2016

Investigating the molecular etiologies of sporadic ALS (sALS) using RNA-Sequencing

David G. Brohawn
Virginia Commonwealth University

Investigating the molecular etiologies of sporadic ALS (sALS) using RNA-Sequencing

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

David G. Brohawn
Bachelor of Arts, University of Delaware, 2008

Director: James P. Bennett, MD, PhD
Associate Professor of Neurology

Administrative Dissertation Advisor: Michael S. Grotewiel
Associate Professor of Human Genetics

Virginia Commonwealth University
Richmond, Virginia
April 28th, 2016

Acknowledgement

I would like to first thank Dr. James Bennett for giving me the opportunity to pursue a PhD under his guidance. His open minded, adventurous nature was inspiring to me as a budding scientist. This project was thoroughly enjoyable, and I will always look back on it as a fun learning experience.

Thank you to Amy Ladd, Laura O' Brien, Ravindar Thomas, Ann Rice, and Paula Keeney. You all taught me a ton about handling RNA, cell culture experiments, and data analysis that made my job a lot easier.

Thank you to my committee members for your guidance and feedback throughout my dissertation project. Your words helped shape my analyses and interpretations, and furthered my maturation as a scientist and person.

Thank you to the Human Genetics faculty (especially Mike Grotewiel and Rita Shiang) for teaching me the tenants of genetics, helping place me in Dr. Bennett's lab, and seeing me through graduation.

Thank you to the VCU Core lab and Supply Center for helping me get necessary materials for my experiments.

Thank you to the authors of all the Bioinformatics packages I used, as well as the many users who posted troubleshooting advice for these tools.

Thank you to Dr. Andrew Larner for helping me schedule and book a room for my committee meetings.

Thank you to all my classmates, including Aymen, Derek, Erik, Mohammed, and Uzo.

A huge thank you my closest friends (Maciej, Tyler, Nav, Sam), and my girlfriend Audra for emotional support, good times, laughs, and feedback on my research.

Finally, a massive thank you to my family members who listened as I shared my graduate school experience over the years. You guys kept me (mostly) sane.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ALS	Amyotrophic Lateral Sclerosis
ATP	Adenosine Triphosphate
dATP	Deoxyadenosine Triphosphate
dCTP	Deoxycytidine Triphosphate
dGTP	Deoxyguanosine Triphosphate
dTTP	Deoxythymidine Triphosphate
dUTP	Deoxyuridine Triphosphate
DEG	Differentially Expressed Gene
DNA	Deoxyribonucleic Acid
dscDNA	Double Stranded complementary DNA
dsRNA	Double Stranded RNA
ETC	Electon Transport Chain
fALS	Familial Amyotrophic Lateral Sclerosis
FPKM	Fragments per Kilobase of exon per Million Fragments Mapped
GFP	Green Fluorescent Protein
GWAS	Genome Wide Association Studies
Indel	Insertion or deletion
iPSCs	Induced Pluripotent Stem Cells
IRES	Internal Ribosome Entry Site
LPS	Lipopolysaccharide
lncRNA	Long noncoding RNA

mRNA	Messenger RNA
miRNA	Micro RNA
mg	Milligram
mL	Milliliter
mM	Millimolar
mtDNA	Mitochondrial DNA
mtRNA	Mitochondrial RNA
MTT	3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide
ncRNA	Non-coding RNA
ng	Nanogram
nM	Nanomolar
NPC	Niemann-Pick Type C
NSC	Neural Stem Cell
OXPHOS	Oxidative Phosphorylation
piRNA	Piwi RNA
Pre-mRNA	Precursor Messenger RNA
RNA	Ribonucleic Acid
RNAP	DNA-dependent RNA Polymerase
RNase MRP	Ribonuclease MRP
RNase P	Ribonuclease P
ROS	Reactive Oxygen Species
rRNA	Ribosomal RNA
sALS	Sporadic Amyotrophic Lateral Sclerosis
snRNA	Small Nuclear RNA
snoRNA	Small Nucleolar RNA
ssRNA	Single Stranded RNA
siRNA	Small interfering RNA
SLOS	Smith-Lemli-Opitz Syndrome
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant

TERC	Telomerase RNA Component
TF	Transcription Factor
TNF-α	Tumor Necrosis Factor
TOM	Topological Overlap
tRNA	Transfer RNA
TSS	Transcription Start Site
ug	Microgram
uL	Microliter
uM	Micromolar
WGCNA	Weighted Gene Co-Expressed Network Analysis
7SL RNA	Signal Recognition Particle RNA

Abstract

INVESTIGATING THE MOLECULAR ETIOLOGIES OF SPORADIC ALS (sALS) USING RNA-SEQUENCING

By
David G. Brohawn, B.A

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2016

Director: James P. Bennett, MD, PhD
Associate Professor of Neurology

Administrative Dissertation Advisor: Michael S. Grotewiel
Associate Professor of Human Genetics

Amyotrophic Lateral Sclerosis is a devastating disease involving progressive degeneration of motor neurons in the brain, brainstem, and spinal cord. Life expectancy

after diagnosis is between 3-5 years on average, with current treatments only extending life by several months. Novel therapeutic targets are sorely needed.

We combined RNA-Sequencing, systems biology analyses, and molecular biology assays to elucidate sALS group-specific differences in postmortem cervical spinal sections (7 sALS samples and 8 neurologically healthy controls) that may be relevant to disease pathology. >55 million paired end (2X150) RNA-sequencing reads per sample were generated, processed, and aligned to an hg19 human reference transcriptome then genome.

In the work presented in Chapter 2, we used bioinformatics tools to identify differentially expressed genes (DEGs) between our sALS and control sample groups. Further, we used Weighted Gene Co-Expression Network Analysis (WGCNA), an unsupervised analysis, to identify gene co-expression networks associated with sALS disease status in our sample set. Qiagen's Ingenuity Pathway Analysis revealed our sALS group-specific DEGs and a sALS group-specific gene co-expression network were both associated with inflammatory processes and TNF-α signaling. Further, Tumor Necrosis Factor Alpha Induced Protein 2 (TNFAIP2) was identified as a sALS group-specific upregulated DEG and a network hub gene in that gene co-expression network. We hypothesized its upregulation in our patients' tissues was a result of increased TNF-α signaling and that it functionally contributed to motor neuron death via TNF superfamily apoptotic pathways. Transient overexpression of TNFAIP2 led to decreased cell viability in both neural stem cells and induced pluripotent stem cell (iPSC)-derived motor neurons. Further, inhibition of activated caspase 9 (a protein necessary for TNF superfamily mitochondrial-mediated apoptosis) reversed this effect in neural stem cells.

In the work presented in Chapter 3, we used bioinformatics tools to identify sALS group-specifc mitochondrial DEGs. We did not identify any in our sample set.

In the work presented in Chapter 4, we used DEXSeq to identify sALS group-specific differentially used exons (DUEs). Qiagen’s Ingenuity Pathway Analysis revealed our sALS group-specific DUEs were associated with cholesterol biosynthesis. Cholesterol biosynthesis defects cause several rare neurodegenerative disorders, and may functionally contribute to sALS pathology.

CHAPTER 1: Introduction

I. Perturbations in Amyotrophic Lateral Sclerosis:

Amyotrophic Lateral Sclerosis (ALS) is a devastating disease involving progressive degeneration of motor neurons in the brain, brainstem, and spinal cord. Life expectancy after diagnosis is between 3-5 years on average, with current treatments only extending life by several months. Novel therapeutic targets are sorely needed.

We hypothesize perturbed cellular processes in ALS patients' tissues promote motor neuron death, and these perturbations are caused by aberrant gene expression events. Further, we presume these aberrant gene expression events can be identified using techniques commonly used in gene expression studies.

We combined RNA-Sequencing, systems biology analyses, and molecular biology assays to elucidate sALS group-specific gene and exon expression level differences in postmortem cervical spinal sections (7 sALS samples and 8 neurologically healthy controls) that may be relevant to disease pathology. For each tested gene or exon, a sALS group-specific difference was identified when the sALS sample group's representative expression value was statistically significantly different from the neurologically healthy control sample group’s representative expression value after multiple corrections. The remainder of this introductory section provides background information relevant to our gene expression study.

II. Known RNA species in eukaryotic cells:

A gene’s deoxyribonucleic acid (DNA) sequence is used to generate a complementary ribonucleic acid (RNA) transcript via transcription. In Eukaryotic cells, transcription of each RNA molecule requires an RNA polymerase (RNAP) and general transcription factors (TFs). In eukaryotic cells, known types of RNA can be broadly separated into three major groups based on their functions. These groups are 1) RNAs involved in protein synthesis and localization, 2) RNAs involved in post-transcriptional modification or telomere DNA replication, and 3) regulatory RNAs.

RNAs involved in protein synthesis and localization include pre-messenger RNAs (pre mRNAs), messenger RNAs (mRNAS), ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and signal recognition particle RNAs (7SL RNA). Pre mRNAs undergo splicing prior to becoming mature mRNAs. mRNAs are then used to encode polypeptides via ribosomal-mediated translation in the cytosol, with contributions from rRNAs and tRNAs. 7SL RNA comprises part of the signal recognition particle, a protein-RNA complex that mediates the transport of secretory and membrane proteins to a cell’s plasma membrane or endoplasmic reticulum (Luirink, Sinning 2004).

RNAs involved in post-transcriptional modification or telomere DNA replication include small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNA), Ribonucleases P and MRP (RNase P and RNase MRP), and telomerase RNA component (TERC).

snRNAs comprise part of the spliceosome complex, and functionally contribute to processing pre mRNAs into mature mRNAs. Additionally, U1 snRNA has been shown to regulate RNA Polymerase II's initiation phase (Kwek et al. 2002). snoRNAs are most widely known for their role in chemically modifying rRNAs during their maturation process. snoRNAs also modify other non-coding RNAs (ncRNAs), impact protein translation, and play a role in maintaining genome stability (Matera, Terns, Terns 2007). Ribonucleases P and MRP are essential for the maturation of tRNAs and rRNAs, respectively (Piccinelli, Rosenblad, Samuelsson 2005). Finally, TERC is an RNA component of telomerase, a ribonucleoprotein that extends telomeric DNA repeat sequences at the end of chromosomes. Telomerase reverse transcribes TERC's RNA sequence into telomeric DNA repeat sequences. These are added onto chromosome ends during telomere elongation, and protect those chromosomes' ends from degradation (Artandi 2006).

Regulatory ncRNAs include piwi-interacting RNAs (piRNAs), microRNAs (miRNAs), small interfering RNAs (siRNAs), and long noncoding RNAs (lncRNAs). piRNAs bind piwi proteins to form RNA-protein complexes, and prevent translation of RNA transcripts from mobile elements in germ line cells across various species (Weick, Miska 2014). Mobile elements are DNAs that insert themselves into various parts of the genome, and their RNA transcripts encode proteins that mediate their movement from one genomic location to another. Silencing mobile elements' RNA transcripts in germline cells is important, as it can prevent the transmission of deleterious mutations (caused by the insertion of mobile elements into susceptible genomic regions) to offspring.

miRNAs and siRNAs are single stranded RNAs (ssRNA) derived from transcribed hairpin structures and double-stranded RNAs (dsRNA), respectively. The existence of endogenous siRNAs (or siRNAs encoded by the host's genome) was only recently confirmed in vertebrate species (Piatek, Werner 2014). miRNA synthesis involves two RNA selective endonucleases (Drosha and Dicer), an Argonaute protein, and other species-specific protein factors. siRNA synthesis involves these same components, except it does not involve contributions from Drosha (Piatek, Werner 2014). A miRNA or siRNA binds to an Argonaute protein (forming an RNA-induced silencing complex known as RISC) prior to hybridizing their target mRNA via complementary basepairing. Hybridization typically occurs between the RISC's miRNA (or siRNA) and a portion of the mRNA's 3' untranslated region. After hybridization, the RISC complex reduces translation of the target mRNA by 1) rendering it vulnerable to degradation after shortening its polyA-tail, 2) reducing how efficiently it is translated into a corresponding polypeptide (or polypeptides), or 3) cleaving it into multiple pieces (Fabian, Sonenberg, Filipowicz 2010).

lncRNAs are defined as non-protein coding RNA transcripts longer than 200 nucleotides (a size that was arbitrarily chosen). While genomic analyses have identified >10,000 lncRNAs encoded by the human genome, their functions remain largely unknown. However, a variety of lncRNAs have been shown to 1) regulate transcription and splicing of pre-mRNAs, 2) alter translation of mRNAs, 3) inhibit protein activities, and 4) yield small ncRNAs after they undergo post-transcriptional processing (Chen 2015, Wilusz, Sunwoo, Spector 2009). Mutations and dysregulations of lncRNAs have been linked to cardiovascular diseases, neurological diseases, diabetes, HIV, and various types of cancers (Chen 2015).

III. RNA processing and alternative splicing in eukaryotic cells:

Most pre mRNAs undergo post-transcriptional processing in the nucleus prior to becoming mature mRNAs. Various protein complexes mediate the addition of a 7-methylguanosine cap and a poly-A tail (comprised of linked adenosine monophosphates) to nearly all pre-mRNAs' 5' and 3' ends, respectively. The cap protects the pre-mRNA from degradation, promotes downstream nuclear export of the eventual mRNA, and aids in downstream translation of the eventual mRNA into a polypeptide (Cowling 2009). The poly-A tail is also known to protect the pre-mRNA from degradation, and aids in translation of the eventual mRNA (Subtelny et al. 2014).

The majority of eukaryotic genes' sequences transcribed into pre-mRNAs contain stretches of deoxyribonucleotides called introns and exons. As a pre-mRNA is processed into a mature RNA, introns are generally removed whereas some (or all) exons are retained. Most mammalian pre-mRNAs contain introns that are a few hundred to several thousand nucleotides long, whereas the size of exons are typically around 120 nucleotides long (Will, Luhrmann 2011).

The spliceosome is a biological complex comprised of numerous snRNAs and proteins. It removes intronic (and occasionally exonic) sequences from a pre-mRNA, followed by ligating retained exonic sequences to each other. I will next describe a typical two-step process for spliceosome-mediated removal of an intronic sequence in a pre-mRNA that is illustrated in Figure 1.

First, snRNAs in the spliceosome recognize specific nucleotide sequences (called splice sites and a branch site) in a pre-mRNA's intron. Next, the spliceosome complex carries out consecutive transesterification reactions to remove the intron and connect the exons flanking it. The first reaction involves the 2' hydroxyl group of an adenosine (located in the intron’s branch site) executing a nucleophilic attack on the intron’s 5’ splice site. This results in cleavage at that site, followed by ligation of the 5’ end of the intron to the adenosine in the branch site. Next, the recently liberated 5’ exon’s 3’ hydroxyl group attacks the intron’s 3’ splice site. This leads to ligation of the two exons that were flanking the intron and removal of that intron (Will, Luhrmann 2011).

While researchers initially believed the spliceosome removed each pre-mRNA’s introns and retained its exons to generate a single mature mRNA, we now know this isn’t the only possibility. Alternative splicing events enable the generation of multiple mature mRNAs from a given pre-mRNA, despite the fact that each copy of that pre-mRNA possesses the same introns and exons. Figure 2 shows various alternative splicing events for a given pre-mRNA molecule, and the variety of resultant mature mRNAs.

Alternative splicing events occur in ~95% of eukaryotic genes (Kornblihtt et al. 2013), and lead to a greater diversity of RNA transcripts and polypeptides (transcribed from the variety of resultant mRNAs). These events are regulated by many factors including (but not limited to) pre-mRNA nucleotide sequences, cell signaling cascades, and protein-mediated modifications of spliceosome components.

Alternative splicing likely played an integral role in increasing biological complexity amongst vertebrate species and has been implicated in numerous human diseases. A recent study discovered primates have significantly higher frequencies of alternative splicing events in various organs compared to other vertebrate species when comparing orthologous genes (Barbosa-Morais et al. 2012). The authors proposed these differences

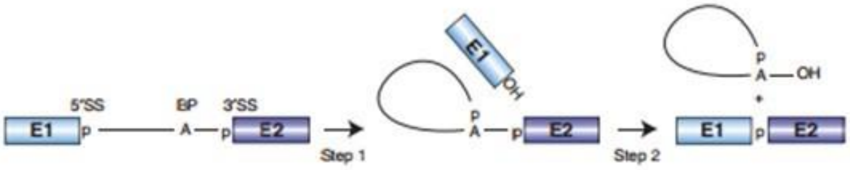


Figure 1. Spliceosome-mediated removal of an intron from a pre-mRNA. This figure shows a cartoon schematic of spliceosome-mediated removal of an intron. E1 and E2 are exons 1 and 2 flanking the intron, respectively. 5'SS and 3'SS are the intron's 5' and 3' splice sites, respectively. A is the adenosine located in the intron's branch point (BP). Adapted from Will, Luhrmann 2011.

likely contribute to primates' increased biological complexity relative to other vertebrates, as vertebrates have comparable numbers of protein coding genes. Aberrant alternative splicing events (and their deleterious impact on biological processes) have been implicated in various human diseases including myotonic dystrophy, dilated cardiomyopathy, autism spectrum disorder, and cancer (Cieply, Carstens 2015).

IV. Transcription, splicing, and translation of mitochondrial genes in eukaryotic cells:

Mitochondria are energy-transducing organelles in eukaryotic cells responsible for synthesizing the majority of cellular adenosine triphosphate (ATP). Proteins use ATP to conduct various essential cellular processes including biosynthetic reactions, cell motility, and cell divison (Taanman 1999). Cells in eukayotic organisms have different numbers of mitochondria depending on energy demands of their respective tissues. Each mitochondria contains multiple copies of maternally derived, ~16 kb circular genomes free of intronic regions (Ladd, Keeney, Govind, Bennett 2014). The mitochondrial genome's two strands (light and heavy) collectively have 37 total genes encoding 2 rRNAS, 22 tRNAs, and 13 polypeptides.

Transcription of mitochondrial genes is less understood than transcription of nuclear genes, but the two processes are thought to share many commonalities. Mitochondrial transcription involves initiation, elongation, and termination stages. Further, mitochondrial transcription requires contributions from a mitochondrial RNAP (POLRMT), a mitochondrial transcription factor A (TFAM), and mitochondrial transcription factors

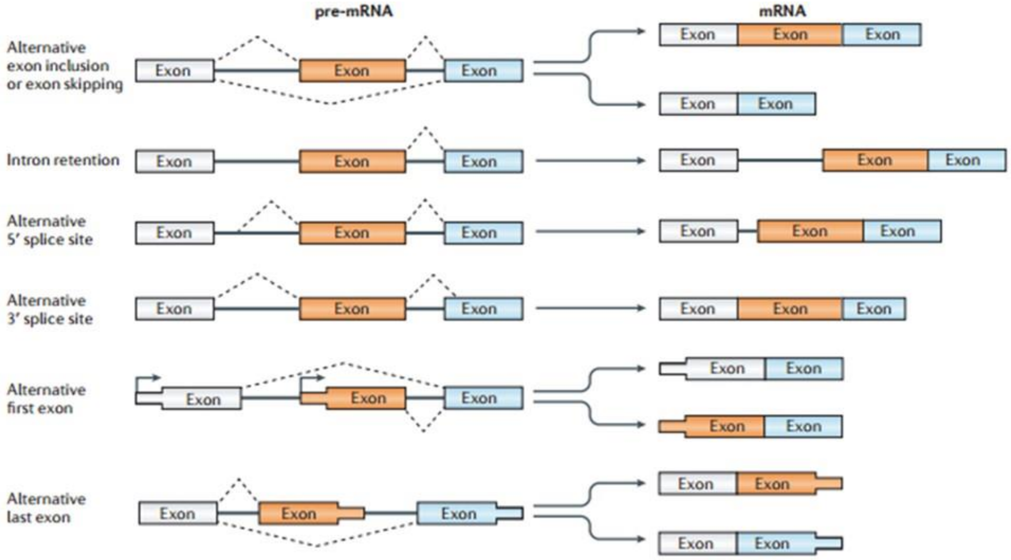

pre-mRNA
mRNA
Alternative exon inclusion or exon skipping
Intron retention
Alternative 5′ splice site
Alternative 3′ splice site
Alternative first exon
Alternative last exon
Exon

Figure 2. Alternative splicing events for a pre-mRNA molecule. This figure shows various alternative splicing events for a single pre-mRNA molecule that can be executed by the spliceosome. Adapted from Carpenter et al. 2014.

TFB1M or TFB2M (Asin-Cayuela, Gustafsson 2007). Other proteins may also contribute to during these steps, though there is less definitive evidence to support this.

During the initiation step, a transcription initiation complex comprised of POLRMT, TFB2M, and TFAM binds to one of three promoter regions (HSP1, HSP2, or LSP) found on the heavy and light strand of the mitochondrial genome. It is not fully understood how the mitochondrial transcription initiation complex recognizes DNA sequences in these promoter regions or begins transcription. However, TFAM is suspected to mediate structural alterations in mtDNA, unwinding it to expose transcription start sites to the initiation complex.

Fewer details are known about the elongation step. Recent finding suggest POLRMT binds a protein called mitochondrial transcription elongation factor to form an elongation complex (Posse et al. 2015). POLRMT is thought to function in the same way RNA Polymerases I-III do during their respective elongation processes, generating complementary ssRNA molecules from deoxyribonucleotide sequences. The initiation complex begins transcription from one of the heavy strand promoters (HSP1 or HSP2), or the light strand promoter (LSP). Different RNA products are generated depending on which promoter is used, as discussed in Figure 3.

When transcription is initiated at the HSP1 site, POLRMT stops transcribing mtDNA upon encountering a termination site (or specific sequence of nucleotides) at the end of the 16s rRNA sequence. This may involve contributions from a protein (mTERF1). It is unclear how POLRMT terminates transcription after beginning at HSP2 or LS, though other termination sites (and interacting proteins) are suspected (Asin-Cayuela, Gustafsson 2007). Splicing of mitochondrial transcripts is thought to require four enzymes

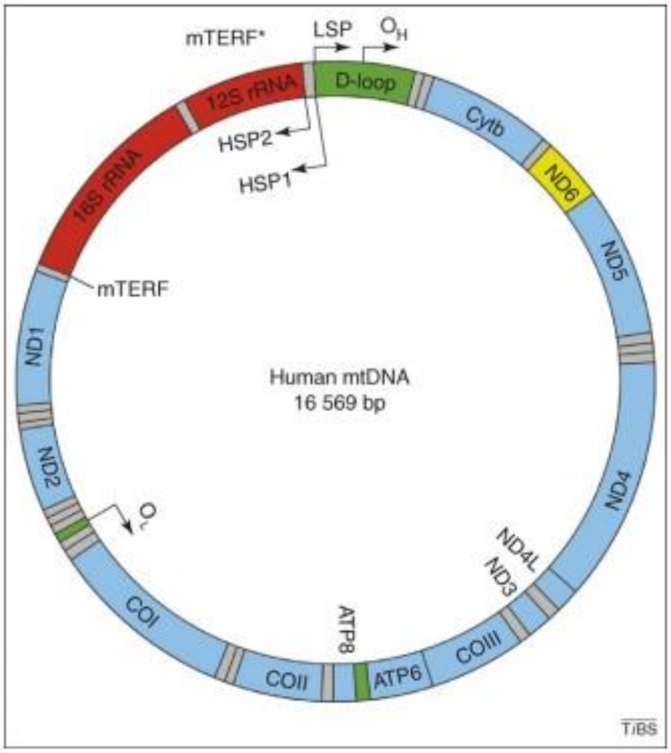


Figure 3. Mitochondrial transcription. This figure shows the mtDNA genome, and transcription initiation sites for the heavy (HSP1 and HSP2) and light (LSP) mitochondrial DNA strands. HSP1's corresponding RNA transcript terminates at the 3' end of the 16S rRNA region. HSP2's corresponding RNA transcript nearly incorporates the entire heavy strand sequence (including all of genes depicted in blue). Finally, LSP's corresponding RNA transcript includes the ND6 gene (in yellow) and primers for initiation of DNA synthesis at the heavy strand origin of replication (OH). Adapted from Asin-Cayuela, Gustafsson 2007.

(including mitochondrial RNase P) that mediate endonucleolytic excision of these pre-mRNA transcripts into smaller pieces (Smits, Smeitink, van den Heuvel 2010). These pieces are mitochondrial mRNAs (encoding rRNAs, tRNAs, mRNAs), or primers involved in initiation of DNA synthesis.

Translation of mitochondrial mRNAs involves initiation, elongation, and termination steps. Further, this process involves contributions from 1) mitochondrial-encoded rRNAs and tRNAs, 2) initiation, elongation, and termination translation proteins, 3) mitochondrial ribosomal proteins, and 4) mitochondrial aminoacyl-tRNA synthetases and methionyl-tRNA transformylases (Smits, Smeitink, van den Heuvel 2010).

During the initiation phase, mitochondrial ribosomes are thought to recognize unique mitochondrial mRNAs with their unstructured 5' sequences (as mitochondrial mRNAs do not have caps). The current model proposes a translation factor (mtIF3) promotes formation of an initiation complex, and the mitochondrial mRNA's start codon (with the nucleotide sequence AUG) is bound to the mitochondrial ribosome's P site. A different translation factor (mtIF2) facilitates entry of a tRNA carrying a methionine residue into the mitochondrial ribosome. This tRNA's anticodon binds the mitochondrial mRNA's start codon in the P site, as the tRNA still carries the methionine residue (Smits, Smeitink, van den Heuvel 2010).

During the elongation step, translation factor proteins assist the mitochondrial ribosome, tRNAS, and rRNAs in generating a polypeptide from the mitochondrial mRNA. At this point, the mitochondrial mRNA's 2nd codon is situated in the mitochondrial ribosome's A site. A tRNA with an anticodon complementary to the mitochondrial mRNA's

2nd codon is recruited into the mitochondrial ribosome's A site, and the tRNA's anticodon binds to the mitochondrial mRNA's 2nd codon. A peptide bond is formed between the amino acids carried by the tRNAs in the mitochondrial ribosome's A and P sites, followed by the mitochondrial mRNA being shifting three nucleotides to the left. This results in the tRNA bound to the mitochondrial mRNA's start codon being ejected from the mitochondrial ribosome, as it no longer carries an amino acid. During this shift, the mitochondrial mRNA's 2nd and 3rd codons were moved into the mitochondrial ribosome's P and A sites, respectively.

An iterative process ensues involving 1) recruitment of a tRNA with an anticodon complementary to the mitochondrial mRNA's codon located in the mitochondrial ribosome's A site, 2) formation of a peptide bond between the amino acids carried by the tRNAs in the mitochondrial ribosome's A and P sites, 3) the mitochondrial mRNA being shifted 3 nucleotides to the left, and 4) ejection of the tRNA hybridized to the mitochondrial mRNA's most 5' codon (as it no longer carries an amino acid). This process is repeated until a stop codon (with a sequence of UAA, UAG, AGA, or AGG) is shifted into the mitochondrial ribosome's A site (Smits, Smeitink, van den Heuvel 2010).

During the termination phase, a mitochondrial release factor (mtRF1a) recognizes the stop codon in the mitochondrial ribosome's A site, and triggers the release of the polypeptide from the mitochondrial ribosome. This involves hydrolysis of the ester bond linking the polypeptide chain to the tRNA in the P site. The ribosome and its associated proteins then dissociate, and the mitochondrial mRNA transcript is freed. The mitochondrial mRNA transcript is either translated into additional polypeptide copies by other mitochondrial ribosomes or degraded (Smits, Smeitink, van den Heuvel 2010).

V. Gene expression levels:

Gene expression is the process by which a gene's deoxyribonucleotide sequence is used to generate structural or functional gene products. These include RNAs and/or proteins. A given gene's expression level in a biological sample can be estimated by measuring the amount of its RNA transcripts in that sample. Gene expression levels can change in response to factors inherent to an organism (such as age, gender, and hormones), environmental factors (such as drugs, temperature, and light) the organism encounters (Arslan-Ergul, Adams 2014, Che, Gingerich, Lall, Howell 2002, Rhodes, Crabbe 2005, Podrabsky, Somero 2004, Rossel, Wilson, Pogson 2002), or an interaction between the internal and external factors. The temporal, developmental, topographical, histological, and physiological patterns in which a gene is expressed can provide clues to its biological role (Shena, Salon, Davis, Brown 1995).

Researchers have long hypothesized multiple genes' expression patterns contribute to phenotypic diversity (Romero, Ruvinsky, Gilad 2012). However, simultaneous measurement of multiple genes' expression levels was not common practice (due its technical infeasibility) until DNA microarrays were created for this purpose in the 1990s (Schena, Salon, Davis, Brown 1995). Next generation RNA-Sequencing has since been created in the late 2000's (Wang, Gerstein, Snyder 2009), and both techniques are now commonly used to measure gene expression levels.

VI. DNA microarrays and next generation RNA-Sequencing:

DNA microarrays are glass slides with covalently bound single stranded oligonucleotide probes (typically 60-mers in length) (Mantione et al. 2014). On a microarray designed to measure gene expression, each probe is complementary in sequence to part (or all) of an annotated gene's known or predicted RNA transcripts (Mantione et al. 2014). There are typically multiple probes assigned to each gene's RNA transcripts. The synthesis of these probes relies on known genomic sequence, and/or known or predicted open reading frames (Malone, Oliver 2011).

When using a DNA microarray to estimate a given sample's gene expression levels, researchers will typically 1) isolate that sample's total RNA or polyA+ RNA, 2) convert it into double stranded complementary DNA (dscDNA) via reverse transcription, 3) label the dscDNA with a fluorescent dye, 4) denature the labeled dscDNA, 5) hybridize the denatured cDNA strands to the DNA microarray's probes via complementary basepairing, 6) shine a laser on the DNA microarray's probes to excite the fluorescent dyes attached to hybridized cDNA strands, 7) record each probe's fluorescence intensity, 8) process and normalize all probes' fluorescence intensities, and 9) use all resultant fluorescence intensity values to estimate each gene's expression level. The presumption is each gene's expression level is positively correlated with its corresponding probes' fluorescence intensities. When a researcher wants to compare two samples' gene expression levels to each other on the same DNA microarray, they follow the steps above with the exception of labeling each sample's dscDNAs with a different colored fluorescent dye in step 3.

DNA sequencing involves determining the identity (and order) of nucleotides in part (or all) of a DNA molecule. Each sequenced read acquired from a DNA molecule reports this information. Current RNA-Sequencing workflows involve isolating a given sample's RNAs, converting them into dscDNAs, and generating sequenced reads from those dscDNAs' denatured strands. These sequenced reads undergo an alignment step, where their nucleotide sequences are compared to known genes' nucleotide sequences. Each sequenced read is aligned (or assigned) to the genomic location with the highest level of similarity to it in that comparison. Each gene's expression level is estimated based on the total number of sequenced reads that aligned to its transcribed regions.

When using RNA-Sequencing to estimate a given sample's gene expression levels, researchers will typically 1) isolate that sample's total RNA or polyA+ RNA, 1a) remove rRNAs from total RNA (in cases where polyA+ RNA is not used), 2) fragment isolated RNA transcripts into a distribution of smaller fragments, 3) convert RNA fragments into dscDNA molecules, 4) ligate sequencing adaptors to those dscDNA molecules, 5) amplify dscDNA molecules that were properly ligated with sequencing adaptors (by targeting sequences in the adaptors) using PCR, 6) denature all residual dscDNA molecules, 7) hybridize a portion of those dscDNA molecules' strands to a sequencing chip (via complementary basepairing between their sequencing adaptors and oligonucleotides on the chip), 8) generate sequenced reads of those dscDNA molecule's bound strands via a sequencer-specific protocol, 9) align sequenced reads to an organism's reference transcriptome and/or genome, and 10) use bioinformatic tools to estimate each gene's expression level using the aligned sequenced reads.

VII. Next generation RNA-Sequencing over DNA microarrays:

Next generation RNA-Sequencing has several advantages over DNA microarrays when measuring gene expression. First, RNA-Sequencing does not rely on probes generated using known (or predicted) RNA transcripts from an organism's annotated reference transcriptome and/or genome. RNA-Sequencing thereby enables simultaneous detection of both annotated and novel RNA transcripts (including novel alternatively spliced RNA transcripts). Second, RNA-Sequencing can detect a much larger range of gene expression levels compared to a standard whole genome microarray (Mantione et al. 2014). Intriguingly, RNA-Sequencing doesn't have known upper or lower limits for detecting gene expression levels (though detection of lowly expressed transcripts is intricately related to the number of sequenced reads generated as discussed later in this dissertation). Third, a given sample's RNA-sequencing reads can be used to: estimate each annotated gene's expression levels, estimate annotated exons' expression levels, identify novel RNA transcripts (and novel alternatively spliced RNA transcripts), detect coding and non-coding RNA transcripts, identify single nucleotide variants (SNVs), detect insertions and deletions (indels), and identify gene fusion events (Mantione et al. 2014). Separate DNA microarrays must be custom-made to separately estimate each annotated exon's expression level, identify SNVs, detect non-coding RNA transcripts, or reveal gene fusion events. Finally, a given sample's RNA-Sequencing data can be re-analyzed as an organism's annotated transcriptome and genome files are updated (Mantione et al. 2014).

There are several disadvantages to using next generation RNA-Sequencing over DNA microarrays. First, most current RNA-Sequencing platforms preferentially generate sequenced reads from longer RNA transcripts relative to shorter RNA transcripts

(Mantione et al. 2014). This can occur even when a longer and shorter RNA transcript have equal expression levels in the studied biological sample. As gene expression estimates rely on the number of sequenced reads that align to each gene's transcribed regions, this technical bias reduces the accuracy of those estimates. Mathematical corrections are often applied to account for this bias, but are unlikely to completely remove its effects. Second, the cost per sample is typically higher for an RNA-Sequencing experiment compared to a DNA microarray experiment, especially when considering laboratory reagents and data storage. For a given sample, a typical RNA-Sequencing data file is generally larger than 5 gigabytes (GB), whereas a typical DNA microarray file is usually <1 megabyte (MB) (Mantione et al. 2014). Third, analysis of RNA-Sequencing data generally requires more training and computer skills compared to analysis of microarray data (Mantione et al. 2014). All of these disadvantages will likely change as RNA-Sequencing technologies improve, much like what has been seen as microarray technologies were refined over the last 15 years.

VIII. Gene expression studies and hypothesis testing:

Gene expression studies compare measured gene expression levels between two or more groups, allowing identification of differentially expressed genes (DEG). DEGs are genes with statistically significantly different expression levels between groups. Over-representation analyses can be used to infer a set of DEGs' likely biological relevance. These analyses detect statistically significant associations between an input set of genes compared to predefined groups of genes known to influence various cellular processes.

Gene expression studies comparing disease and control sample groups can identify DEGs specific to the disease group (or groups), and reveal cellular processes associated with those DEGs using over-representation analyses. This approach can uncover perturbed cellular processes that may reflect disease pathology, and enables researchers to form novel hypotheses about which genes in a list of DEGs may functionally impact those cellular processes.

Selecting a candidate gene for hypothesis testing in downstream molecular biology experiments is rarely trivial, as gene expression studies often yield considerable options. Typically, researchers select a candidate gene that 1) was identified as a DEG, 2) was in a group of DEGs associated with cellular processes relevant to disease pathology, and 3) has known structural or functional properties that could plausibly influence one of these cellular processes.

Systems-level gene co-expression network analyses comparing disease and control sample groups provide separate criteria for candidate gene selection that can be used in conjunction with DEG analysis results. These analyses identify gene co-expression networks, or sets of genes clustered together based on similarities in their measured gene expression levels across samples, associated with disease status. Co-expressed genes are often functionally related, members of the same pathway or protein complex, or modulated by important regulatory transcriptional programs (Weirauch 2011). Genes comprising gene-co expression networks associated with disease status can be input to over-representation analyses, revealing cellular processes associated with those networks that may be relevant to disease pathology.

Gene co-expression network analyses provide more than a list of genes (comprising each identified network) associated with disease status. They also predict highly connected network hub genes most likely to functionally regulate their network's activities and associated cellular processes (Jeong, Mason, Barabasi, Oltvai 2001, Carter, Brechbuhler, Griffin, Bond 2004). Studies using gene co-expression network analyses have revealed networks associated with a polygenic trait and plausibly related cellular processes. Further, several of these networks' hub genes were previously linked to the polygenic trait using separate molecular biology techniques (Kogelman et al. 2014, Maschietto et al. 2015). Arguably more compelling, Horvath et al. showed siRNA-mediated reduction of Abnormal Spindle Microtubule Assembly (ASPM), a hub gene in a gene co-expression network associated with glioblastoma and mitosis, significantly reduced proliferation rates in glioblastoma tumor cells *in vitro* (Horvath et al. 2006).

Hub genes in a gene co-expression network associated with both disease status and cellular processes plausibly related to disease pathology are strong candidates for hypothesis testing. Hypotheses about a given hub gene's functional impact on these cellular processes can be tested using appropriate molecular biology assays in an *in vitro* or *in vivo* model system. Specifically, a researcher can test whether imitating a hub gene's expression level as it was observed in their disease sample group perturbs a cellular process in a manner consistent with what is known about disease pathology.

IX. Amyotrophic Lateral Sclerosis and its genetic epidemiology:

Amyotrophic Lateral Sclerosis (ALS) is a devastating neurodegenerative disease caused by the death of upper and lower motor neurons in the brain stem and spinal cord. The incidence of ALS in European populations is 2-3 people per year per 100,000 of the general population over the age of 15 years, with men at a slightly higher risk than women (Al-Chalabi, Hardiman 2013). Clinical features vary considerably across ALS patients, but always involve progressive muscle weakness and paralysis. Muscles in the hands and feet (as well as those involved in speaking and swallowing) often atrophy early in disease progression. The average life expectancy after diagnosis is between 2-5 years, with patients dying of respiratory failure (as the neurons innervating their diaphrams and other respiratory muscles die) (Al-Chalabi, Hardiman 2013). Unfortunately, current drug treatments cannot stop ALS disease progression and only extend life by a few months.

10% of patients report a first degree relative with ALS and receive a familial ALS (fALS) diagnosis, whereas the remaining 90% of patients report no family history (receiving a diagnosis of sporadic ALS [sALS]). Early familial aggregation studies, twin studies, and epidemiological studies suggested genetic factors contribute to both forms of ALS (Al-Chalabi et al. 2010, Chio et al. 2013, Fang et al. 2009, Wingo et al. 2011). While discerning fALS from sALS is extremely challenging using the traditional El Escorial clinical guidelines (Al-Chalabi 2013), genetic studies have revealed important differences between their molecular etiologies.

fALS is transmitted in a Mendelian fashion. Bonafide causal mutations have been identified within 9 genes using classic linkage and/or next generation DNA sequencing techniques, and there is preliminary evidence for causal mutations in an additional 15 genes to date (Renton, Chio, Traynor 2014). As sequencing costs decrease and more fALS pedigrees are studied, the number of identified causal mutations (and implicated genes) is expected to grow. As of 2014, causal mutations in the 9 thoroughly substantiated ALS genes account for ~67% of fALS in Caucasian patients as of 2014 (Renton, Chio, Traynor 2014).

Interestingly, causal mutations in those same 9 genes only account for ~11% of sALS in Caucasian patients (Renton, Chio, Traynor 2014). While it is likely causal mutations residing in other genes account for some portion of the remaining sALS cases, the majority of sALS is suspected to have polygenic and environmental contributions (Al-Chalabi, Hardiman 2013, Hardiman, Greenway 2007). Various exploratory approaches have been (and continue to be) adopted to elucidate sALS' polygenic contributions.

14 Genomewide association studies (GWAS) comparing allele frequencies at millions of single nucleotide polymorphisms (SNPs) in sALS cases vs. neurologically healthy controls have implicated numerous chromosomal regions, though only one association signal on chromosome 9 has been replicated across studies. That signal was instrumental in cloning Chromosome 9 Open Reading Frame 72 (C9ORF72), a gene that carries variable amounts of intronic repeat sequences across individuals. An excess number of repeats has been shown to cause ALS, and accounts for 7% of sALS in Caucasians as of 2014 (Renton, Chio, Traynor 2014). Larger sample sizes in ALS GWAS will likely increase the number of identified and replicated signals moving forward, as has been observed in other neurodegenerative diseases (Hollingworth et al. 2011, Nalls et al. 2011).

A recent study (Cirulli et al. 2015) compared whole exome sequencing data from 2,874 ALS patients and 6,405 healthy controls in search of rare (likely deleterious) coding variants (SNVs and indels) that may contribute to ALS pathology. 105 of these patients had fALS, whereas the rest of the patients had sALS. An excess of rare coding SNVs were discovered in 1) genes known to harbor fALS causal mutations and 2) genes associated with autophagy and neuroinflammation.

Other studies focused on identifying *de novo* mutations in affected sALS patients that were absent in his/her parents. Early studies using this approach found *de novo* mutations known to cause fALS in sALS-affected offspring via sequencing individual genes of interest (Alexander et al. 2002, Chio et al. 2011). A more recent study (Chesi et al. 2013) surveyed sALS-affected offsprings' protein coding regions for *de novo* mutations using whole exome sequencing. These researchers discovered sALS-affected offspring had a statistically significant excess of *de novo* mutations in chromatin regulator genes. While replication and functional validation is needed, this is an interesting finding and represents a promising approach for identifying rare genetic variants in sALS patients that may be relevant to disease pathology.

X. Scope of this dissertation project:

In the subsequent chapters of this dissertation project, I will detail our lab's ALS gene expression study comparing tissues from sALS patients and neurologically healthy controls. Our overarching goal was to identify sALS group-specific differences and their associated cellular processes that may be relevant to disease pathology.

Broadly, we used RNA-Sequencing and selected bioinformatics analyses to compare gene (and exon) expression levels in postmortem cervical spinal sections from human sALS patients and neurologically healthy controls. This enabled identification of sALS group-specific differences. We relied on systems biology analyses to identify cellular processes associated with these sALS group-specific differences. We presumed sALS group-specific differences in select genes' (or exons') expression levels holistically induced changes in their associated cellular processes, and some of these changes may have functionally contributed to disease pathology. Finally, we used molecular biology techniques to assess whether overexpression of an identified hub gene's expression level (as observed in our sALS sample group) perturbed an associated cellular process in a manner consistent with what is known about ALS pathology. Findings from this study have been used as rationale in the development of an ALS drug screening assay and the design of an ALS clinical trial.

CHAPTER 2: Tumor Necrosis Factor-mediated Inflammation is identified as a major abnormality in Postmortem sALS patients' cervical spinal sections

I. Introduction:

A comprehensive review of ALS gene expression studies dating back to 2001 was recently published (Heath, Kirby, Shaw 2006). Tissue types compared between ALS and control samples included: human bicep muscle, human lymphocytes, human and rodent spinal tissue containing disease-vulnerable motor neurons, rodent gastrocnemius, and isolated human and rodent spinal motor neurons.

There was a broad range in the number (14 to 1,182) and identity of ALS group-specific DEGs discovered in each study. Interestingly, a recurrent set of associated cellular processes emerged across over-representation analyses using these sets of DEGs. These included oxidative stress, mitochondrial dysfunction, apoptosis, cytoskeletal architecture, inflammation, RNA processing, and protein aggregation. Separate molecular biology assays revealed various ALS tissues from human patients and rodents carrying fALS causal mutations showed increased oxidative damage (Ferrante et al. 1997, Andrus, Fleck, Gurney, Hall 1998, Hall et al. 1998, Liu, Wen, Liu, Li

1999, Chang et al. 2008), abnormal mitochondrial morphology (Sasaki and Iwata 2007, Sasaki and Iwata 1996, Hirano, Donnenfeld, Sasaki, Nakano 1984, Higgins, Jung, Xu 2003, Kong and Xu 1998, Damiano et al. 2006, Mattiazzi et al. 2002), and elevated inflammation (Schiffer, Cordera, Cavalla, Migheli 1996, Nagy, Kato, Kushner 1994, Zhao, Beers, Appel 2013, Turner et al. 2004, Corcia et al. 2012, Henkel et al. 2004, Lewis et al. 2014, Alexianu, Kozovska, Appel 2001, Henkel, Beers, Siklos, Appel 2006, Poloni et al. 2000, Babu et al. 2008, Cereda et al. 2008, Elliott 2001, Hensley et al. 2002, Yoshihara et al. 2002). Increased inflammatory tumor necrosis factor (TNF-α) signaling (Poloni et al. 2000, Babu et al. 2008, Cereda et al. 2008, Elliott 2001, Hensley et al. 2002, Yoshihara et al. 2002) in ALS tissues may have therapeutic relevance, as it is known to carry out cell fate decisions that may contribute to motor neuron death (Probert 2015). Further, elevated TNF-α signaling been previously shown to induce motor neuron death (He, Wen, Strong 2002, Robertson et al. 2001, Terrado et al. 2000). Taken together, it is likely aberrations in these cellular processes contribute to ALS onset, progression, and symptoms.

Previous studies identified ALS group-specific gene co-expression networks associated with immune response, stress response, post-translational modifications, and neuroprotective processes (Holtman et al. 2015, Saris et al. 2009). Further, several of these networks' predicted hub genes were shown to functionally impact a cellular process associated with their respective network. Glutathione synthetase (GSS) and Aconitase2 (ACO2) both play important roles in stress response. GSS encodes a protein important for generation of glutathione, an antioxidant that prevents DNA damage from reactive oxygen species. ACO2 encodes an enzyme important for detecting oxidative stress, preserving mitochondrial DNA integrity, and preventing mitochondrial-mediated apoptosis (Kim et al. 2014). AXL receptor tyrosine kinase (AXL) plays an important role in neuroprotective processes, and was recently implicated in microglial phagocytosis of apoptotic cells and myelin (Holtman et al. 2015).

In this chapter, we used a combination of RNA-sequencing, bioinformatics tools, systems biology analyses, and *in vitro* molecular biology experiments to elucidate sALS group-specific nuclear gene expression level differences and assess their biological significance. Specifically, we set out to 1) identify cellular processes associated with sALS group-specific nuclear gene expression level differences that may be relevant to disease pathology, and 2) test hypotheses generated from those results assessing whether a network hub gene's expression level was functionally related to apoptosis in relevant cell types *in vitro*.

II. Methods:

Biological samples used: Human cervical spinal section tissues from 7 ALS patients and 8 neurologically healthy controls were procured from the National Disease Research Interchange (Philadelphia, PA). We defined neurologically healthy controls as individuals that were not diagnosed with ALS or any other neurodegenerative disorder. In addition to ALS disease status, we received information about each individual's age (at time of death), ethnicity, and gender. Unfortunately, we did receive information about any individual's history of medication use or postmortem interval, which may have influenced our measured gene expression levels. We chose to study isolated RNA from postmortem human cervical spinal sections, as they included disease-vulnerable motor neurons (as well as astrocytes and microglia).

Each sample's frozen tissue embedded in OCT was shipped to us on dry ice. NDRI provided age, ethnicity, gender, and disease status data for each sample. On average, ALS patients were 67.71 years old at death (standard deviation of 7.99 years), whereas neurologically healthy controls were 69.75 years old at death (standard deviation of 11.29 years). There was no statistically significant difference in age between the two groups as assessed by a Student's T test (p=0.697). All 7 ALS patients were Caucasian, with 4 Males and 3 Females. 6 of the neurologically healthy controls were Caucasian, and 2 were African American. There were 4 males and 4 females in the neurologically healthy control sample group. Each sample's age, ethnicity, and gender information can be found in Table 1.

Total RNA isolation: For each sample, twenty 20-micron cross sections of cervical spinal tissue embedded in OCT were iteratively cut using a Cryostat. The cross sections were placed in a tube of Qiazol within the Cryostat, and then repeatedly passed through a 19G hypodermic needle attached to a 1 mL sterile syringe to lyse cell membranes. Total RNA was extracted following the miRNeasy Mini (Qiagen) kit workflow, including the optional on-column DNase treatment step to prevent downstream DNA-derived sequencing reads. Eluted total RNA was further purified using the RNeasy Micro kit (Qiagen) to remove organic contaminants. We elected to use total RNA (instead of polyA+ RNAs only) for generating RNA-sequencing libraries. This allowed us to measure pre-mRNAs, mRNAs, lncRNAs, tRNAs, and mitochondrial RNAs (mtRNAs). We selected the Qiagen miRNeasy Mini kit over similar kits, as Qiagen RNA isolation products

Table 1: Sample demographics

	Age	Ethnicity	Gender
Patients			
ALS1	70	Caucasian	Male
ALS2	67	Caucasian	Male
ALS3	80	Caucasian	Female
ALS4	57	Caucasian	Male
ALS9	75	Caucasian	Female
ALS10	64	Caucasian	Female
ALS14	61	Caucasian	Male
Controls			
CTL6	80	African-American	Male
CTL8	67	Caucasian	Female
CTL16	66	Caucasian	Female
CTL22	54	Caucasian	Male
CTL23	65	African-American	Female
CTL24	83	Caucasian	Male
CTL25	59	Caucasian	Male
CTL27	84	Caucasian	Female

performed well at preserving RNA quality in a recent comparative analysis with other RNA isolation products (Sellin, Kiss, Smith, Oris 2014).

For each sample, smaller RNA transcripts (<100 nucleotides) were lost in a purification step using the RNeasy Micro kit columns. This included miRNAs, piRNAs, siRNAs, snRNAs, and snoRNAs. We used this purification step as 1) organic contaminants may interfere with downstream rRNA removal, and 2) sequencing smaller RNA transcripts requires Illumina sequencing parameters (shorter single end reads) incompatible with those that best suited for our scientific questions.

RNA quantification and quality assessments: Each sample's isolated total RNA was next quantified using a Nanodrop 2000c spectrophotometer (Thermo Scientific), and its quality was assessed using the Experion automated electrophoresis system (Bio-Rad). Bio-Rad's Experion calculated an RNA Quality Index (RQI) score for each sample via comparing three portions of the sample's electrophoretic profile to a manufactured standard of degraded RNAs. RNA Quality Index (RQI) values ranged between 1-10, with increasing values representing higher quality RNA with less degradation.

Sample inclusion criteria: We used 500 ng of high quality total RNA (RQI score ≥7) to construct each sample's RNA-sequencing library. The Illumina Truseq® Stranded Total RNA HT kit required between 100-1,000 ng of total RNA per sample, and was designed to accept RNA with any level of degradation (RQI score 1-10).

Using a larger amount of input total RNA buffers the inevitable loss of isolated RNA transcripts across the RNA-sequencing library preparation kit's reaction and purification steps. Using high quality RNA (characterized by less degraded RNA transcripts) increases the chance of generating sequencing reads that preserve longer stretches (more nucleotides) of their originating RNA molecules. In general, these sequencing reads have a higher probability of aligning to the human genome in a downstream step compared to sequencing reads containing shorter stretches of their originating RNA molecules.

Taken together, using a larger amount of high quality RNA for each sample likely increased 1) the number and variety of isolated RNA transcripts represented in their final RNA-Sequencing library, and 2) the proportion of their sequencing reads that aligned to the reference genome. Concurrently, this likely increased our downstream gene expression level estimates.

RNA-Sequencing library preparation kit overview: We used the Illumina Truseq® Stranded Total RNA HT kit to generate each sample's final RNA-sequencing library. This kit converted fragmented pieces of each sample's isolated RNA transcripts into dscDNAs with Illumina sequencing adaptors ligated to their 5' and 3' ends, hereby referred to as sequenceable dscDNA molecules.

Denatured strands from these sequenceable dscDNA molecules were bound to the Illumina NextSeq 500 flowcell and underwent amplification prior to sequencing. All details related to these processes are provided in a subsequent section. Ultimately, the Illumina NextSeq 500 workflow produced two sequencing reads from each bound strand that was amplified and sequenced, hereby referred to as Read 1 and Read 2. Each of these reads reported the identity of 150 nucleotides beginning at opposite ends of the cDNA fragment contained in each sequenced strand. Figure 4 shows where these sequenced reads were generated from in each sequenced strand.

RNA-Sequencing library preparation kit workflow: The Illumina Truseq® Stranded Total RNA HT kit converted fragmented pieces of each sample's isolated RNA transcripts into sequenceable dscDNA molecules. This multi-step process involved 1) removal of rRNAs from total RNA, 2) RNA fragmentation using divalent cations and heat, 3) dscDNA generation, 4) addition of a single A nucleotide to the 3' end of both cDNA strands in each dscDNA molecule, 5) ligation of Illumina i5 and i7 sequencing adaptors with single T nucleotide overhangs to both ends of each dscDNA molecule, and 6) enrichment PCR amplification to increase the ratio of sequenceable dscDNA molecules to dscDNA molecules ligated with one or zero sequencing adaptors. DNA purification steps were performed after steps 3, 5, and 6.

rRNA comprises 80-90% of total RNA in human cell types surveyed (O'Neil, Glowatz, Schlumpberger 2013; Wilhelm, Landry 2009). The Illumina Truseq® Stranded Total RNA HT kit removed rRNAs from each of our sample's total RNA via Ribo-Zero technology shown in Figure 5. This rRNA removal step ensured 80-90% of our downstream RNA-sequencing reads didn't correspond to rRNA transcripts. This involved 1) denaturing each sample's total RNA, 2) hybridizing rRNA molecules to oligonucleotides (attached to magnetic probes) with complementary sequences, 3) placing all sample tubes on a magnet, and 4) allowing time for the magnetized beads (with attached rRNA-probe complexes) to be pulled to the side of each tube. Each sample's supernatant containing unbound RNAs was then collected for proceeding steps.

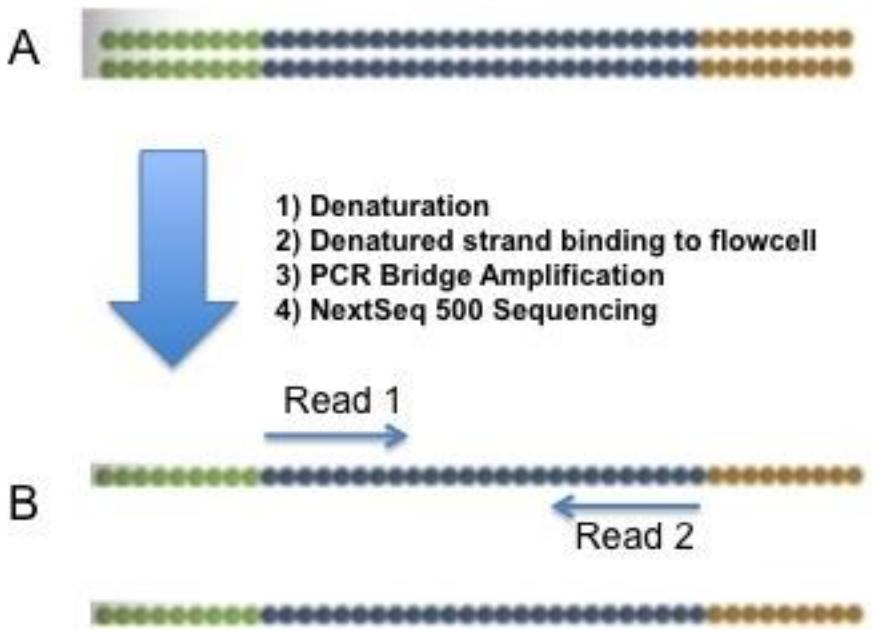

A
1) Denaturation
2) Denatured strand binding to flowcell
3) PCR Bridge Amplification
4) NextSeq 500 Sequencing
Read 1
B
Read 2

Figure 4. Sequenceable dscDNA molecule and sequencing reads. This figure shows what part of the sequenceable dscDNA molecule strand is read to generate Read 1 and Read 2 sequences. Part A shows a sequenceable dscDNA molecule generated by converting a fragment of a single stranded RNA transcript into a dscDNA molecule (blue dots) flanked by double stranded Illumina i5 (green dots) and i7 (yellow dots) sequencing adaptors. The sequenceable dscDNA molecule was denatured, enabling one or both of its denatured strands to bind to the Illumina NextSeq 500 flowcell. A PCR bridge amplification was performed, followed by NextSeq 500 sequencing. Part B shows where Illumina NextSeq 500 sequencing reads 1 and 2 come from in a sequenced strand. Each sequencing read was 150 nucleotides in length.

RNA transcripts were next fragmented to a size range between 120-220 nucleotides (with a median fragment size of 150 nucleotides) using divalent cations and heat. Based on annotated (known) RNA transcripts in the hg19 human genome, the average length of a mature mRNA transcript is 2,227 nucleotides long (Kim et al. 2013). RNA fragmentation ensured sequenceable dscDNA molecules generated from isolated RNA transcripts were amenable to Illumina Nextseq 500 sequencing parameters using 150 nucleotide reads. Fragmented RNAs were subsequently primed with random hexamers for first strand cDNA synthesis.

First strand synthesis was performed using standard components (reverse transcriptase, reaction buffer, nucleotides) and Actinomycin D to prevent DNA-dependent DNA synthesis. This increased the likelihood resultant sequencing data had a minimal number of reads derived from DNA molecules.

Second strand synthesis using standard reagents (DNA polymerase, RNase H, reaction buffer, dATPs, dCTPs, dGTPs) and dUTPs in place of dTTPs was next. The DNA polymerase incorporated dUTPs instead of dTTPs into the second strand, thereby marking that strand. During a downstream enrichment PCR amplification step, that DNA polymerase didn’t incorporate nucleotides past the first dUTP it encountered in the second strand. This ensured only the first strand of cDNAs (antisense to the original RNA transcript it was reverse transcribed from) were ultimately sequenced. This allowed us to align each sequenced read to the DNA strand its corresponding RNA molecule was transcribed from. This was especially important for identifying antisense transcripts, determining which strand of lncRNAs was transcribed, and identifying overlapping genes’ boundaries when they resided on the same strand.

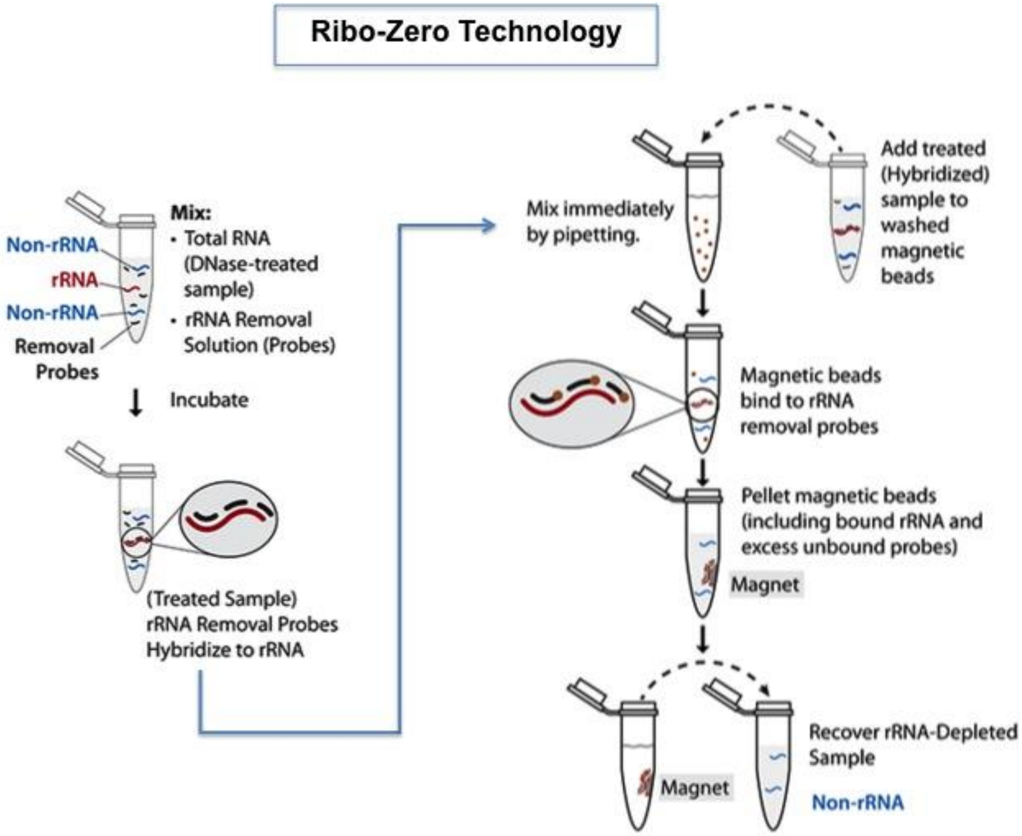

Ribo-Zero Technology
Non-rRNA
rRNA
Non-rRNA
Removal Probes
Mix:
• Total RNA (DNase-treated sample)
• rRNA Removal Solution (Probes)
Incubate
(Treated Sample) rRNA Removal Probes Hybridize to rRNA
Mix immediately by pipetting.
Add treated (Hybridized) sample to washed magnetic beads
Magnetic beads bind to rRNA removal probes
Pellet magnetic beads (including bound rRNA and excess unbound probes)
Magnet
Magnet
Recover rRNA-Depleted Sample
Non-rRNA

Figure 5. Ribo-Zero technology. This figure shows a cartoon schematic of how ribosomal RNAs (rRNAs) are removed from each sample's total RNA. DNase-treated total RNA is mixed together with rRNA removal solution. This solution contains magnetic rRNA removal probes carrying oligonucleotide sequences complementary to rRNA species. After these probes' oligonucleotides hybridize to their complementary rRNAs in the total RNA mixture, magnetic beads are added into the entire solution. These magnetic beads bound the magnetic rRNA removal probes regardless of whether their oligonucleotides were hybridized to rRNAs or not. Tubes were then placed on a magnet, and RNAs unbound to these rRNA removal probes were collected for proceeding steps. Illumina; [accessed 2016 Feb 26]. Adapted from http://www.illumina.com/products/ribo-zero-rrna-removal-human-mouse-rat.html.

dscDNAs next underwent consecutive ligation-based reactions to add i5 and i7 Illumina sequencing adaptors to their 5’ and 3’ ends. In the first reaction, a single dATP nucleotide was ligated to the 3’ end of both cDNA strands in each dscDNA molecule. Next, both ends of each dscDNA molecule were ligated to a double stranded i5 or i7 Illumina sequencing adaptor via complementary dATP-dTTP hybridization. This was successful as these adaptors had single dTTP overhangs that complemented the dscDNA molecule’s single dATP overhangs. Each sample was assigned a different Illumina i5 sequencing containing a unique 8-nucleotide index sequence, and one identical i7 Illumina sequencing adaptor.

As each sample’s sequenceable dscDNA molecules possessed a different 8-nucleotide index sequence in their respective i5 adapters, we were able to mix multiple sample’s sequenceable dscDNA molecules together prior to Illumina NextSeq500 sequencing. We could then denature and sequence denatured dscDNA strands from these multi-sample pools. Most importantly, resultant Read 1 and 2 sequencing reads generated from each sequenced strand could be assigned to their correct sample using that strand’s index sequence.

The final step in this kit’s workflow involved an enrichment PCR amplification. This amplification used standard reagents (DNA polymerase, buffer, dNTPs), and served to increase the number of sequenceable dscDNA molecules relative to dscDNA molecules improperly ligated with one or zero sequencing adaptors. This ensured a negligible amount of unsequenceable denatured strands (from improperly ligated dscDNA molecules) bound the NextSeq 500 flowcell.

DNA purification steps were performed after steps 3, 5, and 6. These relied on Agencourt AMPure XP beads. Agencourt Ampure XP beads are paramagnetic polystyrene beads coated in carboxyl molecules. They are suspended in a solution containing polyethylene glycol and salt, as these components cause DNA to bind to the AMPure XP beads' carboxyl groups provided proper stoichiometry. Specified amounts of the Ampure XP bead solution were added to each sample after the steps listed above. DNA molecules were given time to bind the paramagnetic beads. Sample tubes were next placed on a magnet, allowing the DNA-paramagnetic bead complexes to be pulled to the side of the tube. The supernatant was discarded, and the beads (complexed to DNA molecules) were washed twice with 70% ethanol. DNA molecules were finally eluted from the paramagnetic beads using 1X TE and gathered for the next step.

Quality control: We next ran several quality control steps to ensure we generated high quality RNA-Sequencing libraries, as this increased the likelihood our downstream gene expression estimates were accurate. We assessed whether each sample had 1) the expected size distribution of sequenceable dscDNA molecules (mean fragment size around 260 basepairs) generated by the Illumina Truseq® Stranded Total RNA HT kit, and 2) comparable molar amounts of sequenceable dscDNA molecules across samples.

We determined the size distribution of each sample's sequenceable dscDNA molecules using the BioRad Experion (Bio-Rad). We measured each library's molar amount of sequenceable dscDNA molecules using a CFX96 Touch Real-Time PCR Detection System (Bio-Rad) and a KAPA Library Quantification Kit (Kapa Biosystems).

This kit provided qPCR amplification reagents for absolute quantitation, including 6 standards containing different molar amounts of a 452-basepair DNA fragment ligated with Illumina sequencing adaptors to generate a standard curve.

NextSeq 500 pre-sequencing steps: We employed Cofactor Genomics (St. Louis, MO) to perform two Illumina NextSeq 500 sequencing runs to generate all samples' sequencing data. Prior to the first run, we created a composite RNA-Sequencing pool by combining an equimolar amount of sequenceable dscDNA molecules from 6 of our samples. This same approach was used for the remaining 9 samples prior to the second run several months later.

An aliquot of each pooled RNA-Sequencing library containing sequenceable dscDNA molecules was first denatured before being washed over the surface of a NextSeq 500 flowcell. The flowcell had two types of covalently bound oligonucleotides. The first type was complementary to each strand's Illumina i5 sequencing adapter, and the second type was complementary to each strand's Illumina i7 sequencing adapter. Approximately 400 million denatured strands attached to the flowcell via their i5 or i7 Illumina sequencing adaptor hybridizing to a complementary oligonucleotide covalently bound to the flowcell surface. Figure 6A shows a denatured strand's i5 Illumina sequencing adaptor hybridizing to a complementary oligonucleotide covalently bound to the flowcell surface.

DNA polymerase and unlabeled dNTPs (dATP, dCTP, dGTP, and dTTP) were next added to the flowcell. The polymerase extended each covalently bound oligonucleotide hybridized to a denatured strand, iteratively incorporating nucleotides complementary to those in the denatured strand. This created a complementary copy of each denatured strand that hybridized an oligonucleotide bound to the flowcell. Further, each complementary copy was attached to the flowcell via the covalently bound oligonucleotide at its base. Each resultant double stranded molecule was then denatured, and the original denatured strand was washed away. These events can be seen in Figure 6B and 6C.

Each covalently bound strand next underwent a three-step PCR bridge amplification process that can be seen in Figure 7. First, the free end of each strand physically bent over, allowing its Illumina i5 or i7 sequencing adaptor to hybridize to its complementary oligonucleotide covalently bound to the flowcell. Second, DNA polymerase and unlabeled dNTPs (dATP, dCTP, dGTP, dTTP) were added. The polymerase extended each recently hybridized covalently bound oligonucleotide, iteratively incorporating complementary nucleotides to the strand that just attached to it. Finally, resultant double stranded molecules were cleaved, causing the ends of each strand that were not covalently bound to the flowcell to be released. Both strands took on a vertical position.

These three steps were iteratively repeated to generate each strand's cluster. A cluster is a clonal population of a single bound strand, comprised of up to one thousand covalently bound copies of it. After the final bridge PCR amplification cycle was complete, the reverse strand from every double stranded molecule in each cluster was cleaved and washed away. The remaining forward strands were used for Read 1 generation.

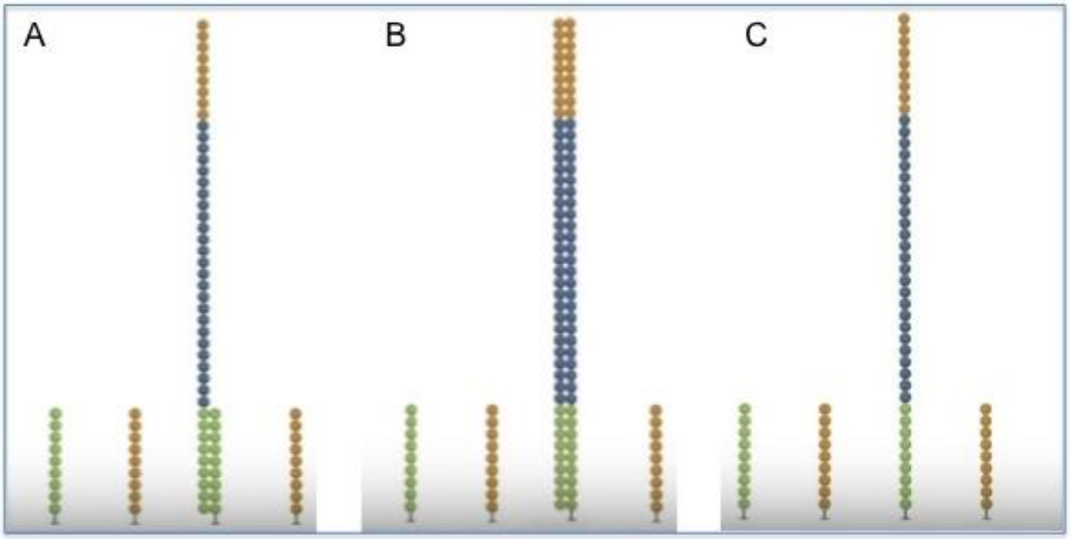


Figure 6. Creation of a complementary copy of each denatured strand. This figure shows a cartoon schematic of how a denatured strand from a sequenceable dscDNA molecule hybridized to the flowcell (A), was copied (B), and was washed away (C). Its complementary copy remained attached to the flowcell via a covalently bound oligonucleotide at its base. Illumina; [accessed 2016 Feb 26]. Adapted from video on http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html.

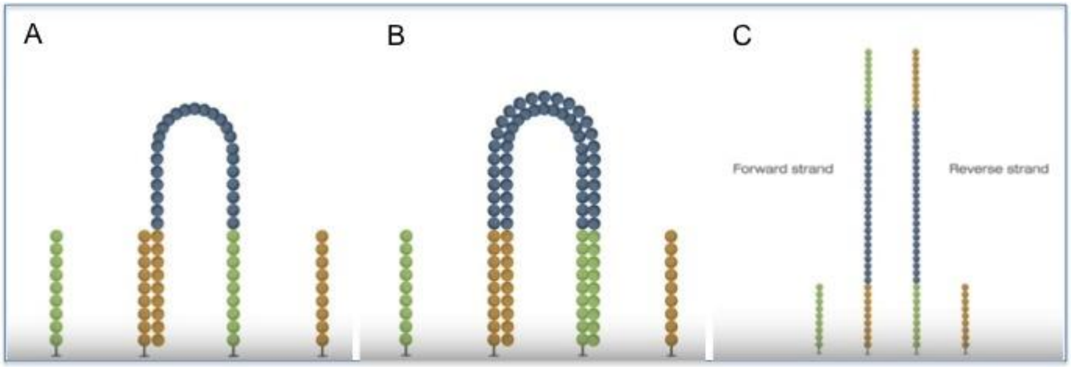


Figure 7. PCR Bridge amplification of each covalently bound strand. This figure shows a cartoon schematic of (A) hybridization between a bound strand's sequencing adaptor and a complementary flowcell oligonucleotide after the strand bent over, (B) a double stranded molecule after a complementary copy of that strand was generated, and (C) two covalently bound strands after their free (non-bound ends) were released from the flowcell. This three-step process was repeated to generate up to one thousand copies of the original bound strand. Illumina; [accessed 2016 Feb 26]. Adapted from video on http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html.

NextSeq 500 sequencing overview and Read 1 preparatory steps: The remaining steps of Illumina NextSeq 500 sequencing ultimately produced two sequencing reads for each bound strand that was clonally amplified to generate each cluster. These two sequencing reads, Read 1 and Read 2, each reported 150 nucleotides beginning at opposite ends of each bound strand's cDNA fragment. A visual depiction of these reads can be seen in Figure 4.

Prior to beginning Read 1's "sequencing by synthesis" process, two types of sequencing primers were added to the flowcell. The first type was complementary to a portion of each strand's Illumina i5 sequencing adapter, and the second type was complementary to a portion of each strand's Illumina i7 sequencing adapter. A sequencing primer hybridized to a portion of the sequencing adapter (i5 or i7) in every strand's unbound end via complementary basepairing. An example of the first type of sequencing primer hybridizing to its complementary i5 sequencing adaptor in a strand's unbound end can be seen in Figure 8A.

Read 1 "Sequencing by Synthesis": Read 1 acquisition involved 150 iterative "sequencing by synthesis" cycles following the same four steps. First, DNA polymerase and 4 unique reversible terminator nucleotides complementary to each covalently bound strand’s dNTPs (dATP, dCTP, dGTP, dTTP) were added to the flowcell. Each reversible terminator nucleotide was unique in that it emitted a different wavelength when excited by a laser in a downstream step. Second, each of the four fluorescently labeled reversible terminator nucleotides competed to extend the chain of nucleotides (beginning adjacent to the recently hybridized sequencing adaptor) towards the flowcell. Only the reversible terminator nucleotide complementary to the covalently bound strand's dNTP at each

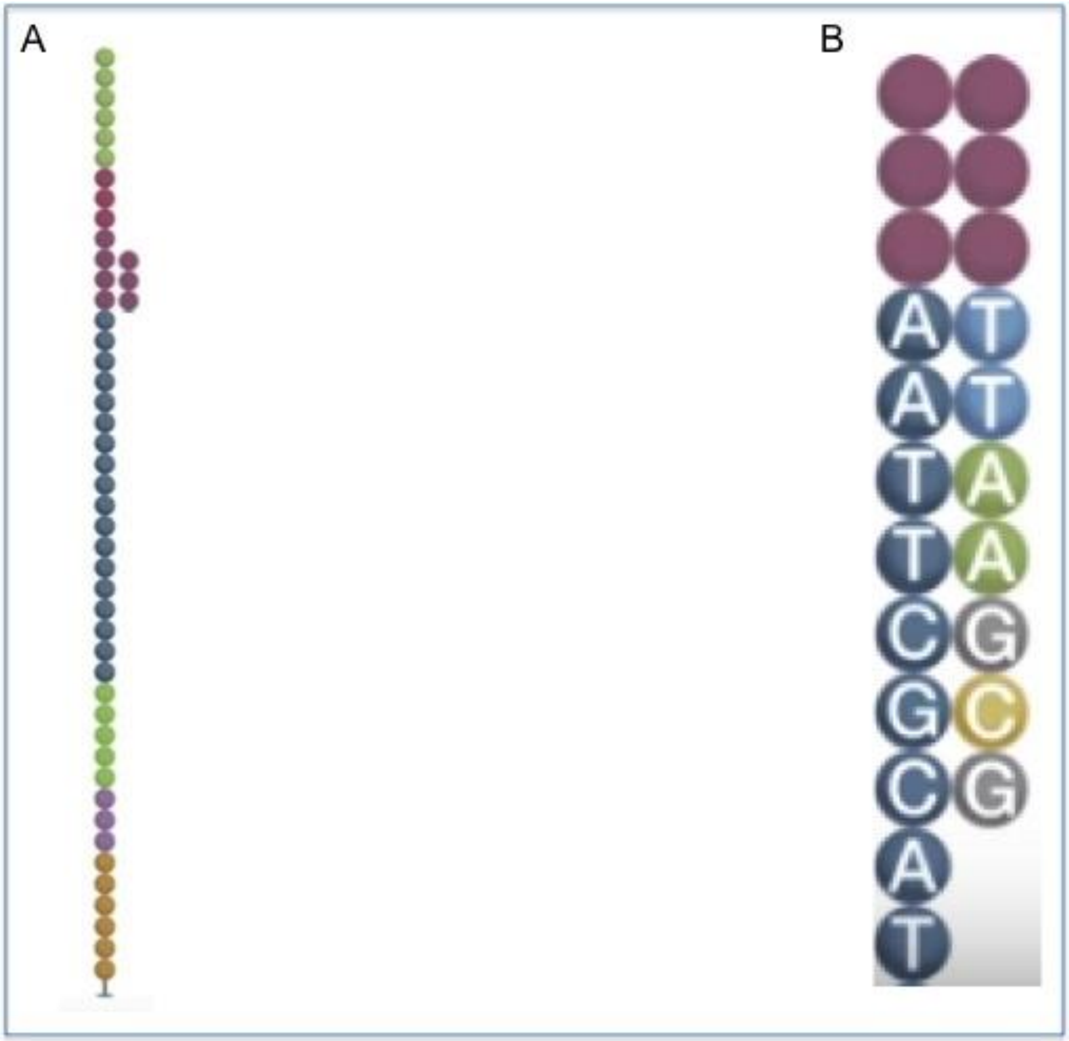


Figure 8. Read 1 “Sequencing by Synthesis”. This figure shows a cartoon schematic of (A) hybridization between a bound strand’s sequencing adaptor and a complementary sequencing primer (purple circles), and (B) a close-up of that strand’s growing Read 1 oligonucleotide chain comprised of complementary reversible terminator nucleotides after the first 7 cycles. These reversible terminator nucleotides are color coded by their unique emission wavelengths when shone with a laser. Illumina; [accessed 2016 Feb 26]. Adapted from video on http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html.

position was incorporated (with the exception of rare errors). Third, a laser was shone at every cluster on the flowcell. A detector recorded the emission wavelength (and signal intensity) from incorporated reversible terminator nucleotides at that position in every strand in every cluster. The signal-to-noise ratio for detection of reversible terminator nucleotides' emitted wavelengths was improved in each cluster as a result of the strand's clonal amplification. Lastly, the inhibitor on each incorporated reversible terminator nucleotide was removed, allowing this iterative process to repeat until the 150th cycle was completed. A cartoon depiction of a Read 1 molecule after the first 7 cycles can be seen in Figure 8B.

Index Read acquisition: Index read acquisition involved 8 iterative "sequencing by synthesis" cycles, enabling us to retrieve the 8-nucleotide index sequence in the i5 sequencing adapter of every strand in every cluster. As mentioned in a previous section, each of our samples was assigned a different Illumina i5 sequencing adapter containing a unique 8-nucleotide index sequence (and an identical Illumina i7 sequencing adapter) following generation of dscDNAs. Retrieving the index sequence in each strand's i5 sequencing adapter allowed us to determine which sample that sequenced strand came from. More importantly, it enabled us to determine which sample Read 1 and Read 2 sequencing reads (generated from that same strand) belonged to.

Prior to the 8 "sequencing by synthesis" cycles, the recently generated chain of 150 nucleotides (corresponding to Read 1) hybridized to each covalently bound strand in each cluster was washed away. An index sequencing primer complementary to a portion of each strand's i5 sequencing adaptor was added to the flowcell, and allowed time to hybridize to its corresponding sequence.

Eight iterative “sequencing by synthesis” cycles were carried out following the same four steps. First, DNA polymerase and 4 unique reversible terminator nucleotides complementary to each covalently bound strand’s dNTPs (dATP, dCTP, dGTP, dTTP) were added to the flowcell. Each reversible terminator nucleotide was unique in that it emitted a different wavelength when excited by a laser in a downstream step. Second, each of the four fluorescently labeled reversible terminator nucleotides competed to extend the chain of nucleotides (beginning adjacent to the recently hybridized index primer) towards the flowcell. Only the reversible terminator nucleotide complementary to the covalently bound strand's dNTP at each position was incorporated (with the exception of rare errors). Third, a laser was shone at every cluster in the flowcell. A detector recorded the emission wavelength (and signal intensity) from incorporated reversible terminator nucleotides at that position in every strand in each cluster. The signal-to-noise ratio for detection of reversible terminator nucleotides’ emitted wavelengths was improved in each cluster as a result of the strand’s clonal amplification. Lastly, the inhibitor on each incorporated reversible terminator nucleotide was removed, allowing this iterative process to repeat until the 8th cycle was completed.

Read 2 preparation and acquisition: Prior to beginning Read 2’s “sequencing by synthesis” process**,** the recently generated chain of 8 nucleotides (corresponding to the index read) hybridized to each covalently bound strand in each cluster was washed away. The remaining steps prior to Read 2 acquisition can be seen in Figure 9.

First, the free end of each strand then physically bent over, allowing its Illumina i5 or i7 sequencing adaptor to hybridize to its complementary oligonucleotide covalently bound to the flowcell. Next, DNA polymerase and unlabeled dNTPs (dATP, dCTP, dGTP, dTTP) were added. The polymerase extended each recently hybridized covalently bound oligonucleotide, incorporating nucleotides complementary to the strand that just attached to it. Finally, resultant double stranded molecules were cleaved, causing the ends of each strand that were not covalently bound to the flowcell to be released. Both strands took on a vertical position.

Each resultant forward strand was cleaved and washed away. The remaining reverse strand remained unaltered. Two types of sequencing primers were added to the flowcell. The first type was complementary to a portion of each strand's Illumina i5 sequencing adapter, and the second type was complementary to a portion of each strand's Illumina i7 sequencing adapter. A sequencing primer hybridized to a portion of the sequencing adapter (i5 or i7) in every reverse strand's unbound end via complementary basepairing.

Read 2 acquisition involved 150 iterative “sequencing by synthesis” cycles following the same four steps as Read 1 acquisition. Each covalently bound reverse strand used for Read 2 acquisition was an inverted complementary copy of the strand used for Read 1 acquisition. Read 2 thereby began at the opposite end of each bound strand’s cDNA fragment (directly adjacent to the Illumina i5 or i7 adaptor) relative to where Read 1 began.

Data processing, FastQ files, and phred scores: Proprietary Illumina software algorithms were used to process the emission wavelength and signal intensity from incorporated reversible terminator nucleotides at every position in every strand in every cluster during the acquisition of Read 1, Index Read, and Read 2. For each strand that

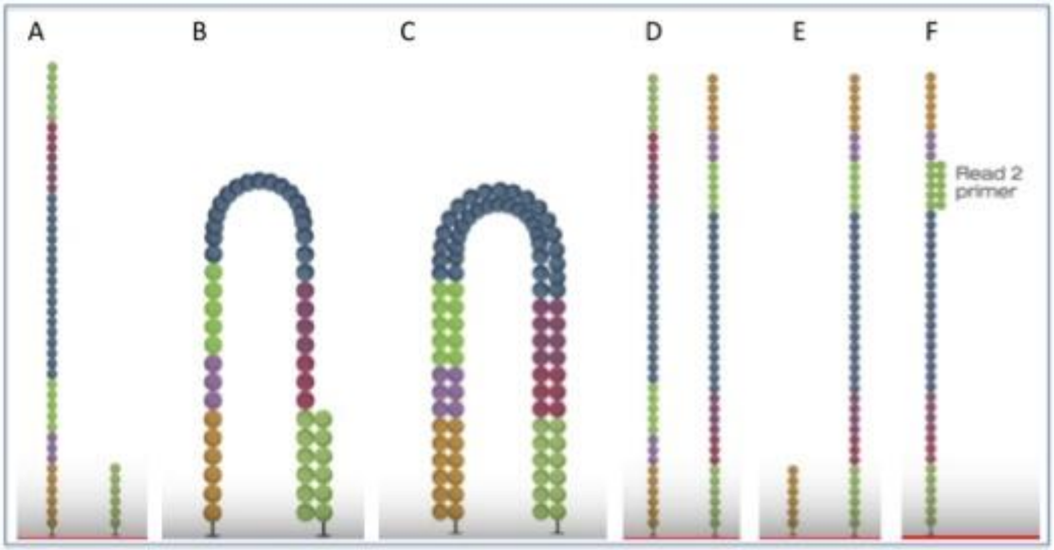


Figure 9. Steps prior to Read 2 “Sequencing by Synthesis”. This figure shows a cartoon schematic of (A) the covalently bound strand used for Read 1 generation after its index read was washed away, (B) hybridization between this strand’s sequencing adaptor and a complementary flowcell oligonucleotide after the strand bent over, (C) a double stranded molecule after a complementary copy was generated, (D) two covalently bound complementary strands after their free (non-bound ends) were cleaved from the flowcell, (E) removal of the forward strand, and (F) hybridization of a Read 2 sequencing primer to the bound strand’s sequencing adaptor prior to Read 2 Sequencing by Synthesis. Illumina; [accessed 2016 Feb 26]. Adapted from video on http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html.

was clonally amplified to generate each cluster, corresponding Read 1 and Read 2 150 nucleotide sequences (hereby referred to as a paired end read) were reported. Each paired end read was assigned to the correct sample using their sequenced strand's index read.

For each sample, all paired end reads' Read 1 and Read 2 sequences were sent in separate FastQ text files. Each Read 1 and Read 2 sequence reported the identity of all 150 sequenced nucleotides and each nucleotide's associated PHRED quality score. This PHRED quality score signified the probability that nucleotide's reported identity was the result of a sequencing error. Table 2 details several possible PHRED scores and their corresponding error probabilities.

FastQC for initial quality check: Each sample's FastQ files were separately input into FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to identify data quality issues pertaining to the sequencer and/or the RNA-Sequencing library itself. Quality was assessed across a range of metrics including average PHRED quality score per position across reads, overall GC content, sequence length distribution, overrepresented sequences, and duplicate read frequency.

Processing of sequencing reads prior to alignment: For each sample, we processed all paired end reads to increase their probability of aligning to the hg19 human reference genome in a proceeding step. A small portion of each sample's paired end reads were generated from sequenced strands containing a cDNA fragment <150 nucleotides in length. 150 nucleotide Read 1 and Read 2 sequencing reads generated from these strands necessarily contained nucleotides from an illumina i5 or i7 adaptor as a result.

Table 2: PHRED quality scores

PHRED Quality Score	Probability of Sequencing Error	Probability Reported Nucleotide is Accurate
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Further, nucleotides towards the 3’ end of Read 2 sequencing reads are known to have lower PHRED scores as a result of Illumina sequencing chemistry degradation towards the end of a run.

For each paired end read, we used Trimmomatic (Bolger, Lohse, and Usadel 2014) to remove nucleotides 1) belonging to Illumina i5 or i7 sequencing adaptors and/or 2) with a PHRED score <20 found in the 3’ end of their Read 1 or Read 2 sequences. This ensured paired end reads didn’t fail downstream alignment to the hg19 human reference genome because they contained 1) Illumina sequencing adaptor nucleotides that didn’t match any sequence in the hg19 human reference genome or 2) an excess of incorrectly identified nucleotides (as a result of sequencing errors) that didn’t match the hg19 human reference genome. We only retained reads that were ≥ 50 nucleotides in length after removing these select nucleotides. Each sample’s processed Read 1 and Read 2 sequences were output into new Read 1 and Read 2 fastQ text files for downstream analyses.

FastQC for pre-alignment quality check: Each sample’s Read 1 and Read 2 fastQ text files output by Trimmomatic were separately input into FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to ensure Illumina sequencing adaptor sequences were successfully removed, and the average PHRED score for nucleotides towards the 3’ end of Read 2 sequencing reads increased.

Alignment of paired end reads: We selected Tophat2 (Kim et al. 2013), an open source program specially designed for RNA-Sequencing data, to attempt alignment of each sample’s paired end reads to the hg19 human reference transcriptome then hg19 human reference genome. We first downloaded hg19 human reference transcriptome (known RNA transcripts) and genome (known primary sequence) text files from the Illumina iGenomes UCSC hg19 directory on the Tophat2 webpage (https://ccb.jhu.edu/software/tophat/igenomes.shtml).

In the context of this RNA-Sequencing experiment, alignment involved comparing the nucleotide sequences of each paired end read (Read 1 and 2) to the nucleotide sequences of all known RNA transcripts, 22 autosomes, X chromosome, Y chromosome, and mitochondrial genome. Each paired end read's alignment reflected what part of the genome that paired end read's corresponding RNA molecule was transcribed from.

A paired end read qualified for alignment if it did not exceed Tophat2's default quality thresholds for the number of nucleotides that 1) didn't match the reference genome (as a result of biological variation and/or sequencing errors), or 2) were not present in the reference genome.

Prior to Tophat2's first step, we used the Burrows-Wheeler Aligner (Li and Durbin 2010), an open source software program, to align 5 million of each sample's paired end reads to the hg19 human reference transcriptome. This allowed us to estimate the average size (and standard deviation) of cDNA fragments in each sample's sequenced strands. Tophat2 creators reported their software aligns a larger number of each sample's paired end reads to the hg19 human reference transcriptome and genome when these calculated metrics are provided relative to when they are not.

For each sample, Tophat2 attempted to align each paired end read (Read 1 and Read 2 from each sequenced strand) to the hg19 human reference transcriptome then hg19 human reference genome using a 5-step process shown in Figure 10. First, Tophat2 attempted to align (or map) each paired end read to all known transcripts in the hg19 human reference transcriptome. If the complete sequences of both reads in a paired end read mapped within the boundaries of a known transcript, that paired end read was aligned to that transcript. Second, Tophat2 attempted to map each unaligned paired end read to all known exons in the reference genome. If the complete sequences of both reads in a paired end read mapped within the boundaries of a known exon, that paired end read was aligned to that exon.

Prior to the next step, Tophat2 generated a list of putative spliceosome splice sites (GT and AG) across introns in the hg19 human reference genome. Third, Read 1 and Read 2 in each unaligned paired end read were broken into smaller segments no greater than 25 nucleotides in length. Tophat2 attempted to align these segments to all known exons in the hg19 human reference genome. If multiple segments from a given paired end read mapped to exons separated by one or more introns flanked with putative splice sites, those segments were aligned to those exons. Fourth, Tophat2 concatenated sequences flanking putative splice sites in the hg19 human reference genome, and then attempted to map all unaligned segments to them. These concatenated sequences did not belong to known exons in the genome. Aligned segments in steps 3 and/or 4 were re-joined to make full-length reads. Fifth, TopHat2 attempted to re-map any portion of an aligned paired end read mapped to an intron to exonic sequence. For each sample, Tophat2 output an accepted_hits.bam file reporting where each aligned paired end read mapped in the genome.

Calculation of sequencing alignment metrics: We used Picard Tools (http://picard.sourceforge.net), an open source software package, to calculate select

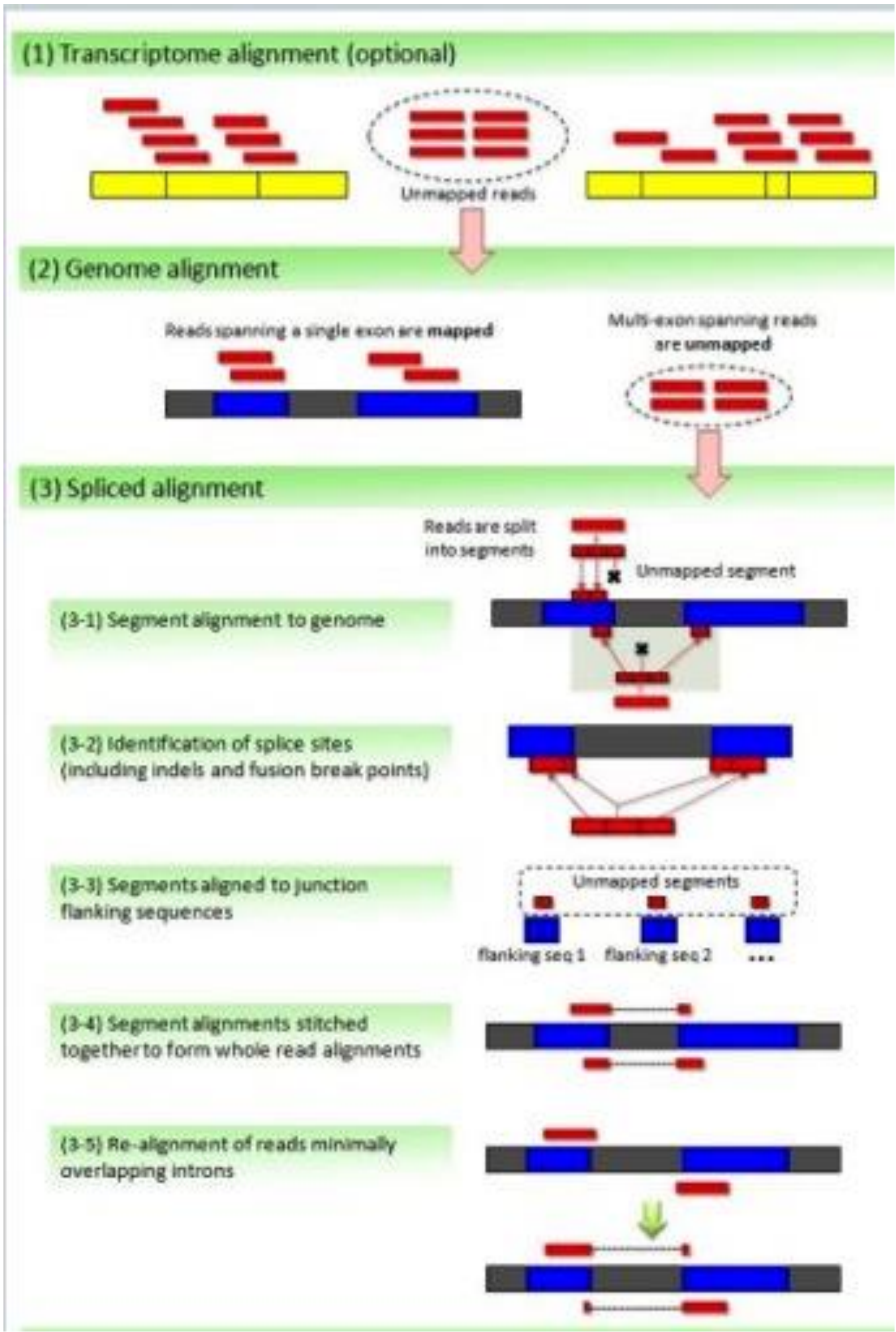

(1) Transcriptome alignment (optional)
Unmapped reads
(2) Genome alignment
Reads spanning a single exon are mapped
Multi-exon spanning reads are unmapped
(3) Spliced alignment
Reads are split into segments
Unmapped segment
(3-1) Segment alignment to genome
(3-2) Identification of splice sites (including indels and fusion break points)
(3-3) Segments aligned to junction flanking sequences
Unmapped segments
flanking seq 1
flanking seq 2
...
(3-4) Segment alignments stitched together to form whole read alignments
(3-5) Re-alignment of reads minimally overlapping introns

Figure 10. Tophat2 alignment of each sequenced read. This figure shows a cartoon schematic of Tophat2's method for aligning a RNA-Sequencing read to a reference transcriptome then genome. This illustration depicts alignment of a single end read (equivalent to a Read 1 in our study), but the alignment process is the same for paired end reads. Adapted from Kim et al 2013.

alignment metrics for each sample's aligned paired end reads. For each sample, Picard Tools reported: the total number of sequenced nucleotides (contained in their paired end reads) that Tophat2 attempted to align to the hg19 human reference genome, the % of sequenced nucleotides that were successfully aligned to the hg19 human reference genome, the % of sequenced nucleotides that aligned to mRNA species, the % of sequenced nucleotides that aligned to rRNA, tRNA, or mtRNA species, and the % of sequenced nucleotides that aligned to intronic or intergenic regions.

For each sample, Picard Tools calculated these metrics via comparing aligned paired end reads to 1) the Illumina iGenomes refFlat text file that contained all known RNA transcripts, introns, and intergenic regions in the hg19 human reference genome, and 2) an interval file I created containing known rRNA, tRNA, and mtRNA transcripts in the hg19 human reference genome.

fALS causal point mutation analysis: We assessed whether any of our sALS samples carried any of 471 known pathogenic coding variants (contained across 21 different genes) mutually reported to cause fALS in three separate databanks (Abel, Powell, Andersen, Al-Chalabi 2013, Landrum et al. 2014, Stenson et al. 2014). All of these genes have been reported to carry mutations that cause ALS in a recent publication (Renton, Chio, Traynor 2014). This involved using a variety of open source bioinformatic software programs including the Genome Analysis Tool Kit (GATK) (McKenna et al. 2010, DePristo et al. 2011, Van der Auwera et al. 2013), STAR (Dobin et al. 2013), and Picard Tools (http://picard.sourceforge.net). For each sALS sample, GATK identified qualifying sequence variants, or alleles that don't match the reference genome, following a 4-step pipeline (https://www.broadinstitute.org/gatk/guide/article?id=3891).

This pipeline involved 1) attempting to align each sALS sample's paired end reads to the hg19 human reference genome using STAR, 2) removing each sALS sample's paired end reads that aligned to an identical genomic location, 3) re-assigning PHRED scores to all nucleotides in each sALS sample's remaining aligned paired end reads, and 4) identifying each sALS sample's sequence variants that passed GATK's default quality filters.

We used STAR, an open source program specially designed for RNA-Sequencing data, to attempt to align each sALS sample's paired end reads to the hg19 human reference genome. STAR used each sALS sample's Read 1 and Read 2 FastQ files from Trimmomatic as input, and output a file containing their aligned paired end reads for downstream analysis. The GATK developers recommended STAR over Tophat, as it yielded a higher proportion of validated sequence variants within their pipeline in an unpublished comparative analysis.

We used Picard Tools to remove each sALS sample's paired end reads that aligned to an identical genomic location (aka duplicate reads). For each qualifying genomic location, one aligned paired end read was retained for downstream analysis. This step was taken as duplicate reads can result from PCR amplification reactions during RNA-Sequencing library preparation as opposed to transcriptional events that occurred in that sample's tissues. Further, sequencing variants identified in such duplicate reads may reflect a PCR artifact (i.e. a DNA polymerase erred and that nucleotide was propagated into multiple PCR copies) as opposed to a biological feature in that sample's tissues.

GATK next re-assigned a PHRED score to every nucleotide in each sALS sample's remaining aligned paired end sequencing reads. GATK reported Illumina sequencing workflows assign biased PHRED scores to select nucleotides depending on what cycle that nucleotide was sequenced and what nucleotides preceded it. GATK re-assigned new PHRED scores to nucleotides that were likely affected by these biases, and did not alter other nucleotides' PHRED scores.

Finally, GATK identified each sALS sample's qualifying sequence variants using all remaining aligned paired end reads. A sequence variant qualified for identification if it 1) had a GATK GQ score of ≥20 (indicating a 99% or better chance the identity of that reported sequence variant was correct at that genomic position), and 2) passed all GATK's default quality filters designed to prevent false positives from technical artifacts. These default quality filters disqualified sequence variants if there was evidence for technical bias related to 1) sequencing depth, 2) where the sequence variant was located in aligned paired end reads containing it, and 3) the quality of alignment for all aligned paired end reads containing that sequence variant. GATK output a text file for each sample listing all their identified sequence variants. We used these lists to determine whether any sALS sample contained a sequence variant that represented any of the 471 pathogenic fALS mutations.

Gene expression estimate overview: For each sample in this study, a given gene's expression level reflected the quantitative amount of its RNA transcripts in that sample's postmortem cervical spinal sections. For a given sample, we presumed each gene's expression level was preserved by a proportionate number of paired end reads that aligned to its transcribed regions in that sample's sequencing data. This concept is illustrated in Figure 11. Our downstream gene expression estimates relied on this presumption. For each sample, we estimated gene expression levels for all annotated genes in the hg19 human reference genome using several bioinformatics workflows described in later sections of this document.

Illumina NextSeq 500 sequencing parameters and accuracy of gene expression estimates: Relative to single end sequencing, paired end sequencing has been shown to increase both the accuracy of downstream gene expression estimates (Salzman, Jiang, Wong 2011) and the number of detected splice sites (Chhangawala, Rudy, Mason, Rosenfeld 2015). Relative to shorter read options, longer reads (≥100 nucleotides) have been shown to increase both the number of aligned sequenced reads (Cho et al. 2014) and detected splice sites (Chhangawala, Rudy, Mason, Rosenfeld 2015). These findings support our decision to use Illumina NextSeq500 paired end 150 nucleotide reads as opposed to single end (or shorter paired end) reads, and likely increased the accuracy of our downstream gene expression estimates.

In an RNA-Sequencing study, the total number of sequenced reads obtained for a given sample is positively correlated with 1) the likelihood genes with lower expression levels are represented in their sequencing data, and 2) how accurate their gene expression estimates are likely to be. These relationships reflect Illumina's (and many other sequencing platform's) sampling procedure, where denatured strands from only a portion of each sample's sequenceable dscDNA molecules are sequenced. These concepts are illustrated in Figure 12.

According to a recent study, 45-65 million RNA-sequencing reads generated from

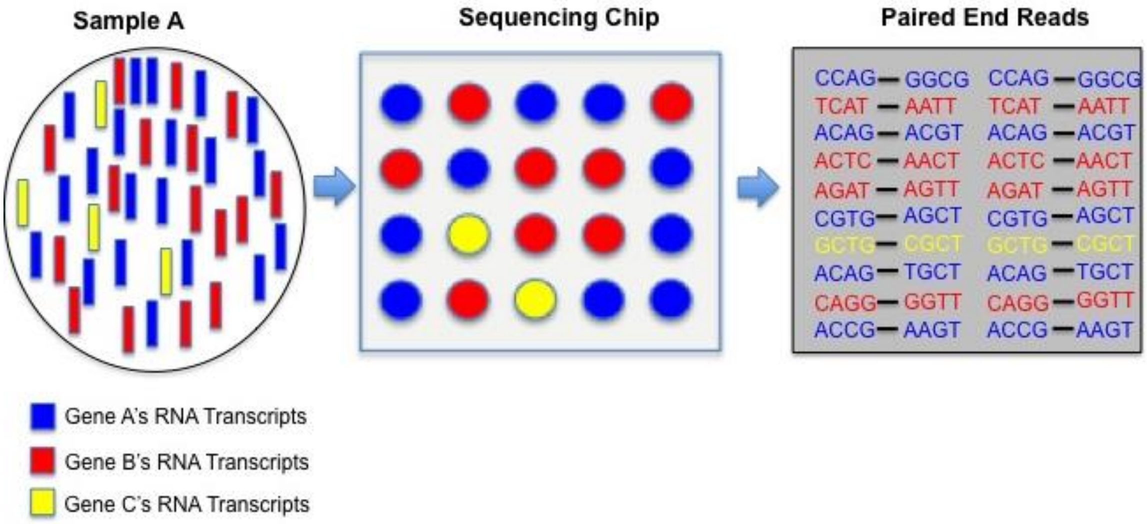


Figure 11. Presumption underlying our gene expression estimates. This figure shows a cartoon schematic of the presumption underlying our gene expression estimates. On the left are a biological sample's RNA transcripts from three different genes (color-coded red, blue, and yellow). Each gene's RNA transcripts were converted into a proportional amount of sequenceable dscDNA molecules. In the middle, these molecules' denatured strands hybridized to all the available spots (or covalently bound complementary oligonucleotides) on the NextSeq500 flowcell. Those spots are color-coded by what gene's denatured strands bound to them. On the right is the sample's paired end reads color-coded by which gene they aligned to, with read 1 and read 2 sequences on the left and right of the black dash marks, respectively. For this sample, each gene's expression level was preserved in their sequencing data. 50%, 40%, and 10% of the total RNA transcripts and aligned paired end reads corresponded to the blue, red, and yellow genes, respectively.

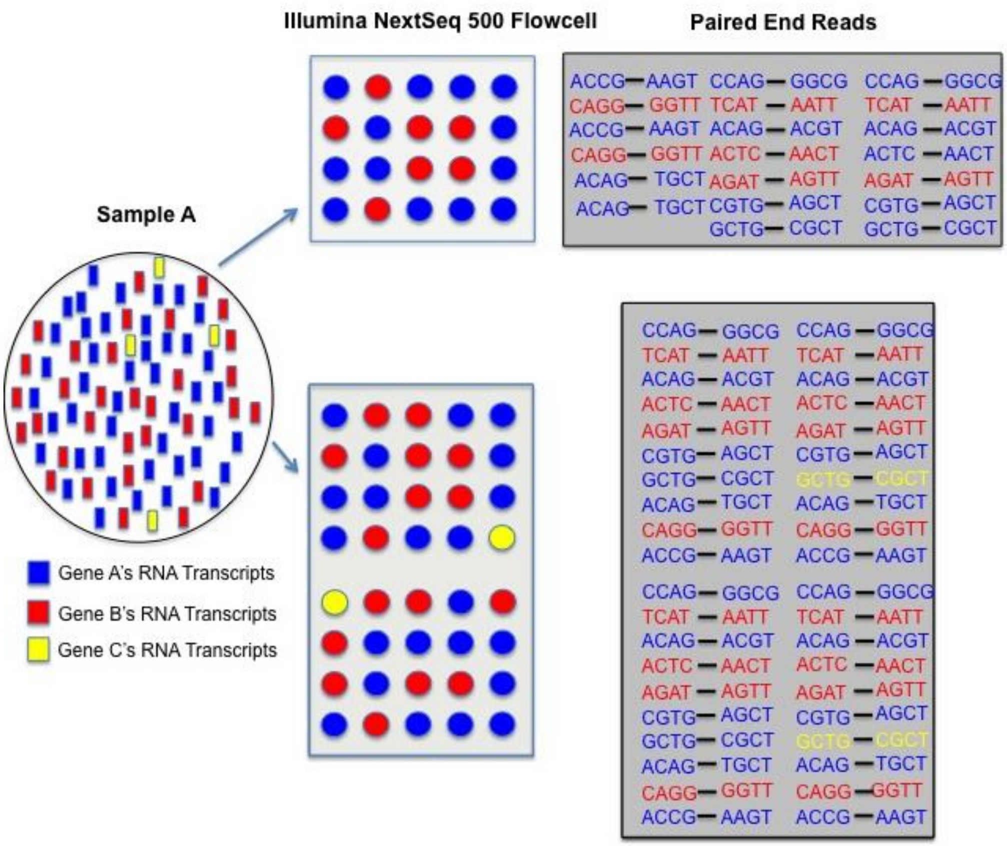


Figure 12. Read count and detection of lowly expressed genes. This figure illustrates how increasing the number of paired end reads obtained for a given sample is positively correlated with 1) the likelihood a gene with a low expression level is represented in their sequencing data, and 2) how accurate gene expression estimates are likely to be. The top and bottom half of the figure shows what happens when a smaller or larger amount of sequencing reads are generated for a given sample, respectively. On the left are a biological sample's RNA transcripts from three different genes (color-coded red, blue, and yellow). Each gene's RNA transcripts were converted into a proportional amount of sequenceable dscDNA molecules. In the middle top half of the figure, denatured strands from the sample's more highly expressed genes (red and blue) hybridized to the smaller number of flowcell spots allotted to this sample. This occurred as a result of their being more prevalent relative to denatured strands from the yellow gene, and thereby being more likely to bind the flowcell. In the middle bottom half of the figure, denatured strands from all three genes hybridized to the larger number of flowcell spots allotted to this sample. On the right are paired end reads from each scenario. Each paired end read is color-coded by the gene it aligned to, and Read 1 and Read 2 sequences are on the left and right of the dash mark, respectively. In the top half of the figure, the lowly expressed yellow gene didn't have any corresponding aligned paired end reads in the dataset. Further, the blue and red genes' aligned paired end reads made up 65% and 35% of their sequencing data, while their RNA transcripts made up 55% and 45% of this sample's total RNA. In the bottom scenario, each gene's expression level was accurately preserved in their sequencing data. 55%, 45%, and 5% of the total RNA transcripts and aligned paired end reads corresponded to the blue, red, and yellow genes, respectively.

input total RNA (after rRNA-depletion) offers a comparable detection level for protein coding genes relative to a standard Agilent DNA microarray (Zhao et al. 2014). There is not a consensus number of RNA-sequencing reads one should appropriate to each sample to achieve highly accurate gene expression estimates for very lowly expressed genes. An early study proposed 200 million reads per sample was necessary to detect the full range of expressed human RNA transcripts, including those from very lowly expressed genes (Tarazona et al. 2011). The ENCODE (Encyclopedia of DNA elements) consortium more recently generated 214 million 100 nucleotide paired end reads per sample from H1 human embryonic stem cells and performed a saturation analysis (Djebali et al 2012). They reported 36 and ~80 million paired end reads per sample were necessary to accurately estimate genes with expression levels corresponding to FPKM values of >10 and <10, respectively. A recent study (Marinov et al. 2014) related the number of RNA transcript copies in single GM12878 cells to reported FPKM values, and found one transcript copy per cell corresponds to an FPKM value of approximately 10. Annotated genes with FPKM values <10 thereby average less than one transcript copy per cell, but are expressed in enough of our spinal cells for detection.

Sequencing depth, biological replicates, and DEG identification: Provided a study goal of identifying DEGs between groups and a limited budget, researchers will often decrease the number of sequenced reads per sample while increasing the number of biological replicates per group. DEG identification analyses model intra-group variability in each gene’s aligned sequenced read counts across samples prior to inter-group comparisons to identify DEGs. Including more biological replicates per group enables more accurate intra-group variability estimates, and has been shown to increase the number of DEGs identified between groups. A recent analysis revealed ~72% more DEGs (6,000 vs. 3,500) were identified when comparing MCF7 breast cancer cell groups (treated or untreated with 17β-estradiol) comprised of 7 samples vs. 3 samples (Liu, Zhou, White 2014). Every MCF7 breast cancer cell line sample had 30 million total sequenced reads in that study. Another study (Zhang et al. 2014) reported a similar increase (~59%) in the number of DEGs identified when comparing lymphoblastoid cell line groups comprised of 8 samples vs. 3 samples. This study showed further expanding each lymphoblastoid cell line groups' size from 8 to 14 biological replicates increased the number of identified DEGs by only ~12.5% (2000 vs. 2250). This suggests diminishing returns in how many additional DEGs are identified when including more than 8 biological replicates per group. Taken together, these findings supported our decision to acquire >55 million paired end reads per sample, and compare 7 sALS samples to 8 neurologically healthy control samples for DEG identification.

HTSeq-Count: We used HTSeq-Count (Anders, Pyl, Huber 2015), an open source program specially designed for RNA-Sequencing data, to report the total number of paired end reads that aligned to each annotated gene's transcribed regions. A paired end read was counted for an annotated gene if the majority (or all) of its Read 1 and Read 2 sequences aligned to that gene's transcribed regions. Figure 13 shows various hypothetical alignments of a sequenced read to a fictitious gene_A, and whether HTSeq-Count would count that sequenced read. For each sample, HTSeq-Count produced a matrix with all annotated genes and their corresponding paired end read count values.

EdgeR and DEG identification: We used EdgeR (Robinson, McCarthy, Smyth 2010), an open source program specially designed for RNA-Sequencing data, to detect sALS-group specific DEGs. EdgeR first normalized each sample's paired end read counts (from HTSeq-Count) for each annotated gene, producing pseudo counts. Annotated genes' pseudo counts in our sALS and neurologically healthy control sample groups were then directly compared to identify sALS group-specific DEGs.

EdgeR followed a multi-step process to normalize each annotated genes' paired end read counts across samples. This process involved mathematical normalizations accounting for 1) differences in the total number of paired end read counts between samples, 2) differences in RNA species represented in each samples' sequencing data, and 3) overdispersion in annotated genes' paired end read counts across samples.

Differences in the total number of paired end read counts between samples could reflect each sample's total number of RNA-Sequencing reads as opposed to differences in annotated genes' expression levels. Assume two of our samples' RNA-Sequencing libraries had an equal number of sequenceable dscDNA molecules corresponding to a given annotated gene (suggesting that gene had an equal expression level in both samples' postmortem cervical spinal sections).

If more total sequencing reads were generated for one of those samples, that sample's sequencing data would likely have more total paired end read counts align to that annotated gene's transcribed regions relative to the other sample. This is because denatured strands from that annotated gene's sequenceable dscDNA molecules were appropriated more Illumina NextSeq500 binding spots (covalently bound oligonucleotides) to potentially hybridize to for sequencing. For each sample, we used 500 ngs of input total RNA to generate their RNA- Sequencing library. However, the RNA

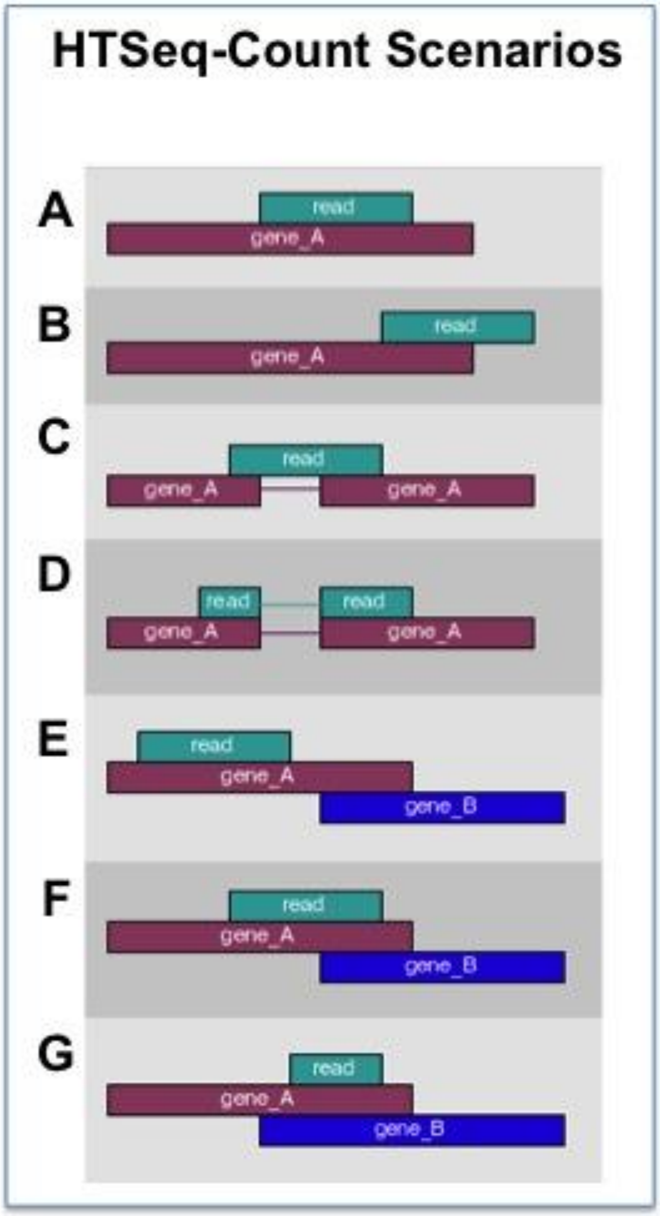


Figure 13. HT-Seq count scenarios. This figure shows a visual schematic of various hypothetical alignments of a sequenced read (in teal) to a fictitious gene_A's transcribed regions. HT-Seq Count would count that sequenced read for gene_A in scenarios A-E. It would not count the sequenced read for gene_A in scenarios F-G, as it is unclear whether that read's corresponding cDNA fragment was generated from gene_A or gene_B. While these scenarios depict a single end sequenced read (equivalent to a Read 1 in our study), the counting process is the same for each paired end read. The program considers the same criteria (the proportion of a sequenced read aligning to a given gene's transcribed regions) for counting. Counting reads in features with htseq-count. Simon Anders; [accessed 2016 Feb 26]. http://www-huber.embl.de/users/anders/HTSeq/doc/count.html.

transcripts comprising each sample's input total RNA varied. Differences in a given annotated gene's aligned paired end read counts across samples could reflect varying RNA compositions in each sample's total input RNA (and Illumina's sampling procedure) as opposed to gene expression differences. This concept is illustrated in Figure 14.

RNA-Sequencing data is often overdispersed. This means the variance of sequenced read counts aligned to each gene across each group's samples often exceeds what is expected using a Poisson distribution. Use of the Poisson distribution to model sequenced read counts across each group's samples could lead to a large number of false positives in downstream DEG analyses. EdgeR used a negative binomial distribution to model each annotated gene's paired end read counts across each group's samples to better account for this variance.

EdgeR then estimated the level of dispersion (using conditional maximum likelihood modeling) for each annotated gene and all annotated genes together. An empirical Bayes' theorem was used to moderate overdispersion via shrinking each gene's level of dispersion towards the consensus dispersion calculated from all annotated genes together.

EdgeR replaced each annotated genes' original paired end read counts across samples with pseudo counts that incorporated all of the above normalizations. An exact test (analogous to a Fisher's Exact test with modifications to suit data modeled using a negative binomial distribution) was used to directly compare each annotated gene's pseudo counts in our sALS and neurologically healthy control sample groups.

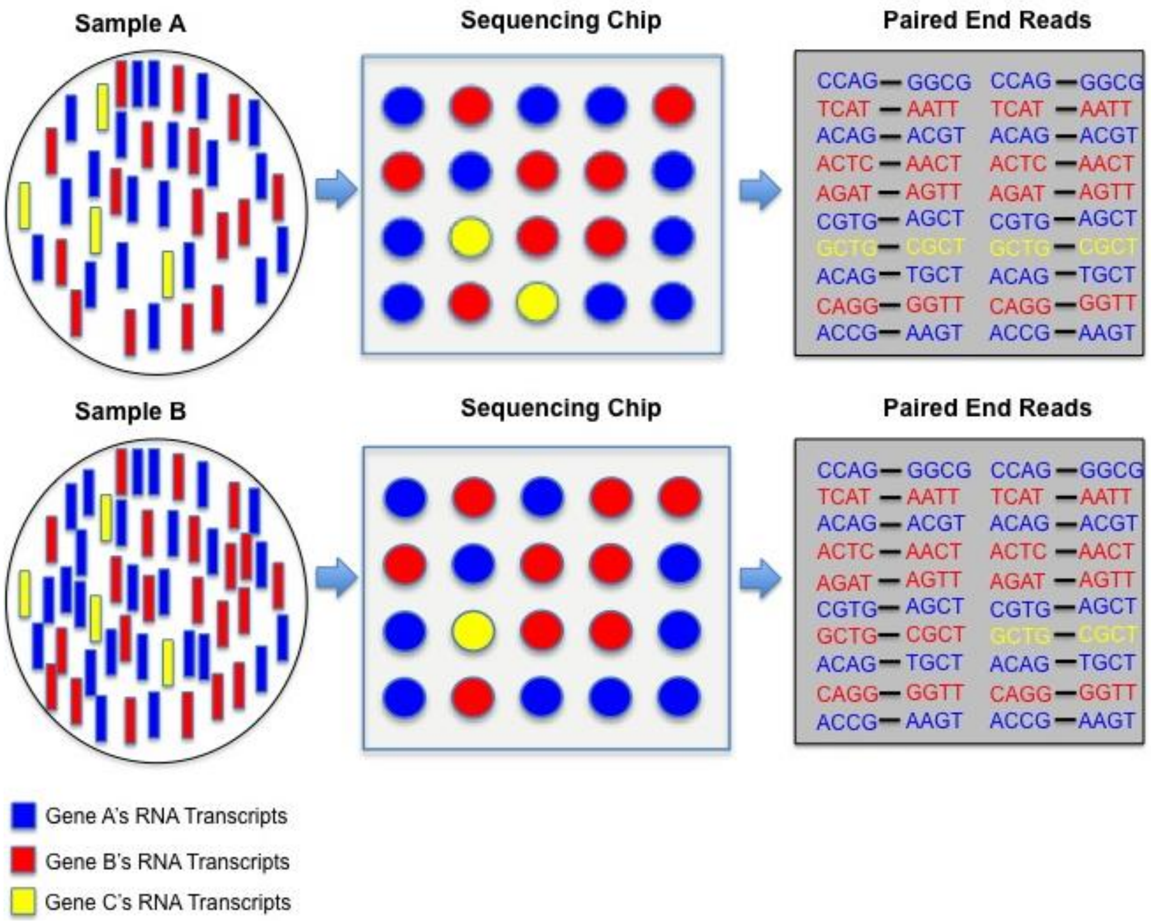


Figure 14. RNA composition differences and gene expression estimates. This figure shows the effects of RNA composition differences between samples on a given gene’s expression estimates. On the left are two biological samples’ RNA transcripts from three different genes (red, blue, and yellow). Each gene’s RNA transcripts were converted into a proportional amount of sequenceable dscDNA molecules. In the middle, these molecules’ denatured strands hybridized to all the available spots (or covalently bound complementary oligonucleotides) on the NextSeq500 flowcell. Those spots are color-coded by what gene’s denatured strands bound to them. On the right are each sample’s paired end reads color-coded by which gene they aligned to. Read 1 and read 2 sequences are on the left and right of the black dash marks, respectively. While both samples had an equal number of RNA transcripts from the yellow gene, sample A would have a higher yellow gene expression estimate as a result of having more paired end reads aligned to it. The blue and red genes were expressed at a higher level in sample B relative to sample A, and the blue gene’s corresponding denatured strands bound more spots on the flowcell. This led to the yellow gene’s denatured strands being under sampled in sample B, and a falsely reported difference in the yellow gene’s expression level estimates between the two samples.

EdgeR reported an associated p-value for each annotated gene tested. We calculated each annotated gene's Benjamini-Hochberg corrected p-value from their EdgeR reported p-value via the R function p.adjust. Each annotated gene with a Benjamini-Hochberg corrected p-value < 0.10 was identified as a sALS group-specific DEG.

We used EdgeR's default workflow for our DEG analysis. We decided to filter out genes with a cpm (counts per million aligned paired end reads) value <1 in 7 samples. We chose a cpm value of 1 as smaller values likely reflected noise. We chose 7 samples as our threshold, as genes that were only expressed in our disease or control group could play an important role in disease pathology.

DESeq2 and DEG identification: We used DESeq2 (Love, Huber, Anders 2014), an open source program specially designed for RNA-Sequencing data, to detect sALS group-specific DEGs. DESeq2 first normalized each sample's paired end read counts (from HTSeq-Count) for each annotated gene. Normalized paired end read counts in our sALS and neurologically healthy control samples were then directly compared to identify sALS group-specific DEGs.

DESeq2 followed a multi-step process to normalize each sample's paired end read counts for each annotated gene. This process involved mathematical normalizations accounting for 1) differences in the total number of paired end read counts between samples, 2) overdispersion in annotated genes' paired end read counts across samples, and 3) high variance in gene expression fold change values for annotated genes with low paired end read counts across samples.

Like EdgeR, DESeq2 used a negative binomial distribution to best account for variance in each annotated genes' paired end read counts across each group's samples. Unlike EdgeR, DESeq2 calculated the level of dispersion for each annotated gene and annotated genes with similar expression levels across samples. DESeq2 applied an empirical Bayes' theorem to moderate overdispersion via shrinking each gene's level of dispersion towards the level of dispersion estimated for genes with a similar expression level.

Each annotated gene's normalized paired end read counts across each group's samples were used to calculate a log transformed fold change value between groups via a maximum likelihood model. Log fold change values for annotated genes with a small number of paired end read counts across both groups' samples are often artificially high, owing to a low signal-to-noise ratio. To mediate this, DESeq2 applied an empirical Bayes procedure to shrink all annotated genes' log fold change estimates towards zero, applying more shrinkage to genes with low paired end read counts across samples. This reduced the chance of false positives in downstream DEG analyses.

To identify sALS group-specific DEGs, DESeq2 applied a Wald test where each annotated gene's shrunken log fold change value was divided by its standard error. Resultant Z-scores were compared to a normal distribution, and a corresponding p-value was generated. We took all annotated genes' corresponding p-values, and calculated corresponding Benjamini-Hochberg corrected p-values using the R function p.adjust. Each annotated gene with a Benjamini-Hochberg corrected p-value < 0.10 was identified as a sALS group-specific DEG.

Cufflinks for gene expression estimates: We separately used Cufflinks (Trapnell et al.

2010), an open source program specially designed for RNA-Sequencing data, to estimate each sample's gene expression levels for all annotated genes in the hg19 human reference genome. Prior to estimating each annotated gene's expression level for each sample, Cufflinks calculated each sample's size distribution of cDNA fragments (in their sequenced strands) using their aligned paired end reads. This data was used in a downstream step. Cufflinks then followed a 5-step process to estimate each annotated gene's expression level for each sample.

For a given annotated gene, Cufflinks first identified a given sample's aligned paired end reads whose corresponding cDNA fragments (in their sequenced strands) were necessarily generated from different RNA transcripts. This step used maximum likelihood statistical modeling. Second, Cufflinks used probabilistic modeling (relying on a proof of Dilworth's theorem) to identify the minimum number of RNA transcripts that accounted for that annotated gene's aligned paired end reads. RNA transcripts identified in this step included annotated transcripts in hg19 human reference transcriptome and/or novel transcripts. Third, Cufflinks used an algorithm to probabilistically assign each aligned paired end read to the RNA transcript its corresponding cDNA fragment was generated from. The algorithm used the sample's calculated distribution of cDNA fragment sizes (in their sequenced strands) as well as the annotated genes' RNA transcript splicing structures to accomplish this. Any aligned paired end read whose corresponding cDNA fragment could have been generated from more than one RNA transcript was assigned to each of those RNA transcripts. This process is illustrated in Figure 15.

Four, Cufflinks estimated that annotated gene’s RNA transcript expression levels. This involved using a statistical model where the probability of observing each transcript’s assigned paired end reads from the previous step was linearly related to each transcript’s expression level. Maximum likelihood equations calculated the most probable quantitative expression levels for that annotated gene’s RNA transcripts. Each transcript’s expression level was reported in normalized gene expression units called FPKMs (Fragments Per Kilobase of exon per Million fragments mapped) whose calculation is explained in a proceeding section. Finally, Cufflinks estimated that annotated gene’s expression level by summing all of its RNA transcripts’ FPKM values together.

We did not include each sample’s paired end reads that aligned to rRNA, tRNA, mtRNA, or unannotated transcripts in our transcript expression estimates. This prevented RNA composition differences between samples leading to false positives or negatives in our downstream Cuffdiff2 DEG analysis. The effects of RNA composition differences between samples on gene expression can be seen in Figure 14.

Cufflinks’ FPKM gene expression units: The Illumina Truseq® Stranded Total RNA HT kit broke each sample’s isolated RNA transcripts into 120-220 nucleotide fragments (with a median size of 155 nucleotides) prior to generating sequenceable dscDNA molecules. For each sample, the length of a given isolated RNA transcript was positively correlated with both 1) the number of fragments generated from it, and 2) the number of corresponding sequenceable dscDNA molecules generated from its fragments. Relative to shorter RNA transcripts, longer RNA transcripts were generally more likely to have a greater number of corresponding paired end reads (and paired end read counts). This was because they had more denatured strands that could hybridize to the limited number

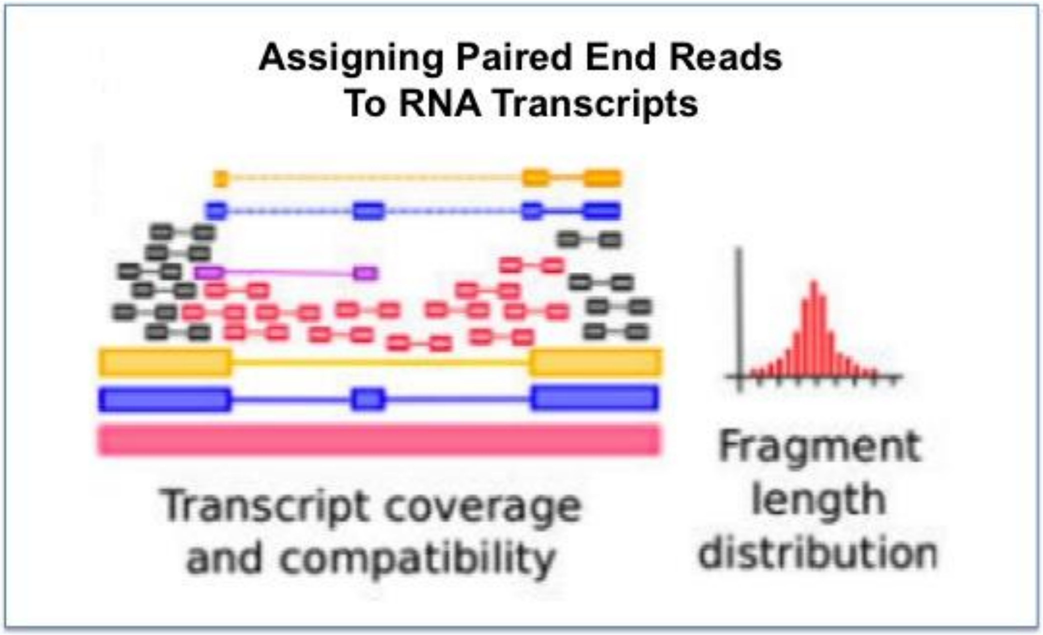


Figure 15. Cufflinks assignment of individual paired end reads to a given annotated gene's RNA transcripts. On the left, this figure shows a visual schematic of Cufflinks assignment of individual paired end reads (dumbbell shaped objects of various sizes) to a given annotated gene's three RNA transcripts (colored yellow, purple, and pink and located just above the words "Transcript coverage and compatibility"). Paired end reads are color-coded by the transcript they aligned to. Black paired end reads could not be exclusively assigned to one RNA transcript. The violet paired end read was originally assigned to either the purple or pink transcript. As that sample did not have many sequenced strands with large cDNA fragments (as indicated by the cDNA fragment length histogram on the right side of this figure), the violet paired end read was ultimately assigned to the purple transcript as opposed to the pink transcript. Adapted from Trapnell et al 2010.

of Illumina NextSeq 500 binding spots (covalently bound oligonucleotides) for sequencing. Estimating each sample's gene expression levels using aligned paired end reads counts alone could thereby lead one to falsely conclude genes with longer RNA transcripts were expressed at a higher level relative to genes with shorter RNA transcripts.

To address this issue, Cufflinks reported all transcript (and gene) expression levels in normalized expression units called FPKM's as opposed to paired end read counts. Any given FPKM value was calculated using the equation FPKM = C/LN. C is the number of paired end reads that aligned to that gene's transcribed regions (or to that transcript). L is the length of that gene's transcribed regions (or the transcript itself) in kilobases. N is the number (in millions) of sequenced reads acquired for that sample. The FPKM measurement accounted for differences in the length of each gene's transcribed regions (or the length of a given transcript), as well as the total number of sequenced reads for each sample. The importance of normalizing for differences in the total number of sequenced paired end reads and its influences on gene expression estimates are described in the EdgeR section. Figure 16 illustrates the advantage of FPKM measurements over paired end read counts for accurate expression estimates.

Cuffdiff2 for DEG identification: We used Cuffdiff2 (Trapnell et al. 2013), an open source program specially designed for RNA-Sequencing data, to detect sALS group-specific DEGs. For each annotated gene, Cuffdiff2 first performed normalizations of its transcripts' FPKM values (from Cufflinks) across samples in our sALS and neurologically healthy control groups. Representative FPKM values were calculated for our sALS and neurologically healthy control sample groups, and then directly compared to identify sALS

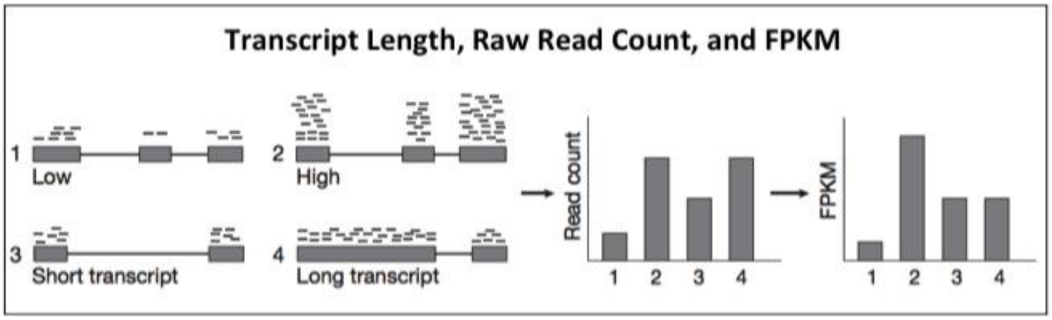


Figure 16. Transcript length, raw read count, and FPKM values. This figure shows paired end read counts vs. FPKM values for four different genes (numbered 1-4) from a hypothetical sample. Above each gene are the sample's aligned sequenced reads. Gene 1 and 2 have equally sized transcribed regions, yet gene 2 was more highly expressed in that sample as evidenced by more aligned sequenced reads. Both paired end read counts and FPKM values accurately reflected the expression difference between genes 1 and 2. Genes 3 and 4 have different sized transcribed regions, but were equally expressed in that sample. However, gene 4 has many more aligned paired end reads as a result of generating many more denatured strands (from sequenceable dscDNA molecules) that bound to the flowcell for sequencing. FPKM values accurately reflected genes 3 and 4 as equally expressed by normalizing for their different sized transcribed regions. If one were to use the raw read count measurements for genes 3 and 4, they would falsely report a difference in their expression levels. Adapted from Garber, Grabherr, Guttman, Trapnell 2011.

group-specific DEGs.

Cuffdiff2 applied mathematical normalizations to RNA transcripts' FPKM values that accounted for 1) differences in the total number of assigned paired end reads to RNA transcripts across samples in each group, 2) uncertainty associated with Cufflinks' assignment of each paired end read to an RNA transcript, and 3) overdispersion in RNA transcripts' FPKM values across each group's samples.

Cuffdiff2 used a mathematical normalization that accounted for differences in the total number of paired end reads assigned to RNA transcripts between samples in each group. This was the same normalization EdgeR used to account for differences in the total number of paired end reads aligned to annotated genes across samples. Rationale for this normalization step and its influences on expression level estimates are the same for transcripts as they are for genes, and can be seen in the EdgeR section.

It has been observed that up to 50% of a given sample's aligned sequenced reads can be assigned to more than one RNA transcript in an annotated gene (Trapnell et al. 2013). This makes sense as human genes' RNA transcripts often share large amounts of sequence, and many genes have paralogs with highly similar sequences. For every sample, Cuffdiff2 used a beta distribution to statistically model the level of uncertainty associated with each paired end read's assignment to an RNA transcript.

Like paired end read counts, FPKM values for each RNA transcript across each group's samples are often overdispersed. This means the variance of each transcript's FPKM values across each group's samples generally exceeds what is expected using a

Poisson distribution. Cuffdiff2 used a negative binomial distribution to model each RNA transcript's FPKM values across each group's samples to better account for this variance.

Cuffdiff2 applied an algorithm mixing results from the beta and negative binomial distributions for each RNA transcript's FPKM values across each group's samples. Resultant beta negative binomial distributions accounted for variability in each RNA transcript's FPKM values owing to 1) uncertainty associated with Cufflinks' assignment of each paired end read to an RNA transcript, and 2) overdispersion in RNA transcripts' FPKM values across each group's samples.

For each annotated gene, Cuffdiff2 calculated representative FPKM values for our sALS and neurologically healthy control sample groups. For each sample group, this involved statistically accounting for the mean, variance, and covariance of all of its transcripts' normalized FPKM values across samples. Cuffdiff2 then calculated an expression ratio (fold change) by dividing the sALS sample group's representative FPKM value by the neurologically healthy control sample group's representative FPKM value. Log-transforming this expression ratio generated a test statistic that followed a standard normal distribution when divided by its variance. Cuffdiff2 then conducted a two-sided t test to assess the significance of the test statistic, and reported an associated p-value for that annotated gene. When an annotated gene had a representative FPKM value of zero in one of our groups, a one-sided t test was conducted. We took all annotated genes' corresponding p-values, and calculated their Benjamini-Hochberg corrected p-values using the R function p.adjust. Each annotated gene with a Benjamini-Hochberg corrected p-value < 0.10 was identified as a sALS group-specific DEG.

Weighted Gene Co-Expression Network Analysis (WGCNA): We used WGCNA (Langfelder, Horvath 2008) to identify gene co-expression networks statistically associated with our sALS sample group. WGCNA followed a 6-step process to predict which genes were co-expressed, cluster them into gene networks, test which gene networks were associated with disease status, and aid our selection of hub genes. The mathematical formulas used in each step are not included in this description, but can be found in an earlier publication (Zhang, Horvath 2005).

First, WGCNA calculated an adjacency matrix (a gene network) that reported a correlation value between every pair of genes' expression values across all 15 samples. This analysis presumes the higher the correlation value between a pair of genes, the more likely they are functionally connected. Once the adjacency matrix was constructed, summation of any individual gene's correlation values to all other genes reflected its level of overall connectedness.

Second, the adjacency matrix was raised to a software-determined exponential power determined by the input dataset. This served to reduce noise by pushing lower pairwise gene correlation values closer to zero relative to higher values. The exponential power used was the lowest value needed to ensure the network approximated scale-free topology. In this context, scale-free topology was satisfied when a small number of genes (hub genes) were highly connected to other genes, whereas the majority of genes were weakly connected to other genes. Many biological (including gene co-expression) networks have demonstrated scale-free topology (Zhang, Horvath 2005), and specifically manipulating these networks' highly connected members modulated cellular processes associated with these networks (Jeong, Mason, Barabasi, Oltvai 2001, Carter, Brechbuhler, Griffin, Bond 2004). This step laid the foundation for identification of hub genes within smaller modules (networks) of interest later in this analysis.

Third, the adjacency matrix was transformed into a topological overlap matrix by calculating topological overlap (TOM) scores for each gene. This score accounted for each pair of genes' connection strength (adjacency value) to each other as well as their connection strengths (adjacency values) to every other gene in the adjacency matrix. Higher TOM scores indicated a pair of genes was more likely connected to each other and a shared set of other genes.

Fourth, WGCNA identified gene co-expression networks (or modules) via average linkage hierarchical clustering using a dissimilarity score (1-TOM score for every gene) as a measure of distance. The resultant dendrogram of clustered genes was segregated into individual modules with at least 30 genes using WGCNA's dynamic tree-cutting algorithm (Langfelder, Horvath 2008).

Fifth, WGCNA calculated each module's eigengene, or first principle component, using all samples' gene expression values for all genes in each module. A module eigengene was considered a summarized expression profile representative of that module. Each module's eigengene was then correlated against every other module's eigengene. If two or more modules' eigengenes had a correlation value >.75, those modules were merged together to form a larger module. Module eigengenes were re-calculated at this stage and the process was repeated until no two modules' eigengenes had a correlation value >.75.

Finally, each module eigengene was tested for statistical association to our provided phenotypic traits. We assessed whether each module's eigenegene was associated with disease status, gender, or age. P-values based on the Student's t-test were reported, and are equivalent to a Wald test in a univariable linear regression model.

For all samples, we input a filtered list of 13,301 genes and their Cufflinks FPKM values. All of these genes had an FPKM value >2 in at least 7 samples. We chose an FPKM of 2 as smaller values more likely reflected noise. We chose 7 for our sample threshold as genes that were only expressed in our sALS or neurologically healthy control sample group could play an important role in disease pathology. We log-transformed these FPKM values using log2 (FPKM value +1) as recommended on the WGCNA FAQ's page (http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/faq.html) prior to analysis.

Our analysis was guided by steps 1, 2b, and 3 in the R Tutorial listed under I. Network analysis of liver expression data from female mice: finding modules related to body weight from the WGCNA website (http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/).

We deviated from the tutorial several times. We generated a signed weighted adjacency matrix as opposed to the default unsigned weighted network. We chose this option to preserve the directions of every pair of genes' correlation, as a positive correlation may indicate biological activation whereas a negative correlation may indicate biological repression. Unsigned networks did not preserve the direction of correlation. We used the bicor (biweight midcorrelation) correlation in place of the Pearson correlation to construct our adjacency matrix and determine the exponential value necessary to approximate scale-free topology. We chose this option as we had a small sample size, and biweight midcorrelations are more robust to outliers compared to Pearson correlations (Langfelder, Horvath 2012). We added the flag corOptions=list(maxPOutliers=0.1)) to further reduce outlier effects.

WGCNA hub gene identification: We identified module hub genes using WGCNA's intramodular connectivity and modular membership scores calculated for every gene in each prioritized module. The intramodular connectivity score reflected the cumulative connection strength a given module gene had with all other module genes. The modular membership score reflected how representative that gene's expression values were of the module as a whole.

Hub genes typically have large values for both of these metrics. We considered each gene's gene significance score to further prioritize one hub gene over another. This score reflected how strongly a given gene's expression values correlated with disease status in our set of samples.

Ingenuity Pathway Analysis: We used QIAGEN's Ingenuity Pathway Analysis to assess what canonical signaling pathways, diseases/disorders, and upstream regulators (molecules known to influence various genes' expression levels) were statistically significantly associated with each of our prioritized gene sets. For each prioritized gene set, IPA used a right-tailed Fisher Exact test to assess the number of genes in that gene set that were separately associated with each canonical signaling pathway, disease/disorder, and upstream regulator in the IPA Knowledge base. Association p-values relating each prioritized gene set to each tested canonical signaling pathway, disease/disorder, and upstream regulator were reported.

***In Vitro* models for hypothesis testing:** For hypothesis testing, we relied on *in vitro* models using neural stem cells and iPSC-derived motor neurons. Neural stem cells are precursors of the cervical spinal cell populations (astrocytes, microglia, and motor neurons) that we isolated RNA from for each sample. We opted for *in vitro* cell models over transgenic animals carrying a fALS causal point mutation. We suspect findings in these cells models would more accurately reflect what occurs in sALS patients' disease-vulnerable cervical spinal cells *in vivo*, as recent studies show the majority of sALS is not caused by a monogenic mutations (Renton, Chio, Traynor 2014).

Neural stem cell derivation and maintenance: Donor blood mononuclear cells were reprogrammed to a pluripotent state and neuralized as previously described (O'Brien, 2015). The resulting neural stem cells (NSC) were maintained in a dividing state on Geltrex (Life Technologies, Grand Island, NY) coated tissue culture ware in NSC Growth Medium [KnockOut DMEM/F-12 containing 2 mM GlutaMAX-I supplement, 20 ng/mL human recombinant basic fibroblast growth factor (bFGF), epidermal growth factor (EGF), 2% StemPro neural supplement, 100 ug/mL pyruvate and 50 ug/ml uridine]. NSC were grown at 37°C in a 5% CO2 incubator with the oxygen concentration held at 5%, and media was changed every 2 to 3 days. Cultures were passaged when 90% confluent.

NSC treatment groups and transfection: 24 hours before the MTT assay, media was exchanged for five groups of NSC cultures with Optimem Media (Thermo-Scientific) containing 1) Fugene HD reagent alone, 2) Fugene HD reagent + GFP plasmid, 3) Fugene HD reagent + 100 ngs TNFAIP2-IRES-GFP plasmid, 4) Fugene HD reagent + 100 ngs TNFAIP2-IRES-GFP plasmid + 75 uM caspase 8 inhibitor, or 5) Fugene HD reagent + 100 ngs TNFAIP2-IRES-GFP plasmid + 75 uM caspase 9 inhibitor. They were then incubated as described in the above section.

The TNFAIP2-IRES-GFP plasmid can be seen in Figure 17. The GFP plasmid only differed in not possessing the TNFAIP2 and internally ribosome entry site (IRES) cassettes. Both plasmids were complexed with FuGENE HD Transfection Reagent (Promega) according to manufacturer instructions prior to transfection. The cell permeable caspase inhibitors used (Sigma Aldrich) were peptides that irreversibly bound to the catalytic sites of their respective activated caspases. Each of the 5 NSC treatment groups consisted of eight independent cell cultures. Experiment done in collaboration with Jim Bennett and Paula Keeney.

Exposing NSCs to TNF-∝, Isolating RNA, and qPCR

48 hours before harvest, medium was exchanged on 80% confluent NSC cultures with 1% DMSO (Sigma, St. Louis) in NSC growth medium and incubated as above. 24 hours before harvest, medium was exchanged with 1% DMSO or 100 ng/mL HumanKine Tumor Necrosis Factor-α (Sigma) in growth medium and incubated. After 24 hours, cells were lifted using TrypLE (Life Technologies), washed, and each pellet sonicated in 350 ul Buffer RLT Plus (Qiagen, Valencia, CA). RNA was isolated using an AllPrep DNA/RNA Mini Kit (Qiagen) with an on-column DNase digestion. RNA quality and quantity were

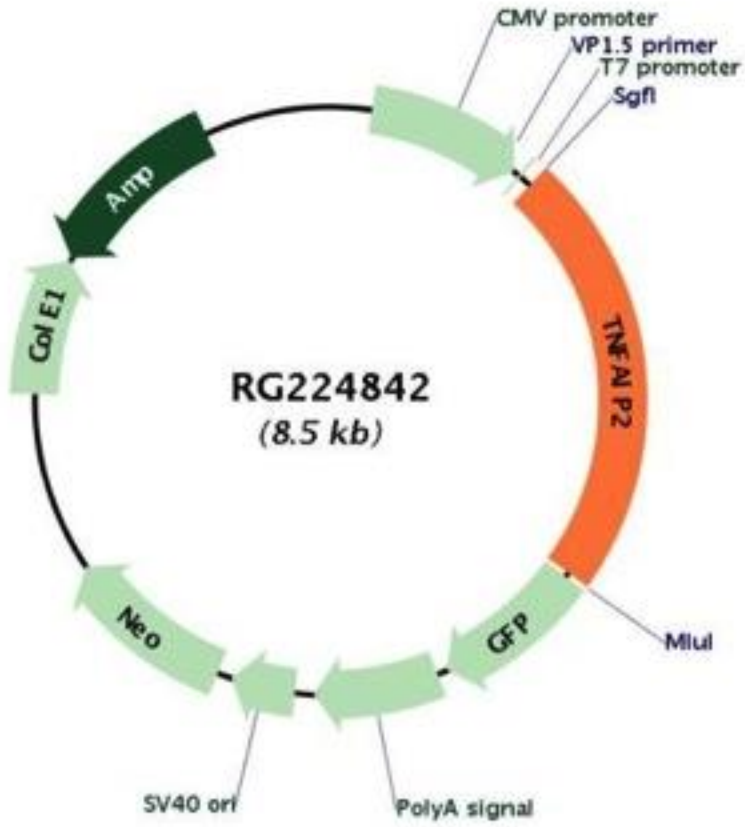


Figure 17. TNFAIP2-IRES-GFP plasmid. This figure shows the TNFAIP2-IRES-GFP plasmid. The internal ribosome entry site (IRES) cannot be seen, but is located between the TNFAIP2 and GFP cassettes. CMV promoter driven transcription encoded the full length TNFAIP2 and GFP proteins.

assessed using a BioRad Experion Electrophoresis Station and a Standard Sensitivity Chip. All samples had a RQI value of 9.5 or higher. 1300 ng of RNA from each sample were reverse transcribed to dscDNA using an iScript kit (Bio-Rad, Hercules, CA) according to directions. Forward (FP) and reverse primers (RP) for SybrGreen detection were designed by Beacon Designer® v 8.12 (Premier Biosoft, www.PremierBiosoft.com) using human gene cDNA sequences downloaded from PubMed, and were synthesized by Operon. qPCR reactions (including forward and reverse primers at 250 nM and standard qPCR reagents) were run on the BioRad CFX96. The BioRad CFX96 automatically detected cycle threshold (Ct) values for each of our 9 DEGs in each tested sample.

Gene normalization was carried out using Qiagen RT2 Profiler Human Housekeeping gene plates, with each cDNA tested in duplicate for expression of the 12 housekeeping genes. qbasePLUS® software (BioGazelle, www.Biogazelle.com) was used to determine the most stable genes (GAPDH, RPL13A, RPL3O) across all experiments. The geometric means of these three genes' expression levels in each cDNA sample were used to normalize expression of the nine DEGs examined by qPCR. 8 independent NSC cultures were run. Experiment done in collaboration with Jim Bennett and Paula Keeney.

iPSC generation and neural Induction: Integration-free iPSCs were generated from peripheral blood mononuclear cells (MNC) from a neurologically healthy, Caucasian male donor (aged 62) using a previously described protocol (Dowey et al. 2012) with modifications (O'Brien, Keeney, Bennett 2015). Briefly, the pEB-C5 and pEB-Tg plasmids (Addgene) were electroporated into cells using an Amaxa Nucleofector 4D system

(Lonza, Allendale, NJ). After three weeks, viable colonies were expanded in mTeSR medium on Geltrex (Life Technologies) coated plates. Neuralization of iPSCs was accomplished using PSC Neural Induction Medium (Life Technologies) according to the protocol with modifications (Amoroso et al. 2013). All cultures were maintained at 37°C in a humidified CO2 incubator with the oxygen level held at 5%.

Motor neuron differentiation and transfection: Neuralized iPSCs were grown in neural induction media containing DMEM/F12 with 0.2 μM LDN-193189 (LDN; Stemgent), 10 μM SB431542 (SB; Stemgent), 10 ng/mL BDNF (R&D systems), 0.4 ug/mL L-ascorbic acid (Sigma), 2 mM GlutaMAX-I supplement, 1% N-2 supplement, and 1% nonessential amino acids (NEAA). Two days later, 1 μM retinoic acid was added. On day four, LDN/SB was stopped and 1 μM smoothened agonist (SAG; Calbiochem) and 0.5 μM PM were added. On day 14, cells were switched to neurobasal media containing 2 mM GlutaMAX-I, 2% B-27, 1% NEAA, 0.4 ug/mL AA, 10 ng/mL GDNF (R&D), 10 ng/mL CNTF (R&D). Media was replaced every 2-3 days. All cell culture materials were purchased from Life Technologies. All cultures were grown at 37°C in 5% oxygen and 5% CO2 conditions. We found motor neurons had a ~15-22 fold increase in expression of motor neuron specific markers HB9 and ISL1 at day 21, suggesting successful differentiation (O'Brien, Keeney, Bennett 2015).

On day 21 of differentiation, iPSC-derived motor neuron cultures were transfected with 100 ngs of TNFAIP2-GFP plasmid or GFP plasmid. The TNFAIP2-GFP plasmid was different from the TNFAIP2-IRES-GFP plasmid mentioned above, in that it did not contain and IRES. It therefore produced a TNFAIP2 protein with a GFP fused to its N terminus.

Both of these plasmids were complexed with FuGENE HD Transfection Reagent (Promega) according to manufacturer instructions prior to transfection.

Real-time quantitative PCR (qPCR): We used qPCR to measure TNFAIP2 expression in our iPSC-derived motor neuron groups transfected with either TNFAIP2-GFP plasmid or the GFP plasmid. RNA was extracted from iPSC-derived motor neuron cultures with the RNeasy Plus Micro Kit (Qiagen) according to manufacturer instructions. Quantification of isolated RNA was performed using a Nanodrop 2000c spectrophotometer (Thermo Scientific). RNA was reverse transcribed into dscDNA using the iScript cDNA synthesis kit (BioRad). For qPCR, 50 ngs of dscDNA per well was loaded into a 96-well plate and analyzed with the CFX96 Real Time PCR Detection System (BioRad). All samples were analyzed in triplicate. Data was normalized to the geometric mean of two reference genes determined to have the greatest stability using the software qbasePLUS-GeNorm (BioGazelle; 14.3.3.Z and CYC1). Statistics were calculated using an unpaired two-sample Welch's t-test in Prism software (GraphPad, Prism).

MTT assay: After 24 hours, cell viability was measured in our transfected NSC and iPSC-derived motor neuron cultures using the In Vitro Toxicology Assay Kit, MTT based (Sigma, TOX1) according to manufacturer instructions. For each tested culture, this colorimetric assay measured the amount of yellow water-soluble substrate 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide (MTT) that was converted to formazen. This served as a proxy for cell viability, as this conversion process is carried out by living cells' mitochondrial dehydrogenases. Experiments done in collaboration with Jim Bennett, Paula Keeney, and Laura O' Brien.

FLICA® 660 Activated Caspase 3/7 Assay: After 24 hours, activated Caspase 3/7 levels, markers of apoptosis, were measured in iPSC-derived motor neuron cultures using the FLICA® 660 Caspase 3/7 Assay Kit (ImmunoChemistry Technologies) according to manufacturer instructions. This involved measuring each culture's amount of fluorescent antibodies specific to activated Caspases 3 and 7 that bound to their respective proteins. For quantification, cells were fixed and 10 representative fields were taken with an Olympus FV1000 confocal microscope. Images were analyzed using MetaMorph image analysis software (Molecular Devices) and pixel intensity was normalized to the number of cells per image. Cells were identified by DAPI nuclear staining.

III. Results

Sequencing metrics: We collected >55 million paired end reads per sample using the Illumina NextSeq500. Picard's CollectRNASeqMetrics (http://picard.sourceforge.net) reported the following averaged metrics across samples: 68,613,940 paired end reads, 65.62% of sequenced nucleotides that passed Tophat2's filters and aligned to the hg19 reference genome, 33.23% of nucleotides that aligned to mRNA species, 29.01% of sequenced nucleotides that aligned to rRNA, tRNA, or mtRNA species, and 37.76% of nucleotides that aligned to intronic/intergenic regions. Individual samples metrics can be found in Table 3.

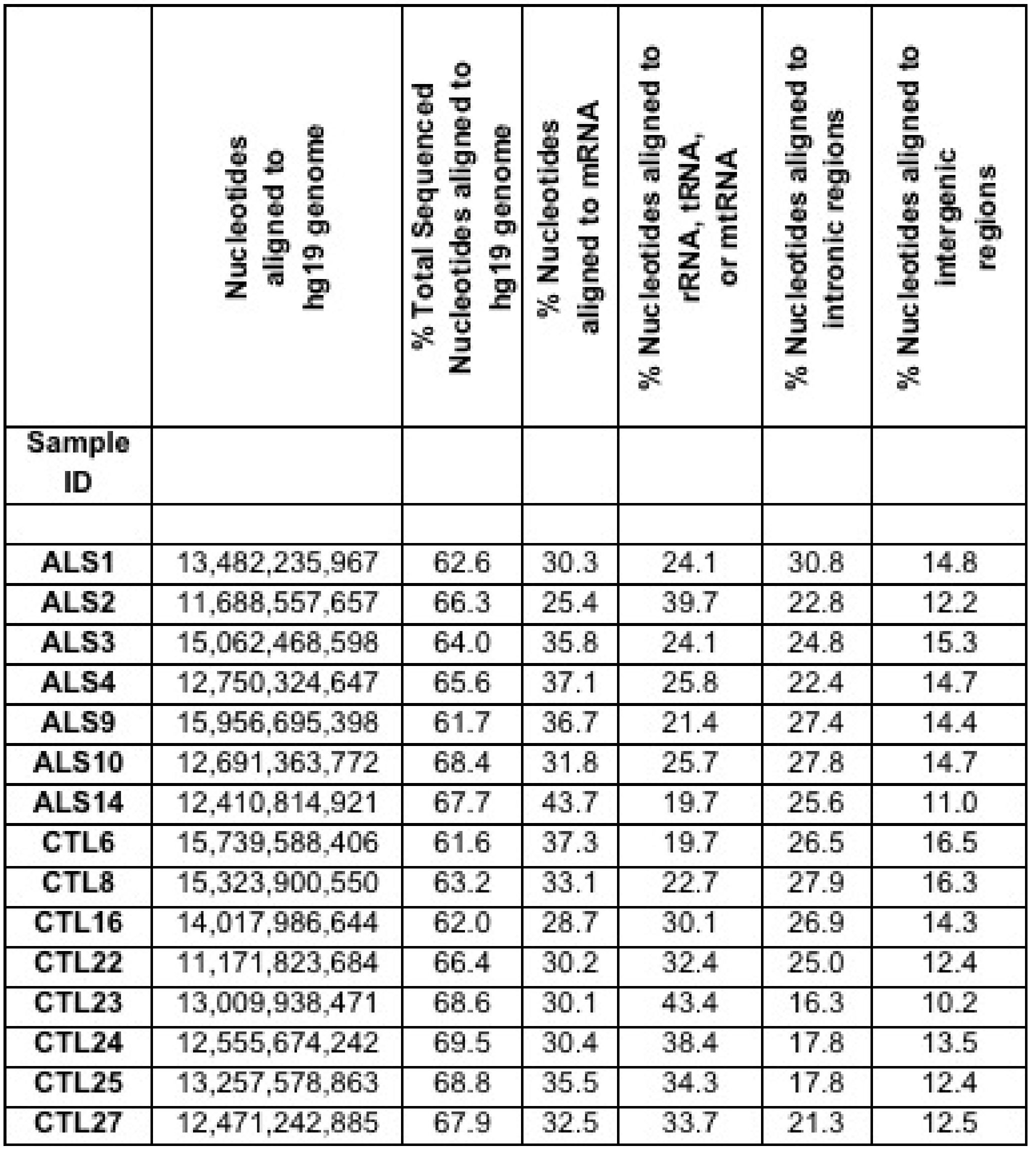
Our samples' averaged % of sequenced nucleotides aligning to mRNA and intronic/intergenic regions was highly similar to what has been observed in other RNA-Sequencing studies using total RNA to construct RNA-Sequencing libraries (Ameur et al. 2011, Shanker et al. 2015).

Sequencing data quality: For each sample, we used FastQC to assess the quality of all paired end reads' Read 1 and Read 2 sequences prior to and after Trimmomatic processing. All samples' paired end reads passed FastQC's quality assessments before and after Trimmomatic processing. Trimmomatic processing successfully increased each sample's average PHRED quality scores for nucleotides towards the 3' end of their Read 2 sequences, and removed all Illumina sequencing adaptor sequences from Read 1 and Read 2 sequences. Figure 18 shows ALS1's analyzed Read 2 sequences before and after Trimmomatic processing. These results are representative of what we observed for all samples.

Presence of known fALS mutations in our sALS samples' sequenced reads: As of 2014, 11% of sALS in Caucasian patients was accounted for by causal mutations in 9 different loci (Renton, Chio, Traynor 2014). Causal fALS mutations have been identified in at least 13 other genes, and may account for a larger proportion of sALS in Caucasian patients (Renton, Chio, Traynor 2014, Abel, Powell, Andersen, Al-Chalabi 2013). We assessed whether any of our sALS samples carried any of 471 known pathogenic coding variants contained in 21 different genes shown in Table 4. ELP3 was not included, as it did not have any qualifying pathogenic variants in any of the three databanks surveyed.

We discovered an sALS sample (ALS4) carried a pathogenic variant from this list. This variant is a missense mutation (A4V) in the Superoxide Dismutase 1 (SOD1) gene, and was found in nearly half of ALS4's paired end reads (754/1576) that aligned to that portion of SOD1. This suggests there was no transcriptional preference for the wildtype or mutant DNA sequence. No other fALS pathogenic variants were found in ALS4 or the other ALS samples. None of these 21 genes were differentially expressed between our

Table 3: Sequencing metrics



	Nucleotides aligned to hg19 genome	% Total Sequenced Nucleotides aligned to hg19 genome	% Nucleotides aligned to mRNA	% Nucleotides aligned to rRNA, tRNA, or mtRNA	% Nucleotides aligned to intronic regions	% Nucleotides aligned to intergenic regions
Sample ID						
ALS1	13,482,235,967	62.6	30.3	24.1	30.8	14.8
ALS2	11,688,557,657	66.3	25.4	39.7	22.8	12.2
ALS3	15,062,468,598	64.0	35.8	24.1	24.8	15.3
ALS4	12,750,324,647	65.6	37.1	25.8	22.4	14.7
ALS9	15,956,695,398	61.7	36.7	21.4	27.4	14.4
ALS10	12,691,363,772	68.4	31.8	25.7	27.8	14.7
ALS14	12,410,814,921	67.7	43.7	19.7	25.6	11.0
CTL6	15,739,588,406	61.6	37.3	19.7	26.5	16.5
CTL8	15,323,900,550	63.2	33.1	22.7	27.9	16.3
CTL16	14,017,986,644	62.0	28.7	30.1	26.9	14.3
CTL22	11,171,823,684	66.4	30.2	32.4	25.0	12.4
CTL23	13,009,938,471	68.6	30.1	43.4	16.3	10.2
CTL24	12,555,674,242	69.5	30.4	38.4	17.8	13.5
CTL25	13,257,578,863	68.8	35.5	34.3	17.8	12.4
CTL27	12,471,242,885	67.9	32.5	33.7	21.3	12.5

patient and control samples. We did not remove ALS4 from our sALS sample group in downstream analyses despite this finding. A recent review (Heath, Kirby, Shaw 2013) of ALS gene expression studies dating back to 2001 revealed various sALS and fALS tissue-specific DEGs were associated with a recurrent set of cellular processes including oxidative stress, mitochondrial dysfunction, apoptosis, cytoskeletal architecture, inflammation, RNA processing, and protein aggregation. Despite varying genetic etiologies, it appears fALS and sALS share a convergent set of perturbed cellular processes, which could explain why distinguishing fALS from sALS using traditional clinical guidelines is extremely challenging (Al-Chalabi, Hardiman 2013). Inclusion of ALS4 is thereby unlikely to hinder our ability to identify cellular processes that may be perturbed in sALS and relevant to disease pathology.

DEG Testing and associated cellular processes: We elected to identify sALS group-specific DEGs using Cufflinks/Cuffdiff2, DESeq2, and EdgeR. In a direct comparison, all three analyses showed different limitations in DEG identification after analyzing the same benchmark datasets (Zhang et al. 2014). EdgeR identified the most DEGs, but also reported the most false positive DEGs. DESeq2 identified the fewest DEGs when input samples had different numbers of total sequencing reads. Cuffdiff2 identified the fewest DEGs when each sample had less than 20 million total sequencing reads. The authors recommended a conservative approach of using at least two (if not all three) analyses to identify DEGs, and proceeding with DEGs mutually reported by multiple analyses to avoid pursuing false positives DEGs. We will likely learn more about which of these DEG tests is most accurate via future comparative analyses using larger benchmark datasets with varying properties. Further, DEG tests that are superior to these three will likely emerge

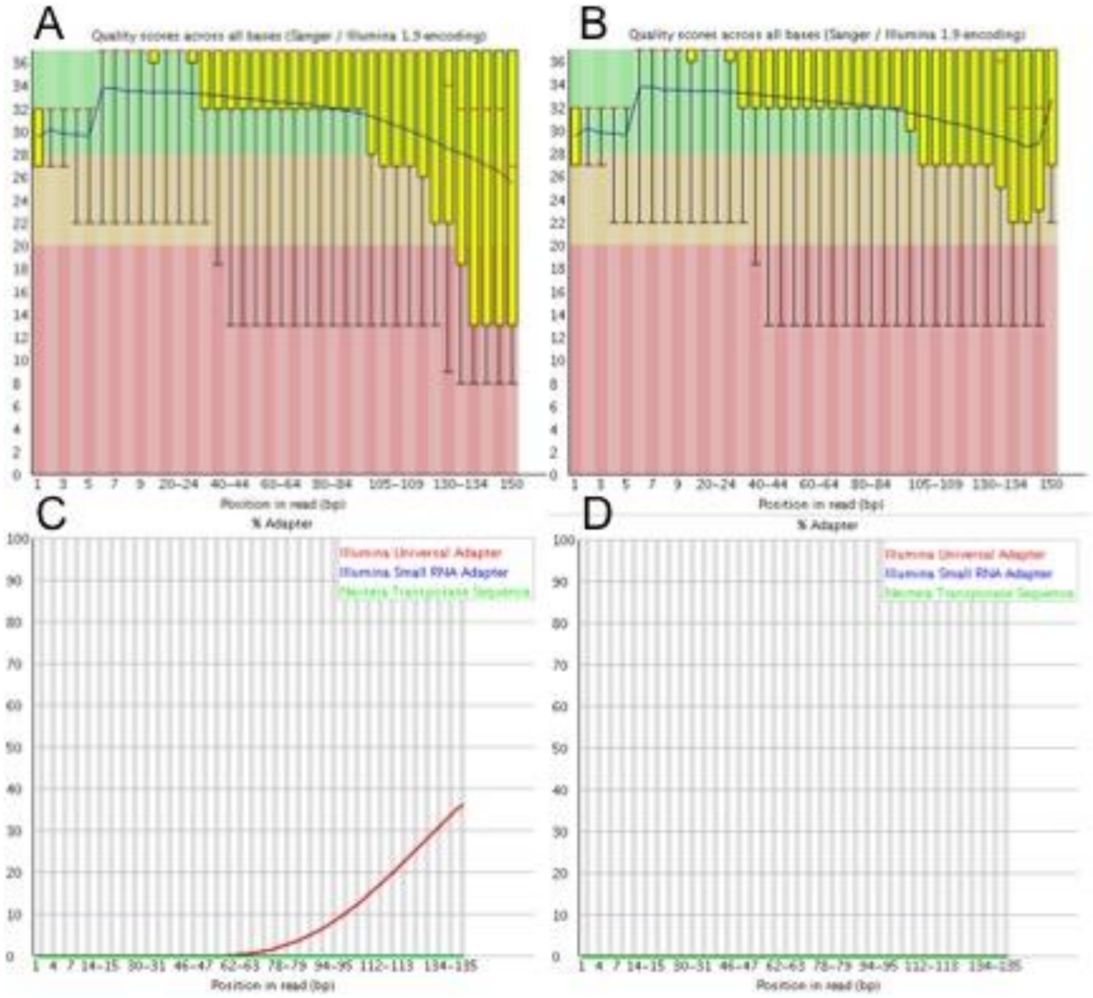


Figure 18. FastQC analyses of ALS1's Read 2 sequences before and after Trimmomatic processing. This figure shows FastQC plots of the average PHRED quality score per position across all of ALS1's Read 2s before (A) and after (B) Trimmomatic processing, and the presence of Illumina sequencing adaptor nucleotides per position across all of ALS1's Read 2s before (C) and after (D) Trimmomatic processing. The x axes in all plots report the nucleotide position in analyzed Read 2s. The Y axes in A and B are PHRED quality scores. The blue lines that extend from left to right in A and B represent the average PHRED quality score across nucleotide positions in analyzed Read 2s. The Y axes in C and D are percentages of nucleotides that correspond to Illumina sequencing adaptors per position in across all analyzed Read 2s.

Table 4: Genes with known fALS mutations

Gene	Chromosomal Position	# of Known Pathogenic Coding Mutations
ALS2	2q33	15
ANG	14q11	36
ATXN2	12q24	3
C9orf72	9p21	1
CHMP2B	3p11	5
DCTN1	2p13	6
FIG4	6q21	5
FUS	16p11	66
HNRNPA1	12q13	2
HNRNPA2B1	7p15	1
NEFH	22q12	9
OPTN	10p13	19
PFN1	17p13	8
SETX	9q34	10
SOD1	21q22	199
SPG11	15q14	1
SQSTM1	5q35	16
TARDBP	1p36	44
UBQLN2	Xp11	12
VAPB	20q13	3
VCP	9p13	10

as RNA-Sequencing technologies and analytics mature.

At an FDR of .10, 74 sALS group-specific DEGs (56 upregulated and 18 downregulated) were mutually identified using Cuffdiff2, DESeq2, and EdgeR. Figure 19 shows a Venn diagram comparing the numbers of DEGs identified across and between analyses at an FDR of .10. Cuffdiff2 identified significantly more DEGs at an FDR of <.10 compared to the other two analyses. We suspect this is largely a result of Cufflinks/Cuffdiff2 employing a considerably different approach to estimating and comparing gene expression levels compared to the other two analyses.

These 74 DEGs, their Cuffdiff2 fold change values (sALS group relative to our neurologically healthy control group), representative FPKM values for the ALS and neurologically healthy control sample groups, and FDR corrected p-values are listed in Table 5. QIAGEN’s Ingenuity Pathway Analysis revealed our 74 DEGs were associated with multiple canonical signaling pathways, disease/disorders, and upstream regulators related to inflammatory cellular processes. TNF-α was identified as an upstream regulator. These IPA results are shown in Table 7.

To avoid false negatives, we identified all DEGs reported at an FDR of <.01 by any of the three analyses. QIAGEN’s Ingenuity Pathway Analysis revealed those 200 DEGs were also associated with multiple canonical signaling pathways and upstream regulators related to inflammatory cellular processes. TNF-α was identified as an upstream regulator. Those IPA results are shown in Table 8.

WGCNA and the black module: We used WGCNA to identify gene modules, or networks, from our dataset. This unsupervised technique identified 37 interconnected gene modules (arbitrarily assigned to different colors) from a filtered list of 13,301 genes without using 1) information about what genes have been shown to interact in previous literature, or 2) information about which samples were from our sALS or neurologically healthy control groups. These modules can be seen in Figure 20. Two of these modules (MEblack and MEsienna4) were associated with sALS disease status at an uncorrected p-value <.01. These modules were not significantly associated with age or gender. They can be seen in Table 6.

Interestingly, QIAGEN's Ingenuity Pathway Analysis revealed the 495 genes comprising the module most strongly correlated to sALS disease status (MEblack, R=0.68, p=0.006) were associated with multiple canonical signaling pathways, disease/disorders, and upstream regulators related to inflammatory cellular processes. TNF-α was identified as an upstream regulator. These results can be seen in Table 7.

As these IPA results were highly similar to those for our sALS group-specific DEGs, we next assessed whether any of those 74 DEGs were found in this module. Intriguingly, we found approximately 57% (42/74) of our DEGs were contained in this module. We decided to prioritize hub genes in this module for candidate gene selection. We found it compelling our sALS group-specific DEGs and a gene co-expression network associated with sALS disease status were both associated with various inflammatory cellular processes and TNF-α signaling. We obtained these results despite discovering both gene sets using independent exploratory approaches. Further, these findings are consistent with previous studies implicating inflammatory processes and TNF-α signaling in ALS pathology (as referenced in the introduction of this chapter).

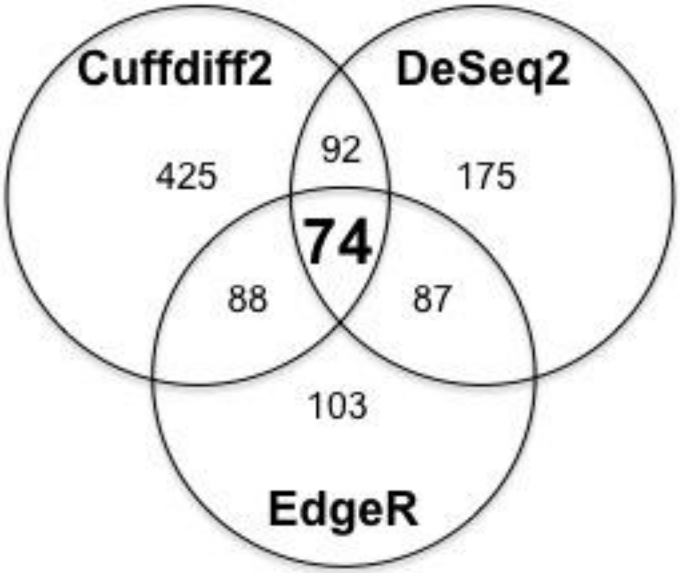


Figure 19. DEG Identification using different algorithms. This figure shows a Venn diagram comparing the number of DEGs identified at an FDR of .10 across and between the DEG analyses.

Table 5: sALS group-specific DEGs identified at an FDR <.10

DEG	Fold Change	ALS FPKM	CTL FPKM	FDR p-value
HTRA4	5.63	1.66	0.30	0.007
PLA2G7	5.59	22.32	3.99	0.007
GPNMB	5.39	237.50	44.03	0.007
OTOA	4.78	0.79	0.16	0.074
APOC1	4.61	348.98	75.64	0.007
LILRA4	4.61	6.40	1.39	0.007
SIGLEC8	4.40	17.79	4.04	0.007
CHAC1	4.30	4.05	0.94	0.007
HLA-DRB1	4.10	80.55	19.64	0.007
KLHL6	3.99	9.14	2.29	0.007
DPEP2	3.96	3.39	0.86	0.007
LILRA2	3.59	6.87	1.91	0.007
CPVL	3.40	44.56	13.10	0.007
CEBPA	3.29	26.50	8.06	0.007
SLC37A2	3.28	7.81	2.38	0.007
APOE	3.25	1785.64	549.01	0.007
SLC7A7	3.01	14.67	4.88	0.007
CAPG	2.99	69.99	23.35	0.007
LILRB4	2.97	9.59	3.22	0.007
FCGR2B	2.97	13.68	4.60	0.007
HPSE	2.96	6.28	2.12	0.007
SELPLG	2.81	15.91	5.64	0.007
HLA-DMB	2.76	49.50	17.91	0.007
BMF	2.73	3.64	1.33	0.007
KCNA5	2.71	8.08	2.97	0.007
THEMIS2	2.71	29.82	10.99	0.007
ITGAX	2.68	13.65	5.09	0.007
CD37	2.66	30.05	11.28	0.007
GK	2.64	15.01	5.68	0.007
FPR3	2.59	10.56	4.07	0.007
ZMYND15	2.56	1.79	0.70	0.083
CD226	2.51	4.16	1.66	0.007
ADAMTS1	2.49	0.71	0.28	0.090
KCNJ5	2.42	4.68	1.93	0.030
CXCL16	2.39	35.65	14.87	0.007
CTSS	2.36	38.81	16.42	0.007
CTSD	2.35	333.56	141.72	0.007
CPM	2.32	15.14	6.52	0.007
HLA-DOA	2.32	15.64	6.74	0.007
TNFAIP2	2.29	12.54	5.47	0.007
ABCG1	2.18	12.55	5.75	0.007
TBXAS1	2.12	17.90	8.43	0.007
HAVCR2	2.07	30.06	14.47	0.007
CD86	2.01	12.74	6.33	0.007
OSBPL11	1.98	30.02	15.10	0.007
CD84	1.98	9.86	4.96	0.007
IL18	1.97	21.02	10.63	0.041
PIK3IP1	1.94	38.02	19.57	0.007
DNASE2	1.89	15.81	8.33	0.007
ASAH1	1.85	305.78	164.90	0.052
GPRIN3	1.84	10.27	5.57	0.007
OTUD1	1.79	9.10	5.08	0.070
CECR1	1.72	15.26	8.83	0.038
WDR91	1.72	7.03	4.08	0.070
LTA4H	1.65	42.38	25.67	0.062
GNB4	1.61	31.09	19.26	0.068
CXCL8	-5.90	4.78	28.23	0.007
WNT16	-4.83	0.44	2.15	0.007
FGF10	-3.40	1.84	6.26	0.023
DCN	-3.17	144.38	458.54	0.007
PTGS2	-3.05	3.08	9.42	0.007
LIPG	-2.72	1.02	2.78	0.018
CFH	-2.58	25.78	66.62	0.007
FHL2	-2.44	3.93	9.61	0.007
COL12A1	-2.42	4.01	9.73	0.007
KDR	-2.26	5.16	11.68	0.007
MSMO1	-2.23	42.05	93.95	0.007
EPHA3	-2.05	2.93	6.02	0.007
NRP1	-2.04	6.36	12.99	0.007
HMGCS1	-2.02	17.63	35.74	0.007
PPP1R3C	-1.99	20.54	40.97	0.007
SQLE	-1.92	22.61	43.53	0.007
ITGA8	-1.90	2.14	4.08	0.007
ABCA8	-1.74	35.29	61.64	0.018

WGCNA hub gene identification: 12 genes in the black module had scores in the top quartile for intramodular connectivity, modular membership, and gene significance metrics. 9 of these genes were separately identified as upregulated sALS group-specific DEGs. TNFAIP2, a gene encoding a TNF-α superfamily protein, was one of these nine. Figure 21 lists all 12 black module hub genes, and contains a graph plotting each black module gene's module membership vs. gene significance score.

Selection of TNFAIP2 as our candidate gene for hypothesis testing: We selected TNFAIP2 as our candidate gene for hypothesis testing for many data-driven reasons. First, TNFAIP2 belonged to the black module associated with sALS disease status, inflammatory cellular processes, and TNF signaling. Second, TNFAIP2 was identified as one of twelve black module hub genes with a score in the top quartile for intramodular connectivity, modular membership, and gene significance metrics. Third, TNFAIP2 was mutually identified as an upregulated sALS group-specific DEG using all three DEG analyses.

Elevated TNF-α signaling plays a known role in cell fate decisions, and induces apoptosis under certain biological circumstances (Probert 2015). Elevated TNF-α signaling has been shown to kill motor neurons in previous literature (He, Wen, Strong 2002, Robertson et al. 2001, Terrado et al. 2000). Elevated TNF-α signaling is also known to increase TNFAIP2 expression in a variety of cell types (Saito et al. 2013, Zhou, Scoggin, Gaynor, Williams 2003, Tian et al 2005), and elevated TNFAIP2 expression has previously been associated with increased apoptosis (Park et al. 2003, Rusiniak et al. 2000, Ma et al. 2003). Studies linking elevated TNFAIP2 expression to increased apoptosis did not assess TNFAIP2's cellular function or whether TNFAIP2

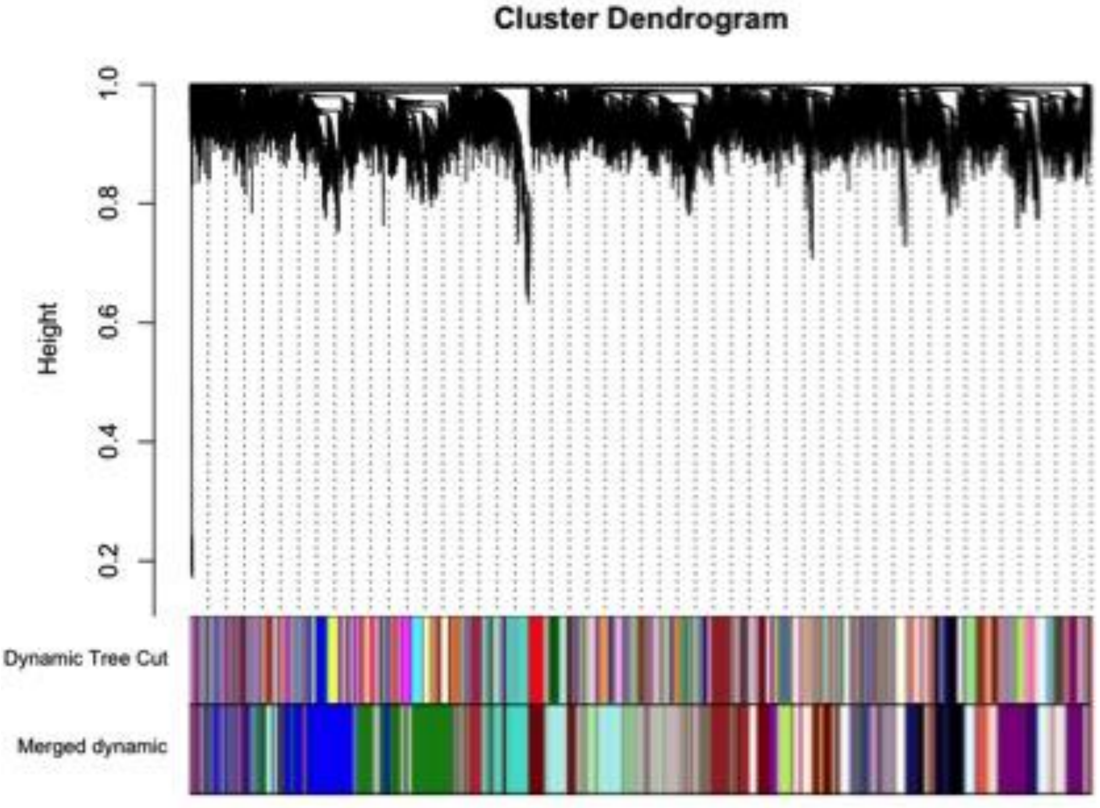


Figure 20. WGCNA module identification. This figure shows all 13,301 genes (individual black lines at top) clustered into different modules based on their topological overlap dissimilarity scores. The multi-colored panel next to "Dynamic Tree Cut" shows 122 identified modules using the Dynamic Tree Cut algorithm. The second multi-colored panel shows 37 larger modules identified after merging smaller modules with highly correlated eigengenes together.

Table 6: Module to phenotype correlation values

Module	Disease Status Correlation	Disease Status Pvalue	Gender Correlation	Gender Pvalue	Age Correlation	Age Pvalue
MEblack	0.676	0.006	-0.370	0.175	-0.014	0.960
MEsienna4	0.662	0.007	-0.023	0.934	0.260	0.349
MEfirebrick3	0.609	0.016	-0.177	0.527	0.027	0.923
MEhoneydew1	0.561	0.030	0.236	0.398	-0.028	0.920
MEhoneydew	0.544	0.036	0.096	0.732	-0.045	0.872
MElightcoral	0.478	0.072	-0.638	0.010	-0.237	0.396
MEmidnightblue	0.402	0.137	-0.305	0.269	-0.135	0.631
MEdarkviolet	0.391	0.149	0.039	0.891	-0.043	0.879
MEdarkred	0.330	0.230	0.115	0.684	0.184	0.513
MEfirebrick4	0.325	0.238	0.307	0.265	0.147	0.602
MEpaleturquoise	0.305	0.269	-0.001	0.997	0.145	0.606
MElavenderblush1	0.269	0.333	-0.003	0.993	0.308	0.264
MEbrown	0.223	0.425	-0.152	0.588	0.064	0.821
MEantiquewhite2	0.214	0.444	-0.061	0.828	-0.158	0.574
MEantiquewhite1	0.110	0.696	0.009	0.973	0.394	0.146
MEturquoise	0.026	0.927	-0.113	0.688	0.175	0.534
MEbisque4	0.023	0.935	0.201	0.472	0.048	0.866
MEdarkolivegreen2	0.009	0.975	0.073	0.797	0.425	0.115
MEcoral2	0.007	0.982	-0.208	0.458	-0.281	0.311
MEindianred3	0.001	0.999	-0.209	0.454	-0.076	0.788
MEdarkmagenta	-0.011	0.970	0.042	0.883	-0.156	0.579
MEnavajowhite1	-0.033	0.907	-0.217	0.437	0.108	0.701
MEdarkseagreen3	-0.033	0.906	0.031	0.914	0.102	0.718
MElightcyan	-0.046	0.872	0.154	0.584	0.108	0.702
MElightyellow	-0.046	0.869	-0.394	0.146	0.114	0.685
MEmagenta3	-0.110	0.697	0.392	0.149	0.462	0.083
MEtan4	-0.169	0.547	-0.128	0.649	0.410	0.129
MElavenderblush3	-0.220	0.432	-0.062	0.825	0.077	0.786
MElightpink3	-0.273	0.325	-0.304	0.270	-0.043	0.878
MEblue	-0.348	0.204	-0.211	0.451	-0.118	0.677
MEdarkseagreen2	-0.436	0.104	-0.127	0.651	0.163	0.562
MEgreen4	-0.510	0.052	-0.138	0.625	0.013	0.963
MElightslateblue	-0.510	0.052	-0.017	0.953	-0.128	0.650
MEantiquewhite4	-0.513	0.050	0.044	0.877	0.354	0.196
MEpink4	-0.569	0.027	0.388	0.153	0.293	0.289
MEbrown4	-0.625	0.013	0.100	0.724	0.040	0.888
MEcoral4	-0.796	0.000	0.090	0.749	0.169	0.548

functionally promoted apoptotic processes directly. We hypothesize elevated TNF-α signaling 1) increased TNFAIP2 expression in our sALS patients' cervical spinal cells, and 2) TNFAIP2 functionally contributed to spinal motor neuron death in our sALS patients via the TNF non-mitochondrial or mitochondrial apoptotic pathway.

TNF-α signaling, TNFAIP2, and other black module hub genes' expression levels:

IPA identified TNF-α as an upstream regulator of genes comprising the black module (Table 7). Elevated TNF-α signaling may have induced changes in black module genes' expression levels, plausibly promoting the black module's associated inflammatory processes (and potentially motor neuron death) in our sALS patients' cervical spinal cells. If this occurred, we suspect TNF-α signaling accomplished this via altering black module hub genes' expression levels. That would be consistent with the theory hub genes functionally regulate their gene co-expression network's activities and associated cellular processes.

Previous literature has already shown elevated TNF-α signaling increases the expression level of TNFAIP2 (a black module hub gene) in a variety of cell types (Saito et al. 2013, Zhou, Scoggin, Gaynor, Williams 2003, Tian et al 2005). For these reasons, we tested whether exposing neural stem cells to TNF-α increased the expression levels of TNFAIP2 and/or the other 8 black module hub genes identified as sALS group-specific upregulated DEGs. qPCR data revealed exposing neural stem cells to TNF-α increased the expression levels of three black module hub genes. These included TNFAIP2, Apoliprotein E (APOE), and Chemokine Ligand 16 (CXCL16). These results can be seen in Figure 22.

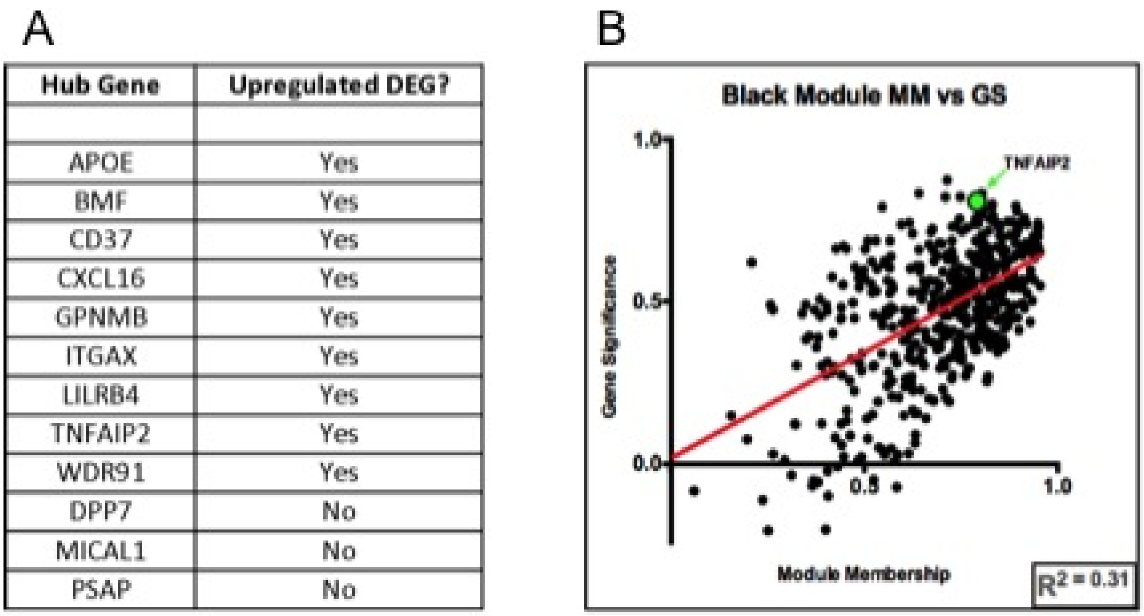
Table 7: Comparing IPA results

74 sALS group-specific DEGs				**495 Black Module Genes**		
Top Canonical Pathways	**Overlapping Genes**	**p-Value**		**Top Canonical Pathways**	**Overlapping Genes**	**p-Value**
Graft-versus-Host Disease Signaling	5/48	8.10E-07		Antigen Presentation Pathway	11/37	3.31E-10
Eicosanoid Signaling	5/64	3.45E-06		TREM1 Signaling	14/75	1.11E-09
Atherosclerosis Signaling	6/124	5.73E-06		Altered T and B Cell Signaling in Rheumatoid Arthritis	14/88	9.72E-09
T Helper Cell Differentiation	5/71	5.77E-06		Role of NFAT in Regulation of the Immune Response	18/171	6.30E-08
B Cell Development	4/33	5.91E-06		CD28 Signaling in T Helper Cells	15/118	6.54E-08
Top Diseases and Disorders		**p-Value**		**Top Diseases and Disorders**		**p-Value**
Endocrine System Disorders		1.54E-03-1.12E-13		Inflammatory Response		1.13E-04-9.32E-25
Gastrointestinal Disease		1.49E-03-1.12E-13		Immunological Disease		1.35E-04-2.33E-19
Immunological Disease		1.53E-03-1.12E-13		Connective Tissue Disorders		7.33E-05-2.79E-19
Metabolic Disease		3.57E-04-1.12E-13		Inflammatory Disease		7.80E-05-2.79E-19
Inflammatory Response		1.68E-03-3.16E-09		Skeletal and Muscular Disorders		7.33E-05-2.79E-19
Upstream Regulators	**Overlapping Genes**	**p-Value**		**Upstream Regulators**	**Overlapping Genes**	**p-Value**
IFNG	24/610	4.78E-13		lipopolysaccharide	95/763	1.98E-20
IL13	14/192	1.43E-11		IFNG	79/610	2.98E-19
cholesterol	10/109	5.41E-10		genistein	42/212	2.82E-18
lipopolysaccharide	23/763	7.68E-10		TNF-α	75/773	4.85E-12
CAMP	7/43	3.59E-09		fluticasone	22/133	1.41E-11
TNF-α	19/773	3.16E-07		IL13	31/192	6.27E-11

Table 8: DEGs identified at an FDR <.01 across analyses

200 sALS group-specific DEGs

Top Canonical Pathways	Overlapping Genes	p-Value
Dendritic Cell Maturation	17/177	3.31E-10
Atherosclerosis Signaling	14/124	1.11E-09
T Helper Cell Differentiation	11/71	9.72E-09
Complement System	8/37	6.30E-08
Graft-versus-Host Disease Signaling	8/48	6.54E-08
Top Diseases and Disorders		**p-Value**
Metabolic Disease		1.13E-04-9.32E-25
Endocrine System Disorders		1.35E-04-2.33E-19
Gastrointestinal Disease		7.33E-05-2.79E-19
Cardiovascular Disease		7.80E-05-2.79E-19
Connective Tissue Disorders		7.33E-05-2.79E-19
Upstream Regulators		**p-Value**
IFNG	61/610	1.98E-20
TGFB1	63/813	2.98E-19
lipopolysaccharide	61/763	2.82E-18
beta-estradiol	61/844	4.85E-12
IL13	30/192	1.41E-11
TNF-α	57/773	6.27E-11



Hub Gene	Upregulated DEG?
APOE	Yes
BMF	Yes
CD37	Yes
CXCL16	Yes
GPNMB	Yes
ITGAX	Yes
LILRB4	Yes
TNFAIP2	Yes
WDR91	Yes
DPP7	No
MICAL1	No
PSAP	No

Figure 21. Black module hub genes and TNFAIP2. This figure lists the 12 black module hub genes and reports whether each was separately identified as an sALS group-specific DEG (A), and shows a graph plotting each black module gene's module membership vs. gene significance score (B). DPP7, MICAL1, and PSAP were not identified as sALS group-specific DEGs. TNFAIP2 is highlighted in green (MM=0.79, GS=0.81) in the plot on the right.

These findings support IPA's prediction of TNF-α as an upstream regulator of genes comprising the black module, and may account for the inflammatory processes associated with those genes. Elevated TNF-α signaling may account for these hub genes' increased expression levels as observed in our sALS patients' RNA-Sequencing data. These findings are also consistent with our hypothesis that elevated TNF-α signaling increased TNFAIP2 expression in our sALS patients' cervical spinal cells.

TNFAIP2 overexpression and apoptosis: If TNFAIP2 functionally contributed to spinal motor neuron death in our sALS patients, we suspect this occurred via a TNF superfamily apoptotic pathway. TNF-α signaling can promote cell survival or cell death depending on cellular and microenvironmental conditions that remain poorly understood (Probert 2015), and elevated TNF-α signaling has been shown to kill motor neurons in previous literature (He, Wen, Strong 2002, Robertson et al. 2001, Terrado et al. 2000). Figure 23 shows a visual of these protein-signaling cascades.

To provide context for the proceeding experiments, I will only describe the TNF superfamily non-mitochondrial and mitochondrial apoptotic pathways. After TNF-α binds Tumor Necrosis Factor Receptor 1 (TNFR1), TNFRSF1A-Associated via Death Domain (TRADD) is recruited. TRADD recruits Fas-Associated Protein With Death Domain (FADD) and pro-forms of Caspases 8 and 10 (Al-Lamki, Mayadas 2015). The pro-form of Caspase 8 undergoes auto-proteolytic activation, and activated Caspase 8 is released into the cytoplasm. The TNF superfamily non-mitochondrial or mitochondrial pathway is then used to induce apoptosis depending on the cell type (Al-Lamki, Mayadas 2015). The non-mitochondrial route involves activated Caspase 8 proteolytically activating Caspases

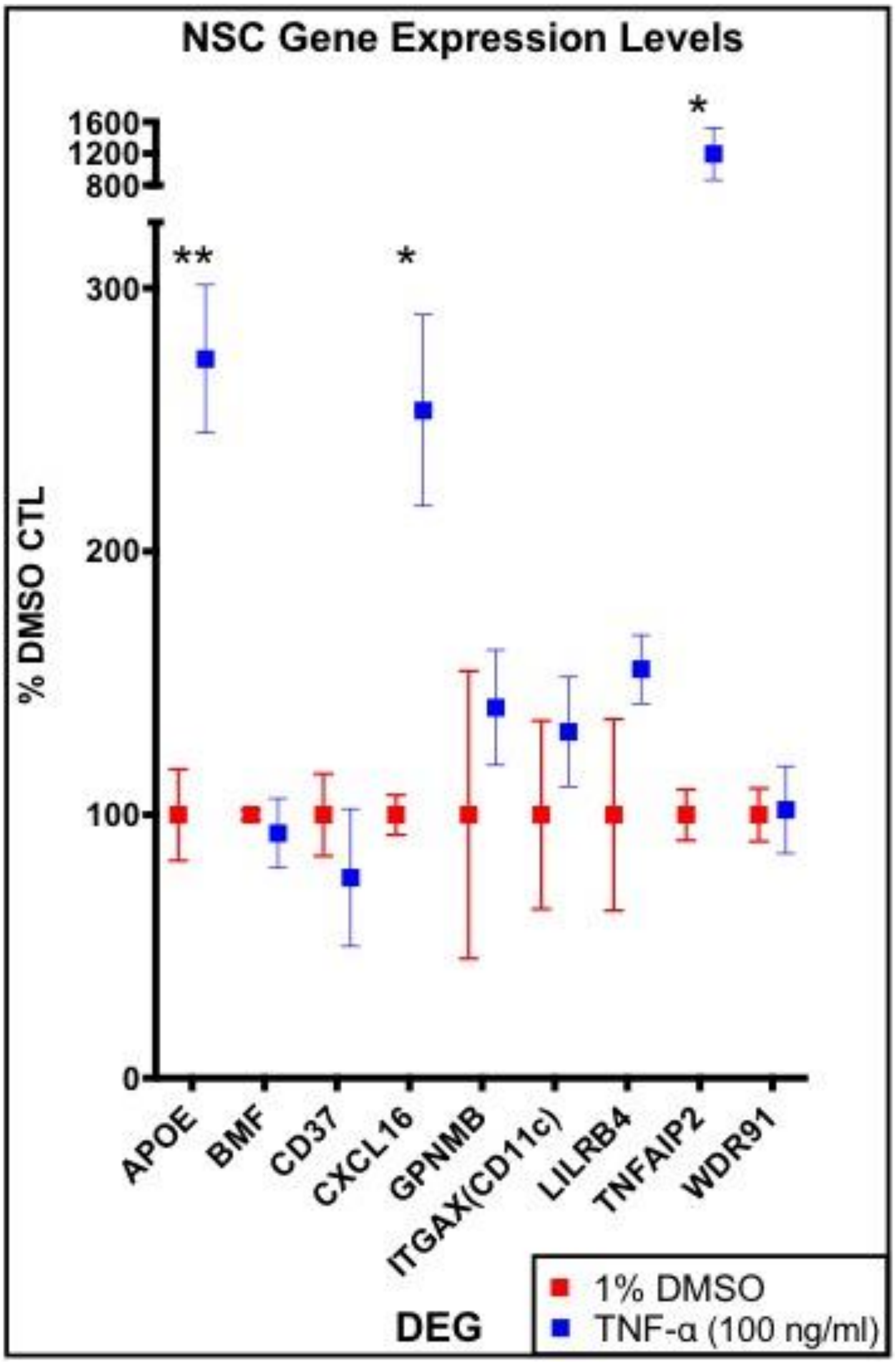


Figure 22. Hub gene expression levels in NSCs after TNF-α exposure. This figure shows each of the 9 hub gene's expression levels in NSCs (n = 8 per group) after treatment with DMSO or TNF-α for 24 hours. Multiple t-tests were run using Prism, which included corrections for multiple comparisons using the Holm-Sidak method. * P < 0.01; ** P < 0.001. Experiment done in collaboration with Jim Bennett and Paula Keeney.

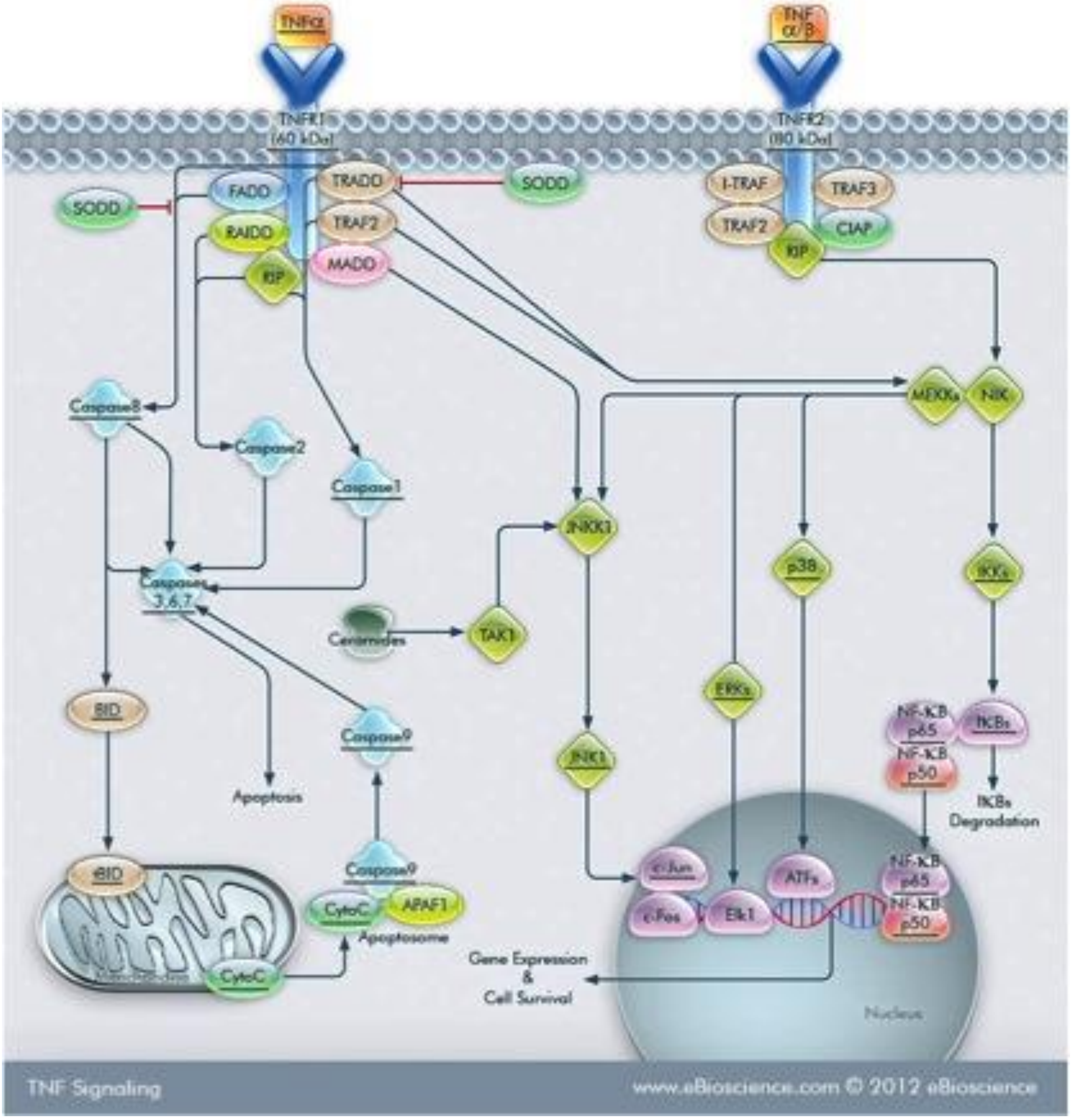


Figure 23. TNF signaling cascades. This figure shows proteins involved in TNF-α signaling pathways that promote apoptosis (left) and cell survival (right). TNF signaling. eBioscience; [accessed 2016 Feb 26].Adapted from http://www.ebioscience.com/resources/pathways/tnf-signaling-pathway.htm.

3 and 7, which then enter the nucleus and initiate heterochromatic formations and DNA fragmentation leading to cell death (Al-Lamki, Mayadas 2015, Falschlehner, Emmerich, Gerlach, Walczak 2007, Matthews, Newbold, Johnstone 2012).The mitochondrial route involves activated Caspase 8 truncating a pro-apoptotic BCL-2 family member, BH3 interacting-domain death agonist (BID). Truncated BID activates BCL-2-Associated X Protein (BAX) and/or BCL-2 Homologous Antagonist Killer (BAK), which move to the mitochondrial membrane and form homo-oligomers. These protein oligomers permeabilize the outer mitochondrial membrane by inserting themselves into it, leading to the release of Cytochrome C. Cytochrome C binds the pro-form of Caspase-9 and Apoptotic Peptidase Activating Factor 1 (APAF1) to form the apoptosome, and the apoptosome proteolytically activates Caspase-9. Activated Caspase-9 then proteolytically activates Caspases 3 and 7, which enter the nucleus and initiate heterochromatic formations and DNA fragmentation leading to cell death (Al-Lamki, Mayadas 2015, Falschlehner, Emmerich, Gerlach, Walczak 2007, Matthews, Newbold, Johnstone 2012).

We conducted two experiments that assessed 1) whether overexpression of TNFAIP2 (observed as a significantly upregulated sALS group-specific DEG) promoted cell death within *in vitro* models of disease-vulnerable cells, and 2) whether TNFAIP2-mediated cell death relied on activated caspases 8 and/or 9. Specifically, our first experiment investigated whether NSCs that transiently overexpressed TNFAIP2 and GFP were significantly less viable than neural stem cells that transiently overexpressed GFP alone. This experiment also investigated whether inhibiting activated caspase 8 or activated caspase 9 reversed any potential TNFAIP2-mediated reduction in NSC viability. Our second experiment investigated whether iPSC-derived motor neuron cultures that transiently overexpressed TNFAIP2-GFP (TNFAIP2 protein with a GFP tag fused to its N-terminus) were significantly less viable and had increased activated caspase 3 and 7 levels relative to iPSC-derived motor neuron cultures that transiently overexpressed GFP alone.

For our first experiment, we compared cell viability in NSCs that 1) were treated with DMSO alone (DMSO), 2) were transfected with Fugene reagent alone (FG), 3) were transfected with Fugene reagent and GFP plasmid to overexpress GFP protein (GFP/FG), 4) were transfected with Fugene reagent and TNFAIP2-IRES-GFP plasmid to overexpress TNFAIP2 and GFP proteins (TIG/FG), 5) were given Caspase 8 inhibitor and were transfected with Fugene reagent and TNFAIP2-IRES-GFP plasmid to overexpress TNFAIP2 and GFP proteins (C8i/TIG/FG), and 6) were given Caspase 9 inhibitor and were transfected with Fugene reagent and TNFAIP2-IRES-GFP plasmid to overexpress TNFAIP2 and GFP proteins (C9i/TIG/FG).

A one way-between subjects ANOVA revealed there was a significant effect of treatment conditions on cell viability at the p<.05 level for the six treatments (F5, 42) = 106.6, p<0.0001). Post hoc comparisons using the Dunnett's test for multiple corrections indicated the mean score for the TIG/FG group (M = 0.207, SD = 0.011) was significantly different than the DMSO group (M = 0.442, SD = 0.030, p=<0.0001), the FG group (M = 0.324, SD = 0.024, p=<0.0001), the GFP/FG group (M = 0.255, SD = 0.023, p=0.0019), and the C9i/TIG/FG group (M = 0.264, SD = 0.031, p=0.0002). Taken together, these results revealed NSCs that transiently overexpressed TNFAIP2 and GFP were significantly less viable than NSCs that overexpressed GFP alone. Further, inhibition of activated caspase 9 reversed this TNFAIP2-mediated reduction in NSC viability, whereas inhibition of activated caspase 8 did not. These results can be seen in Figure 24.

Compared to iPSC-derived motor neurons transfected with GFP alone, iPSC-derived motor neuron cultures that transiently overexpressed TNFAIP2-GFP were 1) significantly less viable and 2) had significantly elevated levels of activated caspases 3 and 7. These results can be seen in Figure 25. qPCR data revealed TNFAIP2 expression increased >300-fold in iPSC-derived motor neuron cultures transfected with TNFAIP2-GFP relative to GFP alone.

These proof of concept experiments demonstrated transient overexpression of TNFAIP2 (an upregulated sALS group-specific DEG) promoted cell death within *in vitro* models of sALS disease-vulnerable cell types. Further, the TNFAIP2-mediated reduction in NSC viability relied on activated Caspase 9, a protein necessary for TNF-α superfamily mitochondrial-mediated apoptosis. Taken together, these results are consistent with our hypothesis that TNFAIP2 functionally contributed to spinal motor neuron death in our sALS patients via the TNF superfamily mitochondrial apoptotic pathway.

IV. Discussion:

In this chapter, we combined deep RNA-Sequencing, systems biology analyses, and molecular biology assays to elucidate sALS group-specific differences in postmortem spinal tissues that may be relevant to disease pathology. To our knowledge, our investigation is the only one that exploits the benefits of next generation RNA-Sequencing

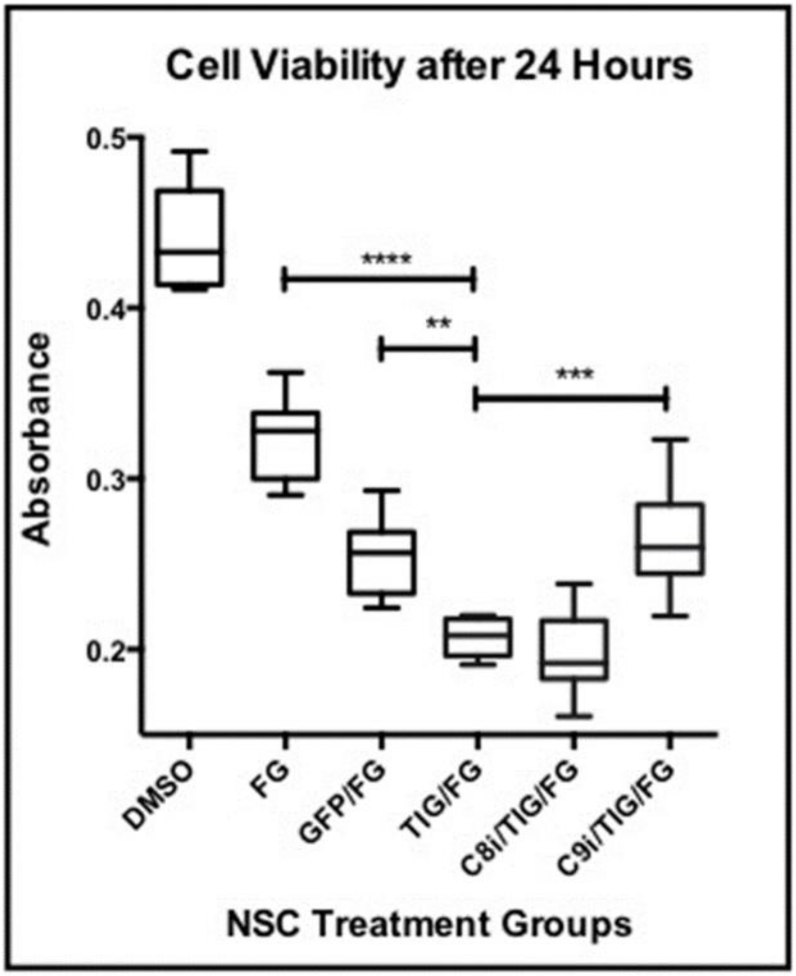


Figure 24. Cell viability measurements for NSC transfection groups. This figure shows absorbance readings reflecting MTT metabolism (a function viable cells perform) for our NSC transfection groups (n = 8 per group). DMSO = DMSO only, FG = Fugene Only, GFP/FG = GFP plasmid with Fugene, TIG/FG = TNFAIP2-IRES-GFP and Fugene, C8i/TIG/FG = Caspase 8 inhibitor, TFNAIP2-IRES-GFP, and Fugene, and C9i/TIG/FG = Caspase 9 inhibitor, TFNAIP2-IRES-GFP, and Fugene. A one-way ANOVA was run using Prism, which included corrections for multiple comparisons using Dunnett's Test. ** P< 0.01, *** P < 0.001; **** P < 0.0001. Experiment done in collaboration with Jim Bennett and Paula Keeney.

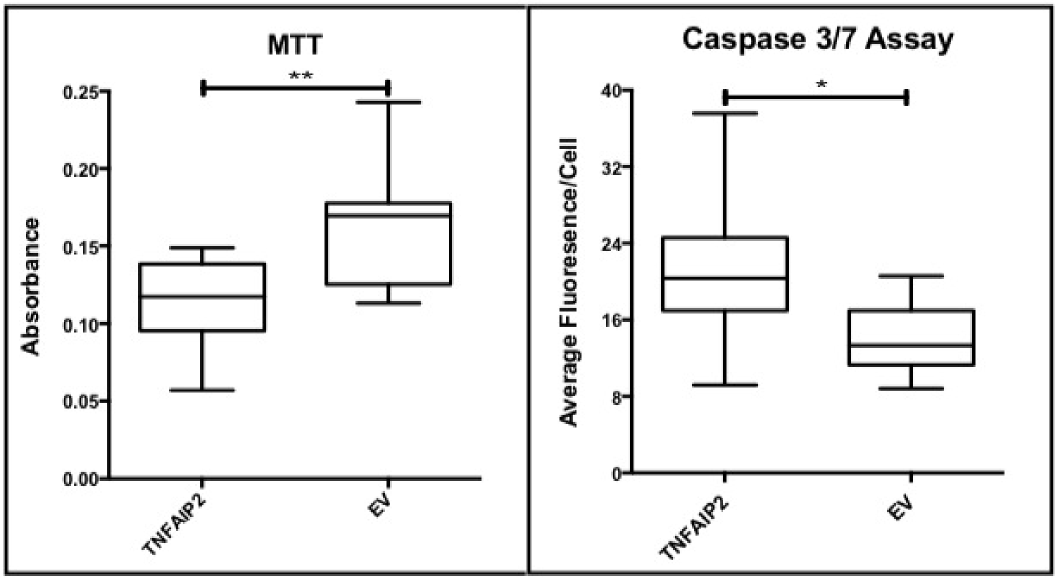


Figure 25. Apoptosis assays in iPSC-derived motor neuron transfection groups. This figure shows absorbance readings reflecting MTT metabolism (a function viable cells perform) on the left, and activated Caspase 3/7 levels measured using fluorescently labeled antibodies on the right for our iPSC-derived motor neuron groups (n = 10 per group). EV = GFP plasmid and Fugene, TNFAIP2 = TNFAIP2-GFP plasmid and Fugene. Unpaired t-tests were run using Prism. * $P < 0.05$, ** $P < 0.01$. Experiment done in collaboration with Laura O' Brien.

(Kratz, Carninci 2014, Wang, Gerstein, Snyder 2009) to measure gene expression differences in sALS patients' postmortem cervical spinal tissues containing disease-vulnerable motor neurons. We chose to study gene expression differences in human sALS patients' postmortem tissues over fALS rodent tissues, as recent findings suggest the majority of sALS is not accounted for by known monogenic ALS causal mutations (Renton, Chio, Traynor 2014). Further, these cervical spinal tissues are inaccessible prior to these patients' deaths. The only other RNA-Sequencing study in human postmortem sALS tissues we are aware of used cerebellar and prefrontal cortex tissues (Pavlou, Dimitromanolakis, Diamandis 2013), and they also found sALS-group specific DEGs associated with inflammatory processes.

Previous studies have used gene network analyses to reveal cellular processes associated with ALS group-specific networks that may be relevant to disease pathology. Studies using gene co-expression network analyses identified ALS group-specific networks associated with immune response, stress response, post-translational modifications, and neuroprotective processes (Holtman et al. 2015, Saris et al. 2009). Several ALS studies modeled gene networks by only connecting genes with known interactions in previous literature. They identified ALS group-specific networks associated with organismal injury, immune response, post-translational modification, regulation of the cytoskeleton, and extracellular matrix repair (Satoh et al. 2014, Figueroa-Romero et al. 2012).

Another group (Izik et al. 2015) used a 2-step approach to identify ALS group-specific gene networks. First, they connected genes based on their co-expression values. Second, they used an information theoretic algorithm to eliminate indirect connections between genes inferred to be connected based on the strength of their co-expression value alone. They then utilized a MARINa algorithm to identify major regulators (such as transcription factors) within an ALS group-specific gene network. The MARINa algorithm predicted 8 network genes were responsible for the elevated rate of apoptosis observed in their *in vitro* motor neuron model of ALS. One of those genes, Nuclear Factor of Kappa Light Polypeptide Gene Enhancer in B-Cells 1 (NFKB1), is a TF with important functions in innate immune responses. Taken together, these findings support the use of systems-level gene network analyses to identify cellular processes that may be perturbed in ALS tissues. Further, they hold potential to unveil therapeutic target genes.

In this study, QIAGEN's IPA revealed inflammatory processes and TNF-α signaling were statistically significantly associated with sALS group-specific gene expression differences identified using independent exploratory DEG analyses (Cuffdiff2, DESeq2, and EdgeR) and an unsupervised gene co-expression network analysis (WGCNA). This is consistent with previous ALS studies' findings, as referenced in the introduction.

qPCR data revealed exposing NSCs to TNF-α increased the expression levels of TNFAIP2, Apoliprotein E (APOE), and Chemokine Ligand 16 (CXCL16). All three of these genes were identified as 1) upregulated sALS group-specific DEGs, and 2) hub genes in a gene co-expression network associated with sALS disease status, inflammatory processes, and TNF-α signaling. These findings support IPA's prediction of TNF-α as an upstream regulator of black module genes, and may account for the inflammatory processes associated with these genes. The observed increase in TNFAIP2 expression after NSCs were exposed to TNF-α corroborates previous findings (Saito et al. 2013, Zhou, Scoggin, Gaynor, Williams 2003, Tian et al 2005), and is consistent with our hypothesis that elevated TNF-α signaling increased TNFAIP2 expression in our sALS patients' cervical spinal cells.

MTT assay results revealed NSCs that transiently overexpressed TNFAIP2 and GFP were significantly less viable than NSCs that overexpressed GFP alone. Further, inhibition of activated Caspase 9 reversed this TNFAIP2-mediated reduction in NSC viability. Relative to iPSC-derived motor neuron cultures that transiently overexpressed GFP, iPSC-derived motor neuron cultures that overexpressed TNFAIP2-GFP were 1) significantly less viable as measured using an MTT assay, and 2) had significantly higher levels of activated caspases 3 and 7.

These proof of concept experiments demonstrated transient overexpression of TNFAIP2 (an upregulated sALS group-specific DEG) promoted cell death within *in vitro* models of sALS disease-vulnerable cell types. This is consistent with previous literature linking elevated TNFAIP2 expression with increased apoptosis (Park et al. 2003, Rusiniak et al. 2000, Ma et al. 2003). The observed TNFAIP2-mediated reduction in neural cell viability relied on activated Caspase 9, a protein necessary for TNF-α superfamily mitochondrial-mediated apoptosis. Taken together, these results are consistent with our hypothesis that TNFAIP2 functionally contributed to spinal motor neuron death in our sALS patients via the TNF superfamily mitochondrial apoptotic pathway.

Modulating TNF signaling activity may be effective in slowing sALS disease progression. TNF-α is a potent inflammatory cytokine that plays an instrumental role in cell fate decisions, and elevated TNF-α signaling has been shown to kill motor neurons in previous literature (He, Wen, Strong 2002, Robertson et al. 2001, Terrado et al. 2000).

TNF signaling-mediated pro-survival processes are largely effected via upregulation of the TFs NFKB1 and Jun Proto-Oncogene (JUN) (Micheau, Tschopp 2003, Walczak 2011), whereas its cell death processes are ultimately carried out by initiator and effector caspases. Bioactive forms of TNF-α commence these processes via two cell surface receptors, TNFR1 and TNFR2. TNFR1 directs cell survival or death, whereas TNFR2 is only known to promote pro-survival effects (Probert 2015).

The extracellular domains of TNFR1 and TNFR2 are shed into general circulation after interacting with bioactive forms of TNF-α, and function in a negative feedback loop as they retain their ability to bind TNF-α (Mohler et al. 1993). Intriguingly, elevated levels of TNF-α and extracellular domains of TNFR1 and TNFR2 have been found in the blood (Poloni et al. 2000) and serum (Babu et al. 2008, Cereda et al. 2008) of human ALS patients compared to controls.

Novel therapies to reduce TNF-α synthesis in human sALS patients could be of great therapeutic value. Non-selective TNF-α inhibitors have proven invaluable in the treatment of chronic diseases with an inflammatory component including rheumatoid arthritis, psoriasis, and inflammatory bowel disease (Probert 2015). Measurement of sALS patients' circulating levels of TNF-α and the extracellular domains of the TNF receptors (TNFR1, and TNFR2) at various treatment timepoints could help establish therapeutic efficacy. They would also serve as non-invasive biomarkers of disease progression.

Two potential therapeutic agents to reduce TNF-α synthesis are Bupropion and curcumin. Bupropion, a drug commonly used to treat clinical depression, decreased TNF-α serum levels in mice likely via increasing intracellular cAMP signaling after binding beta-adrenergic and/or D1 receptors (Brustolim et al. 2006). Curcumin, an anti-inflammatory compound in turmeric, reduced TNF-α transcription in human cancer cells (Han, Keum, Seo, Surh 2002, Surh et al. 2001) and lipopolysaccharide (LPS)-stimulated murine microglia (Jin et al. 2007). Curcumin likely reduced TNF-α transcription via inhibition of NFKB1. NFKB1, a TF that is upregulated by TNF signaling, is known to induce TNF-α and other inflammatory cytokines under certain biological circumstances (Hoesel, Schmid 2013). Curcumin is also predicted to bind and inhibit caspase-3 (Khan et al. 2015), an effector caspase used by the TNF superfamily mitochondrial and non-mitochondrial apoptotic pathways. Curcumin oral bioavailability and brain penetration was substantially increased by micellular formulation (Hagl et al. 2015), setting the stage for clinical testing of it and Bupropion.

Our study has several important limitations. First, we measured and compared gene expression in a small number of postmortem cervical spinal cord section samples. There are ~35,000 persons with ALS in the US. The cost of RNA-sequencing limited the numbers of cases we could examine at the sequencing depth employed. As a result, it is impossible to state to what degree our findings can be generalized to thousands of patients. Second, we used postmortem tissue. As a result, we examined gene expression of cells (mainly astrocytes) that were survivors of the neurodegenerative process. To what extent ALS modifies gene expression over time is not known, and it is not currently possible to examine human CNS tissues across disease progression. It is unclear whether the young motor or other neurons we produced using iPSC approaches approximate changes seen in sALS patients' spinal motor neurons that are present for many years as ALS progresses. Third, we did not explore novel transcripts or smaller ncRNAs (including miRNAs) in this study.

Although we focused on TNF-α signaling and modulated TNFAIP2 expression in NSCs and iPSC-derived motor neurons in this study, we do not claim aberrant inflammatory TNF-α signaling is the sole pathogenic factor in sALS. We identified a second sALS group-specific gene co-expression network that was associated with cell proliferation, cell cycle functions, interleukin-4 (IL4) signaling, and various metabolic compounds' (methyglyoxal and phenylethylamine) degradation processes. We prioritized pursuit of candidate genes in the black module, as that gene co-expression network's associated cellular processes were more plausibly linked to cell death. As sALS patients die after the motor neurons that innervate their lungs degenerate, identifying novel therapeutic targets to prevent motor neuron death is paramount. The black module appeared to be a better option than the sienna4 module for pursuing that goal.

Aside from TNFAIP2, we identified 8 other hub genes that were also upregulated DEGs within the black module. These genes could serve as foci for additional mechanistic studies and therapeutic interventions. Specifically, investigating whether transient overexpression of each of them (as they were all upregulated in our sALS sample group) leads to perturbed cellular processes (such as increased apoptosis) in models of sALS disease-vulnerable cell types would be valuable. Previous literature findings highlight their potential relationship to perturbed cellular processes important to ALS pathology.

Genetic variants in APOE, a gene encoding a protein important for transporting cholesterol and other lipids between cells, have been shown to modify ALS age of onset and features of disease progression (Verghese, Castellano, Holtzman 2011). This may be related to aberrant cholesterol transport processes in sALS disease-vulnerable cells, as accumulation of cholesterol esters has been linked to oxidative stress-induced motor neuron death in ALS tissues previously (Cutler et al. 2002). Cholesterol was identified as a significant upstream regulator of genes comprising the black module. It is possible upregulation of APOE (as observed in our sALS sample group) contributed to aberrant cholesterol transport processes in our sALS patients spinal cells, leading to oxidative stress-induced spinal motor neuron death. Transient overexpression of APOE in our *in vitro* models of sALS disease-vulnerable cell types followed by measurement of cholesterol ester and cell viability levels would be valuable.

Bcl2-Modifying Factor (BMF) binds to Bcl2 and related anti-apoptotic proteins and promotes mitochondrial-mediated apoptosis. Transient overexpression of BMF in our *in vitro* models of sALS disease-vulnerable cell types followed by characterization of cell viability levels could shed light on the potential contributions of mitochondrial-mediated apoptosis in sALS.

CD37, a gene whose expression is restricted to human leukocytes, is integral to T cell proliferation (van Spriel et al. 2004). While the majority of immune surveillance in the CNS (including the spinal cord) is carried out by microglia, T cells do play a role in this process as well (Ousman, Kubes 2012). CD37's observed upregulation in our sALS patients' spinal cells may reflect recruitment of activated T cells to combat deleterious cellular processes induced by the disease. Recruited T cells may have excreted large amounts of TNF-α in this process, as they have been shown to do that in a previous study (Ofotokun et al. 2015). This may have ultimately led to our sALS patients' spinal motor neuron death. Modeling CD37 overexpression in a rodent model followed by measurement of TNF- α levels in their spinal tissues could help assess the likelihood of this connection.

CXCL16, a transmembrane chemokine produced by reactive astroglial cells, plays an important role in immunosurveillance processes and serves as a chemoattractant for macrophages. Its expression is increased by TNF-α (Abel et al. 2004), and has been shown to sensitize cells to TNF-α mediated apoptosis (Kee et al. 2014). It also promotes CXCR6-positive glial cell invasion that favors astrogliosis (Hattermann et al. 2008), a feature seen in ALS CNS tissues. Taken together, its observed upregulation in our sALS patients' spinal cells may have played a role in sensitizing spinal motor neurons to TNF-α mediated apoptosis. This could be tested in our *in vitro* models of sALS disease-vulnerable cell types.

Glycoprotein NMB (GPNMB) was previously identified as an upregulated DEG in the spinal cords of fALS rodents. Interestingly, extracellular fragments of GPNMB released by activated astrocytes lessened the neurotoxicity of mutant SOD1, suggesting it may play a protective role against neurodegeneration (Tanaka et al. 2012). Its observed upregulation in our sALS patients' spinal cells may reflect an attempt to protect motor neurons against ongoing neurodegenerative processes related to sALS pathology.

Integrin Alpha X (ITGAX), a leukocyte-specific integrin, was found as an upregulated DEG in leukocytes that invaded the spinal cords of fALS rodents at different stages of disease progression (Chiu et al. 2008). ITGAX plays a known role in cell-cell interactions during immune responses, and its observed upregulation in our sALS patients’ spinal cells may reflect spinal cells’ recruitment of activated T cells to combat deleterious effects of sALS pathology.

Leukocyte Immunoglobulin-Like Receptor Subfamily B Member 4 (LILRB4), a cell surface receptor in immune cells, binds MHC class 1 molecules to inhibit immune responses. While not directly studied in ALS tissues, LILRB4 expression negatively correlated with pathologic inflammation in a mouse model of allergic pulmonary inflammation (Fanning et al. 2013). Its observed upregulation in our sALS patients’ spinal cells may reflect an attempt to reduce inflammatory processes that may have ultimately led to motor neuron death.

To our knowledge, no study has investigated the function of WDR91 (WD Repeat Domain 91), so it is impossible to speculate on its possible connection to ALS pathology. While more than half (42/74) of our sALS group-specific DEGs were contained in the black module, 32 DEGs were not. These genes could also play an important role in sALS disease pathology and may warrant further study. Chemokine C-X-C Motif Ligand 8 (CXCL8), Decorin (DCN), and Neuropilin 1 (NRP1) are particularly good candidates.

CXCL8 was found to be statistically significantly increased in the cerebrospinal fluid of sALS patients compared to cerebrospinal fluid of patients with other non-inflammatory neurological diseases. Further, its level was negatively correlated with these patients’ scores on the revised ALS functional rating scale (Tateishi 2010). CXCL8 plays an important role in sending neutrophils to a site of infection, as well as inducing phagocytosis. DCN’s mRNA and protein expression levels were greatly increased in both astrocytes and spinal cells of fALS rodents carrying a causal SOD1 mutation. DCN is a proteoglycan with a known role in attenuating glial scar formation and inflammation. (Vargas et al. 2008). Both CXCL8 and DCN downregulation in our sALS patients' spinal cells may have contributed to unresolved inflammatory processes that ultimately led to motor neuron death.

Axon degeneration is often observed in fALS rodent models prior to motor neuron death. Semaphorin 3A (SEMA3A), an important axon guidance cue involved in neural patterning during development, binds to NRP1. This leads to axonal retraction by destabilizing microtubules and microfilament networks. Blocking the interaction between SEMA3A and NRP1 in fALS rodents led to decreased axon degeneration and motor neuron death, suggesting NRP1 may play a role in ALS pathology prior to clinical symptom onset related to motor neuron death (Venkova et al. 2014).

We anticipate future exploratory studies will continue to uncover polygenic contributions, perturbed cellular processes, and potential therapeutic targets in sALS. Genome-wide Association Studies (GWAS) comparing sALS cases vs. neurologically healthy controls were instrumental in the discovery of excess pathogenic non-coding repeats in C9orf72 found in 7% of Caucasian sALS patients (Renton, Chio, Traynor 2014). Another study identified excess de novo mutations in chromatin regulator genes after comparing exome sequencing data from sALS offspring and their neurologically healthy parents (Chesi et al. 2013). This ALS gene expression study joins those preceding it in identifying perturbed cellular processes and corroborating them using separate molecular biology assays.

In this study, we identified sALS group-specific gene expression differences and associated cellular processes that may be relevant to disease pathology. However, it lumped all of our sALS patients together to find commonalities across them without elucidating differences between them.

Emerging findings suggest considerable clinical heterogeneity between patients with either form of ALS. A recent study (Ganesalingam et al. 2009) applied a latent class cluster analysis to 1,467 ALS patients' clinical metrics to assess whether there were multiple disease sub-groups. These metrics included family history of ALS (fALS vs. sALS), sex, ethnicity, site of symptom onset, age of onset, and diagnostic delay after symptom onset. Five different groups emerged, with one group showing no deaths and another exceeding the average median survival time by 12 years. Heterogeneity in disease features has also been observed in fALS patients carrying causal mutations in different genes. Patients carrying causal mutations in the FUS gene show a younger age of onset and rapid disease progression compared to those with the SOD1 Asp90Ala variant (Ganesalingam et al. 2009). Even within the same fALS pedigree, some family members who inherit a fALS-causal mutation do not develop disease features.

Perhaps more striking were study findings where the same fALS causal mutation was modeled into two genetically distinct transgenic mouse lines (Nardo et al. 2013) These researchers demonstrated introducing the SOD1 G93A point mutation into 129Sv and C57 mice led to a rapid and slow disease progression, respectively. They compared measured gene expression values from each transgenic line's spinal motor neurons at multiple disease stages, and identified hundreds of DEGs associated with different cellular processes at 3 different timepoints. DEGs specific to the 129Sv group (showing rapid disease progression) were associated with reduced mitochondrial function and deficient protein degradation. DEGs specific to the C57 group (showing slow disease progression) were associated with upregulated immune system processes. These findings suggest even with a pure monogenic form of ALS and a controlled environment, genetic differences between animals greatly contributed to clinical disease features.

The above findings imply elucidating genetic differences between ALS patients will likely be necessary to fully explain for their disease features. As we only had 7 sALS samples and 8 neurologically healthy controls, we had limited statistical power to identify these differences. However, as sequencing costs decrease, larger sample sizes conferring greater statistical detection power will become feasible. These data sets will likely enable stratification of sALS by its varied molecular phenotypes as has been seen in other diseases like breast cancer (Cancer Genome Atlas Network 2012). These approaches may ultimately lead to therapies against pathways that are universally beneficial to sALS patients, such as TNF signaling, as well as those specifically tailored to an individual patient's pathophysiology.

CHAPTER 3: Mitochondrial gene expression levels were not aberrant in sALS patients' postmortem cervical spinal sections

I. Introduction:

Mitochondrial abnormalities have been identified in ALS tissues in numerous studies dating back to 1994. Early studies found mitochondria isolated from fALS rodents' spinal motor neurons had aberrant morphologies (Higgins, Jung 2003, Kong, Xu 1998). Mitochondria isolated from human sALS and fALS patients' muscle tissues, spinal cells, and postmortem motor neurons have also shown morphological abnormalities (Crugnola et al. 2010, Echaniz-Laguna et al. 2006, Sasaki, Iwata 1996, Hirano, Donnenfeld, Sasaki, Nakano 1984, Sasaki, Iwata 2007). In addition to aberrant mitochondrial morphology, defective electron transport chain (ETC) activity and oxidative phosphorylation (OXPHOS) rates have been observed in various tissues from fALS rodents and human ALS patients. Mitochondria isolated from fALS rodents' brain and spinal cord tissues had significantly decreased rates of OXPHOS compared to mitochondrial isolated from matched control tissues (Mattiazzi et al. 2002). Postmortem spinal cord tissue from both sALS and fALS patients have shown decreased activity of ETC complexes I, II, III, IV, and V (Borthwick et al. 1999, Fujita et al. 1996, Wiedemann et al. 2002), which may reflect selective loss of mitochondria in those spinal cells. Further, skeletal muscle from sALS patients also displayed aberrant ETC activity, specifically in complexes 1 and 4 (Crugnola et al. 2010, Vielhaber et al. 2000, Wiedermann et al. 1998).

A relatively recent study (Cassina et al. 2008) linked defective OXPHOS activity in fALS rodents' spinal cell mitochondria to increased spinal motor neuron apoptosis. Mitochondria isolated from fALS rodent astrocytes had defective OXPHOS characterized by decreased oxygen consumption, lack of ADP-dependent respiratory control, and decreased membrane potential. Interestingly, these fALS rodents' astrocytes (but not wildtype rodents' astrocytes) induced death of spinal motor neurons when both cell types were co-cultured *in vitro*. After treating these fALS rodent astrocytes with mitochondrial-targeted antioxidants (ubiquinone and carboxy-proxyl nitroxide), they showed improved mitochondrial OXPHOS and did not induce motor neuron death when co-cultured with motor neurons *in vitro.* Taken together, these findings suggest defective OXPHOS related to mitochondrial dysfunction in spinal cells may contribute to spinal motor neuron death.

Members of our lab recently published a study (Ladd, Keeney, Govind, Bennett 2014) comparing 76 (72 nuclear-encoded and 4 mitochondrial-encoded) genes' expression levels in various tissues from ALS patients and neurologically healthy controls. These 76 genes encode components of the ETC and OXPHOS complexes I-V. Their expression levels were compared in 16 postmortem cervical spinal section samples (10 ALS and 6 neurologically healthy controls) and 20 peripheral blood mononuclear cell samples (9 ALS and 11 neurologically healthy controls). Postmortem spinal sections from select samples in my dissertation project (all 7 sALS patients and 3 neurologically healthy controls) were analyzed in this referenced study.

Interestingly, the 4 mitochondrial-encoded OXPHOS genes (12s rRNA, COX3, ND2, and ND4) had statistically significantly decreased expression levels in both ALS sample groups (postmortem spinal sections and peripheral mononuclear blood cells) relative to their respective control sample groups. The majority of the 72 nuclear-encoded OXPHOS genes had decreased expression levels in the sALS patients' postmortem spinal section samples relative to the neurologically healthy controls' postmortem spinal section samples. Taken together, reduced expression of these OXPHOS genes may have led to aberrant OXPHOS activity in these ALS patients' spinal cells. Further, that may have promoted spinal motor neuron death.

In this chapter, we used a combination of RNA-sequencing and bioinformatics tools to elucidate sALS group-specific mitochondrial gene expression level differences. We assessed whether any of the 37 annotated mitochondrial genes in the hg19 human reference mitochondrial genome were differentially expressed in our sALS patients' postmortem cervical spinal section samples relative to our neurologically healthy controls' postmortem cervical spinal section samples.

II. Methods:

Input sample datasets: We used the same 7 sALS and 8 neurologically healthy control samples' Read 1 and Read 2 FastQ files output by Trimmomatic (as described in Chapter 2) for our downstream Mitochondrial DEG analysis.

Pre-alignment steps: Prior to attempting alignment of each sample's paired end reads to the hg19 human reference mitochondrial transcriptome then hg19 human reference mitochondrial genome, I needed to generate these files. The hg19 human reference transcriptome file obtained from the Illumina iGenomes UCSC hg19 directory as described in Chapter 2 did not include mitochondrial RNA transcripts.

To accomplish this, I downloaded ENCODE's hg19 human reference mitochondrial transcriptome file (known mtRNA transcripts) from the UCSC Table Browser using their website (http://genome.ucsc.edu/cgi-bin/hgTables). I specified Feb. 2009 (GRCh37/hg19) for assembly, GENCODE Genes V19 for track, and chrM in the position search field under region. All other fields were unchanged from their default entries.

To generate the hg19 human reference mitochondrial genome file, I simply extracted the mitochondrial sequence (ChrM) found in the hg19 human reference genome text file obtained from the Illumina iGenomes UCSC hg19 directory as described in Chapter 2.

Alignment of paired end reads: We used Tophat2 (Kim et al. 2013) to attempt alignment of each sample's paired end reads to the hg19 human reference mitochondrial transcriptome then hg19 human reference mitochondrial genome. Tophat2 followed the same alignment procedure detailed in Chapter 2 and shown in Figure 10.

HTSeq-Count: We used HTSeq-Count (Anders, Pyl, Huber 2015) to report the total number of paired end reads that aligned to each annotated mitochondrial gene's transcribed regions. HTSeq-Count used the same counting procedure detailed in Chapter

2 and shown in Figure 13. For each sample, HTSeq-Count produced a matrix with all annotated mitochondrial genes and their corresponding paired end read count values.

EdgeR and DEG identification: We used EdgeR (Robinson, McCarthy, Smyth 2010), to detect sALS group-specific mitochondrial DEGs. EdgeR followed the same normalization and testing procedures detailed in Chapter 2 to identify sALS group-specific mitochondrial DEGs. EdgeR reported an associated p-value for each annotated mitochondrial gene tested. We calculated each annotated mitochondrial gene's Benjamini-Hochberg corrected p-value using their corresponding EdgeR reported p-value via the R function p.adjust. Each annotated mitochondrial gene with a Benjamini-Hochberg corrected p-value < 0.10 was identified as a sALS group-specific mitochondrial DEG.

We used EdgeR's default workflow for our mitochondrial DEG analysis. We decided to filter out genes with a cpm (counts per million aligned paired end reads) value <1 in 7 samples. We chose a cpm value of 1 as smaller values likely reflected noise. We chose 7 samples as our threshold, as genes that were only expressed in our disease or control group could play an important role in disease pathology.

DESeq2 and DEG identification: We used DESeq2 (Love, Huber, Anders 2014) to detect sALS group-specific mitochondrial DEGs. DESeq2 followed the same normalization and testing procedures detailed in Chapter 2 to identify sALS group-specific mitochondrial DEGs. DESeq2 reported an associated p-value for each annotated mitochondrial gene tested. We calculated each annotated mitochondrial gene's Benjamini-Hochberg corrected p-value using their corresponding DESeq2 reported p-value via the R function p.adjust. Each annotated mitochondrial gene with a Benjamini-Hochberg corrected p-value < 0.10 was identified as a sALS group-specific mitochondrial DEG.

Cufflinks for gene expression estimates: We separately used Cufflinks (Trapnell et al. 2010) to estimate each sample's mitochondrial gene expression levels for all annotated mitochondrial genes in the hg19 human reference mitochondrial genome. Cufflinks followed the same procedure detailed in Chapter 2 to estimate these mitochondrial gene's expression levels.

Cuffdiff2 for DEG identification: We used Cuffdiff2 (Trapnell et al. 2013) to detect sALS group-specific mitochondrial DEGs. Cuffdiff2 followed the same normalization and testing procedures detailed in Chapter 2 to detect sALS group-specific mitochondrial DEGs. Cuffdiff2 reported an associated p-value for each annotated mitochondrial gene. We calculated all annotated mitochondrial genes' Benjamini-Hochberg corrected p-values using their corresponding Cuffdiff2 p-values via the R function p.adjust. Each annotated gene with a Benjamini-Hochberg corrected p-value < 0.10 was identified as a sALS group-specific mitochondrial DEG.

III. Results:

Mitochondrial DEG testing: We elected to identify sALS group-specific DEGs using Cufflinks/Cuffdiff2, DESeq2, and EdgeR. We made this decision based on the findings of a recent study showing these three analyses have different limitations when they were directly compared (Zhang et al. 2014), as described in Chapter 2.

None of the 37 annotated mitochondrial genes were identified as sALS group-specific DEGs in any of the three analyses. However, 3 of the 4 genes identified as statistically significantly decreased in ALS samples in our lab's previous publication (Ladd, Keeney, Govind, Bennett 2014) had decreased expression levels in our sALS samples. Further, we found all but three of the 13 mitochondrial protein-coding genes (whose translated proteins are part of ETC complexes) were expressed at lower levels in our sALS patients relative to our neurologically healthy controls. These results can be seen in Figures 26 and 27.

IV. Discussion:

Evidence for mitochondrial dysfunction has been identified in ALS tissues using a variety of molecular biology techniques across independent studies. This includes findings of aberrant mitochondrial morphology and defective OXPHOS in ALS tissues from human patients (both fALS and sALS) and fALS rodents (Higgins, Jung 2003, Kong, Xu 1998, Crugnola et al. 2010, Echaniz-Laguna et al. 2006, Sasaki, Iwata 1996, Hirano, Donnenfeld, Sasaki, Nakano 1984, Sasaki, Iwata 2007, Mattiazzi et al. 2002, Borthwick et al. 1999, Fujita et al. 1996, Wiedemann et al. 2002, Vielhaber et al. 2000, Wiedermann et al. 1998, Cassina et al. 2008, Ladd, Keeney, Govind, Bennett 2014).

Defective OXPHOS activity has been shown to increase the production of reactive oxygen species (ROS), a marker for oxidative stress, in affected cells (Duffy, Chapman, Shaw, Grierson 2011). Defective OXPHOS activity in ALS patients' spinal cells may

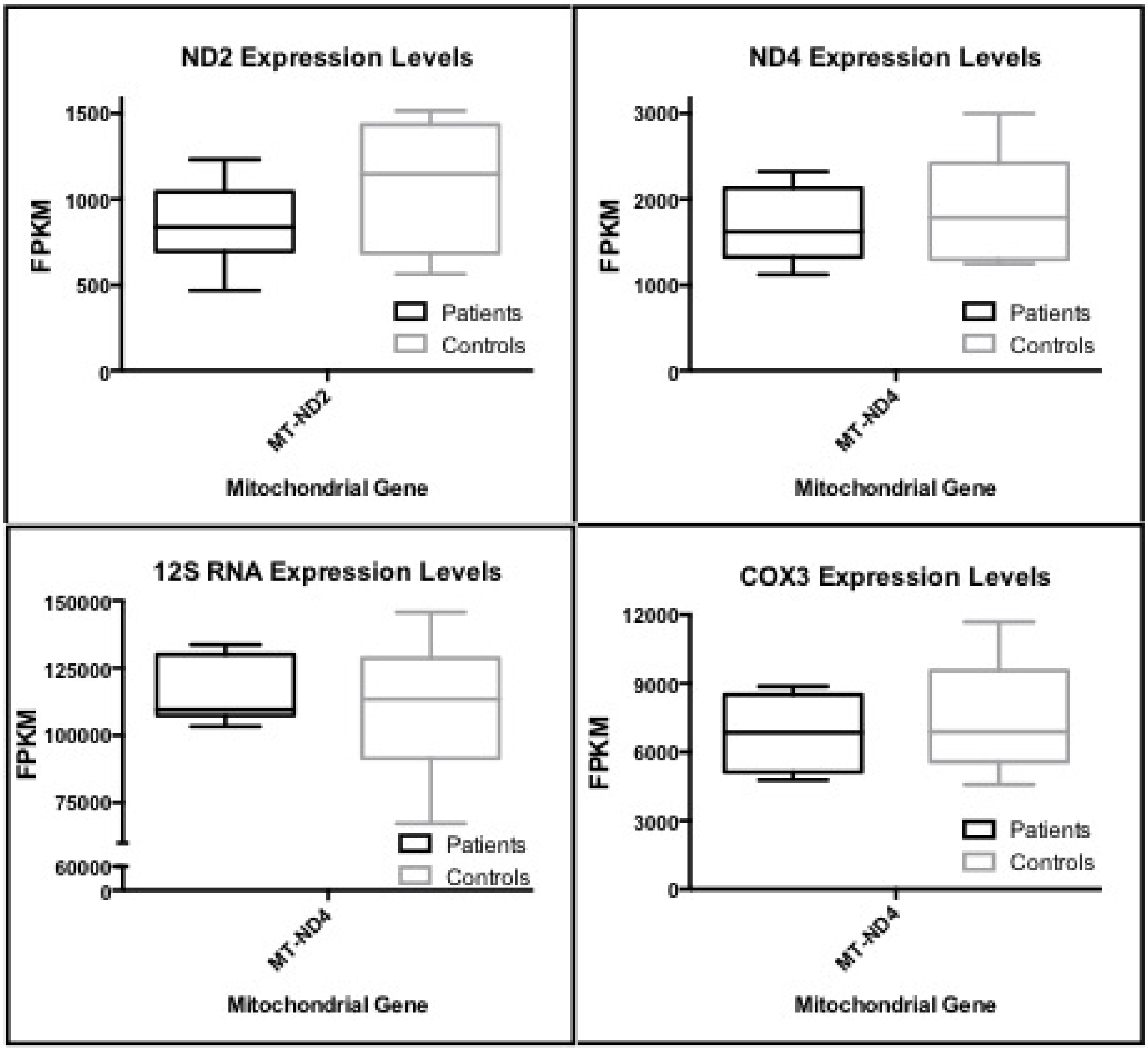


Figure 26. Mitochondrial gene expression levels. This figure shows 4 mitochondrial genes' expression levels (assessed in our lab's previous study) in our sALS sample group vs. our neurologically healthy control sample group. None of these genes were identified as statistically significant DEGs.

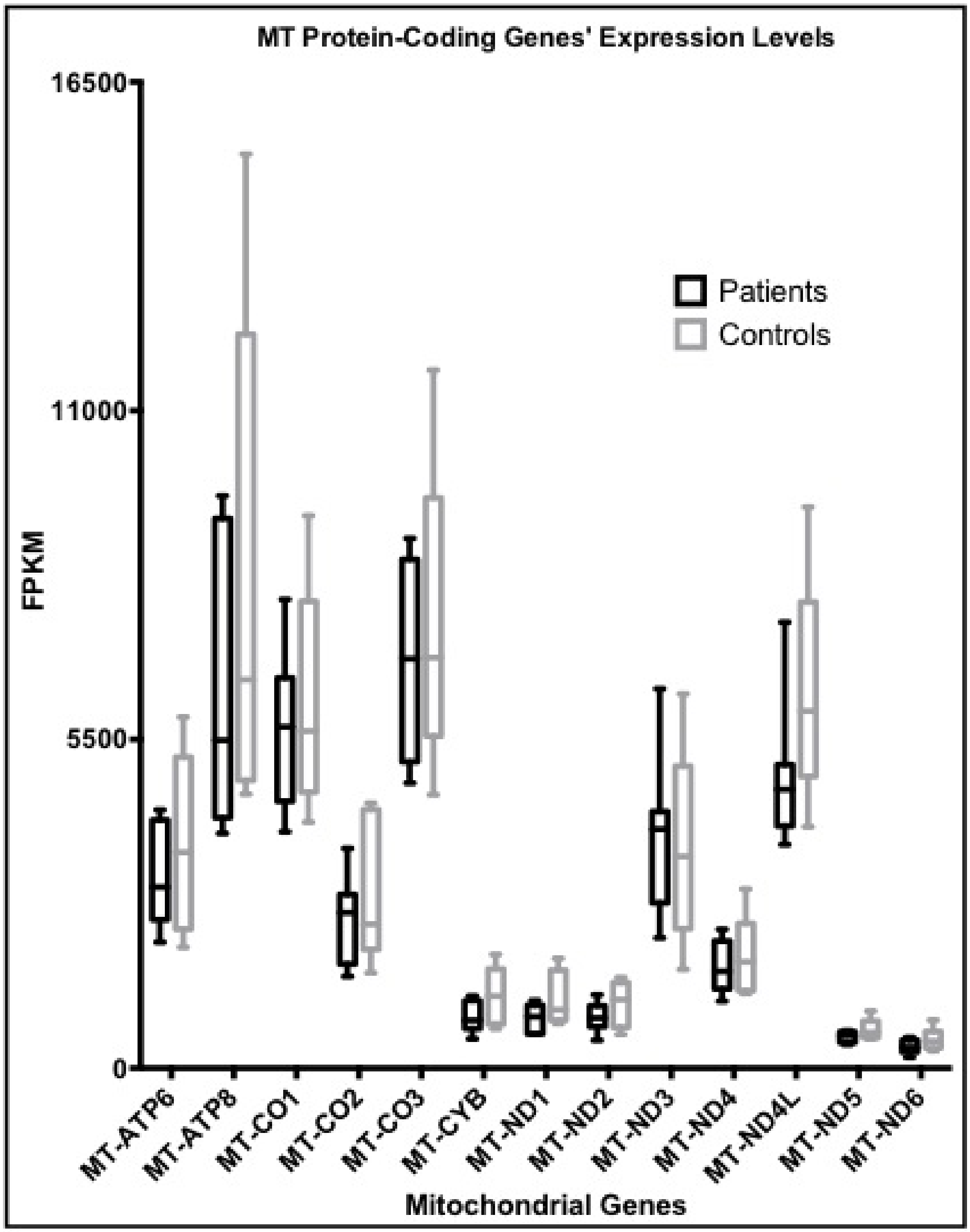


Figure 27. Mitochondrial protein-coding genes' expression levels. This figure shows all 13 mitochondrial protein-coding genes' expression levels in our sALS sample group vs. our neurologically healthy control sample group. None of these genes were identified as statistically significant DEGs.

ultimately promote motor neuron death as a result of elevated oxidative stress. A recent study (Cassina et al. 2008) found mitochondria isolated from fALS rodents' astrocytes 1) showed defective OXPHOS activity, and 2) induced motor neuron death when the two cell types were co-cultured *in vitro.* After treating these astrocytes with mitochondrial-targeted antioxidants, they showed normal mitochondrial respiratory function and did not induce motor neuron death when co-cultured with motor neurons *in vitro.*

Elevated oxidative stress has been observed in ALS tissues from human patients and fALS rodents. One study revealed cerebrospinal fluid from ALS patients had elevated levels of 3-nitrotyrosine, a marker of free radical damage related to oxidative stress, relative to neurologically healthy control samples' cerebrospinal fluid (Duffy, Chapman, Shaw, Grierson 2011). Further, fALS rodents' spinal cord motor neurons have shown evidence of elevated oxidative stress characterized by increased oxyradical production, carbonylation of proteins, and peroxidation of lipids in the mitochondrial membrane. Interestingly, peroxidation of cardiolipin disrupts its interaction with cytochrome C, leading to cytochrome C release from the mitochondrial membrane. It is widely known that cytochrome C release can promote mitochondrial-mediated apoptosis (Al-Lamki, Mayadas 2015, Falschlehner, Emmerich, Gerlach, Walczak 2007, Matthews, Newbold, Johnstone 2012). Aberrant mitochondrial gene expression could plausibly lead to OXPHOS defects (and potentially motor neuron death) in sALS patients' disease-vulnerable tissues.

In this chapter, we combined deep RNA-Sequencing and systems biology analyses to identify sALS group-specific mitochondrial DEGs in postmortem spinal tissues. To our knowledge, this study is the only one that uses RNA-Sequencing to estimate all 37 annotated mitochondrial genes' expression levels in sALS patients' postmortem cervical spinal tissues containing disease-vulnerable motor neurons. We did not identify any sALS group-specific DEGs in this analysis.

Further, we did not replicate our lab's previous findings showing 4 mitochondrial-encoded OXPHOS genes (12s rRNA, COX3, ND2, and ND4) had statistically significantly decreased expression levels in postmortem cervical spinal sections from ALS samples relative to neurologically healthy controls. While there was an overlap in the postmortem cervical spinal section samples examined in this mitochondrial DEG analysis and our lab's previous study, multiple samples were not shared between the two analyses. This likely explains why we did not replicate the previous study's findings.

We did not assess whether any of our sALS patients carried known (or novel) pathogenic variants that lead to defective OXPHOS. Mitochondrial DNA (mtDNA) has an elevated mutation rate, which results in a high frequency of rare variants across individuals (Taylor, Turnbull 2009). Further, the pathogenic mtDNA mutations identified in various mitochondrial diseases invariably lead in defective mitochondrial OXPHOS, resulting in a reduced ability to produce cellular ATP (Tuppen, Blakely, Turnbull, Taylor 2010).

To determine whether any of our sALS patients harbored known or novel pathogenic mitochondrial variants that lead to defects in OXPHOS, I would first identify each patient's mitochondrial SNVs and indels using the GATK pipeline described in Chapter 2. I would next assess whether any of these variants match published pathogenic mitochondrial variants in a database such as MITOMAP (Ruiz-Pesini et al. 2007). For the remaining variants, I would 1) assess whether they were found in healthy individuals' mtDNA sequences catalogued in an online database such as The Human Mitochondrial Genome Database (Ruiz-Pesini et al. 2007), and 2) input them into a tool such as PolyPhen-2 (Adzhubei et al. 2010) to assess whether they are likely deleterious.

There is considerable heterogeneity in sALS patients' clinical features (Al-Chalabi, Hardiman 2013, Ganesalingam et al. 2009), which likely reflects different genetic etiologies underlying different instances of sALS. It is also entirely possible the pathogenesis of sALS in our patients did not involve aberrant mitochondrial gene expression patterns, or defective mitochondrial OXPHOS.

CHAPTER 4: Cholesterol biosynthesis defects may contribute to disease pathology in Postmortem Sporadic ALS patients' cervical spinal sections

I. Introduction:

A role for RNA processing defects in ALS pathology was largely confirmed via the identification of >100 ALS causal mutations in the Fused in Sarcoma/Translocation in Liposarcoma (FUS/TLS) and TAR DNA-binding protein (TARDBP) genes (Renton, Chio, Traynor 2014). These genes encode the RNA-binding proteins FUS/TLS and TDP-43, respectively. FUS/TLS and TDP-43 regulate nuclear RNA processing activities including pre-mRNA splicing, RNA stability, RNA transport, protein translation, and microRNA maturation (Xu 2012, Colombrita et al 2012). Each protein binds >5,000 RNA transcripts (with minimal overlap in which RNA transcripts each binds to), suggesting they are both major regulators of nuclear RNA processing activities (Donnelly, Grima, Sattler 2014). However, it remains unclear how mutations in these two genes promote neurodegenerative processes in ALS patients.

A recent study found iPSC-derived motor neurons carrying a FUS/TLS causal ALS mutation had cytoplasmic aggregates of FUS/TLS (known as FUS/TLS proteinopathy) and elevated levels of apoptosis relative to iPSC-derived motor neurons that did not carry a FUS/TLS causal ALS mutation (Ichiyanagi et al. 2016). Similarly, overexpression of mutant TDP-43 proteins (encoded by the TARDBP gene carrying different causal ALS mutations) led to increased 1) TDP-43 proteinopathy and apoptosis in HEK-293 cells (Mutihac et al. 2015), and 2) TDP-43 proteinopathy and neurodegeneration in a *Drosophila* model (Vanden Broeck et al. 2015).

While <4% of Caucasian sALS patients carry a causal ALS mutation in TARDBP or FUS/TLS as of 2014 (Renton, Chio, Traynor 2014), ~98% of all ALS patients (both fALS and sALS) have TDP-43 proteinopathy in their spinal motor neurons, spinal glial cells, and/or select brain cells (Lagier-Tourenne, Cleveland 2009, Donnelly, Grima, Sattler 2014, Yang et al. 2014). Interestingly, cells exhibiting FUS/TLS or TDP-43 proteinopathy (with or without accompanying FUS/TLS or TARDBP causal ALS mutations) have reduced levels of FUS/TLS or TDP-43 protein in their nuclei, respectively (Kwiatkowski et al. 2009, Yang et al. 2014).

Reduced nuclear levels of FUS/TLS or TDP-43 promote apoptosis in cellular and animals models. iPSC-derived motor neurons with reduced nuclear FUS/TLS had elevated rates of apoptosis compared to iPSC-derived motor neurons with basal nuclear levels of FUS/TLS (Ichiyanagi et al. 2016). Systemic knockdown of TDP-43 in zebrafish led to muscle degeneration, as well as morphological and functional defects in the CNS (Schmid et al. 2013). Similarly, systemic knockdown of TDP-43 in a transgenic mouse line led to an age-dependent neurodegenerative phenotype characterized by motor weakness, paralysis, death of spinal motor neurons and layer V cortical neurons, and premature death (Yang et al. 2014). Taken together, 1) the majority of ALS patients' disease-vulnerable cells likely have reduced nuclear levels of TDP-43 as a result of TDP-

43 proteinopathy, and 2) reduced nuclear levels of TDP-43 (or FUS/TLS) can promote neurodegeneration.

Reduced nuclear levels of FUS/TLS or TDP-43 may lead to apoptosis (or neurodegeneration) via disruption of some or all of their normal RNA processing activities in the nucleus. Systemic knockdown of TDP-43 in a mouse model induced pre-mRNA splicing defects alongside neurodegeneration in affected cells (Yang et al. 2014). Whether those pre-mRNA splicing defects contributed to neurodegenerative processes or simply coincided with them is unknown. However, disruption of either FUS/TLS or TDP-43's pre-mRNA splicing activities could plausibly lead to neurodegeneration.

Within human primary cortical neurons and mouse brains, FUS/TLS is known to splice RNA transcripts from genes associated with neurodegenerative disorders (including Microtubule-Associated Protein Tau [MAPT], Calmodulin-Dependent Protein Kinase II Alpha [CAMK2A], Fragile X Mental Retardation 1 [FMR1], and NDRG Family Member 2 [Ndrg2]) (Masuda, Takeda, Ohno 2016). Further, shRNA-mediated knockdown of TDP-43 in human neuroblastoma cells 1) led to splicing changes for genes with known roles in neuronal development and survival, and 2) increased expression of BCL-2 Interacting Mediator of Cell Death (BIM)'s most cytotoxic RNA isoform (Tollervey et al. 2011).

None of our sALS patients carried a known causal ALS coding mutation in the TARDBP or FUS/TLS genes, as reported in Chapter 2. However, it is likely our sALS patients' cervical spinal cells had TDP-43 proteinopathy, as this feature is seen in 98% of ALS patients (Lagier-Tourenne, Cleveland 2009, Donnelly, Grima, Sattler 2014, Yang et al. 2014). As TDP-43 proteinopathy is often accompanied by reduced nuclear TDP-43 levels (Yang et al. 2014), corresponding TDP-43 splicing defects may have occurred in our sALS patients' cervical spinal cells as a result.

In this chapter, we used a combination of RNA-sequencing, bioinformatics tools, and systems biology analyses to elucidate sALS group-specific exon usage differences in nuclear genes and assess their biological significance. Specifically, we set out to 1) identify cellular processes associated with nuclear genes containing exons that were statistically significantly differentially used in the sALS sample group, as they may be relevant to disease pathology.

II. Methods:

Input sample datasets: We used the same 7 sALS and 8 neurologically healthy control samples' Read 1 and Read 2 FastQ files output by Trimmomatic (as described in Chapter 2) for our downstream DEXSeq (Anders, Reyes, Huber 2012) differential exon usage analysis.

Pre-alignment steps: The DEXSeq analysis' script was originally written using several of Ensembl's annotated human reference transcriptome and genome files. Consequently, DEXSeq is known to have compatibility issues with the hg19 human reference transcriptome and genome files I obtained from the Illumina iGenomes UCSC hg19 directory as described in Chapter 2.

Prior to attempting alignment of each sample's paired end reads to the hg19 human reference transcriptome then genome, I downloaded Ensembl's hg19 human reference transcriptome and hg19 human reference genome files from ftp://ftp.ensembl.org/pub/release-63/gtf/homo_sapiens/ and ftp://ftp.ensembl.org/pub/release-63/fasta/homo_sapiens/dna/, respectively. I combined each chromosome's DNA sequence (contained in individual Fasta files) to create the hg19 human reference genome file.

Alignment: We used STAR, an open source program specially designed for RNA-Sequencing data, to attempt alignment of each sample's paired end reads to Ensembl's hg19 human reference genome. STAR used each sample's Read 1 and Read 2 FastQ files from Trimmomatic as input, and output a file containing their aligned paired end reads for downstream analysis.

DEXSeq counting process: For each annotated gene, DEXSeq first identified every unique counting bin within each of its exons via comparing known RNA transcripts' sequences (in the Ensembl hg19 human reference transcriptome file) to each annotated gene's exon sequences (in the Ensembl hg19 human reference genome file). Figure 28 illustrates this process for a fictional gene.

Ensembl's hg19 human reference transcriptome file listed some exons as belonging to RNA transcripts from multiple genes (rather than one gene) based on observations from previous studies. We decided to exclude these exons' counting bins from the DEXSeq analysis, as determining which of those gene's RNA transcripts contained that exon would be impossible in most instances given the length of our RNA-Sequencing reads. Inputting all of possible genes into the downstream over-representation analyses would likely bias our results.

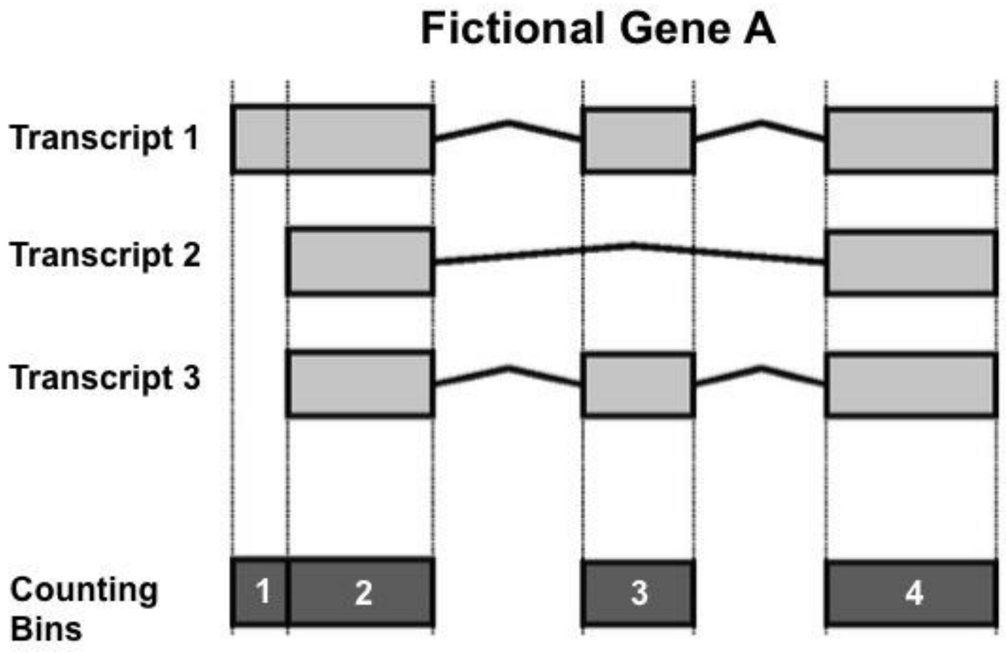


Figure 28. DEXSeq counting bin identification. This figure shows counting bins identified for a fictional gene A by DEXSeq. This fictional gene has three annotated RNA transcripts (transcribed exons are represented by light gray boxes). The majority of the first exon's sequence is contained in all three RNA transcripts, but transcript 1 has additional transcribed sequence on its 5' end relative to the other two RNA transcripts. DEXSeq thereby splits this fictional gene A's first exon into 2 counting bins. The other counting bins' boundaries correspond exactly to known boundaries for those exons in fictional gene A. Four total counting bins (dark shaded boxes) were formed for this gene. Adapted from Anders, Reyes, Huber 2012.

For each sample, DEXSeq counted the number of aligned paired end reads that fell within each eligible counting bin. Reads that overlapped several counting bins were counted for each of those bins.

DEXSeq analysis: We used DEXSeq to identify sALS group-specific DUEs. The creators of DESeq2 made DEXSeq, and DEXSeq applied many of the same mathematical procedures used in DESeq2 (and EdgeR) to appropriately model counting bins' counts across samples. These procedures ultimately reduced the number of identified DUEs that were false positives. DEXSeq relied on generalized linear models to identify sALS group-specific DUEs.

Prior to testing for sALS group-specific DUEs, DEXSeq omitted 1) any gene that only contained one counting bin (or had less than two counting bins with counts), and 2) any counting bin with an extremely low count sum across samples. It would be impossible to identify whether an annotated gene with only one counting bin (or with less than two counting bins with counts) was differentially expressed or if its individual counting bin was differentially expressed. Removing counting bins with a low count sum across samples served to reduce downstream false positive DUEs, as these counting bins had a low signal-to-noise ratio.

To further prevent false positive DUEs, DEXSeq mathematically accounted for 1) differences in the total number of counting bin counts between samples, 2) overdispersion in each counting bin's counts across samples, and 3) whether each counting bin's corresponding annotated gene was differentially expressed between groups.

For a given counting bin, differences in the number of counts between samples could reflect each sample’s total number of RNA-Sequencing reads as opposed to differences in that counting bin’s expression levels between samples. Assume two of our samples’ RNA-Sequencing libraries had an equal number of sequenceable dscDNA molecules corresponding to a given counting bin (suggesting that counting bin had an equal expression level in both samples’ postmortem cervical spinal sections). If more total sequencing reads were generated for one of those samples, that sample’s sequencing data would likely have more total paired end reads corresponding to that counting bin relative to the other sample. This is because denatured strands from that counting bin’s corresponding sequenceable dscDNA molecules were appropriated more Illumina NextSeq500 binding spots (covalently bound oligonucleotides) to potentially hybridize to for sequencing.

Like paired end read counts, counting bin counts across samples are often overdispersed. This means the variance of counting bin counts across each group’s samples often exceeds what is expected using a Poisson distribution. DEXSeq’s DUE analysis relied on negative binomial distributions to best account for variance in each counting bin’s counts across each group’s samples. DEXSeq also estimated the level of dispersion for each individual counting bin (using maximum likelihood modeling) and counting bins with similar expression levels across samples. DEXSeq then applied an empirical Bayes’ theorem to moderate overdispersion via shrinking each counting bin’s level of dispersion towards the level of dispersion estimated for counting bins with a similar expression level.

DEXSeq assessed whether each counting bin’s corresponding annotated gene was differentially expressed between our sALS and neurologically healthy control groups.

If a given annotated gene was differentially expressed between groups, the authors reasoned each of that annotated gene’s counting bins was likely to be differentially used between groups by a similar quantitative factor. In cases where a given annotated gene was differentially expressed, that quantitative factor was accounted for when testing whether each of that annotated gene’s counting bins was differentially used between groups.

For each counting bin, two generalized linear models (full and reduced) were generated. Each of these generalized linear models incorporated the above factors and a log fold change value calculated via comparing the counting bin’s representative count values in our sALS and neurologically healthy control groups. The full generalized linear model included a variable that estimated how much of the difference in each counting bin’s estimated usage levels between groups was explained by group membership (case vs. control status), whereas the reduced generalized linear model omitted that variable.

To assess whether a given counting bin was differentially used between groups, DEXSeq compared the fits of these two generalized linear models using a likelihood ratio test. For each counting bin, this comparison produced a corresponding p-value. We calculated all counting bins’ Benjamini-Hochberg corrected p-values using DEXSeq’s reported p-values via the R function p.adjust. Each counting bin with a Benjamini-Hochberg corrected p-value < 0.10 was identified as a sALS group-specific DUE. Prior to uploading our list of annotated genes containing a DUE to IPA, we converted their Ensembl ID’s to HGNC ID’s using Biomart’s ID conversion tool.

Ingenuity Pathway Analysis: We used QIAGEN's Ingenuity Pathway Analysis to assess what canonical signaling pathways, diseases/disorders, and cellular functions were statistically significantly associated with our 46 annotated genes containing one or more DUEs. IPA used a right-tailed Fisher Exact test to assess the number of these annotated genes that were separately associated with each canonical signaling pathway, disease/disorder, and cellular functions in the IPA Knowledge base. Corresponding association p-values relating these annotated genes carrying one or more DUEs to each tested canonical signaling pathway, disease/disorder, and cellular function were reported.

III. Results:

DEXSeq DUE testing and associated cellular processes: At an FDR of .10, DEXSeq identified 52 sALS group-specific DUEs (from 46 annotated genes). These results are shown in Table 9. A visual schematic of a sALS group-specific DUE can be seen in Figure 29. Among various findings, QIAGEN's Ingenuity Pathway Analysis revealed these DUEs' annotated genes were associated with cholesterol biosynthesis, the mevalonate pathway, and lipid metabolism. These results are shown in Table 10.

Table 9: sALS group-specific DUEs

Annotated Gene	sALS group-specific DUE	Fold Change
LINC00476	CB007	-2.32
SEMA4G	CB044	-1.73
SLCO2B1	CB019	-1.69
CTNND2	CB048	-1.53
FCGR2A	CB024	-1.52
SUSD1	CB013	-1.26
CD27-AS1	CB011	-1.15
PCNT	CB017	-1.15
SLC25A36	CB021	-1.12
LRRTM4	CB006	-1.09
GABPA	CB005	-1.08
HAPLN4	CB001	-1.08
MORC3	CB018	-1.08
MORC3	CB017	-1.07
CCDC91	CB027	-1.06
GRAMD3	CB020	-1.05
CDK17	CB018	-1.05
BARD1	CB018	-1.05
CHCHD2	CB001	-1.05
LYPLA1	CB008	-1.05
MKLN1	CB016	-1.05
DDR1	CB129	-1.04
ENTPD4	CB003	1.04
FAM219B	CB005	1.09
POLI	CB012	1.17
RNF170	CB004	1.17
VEZT	CB086	1.22
MYLK	CB044	1.24
HSD17B7	CB016	1.27
FDFT1	CB025	1.29
RSRC2	CB046	1.30
TF	CB004	1.31
TF	CB003	1.34
ATRNL1	CB048	1.34
ATAD5	CB021	1.35
HMGCS1	CB025	1.38
RSRC2	CB047	1.50
CCDC141	CB004	1.51
RICTOR	CB002	1.51
KALRN	CB071	1.52
USH2A	CB001	1.52
THAP9-AS1	CB009	1.54
NBEAL2	CB083	1.55
IDI1	CB013	1.56

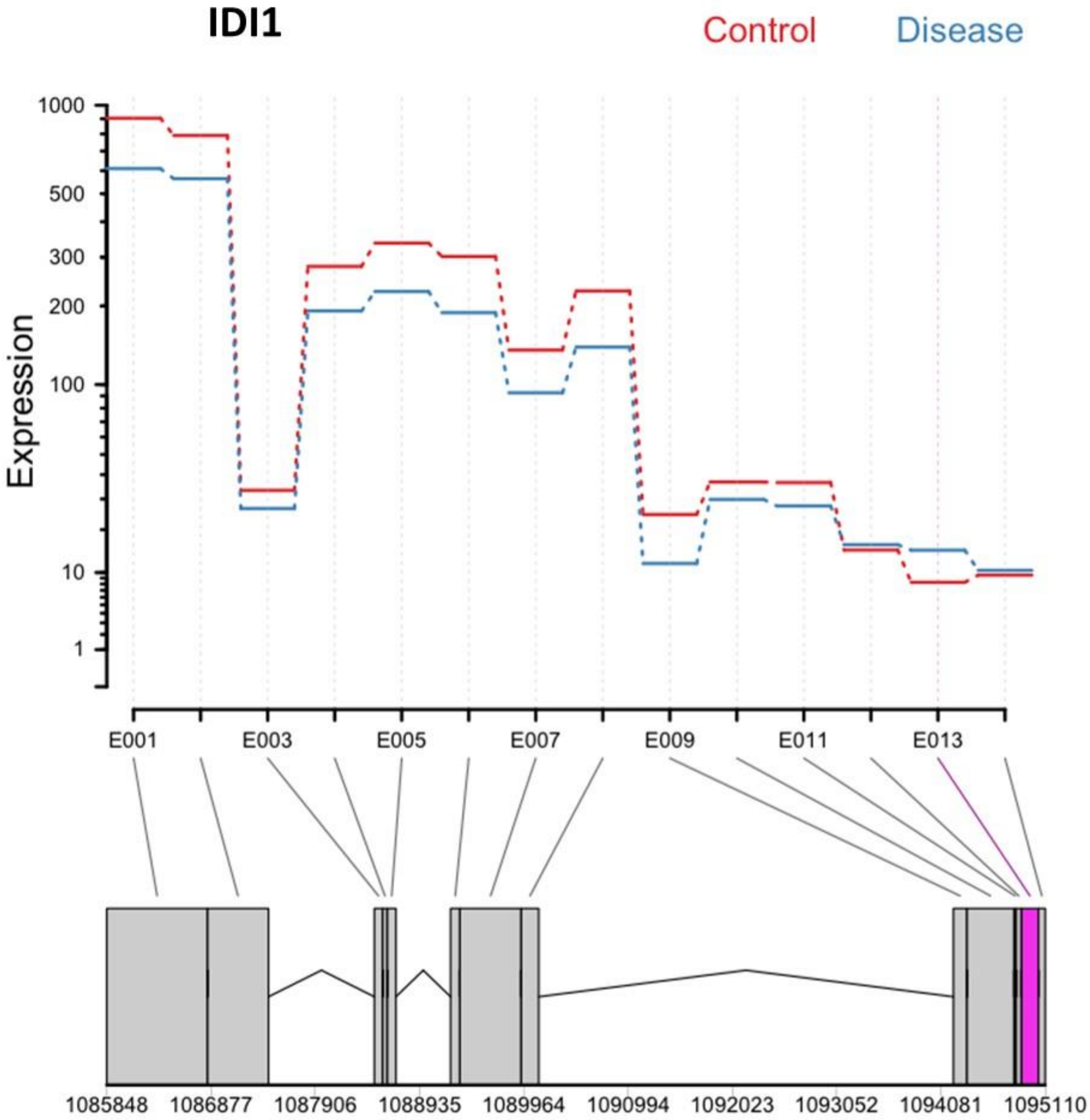


Figure 29. DEXSeq differentially used exon plot. This figure shows the difference in exon usage for counting bin #13 of the IDI1 locus. This counting bin was selected as an illustrative example, as its usage level had the largest positive fold change in our sALS sample group relative to our neurologically healthy control sample group.

Table 10: IPA results for sALS group-specific DUEs

Top Canonical Pathways	**Overlapping Genes**	**p-Value**
Superpathway of Cholesterol Biosynthesis	4/28	2.40E-07
Cholesterol Biosynthesis I	2/13	2.82E-04
Mevalonate Pathway I	2/13	2.82E-04
Cholesterol Biosynthesis II	2/13	2.82E-04
Cholesterol Biosynthesis III	2/13	2.82E-04
Top Diseases and Disorders		**p-Value**
Cardiovascular Disease		4.48E-02 - 1.16E-03
Hematological Disease		3.44E-02 - 1.16E-03
Cancer		4.81E-02 - 1.94E-03
Connective Tissue Disorders		4.80E-02 - 1.94E-03
Developmental Disorder		4.74E-02 - 1.94E-03
Top Upstream Regulators	**Overlapping Genes**	**p-Value**
FOXO4	4/26	1.23E-06
SREBF2	4/37	8.33E-06
Pitavastatin	3/13	9.10E-06
NPPB	3/20	2.37E-05
SREBF1	5/92	2.38E-05

IV. Discussion:

Aberrant RNA processing has been recurrently associated with ALS-group specific DEGs identified in tissues from both human patients (fALS and sALS) and fALS rodents (Heath, Kirby, Shaw 2013). This may reflect a disruption of TDP-43's normal RNA processing activities as a result of reduced nuclear TDP-43 protein in ALS tissues. Reduced nuclear TDP-43 has been observed in cells with TDP-43 proteinopathy, a feature found in ~98% of ALS patients' (both fALS and sALS) spinal motor neurons, spinal glial cells, and/or select brain cells (Lagier-Tourenne, Cleveland 2009, Donnelly, Grima, Sattler 2014, Yang et al. 2014).

Reduced nuclear TDP-43 levels promoted neurodegeneration in various animal models (Schmid et al. 2003, Yang et al. 2014). Further, one of those studies demonstrated reduced nuclear TDP-43 caused TDP-43 splicing defects alongside neurodegeneration (Yang et al. 2014). It is unclear whether TDP-43 splicing defects functionally contribute to neurodegenerative processes. However, it seems plausible as TDP-43 regulates pre-mRNA splicing for genes that 1) encode synaptic proteins (Polymenidou et al. 2011), 2) preserve neuronal integrity (Lagier-Tourenne et al. 2012), and 3) promote neuronal survival (Tollervey et al. 2011).

To date, two ALS studies have used splicing-sensitive microarrays to identify statistically significant splicing differences in spinal motor neurons from sALS patients and neurologically healthy controls (Rabin et al. 2010, Highley et al. 2014). In both studies, researchers confirmed their sALS patients' spinal motor neurons had TDP-43 proteinopathy. The earlier study identified 411 aberrantly spliced genes, and over-representation analyses revealed those genes were associated with cell adhesion, transmembrane receptor protein tyrosine kinase activity, and the extracellular matrix (Rabin et al. 2010). The later study identified 6,449 sALS group-specific DUEs (in 4,311 genes), and over-representation analyses revealed those genes were associated with ribonucleotide binding, cytoskeletal organization, protein localization, and macromolecule catabolic processes (Highley et al. 2014).

In this chapter, we combined deep RNA-Sequencing and systems biology analyses to identify sALS group-specific DUEs in postmortem spinal tissues from sALS patients and neurologically healthy controls. We discovered 52 DUEs in 46 annotated genes. 15 of these DUEs' fold change usage difference between groups was <10%. It is difficult to envision these differences playing a significant role in disease pathology. Nonetheless, over-representation analyses revealed the 46 genes containing DUEs were associated with cholesterol biosynthesis, the mevalonate pathway, and lipid metabolism.

Cholesterol is an essential component of neuronal membranes, and is needed to form membrane lipid rafts necessary for protein anchorage and trafficking. Cholesterol is also used for continued axon growth and synapse remodeling in the mature adult brain, and serves as a precursor of neurosteroids (Anchisi, Dessi, Pani, Mandas 2013). Cholesterol biosynthesis involves a multi-step process beginning with conversion of acetyl-CoA to 3-hydroxy-3-methylglutaryl-CoA, followed by the generation of mevalonate (Martin, Pfrieger, Dotti 2014). After conversion into many other intermediary substrates, cholesterol is produced.

Defects in cholesterol biosynthesis are known to cause several rare neurodegenerative disorders. These include lathosterolosis, desmosterolosis, cerebrotendinous xanthomatosis, congenital hemidysplasia with ichthyosiform erythroderma and limb defects, and Smith-Lemli-Opitz Syndrome (SLOS) (Vance 2012). Dysregulated cholesterol homeostasis has separately been associated with more common neurodegenerative disorders (including Alzheimer's disease, Huntington's disease, and Parkinson's disease). If and how defects in cholesterol biosynthesis or homeostasis functionally contribute to these disorders' neurodegenerative processes remains largely unknown. However, we do know SLOS causal mutations in the 7-Dehyrocholesterol reductase (DHCR7) gene lead to abnormally low levels of cholesterol (and high levels of 7-dehydrocholesterol) in cells, plasma, and the brain (Vance 2012).

Our best understanding of how dysregulated cholesterol metabolism can functionally contribute to neurodegenerative processes comes from studies of Niemann-Pick Type C (NPC), a neurodegenerative disorder affecting 1 in 150,000 people. In NPC, mutant NPC1 and NPC2 proteins fail to transport free cholesterol from lipoproteins into neuronal cells' cytosols, leaving the cholesterol sequestered in late endosomes and/or lysosomes. This results in disproportionately low levels of cholesterol in affected neurons' plasma membranes and axons (Vance 2012).

Interestingly, subcutaneous injection of a cholesterol binding compound (cyclodextrin) into NPC -/- mice slowed neurodegeneration and extended their lifespan by 50% (Liu et al. 2009). A separate analysis showed administrating a low dose of cyclodextrin to NPC -/- neurons released sequestered cholesterol from their late endosomes and/or lysosomes into their cytoplasms (Vance 2012). Researchers have proposed cyclodextrin may stop neurodegeneration in NPC mouse models by re-distributing sequestered cholesterol from affected neurons' late endosomes and/or lysosomes into their plasma membranes (where it was low in concentration) (Martin, Pfreiger, Dotti 2014).

Much less is known about how cholesterol biosynthesis, cholesterol homeostasis, or the distribution of cholesterol in neuronal cells may contribute to sALS pathology. A very early study (Cutler et al. 2002) reported elevated levels of sphingomyelin, ceramides, and cholesterol esters were found in the spinal cords of ALS patients and fALS rodents carrying a SOD1 mutation. Further, they speculated this could promote neurodegeneration via oxidative stress related apoptotic events. Several recent studies have shown ALS patients with higher circulating cholesterol levels (characterized by elevated LDL to HDL ratios) live longer than patients with lower cholesterol levels (D’Amico, Factor-Litvak, Santella, Mitsumoto 2013). It is tempting to think this protective effect is the result of higher cholesterol concentrations in those neurons’ plasma membranes, especially considering low cholesterol levels were linked to neurodegneration in SLOS and NPC. However, the CNS blocks entry of cholesterol-rich lipoproteins circulating in the blood via the blood brain barrier (BBB). Therefore, it is unclear how a higher circulating cholesterol level would protect disease-vulnerable cells in the spinal cord (Martin, Pfreiger, Dotti 2014).

We cannot predict whether cholesterol biosynthesis was increased or decreased in our sALS samples based on our sALS group-specific DUEs alone. Further, it is unclear what significance these findings have for motor neurons, as the majority of RNA isolated from these spinal sections came from astrocytes and microglia. However, the link between cholesterol biosynthesis and neurodegeneration has been established previously. Further, lower levels of cholesterol (found systemically or within neuronal plasma membranes and axons) have been linked to neurodegeneration in SLOS and NPC. For these reasons, I propose future experiments that first generate iPSC-derived astrocytes, iPSC-derived microglia, and iPSC-derived motor neurons using mononuclear blood cells drawn from sALS patients and neurologically healthy controls. I would co-culture these cells using 3-dimensional scaffolding techniques *in vitro,* as described in this paper (Schwartz et al. 2015).

I would then directly compare both membrane bound and free cholesterol levels in iPSC-derived motor neurons from sALS patients and neurologically healthy controls using immunohistochemistry and a colorimetric assay, respectively. My immunohistochemistry experiments would involving staining for Filipin, a highly fluorescent compound that specifically binds to cholesterol. Filipin staining is used to diagnosis NPC in clinical settings by assessing the level of cholesterol found sequestered in late endosomes and/or lysosomes and in the plasma membrane (Vanier, Latour 2015). I would also lyse equal amounts of each 3D culture prior to measuring total cholesterol levels (reflecting cholesterol biosynthesis) using a commercially available colorimetric assay. I would be most interested in determining 1) whether iPSC-derived motor neurons from sALS patients (relative to neurologically healthy controls) had significantly different levels of cholesterol staining in their late endosomes or cell membranes, and 2) whether 3D neuronal models from sALS patients (relative to neurologically healthy controls) had significantly different levels of free cholesterol.

If the sALS group did have significant differences in either measurement, I would assess whether those differences were seen alongside increased rates of motor neuron death as measuring using a cell viability MTT assay. If that relationship was established, I would then begin testing whether different pharmacological agents 1) normalized these sALS-group specific differences in cholesterol distribution in the cell and/or cholesterol biosynthesis, and 2) reduced motor neuron death as measured using the MTT assay.

CHAPTER 5: Summary, What I learned, and future directions

ALS is a disease characterized by degeneration of upper and lower motor neurons in the brain stem and spinal cord. Average life expectancy after diagnosis is between 2-5 years, as the death of motor neurons innervating the lungs ultimately leads to many sALS patients' deaths. Unfortunately, current treatments only extend life by several months. More effective therapies are sorely needed for this devastating illness.

As of 2014, 68% of fALS in Caucasians was accounted for by causal mutations in 9 different genes. However, only 11% of sALS in Caucasians was accounted by mutations in these genes (Renton, Chio, Traynor 2014). Despite evidence for varying genetic etiologies, ALS gene expression studies dating back to 2001 suggest there is a convergent set of perturbed cellular processes germane to both fALS and sALS. ALS tissue-specific DEGs identified from human (fALS and sALS) patient and fALS model rodent samples were associated with oxidative stress, mitochondrial dysfunction, apoptosis, inflammation, RNA processing defects, and protein aggregation (Heath, Kirby, Shaw 2013). Researchers using separate molecular biology assays found these same cellular processes were perturbed in human (fALS and sALS) patient and fALS rodent tissues, as referenced in the introduction of Chapter 2. Taken together, it is likely perturbations in these cellular processes contribute to ALS onset, progression, and symptoms. These findings also support the use of gene expression studies to identify cellular processes likely perturbed in ALS pathology moving forward.

Follow-up molecular biology experiments can be used to test how sALS group-specific differences (identified in gene expression studies) may have functional relevance to disease pathology. Further, these experiments could unveil novel therapeutic targets that may slow this devastating disorder. In the course of this dissertation project, we combined RNA-Sequencing, systems biology analyses, and molecular biology assays to elucidate sALS group-specific differences in postmortem spinal tissues that may be relevant to disease pathology.

In chapter 2, we discovered inflammatory processes and TNF-α signaling were statistically significantly associated with sALS group-specific gene expression differences identified using independent exploratory DEG analyses and an unsupervised gene co-expression network analysis. Increased inflammatory processes and elevated TNF-α signaling have been recurrently reported in ALS tissues, and likely play a role in sALS pathology. We selected TNFAIP2, an upregulated sALS group-specific DEG and network hub gene, for downstream molecular assays. Elevated TNF-α signaling is known to increase TNFAIP2 expression in a variety of cell types (Saito et al. 2013, Zhou, Scoggin, Gaynor, Williams 2003, Tian et al 2005), and elevated TNFAIP2 expression has previously been associated with increased apoptosis (Park et al. 2003, Rusiniak et al. 2000, Ma et al. 2003). Studies linking elevated TNFAIP2 expression to increased apoptosis did not assess TNFAIP2's cellular function or whether TNFAIP2 functionally promoted apoptotic processes directly.

Within *in vitro* models of sALS disease-vulnerable cell types, we wished to test whether 1) elevated TNF-α signaling increased TNFAIP2 expression, and 2) whether transient overexpression of TNFAIP2 increased cell death, potentially through the TNF superfamily mitochondrial apoptotic pathway. We discovered exposing neural stem cells to extracellular TNF-α increased TNFAIP2 gene expression by ~100 fold, along with elevating several other network hub genes' expression levels. Transient overexpression of TNFAIP2 decreased neural stem cell viability, and simultaneous inhibition of activated caspase 9 (a protein necessary for TNF-α superfamily mitochondrial-mediated apoptosis) reversed this effect in these cells. Further, transient overexpression of TNFAIP2 with Green Fluorescent Protein (GFP) fused to its N terminus in iPSC-derived motor neurons led to increased cell death evidenced by decreased cell viability and increased caspase 3/7 levels.

These proof of concept experiments within *in vitro* models of sALS disease-vulnerable cell types demonstrated TNFAIP2 expression increased in response to elevated TNF-α signaling, and transient overexpression of TNFAIP2 promoted cell death. TNFAIP2 may mediate cell death via the TNF-α superfamily mitochondrial-mediated apoptotic pathway, as inhibition of activated Caspase 9 (a protein necessary for TNF-α superfamily mitochondrial-mediated apoptosis) prevented TNFAIP2-mediated cell death in our neural stem cells. Taken together, these findings support our hypotheses that elevated TNF-α signaling 1) increased TNFAIP2 expression in our sALS patients' cervical spinal cells, and 2) TNFAIP2 functionally contributed to spinal motor neuron death in our sALS patients via the TNF mitochondrial apoptotic pathway.

Further research into the potential role of elevated TNF-α signaling (and TNFAIP2 expression) in sALS pathology is needed.

In chapter 3, we did not identify any statistically significant differences in mitochondrial gene expression in our sALS sample group vs. our neurologically healthy controls. However, there were trend level findings of reduced gene expression for the majority of mitochondrial protein coding genes (12 out of 13) in our sALS patients vs. our neurologically healthy controls. These genes encode components of ETC protein complexes, essential for mitochondrial OXPHOS. Defective mitochondrial OXPHOS has been linked to elevated oxidative stress, a phenomenon known to mediate neurodegeneration under certain circumstances. While we did not identify sALS group-specific mitochondrial DEGs, it is possible our sALS patients carried mitochondrial SNVs or indels that perturbed mitochondrial OXPHOS in their disease-vulnerable cells. This may have contributed to their motor neurons dying. Further investigation into the potential role of pathogenic mitochondrial variants in our samples (and in sALS pathology) is needed.

In chapter 4, we identified 52 sALS group-specific DUEs in 46 annotated genes, and over-representation analyses revealed those genes were associated with cholesterol biosynthesis. Defects in cholesterol biosynthesis and metabolism have been linked to numerous neurodegenerative disorders. In NPC, cholesterol sequestration in the late endosomes and/or lysosomes leads to lower levels of cholesterol in neurons' membranes. While the exact mechanism is unknown, this ultimately promotes neurodegeneration. While little is known about whether (or how) defects in cholesterol biosynthesis and/or metabolism play a role in ALS, cholesterol has been linked to neurodegeneration repeatedly in previous literature.

It is not difficult to imagine how perturbations in the level or distribution of cholesterol in neurons could lead to pathological cellular processes. Cholesterol is an essential component of neuronal membranes, necessary for protein anchorage and trafficking. Further, cholesterol is used for continued axon growth and synapse remodeling in the mature adult brain, and serves as a precursor of neurosteroids (Anchisi, Dessi, Pani, Mandas 2013). Research into the potential role of aberrant cholesterol biosynthesis and/or metabolism in sALS pathology is needed.

What I learned:

In the course of this dissertation project, I have learned how many factors inherent to current RNA-Sequencing workflows influence gene expression estimates. Many of these factors are completely unrelated to gene expression levels in the original biological tissue. I will discuss the numerous factors in our workflow that likely influenced our annotated gene expression estimates.

For review, the steps of our Illumina RNA-Sequencing workflow for each sample involved 1) isolating total RNA, 2) removing rRNAs from total RNA, 3) fragmenting isolated RNA transcripts into a distribution of smaller fragments, 4) converting RNA fragments into sequenceable dscDNA molecules, 5) increasing the proportion of properly ligated sequenceable dscDNA molecules via enrichment PCR, 6) binding denatured strands of sequenceable dscDNA molecules to an Illumina NextSeq500 flowcell for generation of paired end reads, 7) aligning paired end reads to the hg19 human reference transcriptome then genome, and 8) using bioinformatics analyses to generate each annotated gene’s expression level estimates. Nearly every one of these steps influences the final gene expression level estimates!

The QIAGEN miRNeasy kit used in step 1 did not recover all RNA transcripts, as some were inevitably retained on its isolation column. Use of the RNeasy Micro kit for sample purification led to our losing various genes’ RNA transcripts <100 nucleotides in length, as that kit is designed to recover RNA transcripts >100 nucleotides in length. Finally, RNA transcripts are known to undergo varying levels of degradation during the RNA isolation process. For any given annotated gene, losing its RNA transcripts would lead to less 1) corresponding sequenceable dscDNA molecules and 2) paired end reads that could have aligned to that gene’s transcribed regions. This would lead to an underestimated gene expression level estimate for that annotated gene relative to its actual expression level in the biological sample.

Paired end reads corresponding to RNA transcripts that incurred higher levels of degradation likely preserved shorter stretches (less nucleotides) of those RNA transcripts relative to RNA transcripts that incurred lower levels of degradation. Paired end reads preserving longer stretches of their corresponding RNA transcripts were generally more likely to accurately align to the hg19 human reference genome relative to paired end reads preserving shorter stretches of their corresponding RNA molecules. While the majority of our RNA transcripts were not likely to be significantly degraded (as we used RNA with an RQI score >7), this effect likely led to underestimated gene expression values for any annotated gene that had significantly degraded RNA transcripts.

The RNA fragmentation process used in step 3 broke all RNA transcripts into a tighter distribution of smaller fragments (between 120-210 nucleotides) prior to dscDNA generation in the next step. The length of a given isolated RNA transcript was positively correlated with both 1) the number of fragments generated from it, and 2) the number of corresponding sequenceable dscDNA molecules generated from its fragments in downstream steps. Relative to shorter RNA transcripts, longer RNA transcripts were generally more likely to have a greater number of corresponding paired end reads (and paired end read counts). This concept is illustrated in Figure 16, and was true even when a long and short RNA transcript had equal expression levels in the biological sample.

As a result of this technical bias, it is highly probable that our annotated genes had disproportionately more or less aligned paired end reads relative to their actual expression levels in each sample’s postmortem tissues. This almost certainly influenced our annotated genes’ expression level estimates, as both HTSeq-Count and Cufflinks used the number of aligned paired end read counts to calculate their gene expression level estimates. While Cufflinks applied a normalize step to correct for this bias, it is unlikely to have completely resolved its effects as discussed later.

The enrichment PCR process in step 5 was necessary to increase the number of sequenceable dscDNA molecules relative to dscDNA molecules improperly ligated with one or zero sequencing adaptors. This ensured a negligible amount of unsequenceable denatured strands (from improperly ligated dscDNA molecules) bound to the NextSeq 500 flowcell. Unfortunately, this necessary step likely altered the amount of input sequenceable dscDNA molecules relative to each other. A previous study has shown DNA fragments ligated with Illumina sequencing adaptors used in Illumina library preparation kits do not amplify equally (Kebschull and Zador, 2015). Depending on whether a given annotated gene's sequenceable dscDNA molecules were disproportionately increased or decreased by enrichment PCR, that annotated gene would have more or less corresponding denatured strands able to hybridize the limited number of Illumina NextSeq 500 binding spots (covalently bound oligonucleotides) for sequencing. This technical property likely led to overestimated or underestimated gene expression values for affected annotated genes.

Use of paired end reads that were >100 nucleotides in length likely increased our total number of aligned sequenced reads relative to if we used single end (or shorter paired end) reads (Salzman, Jiang, Wong 2011, Cho et al. 2014). This likely prevented underestimated gene expression values for at least a portion of our annotated genes, as a greater number of their corresponding paired end reads likely aligned to their transcribed regions.

Perhaps the biggest influence on our downstream gene expression estimates was the number of paired end reads generated for each sample. The total number of sequenced reads obtained for a given sample is positively correlated with 1) the likelihood genes with lower expression levels are represented in their sequencing data, and 2) how accurate their gene expression estimates are likely to be. These concepts are illustrated in Figure 12.

According to a recent study, 45-65 million RNA-sequencing reads generated from input total RNA (depleted of rRNAs) offers a detection level for lowly expressed protein coding genes that is comparable to a standard Agilent microarray (Zhao et al. 2014). If one uses input polyA+ RNA instead, ~13 million reads allows a detection level for lowly expressed protein coding genes that is comparable to a standard Agilent microarray. There is not a consensus number of RNA-sequencing reads one should appropriate to each sample to achieve highly accurate gene expression estimates for very lowly expressed genes. An early study proposed 200 million reads per sample was necessary to detect the full range of expressed human RNA transcripts, including those from very lowly expressed genes (Tarazona et al. 2011). The ENCODE consortium reported ~36 and ~80 million paired end reads per sample were necessary to accurately estimate genes with expression levels corresponding to FPKM values of >10 and <10, respectively.

We obtained >55 million paired end reads for each of our samples. Taken together, we should have a comparable detection level for protein coding genes relative to using standard microarray technology. According to ENCODE's estimates, our gene expression estimates for annotated genes with expression levels that correspond to an FPKM value of >10 should be highly accurate. Our gene expression estimates for annotated genes with expression levels that correspond to an FPKM value of <10 are likely less accurate (with decreasing accuracy tracking with lower gene expression levels).

Use of Tophat2 likely had mixed effects on our downstream gene expression estimates. A group recently compared alignment results from 26 mapping protocols on 4 common RNA-Sequencing read datasets (Engstrom et al. 2013). Tophat2 reported one the smaller numbers of total aligned reads due to its low tolerance for mismatching nucleotides, but one of the higher numbers of identified splice sites when used with a guide annotation. Overall, use of Tophat2 probably led to more underestimated gene expression values across annotated genes as a result of fewer paired end reads that successfully aligned to those genes' transcribed regions. However, use of Tophat2 probably prevented underestimated gene expression values for some annotated genes by using additional identified splice sites to correctly map paired end reads to those genes' transcribed regions across introns.

Finally, properties inherent to the bioinformatics tools used to calculate each sample's annotated gene expression estimates were also likely to affect the accuracy of our gene expression estimates. For each sample, HTSeq-Count reported each annotated gene's total number of paired end reads that aligned to its transcribed regions without accounting for the influences of transcript length or the sample's total number of paired end reads on these counts. This likely led to overestimated and underestimated gene expression values for many of our annotated genes depending on how they were influenced by these technical factors. Cufflinks' annotated gene expression estimates were likely more accurate for each sample, as their FPKM normalization accounted for transcript length, that sample's total number of sequenced reads, and RNA composition biases. However, this is likely an imperfect solution. It is probable a portion of each sample's endogenously smaller RNA transcripts didn't have any corresponding paired end reads generated (despite equal expression levels to larger RNA transcripts) as a result of having less corresponding denatured strands to bind the flowcell for sequencing. Cufflinks can't normalize aligned paired end read counts for a given annotated gene if that gene's endogenously smaller RNA transcripts didn't have any corresponding paired end reads.

Provided a thoughtful experimental design that buffers the influences of these various technical factors on gene expression estimates, RNA-Sequencing is an excellent research tool for estimating gene expression levels and performing downstream analyses. With technological advances, RNA-Sequencing's ability to 1) accurately estimate gene expression levels, 2) detect and estimate lowly expressed genes' expression levels, and 3) identify various RNA isoforms will only improve.

Commonly used RNA-Sequencing workflows for DEG analyses rely on sequenceable dscDNA molecules made from fragmented RNA transcripts. Pacific Biosystem's current RNA-sequencing workflow does not fragment RNA transcripts, and its sequencing technology routinely generates sequencing reads that are (on average) 15,000 nucleotides. Generation of each sample's RNA-Sequencing reads using this platform would greatly reduce (if not altogether eliminate) the influence of transcript length on downstream gene expression estimates, and mitigate Cufflinks' uncertainty in assigning a given sequenced read to its corresponding RNA transcript. Further, gene expression estimates would likely be much more accurate as a result, and widespread use of this platform would produce more trustworthy differential transcript expression analyses.

However, Pacific Biosystem's RNA-Sequencing workflow is rarely used in preparation for DEG analyses today. The average cost per read is drastically higher compared to platforms that generate shorter reads. Typically, gene expression studies with DEG analyses involve generation of >10 million reads per sample (when using polyA+ enrichment input mRNAs) or >30 million reads per sample (when using rRNA-depleted total RNA). It would be astronomically expensive to generate these numbers of reads per sample using the Pacific Biosystem RNA-Sequencing workflow. However, sequencing costs have decreased rapidly in the last decade. It is possible this technology will become more affordable and replace the current workflows.

A potentially superior approach to Pacific Biosystem's RNA-Sequencing workflow in the future would involve generating long sequencing reads (covering the entirety of the RNA transcript) directly from every RNA transcript in each tested sample. That would eliminate influences on gene expression estimates introduced by 1) transcript length, 2) PCR amplification biases, 3) RNA transcripts lost across the library preparation workflow, and 4) the effects of a sampling procedure for sequencing. Numerous companies (including Oxford Nanopore Technologies) are in the process of creating a technology that can generate long sequencing reads (covering the entirety of the RNA transcript) from every RNA transcript in each tested sample.

Future Directions:

Clinical trials that aim to inhibit TNF-α synthesis may prove fruitful in sALS as they have in other chronic diseases with an inflammatory component (Probert 2015). However, the use of non-selective TNF-α inhibitors has exacerbated symptoms in human multiple sclerosis patients, and induced new cases of demyelinating disease and neuropathies in other clinical populations. Side effects have included an elevated risk for bacterial sepsis and invasive fungal infections as a result of suppressed immune function (Probert 2015).

Additionally, elevated TNF-α signaling can lead to increased cell survival processes via NFKB signaling under certain biological conditions, so inhibiting it altogether may prevent those protective processes from occurring in sALS patients' disease-vulnerable cells. For these reasons, inhibition of target proteins in the TNF superfamily apoptotic pathways may prove to be a superior treatment strategy for sALS patients if therapeutic efforts to generally inhibit TNF-α synthesis have significant shortcomings.

In our studies presented in Chapter 2, we found exposing neural stem cells to extracellular TNF-α increased TNFAIP2 gene expression by ~100 fold. We demonstrated transient overexpression of TNFAIP2 promoted cell death within *in vitro* models of sALS disease-vulnerable cell types. Further, we discovered inhibiting caspase 9 (a protein necessary for TNF-α superfamily mitochondrial-mediated apoptosis) decreased TNFAIP2-mediated cell death in our neural stem cells. Taken together, these findings suggest TNFAIP2 may functionally contribute to TNF superfamily mitochondrial-mediated apoptosis in disease-vulnerable cell types in sALS patients.

For follow up experiments, I would like to assess whether TNFAIP2 functionally contributes to apoptosis within iPSC-derived motor neurons *in vitro* and within transgenic mice *in vivo.* The results of these experiments would have important implications for TNFAIP2's potential therapeutic relevance in sALS, as it may represent a good candidate for therapeutic targeting if it does functionally contribute to apoptosis in one or both models.

In preparation for the *in vitro* experiment, I would use CRISPR/Cas9 technology to establish an iPSC-derived motor neuron line without a functional copy of the TNFAIP2 gene (TNFAIP2 -/-). I would then empirically determine what amount of extracellular TNF-α reduces cell viability in iPSC-derived motor neurons. I would then expose wildtype and TNFAIP2 -/- iPSC-derived motor neuron groups to that pre-determined amount of TNF-α for 6 hours, followed by assessing cell viability in both groups using the MTT assay. If wildtype iPSC-derived motor neurons showed a statistically significant reduction in cell viability compared to TNFAIP2 -/- iPSC-derived motor neurons, this would suggest TNFAIP2 functionally contributes to apoptosis mediated by elevated TNF-α signaling.

In preparation for the *in vivo* experiment, I would use CRISPR/Cas9 technology to insert a doxycycline-inducible human TNFAIP2 transgene (coupled to a motor neuron specific promoter) into various transgenic mouse lines (to assess against different genomic backgrounds). I would then raise these mice to adulthood, and administer varying doses of doxycycline to these mice and their littermates (who would serve as controls). I would then assess whether varying levels of TNFAIP2 overexpression led to a neurodegenerative phenotype characterized by muscle weakness, paralysis, muscle wasting, and early death. If it does, I would confirm TNFAIP2 overexpression occurred in spinal motor neurons, and that death of upper and lower motor neurons was observed.

These findings (along with those presented in data chapter 2) could provide the bases for a future clinical trial assessing whether TNFAIP2 knockdown in the CNS ameliorates symptoms related to motor neuron death in sALS patients. Clinical metrics to assess disease progression, such as the revised ALS functional rating scale used in previous studies (Tateishi 2010), could be a preliminary measure of treatment efficacy.

Provided an unlimited budget, I would execute a massive ALS research project involving the enrollment of 20,000 sALS patients and 20,000 healthy controls (without ALS or any other neurological disorder) that are matched for age, ethnicity, and gender.

To better understand the different molecular etiologies of sALS, I would perform multiple experiments that require collection of DNA samples from sALS patients and neurologically healthy controls. To avoid an invasive procedure, I would collect blood samples from sALS patients and neurologically healthy controls upon entry into our study. I would also collect blood samples from each sALS patient’s parents and an unaffected sibling whenever available. For healthy controls, I would contact them every 6 months (for 20 years) to ensure they did not develop a neurological disorder. If they did, I’d remove them from our study.

A recent publication (Kiezun et al. 2012) suggested 1) most of the rare coding variants in the human population are deleterious, 2) increased sequencing sample sizes are positively correlated with the number of rare coding variants identified, and 3) >10,000 cases and controls are likely necessary for sufficient statistical power to identify a single gene harboring an excess number of rare coding variants in a gene burden test. These researchers calculated sample sizes of 10,000 would be needed via simulations that modeled typical numbers of rare variants identified in disease vs. control groups.

For my first experiment, I would aim to identify genes carrying a statistically significant excess of rare (minor allele frequency <1%) protein-coding variants in sALS patients vs. neurologically healthy controls. Variants in these genes could contribute to (or cause) sALS. I would first extract DNA from blood mononuclear cells for all 20,000 sALS patients and 20,000 healthy controls. I would then generate, align, and process whole exome sequencing data for all 40,000 samples using appropriate kits, computing resources, and bioinformatics software programs.

I would next perform a gene burden test to identify any gene in the hg38 human genome that harbors a statistically significant excess of rare coding variants in our sALS patients vs. our neurologically healthy controls. For each annotated gene, the gene burden test compares the total number of rare coding variants in our sALS patients vs. our neurologically healthy controls' exome sequencing data. A corresponding p-value is generated, and a Bonferroni correction is applied to account for multiple testing. A significant p-value for a single gene would be p < 1.04X10^-6. That is a p-value of < 0.05 after applying a Bonferroni correction for 48,000 genes tested, as the hg38 human genome has ~48,000 annotated genes.

I would be particularly interested in any rare variant (in any gene) that was observed multiple times in our sALS samples' exome sequencing data, but not found in our healthy controls' exome sequencing data. If any of these rare variants were not found in publically available exome sequencing datasets generated from individuals that do not have any neurological disorders, I would validate each of them using Sanger sequencing. I would then generate separate transgenic mouse lines (using CRISPR/Cas9 technology) that carry each of these validated rare variants to assess whether any of them cause an ALS-like neurodegenerative phenotype characterized by muscle weakness, muscle wasting, spinal motor neuron death, paralysis, and early death. This could lead to the discovery of novel loci that carry causal ALS mutations, and may provide more insight into neurodegeneration in sALS.

For my second experiment, I would aim to identify *de novo* coding variants that cause sALS. I would first isolate blood mononuclear cells' DNA from both parents and an unaffected sibling for every qualifying sALS patient. I would then generate, align, and process exome sequencing data using these family members' DNA via the same methods proposed in my first experiment. For each trio (2 parents and their sALS-affected offspring), I would identify all coding variants found only in the sALS-affected offspring's exome sequencing data. I would apply this same approach to each trio comprised of each sALS patient's 2 parents and unaffected sibling, identifying all coding variants found only in the unaffected sibling's exome sequencing data compared to their parents' exome sequencing data.

I would be particularly interested in all *de novo* coding variants identified in sALS-affected offspring that were not found in their unaffected siblings. If these *de novo* variants were also not found in publically available exome sequencing datasets generated from individuals that do not have any neurological disorders, I would validate each variant using Sanger sequencing. I would then generate separate transgenic mouse lines (using CRISPR/Cas9 technology) that carry each of these validated rare variants to assess whether it leads to an ALS-like neurodegenerative phenotype characterized by muscle weakness, muscle wasting, spinal motor neuron death, paralysis, and early death.

Literature Cited

Abel, O., Powell, JF. Andersen, PM., Al-Chalabi, A. Credibility analysis of putative disease-causing genes using bioinformatics. *PLoS One* 8(6), e64899 (2013).

Abel, S. *et al*. The transmembrane CXC-chemokine ligand 16 is induced by IFN-gamma and TNF-alpha and shed by the activity of the disintegrin-like metalloproteinase ADAM10. *J Immunol*. 172(10), 6362-72 (2004).

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248-9.

Alexander MD, Traynor BJ, Miller N, Corr B, Frost E, McQuaid S, et al. "True" sporadic ALS associated with a novel SOD-1 mutation. Annals of neurology. 2002;52(5):680-3.

Alexianu, ME., Kozovska, M., Appel, SH. Immune reactivity in a mouse model of familial ALS correlates with disease progression. *Neurology*. 57(7), 1282-9 (2001).

Al-Chalabi A, Fang F, Hanby MF, Leigh PN, Shaw CE, Ye W, et al. An estimate of amyotrophic lateral sclerosis heritability using twin data. Journal of neurology, neurosurgery, and psychiatry. 2010;81(12):1324-6.

Al-Chalabi A, Hardiman O. The epidemiology of ALS: a conspiracy of genes, environment and time. Nature reviews Neurology. 2013;9(11):617-28.

Al-Lamki RS, Mayadas TN. TNF receptors: signaling pathways and contribution to renal dysfunction. Kidney Int. 2015 Feb;87(2):281-96.

Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, et al. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat Struct Mol Biol. 2011 Nov 6;18(12):1435-40.

Amoroso, MW. *et al*. Accelerated high-yield generation of limb-innervating motor neurons from human stem cells. *The Journal of neuroscience*: *the official journal of the society for neuroscience* 33, 574-86 (2013).

Anchisi L, Dessi S, Pani A, Mandas A. Cholesterol homeostasis: a key to prevent or slow down neurodegeneration. Front Physiol. 2013 Jan 4;3:486.

Anders, S., Pyl, PT., Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2), 166-9 (2015).

Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012 Oct;22(10):2008-17.

Andrus, PK., Fleck, TJ., Gurney, ME., Hall, ED. Protein oxidative damage in a transgenic mouse model of familial amyotrophic lateral sclerosis. *J Neurochem*. 71(5), 2041-8 (1998).

Arslan-Ergul A, Adams MM. Gene expression changes in aging zebrafish (Danio rerio) brains are sexually dimorphic. BMC Neurosci. 2014 Feb 18;15:29,2202-15-29.

Asin-Cayuela J, Gustafsson CM. Mitochondrial transcription and its regulation in mammalian cells. Trends Biochem Sci. 2007 Mar;32(3):111-7.

Artandi SE. Telomerase flies the coop: the telomerase RNA component as a viral-encoded oncogene. J Exp Med. 2006 May 15;203(5):1143-5.

Babu, GN. *et al.* Elevated inflammatory markers in a group of amyotrophic lateral sclerosis patients from northern India. *Neurochem Res*. 6, 1145-9 (2008).

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012 Dec 21;338(6114):1587-93.

Bolger, AM., Lohse, M., Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15), 2114-20 (2014).

Borthwick GM, Johnson MA, Ince PG, Shaw PJ, Turnbull DM. Mitochondrial enzyme activity in amyotrophic lateral sclerosis: implications for the role of mitochondria in neuronal cell death. Ann Neurol. 1999 Nov;46(5):787-90.

Brustolim, D., Ribeiro-dos-Santos, R., Kast, RE., Altschuler, EL., Soares, MB. A new chapter opens in anti-inflammatory treatments: the antidepressant bupropion lowers production of tumor necrosis factor-alpha and interferon-gamma in mice. *Int Immunopharmacol*. 6(6), 903-7 (2006).

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418), 61-70 (2012).

Carpenter S, Ricci EP, Mercier BC, Moore MJ, Fitzgerald KA. Post-transcriptional regulation of gene expression in innate immunity. Nat Rev Immunol. 2014 Jun;14(6):361-76.

Carter, SL., Brechbuhler, CM., Griffin, M., Bond, AT. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 20(14), 2242-50 (2004).

Cassina P, Cassina A, Pehar M, Castellanos R, Gandelman M, de Leon A, et al. Mitochondrial dysfunction in SOD1G93A-bearing astrocytes promotes motor neuron degeneration: prevention by mitochondrial-targeted antioxidants. J Neurosci. 2008 Apr 16;28(16):4115-22.

Cereda, C. *et al*. TNF and sTNFR1/2 plasma levels in ALS patients. *J Neuroimmunol*. 194(1-2), 123-31 (2008).

Chang, Y. *et al*. Messenger RNA oxidation occurs early in disease pathogenesis and promotes motor neuron degeneration in ALS. *PLoS One*. 3(8), e2849 (2008).

Che P, Gingerich DJ, Lall S, Howell SH. Global and hormone-induced gene expression changes during shoot development in Arabidopsis. Plant Cell. 2002 Nov;14(11):2771-85.

Chen X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. Sci Rep. 2015 Aug 17;5:13186.

Chesi, A. *et al*. Exome sequencing to identify de novo mutations in sporadic ALS trios. *Nat Neurosci*. 7, 851-5 (2013).

Cheung AC, Cramer P. A movie of RNA polymerase II transcription. Cell. 2012 Jun 22;149(7):1431-7.

Chhangawala S, Rudy G, Mason CE, Rosenfeld JA. The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biol. 2015 Jun 23;16:131,015-0697-y.

Chio A, Calvo A, Moglia C, Ossola I, Brunetti M, Sbaiz L, et al. A de novo missense mutation of the FUS gene in a "true" sporadic ALS case. Neurobiology of aging. 2011;32(3):553 e23-6.

Chio A, Logroscino G, Traynor BJ, Collins J, Simeone JC, Goldstein LA, et al. Global epidemiology of amyotrophic lateral sclerosis: a systematic review of the published literature. Neuroepidemiology. 2013;41(2):118-30.

Chiu, IM. *et al*. T lymphocytes potentiate endogenous neuroprotective inflammation in a mouse model of ALS. *Proc Natl Acad Sci U S A*. 105(46), 17913-8 (2008).

Cho H, Davis J, Li X, Smith KS, Battle A, Montgomery SB. High-resolution transcriptome analysis with long-read RNA sequencing. PLoS One. 2014 Sep 24;9(9):e108095.

Cieply B, Carstens RP. Functional roles of alternative splicing factors in human disease. Wiley Interdiscip Rev RNA. 2015 May-Jun;6(3):311-26.

Cirulli ET, Lasseigne BN, Petrovski S, Sapp PC, Dion PA, Leblond CS, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science. 2015 Mar 27;347(6229):1436-41.

Colombrita C, Onesto E, Megiorni F, Pizzuti A, Baralle FE, Buratti E, et al. TDP-43 and FUS RNA-binding proteins bind distinct sets of cytoplasmic messenger RNAs and differently regulate their post-transcriptional fate in motoneuron-like cells. J Biol Chem. 2012 May 4;287(19):15635-47.

Corcia, P. *et al*. Molecular imaging of microglial activation in amyotrophic lateral sclerosis. *PLoS One*. 7(12), e52941 (2012).

Cowling VH. Regulation of mRNA cap methylation. Biochem J. 2009 Dec 23;425(2):295-302.

Crick, F. On protein synthesis. Symp Soc Exp Biol. 1958;12:138-63.

Cutler RG, Pedersen WA, Camandola S, Rothstein JD, Mattson MP. Evidence that accumulation of ceramides and cholesterol esters mediates oxidative stress-induced death of motor neurons in amyotrophic lateral sclerosis. Ann Neurol. 2002 Oct;52(4):448-57.

Crugnola V, Lamperti C, Lucchini V, Ronchi D, Peverelli L, Prelle A, et al. Mitochondrial respiratory chain dysfunction in muscle from patients with amyotrophic lateral sclerosis. Arch Neurol. 2010 Jul;67(7):849-54.

D'Amico E, Factor-Litvak P, Santella RM, Mitsumoto H. Clinical perspective on oxidative stress in sporadic amyotrophic lateral sclerosis. Free Radic Biol Med. 2013 Dec;65:509-27.

Damiano, M. *et al*. Neural mitochondrial Ca2+ capacity impairment precedes the onset of motor symptoms in G93A Cu/Zn-superoxide dismutase mutant mice. *J Neurochem*. 96(5), 1349-61 (2006).

DePristo, MA. *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 43(5), 491-8 (2011).

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012 Sep 6;489(7414):101-8.

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

Donnelly CJ, Grima JC, Sattler R. Aberrant RNA homeostasis in amyotrophic lateral sclerosis: potential for new therapeutic targets? Neurodegener Dis Manag. 2014;4(6):417-37.

Dowey, SN., X, Huang., BK, Chou., Z, Ye., L, Cheng. Generation of integration-free human induced pluripotent stem cells from postnatal blood mononuclear cells by plasmid vector expression. *Nature Protocol* 7, 2013-21 (2012).

Duffy LM, Chapman AL, Shaw PJ, Grierson AJ. Review: The role of mitochondria in the pathogenesis of amyotrophic lateral sclerosis. Neuropathol Appl Neurobiol. 2011 Jun;37(4):336-52.

Echaniz-Laguna A, Zoll J, Ponsot E, N'guessan B, Tranchant C, Loeffler JP, et al. Muscular mitochondrial function in amyotrophic lateral sclerosis is progressively altered as the disease develops: a temporal study in man. Exp Neurol. 2006 Mar;198(1):25-30.

Elliott, JL. Cytokine upregulation in a murine model of familial amyotrophic lateral sclerosis. *Brain Res Mol Brain Res*. 95(1-2), 172-8 (2001).

Engstrom, PG. *et al*. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* 12, 1185-91 (2013).

Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. Annu Rev Biochem. 2010;79:351-79.

Falschlehner C, Emmerich CH, Gerlach B, Walczak H. TRAIL signalling: decisions between life and death. Int J Biochem Cell Biol. 2007;39(7-8):1462-75.

Fang F, Kamel F, Lichtenstein P, Bellocco R, Sparen P, Sandler DP, et al. Familial aggregation of amyotrophic lateral sclerosis. Annals of neurology. 2009;66(1):94-9.

Fanning, LB., Buckley, CC., Xing, W., Breslow, RG., Katz, HR. Downregulation of key early events in the mobilization of antigen-bearing dendritic cells by leukocyte immunoglobulin-like Receptor B4 in a mouse model of allergic pulmonary inflammation. *PLoS One* 8(2), e57007 (2013).

Ferrante, RJ. *et al.* Evidence of increased oxidative damage in both sporadic and familial amyotrophic lateral sclerosis*.* *J Neurochem*. 69(5), 2064-74 (1997).

Figueroa-Romero, C. *et al*. Identification of epigenetically altered genes in sporadic amyotrophic lateral sclerosis. *PLoS One* 7(12), e52672 (2012).

Fujita K, Yamauchi M, Shibayama K, Ando M, Honda M, Nagata Y. Decreased cytochrome c oxidase activity but unchanged superoxide dismutase and glutathione peroxidase activities in the spinal cords of patients with amyotrophic lateral sclerosis. J Neurosci Res. 1996 Aug 1;45(3):276-81.

Ganesalingam, J. *et al*. Latent cluster analysis of ALS phenotypes identifies prognostically differing groups. *PLoS One* 4(9), e7107 (2009).

Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011 Jun;8(6):469-77.

Grummt I. Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. Genes Dev. 2003 Jul 15;17(14):1691-702.

Hagl, S *et al*. Curcumin micelles improve mitochondrial function in neuronal PC12 cells and brains of NMRI mice - Impact on bioavailability. *Neurochem Int*. doi: 10.1016/j.neuint.2015.07.026 (2015).

Hall, ED., Andrus, PK., Oostveen, JA., Fleck, TJ., Gurney, ME. Relationship of oxygen radical-induced lipid peroxidative damage to disease onset and progression in a transgenic model of familial ALS. *J Neurosci Res*. 53(1), 66-77 (1998).

Han, SS., Keum, YS., Seo, HJ., Surh, YJ. Curcumin suppresses activation of NF-kappaB and AP-1 induced by phorbol ester in cultured human promyelocytic leukemia cells. *J Biochem Mol Biol*. 35(3), 337-42 (2002).

Hardiman O, Greenway M. The complex genetics of amyotrophic lateral sclerosis. The Lancet Neurology. 2007;6(4):291-2.

Hattermann, K., Ludwig, A., Gieselmann, V., Held-Feindt, J., Mentlein, R. The chemokine CXCL16 induces migration and invasion of glial precursor cells via its receptor CXCR6. *Mol Cell Neurosci*. 39(1), 133-41 (2008).

He BP, Wen W, Strong MJ. Activated microglia (BV-2) facilitation of TNF-alpha-mediated motor neuron death in vitro. J Neuroimmunol. 2002 Jul;128(1-2):31-8.

Heath, PR., Kirby, J., Shaw, PJ. Investigating cell death mechanisms in amyotrophic lateral sclerosis using transcriptomics. *Front Cell Neurosci*. 7, 259 (2013).

Henkel, JS., Beers, DR., Siklos, L., Appel, SH. The chemokine MCP-1 and the dendritic and myeloid cells it attracts are increased in the mSOD1 mouse model of ALS. *Mol Cell Neurosci*. 3, 427-37 (2006).

Hensley, K. *et al*. Temporal patterns of cytokine and apoptosis-related gene expression in spinal cords of the G93A-SOD1 mouse model of amyotrophic lateral sclerosis. *J Neurochem*. 82(2), 365-74 (2002).

Henkel, JS. *et al*. Presence of dendritic cells, MCP-1, and activated microglia/macrophages in amyotrophic lateral sclerosis spinal cord tissue. *Ann Neurol*. 55(2), 221-35 (2004).

Higgins, CM., Jung, C., Xu, Z. ALS-associated mutant SOD1G93A causes mitochondrial vacuolation by expansion of the intermembrane space and by involvement of SOD1 aggregation and peroxisomes. *BMC Neurosci.* 4, 16 (2003).

Highley JR, Kirby J, Jansweijer JA, Webb PS, Hewamadduma CA, Heath PR, et al. Loss of nuclear TDP-43 in amyotrophic lateral sclerosis (ALS) causes altered expression of splicing machinery and widespread dysregulation of RNA splicing in motor neurones. Neuropathol Appl Neurobiol. 2014 Oct;40(6):670-85.

Hirano, A., Donnenfeld, H., Sasaki, S., Nakano, I. Fine structural observations of neurofilamentous changes in amyotrophic lateral sclerosis. *J Neuropathol Exp Neurol*. 43(5), 461-70 (1984).

Hoesel, B., Schmid, JA. The complexity of NF-kappaB signaling in inflammation and cancer. *Mol Cancer* 12, 86 (2013).

Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo
MM, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nature genetics. 2011;43(5):429-35.

Holtman, IR. *et al*. Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a co-expression meta-analysis. *Acta Neuropathol Commun*. 3, 31 (2015).

Horvath, S. *et al*. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc Natl Acad Sci U S A* 103(46), 17402-7 (2006).

Ichiyanagi N, Fujimori K, Yano M, Ishihara-Fujisaki C, Sone T, Akiyama T, et al. Establishment of In Vitro FUS-Associated Familial Amyotrophic Lateral Sclerosis Model Using Human Induced Pluripotent Stem Cells. Stem Cell Reports. 2016 Mar 16.

Ikiz, B. *et al*. The Regulatory Machinery of Neurodegeneration in In Vitro Models of Amyotrophic Lateral Sclerosis. *Cell Rep*. 12(2), 335-45 (2015).

Jackson RJ, Hellen CU, Pestova TV. The mechanism of eukaryotic translation initiation and principles of its regulation. Nat Rev Mol Cell Biol. 2010 Feb;11(2):113-27.

Jeong, H., Mason, SP., Barabasi, AL., Oltvai, ZN. Lethality and centrality in protein networks. *Nature* 411(6833), 41-2 (2001).

Jin, CY., Lee, JD., Park, C., Choi, YH., Kim, GY. Curcumin attenuates the release of pro-inflammatory cytokines in lipopolysaccharide-stimulated BV2 microglia. *Acta Pharmacol Sin*. 10, 1645-51 (2007).

Kapp LD, Lorsch JR. The molecular mechanics of eukaryotic translation. Annu Rev Biochem. 2004;73:657-704.

Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res. 2015 Dec 2;43(21):e143.

Kee JY, Ito A, Hojo S, Hashimoto I, Igarashi Y, Tsuneyama K, et al. CXCL16 suppresses liver metastasis of colorectal cancer by promoting TNF-alpha-induced apoptosis by tumor-associated macrophages. BMC Cancer. 2014 Dec 15;14:949,2407-14-949.

Khan, S. *et al*. Implication of Caspase-3 as a Common Therapeutic Target for Multineurodegenerative Disorders and Its Inhibition Using Nonpeptidyl Natural Compounds. *Biomed Res Int*. 2015, 379817 (2015).

Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012 May 29;44(6):623-30.

Kim, D. *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 14, R36 (2013).

Kim SJ, Cheresh P, Williams D, Cheng Y, Ridge K, Schumacker PT, et al. Mitochondria-targeted Ogg1 and aconitase-2 prevent oxidant-induced mitochondrial DNA damage in alveolar epithelial cells. J Biol Chem. 2014 Feb 28;289(9):6165-76.

Kogelman, LJ. *et al*. Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model. *BMC Med Genomics* 7, 57 (2014).

Kong, J., Xu, Z. Massive mitochondrial degeneration in motor neurons triggers the onset of amyotrophic lateral sclerosis in mice expressing a mutant SOD1. *J Neurosci*. 18(9), 3241-50 (1998).

Kornblihtt AR, Schor IE, Allo M, Dujardin G, Petrillo E, Munoz MJ. Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Biol. 2013 Mar;14(3):153-65.

Kratz, A., Carninci, P. The devil in the details of RNA-seq. *Nature biotechnology* 32(9), 882-4 (2014).

Kwek KY, Murphy S, Furger A, Thomas B, O'Gorman W, Kimura H, et al. U1 snRNA associates with TFIIH and regulates transcriptional initiation. Nat Struct Biol. 2002 Nov;9(11):800-5.

Kwiatkowski TJ,Jr, Bosco DA, Leclerc AL, Tamrazian E, Vanderburg CR, Russ C, et al. Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. Science. 2009 Feb 27;323(5918):1205-8.

Ladd AC, Keeney PM, Govind MM, Bennett JP,Jr. Mitochondrial oxidative phosphorylation transcriptome alterations in human amyotrophic lateral sclerosis spinal cord and blood. Neuromolecular Med. 2014 Dec;16(4):714-26.

Lagier-Tourenne C, Polymenidou M, Hutt KR, Vu AQ, Baughn M, Huelga SC, et al. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. Nat Neurosci. 2012 Nov;15(11):1488-97.

Landrum, MJ. *et al*. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 42(Database issue), D980-5 (2014).

Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559 (2008).

Langfelder, P., Horvath, S. Fast R Functions for Robust Correlations and Heirarchical Clustering. *Journal of Statistical Software* 46, 1-17 (2012).

Lagier-Tourenne C, Cleveland DW. Rethinking ALS: the FUS about TDP-43. Cell. 2009 Mar 20;136(6):1001-4.

Lagier-Tourenne C, Polymenidou M, Hutt KR, Vu AQ, Baughn M, Huelga SC, et al. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. Nat Neurosci. 2012 Nov;15(11):1488-97.

Learned RM, Cordes S, Tjian R. Purification and characterization of a transcription factor that confers promoter specificity to human RNA polymerase I. Mol Cell Biol. 1985 Jun;5(6):1358-69.

Lewis, KE. *et al*. Microglia and motor neurons during disease progression in the SOD1G93A mouse model of amyotrophic lateral sclerosis: changes in arginase1 and inducible nitric oxide synthase. *J Neuroinflammation* 11, 55 (2014).

Li, H., Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5), 589-95 (2010).

Liu B, Turley SD, Burns DK, Miller AM, Repa JJ, Dietschy JM. Reversal of defective lysosomal transport in NPC disease ameliorates liver dysfunction and neurodegeneration in the npc1-/- mouse. Proc Natl Acad Sci U S A. 2009 Feb 17;106(7):2377-82.

Liu, D., Wen, J., Liu, J., Li, L. The roles of free radicals in amyotrophic lateral sclerosis: reactive oxygen species and elevated oxidation of protein, DNA, and membrane phospholipids. *FASEB J*. 15, 2318-28 (1999).

Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014 Feb 1;30(3):301-4.

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

Luirink J, Sinning I. SRP-mediated protein targeting: structure and function revisited. Biochim Biophys Acta. 2004 Nov 11;1694(1-3):17-35.

Luirink J, Sinning I. SRP-mediated protein targeting: structure and function revisited. Biochim Biophys Acta. 2004 Nov 11;1694(1-3):17-35.

Ma, Y. *et al*. Microarray analysis uncovers retinoid targets in human bronchial epithelial cells. *Oncogene* 22(31), 4924-32 (2003).

Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 2011 May 31;9:34,7007-9-34.

Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, et al. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. Med Sci Monit Basic Res. 2014 Aug 23;20:138-42.

Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 2014 Mar;24(3):496-510.

Martin MG, Pfrieger F, Dotti CG. Cholesterol in brain disease: sometimes determinant and frequently implicated. EMBO Rep. 2014 Oct;15(10):1036-52.

Maschietto, M. *et al*. Co-expression network of neural-differentiation genes shows specific pattern in schizophrenia. *BMC Med Genomics* 8, 23 (2015).

Masuda A, Takeda JI, Ohno K. FUS-mediated regulation of alternative RNA processing in neurons: insights from global transcriptome analysis. Wiley Interdiscip Rev RNA. 2016 Jan 28.

Matera AG, Terns RM, Terns MP. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol. 2007 Mar;8(3):209-20.

Matthews GM, Newbold A, Johnstone RW. Intrinsic and extrinsic apoptotic pathway signaling as determinants of histone deacetylase inhibitor antitumor activity. Adv Cancer Res. 2012;116:165-97

Mattiazzi, M. *et al*. Mutated human SOD1 causes dysfunction of oxidative phosphorylation in mitochondria of transgenic mice. *J Biol Chem*. 277(33), 29626-33 (2002).

McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(9), 1297-303 (2010).

Micheau, O., Tschopp, J. Induction of TNF receptor I-mediated apoptosis via two sequential signaling complexes. *Cell* 114(2), 181-90 (2003).

Mohler, KM. *et al*. Soluble tumor necrosis factor (TNF) receptors are effective therapeutic agents in lethal endotoxemia and function simultaneously as both TNF carriers and TNF antagonists. *J Immunol*. 151(3), 1548-61 (1993).

Mutihac R, Alegre-Abarrategui J, Gordon D, Farrimond L, Yamasaki-Mann M, Talbot K, et al. TARDBP pathogenic mutations increase cytoplasmic translocation of TDP-43 and cause reduction of endoplasmic reticulum Ca(2)(+) signaling in motor neurons. Neurobiol Dis. 2015 Mar;75:64-77.

Nagy, D., Kato, T., Kushner, PD. Reactive astrocytes are widespread in the cortical gray matter of amyotrophic lateral sclerosis. *J Neurosci Res*. 38(3), 336-47 (1994).

Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet. 2011;377(9766):641-9.

Nardo G, Iennaco R, Fusi N, Heath PR, Marino M, Trolese MC, et al. Transcriptomic indices of fast and slow disease progression in two mouse models of amyotrophic lateral sclerosis. Brain : a journal of neuke2013;136(Pt 11):3305-32.

O'Brien, LC., Keeney, PM., Bennett, JP,Jr. Differentiation of Human Neural Stem Cells into Motor Neurons Stimulates Mitochondrial Biogenesis and Decreases Glycolytic Flux. *Stem Cells Dev*. 24(17), 1984-94 (2015).

O'Neil D, Glowatz H, Schlumpberger M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol. 2013 Jul;Chapter 4:Unit 4.19.

Ofotokun I, Titanji K, Vikulina T, Roser-Page S, Yamaguchi M, Zayzafoon M, et al. Role of T-cell reconstitution in HIV-1 antiretroviral therapy-induced bone loss. Nat Commun. 2015 Sep 22;6:8282.

Ousman SS, Kubes P. Immune surveillance in the central nervous system. Nat Neurosci. 2012 Jul 26;15(8):1096-101.

Park, DJ., Vuong, PT., de Vos, S., Douer, D., Koeffler, HP. Comparative analysis of genes regulated by PML/RAR alpha and PLZF/RAR alpha in response to retinoic acid using oligonucleotide arrays. *Blood* 102(10), 3727-36 (2003).

Pavlou, MP., Dimitromanolakis, A., Diamandis, EP. Coupling proteomics and transcriptomics in the quest of subtype-specific proteins in breast cancer. *Proteomics* 7, 1083-95 (2013).

Piatek MJ, Werner A. Endogenous siRNAs: regulators of internal affairs. Biochem Soc Trans. 2014 Aug;42(4):1174-9.

Piccinelli P, Rosenblad MA, Samuelsson T. Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. Nucleic Acids Res. 2005 Aug 8;33(14):4485-95.

Podrabsky JE, Somero GN. Changes in gene expression associated with acclimation to constant temperatures and fluctuating daily temperatures in an annual killifish Austrofundulus limnaeus. J Exp Biol. 2004 Jun;207(Pt 13):2237-54.

Poloni, M. *et al*. Circulating levels of tumour necrosis factor-alpha and its soluble receptors are increased in the blood of patients with amyotrophic lateral sclerosis. *Neurosci Lett.* 287(3), 211-4 (2000).

Polymenidou M, Lagier-Tourenne C, Hutt KR, Huelga SC, Moran J, Liang TY, et al. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nat Neurosci. 2011 Apr;14(4):459-68.

Posse V, Shahzad S, Falkenberg M, Hallberg BM, Gustafsson CM. TEFM is a potent stimulator of mitochondrial transcription elongation in vitro. Nucleic Acids Res. 2015 Mar 11;43(5):2615-24.

Probert, L. TNF and its receptors in the CNS: The essential, the desirable and the deleterious effects. *Neuroscience* 302, 2-22 (2015).

Rabin SJ, Kim JM, Baughn M, Libby RT, Kim YJ, Fan Y, et al. Sporadic ALS has compartment-specific aberrant exon splicing and altered cell-matrix adhesion biology. Hum Mol Genet. 2010 Jan 15;19(2):313-28.

Rhodes JS, Crabbe JC. Gene expression induced by drugs of abuse. Curr Opin Pharmacol. 2005 Feb;5(1):26-33.

Robinson, MD., McCarthy, DJ., Smyth, GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1), 139-40 (2010).

Renton, AE., Chio, A., Traynor, BJ. State of play in amyotrophic lateral sclerosis genetics. *Nat Neurosci.* 1, 17-23 (2014).

Robertson J, Beaulieu JM, Doroudchi MM, Durham HD, Julien JP, Mushynski WE. Apoptotic death of neurons exhibiting peripherin aggregates is mediated by the proinflammatory cytokine tumor necrosis factor-alpha. J Cell Biol. 2001 Oct 15;155(2):217-26.

Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 2012 Jun 18;13(7):505-16.

Rossel JB, Wilson IW, Pogson BJ. Global changes in gene expression in response to high light in Arabidopsis. Plant Physiol. 2002 Nov;130(3):1109-20.

Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D, et al. An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res. 2007 Jan;35(Database issue):D823-8.

Rusiniak, ME., Yu, M., Ross, DT., Tolhurst, EC., Slack, JL. Identification of B94 (TNFAIP2) as a potential retinoic acid target gene in acute promyelocytic leukemia. *Cancer Res*. 60(7), 1824-9 (2000).

Saito, A. *et al.* An integrated expression profiling reveals target genes of TGF-beta and TNF-alpha possibly mediated by microRNAs in lung cancer cells. *PLoS One* 8(2), e56587 (2013).

Salzman J, Jiang H, Wong WH. Statistical Modeling of RNA-Seq Data. Stat Sci. 2011 Feb;26(1):10.1214/10-STS343.

Saris, CG. *et al*. Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. *BMC Genomics*. 10, 405 (2009).

Sasaki, S., Iwata, M. Dendritic synapses of anterior horn neurons in amyotrophic lateral sclerosis: an ultrastructural study. *Acta Neuropathol*. 91(3), 278-83 (1996).

Sasaki, S., Iwata, M. Mitochondrial alterations in the spinal cord of patients with sporadic amyotrophic lateral sclerosis. *J Neuropathol Exp Neurol*. 66(1), 10-6 (2007).

Satoh, J. *et al*. Molecular network analysis suggests a logical hypothesis for the pathological role of c9orf72 in amyotrophic lateral sclerosis/frontotemporal dementia. *J Cent Nerv Syst Dis*. 6, 69-78 (2014).

Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20;270(5235):467-70.

Schiffer, D., Cordera, S., Cavalla, P., Migheli, A. Reactive astrogliosis of the spinal cord in amyotrophic lateral sclerosis. *J Neurol Sci*. 139, 27-33 (1996).

Schmid, B. *et al.* Loss of ALS-associated TDP-43 in zebrafish causes muscle degeneration, vascular dysfunction, and reduced motor neuron axon outgrowth. Proc Natl Acad Sci USA. 2013 110(13):4986-4991.

Schroder O, Bryant GO, Geiduschek EP, Berk AJ, Kassavetis GA. A common site on TBP for transcription by RNA polymerases II and III. EMBO J. 2003 Oct 1;22(19):5115-24.

Schwartz MP, Hou Z, Propson NE, Zhang J, Engstrom CJ, Santos Costa V, et al. Human pluripotent stem cell-derived neural constructs for predicting neural toxicity. Proc Natl Acad Sci U S A. 2015 Oct 6;112(40):12516-21.

Sellin Jeffries MK, Kiss AJ, Smith AW, Oris JT. A comparison of commercially-available automated and manual extraction kits for the isolation of total RNA from small tissue samples. BMC Biotechnol. 2014 Nov 14;14:94,014-0094-8.

Shanker S, Paulson A, Edenberg HJ, Peak A, Perera A, Alekseyev YO, et al. Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. J Biomol Tech. 2015 Apr;26(1):4-18.

Smits P, Smeitink J, van den Heuvel L. Mitochondrial translation and beyond: processes implicated in combined oxidative phosphorylation deficiencies. J Biomed Biotechnol. 2010;2010:737385.

Stenson, PD. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet*. 133(1), 1-9 (2014).

Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature. 2014 Apr 3;508(7494):66-71.

Surh, YJ. *et al*. Molecular mechanisms underlying chemopreventive activities of anti-inflammatory phytochemicals: down-regulation of COX-2 and iNOS through suppression of NF-kappa B activation. *Mutat Res*. 480, 243-68 (2001).

Taanman JW. The mitochondrial genome: structure, transcription, translation and replication. Biochim Biophys Acta. 1999 Feb 9;1410(2):103-23.

Tanaka, H. *et al*. The potential of GPNMB as novel neuroprotective factor in amyotrophic lateral sclerosis. *Sci Rep*. 2, 573 (2012).

Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Res. 2011 Dec;21(12):2213-23.

Tateishi T, Yamasaki R, Tanaka M, Matsushita T, Kikuchi H, Isobe N, et al. CSF chemokine alterations related to the clinical course of amyotrophic lateral sclerosis. J Neuroimmunol. 2010 May;222(1-2):76-81.

Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. Nat Rev Genet. 2005 May;6(5):389-402.

Terrado J, Monnier D, Perrelet D, Vesin D, Jemelin S, Buurman WA, et al. Soluble TNF receptors partially protect injured motoneurons in the postnatal CNS. Eur J Neurosci. 2000 Sep;12(9):3443-7.

Tian, B., Nowak, DE., Jamaluddin, M., Wang, S., Brasier, AR. Identification of direct genomic targets downstream of the nuclear factor-kappaB transcription factor mediating tumor necrosis factor signaling. *J Biol Chem*. 280(17), 17435-48 (2005).

Tollervey JR, Curk T, Rogelj B, Briese M, Cereda M, Kayikci M, et al. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nat Neurosci. 2011 Apr;14(4):452-8.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5.

Trapnell, C. *et al*. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol*. 1, 46-53 (2013).

Tuppen HA, Blakely EL, Turnbull DM, Taylor RW. Mitochondrial DNA mutations and human disease. Biochim Biophys Acta. 2010 Feb;1797(2):113-28.

Turner, MR. *et al.* Evidence of widespread cerebral microglial activation in amyotrophic lateral sclerosis: an [11C](R)-PK11195 positron emission tomography study. *Neurobiol Dis*. 3, 601-9 (2004).

Van der Auwera, GA. *et al*. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 11(1110), 11.10.1,11.10.33 (2013).

van Spriel, AB. *et al*. A regulatory role for CD37 in T cell proliferation. *J Immunol*. 172(5), 2953-61 (2004).

Vance JE. Dysregulation of cholesterol balance in the brain: contribution to neurodegenerative diseases. Dis Model Mech. 2012 Nov;5(6):746-55.

Vanden Broeck L, Kleinberger G, Chapuis J, Gistelinck M, Amouyel P, Van Broeckhoven C, et al. Functional complementation in Drosophila to predict the pathogenicity of TARDBP variants: evidence for a loss-of-function mechanism. Neurobiol Aging. 2015 Feb;36(2):1121-9.

Vanier MT, Latour P. Laboratory diagnosis of Niemann-Pick disease type C: the filipin staining test. Methods Cell Biol. 2015;126:357-75.

Vargas MR, Pehar M, Diaz-Amarilla PJ, Beckman JS, Barbeito L. Transcriptional profile of primary astrocytes expressing ALS-linked mutant SOD1. J Neurosci Res. 2008 Dec;86(16):3515-25.

Venkova K, Christov A, Kamaluddin Z, Kobalka P, Siddiqui S, Hensley K. Semaphorin 3A signaling through neuropilin-1 is an early trigger for distal axonopathy in the SOD1G93A mouse model of amyotrophic lateral sclerosis. J Neuropathol Exp Neurol. 2014 Jul;73(7):702-13.

Verghese, PB., Castellano, JM., Holtzman, DM. Apolipoprotein E in Alzheimer's disease and other neurological disorders. *Lancet Neurol*. 10(3), 241-52 (2011).

Vielhaber S, Kunz D, Winkler K, Wiedemann FR, Kirches E, Feistner H, et al. Mitochondrial DNA abnormalities in skeletal muscle of patients with sporadic amyotrophic lateral sclerosis. Brain. 2000 Jul;123 (Pt 7)(Pt 7):1339-48.

Viktorovskaya OV, Schneider DA. Functional divergence of eukaryotic RNA polymerases: unique properties of RNA polymerase I suit its cellular role. Gene. 2015 Feb 1;556(1):19-26.

Walczak, H. TNF and ubiquitin at the crossroads of gene activation, cell death, inflammation, and cancer. *Immunol Rev*. 244(1), 9-28 (2011).

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

WATSON JD, CRICK FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953 Apr 25;171(4356):737-8.

Weick EM, Miska EA. piRNAs: from biogenesis to function. Development. 2014 Sep;141(18):3458-71.

Weirauch, Matthew T (2011). "Gene coexpression networks for the analysis of DNA microarray data". *Applied Statistics for Network Biology: Methods in Systems Biology*.

White RJ. Transcription by RNA polymerase III: more complex than we thought. Nat Rev Genet. 2011 May 4;12(7):459-63.

Wiedemann FR, Manfredi G, Mawrin C, Beal MF, Schon EA. Mitochondrial DNA and respiratory chain function in spinal cords of ALS patients. J Neurochem. 2002 Feb;80(4):616-25.

Wiedemann FR, Winkler K, Kuznetsov AV, Bartels C, Vielhaber S, Feistner H, et al. Impairment of mitochondrial function in skeletal muscle of patients with amyotrophic lateral sclerosis. J Neurol Sci. 1998;156(1):65-72.

Wilhelm BT, Landry JR. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods. 2009 Jul;48(3):249-57.

Will CL, Luhrmann R. Spliceosome structure and function. Cold Spring Harb Perspect Biol. 2011 Jul 1;3(7):10.1101/cshperspect.a003707.

Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 2009 Jul 1;23(13):1494-504.

Wingo TS, Cutler DJ, Yarab N, Kelly CM, Glass JD. The heritability of amyotrophic lateral sclerosis in a clinically ascertained United States research registry. PloS one. 2011;6(11):e27985.

Xu ZS. Does a loss of TDP-43 function cause neurodegeneration? Mol Neurodegener. 2012 Jun 14;7:27,1326-7-27.

Yang C, Wang H, Qiao T, Yang B, Aliaga L, Qiu L, et al. Partial loss of TDP-43 function causes phenotypes of amyotrophic lateral sclerosis. Proc Natl Acad Sci U S A. 2014 Mar 25;111(12):E1121-9.

Yoshihara, T. *et al*. Differential expression of inflammation- and apoptosis-related genes in spinal cords of a mutant SOD1 transgenic mouse model of familial amyotrophic lateral sclerosis. *J Neurochem*. 80(1), 158-67 (2002).

Zhang, B., Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 4, Article17 (2005).

Zhang, ZH. *et al*. A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS One* 9(8), e103207 (2014).

Zhou, A., Scoggin, S., Gaynor, RB., Williams, NS. Identification of NF-kappa B-regulated genes induced by TNFalpha utilizing expression profiling and RNA interference. *Oncogene* 22(13), 2054-64 (2003).

Zhao, W., Beers, DR., Appel, SH. Immune-mediated mechanisms in the pathoprogression of amyotrophic lateral sclerosis. *J Neuroimmune Pharmacol*. 4, 888-99 (2013).

Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics. 2014 Jun 2;15:419,2164-15-419.

Vita

David Graf Brohawn was born on November 27, 1985 in Baltimore, Maryland, and is an American citizen. He graduated from Mount Saint Joseph High School, Baltimore, Maryland in 2004. He received his Bachelor of Arts in Psychology from University of Delaware, Newark, Delaware in 2008.