RNA-seq analysis of skeletal muscle in motor neurone disease

Anna Freydenzon¹,*, Shivangi Wani¹, Vanda Bharti¹, Leanne M. Wallace¹,
Anjali K. Henders¹, Pamela A. McCombe²,³, Robert D. Henderson²,³, Frederik
J. Steyn²,⁴, Naomi R. Wray¹,⁵, Shyuan T. Ngo²,³,⁵,⁶,⁺, Allan F. McRae¹,⁺

¹ Institute for Molecular Bioscience; The University of Queensland; St Lucia,
QLD, 4072; Australia
² Centre for Clinical Research (CCR); The University of Queensland; Herston,
QLD, 4029; Australia
³ Department of Neurology; Royal Brisbane and Women’s Hospital; Brisbane,
QLD, 4029; Australia
⁴ The School of Biomedical Sciences; The University of Queensland;
Brisbane, QLD, 4072; Australia
⁵ Queensland Brain Institute; The University of Queensland; Brisbane, QLD,
4072; Australia
⁶ Australian Institute for Bioengineering and Nanotechnology; The University
of Queensland; St Lucia, QLD, 4072; Australia

⁺ These authors contributed equally

* Correspondence: a.freydenzon@uq.edu.au

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

**Background**: Amyotrophic lateral sclerosis (ALS), the most predominant form of Motor Neuron Disease (MND), is a progressive and fatal neurodegenerative condition that spreads throughout the neuromotor system by afflicting upper and lower motor neurons. Lower motor neurons project from the central nervous system and innervate muscle fibres at motor endplates, which degrade over the course of the disease leading to muscle weakness. The direction of neurodegeneration from or to the point of neuromuscular junctions and the role of muscle itself in pathogenesis has continued to be a topic of debate in ALS research.

**Methods**: To assess the variation in gene expression between affected and nonaffected muscle tissue that might lead to this local degeneration of motor units, we generated RNA-seq skeletal muscle transcriptomes from 28 MND cases and 18 healthy controls and conducted differential expression analyses on gene-level counts, as well as an isoform switching analysis on isoform-level counts.
Results: We identified 52 differentially-expressed genes (Benjamini-Hochberg-adjusted \( p < 0.05 \)) within this comparison, including 38 protein coding, 9 long non-coding RNA, and 5 pseudogenes. Of protein-coding genes, 31 were upregulated in cases including with notable genes including the collagenic \( \text{COL25A1} \) \((p = 3.1 \times 10^{-10})\), \( \text{SAA1} \) which is released in response to tissue injury \((p = 3.6 \times 10^{-5})\) as well as others of the SAA family, and the actin-encoding \( \text{ACTC1} \) \((p = 2.3 \times 10^{-5})\). Additionally, we identified 17 genes which exhibited a functional isoform switch with likely functional consequences between cases and controls.

Conclusions: Our analyses provide evidence of increased tissue generation in MND cases, which likely serve to compensate for the degeneration of motor units and skeletal muscle.

Keywords
Amyotrophic lateral sclerosis, ALS, motor neurone disease, transcriptomics, RNA-seq, differential expression, alternative splicing

Background
Motor neuron disease (MND) is a group of several similar neurodegenerative disorders with highly variable patterns of onset, progression, and symptomology. Amyotrophic lateral sclerosis (ALS), the most predominant motor neuron disease, is identified as featuring both upper and lower motor neuron loss. There is active discussion as to whether MND operates as a ‘dying forward’ phenomenon where initial cell death occurs in the cortical motor neurons and extends down into the corticospinal projections (1), or a ‘dying back’ phenomenon where neuromuscular junctions (NMJs) are the first to suffer failure and the pathology proceeds up the axon back into the body of the motor neuron (2). Nevertheless, the progressive loss of upper and lower motor neurons results in the presentation of clinical
features including spasticity and brisk reflexes, and muscle weakness and wasting across multiple body regions (3). In MND, muscle weakness and wasting occurs in response to the dismantling of neuromuscular junctions (NMJs) and axonal retraction that ultimately results in the denervation of skeletal muscle and subsequent die-off (4). Compensatory mechanisms can be seen in surviving motor units; axon terminals undergo sprouting to markedly increase reinnervation of skeletal muscle, and this occurs in parallel with abnormally fast firing (5).

MND is a highly polygenic disease consisting of over 20 known associated genes, although only ~10% of patients have familial history associated with mutations of large effect (4). One functional mechanism implicated by MND-associated genes is disturbance in RNA metabolism, which has contributed to an increasing research focus exploring transcriptomic machinery (6). One of the most notable genes is TDP-43, which undergoes substantial posttranslational modification and aggregation in both MND and frontotemporal dementia (7), leading to targeted RNA instability in its overexpression that results in consequences for ribosome biogenesis and oxidative phosphorylation pathways (8). Similarly, FUS is involved in transcriptional regulation through binding to RNA targets to regulate splicing and the removal of minor introns (9). Recent transcriptome-wide studies have been conducted on a variety of different tissues and models to find aberrant differential gene expression reflected in all tested tissue types, including but not limited to the postmortem motor cortex (10), patient whole blood (11), motor neurons from stem-cell derived spinal motor axons (12), and spinal cords from the recently generated PFN1 mouse model (13). Here, we report a whole transcriptome analysis of skeletal muscle tissue from MND patients, and describe the compensatory mechanisms invoked within MND-affected muscle tissue in response to denervation and local cell death.
Methods

Patient recruitment

Patient recruitment and muscle biopsies were undertaken as described by Steyn et al. 2020 (14). Briefly, MND patients who met the revised El-Escorial criteria for ALS (15) at the time of diagnosis were enrolled from the Royal Brisbane and Women’s Hospital MND clinic for the collection of skeletal muscle biopsies, along with healthy controls in the form of relatives, friends or community volunteers. For all participants, exclusion criteria were history of a metabolic condition (e.g. Hashimoto’s disease) and diabetes mellitus. Participants involved in the study are described in Table S1. Muscle biopsies (~200mg) were collected from the vastus lateralis from one leg (the least affected leg for MND participants) using a hollow Bergstrom biopsy needle modified for suction. Samples were placed in Dulbecco’s Modified EagleMedium/Ham’s F-12 with 0.5% gentamicin for transport to the laboratory. A portion (~50-100mg) of muscle tissue was divided from the primary biopsy, washed three times with room temperature phosphate buffered saline, frozen on dry ice, and stored at -80 degrees Celsius.

RNA sequencing

RNA sequencing was performed on MND and healthy controls for both blood and muscle samples (36 overlapping; 23 cases, 13 controls). RNaseq was performed first on an initial batch with all blood and muscle samples, followed by the same muscle samples being re-run once (split across two additional batches) to increase total sequencing reads.

Muscle

Muscle samples passing QC were contributed by 17 healthy controls and 28 MND cases.
Total RNA was extracted from the muscle tissue using a standard Trizol-chloroform extraction method. Post extraction, the total RNA was quantitated using the Epoch microplate spectrophotometer followed by quality assessment on the Agilent Bioanalyzer using the RNA Nano kit. On average the samples yielded 7.4 micrograms of total RNA with an average RNA Integrity Number (RIN) of 8.2.

One microgram of DNase1 treated total RNA was then used to prepare libraries for sequencing. The libraries were prepared using Illumina’s Truseq stranded total RNA library prep gold kit in conjunction with the TruSeq RNA UD indexes. The libraries were then equimolar pooled and paired end sequencing was carried out on HiSeq 4000 system.

**Blood**

Blood samples passing QC were contributed to by 13 healthy controls and 24 MND cases. Approximately 2.5mL of whole blood was collected in PAXgene Blood RNA tubes. Total RNA was extracted from these tubes using the PAXgene Blood RNA kit from Qiagen. Post extraction, the total RNA was quantitated using the Nanodrop spectrophotometer followed by quality assessment on the Agilent Bioanalyzer using the RNA Nano kit. On average the samples yielded 4.8 micrograms of total RNA with an average RNA Integrity Number (RIN) of 8.2.

500ng of DNase1 treated total RNA was then used to prepare libraries for sequencing. The libraries were prepared using Illumina Truseq stranded total RNA library prep gold kit in conjunction with the TruSeq RNA UD indexes. The libraries were then equimolar pooled and paired end sequencing was carried out on HiSeq 4000 system.

**Read processing**

Initial QC and read trimming was undertaken using fastp (16), setting a threshold of 1 to pass
through all non-empty reads (aberrant reads being filtered in the alignment step). By default, fastp conducts automatic adapter trimming by overlapping read pairs, as well as N-based qualify filtering and polyG tail trimming.

Trimming reads were mapped to the GENCODE v32-indexed genome through STAR 2.7.3 (17) with non-default parameters suggested by the STAR 2.7.0 manual for ENCODE alignment. These parameters specify that splices are only recognized if they fall over indexed, known splice junctions, sets a maximum number of mismatches to be read length* 0.04, limits genomic distance between read mates to 1,000,000 and sets a minimum intron length of 20bp.

The STAR-generated genome alignments were then quantified to gene meta-features through Rsubread featurecounts (18), with the parameter requiring both pairs to be mapped. Samples with alignment counts of < 20,000,000 across summed technical replicates were consequently filtered out of the analysis (4 blood, 1 muscle).

In addition, Salmon 1.0 (19) was run in quasi-mapping mode to the GENCODE v32 transcriptome (indexed over k-mer length 17). This was done to verify signals detected in STAR and act as a definitive source of transcripts as compared to the finer STAR genomic alignment method. This was run with the non-default parameters of validateMappings, seqBias, gcBias and useEM. This automatically estimates the library type (paired), runs in selective alignment mode where it ranks potential alignment sites based on mapping score, corrects for hexamer priming and GC biases and uses the expectation maximization (EM) method for abundance estimates.

Tximport 1.20 (20) was used to aggregate the Salmon transcript counts across samples to acquire both the summarized gene counts for the analysis complimentary to STAR, as well as
the counts of the transcripts for differential isoform analyses.

Differential expression analyses

STAR-generated read counts were read into DESeq2 1.30.1 (21) to collapse replicates into a single measure, and to estimate size factors and dispersion factors. Genes with ≥ 3 samples with normalized counts ≤ 5 were pruned from the individual blood and muscle sets. The resulting subset of genes were then run through DESeq2’s negative binomial GLM with a Wald significance test, increasing maximum iterations to 5,000.

The muscle and blood GLMs were fitted with MND as the condition and age and sex as covariates. Additionally, a binary variable indicating whether the samples were replicated across the second or third batch was added to the muscle model. These preliminary models were designated as the ‘base’ models.

The variation in counts was visualized within the muscle and blood samples through a principal component analysis applied to the 500 genes with highest variance across all samples, as well as through plotting the relative log expression on both centered and non-centered counts. Two control muscle samples were consequently dropped from further analysis due to featuring highly irregular counts and variance.

The genomic inflation factor (GIF) for each analysis (blood and muscle) was calculated as the median of the $\chi^2$ test statistic divided by the median of the $\chi^2$ distribution with one degree of freedom (22). The GIF is a summary statistic that compares the distribution of test statistics against expectation under the null hypothesis of no association of gene expression with MND and has an expected value of 1 when confounding is absent from the data, under the assumption that at least 50% of tests represent truly null associations. The GIF of the muscle models was found to be high, and as such a surrogate variable analysis (SVA) was
conducted individually for blood and muscle using the R sva 3.38 package (23). SVA analysis estimates the probability of each gene being an empirical control gene (i.e., not associated with MND) through an iterative re-weighted least squares regression and captures variation outside of a specified model with known covariates. In this case, the SVA was conducted with MND status as the only covariate to preserve in the non-null model. This function also generated a suggested number of SVs to use in differential expression analysis through a permutation procedure.

Genes with Benjamini-Hochberg-adjusted p-values falling under the significance threshold of 0.05 were deemed as transcriptome-wide significant and were taken forward for further enrichment analyses. The union of genes significant to Salmon and STAR models formed the gene list entered gprofiler2 0.2 (24) gene ontology enrichment and the STRING 11.0 (25) protein-protein interaction network tools using default settings. The p-values returned by this enrichment analysis are adjusted for multiple testing by gprofiler2’s default gSCS algorithm. Additionally, a supplementary analysis was conducted with the Revised Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS-R) (26) lower limb subscore as the outcome, adjusting for age, sex and the estimated surrogate variables. This subscore combined scored answers for questions relating to the ability to walk and the ability to climb stairs, with lower values reflecting decreased ability.

Finally, the isoform-level length-scaled transcript counts generated by Salmon for the muscle data were analyzed through the IsoformSwitchAnalyzeR (27) 1.12 library in R. This package implements visualization and annotation functions on top of a DEXSeq, which currently implements DESeq2’s normalization and estimation approaches to iteratively compare the number of transcripts of a given exon compared to the rest of that gene (28) but in this case for isoforms. Isoforms were annotated to exons through the Gencode v32 primary assembly.
Following default prefiltering, isoforms were tested for differential usage using MND status as its condition, adjusted for age, sex and batch effect, then analyzed for alternative splicing as well as predicted the open reading frame (ORF) of each isoform of significantly switched genes. IsoFormSwitchAnalyzer implements multiple functions for the export of all isoform sequences for genes with a significant switch, and the import of the results from external annotation databases utilizing these. The databases were CPAT (using the suggested human coding threshold of 0.364) (29) for coding potential, SignalP (30) for predicting signal peptide sites, IUPred2A (31) to predict intrinsically disordered regions (IDRs) and Pfam (32) to predict protein domains.

Results

Quality control of the RNAseq reads for each sample using PHRED scores indicated the quality was high, with the main batch containing mean scores at each base pair across reads and samples of >30. However, the samples demonstrated some form of degradation (potentially from long-term storage), particularly for blood. Mean genome alignment over all blood and muscle samples through STAR was 54.4% and 87.6% respectively, while the proportion of those which were assigned to exons were 29.8% and 58.2% respectively. For Salmon, a mean of 74.1% of reads mapped to the transcriptome and were successfully assigned a feature in muscle, whereas this number was only 40% in blood. The proportions of the reads successfully assigned to a feature are plotted in Figure S1 for STAR and Figure S2 for Salmon.

A principal component analysis of the transcript levels revealed a differing pattern of variation between the blood and muscle analyses (Figure 1). The second principal component of both blood and muscle (explaining 7.2% and 8.4% of variance in gene expression, respectively), captured sex differences (correlation $r = 0.90$ with $p = 4.9 \times$
for blood, and $r = -0.89$ with $p = 2.6 \times 10^{-16}$ for muscle). The first principal component in muscle (explaining 22.3% of expression variance) captured differences between cases and controls. The second principal component in muscle is significantly correlated with the lower limb subscore of the ALSFRS-R when estimating the PCA in case counts only ($r = 0.46$, $p = 0.014$). The relative log expression of the counts (centered and non-centered) passing QC is displayed within Figure S3. Similar observations with regards to the principal components and correlations apply to the counts as generated through STAR (Figure S4).

Testing for case-control differences in gene-expression levels in muscle while correcting for age, sex and replicate batch effect identified additional confounding within the data. The genomic inflation factor (GIF) of Salmon and STAR were 2.53 and 3.47 respectively, with 3258 and 6874 transcriptome-wide significant genes detected under the adjusted p-value of 0.05 (Figure S5). This was corrected with the addition of an automatically determined number of surrogate variables to each model as calculated by permutation procedure. With this correction the GIF falls to 1.18 and 1.13, with 38 and 26 significant genes for Salmon and STAR-aligned data respectively. For the blood transcriptome analysis, such inflation was not an issue (a GIF of 0.85 and 1.01 in Salmon and STAR respectively, with two significant genes each) and as such no SVs were added to the model.

Of the 38 and 26 genes detected by Salmon and STAR respectively, there was an intersection of 12 (Table 1) with the same direction of fold change for both methods. The complete list of transcriptome-wide genes for each model in each tissue type as well as their test statistics can be found in individual tabs of Table S2. Of the 52 genes identified by either Salmon and STAR models, 9 were long non-coding RNAs and 5 were pseudogenes. Of the 38 protein-coding genes, 7 had lower expression and 31 had higher expression in MND cases versus
controls. As one of these genes, AHRR, is a known smoking biomarker in blood (33) we conducted an additional iteration of the ‘base’ model for muscle with the addition of ever-smoked status (23 never, 22 ever). MND status was found to be significantly associated with AHRR count in the STAR model (log2FC = 1.98, \( p = 0.0085 \)), whereas smoking history was not (log2FC -0.11, \( p = 0.88 \)).

A gene ontology enrichment analysis conducted on the results of each method (Table 2) revealed shared enrichment for the GO biological processes term of acute phase response (\( p = 1.4 \times 10^{-3} \) and \( p = 5.5 \times 10^{-4} \) for Salmon and STAR respectively), attributable to the components of serum amyloid A within the set. Other notable terms enriched in the significant results for Salmon were cardiac muscle tissue morphogenesis \( (p = 4.4 \times 10^{-3}) \), positive regulation of synapse assembly \( (p = 4.7 \times 10^{-3}) \), and tissue migration \( (p = 7.3 \times 10^{-3}) \). For STAR, notable terms were skeletal muscle tissue growth \( (p = 9.7 \times 10^{-3}) \), axonogenesis involved in innervation \( (p = 8.4 \times 10^{-3}) \) and neutrophil clearance \( (p = 7.0 \times 10^{-3}) \). However, these terms test as enriched due to having one or two significant genes detected from a small pool of other genes in the global set annotated with the term. The unified set of differentially expressed proteins between STAR and Salmon in muscle was identified to have significantly more interactions (9) than expected (4, \( p = 0.01 \)) according to the STRING pathway analysis, indicating a likely biological relation if not interaction between them (Figure S6).

Within the blood sample, the \( ODZ1 \) homolog pseudogene \( AL0087071 \), notably found on the X-chromosome, was the only gene differentially expressed by both Salmon and STAR methods (respective log2FCs of 3.14 and 3.07, with a \( p = 1.9 \times 10^{-6} \) and \( p = 3.5 \times 10^{-6} \), Table S3). \( EDH2 \) was not detected by the STAR aligner though was significantly expressed in Salmon (log2FC of -1.75; \( p = 7.0 \times 10^{-7} \)) and the expression of \( FSD1L \) was only found
significant in STAR (log2FC of 0.57; \(p = 2.0 \times 10^{-6}\)).

The results of the supplementary analyses on the 28 cases in predicting the ALSFRS-R lower limb subscores are reported in **Table S4**. The STAR model identified 11 genes significantly expressed in change of scoring and the Salmon model identified 10, with 3 genes (**IGFNI**, **PBX1**, and **TMEM9**) sharing association in both. Both models were adjusted for with two additional surrogate variables (as the number automatically suggested by the function).

A total of 17 genes were identified with isoform switches within the muscle sample falling underneath the FDR-adjusted p-value (q-value) of 0.05 and having predicted functional consequences to the expressed protein. The list of features, their direction of consequence and the genes falling under each is listed in **Table S5**. The individual per-isoform tests for each gene are available in **Table S6**. No significant (FDR < 0.05) directionality of functional consequences of these isoform differences was observed for any genomic feature, although there was a slight increase in switching to longer transcripts (8:6 longer vs. shorter switches) and an increase in exon number gain (6:3 gain vs. loss of exon count).

From the isoform switches with functional consequences, only **SAA2** was represented within our main gene-level analysis. **SAA2** featured an increase of its ENST00000256733 isoform fraction by 24% (\(q = 5.74 \times 10^{-11}\)) correlated with a decrease of its ENST00000528349 isoform fraction by 25% (\(q = 9.84 \times 10^{-11}\)). The resulting functional consequence was an overall decrease in transcript length by a shortened 3’ UTR and the reduction of intron 3-4 from 3090 to 394 bp, increasing its predicted coding potential (**Figure 2**). Two other isoform switches with predicted gain in coding potential are **PLCE1** (**Figure S7**) and **SAMD4A** (**Figure S8**). **PIP5K1B** (**Figure S9**) and **MYH3** (**Figure S10**) also have gain in domains which indicate that the isoform with increased expression encodes the more complete version of the protein.
Discussion

The majority of studies on muscle changes associated with MND have been performed in mice (34,35) or iPSC-derived cell-culture (36). To our knowledge, our study is one of the first investigations into the transcriptomic machinery in human MND muscle tissue and provides a substantial increase in the number of individuals examined. We identify 38 protein-coding genes differentially expressed between MND cases and controls with 31 of these being upregulated in cases. Muscle biopsies were sampled from the healthier leg in MND cases.

Serum amyloid A transcripts (specifically \textit{SAA1}, \textit{SAA2} and the \textit{SAA2-SAA4} readthrough, contiguously located on chromosome 11) were highly expressed in MND patients. SAA is persistently elevated in other chronic inflammatory diseases and a reliable indicator of inflammation due to an up-to 1000-fold increase in plasma levels during acute phase responses (37). In a heightened response, SAA most often becomes associated with the lipid surface of high-density lipoproteins (HDLs), possibly as a stabilized hub to interact with other cellular components (38). In its association with HDL cholesterol, SAA is indicated to be involved in stimulating the inflammatory response within the innate immune system (39). Other acute phase proteins such as C-reactive protein and CD14 have been detected as elevated in MND sera, and increased levels of these are reported to be associated with disease progression and mortality rate (40). In addition to the aforementioned acute phase proteins, several components of the complement system including C3 (41), C4 (42), and C5a/C5b-9 (43) have been reported to be higher in the whole blood, serum or plasma of ALS cases to controls in previous studies. As such, it may be of value in the future to investigate local accumulation of SAA as a biomarker for disease progression and as a representation of the overall inflammatory response.
Several genes associated with tissue generation are detected as positively expressed in cases. *FST* (follistatin) is a receptor agonist which binds to other transforming growth factor-β ligands, including myostatin and activin-A which would otherwise limit muscle mass without inhibition (44). Additionally, *FST* has an antagonistic effect on several bone morphogenetic proteins, of which it forms a complex with and alters signaling to induce muscle hypertrophy while reducing atrophy (45). *MYH3*, a myosin heavy-chain encoding gene primarily expressed embryonically, was not captured by the gene-based analyses; however, the differential isoform expression analysis indicated a heightened expression of its most complete isoform within MND cases. This is of particular interest as *MYH3* has been indicated as a valid and robust biomarker for skeletal muscle tissue regeneration across all ages in other muscle-wasting diseases such as Duchenne muscular dystrophy (46). Muscle tissue does not exist in the body solely as actin filaments, but bundled within a complex net of connective tissue and infiltrating immune cells (47). Other genes instrumental in tissue development and differentiation found expressed higher in cases are the keratin *KRT80* (48), the actin-binding protein encoding *SHROOM3* which is utilized for epithelial folding in neural tube closure (49), and filamentous actin-binding ectodermal-neural cortex 1 *ENC1* (50).

Several transcripts found over-expressed in MND cases are not principally expressed by skeletal muscle tissues in adulthood. For instance, *ACTC1* was found to be significantly upregulated but not *ACTA1*. *ACTC1* encodes the α-cardiac muscle actin which constitutes < 5% of skeletal muscle by adulthood, whereas *ACTA1* encodes the almost-identical predominant muscular isoform α-skeletal muscle actin (51). Within a small case study of 7 congenital myopathy patients with a complete lack of α-skeletal muscle actin, α-cardiac actin was shown to constitute muscle tissues and the highest amount correlated to the participant with the best muscular function (52). However, as *ACTA1* has been previously identified to
be increased as a function of compensatory hypertrophy in a small real-time PCR study of 5
ALS patients, a larger sample size may capture the increased expression of both it and
ACTC1 (53). ANKRD1/CARP is the ankryn repeat domain most present in cardiac tissue and
is shown to be over-expressed within MND muscle compared to controls unlike its skeletal
muscle paralog, ANKRD2/ARPP. However, ANKRD1 expression has also been noted to
increase in skeletal muscle with denervation-induced atrophy in murine models (54).

The expression of extracellular matrix components and collagens COL19A1 and COL25A1
was also found to be increased with muscle tissue genes such as ACTC1. COL25A1, encoding
the collagen type XXV alpha 1 chain, is a known Alzheimer’s risk gene (55) for its role in
assisting the formation of characteristic senile plaques when over-expressed in the brain (56).
The cleavage of COL25A1 creates the proteolytic CLAC (collagenous Alzheimer amyloid
plaque component), and as such has the alias of CLAC-P (precursor) (57). Collagen XXV is
expressed primarily in skeletal muscle during development, acting as a required factor in
myofiber fusion from myoblasts in early myogenesis (58). COL25A1-KO mice nearly
completely failed to innervate skeletal muscle tissue during development by not being able to
elongate and branch motor axon bundles, indicating it as a required component to initial
intramuscular motor innervation (59). Collagen XIX structures part of the basement
membranes, and is present in a host of different tissues including skeletal muscle, axons,
epithelia and brain intraneurons (60). As collagens are vital in the production of functional
skeletal muscle tissue, their increased expression in MND further supports the evidence of
attempted muscle repair as a compensatory response. Higher COL19A1 expression levels in
MND muscle biopsies have already been linked to faster progression of the disease (61).

A few other highly upregulated genes of note include some that are influenced by the
multiple upregulated transcription factors in the set (SPI and MYCL). The increased
expression of *TDO2* and *AHRR* may indicate the presence of prior *AHR* upregulation, which has been indicated to be invoked in neural cell proliferation (62). Following *AHRR* and the aforementioned *SAA1*, tryptophan catabolism enzyme *TDO2* has the largest positive fold-change in expression. *TDO2* is ordinarily expressed in the liver as a tryptophan regulator and is also known as an oncogene commonly expressed in most cancers (63). While the products of the catabolic kynurenine pathway include metabolites used for neurotransmission and immune regulation, *TDO2* depletes free tryptophan used for proliferation of T-cells and produces kynurenine that inhibits the proliferation of effector T lymphocytes and promoting regulatory T-cell differentiation, and activates the transcription factor *AHR* (64). The kynurenine pathway itself produces metabolites that influence energy homeostasis in skeletal muscle and other tissues, and kynurenic acid (a product of the pathway) increases energy utilization while also acting as an *AHR* ligand (65).

**Conclusions**

In summary, we identify 52 genes whose expression in muscle is significantly associated with MND cases versus controls, with the majority showing increased expression in cases. These genes encode vital structural components of muscle, particularly in regenerative or early developmental processes. This provides novel insight into the repair mechanisms triggered in the presence of neurodegeneration and muscle wasting in MND, as well as the chronic inflammation within these tissues as evidenced by expression of acute-phase proteins.

**Abbreviations**

*ALS-FRS-R*: Revised Amyotrophic Lateral Sclerosis Functional Rating Scale
398  **ALS**: Amyotrophic lateral sclerosis

399  **EM**: Expectation maximalization

400  **GIF**: Surrogate variable analysis

401  **IncRNA**: Long non-coding RNA

402  **MND**: Motor neurone disease

403  **NMJ**: Neuromuscular junction

404  **ORF**: Open reading frame

405  **PC**: Protein coding

406  **PCA**: Principal component analysis

407  **PP**: Processed pseudogene

408  **RIN**: RNA integrity number

409  **TUP**: Transcribed unprocessed pseudogene.

410  **UP**: Unprocessed pseudogene

411  **Declarations**

412  Ethics approval and content to participate

413  All work performed in this study was approved by the Royal Brisbane and Women’s Hospital

414  and University of Queensland human research ethics committees. All participants provided

415  written informed consent.
Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study are available on request from the corresponding author. The raw transcript data are not publicly available due to ethical restrictions; however, gene and transcript counts have been made available. The datasets generated during and/or analysed during the current study are available in the University of Queensland data collection repository, https://doi.org/10.48610/b722f1f.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported through funding from the Scott Sullivan MND Research Fellowship to S.T.N. (2015-2020; MND and Me Foundation and RBWH Foundation), FightMND Mid-Career Fellowship (to S.T.N.), MNDRA Charcot Grant (GIA 1701 to S.T.N. and F.J.S.), the National Health and Medical Research Council Australia (1101085 to S.T.N. and F.J.S., 1113400 & 1173790 to N.R.W.) and the Australian Research Council (Future Fellowship 200100837 to A.F.M.).

Authors’ contributions

A.F., A.F.M., and N.R.W. wrote the paper. A.F. conducted the analyses. A.F.M., N.R.W., and S.T.N. conceptualized this work. S.W and V.B. processed the samples for sequencing. L.M.W. and A.K.H. allocated laboratory resources and supervised sample processing. P.A.M. and R.D.H. provided samples. S.T.N. and F.J.S. performed the muscle biopsies. All authors read and approved the final manuscript.
Acknowledgements

The authors thank all patients and participants who provided biological samples for use in this research.
References


25
**Figures**

![Figure 1](image_url)

**Figure 1.** First two principal components of transcriptome log counts (regularized using Salmon) consisting of the top 500 most variable genes for blood and muscle respectively. PC1 in muscle captures variation between cases in sampled lower limb lower motor neurone scores ($r = 0.40$, $p = 0.041$), though does not fully separate cases and controls. PC2 captures sex differences for both tissue types (blood: $r = 0.90$, $p = 4.9 \times 10^{-14}$; muscle: $r = -0.89$, $p = 2.6 \times 10^{-16}$).

Ellipses added to visualize confidence between cases and controls.
The isoform switch in SAA2 (Control vs Case)

- **Figure 2.** Visualization of isoform expression in the transcriptome-wide significant gene SAA2. The coloured plots at the top show three transcripts (detected via Salmon), annotated for their inclusion of protein domains as well as the site of their transcription start and termination sites. The top transcript is predicted to have high coding potential through CPAT. The bottom graphs individually delineate the difference in mean counts between cases and controls in the expression of the gene and its constituent isoforms. The rightmost graph indicates the isoform expression as a proportion of the total expression across all isoforms within cases and controls separately, indicating a greater proportion of expression for the coding variant and a decreased proportion of expression for the second non-coding variant.
Tables

Table 1. Overlap of transcriptome-wide significant differences between MND cases and controls from STAR and Salmon models with log2 fold change and Benjamini-Hochberg adjusted p-values.

Bolded names are significant in both models. lncRNA: long non-coding RNA; PC: protein coding; PP: processed pseudogene; UP: Unprocessed pseudogene; TUP: transcribed unprocessed pseudogene.

<table>
<thead>
<tr>
<th>Name</th>
<th>Chr</th>
<th>Start</th>
<th>Length</th>
<th>Type</th>
<th>Description</th>
<th>STAR</th>
<th>Salmon</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO014900.1</td>
<td>19</td>
<td>75122</td>
<td>3537</td>
<td>lncRNA</td>
<td>Novel transcript, antisense to det1l associated protein (strap) pseudogene</td>
<td>2.17</td>
<td>2.89</td>
</tr>
<tr>
<td>AO019878.1</td>
<td>2</td>
<td>37816</td>
<td>1947</td>
<td>PC</td>
<td>Spermatogenesis/spermatogenesis-related protein (strap) pseudogene</td>
<td>0.56</td>
<td>0.63</td>
</tr>
<tr>
<td>AC013775.1</td>
<td>2</td>
<td>119476</td>
<td>10919</td>
<td>lncRNA</td>
<td>Sct transcript rna 1</td>
<td>-2.98</td>
<td>5.4</td>
</tr>
<tr>
<td>AC068540.4</td>
<td>11</td>
<td>17548</td>
<td>29134</td>
<td>PC</td>
<td>Novel protein</td>
<td>1.71</td>
<td>1.76</td>
</tr>
<tr>
<td>AC024757.1</td>
<td>15</td>
<td>34755</td>
<td>5822</td>
<td>lncRNA</td>
<td>Novel transcript, antisense to detl</td>
<td>0.62</td>
<td>0.65</td>
</tr>
<tr>
<td>AC033760.1</td>
<td>15</td>
<td>62827</td>
<td>3630</td>
<td>lncRNA</td>
<td>Novel transcript, antisense to laln</td>
<td>0.76</td>
<td>1.32</td>
</tr>
<tr>
<td>AC035780</td>
<td>19</td>
<td>11577</td>
<td>4049</td>
<td>PC</td>
<td>Acid phosphatase 5, transferrin resistant</td>
<td>1.46</td>
<td>3.14</td>
</tr>
<tr>
<td>ACTC1</td>
<td>15</td>
<td>34790</td>
<td>5320</td>
<td>PC</td>
<td>Actin, alpha cardiac muscle 1</td>
<td>2.08</td>
<td>0.74</td>
</tr>
<tr>
<td>AHRR</td>
<td>5</td>
<td>27167</td>
<td>18662</td>
<td>PC</td>
<td>Arylsulphatase-receptor repressor</td>
<td>0.65</td>
<td>0.1</td>
</tr>
<tr>
<td>AL136655.2</td>
<td>14</td>
<td>32141</td>
<td>1304</td>
<td>lncRNA</td>
<td>Novel transcript</td>
<td>0.94</td>
<td>0.1</td>
</tr>
<tr>
<td>AL106750.2</td>
<td>13</td>
<td>11012</td>
<td>2827</td>
<td>PC</td>
<td>Toc</td>
<td>1.35</td>
<td>0.54</td>
</tr>
<tr>
<td>ANKRD1</td>
<td>10</td>
<td>90912</td>
<td>9161</td>
<td>PC</td>
<td>Ankyrin-repeat domain 1</td>
<td>2.48</td>
<td>0.17</td>
</tr>
<tr>
<td>APOE1</td>
<td>17</td>
<td>57628</td>
<td>29919</td>
<td>PC</td>
<td>Apolipoprotein-related protein complex 5 subunit 1</td>
<td>0.57</td>
<td>0.09</td>
</tr>
<tr>
<td>APP23</td>
<td>12</td>
<td>57906</td>
<td>8310</td>
<td>PC</td>
<td>Ap23 metalloprotease and asp synthase assembly factor binding</td>
<td>0.12</td>
<td>0.15</td>
</tr>
<tr>
<td>DMDM4</td>
<td>5</td>
<td>73716</td>
<td>14727</td>
<td>TUP</td>
<td>Tandem pseudogene 4</td>
<td>0.46</td>
<td>0.14</td>
</tr>
<tr>
<td>CH4D1</td>
<td>12</td>
<td>57747</td>
<td>7287</td>
<td>PC</td>
<td>Cyclin dependent kinase 4</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>CNKNS3</td>
<td>9</td>
<td>22002</td>
<td>6403</td>
<td>PC</td>
<td>Cyclin-dependent kinase inhibitor 28</td>
<td>0.94</td>
<td>0.28</td>
</tr>
<tr>
<td>CHD3L1</td>
<td>2</td>
<td>20317</td>
<td>9312</td>
<td>PC</td>
<td>Chordin-like 3</td>
<td>2.12</td>
<td>0.94</td>
</tr>
<tr>
<td>CHRNA1</td>
<td>2</td>
<td>17477</td>
<td>9434</td>
<td>PC</td>
<td>Cholinergic receptor nicotinic alpha subunit</td>
<td>0.94</td>
<td>0.14</td>
</tr>
<tr>
<td>COL5A1</td>
<td>6</td>
<td>86656</td>
<td>34513</td>
<td>PC</td>
<td>Collagen type v alpha 1 chain</td>
<td>2.08</td>
<td>0.16</td>
</tr>
<tr>
<td>COL5A1</td>
<td>4</td>
<td>10886</td>
<td>7257</td>
<td>PC</td>
<td>Collagen type v alpha 1 chain</td>
<td>2.08</td>
<td>0.16</td>
</tr>
<tr>
<td>DCLK1</td>
<td>13</td>
<td>30764</td>
<td>65455</td>
<td>PC</td>
<td>Doublecortin-like kinase 1</td>
<td>1.08</td>
<td>0.23</td>
</tr>
<tr>
<td>DLRUL2</td>
<td>1</td>
<td>63547</td>
<td>3555</td>
<td>lncRNA</td>
<td>Deleted in lymphomatous leukemia 2</td>
<td>1.19</td>
<td>0.27</td>
</tr>
<tr>
<td>EPAC1B</td>
<td>1</td>
<td>63523</td>
<td>9822</td>
<td>PC</td>
<td>Richard calcium binding domain 7</td>
<td>0.87</td>
<td>0.34</td>
</tr>
<tr>
<td>EVC1</td>
<td>5</td>
<td>75517</td>
<td>14727</td>
<td>TUP</td>
<td>Tandem pseudogene 4</td>
<td>0.43</td>
<td>0.14</td>
</tr>
<tr>
<td>FST</td>
<td>5</td>
<td>53480</td>
<td>6209</td>
<td>PC</td>
<td>Follistatin</td>
<td>1.45</td>
<td>0.27</td>
</tr>
<tr>
<td>KRTFP9</td>
<td>12</td>
<td>52180</td>
<td>23193</td>
<td>PC</td>
<td>Keratin 92</td>
<td>2.75</td>
<td>0.94</td>
</tr>
<tr>
<td>LMCT1-LF15</td>
<td>18</td>
<td>25845</td>
<td>26188</td>
<td>lncRNA</td>
<td>Lmct1-like transcript 1</td>
<td>0.87</td>
<td>0.28</td>
</tr>
<tr>
<td>LINC009727</td>
<td>2</td>
<td>37820</td>
<td>8577</td>
<td>lncRNA</td>
<td>Long intergenic non-protein-coding rna 211</td>
<td>0.75</td>
<td>0.92</td>
</tr>
<tr>
<td>LINC026692</td>
<td>1</td>
<td>16784</td>
<td>8587</td>
<td>lncRNA</td>
<td>Long intergenic non-protein-coding rna 626</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>LRKNS</td>
<td>7</td>
<td>11109</td>
<td>3449</td>
<td>PC</td>
<td>Leucine rich repeat neuronal 3</td>
<td>0.27</td>
<td>0.34</td>
</tr>
<tr>
<td>LRTM1</td>
<td>3</td>
<td>54198</td>
<td>4858</td>
<td>PC</td>
<td>Leucine rich repeat neuronal 1</td>
<td>0.50</td>
<td>0.58</td>
</tr>
<tr>
<td>MERTK</td>
<td>11</td>
<td>11184</td>
<td>13075</td>
<td>PC</td>
<td>Mert proto-oncogene, tyrosine kinase</td>
<td>0.63</td>
<td>0.58</td>
</tr>
<tr>
<td>MTOCP1P</td>
<td>4</td>
<td>42246</td>
<td>518</td>
<td>PF</td>
<td>Mto13 pseudogenes 33</td>
<td>0.63</td>
<td>0.58</td>
</tr>
<tr>
<td>MTND1P22</td>
<td>4</td>
<td>42246</td>
<td>315</td>
<td>PF</td>
<td>Mto13 pseudogenes 22</td>
<td>0.63</td>
<td>0.58</td>
</tr>
<tr>
<td>MUC09</td>
<td>12</td>
<td>40392</td>
<td>26715</td>
<td>PC</td>
<td>Mucin 19 oligomer</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>MYCL</td>
<td>3</td>
<td>39856</td>
<td>6831</td>
<td>PC</td>
<td>Myc proto-oncogene, bhlh</td>
<td>1.40</td>
<td>0.93</td>
</tr>
<tr>
<td>NGCN</td>
<td>14</td>
<td>23469</td>
<td>6174</td>
<td>PC</td>
<td>Neuroginin</td>
<td>0.04</td>
<td>0.07</td>
</tr>
<tr>
<td>PDEB9</td>
<td>5</td>
<td>77210</td>
<td>44995</td>
<td>PC</td>
<td>Phosphodiesterase 8</td>
<td>0.23</td>
<td>0.41</td>
</tr>
<tr>
<td>PEG100</td>
<td>7</td>
<td>94654</td>
<td>3216</td>
<td>PC</td>
<td>PEG100</td>
<td>0.62</td>
<td>0.37</td>
</tr>
<tr>
<td>PNPLA3</td>
<td>4</td>
<td>45923</td>
<td>7097</td>
<td>PC</td>
<td>Patisan like phospholipid domain containing 3</td>
<td>0.98</td>
<td>0.65</td>
</tr>
<tr>
<td>RBM14</td>
<td>2</td>
<td>23798</td>
<td>3159</td>
<td>PC</td>
<td>Ras binding motif protein 4</td>
<td>0.79</td>
<td>0.28</td>
</tr>
<tr>
<td>SBD3</td>
<td>9</td>
<td>92514</td>
<td>27513</td>
<td>PC</td>
<td>Ras and raf interaction 3</td>
<td>0.86</td>
<td>0.34</td>
</tr>
<tr>
<td>SAA1</td>
<td>17</td>
<td>1826</td>
<td>3717</td>
<td>PC</td>
<td>Serum amyloid a1</td>
<td>3.05</td>
<td>3.40</td>
</tr>
<tr>
<td>SAA3A-A4</td>
<td>17</td>
<td>18231</td>
<td>2713</td>
<td>PC</td>
<td>Serum amyloid a3</td>
<td>4.01</td>
<td>4.01</td>
</tr>
<tr>
<td>SEHOMD3</td>
<td>6</td>
<td>74241</td>
<td>30825</td>
<td>PC</td>
<td>Scleroma family member 3</td>
<td>2.72</td>
<td>2.72</td>
</tr>
<tr>
<td>SLC47A2</td>
<td>17</td>
<td>19678</td>
<td>60494</td>
<td>PC</td>
<td>Solute carrier family member 2</td>
<td>3.17</td>
<td>3.17</td>
</tr>
<tr>
<td>SPP1</td>
<td>12</td>
<td>53254</td>
<td>36715</td>
<td>PC</td>
<td>Spp1 transcript factor</td>
<td>0.26</td>
<td>0.06</td>
</tr>
<tr>
<td>SYT15</td>
<td>10</td>
<td>46757</td>
<td>21597</td>
<td>PC</td>
<td>Syntaphilin 15</td>
<td>0.18</td>
<td>0.28</td>
</tr>
<tr>
<td>TDO2</td>
<td>15</td>
<td>73573</td>
<td>6868</td>
<td>PC</td>
<td>Tryptophan 2,3-dioxygenase</td>
<td>0.78</td>
<td>0.98</td>
</tr>
<tr>
<td>VN583P</td>
<td>19</td>
<td>21889</td>
<td>5464</td>
<td>UP</td>
<td>Urothelial receptor 3</td>
<td>0.50</td>
<td>0.27</td>
</tr>
</tbody>
</table>
Table 2. Enriched gene ontology (GO) terms for both Salmon and STAR models, as generated from TopGO. The total number of annotated genes from the greater set is followed by the number of those which are significantly differentially expressed, as well as the enrichment of the term as calculated by Fisher’s exact test. Significant/background genes are subset to those with GO terms (27/14475 for salmon and 19/13588 for STAR).

<table>
<thead>
<tr>
<th>GO.ID</th>
<th>Term</th>
<th>Genes</th>
<th>Annotation</th>
<th>Significant</th>
<th>Expected</th>
<th>Fisher p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salmon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006953</td>
<td>Acute-phase response</td>
<td>30</td>
<td>2: SAA1, SAA2</td>
<td>0.06</td>
<td>1.4 x 10^-3</td>
<td></td>
</tr>
<tr>
<td>GO:0005088</td>
<td>Cardiac muscle tissue morphogenesis</td>
<td>53</td>
<td>2: ANKR41, ACTC1</td>
<td>0.10</td>
<td>4.4 x 10^-3</td>
<td></td>
</tr>
<tr>
<td>GO:0019865</td>
<td>Positive regulation of synapse assembly</td>
<td>55</td>
<td>2: LRTM1, LRRN3</td>
<td>0.10</td>
<td>4.7 x 10^-3</td>
<td></td>
</tr>
<tr>
<td>GO:0000130</td>
<td>Tissue migration</td>
<td>271</td>
<td>2: ACTC1, SP1</td>
<td>0.61</td>
<td>7.3 x 10^-3</td>
<td></td>
</tr>
<tr>
<td>GO:0045607</td>
<td>Regulation of inner ear auditory receptor cell differentiation</td>
<td>5</td>
<td>1: MYC1</td>
<td>0.01</td>
<td>9.3 x 10^-3</td>
<td></td>
</tr>
<tr>
<td>GO:0031444</td>
<td>Tryptophan catabolic process to kynurenine</td>
<td>5</td>
<td>1: TDO2</td>
<td>0.01</td>
<td>9.3 x 10^-3</td>
<td></td>
</tr>
<tr>
<td>GO:0036155</td>
<td>Aroylglycerol aeryl-chain remodeling</td>
<td>5</td>
<td>1: PNPLA3</td>
<td>0.01</td>
<td>9.3 x 10^-3</td>
<td></td>
</tr>
<tr>
<td>GO:0070814</td>
<td>Hydrogen sulfide biosynthetic process</td>
<td>5</td>
<td>1: SP1</td>
<td>0.01</td>
<td>9.3 x 10^-3</td>
<td></td>
</tr>
<tr>
<td>STAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006953</td>
<td>Acute-phase response</td>
<td>25</td>
<td>2: SAA1, SAA2</td>
<td>0.03</td>
<td>5.5 x 10^-4</td>
<td></td>
</tr>
<tr>
<td>GO:1903566</td>
<td>Positive regulation of protein localization to cilium</td>
<td>5</td>
<td>1: EPCAB7</td>
<td>0.01</td>
<td>7.0 x 10^-4</td>
<td></td>
</tr>
<tr>
<td>GO:0097350</td>
<td>Neutrophil clearance</td>
<td>5</td>
<td>1: MERTK</td>
<td>0.01</td>
<td>7.0 x 10^-4</td>
<td></td>
</tr>
<tr>
<td>GO:0036155</td>
<td>Aroylglycerol aeryl-chain remodeling</td>
<td>6</td>
<td>1: PNPLA3</td>
<td>0.01</td>
<td>8.4 x 10^-4</td>
<td></td>
</tr>
<tr>
<td>GO:0060385</td>
<td>Axonogenesis involved in innervation</td>
<td>6</td>
<td>1: COUL2A1</td>
<td>0.01</td>
<td>8.4 x 10^-4</td>
<td></td>
</tr>
<tr>
<td>GO:0017798</td>
<td>Positive regulation of hair follicle development</td>
<td>7</td>
<td>1: FST</td>
<td>0.01</td>
<td>9.7 x 10^-4</td>
<td></td>
</tr>
<tr>
<td>GO:0046360</td>
<td>Skeletal muscle tissue growth</td>
<td>7</td>
<td>1: CIRNA1</td>
<td>0.01</td>
<td>9.7 x 10^-4</td>
<td></td>
</tr>
<tr>
<td>GO:1901072</td>
<td>Glucosamine-containing compound catabolic process</td>
<td>7</td>
<td>1: CH3L1</td>
<td>0.01</td>
<td>9.7 x 10^-4</td>
<td></td>
</tr>
</tbody>
</table>
Supplemental Information

Primary supplement. PDF. Figures S1-S10 and Table S1, S3-S5.

Table S2. Excel. Complete list of transcriptome-wide genes for each model in each tissue type as well as their test statistics.

Table S6. Excel. Differentially expressed isoforms as estimated from the IsoformSwitchAnalyzeR implementation of DEXSeq.