambrogio, A., Tollervey, J., Ule, J., Baralle, M., Buratti, E., & Baralle, F. E. (2011b). TDP-43 regulates its mRNA levels through a negative feedback loop. The EMBO Journal, 30(2), 277–288. https://doi.org/10.1038/EMBOJ.2010.310

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

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Filho, W. J., Lent, R., & Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. Journal of Comparative Neurology, 513(5), 532–541. https://doi.org/10.1002/CNE.21974

Balendra, R., & Isaacs, A. M. (2018). C9orf72-mediated ALS and FTD: multiple pathways to disease. Nature Reviews. Neurology, 14(9), 544. https://doi.org/10.1038/S41582-018-0047-2

Baloh, R. H. (2011). TDP-43: The relationship between protein aggregation and neurodegeneration in amyotrophic lateral sclerosis and frontotemporal lobar degeneration. FEBS J. 278, 3539–3549. doi:10.1111/j.1742-4658.2011.08256.x.

Barmada, S. J., Skibinski, G., Korb, E., Rao, E. J., Wu, J. Y., and Finkbeiner, S. (2010). Cytoplasmic mislocalization of TDP-43 is toxic to neurons and enhanced by a mutation associated with familial amyotrophic lateral sclerosis. J. Neurosci. 30, 639–649. doi:10.1523/JNEUROSCI.4988-09.2010.

Barouch, D. H., Santra, S., Schmitz, J. E., Kuroda, M. J., Fu, T. M., Wagner, W., et al. (2000). Control of viremia prevention of clinical AIDS in rhesus monkeys by cytokine-augmented DNA vaccination. Science (80-.). 290, 486–492. doi:10.1126/science.290.5491.486.

Beckers J, Tharkeshwar AK, Van Damme P. C9orf72 ALS-FTD: recent evidence for dysregulation of the autophagy-lysosome pathway at multiple levels. Autophagy. 2021 Nov;17(11):3306-3322. doi: 10.1080/15548627.2021.1872189. Epub 2021 Feb 26. PMID: 33632058; PMCID: PMC8632097.

Bilsland, L. G., Sahai, E., Kelly, G., Golding, M., Greensmith, L., & Schiavo, G. (2010). Deficits in axonal transport precede ALS symptoms in vivo. Proceedings of the National Academy of Sciences of the United States of America, 107(47), 20523–20528. https://doi.org/10.1073/PNAS.1006869107/SUPPL_FILE/SM02.AVI

Birsa N, Ule AM, Garone MG, Tsang B, Mattedi F, Chong PA, Humphrey J, Jarvis S, Pisiren M, Wilkins OG, Nosella ML, Devoy A, Bodo C, de la Fuente RF, Fisher EMC, Rosa A, Viero G, Forman-Kay JD, Schiavo G, Fratta P. FUS-ALS mutants alter FMRP phase separation equilibrium and impair protein translation. Sci Adv. 2021 Jul 21;7(30):eabf8660. doi: 10.1126/sciadv.abf8660. PMID: 34290090; PMCID: PMC8294762.

Boehm, I. et al. Comparative anatomy of the mammalian neuromuscular junction. J. Anat. 237, 827–836 (2020).

Bonfanti, E., Bonifacino, T., Raffaele, S., Milanese, M., Morgante, E., Bonanno, G., et al. (2020). Abnormal upregulation of gpr17 receptor contributes to oligodendrocyte dysfunction in SOD1G93A mice. Int. J. Mol. Sci. 21:2395. doi: 10.3390/ijms21072395

Bosco, D. A., Lemay, N., Ko, H. K., Zhou, H., Burke, C., Kwiatkowski, T. J., et al. (2010). Mutant FUS proteins that cause amyotrophic lateral sclerosis incorporate into stress granules. Hum. Mol. Genet. 19, 4160–4175. doi:10.1093/hmg/ddq335.

Bradl, M., & Lassmann, H. (2010). Oligodendrocytes: biology and pathology. Acta Neuropathologica, 119(1), 37. https://doi.org/10.1007/S00401-009-0601-5

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 2016 345 34, 525. doi:10.1038/nbt.3519.

Brelstaff, J. H., Mason, M., Katsinelos, T., McEwan, W. A., Ghetti, B., Tolkovsky, A. M., & Spillantini, M. G. (2021). Microglia become hypofunctional and release metalloproteases and tau seeds when phagocytosing live neurons with P301S tau aggregates. Science Advances, 7(43). https://doi.org/10.1126/SCIADV.ABG4980/SUPPL_FILE/SCIADV.ABG4980_SM.PDF

Brion, J. P., Couck, A. M., Passareiro, E., & Flament-Durand, J. (1985). Neurofibrillary tangles of Alzheimer's disease: an immunohistochemical study. Journal of Submicroscopic Cytology, 17(1), 89–96. https://europepmc.org/article/med/3973960

Broce, I., Karch, C. M., Wen, N., Fan, C. C., Wang, Y., Hong Tan, C., Kouri, N., Ross, O. A., Höglinger, G. U., Muller, U., Hardy, J., Momeni, P., Hess, C. P., Dillon, W. P., Miller, Z. A., Bonham, L. W., Rabinovici, G. D., Rosen, H. J., Schellenberg, G. D., … Momeni, P. (2018). Immune-related genetic enrichment in frontotemporal dementia: An analysis of genome-wide association studies. PLoS Medicine, 15(1). https://doi.org/10.1371/JOURNAL.PMED.1002487

Broustal, O., Camuzat, A., Guillot-Noël, L., Guy, N., Millecamps, S., Deffond, D., et al. (2010). FUS mutations in frontotemporal lobar degeneration with amyotrophic lateral sclerosis. J. Alzheimer's Dis. 22, 765–769. doi:10.3233/JAD-2010-100837.

Buck, D., Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X. J., & Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research 2017 6:100, 6, 100. https://doi.org/10.12688/f1000research.10571.2

Burgoyne RD, Morgan A. Cysteine string protein (CSP) and its role in preventing neurodegeneration. Semin Cell Dev Biol. 2015 Apr;40:153-9. doi: 10.1016/j.semcdb.2015.03.008. Epub 2015 Mar 21. PMID: 25800794; PMCID: PMC4447612.

Burguillos, M. A. (2013). Use of meso-scale discoveryTM to examine cytokine content in microglia cell supernatant. Methods in Molecular Biology, 1041, 93–100. https://doi.org/10.1007/978-1-62703-520-0_11/FIGURES/00112

Cadieux-Dion M, Andermann E, Lachance-Touchette P, Ansorge O, Meloche C, Barnabé A, Kuzniecky RI, Andermann F, Faught E, Leonberg S, Damiano JA, Berkovic SF, Rouleau GA, Cossette P. Recurrent mutations in DNAJC5 cause autosomal dominant Kufs disease. Clin Genet. 2013 Jun;83(6):571-5. doi: 10.1111/cge.12020. Epub 2012 Nov 7. PMID: 22978711.

Canard, B., and Sarfati, R. S. (1994). DNA polymerase fluorescent substrates with reversible 3′-tags. Gene 148, 1–6. doi:10.1016/0378-1119(94)90226-7.

Chiò, A., Logroscino, G., Traynor, B. J., Collins, J., Simeone, J. C., Goldstein, L. A., et al. (2013). Global epidemiology of amyotrophic lateral sclerosis: A systematic review of the published literature. Neuroepidemiology 41, 118–130. doi:10.1159/000351153.

Chiò, A., Restagno, G., Brunetti, M., Ossola, I., Calvo, A., Mora, G., et al. (2009). Two Italian kindreds with familial amyotrophic lateral sclerosis due to FUS mutation. Neurobiol. Aging 30, 1272–1275. doi:10.1016/j.neurobiolaging.2009.05.001.

Chiò, A., Traynor, B. J., Lombardo, F., Fimognari, M., Calvo, A., Ghiglione, P., et al. (2008). Prevalence of SOD1 mutations in the Italian ALS population. Neurology 70, 533–537. doi:10.1212/01.wnl.0000299187.90432.3f.

Choi, S. Y., Lopez-Gonzalez, R., Krishnan, G., Phillips, H. L., Li, A. N., Seeley, W. W., Yao, W. D., Almeida, S., & Gao, F. B. (2019). C9ORF72-ALS/FTD-associated poly(GR) binds Atp5a1 and compromises mitochondrial function in vivo. Nature Neuroscience, 22(6), 851. https://doi.org/10.1038/S41593-019-0397-0

Clark, D. P., Pazdernik, N. J., and McGehee, M. R. (2018). Molecular biology. Elsevier doi:10.1016/C2015-0-06229-3.

Cleveland, D. W., Hwo, S. Y., and Kirschner, M. W. (1977). Purification of tau, a microtubule-associated protein that induces assembly of microtubules from purified tubulin. J. Mol. Biol. 116, 207–225. doi:10.1016/0022-2836(77)90213-3.

Coady, T. H., and Manley, J. L. (2015). ALS mutations in TLS/FUS disrupt target gene expression. Genes Dev. 29, 1696–1706. doi: 10.1101/gad.267286.115

Colombrita, C., Onesto, E., Megiorni, F., Pizzuti, A., Baralle, F. E., Buratti, E., et al. (2012). 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. J. Biol. Chem. 287, 15635–15647. doi:10.1074/jbc.M111.333450.

Cong, X., Nagre, N., Herrera, J., Pearson, A. C., Pepper, I., Morehouse, R., Ji, H. L., Jiang, D., Hubmayr, R. D., & Zhao, X. (2020). TRIM72 promotes alveolar epithelial cell membrane repair and ameliorates lung fibrosis. Respiratory Research, 21(1), 1–20. https://doi.org/10.1186/S12931-020-01384-2/FIGURES/10

Conlon, E. G., Fagegaltier, D., Agius, P., Davis-Porada, J., Gregory, J., Hubbard, I., Kang, K., Kim, D., Phatnani, H., Shneider, N. A., Manley, J. L., Kwan, J., Sareen, D., Broach, J. R., Simmons, Z., Arcila-Londono, X., Lee, E. B., Van Deerlin, V. M., Fraenkel, E., … Nath, A. (2018). Unexpected similarities between C9ORF72 and sporadic forms of ALS/FTD suggest a common disease mechanism. ELife, 7. https://doi.org/10.7554/ELIFE.37754

Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, 32(Database issue), D258. https://doi.org/10.1093/NAR/GKH036

Corley, S. M., Troy, N. M., Bosco, A., & Wilkins, M. R. (2019). QuantSeq. 3′ Sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis. Scientific Reports. https://doi.org/10.1038/s41598-019-55434-x

Corley, S. M., Troy, N. M., Bosco, A., and Wilkins, M. R. (2019). QuantSeq. 3′ Sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis. Sci. Rep. 9:18895. doi: 10.1038/s41598-019-55434-x

Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).

Cui, Y., Cai, M., & Stanley, H. E. (2017). Comparative Analysis and Classification of Cassette Exons and Constitutive Exons. BioMed research international, 2017, 7323508. https://doi.org/10.1155/2017/7323508

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2015). Ensembl 2015. Nucleic Acids Res. 43, D662–D669. doi:10.1093/nar/gku1010.

Dash, B. P., Freischmidt, A., Weishaupt, J. H., & Hermann, A. (2022). Downstream Effects of Mutations in SOD1 and TARDBP Converge on Gene Expression Impairment in Patient-Derived Motor Neurons. International Journal of Molecular Sciences, 23(17). https://doi.org/10.3390/IJMS23179652/S1

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2), giab008. https://doi.org/10.1093/gigascience/giab008

De Giorgio, F., Maduro, C., Fisher, E. M. C., & Acevedo-Arozena, A. (2019). Transgenic and physiological mouse models give insights into different aspects of amyotrophic lateral sclerosis. DMM Disease Models and Mechanisms, 12(1). https://doi.org/10.1242/DMM.037424/3045

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. Nature biotechnology, 34(5), 518–524. https://doi.org/10.1038/nbt.3423

DeJesus-Hernandez, M., Kocerha, J., Finch, N., Crook, R., Baker, M., Desaro, P., et al. (2010). De novo truncating FUS gene mutation as a cause of sporadic amyotrophic lateral sclerosis. Hum. Mutat. 31, E1377–E1389. doi:10.1002/humu.21241.

DeJesus-Hernandez, M., Kocerha, J., Finch, N., Crook, R., Baker, M., Desaro, P., Johnston, A., Rutherford, N., Wojtas, A., Kennelly, K., Wszolek, Z. K., Graff-Radford, N., Boylan, K., & Rademakers, R. (2010). De Novo Truncating FUS Gene Mutation as a Cause of Sporadic Amyotrophic Lateral Sclerosis. Human Mutation, 31(5), E1377. https://doi.org/10.1002/HUMU.21241

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

Devoy, A., Kalmar, B., Stewart, M., Park, H., Burke, B., Noy, S. J., Redhead, Y., Humphrey, J., Lo, K., Jaeger, J., Mejia Maza, A., Sivakumar, P., Bertolin, C., Soraru, G., Plagnol, V., Greensmith, L., Acevedo Arozena, A., Isaacs, A. M., Davies, B., … Fisher, E. M. C. C. (2017). Humanized mutant FUS drives progressive motor neuron degeneration without aggregation in "FUSDelta14" knockin mice. Brain, 140(11), 2797–2805. https://doi.org/10.1093/brain/awx248

Dickson, D. W., Baker, M. C., Jackson, J. L., Dejesus-Hernandez, M., Finch, N. C. A., Tian, S., Heckman, M. G., Pottier, C., Gendron, T. F., Murray, M. E., Ren, Y., Reddy, J. S., Graff-Radford, N. R., Boeve, B. F., Petersen, R. C., Knopman, D. S., Josephs, K. A., Petrucelli, L., Oskarsson, B., … Van Blitterswijk, M. (2019). Extensive transcriptomic study emphasizes importance of vesicular transport in C9orf72 expansion carriers. Acta Neuropathologica Communications, 7(1), 1–21. https://doi.org/10.1186/S40478-019-0797-0/FIGURES/7

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. doi:10.1093/bioinformatics/bts635.

Dupuis, L. & Loeffler, J. P. Neuromuscular junction destruction during amyotrophic lateral sclerosis: insights from transgenic models. Current Opinion in Pharmacology vol. 9 341–346 (2009).

Ebstein, S. Y., Yagudayeva, I., & Shneider, N. A. (2019). Mutant TDP-43 Causes Early-Stage Dose-Dependent Motor Neuron Degeneration in a TARDBP Knockin Mouse Model of ALS. Cell reports, 26(2), 364–373.e4. https://doi.org/10.1016/j.celrep.2018.12.045

Ederle, H., & Dormann, D. (2017). TDP-43 and FUS en route from the nucleus to the cytoplasm. FEBS Letters, 591(11), 1489–1507. https://doi.org/10.1002/1873-3468.12646

Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., Chuaqui, R. F., Zhuang, Z., Goldstein, S. R., et al. (1996). Laser capture microdissection. Science (80-.). 274, 998–1001. doi:10.1126/science.274.5289.998.

Eréndira Avendaño-Vázquez, S., Dhir, A., Bembich, S., Buratti, E., Proudfoot, N., & Baralle, F. E. (2012). Autoregulation of TDP-43 mRNA levels involves interplay between transcription, splicing, and alternative polyA site selection. Genes & Development, 26(15), 1679. https://doi.org/10.1101/GAD.194829.112

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

Farg, M. A., Konopka, A., Soo, K. Y., Ito, D., and Atkin, J. D. (2017). The DNA damage response (DDR) is induced by the C9orf72 repeat expansion in amyotrophic lateral sclerosis. Hum. Mol. Genet. 26, 2882–2896. doi:10.1093/hmg/ddx170.

Ferland, R. J., Eyaid, W., Collura, R. v., Tully, L. D., Hill, R. S., Al-Nouri, D., Al-Rumayyan, A., Topcu, M., Gascon, G., Bodell, A., Shugart, Y. Y., Ruvolo, M., & Walsh, C. A. (2004). Abnormal cerebellar development and axonal decussation due to mutations in AHI1 in Joubert syndrome. Nature Genetics 2004 36:9, 36(9), 1008–1013. https://doi.org/10.1038/ng1419

Ferrari, R., Hernandez, D. G., Nalls, M. A., Rohrer, J. D., Ramasamy, A., Kwok, J. B. J., Dobson-Stone, C., Brooks William S., B. S., Schofield, P. R., Halliday, G. M., Hodges, J. R., Piguet, O., Bartley, L., Thompson, E., Haan, E., Hernández, I., Ruiz, A., Boada, M., Borroni, B., … Momeni, P. (2014). Frontotemporal dementia and its subtypes: a genome-wide association study. Lancet Neurology, 13(7), 686. https://doi.org/10.1016/S1474-4422(14)70065-1

Fong, J. C., Karydas, A. M., and Goldman, J. S. (2012). Genetic counseling for ftd/als caused by the c9orf72 hexanucleotide expansion. Alzheimer's Res. Ther. 4, 27. doi:10.1186/alzrt130.

Fort-Aznar, L., Ugbode, C., & Sweeney, S. T. (2020). Retrovirus reactivation in CHMP2BIntron5 models of frontotemporal dementia. Human molecular genetics, 29(16), 2637–2646. https://doi.org/10.1093/hmg/ddaa142

Fratta, P. et al. Mice with endogenous <scp>TDP</scp> -43 mutations exhibit gain of splicing function and characteristics of amyotrophic lateral sclerosis. EMBO J. 37, e98684 (2018).

Fratta, P., Sivakumar, P., Humphrey, J., Lo, K., Ricketts, T., Oliveira, H., Brito-Armas, J. M., Kalmar, B., Ule, A., Yu, Y., Birsa, N., Bodo, C., Collins, T., Conicella, A. E., Mejia Maza, A., Marrero-Gagliardi, A., Stewart, M., Mianne, J., Corrochano, S., … Montalcini, R. L. (2018). Mice with endogenous TDP-43 mutations exhibit gain of splicing function and characteristics of amyotrophic lateral sclerosis. The EMBO Journal, 37(11), e98684. https://doi.org/10.15252/EMBJ.201798684

Freibaum, B. D., Chitta, R. K., High, A. A., and Taylor, J. P. (2010). Global analysis of TDP-43 interacting proteins reveals strong association with RNA splicing and translation machinery. J. Proteome Res. 9, 1104–1120. doi:10.1021/pr901076y.

Gao, F. B., Almeida, S., & Lopez-Gonzalez, R. (2017). Dysregulated molecular pathways in amyotrophic lateral sclerosis-frontotemporal dementia spectrum disorder. The EMBO journal, 36(20), 2931–2950. https://doi.org/10.15252/embj.201797568

Gasset-Rosa, F., Lu, S., Yu, H., Chen, C., Melamed, Z., Guo, L., Shorter, J., Da Cruz, S., & Cleveland, D. W. (2019). Cytoplasmic TDP-43 De-mixing Independent of Stress Granules Drives Inhibition of Nuclear Import, Loss of Nuclear TDP-43, and Cell Death. Neuron, 102(2), 339–357.e7. https://doi.org/10.1016/j.neuron.2019.02.038

Geistlinger, L., Csaba, G., & Zimmer, R. (2016). Bioconductor's EnrichmentBrowser: Seamless navigation through combined results of set- & network-based enrichment analysis. BMC Bioinformatics. https://doi.org/10.1186/s12859-016-0884-1

Gendron, T. F., Belzil, V. V., Zhang, Y. J. & Petrucelli, L. Mechanisms of toxicity in C9FTLD/ALS. Acta Neuropathologica vol. 127 359–376 (2014).

Gendron, T. F., Belzil, V. V., Zhang, Y. J., and Petrucelli, L. (2014). Mechanisms of toxicity in C9FTLD/ALS. Acta Neuropathol. 127, 359–376. doi:10.1007/s00401-013-1237-z.

Goldstein, L. D., Cao, Y., Pau, G., Lawrence, M., Wu, T. D., Seshagiri, S., et al. (2016). Prediction and quantification of splice events from RNA-seq data. PLoS One 11, e0156132. doi:10.1371/journal.pone.0156132.

Gong, T., & Szustakowski, J. D. (2013). DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics (Oxford, England), 29(8), 1083–1085. https://doi.org/10.1093/BIOINFORMATICS/BTT090

Gurney, M. E., Pu, H., Chiu, A. Y., Dal Canto, M. C., Polchow, C. Y., Alexander, D. D., Caliendo, J., Hentati, A., Kwon, Y. W., Deng, H. X., Chen, W., Zhai, P., Sufit, R. L., & Siddique, T. (1994). Motor Neuron Degeneration in Mice that Express a Human Cu,Zn Superoxide Dismutase Mutation. Science, 264(5166), 1772–1775. https://doi.org/10.1126/SCIENCE.8209258

Hagenauer, M. H., Schulmann, A., Li, J. Z., Vawter, M. P., Walsh, D. M., Thompson, R. C., et al. (2018). Inference of cell type content from human brain transcriptomic datasets Illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. PLoS One 13, e0200003. doi:10.1371/journal.pone.0200003.

Hagenauer, M. H., Schulmann, A., Li, J. Z., Vawter, M. P., Walsh, D. M., Thompson, R. C., Turner, C. A., Bunney, W. E., Myers, R. M., Barchas, J. D., Schatzberg, A. F., Watson, S. J., & Akil, H. (2018). Inference of cell type content from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. PloS One, 13(7). https://doi.org/10.1371/JOURNAL.PONE.0200003

Hansen, D. V., Hanson, J. E., & Sheng, M. (2018). Microglia in Alzheimer's disease. The Journal of Cell Biology, 217(2), 459. https://doi.org/10.1083/JCB.201709069

Hartmann, H., Hornburg, D., Czuppa, M., Bader, J., Michaelsen, M., Farny, D., Arzberger, T., Mann, M., Meissner, F., & Edbauer, D. (2018). Proteomics and C9orf72 neuropathology identify ribosomes as poly-GR/PR interactors driving toxicity. Life Science Alliance, 1(2). https://doi.org/10.26508/LSA.201800070

Hayes, L. R., Duan, L., Bowen, K., Kalab, P., & Rothstein, J. D. (2020). C9orf72 arginine-rich dipeptide repeat proteins disrupt karyopherin-mediated nuclear import. ELife, 9. https://doi.org/10.7554/ELIFE.51685

Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107(1):1-8. doi: 10.1016/j.ygeno.2015.11.003. Epub 2015 Nov 10. PMID: 26554401; PMCID: PMC4727787.

Henderson, M. X., Wirak, G. S., Zhang, Y. quan, Dai, F., Ginsberg, S. D., Dolzhanskaya, N., Staropoli, J. F., Nijssen, P. C. G., Lam, T. K. T., Roth, A. F., Davis, N. G., Dawson, G., Velinov, M., & Chandra, S. S. (2016). Neuronal ceroid lipofuscinosis with DNAJC5/CSP mutations have PPT1 pathology and exhibit aberrant protein palmitoylation. Acta Neuropathologica, 131(4), 621. https://doi.org/10.1007/S00401-015-1512-2

Hergesheimer, R. C., Chami, A. A., De Assis, D. R., Vourc'h, P., Andres, C. R., Corcia, P., et al. (2019). The debated toxic role of aggregated TDP-43 in amyotrophic lateral sclerosis: A resolution in sight? Brain 142, 1176–1194. doi:10.1093/brain/awz078.

Hicks, G. G., Singh, N., Nashabi, A., Mai, S., Bozek, G., Klewes, L., et al. (2000). Fus deficiency in mice results in defective B-lymphocyte development and activation, high levels of chromosomal instability and perinatal death. Nat. Genet. 24, 175–179. doi:10.1038/72842.

Highley, J. R., Kirby, J., Jansweijer, J. A., Webb, P. S., Hewamadduma, C. A., Heath, P. R., Higginbottom, A., Raman, R., Ferraiuolo, L., Cooper-Knock, J., McDermott, C. J., Wharton, S. B., Shaw, P. J., & Ince, P. G. (2014). Loss of nuclear TDP-43 in amyotrophic lateral sclerosis (ALS) causes altered expression of splicing machinery and widespread dysregulation of RNA splicing in motor neurones. Neuropathology and applied neurobiology, 40(6), 670–685. https://doi.org/10.1111/nan.12148

Hochberg, B. (1995). Summary for Policymakers. Clim. Chang. 2013 - Phys. Sci. Basis, 1–30. doi:10.1017/CBO9781107415324.004.

Humphrey, J., Birsa, N., Milioto, C., Robaldo, D., Bentham, M., Jarvis, S., et al. (2020). FUS ALS-causative mutations impact FUS autoregulation and the processing of RNA-binding proteins through intron retention. Cold Spring Harbor Laboratory doi:10.1101/567735.

Humphrey, J., Emmett, W., Fratta, P., Isaacs, A. M. & Plagnol, V. Quantitative analysis of cryptic splicing associated with TDP-43 depletion. BMC Med. Genomics 10, 38 (2017).

Humphrey, J., Venkatesh, S., Hasan, R., Herb, J. T., de Paiva Lopes, K., Küçükali, F., Byrska-Bishop, M., Evani, U. S., Narzisi, G., Fagegaltier, D., Sleegers, K., Phatnani, H., Knowles, D. A., Fratta, P., & Raj, T. (2022). Integrative transcriptomic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes. Nature Neuroscience 2022 26:1, 26(1), 150–162. https://doi.org/10.1038/s41593-022-01205-3

Hunt, G. J., Freytag, S., Bahlo, M., & Gagnon-Bartsch, J. A. (2019). dtangle: accurate and robust cell type deconvolution. Bioinformatics (Oxford, England), 35(12), 2093–2099. https://doi.org/10.1093/BIOINFORMATICS/BTY926

Hunt, G. J., Freytag, S., Bahlo, M., and Gagnon-Bartsch, J. A. (2019). dtangle: accurate and robust cell type deconvolution. Bioinformatics 35, 2093–2099. doi:10.1093/bioinformatics/bty926.

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

Illumina, Inc (2017). Illumina Considerations for RNA-Seq Read Length and Coverage. Available online at: https://support.Illumina.com/bulletins/2017/04/ considerations-for-rna-seq-read-length-and-coverage-.html (accessed May 21, 2020).

Ishigaki, S., Masuda, A., Fujioka, Y., Iguchi, Y., Katsuno, M., Shibata, A., et al. (2012). Position-dependent FUS-RNA interactions regulate alternative splicing events and transcriptions. Sci. Rep. 2, 1–9. doi:10.1038/srep00529.

Ishigaki, S., Masuda, A., Fujioka, Y., Iguchi, Y., Katsuno, M., Shibata, A., et al. (2012). Position-dependent FUS-RNA interactions regulate alternative splicing events and transcriptions. Sci. Rep. 2:529. doi: 10.1038/srep00529

Jafarinia, H., Van der Giessen, E., & Onck, P. R. (2022). Molecular basis of C9orf72 poly-PR interference with the β-karyopherin family of nuclear transport receptors. Scientific Reports 2022 12:1, 12(1), 1–11. https://doi.org/10.1038/s41598-022-25732-y

Jarvis, S., Birsa, N., Secrier, M., Fratta, P., and Plagnol, V. (2020). A Comparison of Low Read Depth QuantSeq 3′ Sequencing to Total RNA-Seq in FUS Mutant Mice. Front. Genet. 11, 1412. doi:10.3389/fgene.2020.562445.

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., & Wong, P. C. (2017). Tdp-43 cryptic exons are highly variable between cell types. Molecular Neurodegeneration, 12(1). https://doi.org/10.1186/S13024-016-0144-X

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: A brief primer. Behavior Therapy, 51(5), 675. https://doi.org/10.1016/J.BETH.2020.05.002

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

Jones, R. A., Reich, C. D., Dissanayake, K. N., Kristmundsdottir, F., Findlater, G. S., Ribchester, R. R., Simmen, M. W., & Gillingwater, T. H. (2016). NMJ-morph reveals principal components of synaptic morphology influencing structure-function relationships at the neuromuscular junction. Open biology, 6(12), 160240. https://doi.org/10.1098/rsob.160240

Joyce, P. I., Fratta, P., Landman, A. S., Mcgoldrick, P., Wackerhage, H., Groves, M., Busam, B. S., Galino, J., Corrochano, S., Beskina, O. A., Esapa, C., Ryder, E., Carter, S., Stewart, M., Codner, G., Hilton, H., Teboul, L., Tucker, J., Lionikas, A., … Acevedo-Arozena, A. (2016). Deficiency of the zinc finger protein ZFP106 causes motor and sensory neurodegeneration. Human Molecular Genetics, 25(2), 291. https://doi.org/10.1093/HMG/DDV471

Kahlson, M. A., & Colodner, K. J. (2015). Glial Tau Pathology in Tauopathies: Functional Consequences. Journal of Experimental Neuroscience, 9(Suppl 2), 43. https://doi.org/10.4137/JEN.S25515

Kan, A. (2017). Machine learning applications in cell image analysis. Immunology and Cell Biology, 95(6), 525–530. https://doi.org/10.1038/ICB.2017.16

Kapeli, K., Martinez, F. J., and Yeo, G. W. (2017). Genetic mutations in RNA-binding proteins and their roles in ALS. Hum. Genet. 136, 1193–1214. doi:10.1007/s00439-017-1830-7.

Kapeli, K., Pratt, G. A., Vu, A. Q., Hutt, K. R., Martinez, F. J., Sundararaman, B., et al. (2016). Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. Nat. Commun. 7, 12143. doi:10.1038/ncomms12143.

Kelley, K. W., Nakao-Inoue, H., Molofsky, A. v., & Oldham, M. C. (2018). Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. Nature Neuroscience, 21(9), 1171–1184. https://doi.org/10.1038/S41593-018-0216-Z

King, O. D., Gitler, A. D., & Shorter, J. (2012). The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. Brain Research, 1462, 61. https://doi.org/10.1016/J.BRAINRES.2012.01.016

Kino, Y., Washizu, C., Kurosawa, M., Yamada, M., Miyazaki, H., Akagi, T., et al. (2015). FUS/TLS deficiency causes behavioral and pathological abnormalities distinct from amyotrophic lateral sclerosis. Acta Neuropathol. Commun. 3, 24. doi:10.1186/s40478-015-0202-6.

Kono, N., and Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. Dev. Growth Differ. 61, 316–326. doi:10.1111/dgd.12608.

Kraemer, B. C. et al. Loss of Murine TDP-43 disrupts motor function and plays an essential role in embryogenesis. Acta Neuropathol. 119, 409–419 (2010).

Kraemer, B. C., Schuck, T., Wheeler, J. M., Robinson, L. C., Trojanowski, J. Q., Lee, V. M. Y. Y., et al. (2010). Loss of murine TDP-43 disrupts motor function and plays an essential role in embryogenesis. Acta Neuropathol. 119, 409–419. doi:10.1007/s00401-010-0659-0.

Kramer, N. J., Haney, M. S., Morgens, D. W., Jovičić, A., Couthouis, J., Li, A., Ousey, J., Ma, R., Bieri, G., Tsui, C. K., Shi, Y., Hertz, N. T., Tessier-Lavigne, M., Ichida, J. K., Bassik, M. C., & Gitler, A. D. (2018). CRISPR–Cas9 screens in human cells and primary neurons identify modifiers of C9ORF72 dipeptide-repeat-protein toxicity. Nature Genetics 2018 50:4, 50(4), 603–612. https://doi.org/10.1038/s41588-018-0070-7

Krug, L., Chatterjee, N., Borges-Monroy, R., Hearn, S., Liao, W. W., Morrill, K., Prazak, L., Rozhkov, N., Theodorou, D., Hammell, M., & Dubnau, J. (2017). Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. PLoS genetics, 13(3), e1006635. https://doi.org/10.1371/journal.pgen.1006635

Krueger, F. (2012). Trim Galore! Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ [Accessed March 27, 2019].

Kuhn, M. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26 (2008).

Kumar-Singh, S. (2011). Progranulin and TDP-43: Mechanistic links and future directions. in Journal of Molecular Neuroscience (J Mol Neurosci), 561–573. doi:10.1007/s12031-011-9625-0.

Kuo, P. H., Doudeva, L. G., Wang, Y. T., Shen, C. K. J., & Yuan, H. S. (2009). Structural insights into TDP-43 in nucleic-acid binding and domain interactions. Nucleic Acids Research, 37(6), 1799–1808. https://doi.org/10.1093/NAR/GKP013

KURLAND, L. T., and MULDER, D. W. (1955). Epidemiologic investigations of amyotrophic lateral sclerosis. 2. Familial aggregations indicative of dominant inheritance. II. Neurology 5, 249–68. doi:10.1212/WNL.5.3.182.

Kwiatkowski, T. J., Bosco, D. A., LeClerc, A. L., Tamrazian, E., Vanderburg, C. R., Russ, C., et al. (2009). Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. Science (80-.). 323, 1205–1208. doi:10.1126/science.1166066.

Kwon, I. et al. Poly-dipeptides encoded by the C9orf72 repeats bind nucleoli, impede RNA biogenesis, and kill cells. Science (80-.). 345, 1139–1145 (2014).

Kwon, I., Xiang, S., Kato, M., Wu, L., Theodoropoulos, P., Wang, T., et al. (2014). Poly-dipeptides encoded by the C9orf72 repeats bind nucleoli, impede RNA biogenesis, and kill cells. Science (80-.). 345, 1139–1145. doi:10.1126/science.1254917.

Kwon, I., Xiang, S., Kato, M., Wu, L., Theodoropoulos, P., Wang, T., Kim, J., Yun, J., Xie, Y., & McKnight, S. L. (2014). Poly-dipeptides encoded by the C9orf72 repeats bind nucleoli, impede RNA biogenesis, and kill cells. Science (New York, N.Y.), 345(6201), 1139–1145. https://doi.org/10.1126/SCIENCE.1254917

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., Wancewicz, E., Kim, A. S., Watt, A., Freier, S., Hicks, G. G., Donohue, J. P., Shiue, L., Bennett, C. F., Ravits, J., Cleveland, D. W., … Yeo, G. W. (2012). Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. Nature neuroscience, 15(11), 1488–1497. https://doi.org/10.1038/nn.3230

Lee, K. H., Zhang, P., Kim, H. J., Mitrea, D. M., Sarkar, M., Freibaum, B. D., Cika, J., Coughlin, M., Messing, J., Molliex, A., Maxwell, B. A., Kim, N. C., Temirov, J., Moore, J., Kolaitis, R. M., Shaw, T. I., Bai, B., Peng, J., Kriwacki, R. W., & Taylor, J. P. (2016). C9orf72 dipeptide repeats impair the assembly, dynamics and function of membrane-less organelles. Cell, 167(3), 774. https://doi.org/10.1016/J.CELL.2016.10.002

Lee, Y. B., Chen, H. J., Peres, J. N., Gomez-Deza, J., Attig, J., Štalekar, M., et al. (2013). Hexanucleotide repeats in ALS/FTD form length-dependent RNA Foci, sequester RNA binding proteins, and are neurotoxic. Cell Rep. 5, 1178–1186. doi:10.1016/j.celrep.2013.10.049.

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., et al. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods 7, 709–715. doi:10.1038/nmeth.1491.

Levine, T. P., Daniels, R. D., Gatta, A. T., Wong, L. H., and Hayes, M. J. (2013). The product of C9orf72, a gene strongly implicated in neurodegeneration, is structurally related to DENN Rab-GEFs. Bioinformatics 29, 499–503. doi:10.1093/bioinformatics/bts725.

Li W, Yan Q, Ding X, Shen C, Hu M, Zhu Y, Qin D, Lu H, Krueger BJ, Renne R, Gao SJ, Lu C. The SH3BGR/STAT3 Pathway Regulates Cell Migration and Angiogenesis Induced by a Gammaherpesvirus MicroRNA. PLoS Pathog. 2016 Apr 29;12(4):e1005605. doi: 10.1371/journal.ppat.1005605. PMID: 27128969; PMCID: PMC4851422.

Li, Y., Nowak, C. M., Pham, U., Nguyen, K., & Bleris, L. (2021). Cell morphology-based machine learning models for human cell state classification. Npj Systems Biology and Applications 2021 7:1, 7(1), 1–9. https://doi.org/10.1038/s41540-021-00180-y

Liao, Y., Dong, Y., and Cheng, J. (2017). The function of the mitochondrial calcium uniporter in neurodegenerative disorders. Int. J. Mol. Sci. 18:248. doi: 10.3390/ijms18020248

Ling, J. P., Pletnikova, O., Troncoso, J. C., & Wong, P. C. (2015). TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. Science (New York, N.Y.), 349(6248), 650–655. https://doi.org/10.1126/SCIENCE.AAB0983

Ling, S. C., Polymenidou, M. & Cleveland, D. W. Converging mechanisms in als and FTD: Disrupted RNA and protein homeostasis. Neuron 79, 416–438 (2013).

Liu, X., Mei, W., Soltis, P. S., Soltis, D. E., & Barbazuk, W. B. (2017). Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome. Molecular Ecology Resources, 17(6), 1243–1256. https://doi.org/10.1111/1755-0998.12670

Logroscino, G., Traynor, B. J., Hardiman, O., Chió, A., Mitchell, D., Swingler, R. J., et al. (2010). Incidence of amyotrophic lateral sclerosis in Europe. J. Neurol. Neurosurg. Psychiatry 81, 385–390. doi:10.1136/jnnp.2009.183525.

Loi, S. M., Tsoukra, P., Chen, Z., Wibawa, P., Eratne, D., Kelso, W., Walterfang, M., & Velakoulis, D. (2022). Risk factors to mortality and causes of death in frontotemporal dementia: An Australian perspective. International Journal of Geriatric Psychiatry, 37(2). https://doi.org/10.1002/GPS.5668

Lorenzini, I., Alsop, E., Levy, J., Gittings, L. M., Lall, D., Rabichow, B. E., Moore, S., Pevey, R., Bustos, L., Burciu, C., Bhatia, D., Singer, M., Saul, J., McQuade, A., Tzioras, M., Mota, T. A., Logemann, A., Rose, J., Almeida, S., … Sattler, R. (2022). Cellular and molecular phenotypes of C9orf72 ALS/FTD patient derived iPSC-microglia mono-cultures. BioRxiv, 2020.09.03.277459. https://doi.org/10.1101/2020.09.03.277459

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

Love, M. New function lfcShrink() in DESeq2. Available at: https://support.bioconductor.org/p/95695/ [Accessed March 28, 2019].

Luo, W., and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics 29, 1830–1831. doi:10.1093/bioinformatics/btt285.

Ma, F., Fuqua, B. K., Hasin, Y., Yukhtman, C., Vulpe, C. D., Lusis, A. J., & Pellegrini, M. (2019). A comparison between whole transcript and 3’ RNA sequencing methods using Kapa and Lexogen library preparation methods 06 Biological Sciences 0604 Genetics. BMC Genomics. https://doi.org/10.1186/s12864-018-5393-3

Ma, F., Fuqua, B. K.,Hasin, Y., Yukhtman, C., Vulpe, C. D., Lusis, A. J., et al. (2019). A comparison between whole transcript and 3’ RNA sequencing methods using kapa and lexogen library preparation methods 06 biological sciences 0604 genetics. BMC Genomics 20:9. doi: 10.1186/s12864-018-5393-3

MacKenzie, I. R. A., and Neumann, M. (2012). FET proteins in frontotemporal dementia and amyotrophic lateral sclerosis. Brain Res. 1462, 40–43. doi:10.1016/j.brainres.2011.12.010.

Majounie, E. et al. Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: A cross-sectional study. Lancet Neurol. 11, 323–330 (2012).

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

Mamanova, L., and Turner, D. J. (2011). Low-bias, strand-specific transcriptome Illumina sequencing by on-flowcell reverse transcription (FRT-seq). Nat. Protoc. 6, 1736–1747. doi:10.1038/nprot.2011.399.

Mankodi, A., Urbinati, C. R., Yuan, Q. P., Moxley, R. T., Sansone, V., Krym, M., Henderson, D., Schalling, M., Swanson, M. S., & Thornton, C. A. (2001). Muscleblind localizes to nuclear foci of aberrant RNA in myotonic dystrophy types 1 and 2. Human molecular genetics, 10(19), 2165–2170. https://doi.org/10.1093/hmg/10.19.2165

Marques, M. J., Conchello, J. A., & Lichtman, J. W. (2000). From Plaque to Pretzel: Fold Formation and Acetylcholine Receptor Loss at the Developing Neuromuscular Junction. Journal of Neuroscience, 20(10), 3663–3675. https://doi.org/10.1523/JNEUROSCI.20-10-03663.2000

Masters, C. L., Simms, G., Weinman, N. A., Multhaup, G., McDonald, B. L., & Beyreuther, K. (1985). Amyloid plaque core protein in Alzheimer disease and Down syndrome. Proceedings of the National Academy of Sciences, 82(12), 4245–4249. https://doi.org/10.1073/PNAS.82.12.4245

Maurel, C., Madji-Hounoum, B., Thepault, R. A., Marouillat, S., Brulard, C., Danel-Brunaud, V., Camdessanche, J. P., Blasco, H., Corcia, P., Andres, C. R., & Vourc'h, P. (2017). Mutation in the RRM2 domain of TDP-43 in Amyotrophic Lateral Sclerosis with rapid progression associated with ubiquitin positive aggregates in cultured motor neurons. Https://Doi.Org/10.1080/21678421.2017.1349152, 19(1–2), 149–151. https://doi.org/10.1080/21678421.2017.1349152

Meier, S., Bell, M., Lyons, D. N., Rodriguez-Rivera, J., Ingram, A., Fontaine, S. N., et al. (2016). Pathological tau promotes neuronal damage by impairing ribosomal function and decreasing protein synthesis. J. Neurosci. 36, 957–962. doi:10.1523/JNEUROSCI.3029-15.2016.

Mejia Maza, A. et al. NMJ-Analyser: high-throughput morphological screening of neuromuscular junctions 1 identifies subtle changes in mouse neuromuscular disease models 2 3.

Mejia Maza, A., Jarvis, S., Lee, W. C., Cunningham, T. J., Schiavo, G., Secrier, M., et al. (2021). NMJ-Analyser identifies subtle early changes in mouse models of neuromuscular disease. Sci. Rep. 11, 12251. doi:10.1038/s41598-021-91094-6.

Mejia Maza, A., Jarvis, S., Lee, W. C., Cunningham, T. J., Schiavo, G., Secrier, M., Fratta, P., Sleigh, J. N., Fisher, E. M. C., & Sudre, C. H. (2021). NMJ-Analyser identifies subtle early changes in mouse models of neuromuscular disease. Scientific Reports, 11(1), 12251. https://doi.org/10.1038/s41598-021-91094-6

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 5, pdb.prot5448. doi:10.1101/pdb.prot5448.

Milioto, C., Carcolé, M., Giblin, A., Coneys, R., Attrebi, O., Ahmed, M., Harris, S. S., Lee, B. il, Yang, M., Nirujogi, R. S., Biggs, D., Salomonsson, S., Zanovello, M., Oliveira, P. de, Katona, E., Glaria, I., Mikheenko, A., Geary, B., Udine, E., … Isaacs, A. M. (2023). PolyGR and polyPR knock-in mice reveal a conserved neuroprotective extracellular matrix signature in C9orf72 ALS/FTD neurons. *BioRxiv*, 2023.07.17.549331. https://doi.org/10.1101/2023.07.17.549331

Miller, J. W., Urbinati, C. R., Teng-Umnuay, P., Stenberg, M. G., Byrne, B. J., Thornton, C. A., & Swanson, M. S. (2000). Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. The EMBO journal, 19(17), 4439–4448. https://doi.org/10.1093/emboj/19.17.4439

Miller, T. M., Cudkowicz, M. E., Genge, A., Shaw, P. J., Sobue, G., Bucelli, R. C., Chiò, A., Van Damme, P., Ludolph, A. C., Glass, J. D., Andrews, J. A., Babu, S., Benatar, M., McDermott, C. J., Cochrane, T., Chary, S., Chew, S., Zhu, H., Wu, F., Nestorov, I., … VALOR and OLE Working Group (2022). Trial of Antisense Oligonucleotide Tofersen for SOD1 ALS. The New England journal of medicine, 387(12), 1099–1110. https://doi.org/10.1056/NEJMoa2204705

Mizielinska, S. et al. C9orf72 repeat expansions cause neurodegeneration in Drosophila through arginine-rich proteins. Science (80-.). 345, 1192–1194 (2014).

Mizielinska, S., Grönke, S., Niccoli, T., Ridler, C. E., Clayton, E. L., Devoy, A., et al. (2014). C9orf72 repeat expansions cause neurodegeneration in Drosophila through arginine-rich proteins. Science (80-.). 345, 1192–1194. doi:10.1126/science.1256800.

Mizielinska, S., Lashley, T., Norona, F. E., Clayton, E. L., Ridler, C. E., Fratta, P., et al. (2013). C9orf72 frontotemporal lobar degeneration is characterised by frequent neuronal sense and antisense RNA foci. Acta Neuropathol. 126, 845–857. doi:10.1007/s00401-013-1200-z.

Mohammadi, S., Zuckerman, N., Goldsmith, A., and Grama, A. (2017). A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. Proc. IEEE 105, 340–366. doi:10.1109/JPROC.2016.2607121.

Mok, K., Traynor, B. J., Schymick, J., Tienari, P. J., Laaksovirta, H., Peuralinna, T., et al. (2012). The chromosome 9 ALS and FTD locus is probably derived from a single founder. Neurobiol. Aging 33, 209.e3-209.e8. doi:10.1016/j.neurobiolaging.2011.08.005.

Moll, P., Ante, M., Seitz, A. et al. QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods 11, i–iii (2014). https://doi.org/10.1038/nmeth.f.376

Moloney, E. B., de Winter, F., & Verhaagen, J. (2014). ALS as a distal axonopathy: molecular mechanisms affecting neuromuscular junction stability in the presymptomatic stages of the disease. Frontiers in Neuroscience, 8(8 JUL). https://doi.org/10.3389/FNINS.2014.00252

Morris, K. V., and Mattick, J. S. (2014). The rise of regulatory RNA. Nat. Rev. Genet. 15, 423–437. doi:10.1038/nrg3722.

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

National Institute of Health Amyotrophic Lateral Sclerosis (ALS) Fact Sheet | National Institute of Neurological Disorders and Stroke. Natl. Inst. Heal. Available at: https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Amyotrophic-Lateral-Sclerosis-ALS-Fact-Sheet [Accessed April 27, 2021].

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

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., McCluskey, L. F., Miller, B. L., Masliah, E., Mackenzie, I. R., Feldman, H., Feiden, W., Kretzschmar, H. A., Trojanowski, J. Q., & Lee, V. M. Y. (2006). Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science, 314(5796), 130–133. https://doi.org/10.1126/science.1134108

Neumann, M., Sampathu, D. M., Kwong, L. K., Truax, A. C., Micsenyi, M. C., Chou, T. T., et al. (2006). Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science (80-.). 314, 130–133. doi:10.1126/science.1134108.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., & Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nature Methods, 12(5), 453–457. https://doi.org/10.1038/NMETH.3337

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. 37, 773–782. doi:10.1038/s41587-019-0114-2.

NGS Library Preparation Available at: https://emea.Illumina.com/techniques/sequencing/ngs-library-prep.html [Accessed April 15, 2021].

Nguyen, H. P., Van Broeckhoven, C., and van der Zee, J. (2018). ALS Genes in the Genomic Era and their Implications for FTD. Trends Genet. 34, 404–423. doi:10.1016/j.tig.2018.03.001.

Nicolas, A., Kenna, K., Renton, A. E., Ticozzi, N., Faghri, F., Chia, R., Dominov, J. A., Kenna, B. J., Nalls, M. A., Keagle, P., Rivera, A. M., van Rheenen, W., Murphy, N. A., van Vugt, J. J. F. A., Geiger, J. T., van der Spek, R., Pliner, H. A., Shankaracharya, Smith, B. N., … Traynor, B. J. (2018). Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. Neuron, 97(6), 1268-1283.e6. https://doi.org/10.1016/J.NEURON.2018.02.027

Nosková, L., Stránecký, V., Hartmannová, H., Přistoupilová, A., Barešová, V., Ivánek, R., Hlková, H., Jahnová, H., Van Der Zee, J., Staropoli, J. F., Sims, K. B., Tyynelä, J., Van Broeckhoven, C., Nijssen, P. C. G., Mole, S. E., Elleder, M., & Kmoch, S. (2011). Mutations in DNAJC5, Encoding Cysteine-String Protein Alpha, Cause Autosomal-Dominant Adult-Onset Neuronal Ceroid Lipofuscinosis. American Journal of Human Genetics, 89(2), 241. https://doi.org/10.1016/J.AJHG.2011.07.003

O'Rourke, J. G., Bogdanik, L., Yáñez, A., Lall, D., Wolf, A. J., Muhammad, A. K. M. G., Ho, R., Carmona, S., Vit, J. P., Zarrow, J., Kim, K. J., Bell, S., Harms, M. B., Miller, T. M., Dangler, C. A., Underhill, D. M., Goodridge, H. S., Lutz, C. M., & Baloh, R. H. (2016). C9orf72 is required for proper macrophage and microglial function in mice. Science (New York, N.Y.), 351(6279), 1324–1329. https://doi.org/10.1126/SCIENCE.AAF1064

Oh, S., Gim, J., Lee, J. K., Park, H., and Shin, O. S. (2020). Coxsackievirus B3 infection of human neural progenitor cells results in distinct expression patterns of innate immune genes. Viruses 12:325.

Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D., & Ragoussis, J. (2016). Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Scientific Reports 2016 6:1, 6(1), 1–13. https://doi.org/10.1038/srep31602

Olsen, T. K., and Baryawno, N. (2018). Introduction to Single-Cell RNA Sequencing. Curr. Protoc. Mol. Biol. 122. doi:10.1002/cpmb.57.

Olszewska, D. A., Lonergan, R., Fallon, E. M., and Lynch, T. (2016). Genetics of Frontotemporal Dementia. Curr. Neurol. Neurosci. Rep. 16, 1–15. doi:10.1007/s11910-016-0707-9.

Onyike, C. U., and Diehl-Schmid, J. (2013). The epidemiology of frontotemporal dementia. Int. Rev. Psychiatry 25, 130–137. doi:10.3109/09540261.2013.776523.

Papanikolopoulou, K., Roussou, I. G., Gouzi, J. Y., Samiotaki, M., Panayotou, G., Turin, L., et al. (2019). Drosophila tau negatively regulates translation and olfactory long-term memory, but facilitates footshock habituation and cytoskeletal homeostasis. J. Neurosci. 39, 8315–8329. doi:10.1523/jneurosci.0391-19.2019.

Pearson, J. P., Williams, N. M., Majounie, E., Waite, A., Stott, J., Newsway, V., et al. (2011). Familial frontotemporal dementia with amyotrophic lateral sclerosis and a shared haplotype on chromosome 9p. J. Neurol. 258, 647–655. doi:10.1007/s00415-010-5815-x.

Perea, J. R., Bolós, M., & Avila, J. (2020). Microglia in Alzheimer's Disease in the Context of Tau Pathology. Biomolecules, 10(10), 1–26. https://doi.org/10.3390/BIOM10101439

Polymenidou, M., Lagier-tourenne, C., Hutt, K. R., Stephanie, C., Moran, J., Liang, T. Y., et al. (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nat. Neurosci. 14, 459–468. doi:10.1038/nn.2779.Long.

Prudencio, M., Belzil, V. V., Batra, R., Ross, C. A., Gendron, T. F., Pregent, L. J., Murray, M. E., Overstreet, K. K., Piazza-Johnston, A. E., Desaro, P., Bieniek, K. F., DeTure, M., Lee, W. C., Biendarra, S. M., Davis, M. D., Baker, M. C., Perkerson, R. B., Van Blitterswijk, M., Stetler, C. T., … Petrucelli, L. (2015). Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. Nature Neuroscience, 18(8), 1175. https://doi.org/10.1038/NN.4065

Rademakers, R., Cruts, M., & van Broeckhoven, C. (2004). The role of tau (MAPT) in frontotemporal dementia and related tauopathies. Human mutation, 24(4), 277–295. https://doi.org/10.1002/humu.20086

Rascovsky, K., Hodges, J. R., Knopman, D., Mendez, M. F., Kramer, J. H., Neuhaus, J., van Swieten, J. C., Seelaar, H., Dopper, E. G. P., Onyike, C. U., Hillis, A. E., Josephs, K. A., Boeve, B. F., Kertesz, A., Seeley, W. W., Rankin, K. P., Johnson, J. K., Gorno-Tempini, M. L., Rosen, H., … Miller, B. L. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain, 134(9), 2456. https://doi.org/10.1093/BRAIN/AWR179

Ratnavalli, E., Brayne, C., Dawson, K., & Hodges, J. R. (2002). The prevalence of frontotemporal dementia. Neurology, 58(11), 1615–1621. https://doi.org/10.1212/WNL.58.11.1615

Ratti, A., & Buratti, E. (2016). Physiological functions and pathobiology of TDP-43 and FUS/TLS proteins. Journal of Neurochemistry, 138 Suppl 1, 95–111. https://doi.org/10.1111/JNC.13625

RBM18 RNA binding motif protein 18 [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved March 22, 2023, from https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=92400

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., et al. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron 72, 257–268. doi:10.1016/j.neuron.2011.09.010.

Reus, L. M., Jansen, I. E., Mol, M. O., van Ruissen, F., van Rooij, J., van Schoor, N. M., Tesi, N., Reinders, M. J. T., Huisman, M. A., Holstege, H., Visser, P. J., de Boer, S. C. M., Hulsman, M., Ahmad, S., Amin, N., Uitterlinden, A. G., Ikram, A., van Duijn, C. M., Seelaar, H., … van der Lee, S. J. (2021). Genome-wide association study of frontotemporal dementia identifies a C9ORF72 haplotype with a median of 12-G4C2 repeats that predisposes to pathological repeat expansions. Translational Psychiatry 2021 11:1, 11(1), 1–8. https://doi.org/10.1038/s41398-021-01577-3

Ricketts, T., McGoldrick, P., Fratta, P., De Oliveira, H. M., Kent, R., Phatak, V., Brandner, S., Blanco, G., Greensmith, L., Acevedo-Arozena, A., & Fisher, E. M. C. (2014). A Nonsense Mutation in Mouse Tardbp Affects TDP43 Alternative Splicing Activity and Causes Limb-Clasping and Body Tone Defects. PLOS ONE, 9(1), e85962. https://doi.org/10.1371/JOURNAL.PONE.0085962

Rizzu, P., Blauwendraat, C., Heetveld, S., Lynes, E. M., Castillo-Lizardo, M., Dhingra, A., Pyz, E., Hobert, M., Synofzik, M., Simón-Sánchez, J., Francescatto, M., & Heutink, P. (2016). C9orf72 is differentially expressed in the central nervous system and myeloid cells and consistently reduced in C9orf72, MAPT and GRN mutation carriers. Acta Neuropathologica Communications, 4(1), 37. https://doi.org/10.1186/S40478-016-0306-7/TABLES/3

Rogelj, B., Easton, L. E., Bogu, G. K., Stanton, L. W., Rot, G., Curk, T., et al. (2012). Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. Sci. Rep. 2, 1–10. doi:10.1038/srep00603.

Rosen, D. R., Siddique, T., Patterson, D., Figlewicz, D. A., Sapp, P., Hentati, A., et al. (1993). Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature 362, 59–62. doi:10.1038/362059a0.

Rostalski, H., Leskelä, S., Huber, N., Katisko, K., Cajanus, A., Solje, E., Marttinen, M., Natunen, T., Remes, A. M., Hiltunen, M., & Haapasalo, A. (2019). Astrocytes and Microglia as Potential Contributors to the Pathogenesis of C9orf72 Repeat Expansion-Associated FTLD and ALS. Frontiers in neuroscience, 13, 486. https://doi.org/10.3389/fnins.2019.00486

Rot, G., Wang, Z., Huppertz, I., Modic, M., Lenˇce, T., Hallegger, M., et al. (2017). High-resolution RNA maps suggest common principles of splicing and polyadenylation regulation by TDP-43. Cell Rep. 19, 1056–1067. doi: 10.1016/j. celrep.2017.04.028

Sammeth, M., Foissac, S. & Guigó, R. A General Definition and Nomenclature for Alternative Splicing Events. PLoS Comput. Biol. 4, e1000147 (2008).

Scekic-Zahirovic, J., Sendscheid, O., El Oussini, H., Jambeau, M., Sun, Y., Mersmann, S., et al. (2016). Toxic gain of function from mutant FUS protein is crucial to trigger cell autonomous motor neuron loss. EMBO J. 35, 1077–1097. doi:10.15252/embj.201592559.

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

Scotter, E. L., Chen, H. J., & Shaw, C. E. (2015). TDP-43 Proteinopathy and ALS: Insights into Disease Mechanisms and Therapeutic Targets. Neurotherapeutics, 12(2), 352. https://doi.org/10.1007/S13311-015-0338-X

Senol Cali, D., Kim, J. S., Ghose, S., Alkan, C., and Mutlu, O. (2018). Nanopore sequencing technology and tools for genome assembly: Computational analysis of the current state, bottlenecks and future directions. Brief. Bioinform. 20, 1542–1559. doi:10.1093/bib/bby017.

Sephton, C. F. et al. Identification of neuronal RNA targets of TDP-43-containing ribonucleoprotein complexes. J. Biol. Chem. 286, 1204–1215 (2011).

Sephton, C. F. et al. TDP-43 is a developmentally regulated protein essential for early embryonic development. J. Biol. Chem. 285, 6826–6834 (2010).

Sha, S. J., Takada, L. T., Rankin, K. P., Yokoyama, J. S., Rutherford, N. J., Fong, J. C., Khan, B., Karydas, A., Baker, M. C., De Jesus-Hernandez, M., Pribadi, M., Coppola, G., Geschwind, D. H., Rademakers, R., Lee, S. E., Seeley, W., Miller, B. L., & Boxer, A. L. (2012). Frontotemporal dementia due to C9ORF72 mutations: Clinical and imaging features. Neurology, 79(10), 1002. https://doi.org/10.1212/WNL.0B013E318268452E

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

Shen-Orr, S. S., & Gaujoux, R. (2013). Computational Deconvolution: Extracting Cell Type-Specific Information from Heterogeneous Samples. Current Opinion in Immunology, 25(5), 571–578. https://doi.org/10.1016/J.COI.2013.09.015

Slater, C. R. (2017). The Structure of Human Neuromuscular Junctions: Some Unanswered Molecular Questions. International Journal of Molecular Sciences 2017, Vol. 18, Page 2183, 18(10), 2183. https://doi.org/10.3390/IJMS18102183

Slater, C. The Structure of Human Neuromuscular Junctions: Some Unanswered Molecular Questions. Int. J. Mol. Sci. 18, 2183 (2017slater).

Sleigh, J. N., Burgess, R. W., Gillingwater, T. H., & Cader, M. Z. (2014). Morphological analysis of neuromuscular junction development and degeneration in rodent lumbrical muscles. Journal of Neuroscience Methods, 227, 159–165. https://doi.org/10.1016/J.JNEUMETH.2014.02.005

Sleigh, J. N., Grice, S. J., Burgess, R. W., Talbot, K., & Cader, M. Z. (2014). Neuromuscular junction maturation defects precede impaired lower motor neuron connectivity in Charcot–Marie–Tooth type 2D mice. Human Molecular Genetics, 23(10), 2639–2650. https://doi.org/10.1093/HMG/DDT659

Slomovic, S., Laufer, D., Geiger, D., & Schuster, G. (2006). Polyadenylation of ribosomal RNA in human cells. Nucleic Acids Research, 34(10), 2966. https://doi.org/10.1093/NAR/GKL357

Sofroniew, M. V., & Vinters, H. V. (2010). Astrocytes: biology and pathology. Acta Neuropathologica, 119(1), 7. https://doi.org/10.1007/S00401-009-0619-8

Sommer, C., & Gerlich, D. W. (2013). Machine learning in cell biology-teaching computers to recognize phenotypes. Journal of Cell Science, 126(24), 5529–5539. https://doi.org/10.1242/JCS.123604/263567/AM/MACHINE-LEARNING-IN-CELL-BIOLOGY-TEACHING

Soulet, D., & Rivest, S. (2008). Microglia. Current Biology, 18(12), R506–R508. https://doi.org/10.1016/j.cub.2008.04.047

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nat. Rev. Genet. 20, 631–656. doi:10.1038/s41576-019-0150-2.

Stevens, C. H., Guthrie, N. J., Van Roijen, M., Halliday, G. M., and Ooi, L. (2019). Increased tau phosphorylation in motor neurons from clinically pure sporadic amyotrophic lateral sclerosis patients. J. Neuropathol. Exp. Neurol. 78, 605–614. doi:10.1093/jnen/nlz041.

Stribl, C., Samara, A., Trümbach, D., Peis, R., Neumann, M., Fuchs, H., Gailus-Durner, V., De Angelis, M. H., Rathkolb, B., Wolf, E., Beckers, J., Horsch, M., Neff, F., Kremmer, E., Koob, S., Reichert, A. S., Hans, W., Rozman, J., Klingenspor, M., … Floss, T. (2014). Mitochondrial Dysfunction and Decrease in Body Weight of a Transgenic Knock-in Mouse Model for TDP-43 *. Journal of Biological Chemistry, 289(15), 10769–10784. https://doi.org/10.1074/JBC.M113.515940

Suk, T. R. & Rousseaux, M. W. C. The role of TDP-43 mislocalization in amyotrophic lateral sclerosis. Molecular Neurodegeneration vol. 15 45 (2020).

Suk, T. R., & Rousseaux, M. W. C. (2020). The role of TDP-43 mislocalization in amyotrophic lateral sclerosis. Molecular Neurodegeneration, 15(1), 1–16. https://doi.org/10.1186/S13024-020-00397-1/FIGURES/2

Sun, Y., Eshov, A., Zhou, J., Isiktas, A. U., & Guo, J. U. (2020). C9orf72 arginine-rich dipeptide repeats inhibit UPF1-mediated RNA decay via translational repression. Nature Communications 2020 11:1, 11(1), 1–9. https://doi.org/10.1038/s41467-020-17129-0

Sun, Z., Diaz, Z., Fang, X., Hart, M. P., Chesi, A., Shorter, J., et al. (2011). Molecular Determinants and Genetic Modifiers of Aggregation and Toxicity for the ALS Disease Protein FUS/TLS. PLoS Biol. 9, e1000614. doi:10.1371/journal.pbio.1000614.

Sutton, G. J., and Voineagu, I. (2020). Comprehensive evaluation of human brain gene expression deconvolution methods. bioRxiv doi:10.1101/2020.06.01.126839.

Sutton, G. J., Poppe, D., Simmons, R. K., Walsh, K., Nawaz, U., Lister, R., Gagnon-Bartsch, J. A., & Voineagu, I. (2022). Comprehensive evaluation of deconvolution methods for human brain gene expression. Nature Communications 2022 13:1, 13(1), 1–18. https://doi.org/10.1038/s41467-022-28655-4

Tarpey, M. D., Amorese, A. J., Balestrieri, N. P., Ryan, T. E., Schmidt, C. A., McClung, J. M., & Spangenburg, E. E. (2018). Characterization and utilization of the flexor digitorum brevis for assessing skeletal muscle function. Skeletal Muscle, 8(1), 1–15. https://doi.org/10.1186/S13395-018-0160-3/FIGURES/8

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

Tollervey, J. R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A. L., Zupunski, V., Patani, R., Chandran, S., Rot, G., Zupan, B., Shaw, C. E., & Ule, J. (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nature neuroscience, 14(4), 452–458. https://doi.org/10.1038/nn.2778

van Es, M. A., Hardiman, O., Chio, A., Al-Chalabi, A., Pasterkamp, R. J., Veldink, J. H., & van den Berg, L. H. (2017). Amyotrophic lateral sclerosis. The Lancet, 390(10107), 2084–2098. https://doi.org/10.1016/S0140-6736(17)31287-4 van Es, M. A., Veldink, J. H., Saris, C. G., Blauw, H. M., van Vught, P. W., Birve, A., Lemmens, R., Schelhaas, H. J., Groen, E. J., Huisman, M. H., van der Kooi, A. J., de Visser, M., Dahlberg, C., Estrada, K., Rivadeneira, F., Hofman, A., Zwarts, M. J., van Doormaal, P. T., Rujescu, D., Strengman, E., … van den Berg, L. H. (2009). Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. Nature genetics, 41(10), 1083–1087. https://doi.org/10.1038/ng.442

Van Langenhove, T., Van Der Zee, J., Gijselinck, I., Engelborghs, S., Vandenberghe, R., Vandenbulcke, M., De Bleecker, J., Sieben, A., Versijpt, J., Ivanoiu, A., Deryck, O., Willems, C., Dillen, L., Philtjens, S., Maes, G., Bäumer, V., Van Den Broeck, M., Mattheijssens, M., Peeters, K., … Van Broeckhoven, C. (2013). Distinct Clinical Characteristics of C9orf72 Expansion Carriers Compared With GRN, MAPT, and Nonmutation Carriers in a Flanders-Belgian FTLD Cohort. JAMA Neurology, 70(3), 365–373. https://doi.org/10.1001/2013.JAMANEUROL.181

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

Van Mossevelde, S., van der Zee, J., Cruts, M., and Van Broeckhoven, C. (2017). Relationship between C9orf72 repeat size and clinical phenotype. Curr. Opin. Genet. Dev. 44, 117–124. doi:10.1016/j.gde.2017.02.008.

van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., van der Spek, R. A. A., Võsa, U., de Jong, S., Robinson, M. R., Yang, J., Fogh, I., van Doormaal, P. T. C., Tazelaar, G. H. P., Koppers, M., Blokhuis, A. M., Sproviero, W., Jones, A. R., Kenna, K. P., … Veldink, J. H. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nature Genetics, 48(9), 1043. https://doi.org/10.1038/NG.3622

van Rheenen, W., van der Spek, R. A. A., Bakker, M. K., van Vugt, J. J. F. A., Hop, P. J., Zwamborn, R. A. J., de Klein, N., Westra, H. J., Bakker, O. B., Deelen, P., Shireby, G., Hannon, E., Moisse, M., Baird, D., Restuadi, R., Dolzhenko, E., Dekker, A. M., Gawor, K., Westeneng, H. J., … Veldink, J. H. (2021). Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. Nature Genetics 2021 53:12, 53(12), 1636–1648. https://doi.org/10.1038/s41588-021-00973-1

Vance, C., Rogelj, B., Hortobágyi, T., De Vos, K. J., Nishimura, A. L., Sreedharan, J., et al. (2009). Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. Science (80-.). 323, 1208–1211. doi:10.1126/science.1165942.

Vance, C., Scotter, E. L., Nishimura, A. L., Troakes, C., Mitchell, J. C., Kathe, C., et al. (2013). ALS mutant FUS disrupts nuclear localization and sequesters wild-type FUS within cytoplasmic stress granules. Hum. Mol. Genet. 22, 2676–2688. doi:10.1093/hmg/ddt117.

Verduijn, J., Van der Meeren, L., Krysko, D. V., & Skirtach, A. G. (2021). Deep learning with digital holographic microscopy discriminates apoptosis and necroptosis. Cell Death Discovery 2021 7:1, 7(1), 1–10. https://doi.org/10.1038/s41420-021-00616-8 von Bartheld, C. S., Bahney, J., & Herculano-Houzel, S. (2016). The Search for True Numbers of Neurons and Glial Cells in the Human Brain: A Review of 150 Years of Cell Counting. The Journal of Comparative Neurology, 524(18), 3865. https://doi.org/10.1002/CNE.24040

Wang, E. T., Cody, N. A. L., Jog, S., Biancolella, M., Wang, T. T., Treacy, D. J., Luo, S., Schroth, G. P., Housman, D. E., Reddy, S., Lécuyer, E., & Burge, C. B. (2012). Transcriptome-wide Regulation of Pre-mRNA Splicing and mRNA Localization by Muscleblind Proteins. Cell, 150(4), 710. https://doi.org/10.1016/J.CELL.2012.06.041

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470–476. doi:10.1038/nature07509.

Wang, H., Guo, W., Mitra, J., Hegde, P. M., Vandoorne, T., Eckelmann, B. J., et al. (2018). Mutant FUS causes DNA ligation defects to inhibit oxidative damage repair in Amyotrophic Lateral Sclerosis. Nat. Commun. 9, 1–18. doi:10.1038/s41467-018-06111-6.

Wang, T., Li, B., Nelson, C. E., & Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics, 20(1), 1–16. https://doi.org/10.1186/S12859-019-2599-6/TABLES/7

Wang, X., Park, J., Susztak, K. et al. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun 10, 380 (2019). https://doi.org/10.1038/s41467-018-08023-x

Wang, X., Park, J., Susztak, K., Zhang, N. R., & Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature Communications, 10(1). https://doi.org/10.1038/S41467-018-08023-X

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. doi:10.1038/nrg2484.

Wegorzewska, I., Bell, S., Cairns, N. J., Miller, T. M., and Baloh, R. H. (2009). TDP-43 mutant transgenic mice develop features of ALS and frontotemporal lobar degeneration. Proc. Natl. Acad. Sci. U. S. A. 106, 18809–18814. doi:10.1073/pnas.0908767106.

White, K., Yang, P., Li, L., Farshori, A., Medina, A. E., & Zielke, H. R. (2018). Effect of Postmortem Interval and Years in Storage on RNA Quality of Tissue at a Repository of the NIH NeuroBioBank. Biopreservation and Biobanking, 16(2), 148–157. https://doi.org/10.1089/bio.2017.0099

White, M. A., Kim, E., Duffy, A., Adalbert, R., Phillips, B. U., Peters, O. M., Stephenson, J., Yang, S., Massenzio, F., Lin, Z., Andrews, S., Segonds-Pichon, A., Metterville, J., Saksida, L. M., Mead, R., Ribchester, R. R., Barhomi, Y., Serre, T., Coleman, M. P., Fallon, J. R., … Sreedharan, J. (2018). TDP-43 gains function due to perturbed autoregulation in a Tardbp knock-in mouse model of ALS-FTD. Nature neuroscience, 21(4), 552–563. https://doi.org/10.1038/s41593-018-0113-5

Willadt, S., Nash, M., & Slater, C. (2018). Age-related changes in the structure and function of mammalian neuromuscular junctions. Annals of the New York Academy of Sciences, 1412(1), 41–53. https://doi.org/10.1111/nyas.13521

Woollacott, I. O. C., and Rohrer, J. D. (2016). The clinical spectrum of sporadic and familial forms of frontotemporal dementia. J. Neurochem. 138, 6–31. doi:10.1111/jnc.13654.

Wu, L. S. et al. TDP-43, a neuro-pathosignature factor, is essential for early mouse embryogenesis. Genesis 48, 56–62 (2010).

Wu, L. S., Cheng, W. C., Chen, C. Y., Wu, M. C., Wang, Y. C., Tseng, Y. H., Chuang, T. J., & Shen, C. K. J. (2019). Transcriptomopathies of pre- and post-symptomatic frontotemporal dementia-like mice with TDP-43 depletion in forebrain neurons. Acta Neuropathologica Communications, 7(1), 50. https://doi.org/10.1186/S40478-019-0674-X

Yang, C., Wang, H., Qiao, T., Yang, B., Aliaga, L., Qiu, L., Tan, W., Salameh, J., McKenna-Yasek, D. M., Smith, T., Peng, L., Moore, M. J., Brown, R. H., Cai, H., & Xu, Z. (2014). Partial loss of TDP-43 function causes phenotypes of amyotrophic lateral sclerosis. Proceedings of the National Academy of Sciences of the United States of America, 111(12), E1121. https://doi.org/10.1073/PNAS.1322641111/-/DCSUPPLEMENTAL

Yang, H.-S., White, C. C., Klein, H.-U., Schneider, J. A., Bennett, D. A., & De Jager Correspondence, P. L. (2020). Genetics of Gene Expression in the Aging Human Brain Reveal TDP-43 Proteinopathy Pathophysiology. Neuron, 107, 496-508.e6. https://doi.org/10.1016/j.neuron.2020.05.010

Yang, M., Chen, L., Swaminathan, K., Herrlinger, S., Lai, F., Shiekhattar, R., & Chen, J. F. (2016). A C9ORF72/SMCR8-containing complex regulates ULK1 and plays a dual role in autophagy. Science Advances, 2(9). https://doi.org/10.1126/SCIADV.1601167

Yoshizawa, T., Ali, R., Jiou, J., Fung, H. Y. J., Burke, K. A., Kim, S. J., et al. (2018). Nuclear Import Receptor Inhibits Phase Separation of FUS through Binding to Multiple Sites. Cell 173, 693-705.e22. doi:10.1016/j.cell.2018.03.003.

Young, A. L., Bocchetta, M., Russell, L. L., Convery, R. S., Peakman, G., Todd, E., Cash, D. M., Greaves, C. V., van Swieten, J., Jiskoot, L., Seelaar, H., Moreno, F., Sanchez-Valle, R., Borroni, B., Laforce, R., Masellis, M., Tartaglia, M. C., Graff, C., Galimberti, D., … Rohrer, J. D. (2021). Characterizing the Clinical Features and Atrophy Patterns of MAPT-Related Frontotemporal Dementia With Disease Progression Modeling. Neurology, 97(9), e941–e952. https://doi.org/10.1212/WNL.0000000000012410

yWorks GmbH. (2019). yEd. Retrieved from https://www.yworks.com/products/yed

Zaitsev, K., Bambouskova, M., Swain, A., & Artyomov, M. N. (2019). Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. Nature Communications, 10(1). https://doi.org/10.1038/S41467-019-09990-5

Zhang, K., Donnelly, C. J., Haeusler, A. R., Grima, J. C., Machamer, J. B., Steinwald, P., et al. (2015). The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. Nature 525, 56–61. doi:10.1038/nature14973.

Zhang, S., Cooper-Knock, J., Weimer, A. K., Shi, M., Moll, T., Marshall, J. N. G., Harvey, C., Nezhad, H. G., Franklin, J., Souza, C. dos S., Ning, K., Wang, C., Li, J., Dilliott, A. A., Farhan, S., Elhaik, E., Pasniceanu, I., Livesey, M. R., Eitan, C., … Snyder, M. P. (2022). Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis. Neuron, 110(6), 992-1008.e11. https://doi.org/10.1016/J.NEURON.2021.12.019

Zhang, Y. J. et al. Poly(GR) impairs protein translation and stress granule dynamics in C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis. Nat. Med. 24, 1136–1142 (2018).

Zhang, Y. J., Gendron, T. F., Ebbert, M. T. W., O'Raw, A. D., Yue, M., Jansen-West, K., et al. (2018). Poly(GR) impairs protein translation and stress granule dynamics in C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis. Nat. Med. 24, 1136–1142. doi:10.1038/s41591-018-0071-1.

Zhang, Y. J., Gendron, T. F., Ebbert, M. T. W., O'Raw, A. D., Yue, M., Jansen-West, K., Zhang, X., Prudencio, M., Chew, J., Cook, C. N., Daughrity, L. M., Tong, J., Song, Y., Pickles, S. R., Castanedes-Casey, M., Kurti, A., Rademakers, R., Oskarsson, B., Dickson, D. W., … Petrucelli, L. (2018a). Poly(GR) impairs protein translation and stress granule dynamics in C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis. Nature Medicine 2018 24:8, 24(8), 1136–1142. https://doi.org/10.1038/s41591-018-0071-1

Zhang, Y. J., Gendron, T. F., Ebbert, M. T. W., O'Raw, A. D., Yue, M., Jansen-West, K., Zhang, X., Prudencio, M., Chew, J., Cook, C. N., Daughrity, L. M., Tong, J., Song, Y., Pickles, S. R., Castanedes-Casey, M., Kurti, A., Rademakers, R., Oskarsson, B., Dickson, D. W., … Petrucelli, L. (2018b). Poly(GR) impairs protein translation and stress granule dynamics in C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis. Nature Medicine, 24(8), 1136. https://doi.org/10.1038/S41591-018-0071-1

Zhang, Y. J., Guo, L., Gonzales, P. K., Gendron, T. F., Wu, Y., Jansen-West, K., O'Raw, A. D., Pickles, S. R., Prudencio, M., Carlomagno, Y., Gachechiladze, M. A., Ludwig, C., Tian, R., Chew, J., DeTure, M., Lin, W. L., Tong, J., Daughrity, L. M., Yue, M., Song, Y., … Petrucelli, L. (2019). Heterochromatin anomalies and double-stranded RNA accumulation underlie C9orf72 poly(PR) toxicity. Science (New York, N.Y.), 363(6428), eaav2606. https://doi.org/10.1126/science.aav2606

Ziff, O. J., Clarke, B. E., Taha, D. M., Crerar, H., Luscombe, N. M., & Patani, R. (2022). Meta-analysis of human and mouse ALS astrocytes reveals multi-omic signatures of inflammatory reactive states. Genome Research, 32(1), 71–84. https://doi.org/10.1101/GR.275939.121/-/DC1

Zinchuk, V., & Grossenbacher-Zinchuk, O. (2020). Machine Learning for Analysis of Microscopy Images: A Practical Guide. Current Protocols in Cell Biology, 86(1), e101. https://doi.org/10.1002/CPCB.101