THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Understanding the Mechanisms of ALS through RNA Sequencing

Jayden Beckwith

Submitted in partial fulfillment for the degree of Master of Bioinformatics at

The University of Queensland in 2023

BIOX7008: Major Research Project & Seminar

Institute for Molecular Bioscience

Semester 1 2023

46267539

Abstract

Background: Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease with a lifetime risk of ~1:400. Degeneration of the motor neurones that innervate skeletal muscle results in progressive disability. TDP-43, a protein critical for RNA, forms aggregates and is a hallmark of disease, occurring in almost all ALS cases. Recently an ALS GWAS risk locus (in *UNC13A*) described a new mechanism of disease - aggregated TDP-43 caused altered splicing of *UNC13A* (enhanced by risk-SNP), reducing UNC13A synaptic ability. Splicing is a critical control mechanism for normal function and disruption can impact essential processes. Altered splicing discovery in ALS has been limited - restricted to brain/neuronal tissue. Thus, it is unclear how widespread these changes might be. Using RNA from matched blood and skeletal muscle, I carried out preliminary analyses to explore how gene expression and aberrant splicing (AbS) may contribute to disease.

Methods: RNA-seq (Illumina, HiSeq 4000) was generated in skeletal muscle (n_{cases} = 30, $n_{controls}$ = 20) and as available, their matched blood (n_{cases} = 30, $n_{controls}$ = 19). Differential gene expression analysis was performed on quality-controlled gene-level counts (DESeq2/ClusterProfiler) followed by splice analyses. Four complementary tools/approaches were used: FRASER (local "sample-by-sample"), rMATS (group, control vs case), LeafCutterMD (candidate) and SpliceAI (variant-based).

Results: Gene expression changes in ALS detected 58 differentially expressed (DE) protein-coding genes (FDR < 0.05 and log_2FC = -1/+1) enriched in pathways such as inflammation (muscle) and tissue regeneration (blood). Splice results focused on events specific to cases. Significant local changes (FRASER) were detected (AbS events n_{cases} = 21, FDR < 0.05). Similarly, AbS sites were observed at the group level (AbS events n_{cases} = 11, shorter isoform, FDR < 0.01). This included *MACF1* in ALS muscle and a known ALS gene, *HNRNPA2B1* in blood. At a variant level (SpliceAI), 9 SNVs with a high probability (≥ 80% chance) of AbS were found (at a threshold based on exclusivity in n_{cases} ≥ 3), including the *SAA1* missense variant (muscle) and *SORL1* variants in ALS blood.

Conclusions: This unique matched transcriptome analysis demonstrated DE and AbS changes in ALS. Potentially functionally relevant findings were identified including one in a pathogenic ALS gene (*HNRNPA2B1*). AbS events demonstrated complex variation in individuals and tissues. ALS relevant AbS is not necessarily widespread i.e., *UNC13A* expression was low, we were unable to detect splice events in muscle or blood. Our results suggest the ALS transcriptome differs across tissues, without a clear priority tissue (i.e., we find disease relevant events in blood and muscle for follow-up). Exploring this variation together with a larger cohort is needed. With limited treatment options, targeting splicing could be therapeutically targeted to reduce disease risk or progression.

Acknowledgements

First and foremost, I'd like to express my gratitude to my research supervisor, Dr Fleur Garton, for her patience and feedback throughout the semester. She has provided me with an invaluable breadth of knowledge and support for my first major research project, and for that I thank her deeply.

I would also like to express many thanks to Dr Anna Freydenzon and colleagues of McRae group, for additional support in providing feedback and for supplying the raw biological dataset used in this body of work.

I would be remiss in not mentioning my friends and family, especially my mother. Their belief in me has kept my motivation high during this process.

Lastly, this endeavour would not have been possible without the participants giving their time and biological samples for the use of research and fight against MND.

Contents

Abbreviations

Aβ: Amyloid beta

AbS: Aberrant splicing

ALSFRS: Revised Amyotrophic Lateral Sclerosis Functional Rating Scale

ALS: Amyotrophic lateral sclerosis

AS: Alternative splicing

AUC-PR: Area under the precision recall curve

A3SS: Alternative 3' splice site

A5SS: Alternative 5' splice site

BAM: Binary alignment map

CE: Cryptic exon

CNS*:* Central nervous system

C9orf72: Chromosome 9 open reading frame 72

EM: Expectation maximisation

ER*:* Endoplasmic reticulum

ES: Exon skipping

fALS: Familial amyotrophic lateral sclerosis

FDR: False discovery rate

FGF: Fibroblast growth factor

FRASER: Find rare splicing events in RNA-seq

FTD: Frontotemporal dementia

GLM: Generalised linear model

GO: Gene ontology

GWAS: Genome-wide association studies

iPSC MNs: iPSC-derived motor neurones

IR: Intron retention

J Index*:* Jaccard index

JCEC: Junction-exon counts

Logit: Logistic regression

Ls: Latent space

MND: Motor neurone disease

MXE: Mutually exclusive exon

PCA: Principal component analysis

PSI: Percent spliced in

QC: Quality control

RIN: RNA integrity number

rMATS: Replicate-based multivariate analysis of transcript splicing

RNA-seq*:* RNA sequencing

ROS**:** Reactive oxidative stress

RT: Room temperature

SAA*:* Serum Amyloid A

sALS**:** Sporadic amyotrophic lateral sclerosis

SOD1: Superoxide dismutase

TARDBP: Tar DNA-binding protein

TDP-43: Tar DNA-binding protein 43

UPS: Ubiquitin-proteosome system

Introduction

Amyotrophic lateral sclerosis

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterised by the progressive loss of motor neurones, which are responsible for controlling muscle movement[3]. ALS is the most common form of all Motor Neurone Disease (MND). It accounts for about 80-90% cases, with death from respiratory compromise usually just 2-5 years after symptom onset[54]. The incidence of ALS is rare, with just approximately 2-3 cases per 100,000 people per year across different populations[2]. While, no clear geographic pattern exists, aging is a significant risk factor[11]. Onset of the disease is typically after the fifth decade and an ALS diagnosis is made once there is progressive degeneration of both upper (symptoms such as hyperreflexia, muscle weakness and spasticity) lower (symptoms such as muscle twitching and atrophy) motor neurones[43,67]. This usually takes approximately 12 months as the presentation of ALS tends to be insidious, manifesting as subtle weakness in specific muscle groups, such as the hands, feet, or vocal cords. As the disease progresses, the weakness will spread resulting in muscle atrophy, slurred speech, difficulties swallowing and fatigue[44]. The high degree of variation means it is challenging to diagnose and anticipate care for individuals.

Limb onset is the predominant variant and comprises about 70% of cases, while the remaining 30% are associated with bulbar onset (muscles involved in speaking, swallowing and breathing) which have been associated with a more rapid deterioration[29]. It has also been identified that approximately 50% of patients will suffer from further neuropsychological manifestations including behavioural and cognitive changes. Furthermore, a proportion of these patients will go on to develop Frontotemporal Dementia (FTD)[43]. This has led to the concept in combination that ALS and FTD can be observed as divergent ends of a spectrum as they form a comprehensive continuum due to the overlaps in symptoms, pathophysiology, and genetic similarities.

Genetic architecture and mechanisms of disease

Substantial strides have been made in understanding the aetiology and pathogenesis of ALS, by investigating the genetic architecture of the disease. ALS was initially categorised as familial (typically 1st or 2nd degree relative with ALS and/or FTD) or sporadic. Familial ALS (fALS), comprises ~10-15% of cases, typically follows a Mendelian autosomal pattern of inheritance, while sporadic ALS (sALS) ~85-90% of cases, is considered complex aetiology[28]. This conventional subdivision has been blurred as ~5% sporadic cases can be found with a Mendelian penetrant variant and family

trees in fALS will show incomplete penetrance or skipped generations[25]. Therefore, it is increasingly recognised that ALS is a heterogenous disease both its clinical presentation and genetic architecture. Heritability estimates (or the genetic contribution to disease liability) based on population studies are about 40%[75]. While Genome-wide association studies (GWAS) has discovered 10 loci for the disease (with current case sample size, n = ~30k), the majority of research findings for disease mechanisms have focussed on Mendelian, fALS genes and variants (with some notable overlaps)[44]. This includes about 30 genes, with four accounting for most cases (based on founder effects). The most common is a hexanucleotide repeat expansion in chromosome 9 open reading frame 72 (*C9orf72*) (accounting for approximately 40% of fALS cases and 5% sALS cases)[68]. Point mutations in TAR DNA-binding protein (*TARDBP),* fused in sarcoma (*FUS*) and superoxide dismutase (*SOD1*) have been highlighted to have critical involvement in familial cases. Investigating the mechanisms of these genes has implicated several pathways in the pathogenesis of ALS. The pathways they impact include reactive oxidative stress (ROS), neuroinflammation, proteostasis failure, mitochondrial dysfunction, excitotoxicity, cytoskeletal disturbances and axonal transport defects, impaired DNA repair, oligodendrocyte dysfunction and nucleocytoplasmic transport deficits[5,41]. Indeed, some genetic cases, such as the identification of mutations in *TARDBP*, have provided crucial insight into the common pathogenic themes of ALS[69].

A critical RNA protein TDP-43 is mislocalised in ALS

In tissues, aggregation of cytoplasmic TAR DNA-binding protein 43 (TDP-43, encoded by *TARDBP*) has been found in ~97% of ALS cases (and ~45% in FTD cases), which has remained a central focus of disease pathogenesis since its discovery in 2006[8,28,45]. TDP-43 is a highly conserved and ubiquitously expressed RNA binding protein. It is typically localised to the nucleus in healthy neurones to maintain RNA transcription, splicing and mRNA stabilisation. Mislocalisation and aggregation of TDP-43 in the cytoplasm is a common characteristic found in both familial (i.e., pathogenic mutations in *TARDBP*) and sporadic (i.e., no mutations in *TARDBP*). It is proposed that this results from multiple mechanisms rather than a single process (Figure 1). It wasn't until 2008, that mutations in *TARDBP* were identified (particularly in the C-terminal region where the majority are found) that result in protein structure alterations leading to misfolding and gain of toxic function[55,61]. Less is understood about the changes in sALS[24]. Recently, TDP-43 has been demonstrated to have a key regulatory role as a splicing repressor to prevent cryptic exon (CE) inclusion, which are typically excluded from mature mRNAs[41]. Unlike normal exons, CEs are non-coding sequences that are a biproduct of aberrant RNA metabolism leading to the formation of

abnormal isoforms. Previous studies have suggested that when TDP-43 becomes depleted in nuclear cells, these CEs are spliced into mRNA, inducing premature stop codons, and leading to frameshift mutations. This frameshift results in protein truncation and can trigger a cellular process known as nonsense-mediated decay[8]. Additionally, this creates a cascading effect that proceed to cause other deleterious events, such as mitochondrial dysfunction and aberrant stress granule formation that drive disease progression. The truncated formation of TDP-43 aggregates has been consistently identified in post-mortem neurones from ALS patients, predominately in the frontal cortex of the brain, but also in the spinal cord (Figure 1).

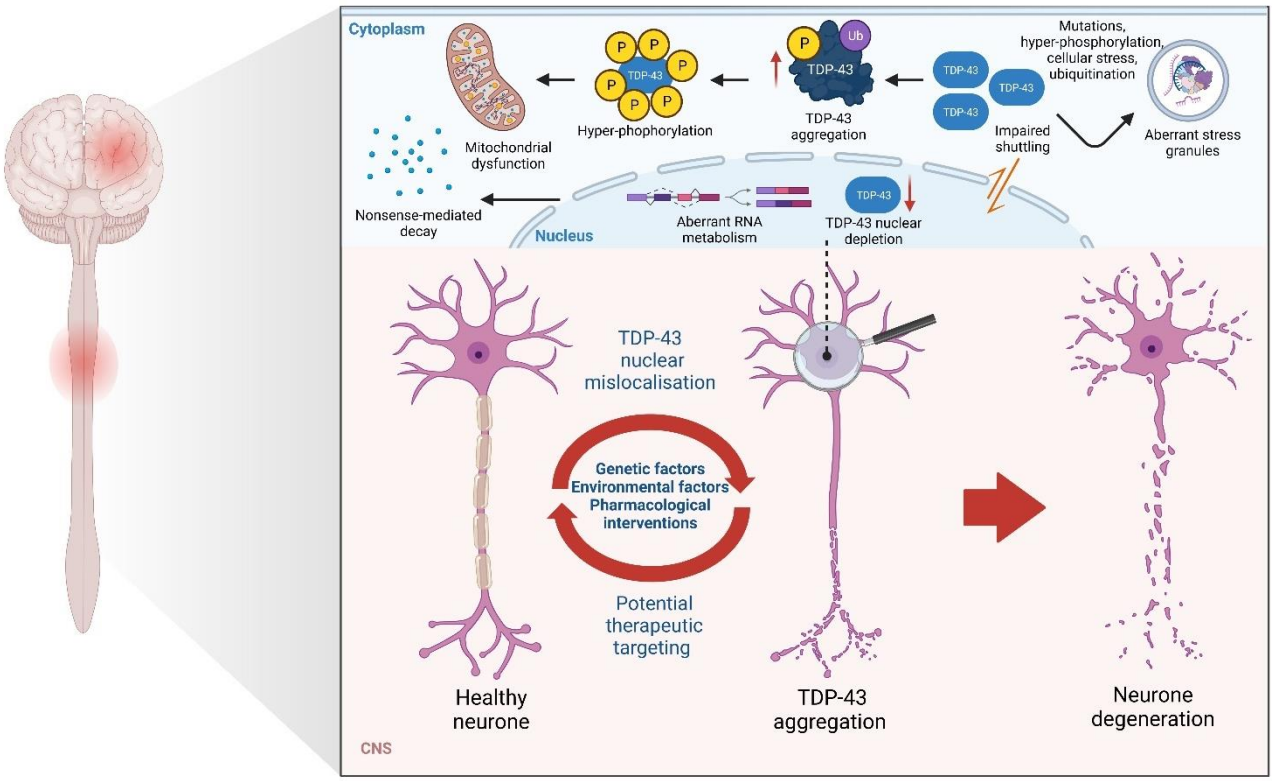


Figure 1. TDP-43 mislocalisation with nuclear depletion and cytoplasmic aggregation due to genetic and/or environmental factors, inducing motor neurone deterioration via multiple downstream pathways in the central nervous system of ALS patients. Although cytoplasmic TDP-43 aggregates have been identified within ~97% of patient neurones, it is not well understood how genetic and environmental factors contribute to TDP-43 mislocalisation. This has limited the development of therapeutics to treat TDP-43 proteinopathies that cause ALS due to its heterogeneous nature. However, early interventions normalising TDP-43 localisation may hold the potential to prevent neuronal degradation.

There are several features that are commonly observed in cytoplasmic inclusions of TDP-43 in ALS patients. These include the accumulation of post-translational modifications such as hyperphosphorylation, ubiquitination and oxidative stress that also contribute to the formation of aggregates in a "prion-like" propagation[64]. Although TDP-43 mislocalisation is a common histopathological hallmark of ALS, it is not unique to patients with only *TARDBP* mutations. This reinforces the complex genetic architecture involved in TDP-43 and the implications of several other

genes within the pathogenesis of the disease. This draws to the possibility that there may be other aberrant events that occur in other genes, due to *TARDBP* not performing its RNA splicing role in the nucleus that have yet to be encountered and may promote TDP-43 proteinopathies in the plight of ALS.

UNC13A is an ALS GWAS risk locus now relevant for therapeutic targeting

The UNC-13 homolog A (*UNC13A*) locus was first linked to ALS in genome-wide association study (GWAS) of ALS cases and controls, but no causal mechanism linking risk SNPs in region was known[18]. Two studies published in 2022 found *UNC13A* to have the most significant levels of alternative splicing in neurones with nuclear TDP-43 depletion[41]. Independent groups, Ma et al. (2022) and Brown et al. (2022) identified a novel CE regulated by TDP-43, specifically between exons 20 and 21 (Hg38; chr19: 17642414 – 17642541). This CE has since been recognised as key factor in understanding the pathophysiology of ALS mediated by *UNC13A*[73]. Both studies performed an RNA sequencing analysis in human neuronal cell lines and iPSC-derived motor neurones (iPSC MNs) with reduced TDP-43 levels. TDP-43 depletion caused inclusion of the CE in *UNC13A,* confirming a role in TDP-43 suppressing the CE[8]. Additionally, the *UNC13A* CE was found in samples exclusively from patients with TDP-43 pathology. The levels of *UNC13A* CE were also higher in those that carried the GWAS risk variant/s (rs12608932 (A>C) and rs12973192 (C>G)) and a third novel mutation (rs56041637), are located within the *UNC13A* CE locus [73]. *UNC13A* plays a role in synaptic regulation and neurotransmitter release via regulating the synaptic active sites[58]. It is therefore proposed that loss of *UNC13A* activity via TDP-43 nuclear depletion and CE in *UNC13A* can compromise neuronal function by interfering with vesicle maturation during exocytosis and neurotransmitter release (Figure 2)[43]. The results from these two studies showed a direct link between TDP-43 pathology and aberrant splicing in *UNC13A*, which could be therapeutically targeted in ALS. However, it is not yet clear the extent of the aberrant splicing, i.e., whether this mechanism is only impacted in neuronal cells – or whether this is also in other tissues and genes. Given TDP-43 is an RNA-binding protein relevant for regulating several processes, it is expected that other aberrant splicing impacts are expected to be discovered that may contribute to disease.

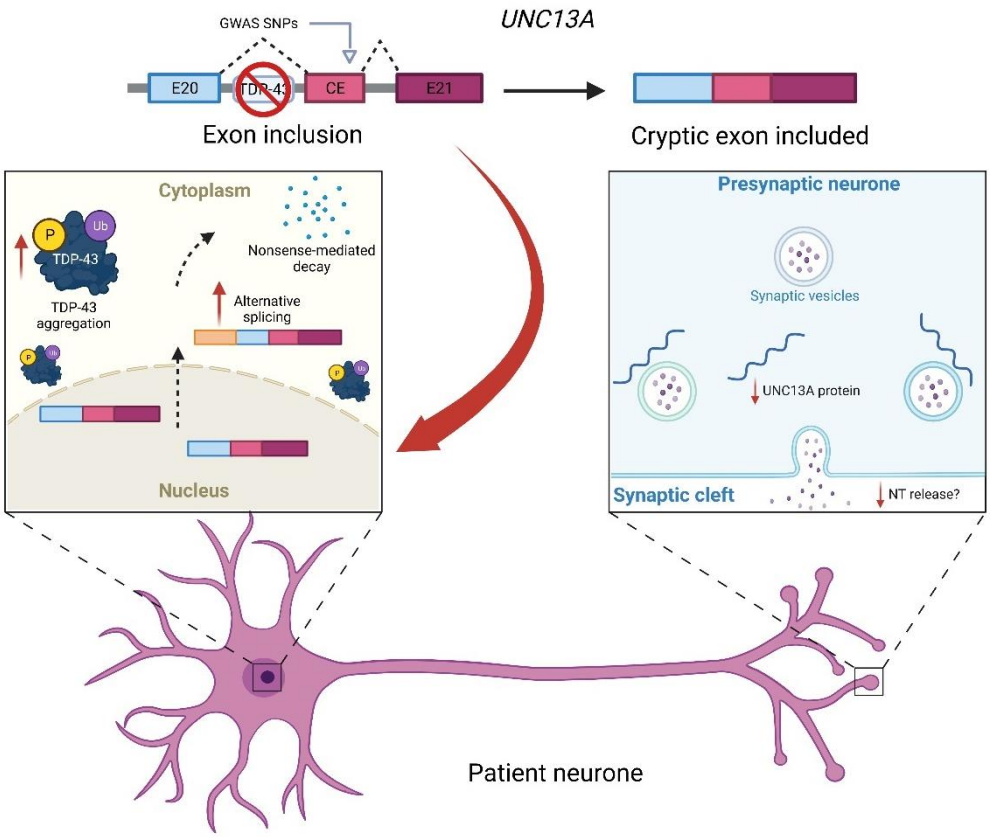


Figure 2. Pathological CE inclusion has been found between exon 20 and 21 as the basis for *UNC13A* associated ALS. As a result of TDP-43 mislocalisation into the cytoplasm, transcripts of *UNC13A* are subsequently mis-spliced leading to inclusion of CEs in mature transcripts, thus, leading to nonsense-mediated decay and downstream cellular dysfunction. Pathological SNPs identified via GWAS have continuously been linked to *UNC13A* dysfunction within the CE locus. Additionally, *UNC13A* has been found to play a role in synaptic transmission and neurotransmitter release. The disruption in *UNC13A* protein formation may result in changes in synaptic vesicle formation and neurotransmitter release.

A transcriptomic-driven approach to understanding ALS

Examining changes in gene expression and alternative splicing with RNA sequencing (RNA-seq) can help identify potential molecular mechanisms implicated in disease pathogenesis. Gene expression results can be integrated with genetic studies (i.e., summary Mendelian randomisation) to better understand causal effects[7]. While this is typically achieved with large sample sizes and healthy tissue, examining disease tissues, particularly over the course of disease may also help discovery and therapeutic design[31].

The transcriptome is of particular interest in ALS, as the hallmark of disease, TDP-43, has a primary role in facilitating RNA-binding and splicing homeostasis[41]. Disturbances of RNA metabolism has been consistently demonstrated within models of ALS, and more recently they have been associated with disease risk[19,74]. RNA-seq is a high-throughput technique that can detect *de novo* aberrant splicing events in Mendelian disorders, which was not possible with earlier transcriptome

approaches (i.e., microarray). RNA-seq pipelines typically involve quality control of sequenced reads, aligning them to a reference genome, and performing downstream analyses. One common downstream analysis is differential gene expression analysis, which identifies genes with varying expression levels across multiple conditions. Differentially expressed genes then undergo further statistical analyses to determine significance, and enrichment analysis is used to explain the function and implication of these significant genes, such as gene set enrichment and gene ontology (GO) terms related to biological processes, molecular functions, and cellular components.

Differential expression (DE) is a common analysis technique of RNA-seq but does not fully utilise the added value that RNA-seq provides – which are isoform analyses. Various splicing tools have been developed to fully utilise this data (accounting for increased testing) to help identify splicing events that may contribute to disease. A comprehensive study, looking for splicing events has not yet been conducted in a large ALS cohort despite two recent studies being recently being published. This included RNA-seq on induced pluripotent stem cells (such as the large Answer ALS iPSC[4]) (n_{cases} = ~1k) and a meta-analysis DE-transcriptome study[79], included the Answer ALS dataset with other iPSC MNs (n_{cases} = 323, $n_{controls}$ = 106), and central nervous tissue analyses (n_{cases} =214, $n_{controls}$ = 57). The lack of RNA splicing analyses, and use of other tissues (i.e., skeletal muscle, which is highly involved in disease progression) may highlight critical splicing events and differential gene expression that have yet to be documented in an ALS cohort. To our knowledge, no study outside of our group has investigated the effect of splicing in ALS RNA-seq using affected skeletal muscle. Importantly, the most comparable study to ours has primarily utilised mouse models. These are limited to certain gene mutations, despite most ALS cases being not having a genetic mutation identified (as described above) as may not fully replicate the complexity of the disease[62].

Approaches to aberrant alternative splicing detection and quantitation

Alternative splicing (AS) is a fundamental cellular process of post-transcriptional regulation of gene expression[48]. In this process, exons of the same gene are joined in different combinations, leading to mRNA transcripts that can encode for proteins that are diverse in terms of sequence and function. Approximately 15% of hereditary diseases and cancers are associated with AS[34]. The detection of disease relevant splicing is non-trivial as there have been thousands of annotated pathological splice sites over the past several years. Subsequently, the classification of different splicing events in genes is characterised by six main events (Figure 3).

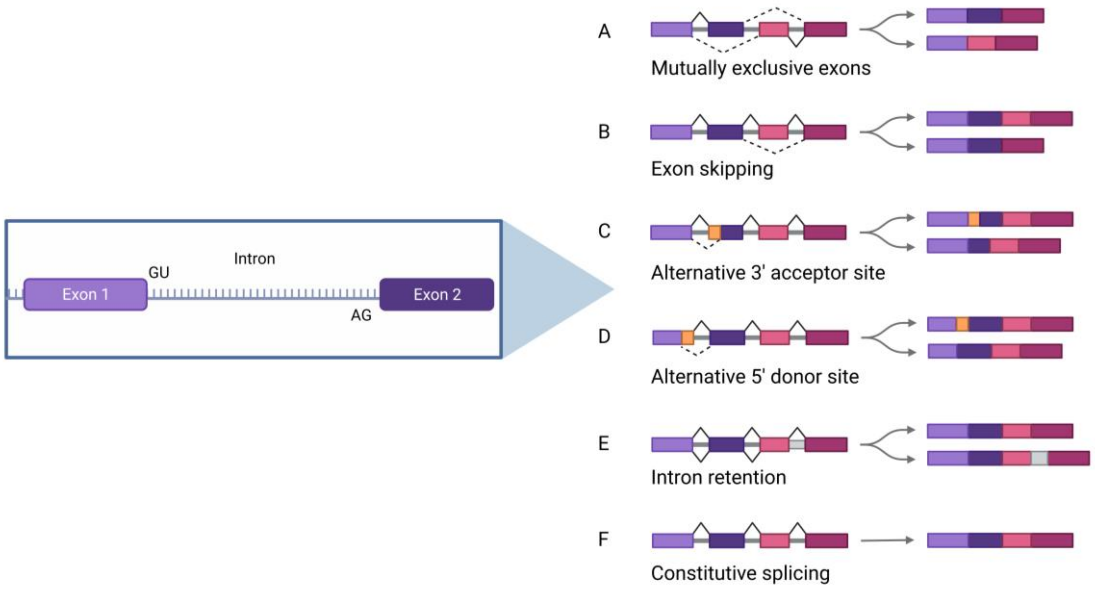


Figure 3. Illustration of the six main splicing events that contribute to the processing of mRNA, including: (A) mutually exclusive exons; (B) exon skipping; (C) alternative acceptor site, where the 3' end of the exon is alternatively spliced; (D) alternative donor site, where 5' end of the exon is alternatively spliced; (E) intron retention and (F) constitutive splicing. A-E are the classical types of AS, while F is the default process that occurs in most genes and is not considered abnormal. The enhanced zoom shows the boundary between exons with the upstream exon at the 5' end. It is the donor site since it "donates" or provides the end of the intron for splicing and is almost always "GU". Conversely, the alternative 3' splice is called the acceptor site because it "accepts" or receives the end of the intron during splicing. The sequence at this site is often "AG", and other nearby sequences contribute to the recognition process.

Genetic variation within splice site sequences frequently causes aberrant splicing in hereditary diseases. Single nucleotide substitutions affecting the 5' and 3' being amongst the most common cause of pathogenicity, which are decomposed into typically five different events[14]. Firstly, mutually exclusive exons (MXEs) are a rare subtype that are categorised by splicing of exons in a coordinated manner, such that two or more events are not independent[53]. The "mutually exclusive" description also dictates that one out of the two events (or one exon out of the two exon groups) are retained, while the other one is spliced out. A common assumption is that the formation of MXEs originate from exon duplication, and hence are highly similar[53]. Subsequently, mutations or genetic variation in MXEs can potentially lead to pathogenic consequences. Exon skipping (ES) on the other hand, can result in a loss-of-function of mutation that is known to contribute to pathogenesis, such as exon skipping seen in *FUS* (exon 7) in ALS and the *DMD* gene loci – commonly seen in Duchenne muscular dystrophy[1,77]. This category is generally the most common type of pathological splicing, accounting for approximately 40% of all AS events in human tissues; while alternative 5', 3' and intron retention (IR) account for about 15-20%, respectively[20,72]. Both 5' (donor, at the start of the exon and denoted as 5ASS) and 3' (acceptor, end of the exon and denoted as 3ASS) AS result in the inclusion or exclusion of different portions of the exon in mature mRNA transcripts. These alterations can occur in both constitutive and alternative exons or result in changes in the ratios between spliced

isoforms. As a result, the quantitation and detection of AS events is intrinsically more complicated than gene expression.

Assessing alternative splicing with the Percent Spliced In metric

Advances in studies of AS have been aligned with improvements in RNA-seq protocols and bioinformatic methods. Historically, the main detection method consisted of performing RT-PCR on RNA extracted from affected tissues, which was often laborious and time consuming. To detect AS and quantify it as a metric involved in rare diseases, the most common modern computational tools are based on different criteria, including event, transcript, variant or graph-based statistical models. One of the most popular methods is to quantify AS is based on the percent spliced in (PSI, Ψ), which captures local information pertaining to the splicing of each exon[46]. The PSI metric is defined as the number of reads supporting each exon inclusion divide by the total number of reads that span the exon. This is denoted as:

$$\mathrm{PSI} = \frac{adjusted\ spliced\ in\ reads}{(adjusted\ spliced\ in\ reads + adjusted\ spliced\ out\ reads)}$$

Consequently, it is also possible to measure the PSI between two groups/samples, denoted by ΔPSI. For example, in control (set as the reference) verses case grouped cohorts, where:

$$\Delta\mathrm{PSI} = PSI(control) - PSI(case)$$

A negative ΔPSI is indicative of the short-form isoform being favoured in cases, while a positive ΔPSI is suggestive of long-form isoform being favoured in controls. This metric dictates whether the splicing is favoured in one group from another and can pinpoint whether the splice is implicated in pathogenesis by further enrichment analysis.

The drop in sequencing costs and increased data generation has resulted in a rapid increase in bioinformatic tools to measure AS and detect CEs. Currently, some of the most popular and well-documented tools to detect differential AS in RNA include rMATS[58] and LeafCutterMD[40] which use an event-orientated statistical models based on the ΔPSI. FRASER[40] and SpliceAI[32] are also gaining popularity due to their ability to detect rare events for rare disease (FRASER) and causal variants (SpliceAI). These tools have different strengths and weaknesses (Table 1) and thus can be used in parallel to perform complementary analyses (each is described in detail below).

Table 1. Summary of bioinformatic tools used in this project for aberrant alternative splicing analyses

Tool	Methodology	Event level	Reference/s	Strengths	Limitations
FRASER	➢ Focuses on finding rare aberrant splicing events, accounting for biological variability and technical noise	Local	46,56	➢ Uses read counts to estimate splice junction usage to detect deviations from the expect distribution. ➢ Consists of 2 model modes – Intron Jaccard Index and PSI metrics ➢ Detect rare and potentially pathogenic splicing events	➢ Computationally intensive ➢ Relies heavily on the reference set, which may influence results ➢ Requires large datasets to reliably detect rare splicing events
rMATS	➢ Models off the basis of replicate data, using multivariate uniform prior and Expectation-Maximisation	Group (paired)	58	➢ Effectively models differential alternative splicing between 2 groups (i.e., case-control studies) ➢ Can detect and differentiate between the 5 major types of alternative splicing events ➢ Doesn't assume equal sample sizes for statistical power	➢ Requires pre-defined annotations ➢ May not be as effective for detecting rare splicing events ➢ Replicates needed for accurate results
LeafCutterMD	➢ PSI is determined using a Dirichlet-Multinomial distribution ➢ Relies on RNA-seq split reads	Group	33,40	➢ Does not rely on annotations to quantify splicing events ➢ Handles both short and long reads ➢ Map differential splicing between different groups	➢ Does not predict the impact of splicing events ➢ Miss small exons due to reliance of intron clustering ➢ Heavily influenced by read depth ➢ Statistical significance is not assessed
SpliceAI	➢ Pre-trained deep convolutional neural network that models the relationship between sequences and splicing outcomes	Variant prediction	32	➢ Uses deep learning to predict the effects of genetic variants on splicing ➢ Detects both intronic and exonic variants that affect splicing ➢ Predicts the impact of novel mutations on splicing, including intron regions	➢ Does not quantify differences in splicing, rather predicts the impact of genetic variants on splicing ➢ Overestimates the impact of certain mutations due to reliance on machine learning modelling ➢ Predictions are based on training data and may not always generalise well to novel variants

FRASER

Find Rare Aberrant Splicing in RNA-seq (FRASER) is a R/Bioconductor tool that consists of three steps: (1) counting; (2) modelling; and (3) detection (Table 1). The current statistical framework (version 2.0) builds upon the initial (version 1.0) metrics described by Mertes et al. (2022) (Table 2).

Table 2. Summary of initial AS metrics from FRASER 1.0

Splicing metric	Description
$\boldsymbol{\Psi_5}$	PSI at the 5'-splice site (donor)
$\boldsymbol{\Psi_3}$	PSI at the 3'-splice site (acceptor)
θ	Splicing efficiency that measures percent of introns retained

FRASER 1.0 resulted in the number of outlier calls per sample being excessively large which led to outliers that reflect local aberrations to only have minor effects. To address this issue, FRASER 2.0 (2023) builds on the model to integrate an intron-based metric the Jaccard Index (J index). The J index is defined as the proportion of reads supporting the splicing of an intron of interest amongst all reads associated with either splice site of the intron[56]. It combines the three previously described variables in the original model, Ψ_5, $\Psi_3$and θ into one metric. It is conceptually similar to the intron excision ratio that LeafCutterMD uses. However, it differs in statistical modelling, since LeafCutterMD uses a clustering approach between multiple introns at the same locus, which leads to modelling complications and does not account for sample co-variation[56]. The authors of FRASER also benchmarked their latest version against LeafCutterMD, which showed higher enrichment of intron-centric splicing outliers[56]. This new approach was designed to be less sensitive to local splice aberrations that do not have a strong effect on canonical splice isoforms. This is to capture several types of aberrant AS, including partial or full intro retention. Splicing quantification using the J index is represented mathematically as:

$$\mathrm{J}_{i,j} = \frac{|\, D_{i,j} \cap A_{i,j} \,|}{|\, D_{i,j} \cup A_{i,j} \,|}$$

Where $\mathrm{J}_{i,j}$ is the J index of the set of donor-associated reads $D_{i,j}$ and the set of acceptor-associated reads $A_{i,j}$ for a given set sample *i* and intron *j* [56]. FRASER controls for confounders by optimising simulated outliers through training a denoising autoencoder, consisting of an encoder and decoder matrix. To determine the statistical significance of outliers, it uses a beta-binomial distribution to calculate and adjust for *p*-values. Subsequently, the new iteration of FRASER stands out as a novel tool that could assist in the identification of rare splicing events that have yet to be identified in ALS.

rMATS

Replicate-based Multivariate Analysis of Transcript Splicing (rMATS) is a statistical command line tool used for robust and flexible detection of differential AS from replicate RNA-seq, such as a case-control samples at a group level (Table 1)[58]. Unlike other splicing tools which are often intron-oriented, rMATS reports differential splicing of canonical events, including ES, MXE, A5SS, A3SS and IR between two sample groups by performing a pairing analysis (i.e., control verse case). Shen et al. (2014) described that it uses a hierarchical model to simultaneously account for estimation uncertainty in individual replicates and variability across exon-exon junctions. In paired analysis mode, the model assumes unequal sample sizes, by considering the number of matched samples between the two groups, rather than group sizes. The model utilises a multivariate uniform prior to represent the between-replicate variability. Subsequently, it uses Expectation-Maximisation (EM) to estimate the model parameters between the two group conditions[52]. The variability within sample groups is then modelled as random effects in a linear mixed model and then mean exon inclusion levels are estimated as fixed effects[58]. This tool offers a significant advantage compared to other techniques (as previously benchmarked against MISO and SUPPA2) by incorporating the variability among replicates caused by both biological and technical factors[48].

LeafCutterMD

Another well-curated command line tool called LeafCutterMD (Table 1) is typically used to detect aberrant splicing events at the group level in rare Mendelian diseases. It uses a Dirichlet-Multinomial generalised linear model to test for differential splicing across two individual conditions or groups[40]. The PSI is determined by using an intron excision analysis that clusters introns that share at least one splice site and are a part of the same event. LeafCutterMD does indeed have shortcomings when there is a limited amount of sequencing depth, which often may hinder the ability to detect novel events[46]. However, a new innovative prediction-based tool to measure the impact of aberrant splicing in SNVs has recently been developed, which has not been previously described in other conventional tools.

SpliceAI

Variants impacting vital splice sites are almost guaranteed to cause AS. However, a recent command line tool demonstrates that by knowing what AS events most commonly happen around a particular splice site, we can predict what events will happen when a genetic variant occurs. SpliceAI is a deep learning model that predicts the location of splice sites, using up to 10,000 nucleotides of mRNA sequence context for each prediction (Table 1)[15]. The algorithm is trained on pre-mRNA sequences

surrounding 130,796 annotated donor-acceptor pairs using a 32-layer deep neural network (through using TensorFlow (Python deep learning library) backend as a dependency for neural network construction)[32]. The model can make predictions about global splicing outcomes, such as exon skipping and CE splice sites. Predictions and relative positions of the altered splice sites are displayed as numeric values, which are denoted as delta scores (Δ-scores) between the raw score of the reference to variant allele[16]. The Δ-score provided by SpliceAI accounts for the maximal differences between the predictions of the variant and the reference allele using the four predictor variables, as represented in Table 3.

Table 3. Summary of SpliceAI delta score metrics

Delta (Δ) scores	Prediction outcome
acceptor gain (AG)	Closest AG (3') acceptor site **upstream** of the variant
acceptor loss (AL)	Closest AG (3') acceptor site **downstream** of the variant
donor gain (DG)	Closest GU (5') acceptor site **upstream** of the variant
donor loss (DL)	Closest GU (5') acceptor site **downstream** of the variant

The higher the Δ-score stipulates a greater difference in splicing probability between the variant and reference sequence, suggesting a greater splicing strength to create a splice site[32]. The authors of SpliceAI describe three optimal Δ-score thresholds, that include Δ-score ≥ 0.2, 0.5 and 0.8, where 0.2 and 0.5 are relatively permissive thresholds. It has been characterised that to retain splice altering variants with high precision, a Δ-score ≥ 0.5 should be used[32]. The authors of SpliceAI demonstrated that the higher threshold accounts for greater likelihood of variants that cause CE inclusion and ensures that the predicted variants are deleterious in the human population[32]. However, factors such as frequency in the population and function consequence of the variant need to also be considered in determining risk for disease.

Bioinformatic tools are evolving with advances in sequencing technology. To date, all tools rely heavily on short-read sequencing which do not typically span a full isoform. Hence, deep sequencing and long-read sequencing will aid in improving accuracy. At this stage, we are limited to using each tool's "threshold" to focus on relevant detectable events. Nevertheless, the need to develop effective RNA-seq pipelines to accurately identify novel splice sites affected by TDP-43 mislocalisation may provide further insight for therapeutic targeting.

Research aims and hypothesis

ALS is a complex disease with aetiology that is not well understood. It is a neurodegenerative disease but multiple systems in the body are impacted. Unclear is how changes in the nervous system or skeletal muscle might be reflected in the blood. Here in this project, I will have two sample matched sources of RNA from ALS cases and controls. Blood, an easily accessible tissue that can be used in large samples, and skeletal muscle, which is a largely overlooked tissue and is heavily affected in disease. Using a preliminary cohort of cases and controls with RNA-seq, I aim to build a thorough transcriptomic analysis pipeline to examine disease processes that are both known and novel. Specifically, this investigation will aim to:

1. Analyse differentially expressed genes and potential pathogenic splicing in ALS patients from blood and muscle collections compared to controls.
2. Look for evidence of AS within blood and muscle tissue of ALS patients (including the CE inclusion in *UNC13A*).
3. Identify novel splice sites and variants that can be prioritised for follow-up and potential therapeutic targeting.

The design of the study, consisting of two sources of RNA, will provide insight into whether any detected changes are unique to the tissue source and disease. I hypothesise that I will detect aberrant splice events that could play a role in disease. While I will be unable to determine whether changes are due to disease cause or a downstream effect of the disease, prioritised findings can be further examined in a larger sample size or longitudinal design. Identifying genes that have significant aberrant splice events may highlight overlooked loci that could provide a potential avenue for therapeutic targeting.

Methods

Participant enrolment

Study recruitment was carried out prior to the start of this project. ALS patients who met the revised El Escorial criteria at the time of diagnosis were enrolled at the Royal Brisbane and Woman's Hospital MND clinic, along with healthy controls (typically partners and/or carers). All participants were subjected to the same exclusion criteria which included the existence of a metabolic condition, such as Hashimoto's disease and diabetes mellitus (exclusion criteria unrelated to the aim of this study). Past medical history and ALS diagnosis criteria for the cohort (prior to this project) is described in Supplementary Table 1 and 2. Patient demographics and study participant statistical analysis were explored by Freydenzon et al. (2023).

Sample collection, handling, and RNA library preparation

Sample collection took place from mid-2017 to early 2018. RNA-seq was performed on both patients with ALS and healthy controls (predominately relatives) for both muscle and blood samples. Muscle punch biopsies of approximately 200mg were collected from the vastus medialis of one leg (the least affected leg for ALS study participants) using a modified hollow Bergstrom biopsy needle with local anaesthetic. Muscle samples were placed in Dulbecco's Modified EaglMedium/Ham's F-12 with 0.5% gentamicin for transport to the processing laboratory at the OTTO building, University of Queensland. The Human Studies Unit (HSU) at the Institute for Molecular Bioscience performed the initial steps in sample processing, RNA preparation and sequencing. A proportion (~50-100mg) of each sample was separated from the primary collection, washed three times with phosphate buffered saline, frozen on dry ice and stored in laboratory freezers at -80^{o}C. Total RNA was extracted from muscle samples using a standard Trizol-chloroform extraction method and then quantified using the Epoch microplate spectrophotometer. Quality assessment was performed using the Agilent Bioanalyzer and the RNA Nano kit. An average yield of 7.4mg was obtained with an average RNA Integrity Number (RIN) of 8.2. Additionally, library preparation for the muscle samples consisted of using 1mg of DNAse1 treated total RNA and Illumina's TruSeq RNA library preparation gold kit concurrently with TruSeq RNA UD indices prior to sequencing.

Peripheral blood samples were collected from a single draw of 2.5ml whole blood in PAXgene Blood RNA tubes. Samples were stored at room temperature (RT) for 24 hours before transferring to -20^{o}C freezers. Following RT thaw, Total RNA was extracted using the PAXgene Blood RNA kit (Qiagen) as per manufacturer's instructions. Post extraction, the total RNA was quantitated using the Nanodrop

spectrophotometer followed by quality assessment on the Agilent Bioanalyzer (RNA Nano kit). On average the samples yielded 4.8 mg of total RNA with an average RIN of 8.2. To prepare the blood library, 500ng of DNAse1 was used to treat the total RNA for sequencing. Subsequently, both libraries in each collection group were pooled and sequenced as paired-end in-house (Illumina HiSeq 4000).

RNA sequencing, data pre-processing and quality control

RNA-seq was performed across case and controls for both skeletal muscle ($n_{cases} = 30$, $n_{controls} = 20$) and blood ($n_{cases} = 30$, $n_{controls} = 19$) samples consisting of three sequencing batches (one in blood, two additional runs in muscle samples). Initial quality control (QC) of short paired-end reads obtained in FASTQ format that underwent trimming with Fastp, using default settings to filter abnormal reads prior to alignment. Fastp (v0.22.0) was used due to automatically trimming adapter sequences by overlapping read pairs, in addition to N-based quality filtering and polyG tail trimming. Trimmed reads of 75bp were mapped with STAR RNA-aligner (v2.7.10b) to the human annotation indexed reference, GENECODE v43 (Hg38). The input parameters of STAR specified that splices were only recognised if they fall over indexed of known splice junctions, which set a maximum number of mismatches to be read length of 0.04. This limited genomic distance between read mates to 1 million and set a minimum intron length of 20bp. Samples with alignment counts of <20 million across technical replicates were filtered out of the analysis (n = 4 blood and n = 1 muscle). Each sample underwent rigorous QC with FastQC (v0.12.1) and subsequently collated with MultiQC (v1.0) to observe quality scores. Next, the average percentage of reads mapped was also calculated using SAMtools (v1.9) across tissue collections. The following output consisted of approximately 8 technical replicates per sample that passed QC, which were then quantified using FeatureCounts (v2.12.3) (from RSubread) at the gene level. After differential expression analysis, replicates were merged using SAMtools for downstream splicing analyses to increase sequencing depth for samples.

Differential gene expression analysis

Differential gene expression analysis was performed using DESeq2 (v1.30.6) with the STAR-generated read counts. Technical replicates were summed using DESeq2's *CollapseReplicates* function to ensure counts, dispersion factors and estimated size factors were collated as a single measure. Low read counts were filtered out across samples to enhance relative power for downstream analyses based on genes with ≥ 3 samples with normalised counts ≤ 10. Variation between samples were observed by performing principal component analysis (PCA) with a variance

stabilisation transformation (vst) on both collection types. Pre and post observations were conducted to visualise any batch effects in the collection types with the use of the Limma (v3.52.4) function, *removeBatchEffects.* A log2 transformation and normalisation was also performed on the data to observe the distribution of expression across samples. DESeq2 uses raw counts as input and normalises data via the median of ratios, accounting for sequencing depth and RNA composition. Muscle and blood generalised linear models (GLMs) were fitted to the data while accounting for batch (encoded as integer variables) and sex covariates, with controls set at the reference. Age was not included as a covariate to include more genes for analysis, thus we opted for a less stringent approach (since age (which had no statistical difference between case-controls) was used as a covariate factor in GLMs from a previously carried out investigation by Freydenzon et al. 2023). Differentially expressed genes were annotated as a HGNC symbol using the biomaRt (v2.52.0) package via their corresponding Ensembl IDs.

Aberrant splicing analyses

For splicing analyses, we took a subset to decrease processing duration due to the long computational runtimes of the statistical models (as it was observed that some samples impeded the splice counting process, particularly when using FRASER) for skeletal muscle ($n_{cases} = 27$, $n_{controls} = 19$) and blood ($n_{cases} = 27$, $n_{controls} = 17$).

Detection of rare (local) aberrant splice sites

Both iterations of FRASER 1.0 (v1.8.1) and 2.0 (v1.99.0) were used to detect rare aberrant splice sites in skeletal muscle and blood samples (FRASER 2.0 was released May 2023). For this report we limit results and reporting to FRASER 2.0 (hereafter FRASER 1.0) given its refined method (utilising RSubread's FeatureCounts for counts of splice junctions/sites and in-built functions, *CountRNAdata* and *CalculatePSIvalues*). A low intron and junction expression filter was performed on the RNA-seq data using recommended $|\Delta PSI| > 0.1$, with minimum expression of 20 reads per site. Intron-centred PSI and Jaccard correlation heatmaps were developed for all splicing metrics based on sex, ALS status and age to observe covariate interactions before (and after) running FRASER. The fitted latent space (*ls*) for autoencoders were estimated using FRASER's hyperparameter tuning tool (*optimHyperParams*). It works by artificially injecting outliers into the data and then comparing the AUC-PR (area under the precision recall curve) of recalling these outliers for different values of *ls*. Hyperparameter tuning of tissue specificity autoencoders were estimated, *ls = 2* nodes for J index (and *ls = 5* nodes for Ψ_5, Ψ_3 and θ) for both blood and muscle collection types. PCA was utilised to estimate the encoder and decoder matrix outliers and then using a beta-binomial loss function to

optimise the weights of the matrices. Since FRASER offers other implementations for analysis, just the recommended PCA approach was carried (faster processing speed over other methods). For all splice sites and junctions, the Hg38 UCSC annotation was used (consistent with alignment). Aberrant splice detection thresholds (|ΔJ index (and ΔPSI) | > 0.3 and adjusted $p < 0.05$) were used in conjunction with multiple-testing correction across all samples (Benjamini-Yekutieli FDR method).

Differential (grouped) splicing feature analysis

For differential splicing feature analysis, rMATS (v4.1.2) was used to detect the type of AS events to compliment the results from FRASER. The following specification flags used on the command line included: the text files containing directory locations to case and control BAMs for each collection type (with controls set as the reference; --b1 control.txt and --b2 case.txt), as paired-end (-t paired), read length of 75bp (--readLength 75). Furthermore, the annotation reference – Hg38 with a default statistical splicing threshold (--cstat 0.0001) was used. Additionally, the novel splice site (--novelSS) flag was also utilised to reveal potential sites that may account for differential splicing. Next, we used a statistical toolkit for AS visualisation called Jutils (v1.0). Jutils collapses the 14 outputs from rMATS into a comprehensive summary file based on rMATS' differential splicing metrics of junction-exon counts (JCEC). Next, we also used ggSashimi (v1.1.5) and Maser (v3.17) to visualise the alternatively spliced transcripts under the Hg38 reference. Statistical significance of differential splicing events was defined by adjusted *p*-value < 0.05 and |ΔPSI| > 0.1, as recommended by rMATS authors. Inspection of differentially spliced transcripts was also checked using the Ensembl database.

Prediction of pathogenic splicing variants

To detect splice variants, variant calling was performed followed by the implementation of SpliceAI. Both blood and skeletal muscle samples were run independently. RNA-sequencing was first processed by Bcftools (v1.9) for diploid variant calling on indexed binary alignment map files (BAMs). Then variants were filtered using standard thresholds (Phred scores > 20 (>99% accuracy for base calling) between reference and alternative alleles) and subsequently run in SpliceAI. SpliceAI (v1.3) was run using standard recommended parameters, but with a high precision threshold cut off (Δ-score ≥ 0.8). This localised genes that had a high probability of splicing. These were subsequently annotated using SpliceAI Lookup[16], which uses reference databases such as Genotype-Tissue Expression (GTEx)[13], Single Nucleotide Polymorphism database (dbSNP)[59], Genome Aggregation Datbase (gnomAD)[26] and ClinVar[39] to support prioritisation (pathogenicity and variant frequency). Hg38 was used for variant annotation. A second analysis was executed, using the chromosomal

locus of *UNC13A* (Hg38; chr19:17601336-17688365) to detect relevant variants within case and controls. However, this analysis differed in prediction parameters, utilising no splicing probability threshold to maximise the number of hits.

Functional enrichment analysis

Genes differentially expressed (Benjamini-Hochberg adjusted *p*-value <0.05) or aberrantly spliced in cases (any tool, Ensembl database) were taken forward for further pathway enrichment analysis. A gene set enrichment analysis consisting of the top 200 most statistically significant genes based on log2 fold change were utilised to identify genes that were over-represented in pathways between tissues using ClusterProfiler (v1.62). Enrichment analyses were performed separately on both tissues individually and then pooled to visualise the top 20 in a single heatmap.

Computing environment and statistical analysis

All sequencing analyses and tools were run on one of two in-house high-performance computer clusters (Tinaroo and Bunya). Statistical analyses and data visualisations were performed in a local environment in R (v4.2.1) with the specific test described in each result (figure legend, main text, methods and supplementary). To minimise false positives given the small sample size, the highest recommended thresholds for each published tool was used. A Benjamini-Hochberg's adjusted *p*-value method was used to adjust for multiple comparison, with an adjusted *p*-value < 0.05 was used as a cut-off to call significance except when stated otherwise.

Results

Sequencing quality and read alignment

RNA-seq reads for each sample using Phred scores indicated that the read quality was high, with a score of >30 (99.9% base call accuracy) across all batches in both collection types). Mean genome alignment over all skeletal muscle and blood samples through STAR was 87% and 53% respectively (rounded to nearest whole number). PCA based on sex and ALS status in blood revealed 44% of the variance was comprised within PC1, while 19% was explained by PC2. It was evident that there was little separation based on gender effects and ALS status. Conversely, for muscle samples PC1 explained 30% and PC2 showed 19% of the total variance in muscle between case and controls (Figure 4). It was evident that there was a more definitive separation between PC1 and PC2 in muscle samples based on gender, with little separation between cases and controls based on ALS status.

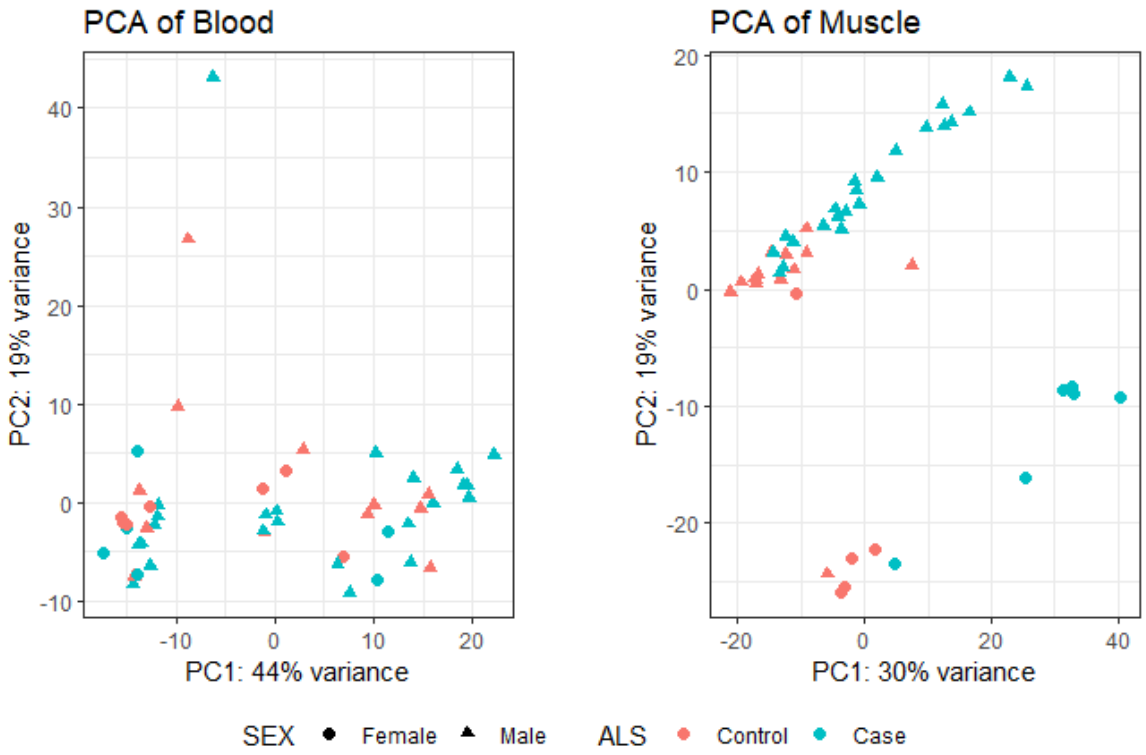


Figure 4. Principal component analysis of blood and muscle samples using a variance stabilisation transformation (vst), with samples coloured by ALS status and shaped by sex. In blood, the largest variation was explained by PC1 which made up 44% of the total variance, while PC2 made up 19% of the total variance. Conversely, it was evident that PC1 in muscle explained 30% of the total variance, while PC2 was comprised of 19%. Although there is no clear separation between cases and controls, it was evident that there was a clear gender effect in muscle samples, while there was little variation explained between gender in blood.

Differential gene expression was detected in ALS and unique to tissue type

Differential gene expression was carried out independently in blood and muscle to compare cases and controls accounting for batch and sex effects. In blood, 19 upregulated and 10 downregulated protein coding genes were identified (log2 fold change (log_2FC) = -1/+1 and adjusted *p*-value (q) < 0.05, 8510 protein coding genes detected). The topmost significant differentially expressed genes consisted of *FGFBP2* ($log_2FC = 1.73$; $q = 1.9 \times 10^{-6}$), *CDKN1A* ($log_2FC = 1.48$; $q = 1.9 \times 10^{-6}$), *COL19A1* ($log_2FC = 2.23$; $q = 7.6 \times 10^{-6}$) and *STYXL2* ($log_2FC = -2.56$; $q = 4 \times 10^{-3}$) (Figure 5A).

In skeletal muscle, 25 upregulated and 4 downregulated protein coding genes (out of 8088) were identified. The topmost differentially expressed genes – *SAA1* ($log_2FC = 3.37$; $q = 2.1 \times 10^{-11}$), *SAA2* ($log_2FC = 3.31$; $q = 9.7 \times 10^{-8}$), *CERKL* ($log_2FC = 1.69$; $q = 2.7 \times 10^{-5}$) and *CLEC12A* ($log_2FC = -1.48$; $q = 6.8 \times 10^{-4}$) (Figure 5B). *SAA* genes were also further explored given their highest level of log fold change, to compare differences between sex and ALS status based on pairwise paired t-test of log2 normalised expression. The rationale was that these genes were chosen based on prior implication by Freydenzon et al. (2023), which did not check for statistical changes based on gender and status of expression levels. Here, we observed a statistically significant difference that was evident in *SAA1* and *SAA2* ($p < 0.05$) based on ALS status, however, there was no supporting evidence of gender effect in expression across cases ($p > 0.38$) (Supplementary Figure 2).

Next, hierarchical clustering was used to visualise the top 50 differentially expressed genes in collection types based on z-score and to observe the interaction between covariates via a combined GLM (Supplementary Figure 1). In blood samples, it was identified that *CLEC12A* exhibited a high z-score within ALS cases of blood samples (z-score > 2.5) (this was downregulated in muscle samples based on log_2FC). Other genes in blood that exhibited high z-scores > 2.5 were *RUNX1, CDC42EP3* and *MYCL*. While in muscle, genes with the highest z-score were in the *SAA* family (*SAA1* and *SAA2*) (z-score > 3.5), with *MYH3, MSC and MYL5* also contributing significantly high z-scores (>3).

An individual and grouped genset and gene ontology enrichment analyses consisting of the top 200 most statistically significant genes (all genes within the top 200 had an $q \leq 0.02$) were taken forward as input from blood and skeletal muscle collections. Genes were ranked based on statistical significance and log2 fold change. ALS skeletal muscle genes were identified to be enriched in pathways such as extracellular organelle machinery pathways ($q < 0.001$) (Figure 5C). Conversely, in ALS blood, genes were enriched in tissue regeneration and response to cytokine release ($q < 0.05$) (Figure 5D).

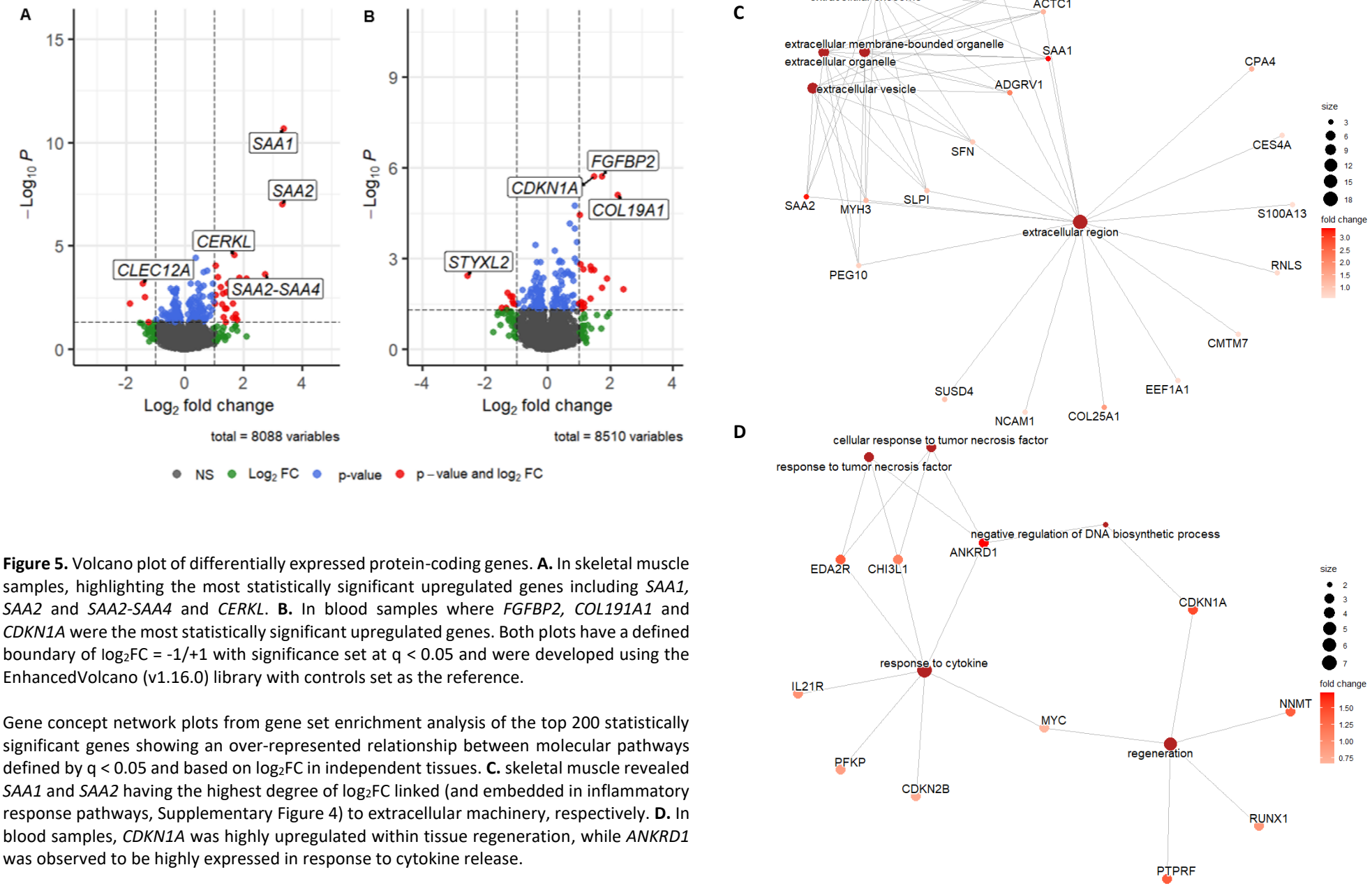


Figure 5. Volcano plot of differentially expressed protein-coding genes. **A.** In skeletal muscle samples, highlighting the most statistically significant upregulated genes including *SAA1, SAA2* and *SAA2-SAA4* and *CERKL*. **B.** In blood samples where *FGFBP2, COL191A1* and *CDKN1A* were the most statistically significant upregulated genes. Both plots have a defined boundary of log_2FC = -1/+1 with significance set at $q < 0.05$ and were developed using the EnhancedVolcano (v1.16.0) library with controls set as the reference.

Gene concept network plots from gene set enrichment analysis of the top 200 statistically significant genes showing an over-represented relationship between molecular pathways defined by $q < 0.05$ and based on log_2FC in independent tissues. **C.** skeletal muscle revealed *SAA1* and *SAA2* having the highest degree of log_2FC linked (and embedded in inflammatory response pathways, Supplementary Figure 4) to extracellular machinery, respectively. **D.** In blood samples, *CDKN1A* was highly upregulated within tissue regeneration, while *ANKRD1* was observed to be highly expressed in response to cytokine release.

We further visualised the top 20 differentially expressed genes using heatmaps in each tissue individually and at a combined level, that passed q < 0.05 threshold (Supplementary Figure 3 and 4 for individual heatmaps). In blood, the most intriguing results we identified was in *ANKRD1* to be enriched ($log_2FC > 1$) in response to cytokine release and cellular death pathways. As seen in the gene concept model, *CDK1NA* was also observed to be enriched in apoptotic signalling and tissue regenerative pathways. Conversely, in muscle, most statistically significant genes were differentially expressed in extracellular machinery (vesicles, organelles, and extracellular space). Notably, *ACTC1* was also enriched in actomyosin structure organisation ($log_2FC > 1$). At the combined level, embedded within these pathways that were significantly upregulated based on log2FC (Figure 6). *SAA1* and *SAA2* had the highest positive ($log_2FC > 3$) involvement particularly within extracellular machinery and inflammatory pathways (q < 0.001). Interestingly, *ADGRV1*, *CERKL, CHRNG, IPTKA and PTPRF* showed a positive log fold change (log2FC > 1), suggesting these genes have upregulated expression within neurone projection pathways (q < 0.05) in ALS patients.

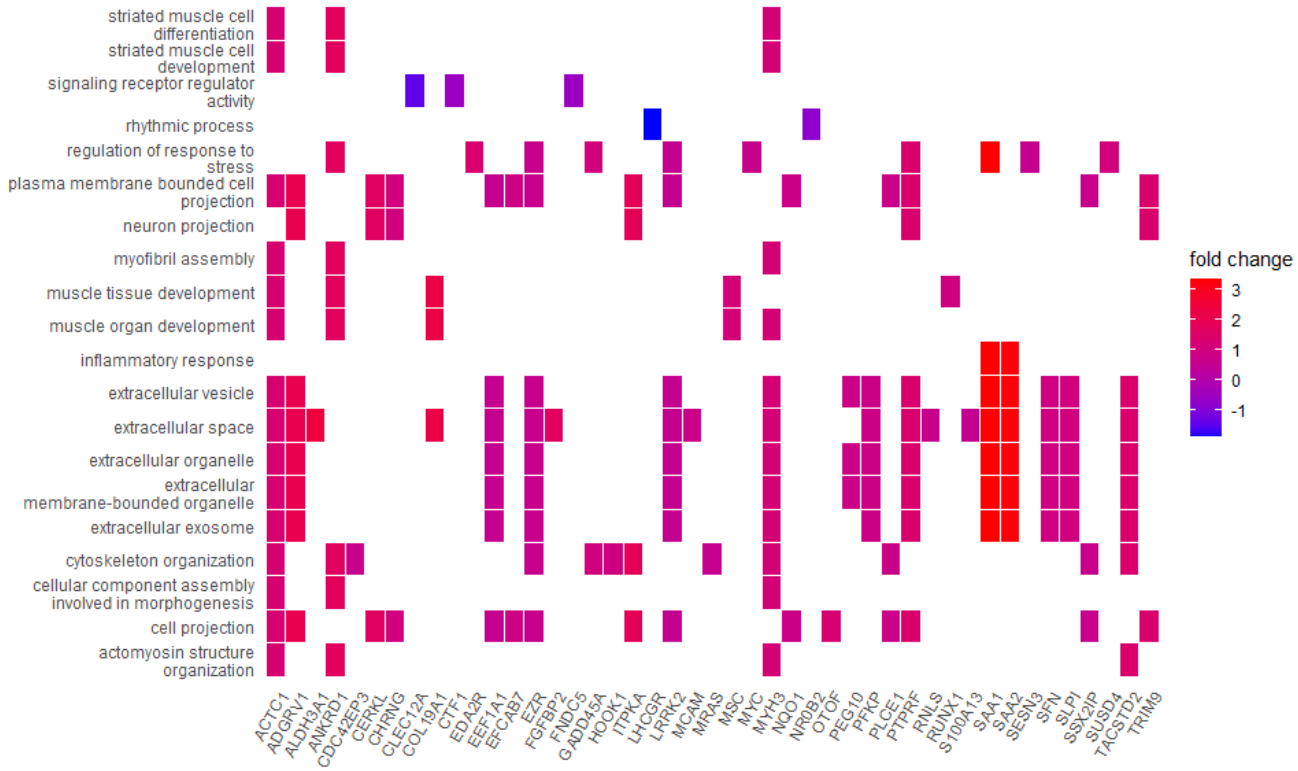


Figure 6. Heatmap of the top 20 statistically significant expressed genes combined from both tissues, enriched in pathways based on log2FC = -1/+1 and q < 0.05 thresholds. Large positive Log2FC was evident in several genes, particularly in the *SAA* family that were enriched in inflammatory response. However, other notable genes enriched in neuronal pathways included *ADGRV1*, *CERKL, CHRNG, IPTKA and PTPRF*. For an individualised presentation of results in skeletal muscle and blood – please see Supplementary Figures 3 and 4.

Detection of rare aberrant splicing events implicated in ALS pathogenesis

FRASER

In FRASER[56], 113,446 splice sites and 57,657 junctions were identified in the blood, while 168,259 splice sites and 87,470 junctions were counted in muscle samples after filtering based on |ΔPSI| > 0.1 (Supplementary Figure 5, FRASER 1.0) (for previous obtained results pertaining to the previously described metrics – Supplementary Table 4 and 5). Since FRASER detects splicing events based on a "sample-by-sample" basis, post-filtering clustering analysis was performed based on logistic regression (logit) correlation of the J index between covariates (sex and age) revealed subtle differences. It was evident that in skeletal muscle samples that the highest positive logit correlation ranging from r = 0.15 to 0.2 was observed predominately in male cases, while majority of logit correlations ranged from r = -0.1 to -0.2 (Supplementary Figure 6A). Alternatively, in blood samples the highest positive logit correlation of r = 0.7 was observed in a female control. All other correlations ranged from r = 0.2 to -0.3 within blood samples (Supplementary Figure 6B). Based on the set pathological thresholds as recommended by the authors of FRASER, we identified 13 intron-centric aberrant events in affected muscle samples and 8 in blood samples (defined by the J index metric), while controlling for covariate effects. The top 5 most statistically significant events in cases are represented below (Table 4). To better understand the detected effect, we prioritised visualising the most statistically significant events based on the J index. This was found in skeletal muscle (*RPL3L* and *PSMD2*) with most results favouring the shorter isoform (Figure 6).

Table 4. Top 5 most statistically significant AS events within skeletal muscle and blood collections at the local level in cases (FRASER 2.0)

Skeletal Muscle

Gene	Sample ID	ALS type	Significance (q)	ΔJ Index
RPL3L	HM-M4	sALS	2.2×10^{-9}	-0.53
FLNC	HM-M28	sALS	2.7×10^{-5}	-0.44
PSMD2	HM-M34	sALS	3.2×10^{-4}	-0.36
CUTC	HM-M27	-	3.9×10^{-4}	-0.43
STX8	HM-M9	sALS	1.9×10^{-3}	0.34

Blood

Gene	Sample ID	ALS type	Significance (q)	ΔJ Index
ASAH1	5197756	-	1.6×10^{-4}	-0.49
RAB31	1001953	fALS	6.7×10^{-3}	-0.91
IRF2	1001541	-	2.6×10^{-2}	-0.83
LYZ	1001689	fALS	2.9×10^{-2}	0.35
SSR2	8004754	sALS	3.5×10^{-2}	-0.59

Note: Top statistically significant splicing events identified in skeletal muscle and blood samples are based on q < 0.05 and |ΔJ index| > 0.3, as per documentation by the authors of FRASER. ALS types were matched based on SALSA IDs of known past medical history of patients, some of which were not obtainable for this project (e.g., M27).

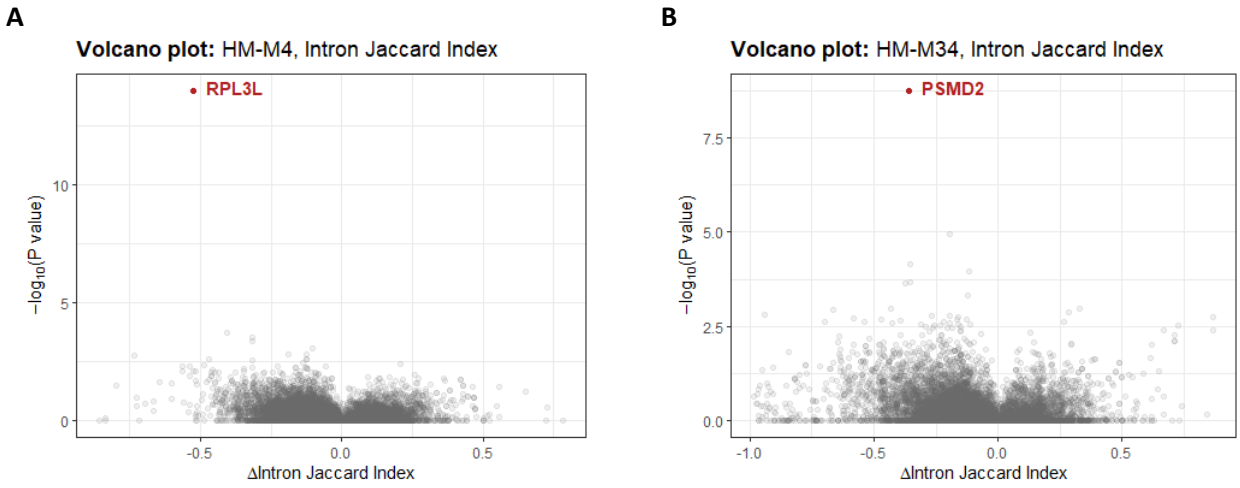


Figure 6. Volcano plots of some of the most statistically significant spliced genes based on the ΔJ index within skeletal muscle cases **A.** *RPL3L* (HM-M4) and **B.** *PSMD2* (HM-M34) were some of the top significant splice sites within affected muscle. Both plots show splice site-level significance ($-\log_{10}(p)$, y-axis) versus the change in J index. Highlighted in red is the statistically significant splice site observed in case patients, while in grey are non-statistically significant sites that did not meet the thresholds, $|\Delta J \text{ Index}| > 0.03$ and $q < 0.05$.

rMATS

In the group-level rMATS analyses, 182,494 events were identified in blood and 203,940 across skeletal muscle with top results listed in Supplementary Table 6. In skeletal muscle, cases vs. controls, 3 genes exhibited a significant degree of AS. *NEB* (most significant, $q = 4.1 \times 10^{-8}$) had an exon skipping event (increased in controls), *TNNT3* (lowest, $q = 6.7 \times 10^{-7}$) had two mutually exclusive exon/expand events (MXE), and *MACF1* ($q = 2.5 \times 10^{-6}$) had a single intron retention event with the largest change in mean PSI (Table 5). In blood, statistically significant genes between case-controls included *WDR26* ($q = 5.9 \times 10^{-12}$) and *SNRPB* ($q = 1.6 \times 10^{-6}$). Both consisting of intron retention (IR) events within ALS patients (Table 5).

Table 5. Top 5 most statistically significant genes based on alternative splicing events within skeletal muscle and blood collections in cases vs. controls level (rMATS)

Skeletal Muscle

Gene	Strand	Feature types	Significance (q)	Mean ΔPSI
NEB	-	ES	4.1×10^{-8}	0.19
NEB	+	MXE	3.2×10^{-7}	-0.10
TNNT3	+	MXE	6.7×10^{-7}	-0.12
MACF1	+	IR	2.5×10^{-6}	-0.24
TNNT3	+	MXE	2.1×10^{-6}	0.12

Blood

Gene	Strand	Feature types	Significance (q)	Mean ΔPSI
WDR26	-	IR	5.9×10^{-12}	-0.16
SNRPB	-	IR	1.6×10^{-6}	-0.11
AIF1	+	A3SS	2.8×10^{-5}	0.13
MFSD8	-	IR	2.2×10^{-5}	-0.11
HNRNPA2B1	-	ES	3.9×10^{-4}	-0.12

Note: Top statistically significant splicing events identified in skeletal muscle and blood samples are based on $q < 0.05$ and mean $|\Delta PSI| > 0.1$, as per documentation by the authors of rMATS. Due to the high frequency of genes occurring based on cases vs controls, here we record the number of events for each gene locus based on significance of adjusted p-value and mean ΔPSI. Negative ΔPSI favours ALS for events, while positive ΔPSI favours controls. Full coordinate output from rMATS is represented in Supplementary Table 6. Events are represented by: IR - intron retention; ES – exon skipping; MXE – mutually exclusive exons; A3SS – alternative 3' splice site.

The top results were visualised in skeletal muscle and blood using sashimi and transcript plots. Since rMATS reports at the group level, case-controls were visualised for aberrant splicing with Jutils and ggSashimi statistical packages, as it is often difficult to discern the effect of splicing without visual inspection. As aforementioned from Table 5, our results showed that the highest absolute mean ΔPSI was observed in *MACF1* with an intron retention (which was the only intron retention event observed in skeletal muscle) splicing event (chr1:39463611-39468732) between exon 95 and 96 of *MACF1.* Subsequently, we used Maser to visualise which transcript was being alternatively spliced in. It was found that with reference to Ensembl database that the intron retention event was consistent with MACF1-242 (ENST00000686657) (Figure 7). Further inspection of the transcript, we identified 3 splice region variant SNPs on exon 2 and 3 that are involved in alternative splicing (rs1198417183, rs774973543 and rs1183719996) using the Ensembl database.

Next, we also visualised which cases may be contributing the largest change in PSI in *MACF1* within ALS patients. Inspecting each sample via sashimi plot and with reference to Supplementary Table 1, it was evident that there were considerably more arcs connecting different regions of the sashimi plots, which is indicative of splice junctions between the intron loci. Visually, it was evident that ALS samples pertaining to patients with bulbar-related symptoms had the greatest number of arcs and higher read counts between splice junctions (Figure 8). Of several identified transcripts at the loci, specific transcripts following the pattern specific to the bulbar onset ALS splice junctions were intron retention events observed to be MACF1-247 (ENST00000687997) and MACF1-246 (ENST00000687885). Both transcripts were checked on the Ensembl database, revealing no protein function consequence.

Figure 7. Maser transcript visualisation revealed a statistically significant difference between case-controls at exons 95 and 96 on the positive strand of *MACF1*, with a mean ΔPSI = -0.24. The visualisation shows that MACF1-242 (ENST00000686657) exhibits an intron retention event. An enhanced view outlined in the green textbox further demonstrates a statistically significant difference in PSI, which is depicted by boxplots between cases vs. controls.

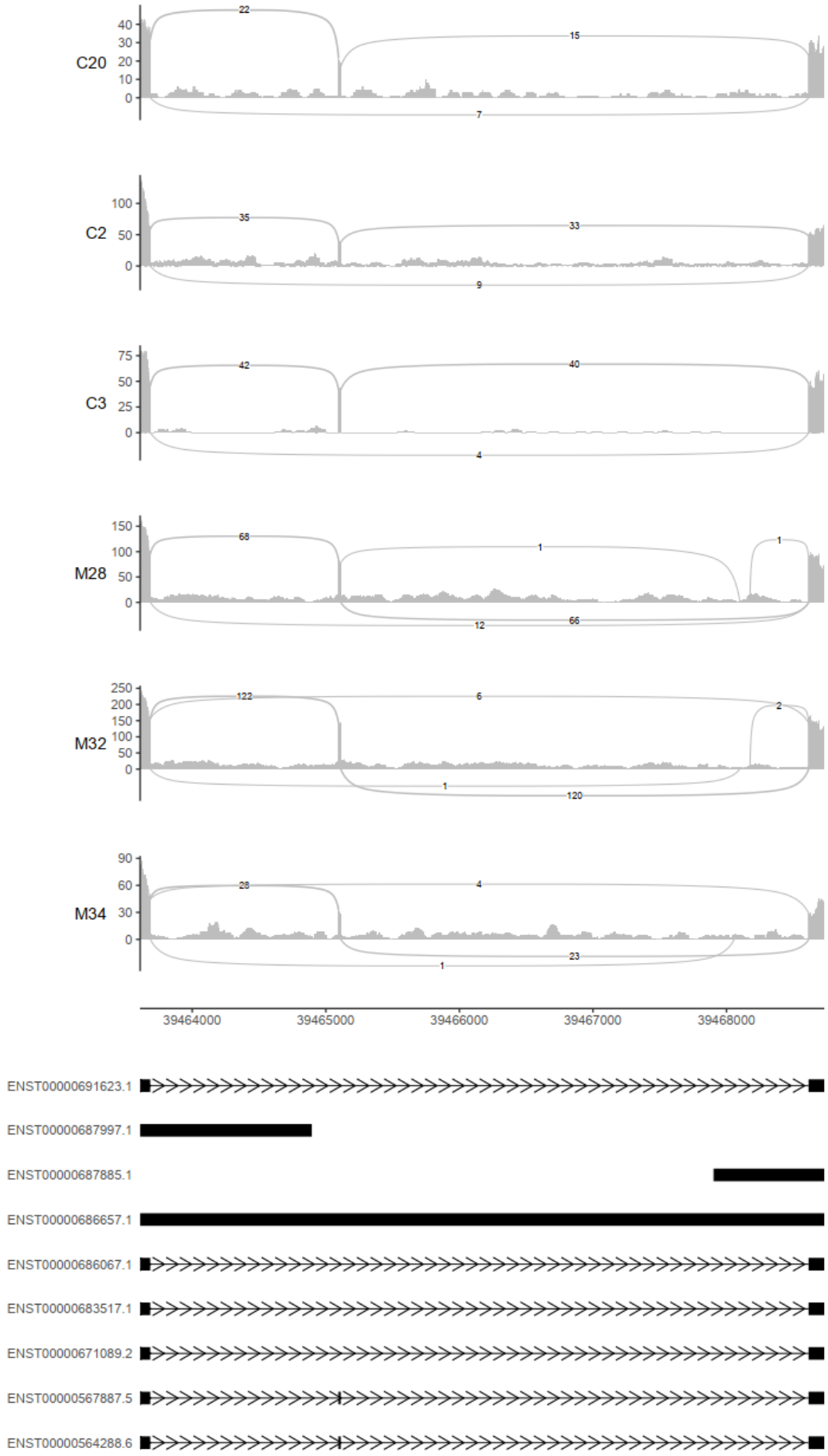


Figure 8. Sashimi plots between exon 95 and 96 of *MACF1*. Subset of bulbar-onset ALS muscle samples (M28, M32, M34) vs. random controls (C20, C2, and C3) show increased number of splice junction arcs (with corresponding values over the arcs indicating read support at that junction) in bulbar cases. The y-axis of each plot represents read coverage and the x-axis is the position loci of *MACF1* (chr1:39463611-39468732). Below the sashimi plots represent alternatively spliced in transcripts from control vs. ALS bulbar samples.

Aberrant splicing variant detection between tissue collections

SpliceAI

Variant calling was initially performed with an average of 32,722 variants in blood and 45,607 variants in muscle per sample (passed the filter threshold, a Phred score > 20). Utilising SpliceAI AS variant prediction, the top 50 most frequent alternative splicing variant events in muscle and blood were examined (Supplementary Figure 7 and 8). In skeletal muscle, *TTN* was identified to be the most persistent – 118 identified in cases and 88 in controls. The most frequent *TTN* variant was observed at position chr2:178593172, C>G, predominately in ALS patients (n = 16) (Supplementary figure 7). In blood, *PHIP* was the most frequent splice site in cases, with the most common observed variant at position chr6:78998283, A>C (n = 15). Focusing on variants present in cases, but not in controls the top six results are reported (Table 6). High-confidence alternative splicing variants were observed in at least 3 different case samples (based on SpliceAI Δ-score ≥ 0.8). Population frequency was obtained from reference databases (European). Interestingly, we identified SNVs that have not been seen before in large population databases. These rare variants, located within the *NBAS*, *SORL1,* and *HERPUD1* genes (Table 6), were classified as novel due to their absence in these databases. The frequency (MAF > 0.1) in our cohort (seen in n = 3, 2, 5 cases respectively) was relatively high (n= 0 controls).

Table 6. Case specific alternative splicing variants within skeletal muscle and blood collections

Skeletal Muscle

Gene	Position	Allele	Variant type	No. Cases (n = 27)	Population frequency
SAA1	chr11: 18269333	C > G	Missense	22	0.995
CAMPSAP1	chr9: 135881632	C > T	Synonymous	7	6×10^{-6}
UBE2G1	chr17: 4296816	T > C	Missense	6	4×10^{-6}
HLA-C	chr6: 31271154	G > C	Missense	5	0.12
HERPUD1	chr16: 56935233	A > C	-	5	Novel
FGD5	chr3: 14924000	A > G	Splice acceptor	3	0.92

Blood

Gene	Position	Allele	Variant type	No. Cases (n = 27)	Population frequency
SORL1	chr11: 121621062	A > T	Splice acceptor	9	0.001
SORL1*	chr11: 121490111	G > C	Missense	2	8×10^{-6}
SORL1*	chr11: 121490112	T > C	-	2	Novel
HLA-C	chr6: 31271154	G > C	Missense	5	0.12
PPIL2	chr22: 21695494	C > T	Missense	5	0.34
NBAS	Chr2: 15551492	C > G	-	3	Novel

Note: * highlights splicing variants of high significance and of interest given their population frequency in comparison to other variants of the same gene.

Discussion

ALS has a complex aetiology for which genetic mechanisms associated with disease risk and progression are still being understood. The aim of this preliminary investigation was to use transcriptomes from blood and muscle to see if we could detect disease relevant changes in expression and/or AS events. This project utilises the only muscle RNA-seq data from ALS cases that we know of, together with matched blood samples. A design that can provide insight into how a neurodegenerative disease can affect or be influenced by other tissues. One source that is easily accessible (blood) and one source that is impacted by disease progression (skeletal muscle).

Despite the small sample size, a total of 58 protein-coding genes had altered expression in cases compared to controls. This number surpasses findings from other larger studies of iPSC MNs[4] – demonstrating systemic disease contributions/impacts that could be relevant to disease. We note that our gene expression results differed between muscle and blood (despite being matched samples i.e., the same individuals) to suggest that the *in-situ* expression changes are tissue specific. As such, we discuss each tissue/source separately.

In muscle samples, the most notable expression change was in Serum Amyloid A (SAA) genes (*SAA1, SAA2* and *SAA2-SAA4*). These were all highly expressed in ALS cases compared to controls (replicating Freydenzon et al. (2023)). Briefly, it is well-known that SAA is embedded in inflammatory response pathways, as well as being involved in extracellular vesicle and organelle machinery with all three genes, contiguously located on chromosome 11. SAA is predominately synthesised by the liver, which are elevated in an acute-phase response and continually expressed in inflammatory rheumatic diseases[57,60]. Therefore, it is surprising that this was not observed in blood. Interestingly, SAA interacts with interlukin-6 (IL6) to mediate angiotensin-II induced skeletal muscle atrophy within heart failure patients[27]. This may explain the high levels of expression seen in our ALS tissue samples, that are likely impacted by lack of activation and muscle wasting/atrophy. Three other notable increased expression changes in ALS muscle were *CERKL*, *ACTC1* and *ANKRD1*. *CERKL* is involved in the regulation of post-transcriptional processes such as mitochondrial trafficking and autophagy regulation. It has also been associated with retinal dystrophy, characterised by the progressive degeneration of retinal neurones and vision loss. Its high expression in hippocampal neurones; a common site affected in ALS-FTD[12,23]. This suggests that this expression might also be altered in other tissues. For *ACTC1* and *ANKRD1* – both had direct involvements in the actomyosin

organisation structure which was detected in our gene-set enrichment. Both have been described to have critical involvement in diseased skeletal muscle[6].

In blood samples, we observed fewer overall changes in comparison to those found in muscle samples. This disparity could be attributed to several factors, including a marginally smaller sample size, decreased sequencing coverage, or potentially fewer changes induced by the disease in the blood. The most significant findings were found in *FGFBP2*, *CDK1NA* and *COL19A1,* which were highly expressed in ALS cases. Briefly, *FGFBP2* encodes for a fibroblast growth factor (FGF) protein that is selectively secreted by cytotoxic lymphocytes and is involved in response to tissue injury in autoimmune diseases[50]. While *FGFBP2* has not been implicated in neuromuscular dysfunction, a close FGF binding protein gene, *FGFBP1* has been demonstrated to modulate the actions of FGF in the wiring of neural circuits in the CNS and is also secreted from muscle fibres[63]. Therefore, given the crucial roles that FGF play in the innate immune system, such as response to tissue injury, it may be possible that other members of the family, like *FGFBP2*, could also share similar characteristics and involvement in the CNS.

Other notable genes including *COL19A1* and *CDK1NA,* have been identified within literature to be upregulated in ALS patients. Briefly, a larger study by Calvo et al. (2019) (n = 268, consisting of 3 cohorts of blood from ALS patients) suggested that Collagen XIX Alpha 1 (*COL19A1*) may be a prognostic biomarker in ALS. From their study, higher (1.75x) *COL19A1* expression levels were observed in ALS affected blood samples in comparison to controls, in combination with an increased mortality risk (17.9%, $p < 0.01$) and associated worse prognosis based on ALSFRS-r (Revised Amyotrophic Lateral Sclerosis Functional Rating Scale) scores ($r = -0.349$, $p = 0.025$)[9]. Our findings, in parallel with Calvo et al. (2019), may reinforce the potential role of upregulated *COL19A1* levels as strong predictors of disease progression. Furthermore, another gene that was identified to be upregulated in blood was *CDK1NA.* Similarly to our investigation, a larger study by Ziff et al. (2022) (n = 429, consisting of ALS iPSC MNs) identified upregulation of *CDK1NA*, a gene in the control of tumour suppressor protein p53 and DNA damage response pathways. Drawing further parallels, a smaller study (n = 7, consisting of ALS iPSC MNs) found upregulation of this gene particularly within familial cases[42]. Further follow-up is warranted in a larger longitudinal study.

These results closely align with Freydenzon et al. (2023) (as expected, same data) with minor pipeline differences (additional covariates and subtle differences in sampling pre-processing was included). The gene expression changes in blood and muscle highlight both the systemic impacts of

the disease – that can be found in a tissue-specific manner. The pathways implicated, particularly inflammatory response and neuronal projection in muscle and blood, provides crucial insight into the pathogenicity of ALS. Larger longitudinal studies, with deeper coverage, may find potential biomarkers of disease progression. In accordance with our project aims, we looked for and identified significant numbers of aberrant splicing events within blood and skeletal muscle. These were not exclusive to cases and so it is expected that a proportion of findings will not be related to disease and will be part of normal human variation. As such we have only focused on events that might have disease relevance (using converging evidence from the literature) that can help be prioritised in future studies. We discuss outcomes from each platform below.

Based on aberrant splicing analysis in FRASER 2.0 – a tool targeting rare events that may cause disease, we observed 21 significant aberrant splicing events at the individual sample level in cases. In skeletal muscle, *RPL3L* was identified as the most statistically significant aberrant splicing event (in sALS case) based on the change in Intron Jaccard index. *RPL3L* (located on chromosome 16) is specifically expressed in skeletal muscle and the heart, with the observed splicing event occurring between exon 10 and 11. It has been reported that a splice-donor variant (rs140192228, chr16:1945498) occurring at c.1167+1G>A produces an alternative isoform that skips exon 9, increasing the risk for atrial fibrillation[66]. This evidence provides good confidence in the event occurring and a downstream impact. Reviewing the clinical data available for this ALS case, it did not indicate this comorbidity, but given these findings we suspect the patient is at increased risk for atrial fibrillation. As this information is generated on a research basis, it will not be fed back to the clinic, but does suggest that these types of analyses do have implications (like other types of genetic data) to be considered in the future, particularly if disease longevity in ALS is improved.

In blood, *ASAH1* was identified to be the most statistically significant at the local level in cases. *ASAH1* (located on chromosome 8) is widely expressed throughout many tissues, but mainly in the heart, blood, and thyroid. We observed an intronic splicing event between exon 8 and 9 (junction 32, based on GTEx). At the given locus, a rare splice-acceptor variant (rs1800170812, chr8:18071391) produces an alternative shorter isoform. Although there is currently no population or clinical statistics on this splicing variant, the subsequent transcript (RefSeq: NM_177924.5) has been previously associated with spinal muscular atrophy and progressive myoclonic epilepsy[38]. Another notable local aberrant splicing event observed in muscle cases locally was *PSMD2* (in sALS case). *PSMD2* encodes for a non-ATPase subunit of the 19S regulatory base, which has been

previously implicated in impaired proteosome function within the pivotal ubiquitin-proteosome system (UPS)[65]. Our results indicated intronic splicing between exon 2 and 3. Several studies have implicated that aberrations within UPS contribute to the pathogenesis of multiple neurodegenerative disorders[76]. A case study by Kumar and Haider (2022) found that *PSMD2* has been identified to have association in DNA damage pathways with *SOD1*-related ALS and FTD. Subsequently, further investigation into AS in *PSMD2* may be worthwhile. Nevertheless, our local results give valuable insight into some unique events that may play a role in the disease of individuals. As mentioned, controls also exhibiting "rare" aberrant splicing which may be due to other underlying predispositions to medical conditions or benign / normal variation. Large reference data of AS is needed and may reveal clearer patterns to support prioritisation at a local level.

At a group level, using rMATS, we identified 11 different splicing events associated with cases. In skeletal muscle, 3 critical genes were identified in genes that previously have been described in nemaline myopathy and neuromuscular dysfunction. These genes included *MACF1, NEB* and *TNNT3.* Our study emphasised that *MACF1* had the most significant shift in PSI across both tissue types, and it was the only gene to exhibit an intron retention (occurring between exon 95 and 96) event in skeletal muscle. The occurrence of intron retention events is crucial because if the retained introns are translated, it could lead to modifications in the amino acid sequence and subsequently change the protein structure, potentially interfering with its normal function. *MACF1* plays an essential role in stabilising the cytoskeleton structure, as it encodes an F-actin protein that is involved in the cross-linking of microtubules and F-actin microfilaments. An *ex-vivo* study by Wang et al. (2017) investigated *MACF1* and its implication in Parkinson's disease and complex brain malformations. It was found that alternatively spliced isoforms may contribute to the genetic aetiology in Parkinson's disease[71]. Additionally, it has been identified to play a critical role in the neuronal migration of pyramidal neurones via microtubule dynamics[47]. Although there specifically is no literature supporting the intron retention event observed in our study to be associated with neurodegenerative conditions, the elevated levels of reads at splice junctions observed in our study in cases of bulbar-associated ALS could be evidence of dysfunction. In our study, the alternatively spliced transcripts in bulbar-associated ALS patients did not demonstrate substantial consequences on protein function (via Ensembl database). Nevertheless, a more comprehensive investigation is warranted due to the low read support at some of these splice junctions. These observations could potentially stem from technical artefacts, highlighting the need for meticulous scrutiny in future investigations.

In blood, among the top five most significant AS events identified in the group analysis, one gene stood out: *HNRNPA2B1.* Mutations in this gene are associated with variable disease presentations including myopathy, dementia, and ALS within the same family[35]. Notably, we observed an exon skipping event in *HNRNPA2B1* affecting on exon 12 (Supplementary Figure 7). Previous studies have suggested that exon skipping within *HNRNPA2B1* is associated with the shorter isoform (resulting in losing part of the glycine-rich region – responsible for critical function and where mutations are found), which has been implicated in the pathogenesis of chronic degenerative myopathy and ALS[36]. Similarly, to *TARDBP, HNRNPA2B1* also encodes for RNA-binding protein associated with neurodegeneration. *HNRNPA2B1* encodes for heterogenous ribonucleoprotein A2 and have been reported to promote protein aggregate formation in both fALS and sALS[17]. Although the exact function remains elusive, it has been demonstrated that TDP-43 binds and modulates the splicing of *HNRNPA2B1*, which may be affected by exon skipping events as they are often induced by stress changes leading to downstream cellular dysfunction[17,30]. Hence, the underlying mechanism has been suggested to mimic the mislocalisation of TDP-43 in a "prion-like" state. Our results showed skipping of exon 12, which specifically has not yet been previously described in literature to be associated with ALS. Therefore, it would be a worthwhile to further validate this event in a larger cohort, given its prior implication in the pathogenesis of ALS.

A previously reported AS event in *UNC13A*, that has been found in iPSC MNs and brain tissue was not detected in any of our transcriptome-wide using tools that we would have expected detection if it was occurring even in just ALS case (FRASER). To examine this more closely, we used a candidate approach to identify any abnormal splicing events within the *UNC13A* locus using LeafCutterMD. There was minimal evidence of reads mapping to junctions for intron excision analysis, with only two patients in the blood sample having a single read mapped to each junction. This resulted in insufficient junction counts to perform intron clustering analysis. In general, the *UNC13A* locus exhibited low read counts, with less than 30 reads per sample mapped in both muscle and blood samples). Consequently, no discernible splicing patterns were identified within the *UNC13A* locus. This lack of detection could be attributable to low expression in these tissues, limited read coverage, limited sequencing depth. Thus, it is unclear whether CE in *UNC13A* locus is a systemic phenomenon – with our current results suggesting it is neuronal specific given lack of expression detected in both skeletal muscle and blood (despite having *UNC13A* risk allele carriers in the cohort).

Splicing variants from SpliceAI that were exclusively seen in at least three cases (and not seen in controls) have been prioritised for discussion, but we note other less frequent variants may also be relevant to disease. In muscle collections, rs1671926 within the *SAA1* locus was consistently seen in 78% of cases within our sample group. Currently, there is no reported clinical significance for this variant in well-curated databases. Given the high population frequency of the alternative allele (C>G) seen in this missense variant (Thr77Ser) and elevated levels of expression within ALS patients, it is likely related to skeletal muscle tissue injury given its implication in inflammatory response. Further investigation into the role of SAA accumulation as an inflammatory biomarker could provide meaningful insight into disease progression.

One variant (rs776188045) in *CAMPSAP1* was identified in 26% of our muscle cases – despite typically being very rare in population databases (< 0.0001). It has not been implicated in ALS other diseases but could be relevant to follow-up as it is predicted to be deleterious by *in-silico* algorithms despite being annotated as a synonymous variant (amino acid unchanged). *CAMPSAP1,* encodes a microtubule minus-end binding protein, which drives neuronal polarisation, migration, and cortical lamination[78]. This might be an interesting investigation into the functional consequence of this variant within ALS patients. Furthermore, we also identified two other variants not seen in population databases – *HERPUD1* (skeletal muscle) and *NBAS* (blood).

HERPUD1 was identified in 19% of our cases. Briefly, it is an endoplasmic reticulum (ER) membrane protein that has been found to maintain intracellular calcium homeostasis under stress conditions in muscle[49]. Though the precise roles of *HERPUD1* in disease conditions are yet to be fully understood. It has been found that the encoded protein interacts with presenilin proteins, leading to an elevation in amyloid-beta (Aβ) protein levels[22]. Although we are unable to confirm the clinical significance of this event, we can postulate that the aberrant splicing variant we identified in *HERPUD1* may be linked to *A*β protein expression, since it is known to accumulate in tissues affected by neurodegenerative diseases. Conversely, *NBAS* was also identified in 11% of our blood cases. It encodes for a protein involved in several cellular processes, including transport pathways in the ER. It has been reported to cause rare diseases, such as short stature optic nerve atrophy[51]. Although current literature does not show any direct evidence of *NBAS* and *HERPUD1* in ALS, further investigation is warranted given their absence from well-curated databases.

Using SpliceAI, three rare *SORL1* variants were identified in several patients within blood collections. We observed high prediction delta scores ranging from 0.97-1 across three separate variants. *SORL1* encodes for SorLA, a key protein involved in the processing of the Aβ precursor protein (APP) and the secretion of Aβ peptide, which has been identified has the aggregation in Alzheimer's disease (AD) pathophysiology[10]. In 9 individual patients, rs1285562954 (splice-acceptor variant) was identified at the boundary of intron 43 and exon 44 of *SORL1*. Two other rare variants were observed on the boundary of exon 5 and intron 5, rs779590215 a missense variant (Trp253Ser) and one other novel *SORL1* variant located 1bp downstream from rs779590215. Among the 9 patients with this splice-acceptor variant, there was no clear pattern in relation to ALS region of onset and family history within this sample. It is worth noting that we identified all three *SORL1* variants in one sporadic ALS case with bulbar onset symptoms, which may play a role in the divergent continuum between AD and FTD. In summary, the frequent occurrence of *SORL1* AS observed exclusively in ALS patients presents compelling insight, however, further AS patterns would need to be observed in a larger dataset to confirm these findings.

The SpliceAI variant detection was extensive and some of the results reporting relied on the most frequent that occurred (Supplementary Figure 8 and 9). As such, this might have enriched results for large genes, such as *TTN*, which is the top ranked and one of the largest in the human genome. Future analyses or newer tools should aim to correct for this, so that a less biased approach that focused on disease is carried out. Its inherent gene annotation relying on a pre-established deep learning model, can be susceptible to erroneous interpretation of genetic variants, and thus each finding is still necessary to confirm its annotations and available information. Notably, SpliceAI does not yet include UK BioBank data and so novel variants might have been annotated.

In *UNC13A* splice variant analysis, we identified several genetic SNPs that could be splice alterations (Supplementary Table 8), but we were not able to see this in our transcriptome data due to low coverage and expression of the *UNC13A* region. Therefore, in this preliminary study we show that to further validate our findings and show that *UNC13A* has neuronal specificity within a larger sample size. Overall, we did not observe much overlap in aberrant splicing between splicing tools, at the local or cohort level. This was somewhat expected, due to using different statistical algorithms that utilise different metrics to capture various types of aberrant splicing in ALS patients. Additionally, it is relevant to note that AS tools fundamentally rely on the prediction of splicing events and short-read sequencing. These approaches, while valuable, are potentially susceptible to

the introduction of errors, thereby necessitating judicious interpretations of results. Subsequently, independent confirmation should be carried out to confirm our findings (particularly if the observed isoform is longer than 120bp). Long-read sequencing may allow the detection of splicing variants that are difficult to detect with short-reads. This technique may uncover previously undetectable isoforms, thereby compensating for the inherent limitations associated with short-read sequencing and augmenting our understanding of the transcriptomic landscape in ALS.

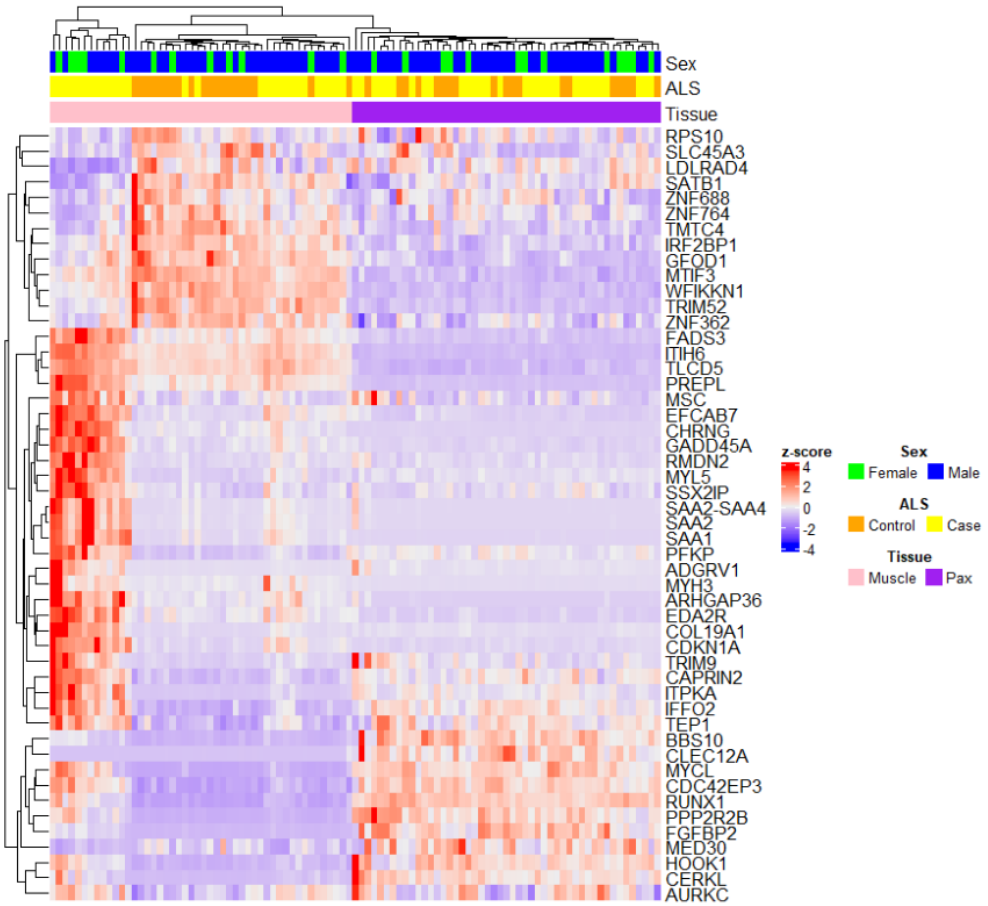
While our investigation into RNA-seq of skeletal muscle and blood is a preliminary study, it has nonetheless revealed potential genetic targets through differential gene expression and aberrant splicing analyses. Of several genes, three key splicing events identified in *SORL1*, *HNPRNAB1* and *MACF1* are warranted for further follow-up. In view of these promising findings, our research group is planning a more comprehensive RNA-seq study using a larger dataset (more ALS blood samples). This follow-up investigation will be critical for corroborating the results of our initial study and identifying common events. Our current results already hint at tissue-specific expressions, highlighting the importance of this forthcoming larger study. Furthermore, combining newly sequenced samples with publicly available data may also provide further validation and enhanced study depth. RNA-seq is a rapidly evolving area, current bioinformatic tools accounting for multiple-testing remain limited as there currently is no “ground truth” reference set. Newer prediction-based tools (such as AbSplice + SpliceMap[70], released May 2023) and robust pipelines may provide further insight into gene expression and aberrant splicing patterns in diseased tissues.

Conclusions

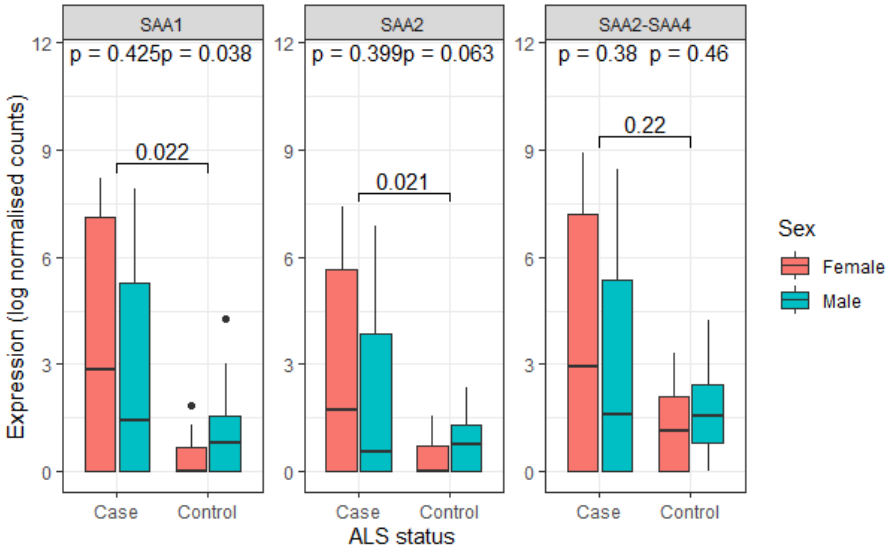
In this preliminary investigation, using two sources of RNA from matched samples, we were able to carry out a thorough investigation of the transcriptome in ALS. We saw significant gene expression changes within blood and muscle, two sources that are typically overlooked, but that might still provide insight into disease processes – from an onset perspective and disease progression. Additionally, using four complementary splicing tools we identified a significant number of aberrant AS events at a local and case-control level, for which a small number of findings could be prioritised for follow-up. Given the extent of variation, it remains relevant to carry out a much larger study, using multiple timepoints and tissues. This may help determine what splicing is normal human variation to help prioritise splice events relevant to pathogenic mechanisms. This of course will be incredibly difficult to do, because of the invasive nature of tissue sampling. We find that aberrant splicing is prevalent but heterogenous and thus to investigate it as a mechanism in disease pathogenesis larger studies are needed. As bioinformatic tools and technologies are evolving rapidly, more enhanced models and robust pipelines may indicate clearer patterns in gene expression and aberrant splicing. These findings lay the foundation for further investigation and provide important perspective on identifying and understanding aberrant splicing in ALS for potential therapeutic targeting.

Supplementary materials

Supplementary figures

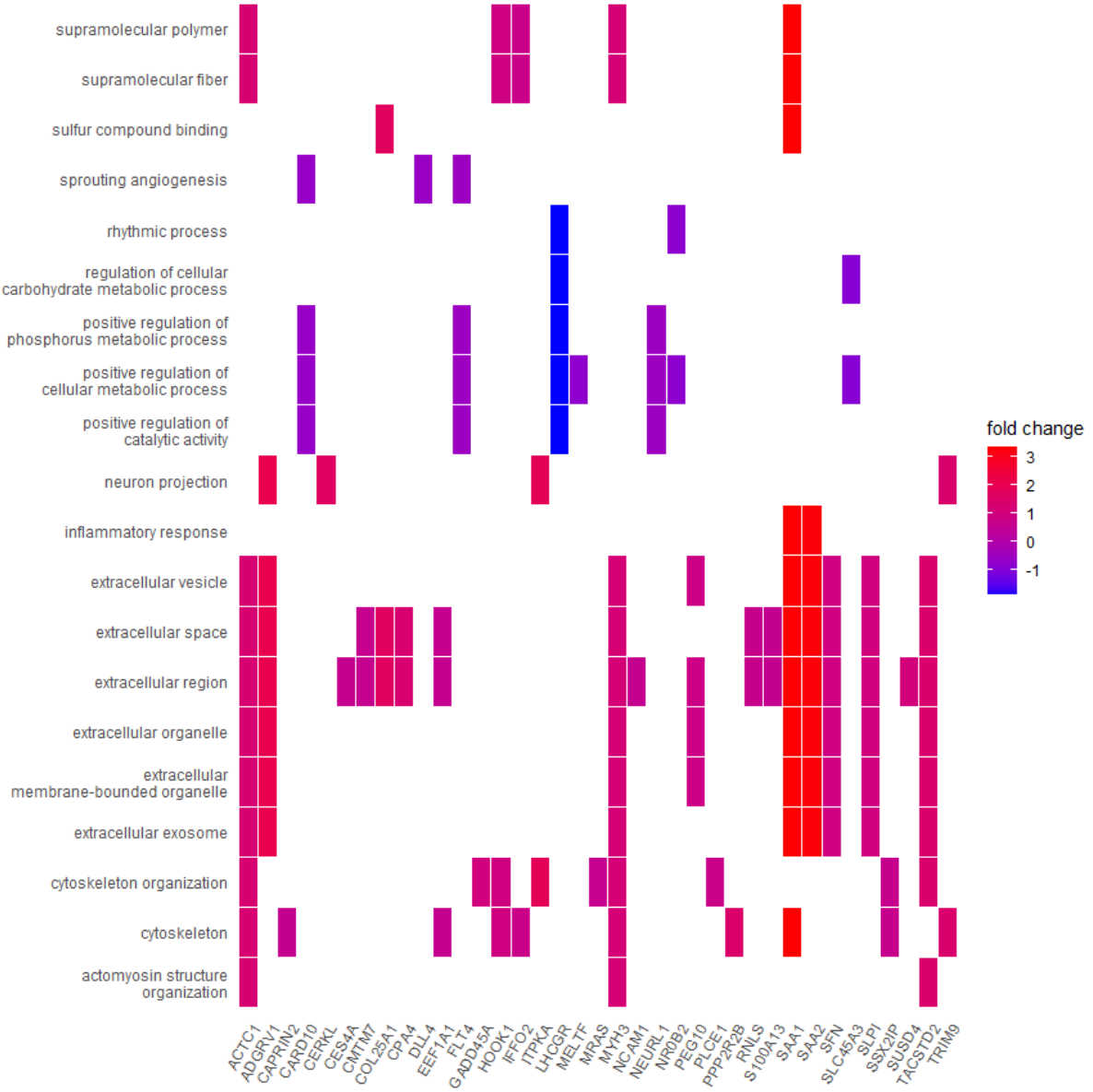


Supplementary Figure 1. Top 50 differentially expressed genes based on z-score of ALS status (with controls set as the reference and pooled into a combined GLM just for visualisation), tissue type and sex. Positive z-scores represent upregulation in ALS patients. Based on the legend, tissues are colour coordinated by Pax (blood, purple) and muscle (pink). Along with gender (male – blue, female – green) and ALS status (case – yellow and control – orange).



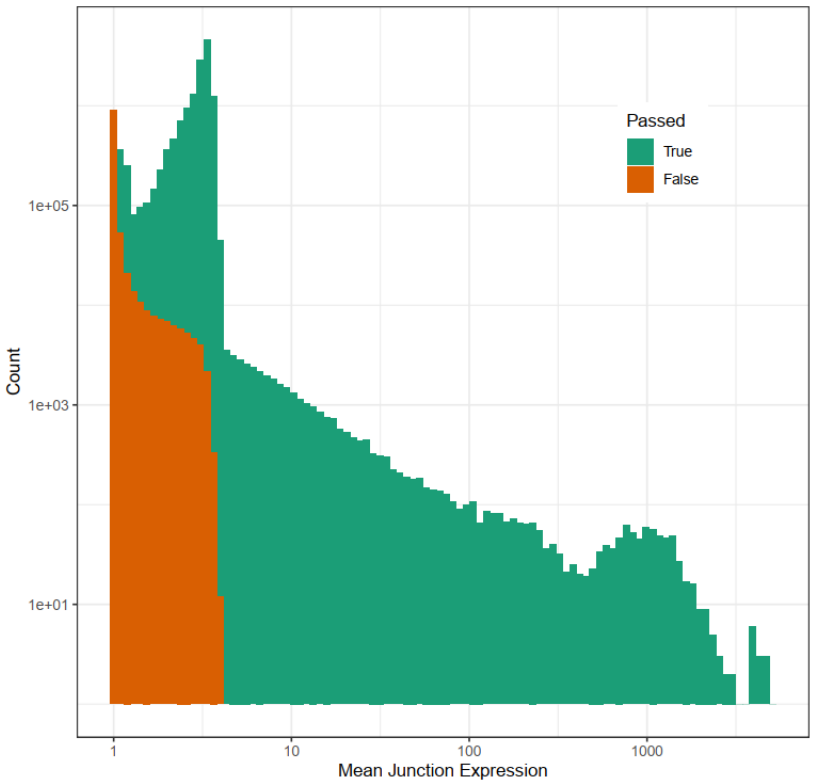
Supplementary Figure 2. Pairwise pair t-test between sex and ALS status based on paralogues of Serum Amyloid A - *SAA1*, *SAA2, SAA2-SAA4* genes. It was evident that there was no significant difference between sex in cases, however, there is statistically significant difference ($p < 0.05$) between cases and controls in *SAA1* and *SAA2*. The arrow between cases and controls represents the comparison based on case status, while p-values above represent significance based on gender.

Supplementary Figure 3. Heatmap of the top 20 statistically significant genes enriched in pathways based on log2FC = -1/+1 and q < 0.05 thresholds in blood. Large positive Log2FC was evident in several genes, particularly in *CDK1NA* and *ANKRD1*.

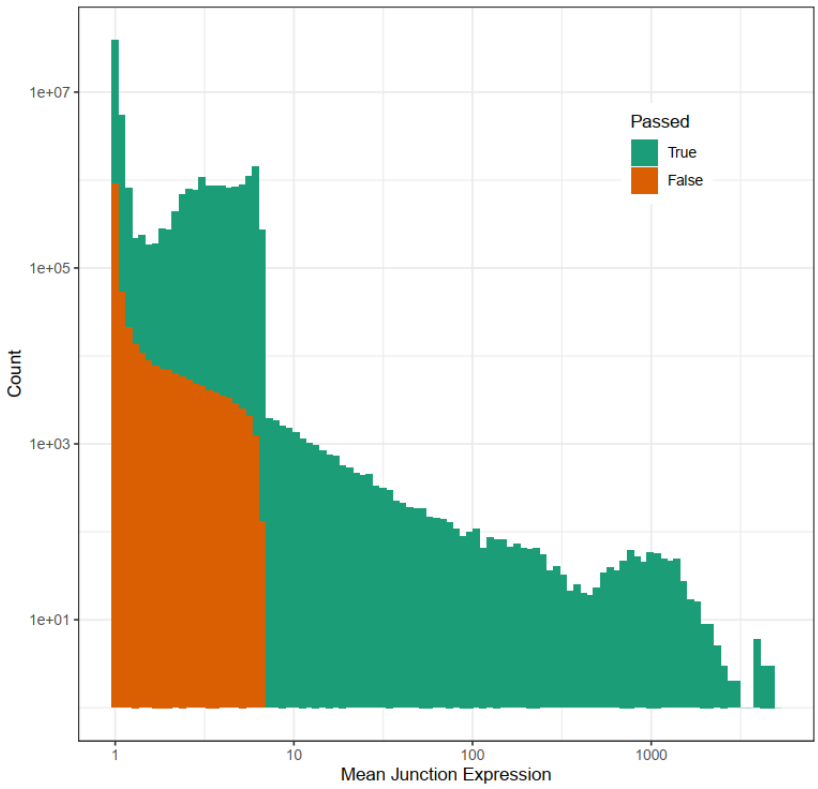


Supplementary Figure 4. Heatmap of the top 20 statistically significant genes enriched in pathways based on log2FC = -1/+1 and q < 0.05 thresholds in muscle. Large positive Log2FC was evident in several genes, particularly in the *SAA* family that were enriched in inflammatory response.

A

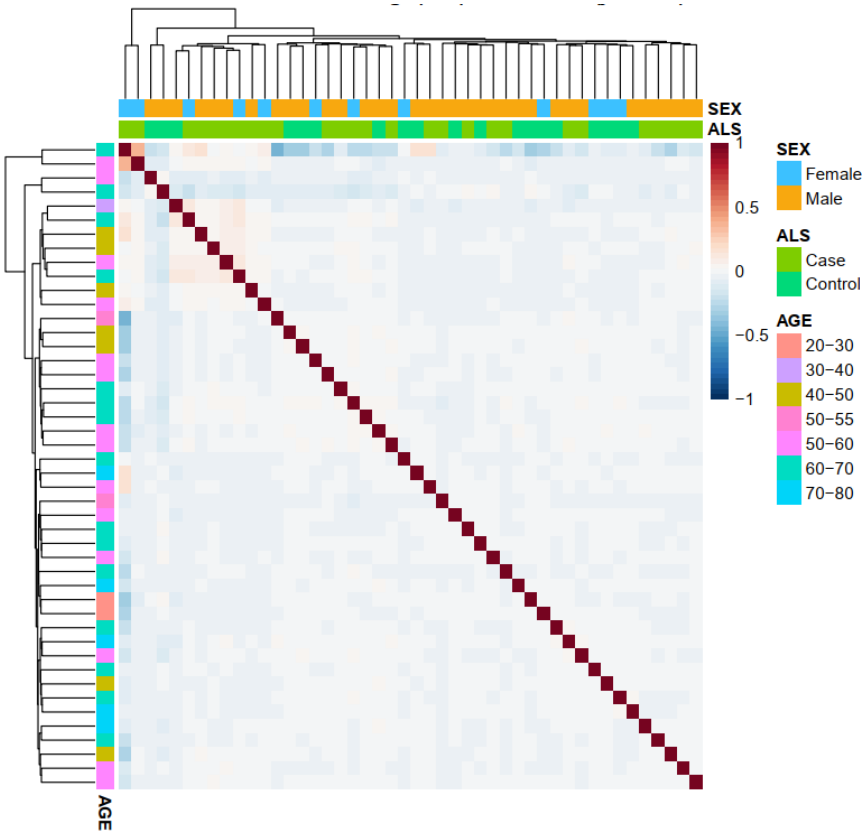


B

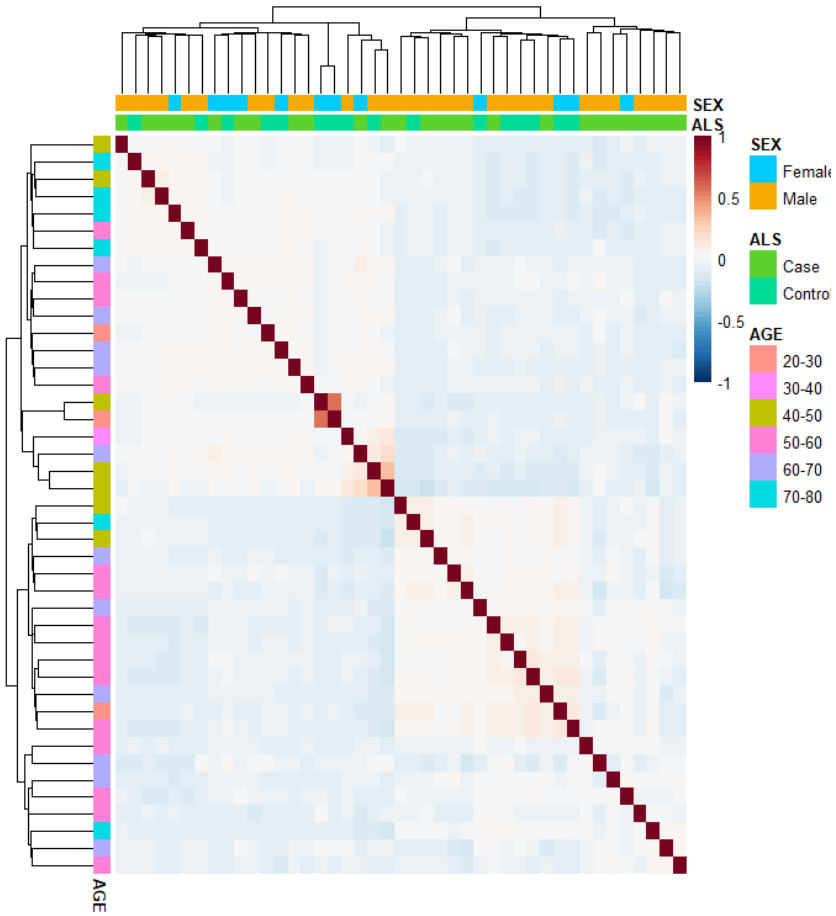


Supplementary Figure 5. FRASER filtering (filtered expression (based on ΔPSI > 0.1)) visualisation showing the distribution of counts of splice sites based on mean junction expression that passed (TRUE) and did not pass (FALSE) the threshold. **A.** plot of counts at junctions in blood **B.** and vice versa in muscle.

A

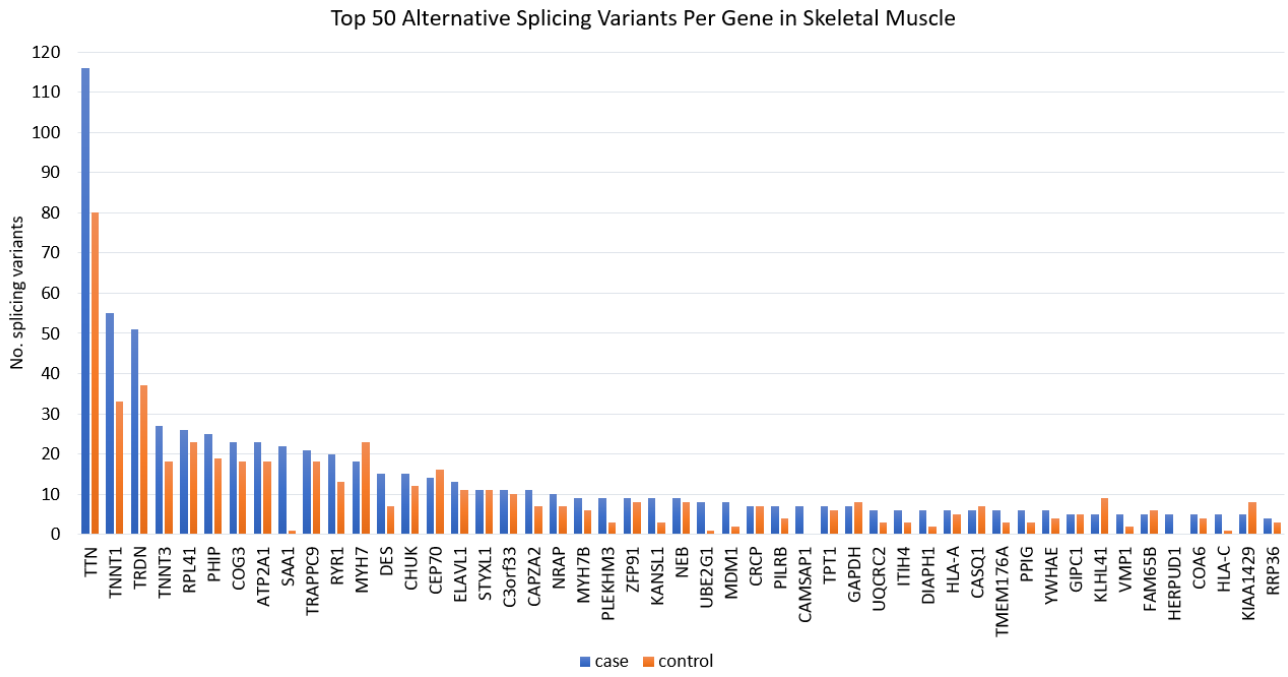


B

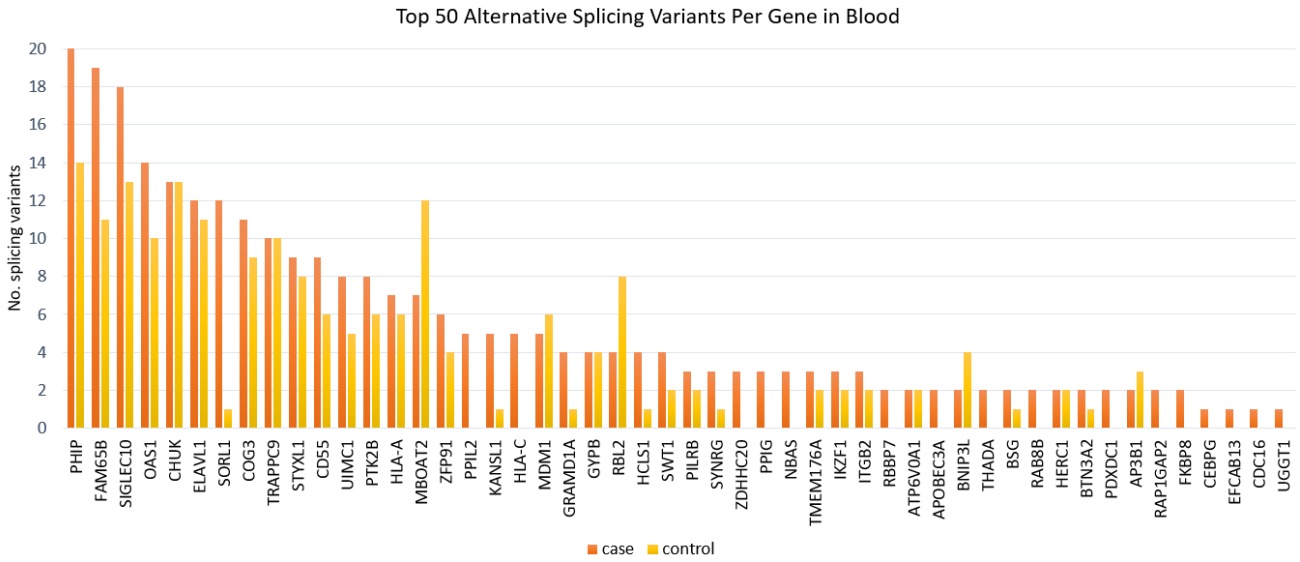


Supplementary Figure 6. Raw intron clustering of logistic regression correlation between age, sex and ALS status for case and controls within **A.** skeletal muscle samples and **B.** blood samples.

Supplementary Figure 7. *HNRNPA2B1* was identified as one of the most critical alternative splicing events at the group level in rMATS. Visualising using Maser, dysregulated alternative splicing resulting in exon skipping of *HNRNPA2B1* with 5 possible transcripts involved. Additionally, showing statistically significant difference based on mean ΔPSI in cases vs. controls.



Supplementary Figure 8. Top 50 alternative splicing variant based on genes with the highest frequency in skeletal muscle samples of cases vs controls.



Supplementary Figure 9. Top 50 alternative splicing variant based on genes with the highest frequency of splicing variants detected from SpliceAI based on Δ-score > 0.8 in blood samples of cases vs controls.

Supplementary tables

Supplementary Table 1. Data matching via SALSA IDs of past medical history of known ALS case and controls within muscle collections

ID	ALS	SEX	AGE (at collection)	COLLECTION TYPE	Genetics	Gene	Family history	Date of diagnosis	Region of onset	Disease type	Deceased	Riluzole
HM-M9	Case	Male	57.2	Muscle	genes neg, c9orf 72 neg	ZA NVD	sporadic	12-Mar-16	bulbar	ALS	T	T
HM-M14	Case	Male	62.5	Muscle			sporadic	1-Feb-16	upper limb	UMN	F	F
HM-M12	Case	Male	46.8	Muscle	genes neg, c9orf neg nmusc normal, c9orf normal at Pathwest	ZA NVD	sporadic	1-Sep-16	upper limb	ALS	T	F
HM-M4	Case	Female	66.8	Muscle	c9orf neg at UQ, UNCrs pos	UNC	sporadic	19-Feb-17	lower limb	UMN	F	F
HM-M18	Case	Male	69.2	Muscle	c9orf and Nmusc panel neg	ZA NVD	sporadic	1-Dec-13	upper limb	flail	T	F
HM-M6	Case	Male	52.5	Muscle	Nmusc- N c9orf neg	ZA NVD	sporadic	1-Jun-13	lower limb	UMN	F	F
HM-M7	Case	Female	57.9	Muscle	genes neg, C90rf neg	ZA NVD	sporadic	5-Nov-14	lower limb	ALS	T	F
HM-M10	Case	Male	46.9	Muscle	c9orf72 expansion, UNCrs pos	c9orf72, UNC	familial - father and children	1-Jun-16	lower limb	ALS	T	T
HM-M8	Case	Male	65	Muscle	genes neg, c9orf neg at UQ none in Auslab	ZA NVD	sporadic	10-Feb-17	upper limb	ALS	T	T
HM-M11	Case	Male	52.1	Muscle	c9orf72 neg, Nmusc panel neg, UNCrepos	UNC	sporadic	1-May-16	upper limb	ALS	F	T
HM-M13	Case	Male	54.1	Muscle	genes neg, c9orf 72 neg,, UNCrs pos	UNC	sporadic	24-Feb-17	lower limb	ALS	F	F
HM-M17	Case	Male	46.3	Muscle	c90rf72 repeat expansion, UNC rs pos	c9orf72 (UQ), UNC	sporadic	1-Feb-16	upper limb	ALS	T	F
HM-M16	Case	Female	62.4	Muscle	UNCrs pos	UNC	sporadic	23-Feb-17	lower limb	ALS	F	F
HM-M19	Case	Male	68.8	Muscle	genes neg, c9orf neg	ZA NVD	sporadic	20-Dec-16	lower limb	ALS	T	F
HM-M20	Case	Male	51.4	Muscle	c229G>T SOD1 genes neg, c9orf 72 neg at UQ, UNCrspos	SOD1, UNC	familial	1-Mar-16	lower limb	ALS	T	T
HM-M21	Case	Female	70.3	Muscle			sporadic	1-Apr-17	bulbar	ALS	T	F
HM-M24	Case	Male	54.2	Muscle	genes neg, c9orf neg, UNC rs pos	UNC	sporadic	24-Jan-14	lower limb	UMN	F	F
HM-M26	Case	Female	59.8	Muscle	geens neg, c9orf neg	ZA NVD	sporadic	28-Apr-17	lower limb	ALS	T	T
HM-M25	Case	Female	66.1	Muscle	genes neg, c9orf rs3849942 pos, UNCrs pos	c9orf, UNC	sporadic	1-Sep-16	lower limb	ALS	T	F
HM-M22	Case	Male	76.3	Muscle	c9orf neg, nmusc neg	ZA NVD	familial - cousin had MND	20-Aug-16	upper limb	flail	F	F
HM-M28	Case	Male	57	Muscle	genes neg, c9orf neg none in Auslab	ZA NVD	sporadic	4-Apr-17	bulbar	ALS	T	T
HM-M27	Case	Male	78.8	Muscle	genes neg, c9orf neg, UNC repos	UNC	sporadic	30-Mar-15	upper limb	flail	F	F
HM-M29	Case	Male	66.6	Muscle	cG1144>A TARDBP c9orf neg UNCrspos	TARDBP, UNC	familial- brother affected	12-Sep-14	lower limb	UMN	F	T
HM-M30	Case	Male	69.1	Muscle	gene neg , C9 neg, UNCrspos	UNC	sporadic	1-Mar-17	bulbar	ALS	T	F
HM-C17	Control	Male	63.6	Muscle			sporadic	15-Jul-21	lower limb	ALS	F	F
HM-C20	Control	Female	63.7	Muscle			sporadic	15-Jul-21	lower limb	ALS	F	F
HM-M23	Case	Male	53.9	Muscle	genes neg, c9orf neg, UNC rs pos	UNC	sporadic	25-Aug-17	upper limb	flail	F	T
HM-M32	Case	Male	57.9	Muscle	gene neg, c9neg nothing on Auslab	ZA NVD	sporadic	7-Sep-17	bulbar	ALS	T	T
HM-M33	Case	Male	58.7	Muscle	gene neg, c9orf neg nothing in Auslab	ZA NVD	sporadic	1-Jan-18	bulbar	ALS	T	F
HM-M34	Case	Male	56.1	Muscle	gene neg, c9orf neg, UNC rs pos	UNC	sporadic	4-Apr-18	bulbar	ALS (saines)	T	T

Supplementary Table 2. Data matching via SALSA IDs of past medical history of known ALS case and controls within blood collections

ID	ALS	SEX	AGE (at collection)	COLLECTION TYPE	Genetics	Gene	Family history	Date of diagnosis	Region of onset	Disease type	Deceased	Riluzole
600985	Case	Male	57.2	Pax	genes neg, c9orf 72 neg	ZA NVD	sporadic	12-Mar-16	bulbar	ALS	T	T
1000415	Case	Male	61.9	Pax			sporadic	1-Feb-16	upper limb	UMN	F	F
1001476	Case	Male	46.5	Pax	genes neg, c9orf neg	ZA NVD	sporadic	1-Sep-16	upper limb	ALS	T	F
8004754	Case	Male	46.5	Pax	genes neg, c9orf neg nmusc normal, c9orf normal at Pathwest	ZA NVD	sporadic	1-Sep-16	upper limb	ALS	T	F
1001530	Case	Female	66.7	Pax	c9orf neg at UQ, UNCrs pos	UNC	sporadic	19-Feb-17	lower limb	UMN	F	F
1001559	Case	Male	69	Pax	c9orf and Nmusc panel neg	ZA NVD	sporadic	1-Dec-13	upper limb	flail	T	F
1001664	Case	Male	52.4	Pax	Nmusc- N c9orf neg	ZA NVD	sporadic	1-Jun-13	lower limb	UMN	F	F
1001665	Case	Female	57.8	Pax	genes neg, C90rf neg	ZA NVD	sporadic	5-Nov-14	lower limb	ALS	T	F
1001689	Case	Male	46.8	Pax	c9orf72 expansion, UNCrs pos	c9orf72, UNC	familial - father and children	1-Jun-16	lower limb	ALS	T	T
1001695	Case	Male	65	Pax	genes neg, c9orf neg at UQ none in Auslab	ZA NVD	sporadic	10-Feb-17	upper limb	ALS	T	T
1001701	Case	Male	52.1	Pax	c9orf72 neg, Nmusc panel neg, UNCrepos	UNC	sporadic	1-May-16	upper limb	ALS	F	T
1001802	Case	Male	54	Pax	genes neg, c9orf 72 neg,, UNCrs pos	UNC	sporadic	24-Feb-17	lower limb	ALS	F	F
1001831	Case	Male	46.3	Pax	c90rf72 repeat expansion, UNC rs pos	c9orf72 (UQ), UNC	sporadic	1-Feb-16	upper limb	ALS	T	F
1001835	Case	Female	62.4	Pax	UNCrs pos	UNC	sporadic	23-Feb-17	lower limb	ALS	F	F
1001841	Case	Male	68.8	Pax	genes neg, c9orf neg	ZA NVD	sporadic	20-Dec-16	lower limb	ALS	T	F
1001859	Case	Male	51.3	Pax	c229G>T SOD1 genes neg, c9orf 72 neg at UQ, UNCrspos	SOD1, UNC	familial	1-Mar-16	lower limb	ALS	T	T
1001865	Case	Female	70.3	Pax			sporadic	1-Apr-17	bulbar	ALS	T	F
1001877	Case	Male	54.1	Pax	genes neg, c9orf neg, UNC rs pos	UNC	sporadic	24-Jan-14	lower limb	UMN	F	F
7336734	Case	Male	54.1	Pax	genes neg, c9orf neg, UNC rs pos	UNC	sporadic	24-Jan-14	lower limb	UMN	F	F
1001897	Case	Female	59.7	Pax	geens neg, c9orf neg	ZA NVD	sporadic	28-Apr-17	lower limb	ALS	T	T
1001900	Case	Female	66	Pax	genes neg, c9orf rs3849942 pos, UNCrs pos	c9orf, UNC	sporadic	1-Sep-16	lower limb	ALS	T	F
1001905	Case	Male	76.3	Pax	c9orf neg, nmusc neg	ZA NVD	familial- cousin had MND	20-Aug-16	upper limb	flail	F	F
1001917	Case	Male	57	Pax	genes neg, c9orf neg none in Auslab	ZA NVD	sporadic	4-Apr-17	bulbar	ALS	T	T
1001920	Case	Male	78.7	Pax	genes neg, c9orf neg, UNC repos	UNC	sporadic	30-Mar-15	upper limb	flail	F	F
1001953	Case	Male	66.5	Pax	cG1144>A TARDBP c9orf neg UNCrspos	TARDBP, UNC	familial- brother affected	12-Sep-14	lower limb	UMN	F	T
4267064	Case	Male	66.5	Pax	cG1144>A TARDBP c9orf neg UNCrspos	TARDBP, UNC	familial- brother affected	12-Sep-14	lower limb	UMN	F	T
1001991	Case	Male	69.1	Pax	gene neg , C9 neg, UNCrspos	UNC	sporadic	1-Mar-17	bulbar	ALS	T	F

Supplementary Table 3. FRASER 2.0 results of rare alternatively spliced events in blood samples based on Jaccard Index

seqnames	start	end	width	strand	sampleID	status	hgncSymbol	type	pValue	padjust	psiValue	deltaPsi	meanTotalCounts
chr8	18071391	18075540	4150	*	5197756	Case	ASAH1	jaccard	8.34E-09	0.001612	0.45	-0.49	27.71
chr8	22067160	22067526	367	*	1001684	Control	DMTN	jaccard	2.52E-08	0.004875	0.68	-0.31	45.34
chr18	9814092	9815115	1024	*	1001953	Case	RAB31	jaccard	3.48E-08	0.006723	0.06	-0.91	6.57
chr6	6250930	6266557	15628	*	5092533	Control	F13A1	jaccard	5.22E-08	0.010087	0.32	-0.67	17.93
chr15	60382442	60386027	3586	*	1001684	Control	ANXA2	jaccard	6.18E-08	0.011942	0.28	-0.68	21.82
chr12	49752550	49752972	423	*	9957921	Control	TMBIM6	jaccard	6.34E-08	0.012243	0.6	-0.38	21.23
chr4	1.84E+08	1.84E+08	760	*	1001541	Case	IRF2	jaccard	1.32E-07	0.025516	0.08	-0.83	9.74
chr12	69353609	69353742	134	*	1001689	Case	LYZ	jaccard	1.54E-07	0.029709	0.38	0.35	340.71
chr16	57172875	57173727	853	*	1001915	Control	PSME3IP1	jaccard	1.59E-07	0.030627	0.2	-0.73	12.51
chr5	1.39E+08	1.39E+08	433	*	9957921	Control	MATR3	jaccard	1.61E-07	0.031128	0	-0.96	12.04
chr1	1.56E+08	1.56E+08	1644	*	8004754	Case	SSR2	jaccard	1.80E-07	0.034857	0.39	-0.59	15.47
chr12	1.09E+08	1.09E+08	30	*	1001835	Case	SELPLG	jaccard	1.86E-07	0.035875	1	0.67	89
chr19	45522582	45522720	139	*	600985	Case	VASP	jaccard	2.44E-07	0.047113	0.92	0.79	32.15
chr20	58895685	58898940	3256	*	1001695	Case	GNAS	jaccard	2.52E-07	0.048645	0.74	0.61	62.18

Supplementary Table 4. FRASER 2.0 results of rare alternatively spliced events in muscle samples based on Jaccard Index

seqnames	start	end	width	strand	sampleID	hgncSymbol	type	pValue	padjust	psiValue	deltaPsi	meanTotalCounts
chr16	1944894	1945498	605	*	HM-M4	RPL3L	jaccard	2.12E-15	2.22E-09	0.47	-0.53	1155.39
chr16	1945619	1945834	216	*	HM-M4	RPL3L	jaccard	2.65E-14	1.38E-08	0.54	-0.45	1479.43
chr7	1.29E+08	1.29E+08	88	*	HM-M28	FLNC	jaccard	2.53E-11	2.64E-05	0.54	-0.44	1048.5
chr19	48918785	48919029	245	*	HM-C14	NUCB1;NUCB1-AS1	jaccard	2.35E-10	0.000246	0.66	-0.34	212.11
chr3	1.84E+08	1.84E+08	372	*	HM-M34	PSMD2	jaccard	3.01E-10	0.000315	0.63	-0.36	217.72
chr10	99754635	99755624	990	*	HM-M27	CUTC	jaccard	7.41E-10	0.000388	0.5	-0.43	434.28
chrX	48898094	48898491	398	*	HM-C11	PQBP1;TIMM17B	jaccard	1.62E-09	0.000848	0.92	0.75	86.93
chr5	44813283	44814912	1630	*	HM-C18	MRPS30	jaccard	1.70E-09	0.001778	0.51	-0.46	152.48
chr17	9557529	9575791	18263	*	HM-M9	STX8	jaccard	1.85E-09	0.001937	0.35	0.34	139.87
chr22	25895281	25898306	3026	*	HM-C10	MYO18B;MYO18B-AS1	jaccard	1.87E-09	0.001958	0.66	-0.31	411.54
chr1	51406109	51408134	2026	*	HM-C4	EPS15	jaccard	2.87E-09	0.001785	0.51	-0.46	147.07

chr2	33511843	33515009	3167	*	HM-C15	RASGRP3	jaccard	3.68E-09	0.003844	0.36	-0.6	103.65
chr8	73292798	73293598	801	*	HM-C18	RPL7	jaccard	5.63E-09	0.002946	0.56	-0.43	1459.11
chr8	73292798	73293598	801	*	HM-M15	RPL7	jaccard	6.04E-09	0.00632	0.56	-0.43	1459.11
chr2	27131832	27132009	178	*	HM-C7	PREB	jaccard	1.24E-08	0.013012	0.21	-0.66	110.02
chr8	73292796	73293598	803	*	HM-C18	RPL7	jaccard	1.61E-08	0.005602	0.44	0.44	1456.35
chrX	48898083	48898491	409	*	HM-C11	PQBP1;TIMM17B	jaccard	1.74E-08	0.004539	0.01	-0.38	99.35
chr8	73292796	73293598	803	*	HM-M15	RPL7	jaccard	1.82E-08	0.009509	0.44	0.44	1456.35
chr11	61873866	61875850	1985	*	HM-C13	FADS3	jaccard	2.71E-08	0.028342	0.27	-0.72	34.04
chr17	9568471	9575791	7321	*	HM-M9	STX8	jaccard	3.51E-08	0.018343	0.35	-0.61	46.74
chr1	51403533	51405904	2372	*	HM-C4	EPS15	jaccard	4.25E-08	0.014821	0.59	-0.39	128.85
chr1	24108655	24112030	3376	*	HM-M5	MYOM3	jaccard	2.97E-07	0.023908	0.37	-0.52	198.46
chr3	1.57E+08	1.57E+08	429	*	HM-M5	CCNL1	jaccard	7.01E-07	0.03839	0.3	-0.6	61.28
chr2	33361516	33363389	1874	*	HM-M5	LTBP1	jaccard	7.13E-07	0.03839	0.32	-0.62	36.04
chr4	1.69E+08	1.69E+08	3679	*	HM-M5	CBR4	jaccard	9.73E-07	0.048455	0.36	0.35	64.57

Supplementary Table 5. FRASER 1.0 results of rare alternatively spliced events in muscle samples based on theta, psi5 and psi3

seqnames	start	end	width	strand	sampleID	hgncSymbol	type	pValue	padjust	zScore	psiValue	deltaPsi	meanTotalCounts
chr16	1944894	1945834	941	*	HM-M4	RPL3L	psi3	1.21E-19	1.25E-13	5.99	0.45	0.45	1467.78
chr22	25895281	25898306	3026	*	HM-C10	MYO18B	psi5	4.52E-15	4.69E-09	-6.07	0.69	-0.3	402.67
chr3	1.84E+08	1.84E+08	372	*	HM-M34	PSMD2	psi5	8.02E-14	8.33E-08	-6.31	0.64	-0.35	214.39
chr10	99754635	99755624	990	*	HM-M27	CUTC	psi3	5.11E-12	5.31E-06	-5.57	0.57	-0.41	413.2
chr8	73292798	73293598	801	*	HM-M15	RPL7	psi3	3.08E-11	3.19E-05	-4.43	0.56	-0.44	1448.93
chr8	73292798	73293598	801	*	HM-C18	RPL7	psi3	4.72E-11	4.90E-05	-4.12	0.56	-0.44	1448.93
chr12	1.25E+08	1.25E+08	684	*	HM-M6	UBC	psi3	5.16E-11	5.36E-05	-4.86	0.05	-0.84	848.28
chr4	1.52E+08	1.52E+08	2	*	HM-M16	FHIP1A	theta	1.08E-10	0.000236	-4.21	0	-0.96	87.76
chr2	33511843	33515009	3167	*	HM-C15	RASGRP3	psi3	4.21E-10	0.000437	-4.66	0.51	-0.47	99.63
chrX	48898083	48898491	409	*	HM-C11	PQBP1	psi3	4.36E-10	0.000452	-5.77	0.01	-0.43	81.72
chr1	51403533	51408134	4602	*	HM-C4	EPS15	psi3	4.48E-10	0.000464	4.46	0.44	0.42	144.11
chr20	34978097	34979660	1564	*	HM-C4	MYH7B	psi5	5.33E-10	0.000554	4.89	0.31	0.3	120.8
chr8	1.02E+08	1.02E+08	2	*	HM-M30	RRM2B	theta	5.99E-10	0.000732	-4.66	0	-0.94	49.24

chr12	28452655	28484051	31397	*	HM-M17	CCDC91	psi5	1.27E-09	0.001323	-5.98	0.64	-0.35	73.52
chr2	85351863	85354335	2473	*	HM-C4	RETSAT	psi3	1.33E-09	0.000688	-4.71	0.34	-0.56	88.57
chr5	44814912	44814913	2	*	HM-C18	MRPS30	theta	1.90E-09	0.00418	4.23	0.39	0.35	150.96
chr21	42904277	42908863	4587	*	HM-M15	NDUFV3	psi5	3.10E-09	0.003216	-4.76	0.51	-0.48	76
chr17	9557529	9568370	10842	*	HM-M9	STX8	psi5	3.11E-09	0.003229	-4.93	0.57	-0.41	96.17
chr19	4511368	4511369	2	*	HM-M32	PLIN4	theta	3.36E-09	0.007375	-4.33	0.04	-0.62	220.78
chr2	27131831	27131832	2	*	HM-C7	PREB	theta	3.41E-09	0.007487	4.69	0.71	0.63	101.11
chr6	1.57E+08	1.57E+08	951	*	HM-M17	TMEM242	psi3	3.85E-09	0.003991	4.76	0.69	0.65	43.3
chr17	10459422	10459951	530	*	HM-M22	MYHAS	psi5	5.35E-09	0.001852	-4.42	0.12	-0.79	28.11
chr21	8214793	8259040	44248	*	HM-C14	LOC107987293	psi3	5.48E-09	0.001422	5.2	0.77	0.72	25.96
chr22	18906564	18906565	2	*	HM-M17	DGCR6	theta	7.49E-09	0.008226	3.68	0.44	0.43	309.26
chr6	83352140	83398366	46227	*	HM-C5	ME1	psi5	9.80E-09	0.009125	-5.75	0.63	-0.36	60.61
chr5	1.16E+08	1.16E+08	839	*	HM-C10	FEM1C	psi3	1.36E-08	0.007051	-5.65	0.54	-0.44	41.24
chr2	1.87E+08	1.87E+08	1236	*	HM-C7	ZC3H15	psi5	1.46E-08	0.015148	-4.28	0.53	-0.44	75.7
chr1	36179316	36179511	196	*	HM-C6	MAP7D1	psi5	1.66E-08	0.008605	4.61	0.51	0.43	154.83
chr1	77929504	77929505	2	*	HM-M15	NEXN	theta	1.82E-08	0.039905	4.57	0.58	0.48	475.13
chr1	2.26E+08	2.26E+08	2	*	HM-M5	ENAH	theta	2.55E-08	0.028028	3.84	0.81	0.7	60.04
chr22	29138783	29140281	1499	*	HM-M5	KREMEN1	psi5	2.79E-08	0.028963	-5.26	0.5	-0.47	39.41
chr6	1.57E+08	1.57E+08	68	*	HM-M23	TMEM242	psi3	3.01E-08	0.031259	3.47	0.88	0.65	33.54
chr19	6381616	6381696	81	*	HM-M19	GTF2F1	psi3	3.98E-08	0.04126	-4.82	0.46	-0.5	37.83
chr1	24112030	24112031	2	*	HM-M5	MYOM3	theta	6.77E-08	0.044669	3.82	0.57	0.48	192.7
chr3	1.57E+08	1.57E+08	2	*	HM-M5	CCNL1	theta	1.02E-07	0.044669	4.3	0.51	0.46	55

Supplementary Table 6. FRASER 1.0 results of rare alternatively spliced events in blood samples based on theta, psi5 and psi3

seqnames	start	end	width	strand	sampleID	hgncSymbol	type	pValue	padjust	zScore	psiValue	deltaPsi	meanTotalCounts
chr8	18069879	18075540	5662	*	5197756	ASAH1	psi3	5.25E-14	1.34E-08	6.25	0.55	0.5	25.67
chr6	6250930	6266557	15628	*	5092533	F13A1	psi3	2.80E-13	7.13E-08	-6.09	0.32	-0.59	17.8
chr2	95887998	95888091	94	*	9957921	ANKRD36C	psi5	2.33E-12	5.94E-07	-7.88	0	-0.73	5.82
chr17	10019776	10019777	2	*	1001530	GAS7	theta	6.75E-12	3.67E-06	5.61	0.83	0.64	6.04
chr19	48918785	48919029	245	*	1002051	NUCB1-AS1	psi5	8.47E-11	2.16E-05	-5.85	0.62	-0.36	19.28

chr4	10088742	10097710	8969	*	1001973	WDR1	psi5	2.15E-10	5.50E-05	-5.04	0.44	-0.49	17.91
chr5	1.34E+08	1.34E+08	3900	*	1001953	CDKL3	psi5	5.18E-10	0.000132	-5.32	0.2	-0.65	8.53
chr1	1.47E+08	1.47E+08	709	*	1001867	NBPF12	psi5	8.00E-10	0.000204	-5.63	0.08	-0.73	4.41
chr8	18069879	18075540	5662	*	5092533	ASAH1	psi3	1.01E-09	0.000258	4.88	0.36	0.32	25.67
chr1	1.49E+08	1.49E+08	720	*	1001859	NBPF14	psi5	1.12E-09	0.000287	-3.82	0.64	-0.34	17.59
chr7	55473046	55492281	19236	*	1001917	VOPP1	psi5	1.71E-09	0.000436	-5.94	0.57	-0.39	8.69
chr5	1.34E+08	1.34E+08	3900	*	1001953	CDKL3	psi5	1.88E-09	0.000481	-5.01	0.24	-0.62	8.53
chr14	1.06E+08	1.06E+08	274803	*	1001835	MIR4507	psi3	3.34E-09	0.000852	4.94	0.88	0.71	4.79
chr1	2.08E+08	2.08E+08	608	*	5092533	CR1	psi3	5.37E-09	0.00137	-3.92	0.4	-0.54	40.66
chr1	1.62E+08	1.62E+08	4042	*	1001957	FCGR2A	psi5	6.88E-09	0.001758	-3.3	0.33	-0.63	161
chr21	8260502	8399307	138806	*	1001953	LOC105379508	psi5	1.20E-08	0.003061	-5.54	0	-0.34	26.47
chr12	68824652	68828770	4119	*	1000415	MDM2	psi5	1.80E-08	0.004584	-4.09	0.44	-0.47	7.35
chr19	48919029	48919030	2	*	1002051	NUCB1	theta	1.89E-08	0.010264	5.05	0.45	0.42	19.56
chr21	8260502	8399307	138806	*	1001953	LOC105379508	psi5	2.73E-08	0.003482	-5.47	0	-0.34	26.47
chr1	1.56E+08	1.56E+08	3073	*	8004754	LOC105371729	psi5	2.88E-08	0.007343	-5.46	0.33	-0.57	19.8
chr1	1.49E+08	1.49E+08	2	*	1001684	NBPF14	theta	3.28E-08	0.017828	4.72	1	0.76	21.57
chr8	70296743	70296744	2	*	8812251	NCOA2	theta	3.35E-08	0.018211	3.29	0.72	0.49	5.97
chr1	1.62E+08	1.62E+08	4042	*	1001830	FCGR2A	psi5	3.89E-08	0.009942	-4.4	0.46	-0.53	161
chr18	9815115	9815116	2	*	1001953	RAB31	theta	5.15E-08	0.027969	6.75	0.94	0.76	6.41
chr12	1.25E+08	1.25E+08	228	*	1001917	UBC	psi5	6.72E-08	0.017151	4.8	1	0.83	29.46
chr21	8400671	8444903	44233	*	1001701	LOC105379508	psi5	9.20E-08	0.023496	-5.25	0.3	-0.55	5.28
chr1	6040645	6040646	2	*	1001877	KCNAB2	theta	9.46E-08	0.051407	4.56	0.83	0.64	6.5
chr19	7643049	7643164	116	*	1001866	STXBP2	psi5	9.85E-08	0.025157	-4.95	0.41	-0.52	9.28
chr4	1.84E+08	1.84E+08	2	*	1001867	IRF2	theta	1.17E-07	0.063327	4.88	0.67	0.56	8.98
chr2	2.18E+08	2.18E+08	133	*	1001905	CTDSP1	psi5	1.25E-07	0.031875	-6.07	0.57	-0.39	10.56
chr1	1.49E+08	1.49E+08	3903	*	1001957	NBPF19	psi5	1.52E-07	0.038762	-4.57	0.19	-0.66	16.41
chr3	1.59E+08	1.59E+08	2	*	1001866	MFSD1	theta	1.65E-07	0.089719	4.3	0.5	0.42	8.1
chr4	1.84E+08	1.84E+08	760	*	1001541	IRF2	psi3	1.72E-07	0.043766	-5.26	0.08	-0.65	9.29
chr10	14522020	14530331	8312	*	1001957	FAM107B	psi3	2.24E-07	0.057008	-4.66	0.54	-0.39	18.11
chr21	8259491	8397144	137654	*	8004754	LOC105379508	psi5	2.39E-07	0.061134	-6.03	0.24	-0.37	19.26
chr1	1.62E+08	1.62E+08	4042	*	1001957	FCGR2A	psi5	2.80E-07	0.071439	-3.05	0.4	-0.56	161

chr1	1.56E+08	1.56E+08	2	*	1001973	ARHGEF2	theta	3.02E-07	0.082126	3.43	0.53	0.4	17.79
chr1	1.62E+08	1.62E+08	4042	*	1001957	FCGR2A	psi5	3.21E-07	0.081952	-3.05	0.39	-0.57	161
chr1	1.49E+08	1.49E+08	719	*	1001917	NBPF19	psi5	3.31E-07	0.042239	-4.05	0.08	-0.75	8.34
chr8	1E+08	1E+08	12339	*	1001973	RNF19A	psi5	3.57E-07	0.045517	-5.93	0.17	-0.66	4.76
chr1	1.62E+08	1.62E+08	4042	*	1001957	FCGR2A	psi5	4.56E-07	0.058169	-3.05	0.39	-0.56	161

Supplementary Table 7. Statistically significant alternative splicing events within skeletal muscle and blood collections at the cohort level from rMATS and Jutils output

Skeletal Muscle

Gene	Event Locus	Strand	Feature types	Significance (q)	Mean ΔPSI
NEB	chr2:151494253,151497625-151497718,151505477	-	ES	1.1×10^{-9}	0.10
TNNT3	chr11:1925116,1925255-1925276,1926694	+	ES	1.7×10^{-7}	0.11
NEB	chr2:151531896,151533441-151533819:151534214-151534319,151535690	-	MXE	2.6×10^{-7}	-0.11
NEB	chr2:151531896,151533329-151533581:151534214-151534429,151535690	-	MXE	3.2×10^{-7}	-0.10
TNNT3	chr11:1923061,1923554-1923572:1925255-1925276,1926469	+	MXE	6.7×10^{-7}	-0.12
TNNT3	chr11:1925116,1925255-1925276:1926694-1926709,1929119	+	MXE	2.1×10^{-6}	0.12
MACF1	chr1:39463611-39463686:39468614-39468732	+	IR	2.6×10^{-6}	-0.24
NEB	chr2:151531896,151533117-151533546:151534214-151534319,151535690	-	MXE	1.1×10^{-5}	-0.13

Blood

Gene	Event Locus	Strand	Feature types	Significance (q)	Mean ΔPSI
WDR26	chr1:224411426-224411565:224413252-224413335	-	IR	5.9×10^{-12}	-0.16
SNRPB	chr20: 2465819-2467191	-	IR	1.6×10^{-6}	-0.11
AIF1	chr6:31615736,31616018-31616506:31616343-31616506	+	A3SS	2.9×10^{-5}	0.13
HNRNPA2B1	chr7:26191128,26191421-26191650,26192494	-	ES	4.0×10^{-4}	-0.12
OSTM1	chr6:108041408-108042264:108042566-108044840	-	IR	7.6×10^{-4}	-0.13
CLEC7A	chr12:10124745-10125029:10125296-10125448	-	IR	7.7×10^{-4}	0.15
OSTM1	chr6:108041408-108042264:108042566-108044840	-	IR	7.6×10^{-4}	-0.13

GNLY	chr2:85694752-85694923:85695319-85695423	+	IR	2.3×10^{-3}	0.14
ECHDC2	chr1:52905033-52905090:52906518-52906611	-	IR	2.4×10^{-3}	-0.13
CENATAC	chr11:119011220-119011311:119011220-119011283,119011938	+	A5SS	2.7×10^{-3}	-0.11
PIGG	chr4:523913,525170-525575,527038	+	ES	3.9×10^{-2}	0.10
ELOVL5	chr6:53276256,53277645-53277743:53290400-53290535,53291775	-	MXE	9.0×10^{-3}	0.11
EMC8	chr16:85779867,85780378-85780473,85781210	-	ES	1×10^{-2}	0.13

Note: Top statistically significant genes identified in skeletal muscle and blood samples are based on adjusted *p*-value < 0.05 and mean |ΔPSI| > 0.1, as per documentation by the authors of rMATS. Due to the high frequency of genes occurring based on cases vs controls, here we record the number of events for each gene locus based on significance of adjusted p-value and mean ΔPSI. Event loci consist of events of up and downstream exons it occurs between.

Supplementary Table 8. *UNC13A* splicing variant analysis in case-controls of intronic variants within tissue collections

Skeletal Muscle

Status	Sample/s	Position	Allele	Splice metric	Splice prediction	CE related	Population frequency
Control	C16	17632104	T > C	AL	0.01	N	7×10^{-5}
Case	M32	17677521	G > A	DG	0.01	N	0.45
Case	M11	17667138	T > C	DL	0.01	N	0.88
Case	M11	17667142	C > T	DL	0.01	N	7×10^{-6}

Blood

Status	Sample/s	Position	Allele	Splice metric	Splice prediction	CE related	Population frequency
Control	1001531	17641880	A > C	AL	0.01	Y	0.34
Control	1001531	17642430	C > G	DG	0.02	Y	0.22
Case	1001530	17641880	A > C	AL	0.01	Y	0.34
Case	1001689	17619357	T > C	DG	0.01	N	0.46
Case	1001689	17667138	T > C	DL	0.01	N	0.88
Case	1001689	17668023	A > G	AG	0.01	N	0.87

Note: Variants depicted above were the only documented variants consisting of splicing predictions. Population frequency was obtained from dbSNP and ClinVar database, based on the European sample group.

References

1. Aartsma-Rus, A. M. (2023). The future of exon skipping for Duchenne muscular dystrophy. *Hum Gene Ther*. DOI: 10.1089/hum.2023.026.
2. Alonso, A., Logroscino, G., Jick, S. S., & Hernan, M. A. (2009). Incidence and lifetime risk of motor neuron disease in the United Kingdom: a population-based study. *Eur J Neurol*, *16*(6), 745-751. DOI: 10.1111/j.1468-1331.2009.02586.x.
3. Arnold, E. S., Ling, S. C., Huelga, S. C., Lagier-Tourenne, C., Polymenidou, M., Ditsworth, D., Kordasiewicz, H. B., McAlonis-Downes, M., Platoshyn, O., Parone, P. A., Da Cruz, S., Clutario, K. M., Swing, D., Tessarollo, L., Marsala, M., Shaw, C. E., Yeo, G. W., & Cleveland, D. W. (2013). ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *Proc Natl Acad Sci U S A*, *110*(8), E736-745. DOI: 10.1073/pnas.1222809110.
4. Baxi, E. G., Thompson, T., Li, J., Kaye, J. A., Lim, R. G., Wu, J., Ramamoorthy, D., Lima, L., Vaibhav, V., Matlock, A., Frank, A., Coyne, A. N., Landin, B., Ornelas, L., Mosmiller, E., Thrower, S., Farr, S. M., Panther, L., Gomez, E., . . . Rothstein, J. D. (2022). Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical and multi-omics data from induced pluripotent cell lines. *Nat Neurosci*, *25*(2), 226-237. DOI: 10.1038/s41593-021-01006-0.
5. Benson, B. C., Shaw, P. J., Azzouz, M., Highley, J. R., & Hautbergue, G. M. (2021). Proteinopathies as Hallmarks of Impaired Gene Expression, Proteostasis and Mitochondrial Function in Amyotrophic Lateral Sclerosis. *Front Neurosci*, *15*, 783624. DOI: 10.3389/fnins.2021.783624.
6. Bernardini, C., Censi, F., Lattanzi, W., Barba, M., Calcagnini, G., Giuliani, A., Tasca, G., Sabatelli, M., Ricci, E., & Michetti, F. (2013). Mitochondrial network genes in the skeletal muscle of amyotrophic lateral sclerosis patients. *PLoS One*, *8*(2), e57739. DOI: 10.1371/journal.pone.0057739.
7. Boehm, F. J., & Zhou, X. (2022). Statistical methods for Mendelian randomization in genome-wide association studies: A review. *Comput Struct Biotechnol J*, *20*, 2338-2351. DOI: 10.1016/j.csbj.2022.05.015.
8. Brown, A. L., Wilkins, O. G., Keuss, M. J., Hill, S. E., Zanovello, M., Lee, W. C., Bampton, A., Lee, F. C. Y., Masino, L., Qi, Y. A., Bryce-Smith, S., Gatt, A., Hallegger, M., Fagegaltier, D., Phatnani, H., Consortium, N. A., Newcombe, J., Gustavsson, E. K., Seddighi, S., . . . Fratta, P. (2022). TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature*, *603*(7899), 131-137. DOI: 10.1038/s41586-022-04436-3.
9. Calvo, A. C., Cibreiro, G. A., Merino, P. T., Roy, J. F., Galiana, A., Rufian, A. J., Cano, J. M., Martin, M. A., Moreno, L., Larrode, P., Vazquez, P. C., Galan, L., Mora, J., Munoz-Blanco, J. L., Munoz, M. J., Zaragoza, P., Pegoraro, E., Soraru, G., Mora, M., . . . Redondo, A. G. (2019). Collagen XIX Alpha 1 Improves Prognosis in Amyotrophic Lateral Sclerosis. *Aging Dis*, *10*(2), 278-292. DOI: 10.14336/AD.2018.0917.
10. Campion, D., Charbonnier, C., & Nicolas, G. (2019). SORL1 genetic variants and Alzheimer disease risk: a literature review and meta-analysis of sequencing data. *Acta Neuropathol*, *138*(2), 173-186. DOI: 10.1007/s00401-019-01991-4.
11. Chio, A., Logroscino, G., Hardiman, O., Swingler, R., Mitchell, D., Beghi, E., Traynor, B. G., & Eurals, C. (2009). Prognostic factors in ALS: A critical review. *Amyotroph Lateral Scler*, *10*(5-6), 310-323. DOI: 10.3109/17482960802566824.
12. Christidi, F., Karavasilis, E., Velonakis, G., Ferentinos, P., Rentzos, M., Kelekis, N., Evdokimidis, I., & Bede, P. (2018). The Clinical and Radiological Spectrum of Hippocampal Pathology in Amyotrophic Lateral Sclerosis. *Front Neurol*, *9*, 523. DOI: 10.3389/fneur.2018.00523.
13. Consortium, G. T. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, *45*(6), 580-585. DOI: 10.1038/ng.2653.
14. Daguenet, E., Dujardin, G., & Valcarcel, J. (2015). The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO Rep*, *16*(12), 1640-1655. DOI: 10.15252/embr.201541116.

15. Dawes, R., Bournazos, A. M., Bryen, S. J., Bommireddipalli, S., Marchant, R. G., Joshi, H., & Cooper, S. T. (2023). SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nat Genet*, *55*(2), 324-332. DOI: 10.1038/s41588-022-01293-8.
16. de Sainte Agathe, J. M., Filser, M., Isidor, B., Besnard, T., Gueguen, P., Perrin, A., Van Goethem, C., Verebi, C., Masingue, M., Rendu, J., Cossee, M., Bergougnoux, A., Frobert, L., Buratti, J., Lejeune, E., Le Guern, E., Pasquier, F., Clot, F., Kalatzis, V., . . . Baux, D. (2023). SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. *Hum Genomics*, *17*(1), 7. DOI: 10.1186/s40246-023-00451-1.
17. Deshaies, J. E., Shkreta, L., Moszczynski, A. J., Sidibe, H., Semmler, S., Fouillen, A., Bennett, E. R., Bekenstein, U., Destroismaisons, L., Toutant, J., Delmotte, Q., Volkening, K., Stabile, S., Aulas, A., Khalfallah, Y., Soreq, H., Nanci, A., Strong, M. J., Chabot, B., & Vande Velde, C. (2018). TDP-43 regulates the alternative splicing of hnRNP A1 to yield an aggregation-prone variant in amyotrophic lateral sclerosis. *Brain*, *141*(5), 1320-1333. DOI: 10.1093/brain/awy062.
18. Diekstra, F. P., Van Deerlin, V. M., van Swieten, J. C., Al-Chalabi, A., Ludolph, A. C., Weishaupt, J. H., Hardiman, O., Landers, J. E., Brown, R. H., Jr., van Es, M. A., Pasterkamp, R. J., Koppers, M., Andersen, P. M., Estrada, K., Rivadeneira, F., Hofman, A., Uitterlinden, A. G., van Damme, P., Melki, J., . . . Veldink, J. H. (2014). C9orf72 and UNC13A are shared risk loci for amyotrophic lateral sclerosis and frontotemporal dementia: a genome-wide meta-analysis. *Ann Neurol*, *76*(1), 120-133. DOI: 10.1002/ana.24198.
19. Donnelly, C. J., Grima, J. C., & Sattler, R. (2014). Aberrant RNA homeostasis in amyotrophic lateral sclerosis: potential for new therapeutic targets? *Neurodegener Dis Manag*, *4*(6), 417-437. DOI: 10.2217/nmt.14.36.
20. Florea, L., Song, L., & Salzberg, S. L. (2013). Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. *F1000Res*, *2*, 188. DOI: 10.12688/f1000research.2-188.v2.
21. Freydenzon, A., Wani, S., Bharti, V., Wallace, L. M., Henders, A. K., McCombe, P. A., Henderson, R. D., Steyn, F. J., Wray, N. R., Ngo, S. T., & McRae, A. F. (2023). RNA-seq analysis of skeletal muscle in motor neurone disease cases and controls. *medrxiv*. DOI: 10.1101/2023.03.13.23287229.
22. Gao, F., Zhang, J., Ni, T., Lin, N., Lin, H., Luo, H., Guo, H., & Chi, J. (2020). Herpud1 deficiency could reduce amyloid-beta40 expression and thereby suppress homocysteine-induced atherosclerosis by blocking the JNK/AP1 pathway. *J Physiol Biochem*, *76*(3), 383-391. DOI: 10.1007/s13105-020-00741-5.
23. Garcia-Arroyo, R., Marfany, G., & Mirra, S. (2022). CERKL, a Retinal Dystrophy Gene, Regulates Mitochondrial Transport and Dynamics in Hippocampal Neurons. *Int J Mol Sci*, *23*(19). DOI: 10.3390/ijms231911593.
24. Gendron, T. F., Rademakers, R., & Petrucelli, L. (2013). TARDBP mutation analysis in TDP-43 proteinopathies and deciphering the toxicity of mutant TDP-43. *J Alzheimers Dis*, *33 Suppl 1*(Suppl 1), S35-45. DOI: 10.3233/JAD-2012-129036.
25. Goutman, S. A., Hardiman, O., Al-Chalabi, A., Chio, A., Savelieff, M. G., Kiernan, M. C., & Feldman, E. L. (2022). Emerging insights into the complex genetics and pathophysiology of amyotrophic lateral sclerosis. *Lancet Neurol*, *21*(5), 465-479. DOI: 10.1016/S1474-4422(21)00414-2.
26. Gudmundsson, S., Singer-Berk, M., Watts, N. A., Phu, W., Goodrich, J. K., Solomonson, M., Genome Aggregation Database, C., Rehm, H. L., MacArthur, D. G., & O'Donnell-Luria, A. (2022). Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat*, *43*(8), 1012-1030. DOI: 10.1002/humu.24309.
27. Hahn, A., Kny, M., Pablo-Tortola, C., Todiras, M., Willenbrock, M., Schmidt, S., Schmoeckel, K., Jorde, I., Nowak, M., Jarosch, E., Sommer, T., Broker, B. M., Felix, S. B., Scheidereit, C., Weber-Carstens, S., Butter, C., Luft, F. C., & Fielitz, J. (2020). Serum amyloid A1 mediates myotube atrophy via Toll-like receptors. *J Cachexia Sarcopenia Muscle*, *11*(1), 103-119. DOI: 10.1002/jcsm.12491.
28. Halim, D., & Gao, F. B. (2022). RNA targets of TDP-43: Which one is more important in neurodegeneration? *Transl Neurodegener*, *11*(1), 12. DOI: 10.1186/s40035-021-00268-9.
29. Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E. M., Logroscino, G., Robberecht, W., Shaw, P. J., Simmons, Z., & van den Berg, L. H. (2017). Amyotrophic lateral sclerosis. *Nat Rev Dis Primers*, *3*, 17085. DOI: 10.1038/nrdp.2017.85.

30. Honda, H., Hamasaki, H., Wakamiya, T., Koyama, S., Suzuki, S. O., Fujii, N., & Iwaki, T. (2015). Loss of hnRNPA1 in ALS spinal cord motor neurons with TDP-43-positive inclusions. *Neuropathology*, *35*(1), 37-43. DOI: 10.1111/neup.12153.
31. Indrayan, A., & Mishra, A. (2021). The importance of small samples in medical research. *J Postgrad Med*, *67*(4), 219-223. DOI: 10.4103/jpgm.JPGM_230_21.
32. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., & Farh, K. K. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, *176*(3), 535-548 e524. DOI: 10.1016/j.cell.2018.12.015.
33. Jenkinson, G., Li, Y. I., Basu, S., Cousin, M. A., Oliver, G. R., & Klee, E. W. (2020). LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics*, *36*(17), 4609-4615. DOI: 10.1093/bioinformatics/btaa259.
34. Jiang, W., & Chen, L. (2021). Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. *Comput Struct Biotechnol J*, *19*, 183-195. DOI: 10.1016/j.csbj.2020.12.009.
35. Kim, H. J., Kim, N. C., Wang, Y. D., Scarborough, E. A., Moore, J., Diaz, Z., MacLea, K. S., Freibaum, B., Li, S., Molliex, A., Kanagaraj, A. P., Carter, R., Boylan, K. B., Wojtas, A. M., Rademakers, R., Pinkus, J. L., Greenberg, S. A., Trojanowski, J. Q., Traynor, B. J., . . . Taylor, J. P. (2013). Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature*, *495*(7442), 467-473. DOI: 10.1038/nature11922.
36. Kim, H. J., Mohassel, P., Donkervoort, S., Guo, L., O'Donovan, K., Coughlin, M., Lornage, X., Foulds, N., Hammans, S. R., Foley, A. R., Fare, C. M., Ford, A. F., Ogasawara, M., Sato, A., Iida, A., Munot, P., Ambegaonkar, G., Phadke, R., O'Donovan, D. G., . . . Taylor, J. P. (2022). Heterozygous frameshift variants in HNRNPA2B1 cause early-onset oculopharyngeal muscular dystrophy. *Nat Commun*, *13*(1), 2306. DOI: 10.1038/s41467-022-30015-1.
37. Kumar, R., & Haider, S. (2022). Protein network analysis to prioritize key genes in amyotrophic lateral sclerosis. *IBRO Neurosci Rep*, *12*, 25-44. DOI: 10.1016/j.ibneur.2021.12.002.
38. Kyriakou, K., Lederer, C. W., Kleanthous, M., Drousiotou, A., & Malekkou, A. (2020). Acid Ceramidase Depletion Impairs Neuronal Survival and Induces Morphological Defects in Neurites Associated with Altered Gene Transcription and Sphingolipid Content. *Int J Mol Sci*, *21*(5). DOI: 10.3390/ijms21051607.
39. Landrum, M. J., & Kattman, B. L. (2018). ClinVar at five years: Delivering on the promise. *Hum Mutat*, *39*(11), 1623-1630. DOI: 10.1002/humu.23641.
40. Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., & Pritchard, J. K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet*, *50*(1), 151-158. DOI: 10.1038/s41588-017-0004-9.
41. Ma, X. R., Prudencio, M., Koike, Y., Vatsavayai, S. C., Kim, G., Harbinski, F., Briner, A., Rodriguez, C. M., Guo, C., Akiyama, T., Schmidt, H. B., Cummings, B. B., Wyatt, D. W., Kurylo, K., Miller, G., Mekhoubad, S., Sallee, N., Mekonnen, G., Ganser, L., . . . Gitler, A. D. (2022). TDP-43 represses cryptic exon inclusion in the FTD-ALS gene UNC13A. *Nature*, *603*(7899), 124-130. DOI: 10.1038/s41586-022-04424-7.
42. Mamoor, S. (2022). Differential Expression of CDKN1A in Amyotrophic Lateral Sclerosis. *OSF Preprints*. DOI: 10.31219/osf.io/zj3bc.
43. Mead, R. J., Shan, N., Reiser, H. J., Marshall, F., & Shaw, P. J. (2023). Amyotrophic lateral sclerosis: a neurodegenerative disorder poised for successful therapeutic translation. *Nat Rev Drug Discov*, *22*(3), 185-212. DOI: 10.1038/s41573-022-00612-2.
44. Mejzini, R., Flynn, L. L., Pitout, I. L., Fletcher, S., Wilton, S. D., & Akkari, P. A. (2019). ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Front Neurosci*, *13*, 1310. DOI: 10.3389/fnins.2019.01310.
45. Meneses, A., Koga, S., O'Leary, J., Dickson, D. W., Bu, G., & Zhao, N. (2021). TDP-43 Pathology in Alzheimer's Disease. *Mol Neurodegener*, *16*(1), 84. DOI: 10.1186/s13024-021-00503-x.
46. Mertes, C., Scheller, I. F., Yepez, V. A., Celik, M. H., Liang, Y., Kremer, L. S., Gusic, M., Prokisch, H., & Gagneur, J. (2022). Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat Commun*, *13*(1), 3474. DOI: 10.1038/s41467-022-31242-2.
47. Moffat, J. J., Ka, M., Jung, E. M., Smith, A. L., & Kim, W. Y. (2017). The role of MACF1 in nervous system development and maintenance. *Semin Cell Dev Biol*, *69*, 9-17. DOI: 10.1016/j.semcdb.2017.05.020.

48. Muller, I. B., Meijers, S., Kampstra, P., van Dijk, S., van Elswijk, M., Lin, M., Wojtuszkiewicz, A. M., Jansen, G., de Jonge, R., & Cloos, J. (2021). Computational comparison of common event-based differential splicing tools: practical considerations for laboratory researchers. *BMC Bioinformatics*, *22*(1), 347. DOI: 10.1186/s12859-021-04263-9.
49. Navarro-Marquez, M., Torrealba, N., Troncoso, R., Vasquez-Trincado, C., Rodriguez, M., Morales, P. E., Villalobos, E., Eura, Y., Garcia, L., Chiong, M., Klip, A., Jaimovich, E., Kokame, K., & Lavandero, S. (2018). Herpud1 impacts insulin-dependent glucose uptake in skeletal muscle cells by controlling the Ca(2+)-calcineurin-Akt axis. *Biochim Biophys Acta Mol Basis Dis*, *1864*(5 Pt A), 1653-1662. DOI: 10.1016/j.bbadis.2018.02.018.
50. Ogawa, K., Tanaka, K., Ishii, A., Nakamura, Y., Kondo, S., Sugamura, K., Takano, S., Nakamura, M., & Nagata, K. (2001). A novel serum protein that is selectively produced by cytotoxic lymphocytes. *J Immunol*, *166*(10), 6404-6412. DOI: 10.4049/jimmunol.166.10.6404.
51. Ono, S., Matsuda, J., Watanabe, E., Akaike, H., Teranishi, H., Miyata, I., Otomo, T., Sadahira, Y., Mizuochi, T., Kusano, H., Kage, M., Ueno, H., Yoshida, K., Shiraishi, Y., Chiba, K., Tanaka, H., Miyano, S., Ogawa, S., Hayashi, Y., . . . Ouchi, K. (2019). Novel neuroblastoma amplified sequence (NBAS) mutations in a Japanese boy with fever-triggered recurrent acute liver failure. *Hum Genome Var*, *6*, 2. DOI: 10.1038/s41439-018-0035-5.
52. Park, J. W., Tokheim, C., Shen, S., & Xing, Y. (2013). Identifying differential alternative splicing events from RNA sequencing data using RNASeq-MATS. *Methods Mol Biol*, *1038*, 171-179. DOI: 10.1007/978-1-62703-514-9_10.
53. Pohl, M., Bortfeldt, R. H., Grutzmann, K., & Schuster, S. (2013). Alternative splicing of mutually exclusive exons--a review. *Biosystems*, *114*(1), 31-38. DOI: 10.1016/j.biosystems.2013.07.003.
54. Rojas, P., Ramirez, A. I., Fernandez-Albarral, J. A., Lopez-Cuenca, I., Salobrar-Garcia, E., Cadena, M., Elvira-Hurtado, L., Salazar, J. J., de Hoz, R., & Ramirez, J. M. (2020). Amyotrophic Lateral Sclerosis: A Neurodegenerative Motor Neuron Disease With Ocular Involvement. *Front Neurosci*, *14*, 566858. DOI: 10.3389/fnins.2020.566858.
55. Rutherford, N. J., Zhang, Y. J., Baker, M., Gass, J. M., Finch, N. A., Xu, Y. F., Stewart, H., Kelley, B. J., Kuntz, K., Crook, R. J., Sreedharan, J., Vance, C., Sorenson, E., Lippa, C., Bigio, E. H., Geschwind, D. H., Knopman, D. S., Mitsumoto, H., Petersen, R. C., . . . Rademakers, R. (2008). Novel mutations in TARDBP (TDP-43) in patients with familial amyotrophic lateral sclerosis. *PLoS Genet*, *4*(9), e1000193. DOI: 10.1371/journal.pgen.1000193.
56. Scheller, I. F., Mertes, C., Yepez, V. A., & Gagneur, J. (2023). Improved detection of aberrant splicing using the Intron Jaccard Index. *medrxiv*. DOI: 10.1101/2023.03.31.23287997
57. Schimmelpennink, M. C., Vorselaars, A. D. M., & Grutters, J. C. (2019). Chapter 19 - Biomarkers in Sarcoidosis. In R. P. Baughman & D. Valeyre (Eds.), *Sarcoidosis* (pp. 219-238). Elsevier. DOI: 10.1016/B978-0-323-54429-0.00019-7
58. Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., & Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*, *111*(51), E5593-5601. DOI: 10.1073/pnas.1419161111.
59. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, *29*(1), 308-311. DOI: 10.1093/nar/29.1.308.
60. Soric Hosman, I., Kos, I., & Lamot, L. (2020). Serum Amyloid A in Inflammatory Rheumatic Diseases: A Compendious Review of a Renowned Biomarker. *Front Immunol*, *11*, 631299. DOI: 10.3389/fimmu.2020.631299.
61. Suk, T. R., & Rousseaux, M. W. C. (2020). The role of TDP-43 mislocalization in amyotrophic lateral sclerosis. *Mol Neurodegener*, *15*(1), 45. DOI: 10.1186/s13024-020-00397-1.
62. Susnjar, U., Skrabar, N., Brown, A. L., Abbassi, Y., Phatnani, H., Consortium, N. A., Cortese, A., Cereda, C., Bugiardini, E., Cardani, R., Meola, G., Ripolone, M., Moggio, M., Romano, M., Secrier, M., Fratta, P., & Buratti, E. (2022). Cell environment shapes TDP-43 function with implications in neuronal and muscle disease. *Commun Biol*, *5*(1), 314. DOI: 10.1038/s42003-022-03253-8.

63. Taetzsch, T., Brayman, V. L., & Valdez, G. (2018). FGF binding proteins (FGFBPs): Modulators of FGF signaling in the developing, adult, and stressed nervous system. *Biochim Biophys Acta Mol Basis Dis*, *1864*(9 Pt B), 2983-2991. DOI: 10.1016/j.bbadis.2018.06.009.
64. Tamaki, Y., Ross, J. P., Alipour, P., Castonguay, C. E., Li, B., Catoire, H., Rochefort, D., Urushitani, M., Takahashi, R., Sonnen, J. A., Stifani, S., Dion, P. A., & Rouleau, G. A. (2023). Spinal cord extracts of amyotrophic lateral sclerosis spread TDP-43 pathology in cerebral organoids. *PLoS Genet*, *19*(2), e1010606. DOI: 10.1371/journal.pgen.1010606.
65. Tan, Y., Jin, Y., Wu, X., & Ren, Z. (2019). PSMD1 and PSMD2 regulate HepG2 cell proliferation and apoptosis via modulating cellular lipid droplet metabolism. *BMC Mol Biol*, *20*(1), 24. DOI: 10.1186/s12867-019-0141-z.
66. Thorolfsdottir, R. B., Sveinbjornsson, G., Sulem, P., Nielsen, J. B., Jonsson, S., Halldorsson, G. H., Melsted, P., Ivarsdottir, E. V., Davidsson, O. B., Kristjansson, R. P., Thorleifsson, G., Helgadottir, A., Gretarsdottir, S., Norddahl, G., Rajamani, S., Torfason, B., Valgardsson, A. S., Sverrisson, J. T., Tragante, V., . . . Stefansson, K. (2018). Coding variants in RPL3L and MYZAP increase risk of atrial fibrillation. *Commun Biol*, *1*, 68. DOI: 10.1038/s42003-018-0068-9.
67. Turner, M. R., & Group, U. M. C. S. (2022). Diagnosing ALS: the Gold Coast criteria and the role of EMG. *Pract Neurol*, *22*(3), 176-178. DOI: 10.1136/practneurol-2021-003256.
68. Umoh, M. E., Fournier, C., Li, Y., Polak, M., Shaw, L., Landers, J. E., Hu, W., Gearing, M., & Glass, J. D. (2016). Comparative analysis of C9orf72 and sporadic disease in an ALS clinic population. *Neurology*, *87*(10), 1024-1030. DOI: 10.1212/WNL.0000000000003067.
69. Van Deerlin, V. M., Leverenz, J. B., Bekris, L. M., Bird, T. D., Yuan, W., Elman, L. B., Clay, D., Wood, E. M., Chen-Plotkin, A. S., Martinez-Lage, M., Steinbart, E., McCluskey, L., Grossman, M., Neumann, M., Wu, I. L., Yang, W. S., Kalb, R., Galasko, D. R., Montine, T. J., . . . Yu, C. E. (2008). TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: a genetic and histopathological analysis. *Lancet Neurol*, *7*(5), 409-416. DOI: 10.1016/S1474-4422(08)70071-1.
70. Wagner, N., Celik, M. H., Holzlwimmer, F. R., Mertes, C., Prokisch, H., Yepez, V. A., & Gagneur, J. (2023). Aberrant splicing prediction across human tissues. *Nat Genet*, *55*(5), 861-870. DOI: 10.1038/s41588-023-01373-3.
71. Wang, X., Li, N., Xiong, N., You, Q., Li, J., Yu, J., Qing, H., Wang, T., Cordell, H. J., Isacson, O., Vance, J. M., Martin, E. R., Zhao, Y., Cohen, B. M., Buttner, E. A., & Lin, Z. (2017). Genetic Variants of Microtubule Actin Cross-linking Factor 1 (MACF1) Confer Risk for Parkinson's Disease. *Mol Neurobiol*, *54*(4), 2878-2888. DOI: 10.1007/s12035-016-9861-y.
72. Wang, Y., Liu, J., Huang, B. O., Xu, Y. M., Li, J., Huang, L. F., Lin, J., Zhang, J., Min, Q. H., Yang, W. M., & Wang, X. Z. (2015). Mechanism of alternative splicing and its regulation. *Biomed Rep*, *3*(2), 152-158. DOI: 10.3892/br.2014.407.
73. Willemse, S. W., Harley, P., van Eijk, R. P. A., Demaegd, K. C., Zelina, P., Pasterkamp, R. J., van Damme, P., Ingre, C., van Rheenen, W., Veldink, J. H., Kiernan, M. C., Al-Chalabi, A., van den Berg, L. H., Fratta, P., & van Es, M. A. (2023). UNC13A in amyotrophic lateral sclerosis: from genetic association to therapeutic target. *J Neurol Neurosurg Psychiatry*. DOI: 10.1136/jnnp-2022-330504.
74. Williams, T. L. (2013). Motor neurone disease: diagnostic pitfalls. *Clin Med (Lond)*, *13*(1), 97-100. DOI: 10.7861/clinmedicine.13-1-97.
75. Wingo, T. S., Cutler, D. J., Yarab, N., Kelly, C. M., & Glass, J. D. (2011). The heritability of amyotrophic lateral sclerosis in a clinically ascertained United States research registry. *PLoS One*, *6*(11), e27985. DOI: 10.1371/journal.pone.0027985.
76. Zheng, Q., Huang, T., Zhang, L., Zhou, Y., Luo, H., Xu, H., & Wang, X. (2016). Dysregulation of Ubiquitin-Proteasome System in Neurodegenerative Diseases. *Front Aging Neurosci*, *8*, 303. DOI: 10.3389/fnagi.2016.00303.
77. Zhou, Y., Liu, S., Liu, G., Ozturk, A., & Hicks, G. G. (2013). ALS-associated FUS mutations result in compromised FUS alternative splicing and autoregulation. *PLoS Genet*, *9*(10), e1003895. DOI: 10.1371/journal.pgen.1003895.
78. Zhou, Z., Xu, H., Li, Y., Yang, M., Zhang, R., Shiraishi, A., Kiyonari, H., Liang, X., Huang, X., Wang, Y., Xie, Q., Liu, S., Chen, R., Bao, L., Guo, W., Wang, Y., & Meng, W. (2020). CAMSAP1 breaks the homeostatic

microtubule network to instruct neuronal polarity. *Proc Natl Acad Sci U S A*, *117*(36), 22193-22203. DOI: 10.1073/pnas.1913177117.

79. Ziff, O. J., Neeves, J., Mitchell, J., Tyzack, G., Martinez-Ruiz, C., Luisier, R., Chakrabarti, A. M., McGranahan, N., Litchfield, K., Boulton, S. J., Al-Chalabi, A., Kelly, G., Humphrey, J., & Patani, R. (2023). Integrated transcriptome landscape of ALS identifies genome instability linked to TDP-43 pathology. *Nat Commun*, *14*(1), 2176. DOI: 10.1038/s41467-023-37630-6.

80. Ziff, O. J., Neeves, J., Mitchell, J., Tyzack, G., McGranahan, N., Luisier, R., Chakrabarti, A. M., Boulton, S. J., Kelly, G., Humphrey, J., & Patani, R. (2022). Meta-analysis of the amyotrophic lateral sclerosis spectrum uncovers genome instability. *medrxiv*. DOI: 10.1101/2022.08.11.22278516.