Characterization of Amyotrophic Lateral Sclerosis Patient Heterogeneity

Using Postmortem Gene Expression

by

Jarrett Eshima

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2024 by the
Graduate Supervisory Committee:

Barbara S. Smith, Chair
Christopher L. Plaisier
Xiaojun Tian
John Fricks
Robert Bowser

ARIZONA STATE UNIVERSITY

May 2024

ABSTRACT

Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disease characterized by the progressive loss of motor function. Pathological mechanisms and clinical measures vary extensively from patient to patient, creating a spectrum of disease phenotypes with a poorly understood influence on individual outcomes like disease duration. The inability to ascertain patient phenotype has hindered clinical trial design and the development of more personalized and effective therapeutics. Wholistic analytical methods ('-omics') have provided unprecedented molecular resolution into cellular and system level disease processes and offer a foundation to better understand ALS disease variability. Building off initiatives by the New York Genome Center ALS Consortium and Target ALS groups, the goal of this work was to stratify a large patient cohort utilizing a range of bioinformatic strategies and bulk tissue gene expression (transcriptomes) from the brain and spinal cord. Central Hypothesis: Variability in the onset and progression of ALS is partially captured by molecular subgroups (subtypes) with distinct gene expression profiles and implicated pathologies. Work presented in this dissertation addresses the following: (Chapter 2): The use of unsupervised clustering and gene enrichment methods for the identification and characterization of patient subtypes in the postmortem cortex and spinal cord. Results obtained from this Chapter establish three ALS subtypes, identify uniquely dysregulated pathways, and examine intra-patient concordance between regions of the central nervous system. (Chapter 3): Patient subtypes from Chapter 2 are considered in the context of clinical outcomes, leveraging multiple survival models and gene co-expression analyses. Results from this Chapter establish a weak association between ALS subtype and clinical outcomes including disease duration and age at symptom onset. (Chapter 4): Utilizing differential expression analysis, ‘marker’ genes are defined and leveraged with supervised classification (“machine learning”) methods to develop a suite of classifiers design to stratify patients by subtype. Results from this Chapter provide postmortem marker genes for two of the three ALS subtypes and offer a foundation for clinical stratification. Significance: Knowledge gained from this research provides a foundation to stratify patients in the clinic and prior to enrollment in clinical trials, offering a path toward improved therapies.

DEDICATION

This dissertation is dedicated to my father, Dr. Dennis Eshima, who passed during my undergraduate degree. My father, a nuclear pharmacist, ignited my passion for science and was relentless in his support of my learning and growth – sharing his perspectives, creativity, enthusiasm, love, encouragement, and guidance. I am forever grateful for you.

To my family, Grayson, Erik, Lorie, Andrew, Matthew, Shigeko, Nobuo, Duane, Carole, Azaria, Mikey, Hannah, Kristen, Gary, Sam, Nolan, Audra, and Grace.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my committee chair and long-time research mentor, Barbara Smith. As a wide-eyed, overeager undergraduate in my first year, she took a risk and offered me a position in her lab. Beyond that, she gave me entirely unique opportunities like funding my travel around the United States to learn GC-MS as a second-year undergraduate, allowing me to lead a research study without a bachelor’s degree, and offering nearly endless time to ensure I felt as confident as I could with manuscript revisions, fellowship applications, and such. During my graduate studies, she was persistent in her encouragement of my work, gave me the freedom to pursue my own question, and fully supported the pursuit of lofty ideas. Her unwavering belief, interest, patience, passion, and day-to-day positivity over nine years is remarkable and she has fundamentally shaped my growth as a scientist. I am sincerely grateful for my time in your research lab and the extensive opportunities and support you have given me.

In addition to my committee chair, I consider myself exceptionally lucky to count a number of other faculty as mentors. As an undergraduate and graduate, Heather Bean always offered her incredible depth of knowledge related to GC-MS and volatile metabolomics. She helped establish my foundation in the field of metabolomics and showed remarkable patience and support at the beginning of my research journey. Our unexpected encounter and conversation on a campus bus on the other side of the country is a fond memory of mine and helped me overcome homesickness at the time. Equally supportive, John Fricks helped guide me through the field of statistics and provided a crucial foundation for my success. You enabled me to carry out more complicated (and exciting) modeling with the confidence and understanding to make sense of the results and validity. Lastly, and fundamental to my growth and development as a scientist, Christopher Plaisier drastically widened my perspective and appreciation for biological complexity. In the first semester of my PhD, you formally introduced me to the field of systems biology and taught the course in a way where I knew I had found what I was both interested in and enjoyed doing. You were consistently supportive of my research and introduced me to ideas that challenged me and enabled me to grow continuously throughout. Even though I wasn't in your lab, you always made time to meet with me and provide extra support and guidance which I am beyond grateful for. You helped shaped the way I approach research and showed me how creativity and science are complementary. Thank you for all the hours you spent discussing my research, pointing me in the right direction and reviewing "interesting" (garbage) models.

Formatting guidelines required placement of the remaining written material as Appendix A.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

The overarching goal of this dissertation is to characterize Amyotrophic Lateral Sclerosis (ALS) patient variability by stratifying a large, publicly available cohort (NCBI GEO Accession: GSE153960) using transcript expression from the postmortem cortex and spinal cord. Findings from this dissertation provide new insight into mechanistic heterogeneity in ALS patients through the discovery of unique molecular subtypes and support the discovery and validation of molecular features to stratify a living cohort. This is demonstrated through the application of a suite of bioinformatic strategies, including: unsupervised clustering, enrichment, survival, correlation-based networks, differential expression, and supervised clustering – presented in subsequent chapters. The resulting insight into ALS patient heterogeneity outlined in this work has the potential to improve clinical trial outcomes by stratifying patients prior to enrollment and guide the design of novel therapies or repurposing of existing drugs. Further, subtype-dependent differences in survival provides a foundation to inform patients of their prognosis and guide clinical decision making.

1.1 Overview

Previous studies provide an important foundation for this work and offer insight into potential drivers of disease variability. Discovery of many genetic mutations in the *SOD1* protein have spurred the development of cell and animal models to better understand the role of the reactive oxygen species (ROS) mediating protein in ALS neurodegeneration. While no consensus on the mechanism of *SOD1* fALS have emerged, more evidence points to a spectrum of toxic 'gain-of-functions' rather than directly due to the loss of ROS attenuating activity[321] – although oxidative and mitochondrial stress are commonly reported[106,131,288]. Even as the number of known mutant proteins and genetic aberrations causing ALS has grown, similar mechanistic disease variability has been observed across disease models[223]. A comprehensive overview of this heterogeneity is provided by Drs. Taylor, Brown Jr., and Cleveland, however in brief, cell and animal models provide evidence for the presence of (i) proteotoxic stress and disturbances in protein quality control, (ii) activated microglia and astrocytes, (iii) negative metabolic regulation reducing energy availability, (iv) dysregulated transcription and RNA metabolism, (v) glutamate associated excitotoxicity, and (vi) defects in the cytoskeleton and axonal transport[321]. Maturing '-omic' approaches offer a platform to stratify patients using quantifiable molecular features when paired with bioinformatics, enabling a better understanding of mechanistic heterogeneity in ALS and how it influences disease outcomes. Benefiting from the global nature of the analysis, large cohort size, extensive transcriptomic quantification throughout the postmortem central nervous system, and detailed patient records, this work 1) identifies three molecular subtypes in the cortex and spinal cord and considers the coherence of intra-patient subtype presentation, 2) links the subtypes to variability in patient outcomes, including weak associations with survival, and 3) leverages differential expression analysis to uncover marker genes for the ALS-Ox subtype that are consistently upregulated in the cortex and spinal cord and demonstrate clinically appreciable stratification accuracies.

1.2 Amyotrophic Lateral Sclerosis

A typical definition of ALS involves the progressive onset of muscle weakness with degeneration primarily affecting motor neurons in the brain and spinal cord of patients[321], with early symptoms often including cortical hyperexcitability, muscle spasticity and fasciculations, and difficulties chewing, swallowing and with speech. ALS is a fatal disease, typically concluding with respiratory failure after an average of 3-5 years[321]. Roughly 5-10% of cases have a known hereditary association (familial ALS), while the remaining 90% do not, known as sporadic ALS. Disease onset occurs in mid-adulthood with a mean age of 55 years[321]. Men are reported to have a slightly higher incidence than women in populations of European and Japanese descent (1.2–1.5:1)[89,138]. Diagnostic timelines vary but often take roughly a year with more than half of patients receiving a misdiagnosis[254], stemming from the lack of disease-specific biomarkers and symptom presentation that overlaps with other neurodegenerative diseases[40]. Given the absence of disease-specific biomarkers, the El Escorial criteria serves to guide clinical diagnosis while the revised ALS functional rating scale (ALSFRS-R) has been developed to approximate rate of progression and severity, but remains limited by subjectivity of scoring and patient variability[45,58]. Riluzole is the only effective therapy and generally offers patients a meager 2-6 months of prolonged survival[138,232], highlighting a clear need for continued work.

Offering insight into the therapeutic gap, a persistent finding in ALS research has been the heterogeneity in phenotypes and outcomes observed in cell models, animal models, and human patients. Broadly interpreted here as ‘atypical’ ALS, evidence for disease onset and progression that falls outside the typical definition is extensive. Age related variability can be directly observed in juvenile ALS cases[62,251], while as many as 10% of patients survive longer than a decade with the disease[40,315,329]. Site of symptom onset is commonly heterogeneous, often categorized as ‘limb’ or ‘bulbar’, although other classifications exist including ‘axial’, ‘generalized’ or a combination thereof[266]. Disease comorbidities, like Alzheimer’s (AD) and frontotemporal dementia (FTD), present in as many as 5%, and 15–50%, of ALS patients respectively[93,99,138,173,216]. Interestingly, despite having a shared protein pathology (TDP-43) in ALS and FTD, not all ALS-TDP patients present with FTD, and FTD symptoms can occur without motor neuron degeneration[50]. The consistent observation of ‘atypical’ ALS in a variety of patient populations has driven a shift towards viewing ALS as a neurodegenerative spectrum that shares both mechanistic and phenotypic characteristics with a range of other diseases including AD, FTD, Primary Lateral Sclerosis (PLS), Progressive Muscular Atrophy (PMA), Progressive Bulbar Palsy, and Spinal Muscular Atrophy (SMA), among others[40,84,138,173,321].

1.2.1 Genetics

A genetic component to ALS has long been recognized, despite the fact that hereditary patients only comprise 5–10% of all cases. Early studies investigating genetic linkage in families with histories of ALS found associations with mutations to chromosome 21, subsequently discovered to encode the superoxide dismutase protein, *SOD1*[281,296]. It was immediately recognized that the variability in the location of the genetic mutations was extensive, despite being restricted to the *SOD1* gene locus[281,296]. Confirming this initial observation, more than 150 additional mutations to the *SOD1* protein have been reported since the initial discovery, effectively covering all regions of the 153 amino acid protein[321]. Complicating the interpretation, *SOD1* mutations only present in 12–35% of fALS and 1–3% of sporadic patients[275,321,371], and no correlations are seen between the reduction in enzymatic activity of mutant *SOD1* and clinical parameters like age of onset or disease duration[70]. Moreover, despite these mutations affecting the same protein, the gambit of genetic aberrations to *SOD1* have been shown to yield vastly different disease phenotypes, clearly exemplified in A4V and H46R genotypes, with more than 15 years separating the average life expectancy of the two groups[270].

Further complicating the biochemical and mechanistic interpretation of mutant *SOD1*, numerous additional gene loci have been linked to fALS, namely *VAPB, TARDBP*, *FUS, ATXN2, UNC13A, UBQLN2,* and a hexanucleotide repeat expansion (GGGGCC) on chromosome 9, open reading frame 72 (*C9orf72*) – among others[2,217]. Even still, known ALS genes only present in roughly 50% of fALS patients[371], indicating continued work is needed to better understand the genetic component of ALS. Despite these limitations, novel ALS genes and loci have shed additional light on disease heterogeneity and offer new perspectives on centralized disease mechanisms. Of these genetic associations, the *C9orf72* repeat expansion has emerged as a dominant genetic link, as it is reported in a third of fALS[274] patients and in 5% of the sporadic population[73], but strongly depends on ethnicity[371]. Elucidated by Renton *et al*.[274], a naturally occurring six-nucleotide repeat is observed in the first intron of chromosome 9 open reading frame 72, comprised of four guanine nucleotides followed by two cytosine nucleotides (GGGGCC). The average repeat length in controls was found to be 2–3, regardless of ethnicity, compared to 53 in Finnish patients with fALS, with a bimodal distribution that reliably separates controls (<25 repeats) from ALS cases (>30 repeats). Similar to findings from *SOD1* ALS patients, phenotypic heterogeneity in *C9orf72* associated ALS is extensive, even when considering patients from the same family[303,315]. Further, the length of the repeat expansion and zygosity are not seen to correlate with clinical parameters like onset and survival[73,315] making it difficult to uncover mechanisms driving heterogeneity. Yet, when comparing phenotypes in *SOD1* and *C9orf72* linked ALS, insightful differences emerge in site of onset and cognitive comorbidity[315]. The collective consideration of known genetic associations in the wider neurodegenerative spectrum continues to reinforce the heterogeneous nature of the phenomenon[200,315]. Despite limitations stemming from phenotypic variability, recently discovered genetic associations provide a multitude of platforms to gain additional disease insight[66,231].

1.2.2 Cell and Animal Models

Building on discoveries linking genetic aberrations to ALS, numerous cell and animal models harboring disease causing mutations have been developed, providing immense insight into both mechanistic and phenotypic heterogeneity. Works using cell models harboring the *C9orf72* repeat expansion commonly report the non-traditional translation of the GGGGCC repeat, producing five unique dipeptides (GA, GR, GP, PR, PA)[11,113,117]. These dipeptides have been observed to accumulate in both the nucleus and cytoplasm with localization depending on amino acid composition[113,214]. *In vitro*, arginine containing peptides are often reported in the nucleus, however these findings do not persist in patients, as GR and PR dipeptides are predominantly found in the cytoplasm[113]. Conversely, the absence of GA inclusions in the nucleus of cell models is also observed in patients[214], demonstrating the utility of genetic models to study ALS heterogeneity. In line with localization variability, studies linking dipeptide expression to toxic disease mechanisms find evidence for a plethora of defects, including complex formation affecting protein induced silencing and splicing[75,90], trafficking between the nucleus and cytoplasm[113,142], impairment of protein translation and RNA metabolism[166,186], and proteasome recruitment[132]. Groups leveraging *SOD1* cell models arrive at similar conclusions, implicating excitotoxicity, endoplasmic reticulum stress, mitochondrial dysfunction and oxidative stress, disruptions to cytoplasmic and axonal transport, protein aggregation, disruptions to RNA metabolism and processing, and activation of supporting glial cells in disease progression[140,321]. Thus, mechanisms inhibiting normal function in ALS cell models propagate to a spectrum of possible phenotypes, sometimes occurring in tandem, highlighting a need to better understand the influence in more complex biological systems and importance of identifying biological factors responsible for driving the expressed phenotype.

The development of ALS animal models, including mice, rats, zebrafish, worms, and fruit flies, have spurred additional insight into mechanistic and outcome heterogeneity[37]. The first animal models were transgenic and included mutations to the human *SOD1* protein (G93A, G37R, G85R), and mouse *SOD1* protein (G86R)[48,133,210,262,276,351]. These mice generally presented with motor neuron loss, axonal denervation, protein aggregation, and progressive paralysis, recapitulating many aspects of human ALS – although age and duration variability were extensive given the genetic similarity of the models[48,133,210,262,276,351]. The utilization of mouse models has helped elucidate the role of glial cell involvement, and allowed for the probing of disease phenotypes and mechanisms throughout the disease course, enhancing understanding for region-specific motor neuron loss, protein aggregation, cellular transport defects, and mitochondrial abnormalities[155,282]. Beyond enhancing disease understanding, the G93A *SOD1* mouse model provides an important foundation for preclinical studies, including its use to evaluate efficacy in the two FDA-approved drugs, Riluzole and Edavarone[134,157]. Outside of mouse models, rat models have been leveraged to identify new therapies and offer insights into shared pathological mechanisms between genetic models[227,309], while the more simplistic *Drosophila* model has allowed groups to identify new associations with neurotransmission[294] and disease-associated factors like Ubiquilin 1[137]. Despite the insight gained from animal models, variability in disease mechanisms continues to persist as a dominant theme[100,101,263].

Generally, the predictable onset of disease phenotypes has driven widespread adoption of genetic models as a means to study ALS. And, while much insight has been gained from the use of these models, gaps remain in the translation of these findings to the wider, sporadic patient population and point to the added necessity of directly studying human patients to better understand disease heterogeneity[223,263]. Recent initiatives, like Target ALS and Answer ALS[21], aim to provide researchers with crucial access to large human cohorts, temporal bio-samples, multi-system level measures (genes, epigenetics, transcripts, and proteins), and extensive clinical data – promising to bridge the gap in understanding and translation.

1.2.3 Common Disease Features and the Pursuit of Biomarkers

Despite the general heterogeneity reported from cell and animal models of genetic ALS, common disease features have been identified and offer additional insight into mechanisms underlying the neurodegenerative pathology. In ALS patients, the observation of aggregated TDP-43 in the cytoplasm of neurons is a near ubiquitous (~97%) phenomenon and is considered a pathological hallmark of the disease[266]. TDP-43 aggregates were first reported by Neumann *et al.* and linked to the posttranslational hyperphosphorylation and ubiquitination of the protein, causing mislocalization from the nucleus to the cytoplasm[243]. Since then, it has been shown that truncated C-terminal fragments from the same protein are sufficient to initiate disease pathology[153]. In both cases, the resulting aggregates are observed in the brain and spinal cord of nearly all ALS patients and up to half of patients with a frontotemporal lobar degeneration (FTLD) exclusively, the typically behavioral variant of FTD[42,54,105,243]. Given the reliable formation of TDP-43 inclusions in ALS, and similarity of phenotypes between hereditary and sporadic patients, many groups have examined the utility of the protein as a biomarker[105]. In addition to the lack of specificity for motor neuron degeneration, given the observation in FTLD patients, the detection of pathological forms of TDP-43 (hyperphosphorylated and C-terminal fragments) in patient biofluid has faced multiple challenges[105]. Rooted in these challenges is the ubiquitous cellular expression of nuclear TDP-43, regardless of cell type, and the tightly regulated exchange between the blood-brain barrier limiting opportunity for plasma to be a reservoir for pathological protein[105]. Antibodies designed to target the endmost C-terminus of pathological TDP-43, the optimal motif for detection of all forms of the disease causing protein[105], were still shown to lack specificity for cytoplasmic inclusions in postmortem tissue and remained reactive for functional TDP-43 in the nucleus[105,187]. Similar specificity concerns were encountered in patient cerebrospinal fluid (CSF)[246,308] and plasma[332], yet recent findings suggest promise in the use of exosomes to address specificity concerns for pathological TDP-43[154].

Limitations surrounding the specific detection of pathological TDP-43 have spurred the search for other ALS-specific biomarkers. Supported by advances in technologies, computation, and methodologies, more ‘wholistic’ measures of patient gene expression (transcriptome) and protein abundance (proteome) have offered numerous additional markers associated with ALS neurodegeneration. The central role of TDP-43 in transcriptional regulation has resulted in the identification of truncated forms of *STMN2* and cryptic exon expression in *UNC13A* associated with ALS-TDP pathology[46,179,212,266]. While both transcripts show promise relative to controls, limited expression in patient tissue and variability amongst patients, respectively, represent challenges for their implementation as biomarkers[46,179,212]. Similarly, promising findings occur at the protein level, with, among others, phosphorylated neurofilament heavy chain (pNfh), MCP-1, IL-8, CD14 to S100β and C3 to pNfh ratios, shown to be elevated and transthyretin and cystatin C decreased in ALS patients relative to non-neurological controls – but specificity is generally limited when compared to disease mimics[43,115,184,269,272,312]. Thus, combinations of protein biomarkers may present an opportunity to increase sensitivity for ALS, although longitudinal and multi-site validation are ultimately necessary for successful clinical translation.

1.3 RNA-Sequencing

Since its formal conception[286], RNA-sequencing technologies have benefited from advances in both methodology and computational power, lowering costs and increasing throughput, resulting in widespread adoption of the technology as a means of assessing gene expression levels in an increasingly diverse set of organisms and contexts. In brief, ‘next-generation’ sequencing leverages a variety of methodologies, each with the goal of ‘reading’ the sequence of DNA bases comprising a given oligonucleotide. Reading DNA sequencing from human tissues requires the extraction and purification of DNA, or RNA, and subsequent steps often reflect the type of instrumentation used. In nanopore sequencing, DNA or RNA fragments are passed through a nanoscale opening under high voltage, producing characteristic current responses dependent on the nucleobase, and is performed in parallel to achieve high throughput. Fluorescent-based methodologies, like Illumina's sequencing by synthesis, involve ligating adapter sequences to purified fragments to allow for massively parallel sequencing on their custom designed flow cell. Sequencing of RNA requires additional steps, including reverse transcription to obtain cDNA and allow method compatibility. The dataset utilized throughout this dissertation leverages the latter approach from Illumina, and advantages and disadvantages of this technology are considered in *Chapter 1.3.4*. In all cases, the sequence(s) of DNA is then compared to a reference human genome using a computer to aid in the identification and quantification of all DNA fragments from a single sample – a step known as alignment. During alignment, fragment characteristics like transcript length, GC content (related to hydrogen bonding between base pairs), strandedness (3` or 5`), total number of transcripts observed (reads), and length of the sequencing read become important considerations when leveraging next-generation sequencing technologies as they can confound with and influence the observed transcript copy number. The complete sequencing of available transcripts through next-generation methods is typically referred to as transcriptomics, with the '-ome' or '-omic' suffix indicating wholistic measure of most or many available features at a given system level (e.g. gene, transcript, protein, metabolite, etc.).

1.3.1 Protein-coding RNA

The central dogma of biology teaches the replication of DNA, the transcription of DNA to RNA, and the translation of processed RNA to proteins. This 'traditional' flow of genetic information ultimately results in the synthesis of new proteins from individual amino acid subunits bound to transfer RNAs (tRNA). This process occurs at the ribosome – a large multi-protein complex found on the endoplasmic reticulum – and requires a tRNA displaying a sequencing complementary to the three-nucleotide codon presented on the processed RNA. These nascent proteins frequently undergo further processing in the Golgi apparatus before performing enzymatic, structural, signaling, transport, or regulatory functions within or outside of the cell. Before the RNA can be translated to proteins, a series of modifications occurs to the newly synthesized RNA sequences following transcription, producing messenger RNA (mRNA). These post-transcriptional modifications include the addition of 7-methylguanosine to the 5` end of the RNA sequence (5` cap), the addition of roughly 200 adenine nucleotides (in mammals) forming a poly(A) tail, and the removal of intronic sequences by the spliceosome – a ribonucleoprotein complex – to produce a contiguous segment of exons that can vary in their inclusion in the final mRNA sequence, a phenomenon known as alternative splicing[79,126,258,348].

Of these post-transcriptional modifications, the alternative splicing of RNA transcripts contributes the most to variability in the encoded protein sequence on the mature mRNA and drives differences in the ultimate structure and function of the resulting protein. Thus, while the same DNA sequence may be transcribed to produce RNA copies, the same pre-processed RNA transcript may produce multiple proteins of similar or dissimilar structure and function[126]. This can create challenges during the analysis of large transcriptomic datasets generated by next-generation RNA-sequencing technologies and for the translation of findings from the mRNA level to the protein level[30]. Large databases like Ensembl[139] provide an important resource for linking mRNA splice variants to a single protein in an effort to infer the ultimate function of the encoded protein – although additional work is needed to understand the extent to which splice variants perform the same function and with what efficiency[290,344].

1.3.2 Non-coding RNA

Counterintuitively, given the importance of proteins in the function of both healthy and diseased cells, protein-coding genes comprise fewer than 3% of the entire human genome (<1.5% when considering exons exclusively)[95,236]. Although the remaining ~97% are never translated by the ribosome to create a protein, groups have shown that the transcription of these regions still occurs (over 80%), generally at low levels and only partially active in any given cell type[3,63,95,236]. Importantly, it has been demonstrated that many of these resulting RNA transcripts retain function, participating in a range of processes including epigenetic memory, development, transcriptional regulation, RNA splicing, translation, and RNA metabolism[3,95,222,236]. These transcripts are broadly classified by length and function, and include: (i) microRNAs, (ii) piRNA, (iii) siRNA, (iv) promoter associated RNA, (v) enhancer RNA, (vi) long non-coding RNA (lncRNA) – further subset by genetic locus and includes intergenic, antisense, intronic, and pseudogenes, (vii) small nuclear (snRNA), (viii) small nucleolar RNA (snoRNA), among others[163,221,367]. Although the mechanistic function of individual non-coding transcripts is generally unknown[221], at present, the diversity of expressed transcripts provides insight into the complexity and importance of non-protein coding RNA in the health and maintenance of normal cell function.

While the study and elucidation of non-coding RNA function is ongoing, numerous works provide key insight into the role of non-coding RNA in health and disease. In female mammals, two copies of the X chromosome require additional regulation of chromosomal gene expression, and the well-studied non-coding RNA transcript *XIST* has been shown to play an important role in the necessary inactivation of one copy[108,360]. In mice, the deletion of the opposite strand of *Hand2* results in abnormal heart function and embryonic death, while the deletion of non-coding *Flicr* decreases susceptibility for autoimmunity in non-obese diabetic mice[108,136,364]. In human cancer, the expression non-coding RNA has been shown to influence the progression, metastasis, severity, and therapeutic responsiveness, reflecting a poorly understood mechanism partially responsible for patient phenotype[28,76,280,320]. More relevantly, in the brain, alternative splicing is known to occur at high frequency, and antisense transcript BACE1-AS has been associated with amyloid β production in Alzheimer's disease, while a long non-coding transcript on the long arm of chromosome 22 has been linked to schizophrenia pathogenesis[18,104,358,362,365]. More recently, in ALS, longitudinal measures of serum microRNAs identified transcripts associated with progression[87]. Although collective knowledge reflects early stages of understanding, the relevance and importance of these non-protein coding features continues to grow, and associations with dysregulated RNA metabolism in ALS[321] hint at new insights to be gained from the consideration of these transcripts.

1.3.3 Transposons

Transposons ("Transposable Elements", TEs) are a class of non-coding RNA and these transcripts fall broadly into two categories depending on the mechanism utilized to reinsert itself into the genome[39,225]. DNA transposons are directly 'copy-and-pasted' into a new location in the genome without transcription to RNA and require a transposase enzyme to complete the duplication[39,145]. RNA transposons, also called retrotransposons, use a separate mechanism to insert itself back into the genome that involves transcription to RNA before it is reverse transcribed back to cDNA and integrated into a new location in the genome[33,39]. Transposons are grouped into subclasses with finer resolution related to the mechanism of genome integration, and further divided into 'superfamilies' that include long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), Penelope-like elements (PLEs), Ty1/copia, Ty3/gypsy, and endogenous retroviruses (ERVs) in retrotransposons and hAT and Tc1/mariner in DNA transposons[39]. TE superfamilies are further grouped by phylogeny to produce 'families' that include highly expressed *Alu*, L1, and mammalian-wide interspersed repeats (MIRs) in humans[39,264] that continue to influence the human genome[102]. Transposon integration into the genome occurs infrequently enough that these DNA features can be used as a method to assess mammalian phylogenetics[162,253,300,301]. Despite the low rate of integration, a portion of these non-coding regions remain actively transcribed in humans, spurring the development of reference databases and new computational methods to aid in the wholistic identification and analysis of these RNA features[16,160,359]. Generally, understanding for the function and role that individual TEs play in human health and disease is limited at present, however TE expression is known to be elevated in the brain[94], associated with cryptic exon expression in native genes[94], an extensive regulator of gene expression[39,69], capable of inducing DNA damage[143] and activation of innate immunity[125], and linked to aging[125], cancer[13], and neurodegeneration via TDP-43[183,197,206].

As may be inferred from mechanisms of genome integration, high sequence similarity amongst retrotransposons from the same family coupled with short read lengths pose challenges during the quantification of these transcripts by next-generation RNA-sequencing technologies. Existing computational methods handle this hurdle differently, either aggregating counts by TE family[160] or maintaining gene locus resolution[359], after applying the expectation maximization (EM) algorithm to aid the allocation of non-specific (multi-mapping) reads.

1.3.4 Advantages and Limitations

Following the appropriate preparation steps, next-generation RNA sequencing provides a means to wholistically measure both coding and non-coding transcripts present in biofluid and tissue. When paired with carefully curated databases and computational methods that account for biological variability, a ‘global’ snapshot of active gene expression is obtained. The quantitative nature of sequencing (integer counts) allows for collective study of a multitude of gene processes in response to perturbation or disease. RNA-sequencing has already demonstrated success in elucidating unforeseen or unexpected mechanisms in disease that often arise from complex interactions, associations with non-coding RNA, and alternative and cryptic splicing[46,52,77,212,266,305].

While providing a platform to consider human health and disease using wholistic gene expression, the nascent technology is not without limitations and challenges. Illumina instruments generally leverage short read lengths (150-200 base pairs) and correctly assigning the read to one of many gene isoforms (splice variants) is prone to error[305]. Considering expression at the gene level rather than isoform level can circumvent this challenge, although there is a growing need to reach this resolution[305]. Although not directly related to the instrumentation, methodologies that extract RNA from bulk tissue homogenate encounter additional drawbacks. Given that any tissue biopsy is comprised of a variety of unique cell types, with different regions of the genome actively expressed, counts obtained by bulk tissue RNA-sequencing are confounded with the proportion of cell types present – generally creating unwanted effects when sampling the same region from different patients. Bioinformatic strategies like cellular deconvolution have been developed to help ascertain effects due to varying cell type proportions, however alternative methodologies like single cell or single nucleus RNA-sequencing can avoid bulk tissue bias by quantifying expression individually in each cell. Further, procedures to extract and maintain RNA integrity from tissue vary substantially in practice, and can drive bias during quantification of expression – a particularly potent problem when looking at postmortem tissue[107]. The implementation of RNA Integrity Number (RIN)[292] as a covariate in RNA-sequencing studies can help adjust for effects due to the degradation of transcripts that occur following death, but literature suggests this measure is incomplete[304,345].

The work in this dissertation leverages RNA expression from bulk tissue sequencing. Future works considering heterogeneity from a single-cell perspective should provide additional insight into phenotype contribution from individual cell types, interactions between cells in the context of subtype-specific mechanisms, and the extent to which the observed subtype is driven by bulk tissue effects.

1.4 Bioinformatics

Data generation from bulk RNA-sequencing experiments is extensive (typically gigabytes per run) and the raw sequence data is generally unfriendly to work with as millions of individual reads can make direct interpretation highly challenging. The implementation of computers to aid in the systematic quantification of the transcriptome becomes an inherent necessity, especially as the amount of sequencing information grows increasingly large (i.e. paired-end reads). Numerous software programs and workflows have been developed by the bioinformatics community aimed at providing standardized methods for aligning sequence reads to a reference human genome, allocating multi-mapping reads, and data-driven normalization. Stemming from the complexity of the desired task, groups often weave multiple coding languages into a coherent pipeline with the goal of enhancing computational speed, analytical accuracy, and reducing the required resources. These software packages are typically made available to the wider community through code repositories like Github and are often most easily utilized in a Linux environment. When combined with parallel computing, often in a high-performance computational architecture, the processing and alignment of raw RNA-sequencing data becomes a far more feasible and efficient task, even for large cohorts.

Beyond alignment and quantification of raw sequencing reads, bioinformatics reflects a wide variety of mathematical, statistical, and computational concepts that are often adapted to better address known biological phenomenon. For example, when comparing transcript expression, the use of the negative binomial distribution over a normal distribution reflects the intuitive assumption that gene expression is non-negative and generally expected to be low relative to a few active processes. In other cases, the expectation maximization algorithm can be leveraged to aid the assignment of multi-mapping transcript sequences, while the cox regression framework serves as the basis for multivariate survival analysis. Elsewhere, sequence alignment algorithms allow for imperfect matches to account for single nucleotide polymorphisms and other naturally occurring and expected genetic variations relative to the reference genome. Thus, many concepts leveraged throughout this dissertation have methodological origins in other disciplines and reflect a substantial amount of work from the bioinformatics community to extend their usage to biological cases.

1.4.1 Databases and Data Repositories

Although data repositories are not bioinformatic methods themselves, they certainly support the field of bioinformatics and are worth discussing. Government-led intuitions like the National Center for Biotechnology Information (NCBI) in the United

States or the European Bioinformatics Institute (EMBL-EBI) are responsible for storing and managing large biological datasets that are becoming increasingly commonplace. These repositories accept a wide variety of raw data formats and aim to provide the wider scientific community with straightforward access to these files in order to answer new biological questions. Available information is typically grouped by system level (e.g. transcripts, proteins, metabolites), organism, instrumentation, and methodology and currently exceeds 10 petabytes in NCBI's Sequencing Read Archive (SRA) alone. More recent repositories like EBI's PRIDE Archive (a part of ProteomeXchange[334]) provides mass spectral data from proteomics, while the MetaboLights repository offers the same raw spectral information but for metabolites. Beyond data management, these institutions also provide the bioinformatics community with a suite of tools and curated databases to aid in the analysis of complex genomic (and non-genomic) information[38,139,228,249].

Outside of these dedicated bioinformatics institutions, numerous initiatives have worked to expand access to -omics data and develop databases to aid in analysis, including: the National Cancer Institute's The Cancer Genome Atlas (TCGA) Program, the Genetic Information Research Institute's Repbase[16], the ENCODE consortium's GENCODE database[95], the Canadian-led human metabolome database[349], sequencing archives like the DNA DataBank of Japan[180], European Nucleotide Archive[194], and GenBank[27], the National Institute of Health's (NIH) Genotype-Tissue Expression Program (GTEx)[128], the University of California Santa Cruz (UCSC) Genome Browser[168], non-profit Sage Bionetworks' Synapse database (among other resources), the University of California San Diego's Metabolomics Workbench[311], and pathway databases that synthesize findings from hundreds of individual publications, like the Kyoto Encyclopedia for Genes and Genomes (KEGG)[165], Reactome[121], and BioCyc[57]. Furthermore, groups have improved file transfer protocols and storage to enhance download speed and security, which include software like IBM Aspera and University of Chicago's Globus[111]. Collectively, these critical and fundamental undertakings serve as the bioinformatics backbone, directly accelerating new discoveries, enhancing biological insight, and supporting the proposal of novel questions or methodologies.

1.4.2 RNA-Sequencing Alignment and Quantification

Raw sequencing files obtained from Illumina instrumentation require multiple processing steps before quantitative information can be obtained. Pre-processing usually involves the trimming of Illumina adapter sequences, accomplished by software like Trimmomatic[35], but are not always necessary depending on the downstream alignment method used[86]. The first major step is referred to as alignment, and involves comparing all sequences of nucleotides against a reference human genome, offered by a number of databases[95,139,249]. Matching sequences are assigned as 'counts' to the corresponding gene or non-coding transcript, and genomic depth can vary depending on the reference genome provided. Multiple software pipelines have been developed to accomplish alignment and include, STAR[86], Bowtie2[189], Tophat2[174], and GSNAP[352], among others. Comparative studies report a substantial variability in alignment performance, with STAR frequently reported on the upper end of alignment procedures[19,96]. Following alignment, integer counts are obtained for genes comprising the reference genome, although a minority of sequences typically show ambiguity during mapping and cannot be assigned to a single feature specifically, known as multi-mapping reads. To help with the data-driven allocation of these reads, the expectation maximization algorithm can be applied to derive count distributions adjusted for parent transcript, length, start position, and strandedness that are then used to determine the probability that a specific fragment originates from a given transcript[196]. The probabilities are then used to assign ambiguous transcripts as a fraction of the total number of reads available for allocation[196]. Following the utilization of RSEM[196], count values may be rounded to integers to allow for compatibility with downstream applications like differential expression[208].

1.4.3 Correcting for Multiple Hypothesis Testing

Massively parallel sequencing enables the quantification of gene expression for most available transcripts, and is not restricted to protein coding regions, often yielding tens of thousands of unique features. Individual hypothesis testing becomes an immediate pitfall, and stems from the rather arbitrary definition of statistical significance. Using an alpha value of 0.05, results are determined to be 'statistically significant' if the probability of observing such an outcome is less than 1 in 20 under the null hypothesis, leading to an acceptance of the alternative hypothesis. Thus, when considering expression differences in 20 different genes, one is likely to be observed as statistically significant entirely by chance and unrelated to the hypothesis being tested – by definition. To correct for the multiple hypothesis testing problem, numerous statistical methods have been developed centered around controlling type I errors, also known as false positives. The most stringent, known as the Bonferroni[36] correction, simply divides the chosen alpha value (typically 0.05) by the number of hypothesis tests being performed, providing the corrected alpha value to test against. The Bonferroni correction, along with the Sidak correction[295], Holm-Bonferroni method[150], and Hochberg's procedure[149] aim to control the family-wise error rate (FWER) or the probability that any given rejected null hypothesis is falsely reported. A different approach, often called the false discovery rate (FDR) developed by Benjamini and Hochberg[26], involves controlling the proportion of all rejected null hypotheses that are falsely reported, ensuring that some are false positives. Typical false discovery rates are 5% or 10%. Although this procedure generally leads to a higher number of false positives compared to FWER controlling methods, the FDR is often preferred in large exploratory studies where follow-up experiments may be performed to confirm the initial findings. Other strategies have been developed to address the problem of multiple hypothesis comparisons but are not utilized in this dissertation.

1.4.4 Clustering

Depending on the research question, interesting biological insight can often be obtained from grouping similar samples or observations together and probing the differences between groups, with notable successes in cancer[72,129,299]. Broadly, clustering algorithms aim to group similar observations into 'clusters' which 'look' more similar to observations within the cluster as compared to other clusters. Clustering can be susceptible to outliers and sensitive to covariates with large effects, such as sex, and expression data typically requires normalization (i.e. regularized log or variance stabilizing transformation, VST) and filtering before interesting clusters can be extracted[208]. Numerous methods have been developed to accomplish this task, ranging in complexity, measures of distance, and approach to assigning clusters. Hierarchical and K-means clustering calculate the distance between points and assign centroids (cluster locations) in a way that minimizes the intra-cluster sum of squares[110,241,260]. For hierarchical methods, alternative clustering schemes have been developed including the minimization of the max or average distance between points and density-based cluster assignment[240,260]. Consensus clustering builds on hierarchical or k-means algorithms by iteratively repeating the clustering process to obtain robust sample labels and parameter estimates[235,347].

Alternatively, matrix factorization can achieve the goals of clustering by decomposing a bulk signal into individual components, reducing the dimensionality of the parameter space. Principal component analysis (PCA) is a well-known matrix factorization technique and involves the decomposition of the covariance matrix to identify orthogonal components (eigenvectors), scaled by the eigenvalue, that maximizes the variance explained[350]. Groups within the sample population can typically be visualized by plotting the first few components capturing the largest variance. Another technique, non-negative matrix factorization (NMF), can be applied to matrices that are strictly non-negative, originally developed for image processing[192,252]. With the understanding that gene expression meets this criterion of non-negativity, the approach has been adapted for biological clustering with numerous additional variants developed[49,148,175,255]. The constraint on negative observations allows for part-based representations in the resulting clusters, rather than wholistic ones (as in PCA), because additive combinations are strictly possible[192]. With gene expression data, part-based representation often yields more intuitive components and aids cluster identification given wholistic components (as in PCA) encoding negative expression are biologically unintelligible. Further differing from PCA, NMF provides probabilities that a given sample belonging to a specific cluster, known as soft clustering[116]. Using the highest cluster probability, a single group label can be assigned to a sample, supporting hard clustering. Finally, gene scores[175] from NMF clustering provide additional information regarding cluster specificity, aiding feature selection for downstream analyses. Relatedly, but outside the scope of this dissertation, non-linear dimensionality reduction techniques like t-SNE[330] and UMAP[22] offer additional clustering frameworks for high dimensional data and are commonly leveraged in single cell sequencing datasets for cell type identification, among other applications[289].

1.4.5 Gene Enrichment

Following clustering analysis, and others like differential expression[208] and gene co-expression[188], one typically arrives at a list of important genes and transcripts related to the disease, group, or biological perturbation being studied. Gene enrichment encompasses a family of methods designed to provide additional information about the pathways and cellular contexts that genes of interest are known to participate in. Databases like KEGG[165] and Reactome[121] enable gene enrichment analyses by maintaining curated gene sets for known pathways to compare the interesting features against. Additionally, the Gene Ontology Consortium has offered a wealth of gene annotations that can be leveraged during enrichment analysis and are separated into three categories, (i) biological processes, (ii) molecular functions, and (iii) cellular compartments, offering both functional and spatial insight into over a million coding and non-coding transcripts across >5,000 different species[118]. Building on these important knowledge resources, multiple procedures and strategies for enrichment have been developed. Hypergeometric enrichment analysis determines the probability of observing the provided genes within the numerous lists offered by pathway and gene ontology databases, using the Fisher exact test, which assumes a binomial distribution[59,277]. Alternatively, Gene Set Enrichment Analysis (GSEA) is an important and popular bioinformatics software which performs gene enrichment analysis by first ranking genes by correlating expression data with two phenotype levels[310] (e.g. disease vs control) – often performed in a pairwise manner for cases with >2 phenotypes. After obtaining a ranked set of genes, an enrichment score is calculated using a running-sum statistic with magnitude determine by the rank, which is followed by randomly resampling the phenotype for empirical estimation of the *p*-value[310]. Normalization is performed by dividing the true enrichment score by mean enrichment score from resampling, which is subsequently used to control the rate of false positives and in the calculation of FDR-adjusted *p*-values[310]. Challenges with enrichment analysis typically stem from the high degree of gene overlap between lists leading to an undesirable number of false positives, and more recently developed approaches work to address this hurdle[123,297]. Collectively, these methods, and others, enable a deeper biological understanding for the genes of interest associated with the research question, and directly complement wholistic analytical strategies like RNA-sequencing where a relatively unbiased snapshot of transcription can provide unexpected insight into diseases processes and associated mechanisms.

1.4.6 Bootstrapping

Bootstrapping is a statistical method that involves the resampling of observations with replacement. The technique is particularly helpful for estimating parameter distributions to assign confidence intervals, errors, variances, and probabilities. For example, Patel *et al*. leverage bootstrapping and glioblastoma subtype marker genes to establish confidence intervals on the expected expression for each phenotype following single cell RNA-sequencing, and utilize the derived distribution to show that individual cells can effectively express hybrid disease states[259]. The use of bootstrapping in bioinformatics generally stems from the complex nature or structure of large '-omic' datasets, often making it difficult to determine parameter values directly, and the relative time and cost effectiveness of estimation through random resampling.

1.4.7 Cell Deconvolution

Cellular deconvolution is a bioinformatic method that aims to decompose bulk tissue expression data into the individual contribution from each cell type provided in a reference gene expression dataset[244,342], although reference-free methods have also been developed[268,341]. Cell deconvolution derives its name from linear algebra and signal processing concepts, namely signal deconvolution, and can help determine if the effects or phenomenon observed in bulk tissue expression may be explained by cell type composition and differences therein. With the understanding that the bulk expression matrix (B) is the observation of the proportion of different cell types (X) that has been convolved with the individual expression profiles of each cell type (A), linear algebra can be used to extract the cell type percentages needed to produce the bulk expression profile. In reference-based approaches, single cell RNA-sequencing data is typically used to define the individual expression. Although powerful and informative, cell deconvolution analysis can only provide estimates and can be biased and limited by currently available reference datasets, given relevant covariates like age, genetic background, exposure, or lifestyle may not align with the cohort being studied. Single cell RNA-sequencing approaches offer more precise determination of cell type and should continue to provide additional insight into the individual contributions to the overall disease phenotype[135,171].

1.4.8 Survival Analysis

Survival analysis serves as a powerful method to associate treatments, interventions, subtypes, and diseases with differences in patient outcomes often measured in, or reported by, the clinic[238]. In the simplest case, the analysis aims to determine the influence of a single variable on patient survival (alive vs dead at time, *t*). Using the statistical framework provided by the log rank test and chi-squared distribution, survival analysis involves determining the probability that an event (death, in the case of survival) occurs at a given point in time, with the null hypothesis assuming no difference between two or more groups[124,167]. During analysis, at each time point where an event occurs, the observed number of events is compared against the expected number – for each group – and used to calculate the survival probability and log-rank test statistic[124]. The resulting plot presents the length of time along the x-axis and the survival probability along the y-axis, with a step function joining points, given the proportion of surviving patients does not change between event times.

To address commonly encountered scenarios resulting in missing data, a technique known as censoring is often applied, where ‘left censoring’ involves the exclusion of patients where the timing of initiating event (e.g. symptom onset) is unknown and ‘right censoring’ involves the exclusion of patients with unexpected study dropout, unrelated to the question or disease being examined, or does not reach some event or milestone of interest (e.g. death)[195]. In the R programming language (The R Foundation for Statistical Computing, Vienna, Austria), the ‘survival’ library[325] implements a ‘status’ argument to indicate whether censoring is to be applied, which allows use of observational information from censored patients without influencing the determination of the survival probability at any given point in time. In cohorts where patient observations are exclusively postmortem, the model is modified to reflect all patients experiencing the death event, eliminating the need for right censoring, but may still require left censoring depending on the informational availability of the initiating event. More complicated cases are also supported, where multiple states are possible (e.g. healthy, diseased, death)[325]. Yet, despite the capabilities of survival analysis, the log rank test is limited in its ability to account for the effects of more than one covariate.

1.4.9 Cox Regression

The multivariate extension of survival analysis was established through a regression statistical framework developed by Dr. David Cox and extended by others[5,80,324,325]. Assuming that the hazard (age-specific failure rate) is proportional between subjects or observations over the timeframe being considered, the resulting hazard ratio becomes constant, and it is possible to simultaneously estimate the effects of all model parameters[80]. The method is commonly reported as the Cox proportional hazard model, reflecting the assumption necessary to allow for multivariate parameter estimation[12,80]. The model incorporates a baseline hazard, $\lambda_0(t)$, which is expected to be equivalent for all patients, and regression parameters are exponentiated. A reference level is specified to allow calculation of the hazard ratio relative to the reference. The hazard ratio is then obtained by exponentiating the parameter coefficients, with hazards > 1 indicating increased risk of death relative to the reference level, and < 1 showing decreased relative risk of death.

Within the proportional hazard framework, increasingly powerful computational strategies have been developed over time by a handful of individuals to extend the types of valid cases, with statistical research ongoing[12,323,324,325]. The inclusion of a single random effect term, often to account for patient-specific differences in the baseline hazard, can be a powerful way to adjust for fundamental individualistic differences in the failure rate and can be identified in peer-reviewed literature as a 'frailty' or Cox Mixed Effects models[12,323]. The latter is considered more robust by the R software developer and a move away from the 'frailty' framework has been recommended. For models that seek to incorporate more than one random effect term, the mixed effects approach becomes the only option. Alternative computational strategies can be leveraged with the proportional hazard framework to account for patient-specific effects, including the use of the 'cluster()' argument which more closely resembles generalized estimating equations. Quantitative testing of the proportional hazard assumption using the score test is offered by the R 'survival' software package[127,325] but can be qualitatively assessed through plots of the scaled martingale and Schoenfeld residuals or the log, –log transformed hazard function against logarithmic time[127,178]. Ultimately, Cox regression for multivariate survival represents one of the better strategies for isolating the effects of a single treatment, phenotype, or covariate after correcting for other relevant drivers of patient survival like age and sex[25,164].

1.4.10 Weighted Gene Co-Expression Network Analysis (WGCNA)

Driven by major advances in the molecular and cellular biology, an understanding for the complex, redundant, and contrasting interactions between genes has been established. Weighted Gene Co-Expression Network Analysis (WGCNA) builds upon this foundation by working to identify correlated gene subsets using transformed expression data, often from next-generation sequencing platforms or microarrays[188]. The basis of WGCNA centers around network adjacency measures and the frequency of expected connectivity between genes in a network with scale-free topology. Two approaches for the calculation of weighted gene correlation networks are available, unsigned and signed, where an unsigned network treats positive and negative correlation the same during determination of adjacency, whereas a signed network does not. Signed networks are advantageous for uncovering biologically relevant or cell type gene clusters, but won't group features with negative correlations[188]. After determining the adjacency for all gene pairings, the connectivity is then defined as the row sum of the adjacency matrix. The adjacency measure incorporates a power, β, which helps to satisfy the underlying assumption of scale-free topology and is assessed through the R^2 value between the log of the connectivity and frequency[188]. Scale-free topology refers to the fraction of nodes with degree k that follows the power law k^{-a}, which is an important feature for biological networks where many genes are expected to be uncorrelated with each other and unrelated to the disease or perturbation[17,188].

Once the soft thresholding power, β, has been selected to roughly satisfy the assumption of scale-free topology, the topological overlap matrix[271] is calculated from the adjacency matrix and hierarchical clustering is then used to define gene clusters[188]. Subsequently, a minimum module size is typically specified and the dendrogram is then ‘cut’ (regrouped) to ensure module size parameter is met[188]. A dissimilarity threshold is also specified and used to filter weakly connected modules[188]. Module (cluster) eigengenes are then calculated by singular value decomposition after standardizing expression and represent a weighted average of all genes comprising the eigengene[188]. Calculating eigengenes offers a number of advantages, including the ability to relate modules to each other and relevant clinical traits, as well as define module membership (co-expression) of individual features. When paired with enrichment strategies and pathway databases, the biological relevance of co-expressed gene subsets is often uncovered and can be associated with phenotypic outcomes of interest – which often represents a central goal of the bioinformatic analysis.

1.4.11 Differential Expression

Differential expression, as the name implies, aims to determine if a particular gene or transcript is significantly upregulated or downregulated relative to a specified reference level. Most commonly, controls are provided as the reference level, and differentially expressed genes are identified from an experimental or disease cohort. Accounting for the inherent nature of gene expression data (i.e. non-negative and large dynamic range), differential expression often assumes a non-normal distribution, either a Poisson[278] or negative binomial[4,208,328] which capture the non-negativity in expression observations, but is not strictly necessary[302].

Of these methods, DESeq2[208] is an R software package that has emerged as one of the more optimal methods for count normalization and identification of differentially expressed genes between samples or groups[78,85,370]. DESeq2 first works to fit a generalized linear model (GLM) to each gene and estimates the dispersion (i.e. variability between replicates or observations) by maximum likelihood estimation[208]. The mean expression of the *i*th gene is scaled by a sample-dependent normalization factor, called size factors, which accounts for differences in sequencing depth between samples[208]. A curve is then fit to define the overall dispersion trend as a function of the mean of normalized counts[208]. The fit is then used to define a prior distribution which is leveraged in a second round of estimation to provide the maximum *a posteriori* (MAP) as the final count estimate[208]. The GLM statistical framework offered by DESeq2 is advantageous when more than one factor or variable is known to influence gene expression, which is often the case in biological experiments. Covariates with more than two levels can be sequentially contrasted in a pairwise manner to obtain a complete view of differentially expressed genes between levels or groups. DESeq2 implements the FDR multiple hypothesis correction, and sets the rate at 10% by default[208]. Differential expression is often necessary when working with large transcriptomic datasets generated by RNA-sequencing platforms, and can be paired with enrichment and classification analyses to elucidate significant biological insight.

1.4.12 Supervised Classification

Classification ('learning') algorithms aim to predict or assign a categorical label to an observation in which the true label may or may not be known. Classification methods fall into two categories, supervised and unsupervised, where the difference lies in knowledge of the true cluster labels prior to running the analysis. Unsupervised classification strategies work to extract labels from patterns in the observed data, with some approaches previously discussed in *Chapter 1.4.4*. Supervised classification requires knowledge of the sample labels prior to the analysis and generally performs iterative optimization of the underlying mathematical structure (decision function) to minimize or maximize a relevant performance metric – often referred to as "machine learning". The resulting mathematical structure needed to provide the predicted label from a set of input observations is called the classifier. The development of a supervised classifier generally starts with randomly splitting a discovery cohort into a training and testing dataset, with allocations to the training cohort often between 60-80%, and repeating the process *k* times for cross validation. The training cohort is used to optimize the classifier, which is subsequently applied to the test cohort to estimate performance. Classifier performance is almost always overestimated in the test cohort due to the fact that all training and testing observations come from the same distribution. To address this common problem, the classifier is tested on an independent 'validation' cohort, often with different noise distributions and batch effects by default, which provides better estimates for future classifier performance. Supervised classifier performance is quantified in terms of true positives, true negatives, false positives, and false negatives which are derived from the true and predicted categorical labels. From the frequency of each classification event, the precision (also known as positive predictive value) can be defined as the number of true positives over the number of true positives and false positives. Similarly, the recall (also known as sensitivity) can be defined as the number of true positives over the number of true positives and false negatives. The precision and recall are often leveraged to produce the F1 score, a symmetric representation of both precision and recall, which ranges between 0 and 1, with 1 indicating perfect classification accuracy. Generally, classifier accuracy can be improved by increasing the number of observations used to train the classifier but may also lead to model overfitting in cases where the underlying batch effects or noise distribution is constant. Often, the class or label being predicted has more than two levels, a problem known as multi-class classification. Strategies to perform multi-class classification typically involve constructing classifiers for each 'one-versus-rest' or 'one-versus-one' combination in the class labels.

Supervised classification algorithms are extensive, ranging from linear models like lasso and ridge regression to non-linear procedures like the multilayer perceptron (also known as neural networks) and decision trees. Software to implement classification algorithms is primarily available through the Python programming language and include TensorFlow, Keras, and Scikit-learn[1,260]. Often, the optimal algorithm to achieve the highest classification accuracy is not clear or easily determined. To circumvent this challenge and aid in selection of the best mathematical framework for the specific classification problem, multiple classifiers can be developed for the same dataset and compared using precision, recall, and F1 scores[248]. This dissertation leverages scikit-learn's linear discriminate analysis (LDA), support vector machines (SVM), decision trees (random forest; RF), nearest neighbors (KNN), and multilayer perceptrons (MLP)[260] – with other algorithms falling outside the scope of this work. In LDA classification, a linear decision boundary is defined by fitting class conditional densities (Gaussian) to the data using Bayesian statistics[260]. In SVM classification, a kernel function is specified and used to define decision boundaries in the high dimensional space from a subset of training points[260]. In RF classification, simple binary (true/false) decisions are inferred from the data to produce a decision tree[260]. The process is repeated using bootstrapping to produce the 'forest', with decision metrics defined by the mean decrease in (Gini) impurity following the split of a node. In KNN classification, a simple majority vote of the *k* nearest neighbors is used to assign the class or label to the given observation, where a larger *k* value is generally less susceptible to noise but produces less distinct boundaries between classes[260]. Lastly, in MLP classification, input features are used to approximate a non-linear decision function by passing linearly weighted information to each neuron comprising one or more hidden layers before transformation by a non-linear activation function like the hyperbolic tangent[260]. The complex structure is then passed to an output layer where the sample label can then be assigned via class probabilities. Guidelines for selecting the number of neurons in each layer and number of hidden layers are scarce, although a low number of hidden layers (≤ 3) is commonly advised[203,211,306]. A wealth of additional information is available at: https://scikit-learn.org/stable/[260].

1.5 Systems Biology as a Unifying Framework

The interdisciplinary space intersecting molecular biology, next-generation sequencing and mass spectrometry instrumentation (at present), statistics, and computation is commonly referred to as systems biology. While many research areas incorporate interdisciplinary perspectives, systems biology differs from traditional research strategies in that there are limited biases or preconceived hypotheses that guide and influence the measured biological variables (i.e. wholistic versus reductionist). The lack of a specific hypothesis to be tested leads to natural pitfalls, including ideas covered in *Chapter 1.4.3*, but can offer advantages for the interpretation of complex biological problems where the sum of the parts does not equal the whole (i.e. emergent properties). Evidence for the existence of emergent properties in biology is extensive and ranges from the macroscopic pattern of a flock of birds or plant fractals to bi-stable and oscillatory gene networks[61,91,265,368].

Two broad perspectives on the field of systems biology exist, which are occasionally referred to as 'top-down' and 'bottom-up'. In a bottom-up approach, a network of interacting features is first defined and subsequently used to generate mathematical models which aim to predict cell behavior – with confirmation often in a synthetic *in vitro* setting[61,265,368]. Bottom-up systems biology has consistently demonstrated that emergent properties in biological systems are present but are typically limited in the number of features that can be considered[61,265,368]. Conversely, driven by advancements across disciplines, a top-down systems biology perspective works to measure many or most features from one informational level (genomics, epigenomics, transcriptomics, proteomics, metabolomics, etc.) or more than one level (multi-omics)[363]. Top-down systems biology relies more heavily on instrumentation, methodology, and statistics and study interpretation can be severely hindered by batch effects, known or unknown methodological bias, and selection of reference data. Despite these limitations, increasingly powerful bioinformatic strategies help correct for quantification and other analytical bias in top-down studies – enabling the genuine identification of novel or unforeseen effects[182,322]. Although top-down approaches are generally not interested in directly identifying emergent properties, the impetus for formulating research questions in this wholistic framework remains the same.

The work in this dissertation leverages a top-down systems biology perspective, which encompasses the use of Illumina next-generation RNA-sequencing for quantification of ALS patient transcriptomes from the central nervous system (CNS). Numerous bioinformatic strategies are leveraged to correct for, or assess, technical and analytical bias. The use of a top-down framework broadly expands upon insight gained from genetic models of ALS and naturally complements the proposed goal, namely to elucidate unknown and complex drivers of ALS patient heterogeneity.

1.5.1 Cellular Heterogeneity

Studies leveraging a systems biology approach have already demonstrated success in the discovery of novel and emergent neurodegenerative disease phenotypes at the cellular level. Using single-cell RNA-sequencing, groups have shown that the activation and phenotypic transition of both microglia and astrocytes occurs over time in Alzheimer's disease[135,171]. Keren-Shaul *et al*. report a subpopulation of microglia following t-SNE non-linear clustering analysis[171]. Differential expression analysis was then applied and led to the identification of marker genes defining the native (homeostatic) and disease-associated states, and co-localization was subsequently confirmed *ex vivo* by fluorescent imaging[171]. Through enrichment analysis, the authors find elevated expression of genes participating in phagocytic pathways in the disease-associated state, again verified by imaging[171]. The proportion of microglia expressing the disease-associated state was found to increase throughout the duration of the disease, lending strength to claims of a neurodegenerative associated phenotype. Similarly, Habib *et al*. leverage many of the same systems biology strategies, leading to the identification of a subpopulation of disease-associated astrocytes in Alzheimer's disease[135]. The disease-associated cell phenotype was found to co-localize with amyloid β plaques in the brains of mice modeling AD and the proportion of astrocytes expressing the disease-associated phenotype increased with age and disease severity[135]. In both studies, the consideration of the entire cellular transcriptome enabled the authors to recover the rare and emergent disease-associated phenotype with limited prior knowledge for relevant factors influencing progression. Supporting the validity of wholistic discovery, these disease-associated phenotypes have been independently observed across the neurodegenerative disease spectrum, including AD, ALS, and FTD[44,83]. Further emphasizing the broad advantages of applying a wholistic perspective to complex biological problems, single cell transcriptome profiling has revealed intra-tumoral heterogeneity in cancer patients and demonstrated the expression of hybrid phenotypes at a single cell level[259].

1.5.2 Patient Heterogeneity

The application of systems biology to stratify ALS patients and characterize poorly understood disease heterogeneity is not unique to this work. Before the development and widespread adoption of next-generation sequencing technologies, DNA microarrays showed early promise for transcriptome profiling and elucidating features distinguishing ALS patients from controls[191,215]. However, the formal consideration of disease heterogeneity within an ALS population leading to the discovery of phenotypic subgroups was not realized for another ~decade[8]. In this important work, the authors identify two ALS phenotypes in a sporadic population (*n*=31) following expression profiling by microarray and application of hierarchical clustering[8,241]. The authors termed the two subtypes SALS1 and SALS2, where the primary differentiating phenotypes involved synaptic signaling, cytoskeletal organization, and neuroinflammation[8]. The same group later combines transcript expression data with copy number variants to find additional support for molecularly distinct phenotypes distinguishing the two SALS subtypes[237]. Building on these findings and providing an important foundation for this dissertation, Tam *et al.* leverage more advanced sequencing and clustering methodologies to identify three subtypes from the frontal and motor cortex of 77 ALS patients[317]. They re-capture many of the subtype-specific phenotypes observed in SALS1 and SALS2 patients and uncover a third subtype defined by the expression of non-coding transposable elements and a hyperactive TDP-43 pathology[160,317]. In addition to the characterization of molecular heterogeneity, Tam *et al*. apply eCLIP-seq to identify TDP-43 binding sites on an expansive set of non-protein coding transcripts including LINE, SINE, and LTR transposable elements, introns, antisense, and long non-coding RNA, among others[317] – providing a direct link to the pathology observed in transcriptome profiling. Despite the importance of these works, no associations had been found with the presentation of ALS subtype and heterogeneity in clinical outcomes like age of onset or survival – representing the primary motivation for this work.

Outside of sequencing-based approaches, but within the field of systems biology, multiple recently published studies leverage proteomics to stratify a neurodegenerative cohort and identify phenotypic subtypes in ALS[337] and Alzheimer's[327]. In the ALS cohort, 59 candidate protein markers from CSF were found to stratify patients with fast (>1/month) and slow (<0.5/month) rates of disease progression (*n*=11), defined by the change in ALSFRS-R per month[337]. Following enrichment, the candidate protein markers were primarily upregulated in fast progressors and associated with neuroinflammation, and continued to show predictive power in an independent validation cohort[337]. In the Alzheimer's cohort, over 1,000 AD-associated proteins identified from >400 individuals with AD and 187 controls were selected for clustering analysis by non-negative matrix factorization[192,327]. Following enrichment, and supplemented by MRI imaging, the authors define five molecular subtypes of AD at the protein level distinguished by: (i) hyperplasticity (32.7%), (ii) immune activation (29.6%), (iii) RNA dysregulation (5.7%),

(iv) choroid plexus dysfunction (18.6%), and (v) blood-brain barrier dysfunction (13.4%)[327]. Subtype-specific AD phenotypes were identified in multiple independent datasets and linked to differences in survival times, dementia comorbidity, sex, and age at onset[327]. The observation of distinct neurodegenerative subtypes at multiple system levels (i.e. transcripts and proteins) lends strength to their biological relevance and points to a need to better map the propagation of phenotypes from the transcript to protein level.

1.6 Outline of the Dissertation

In Chapter 2, transcript alignment and quantification are performed enabling unsupervised clustering and enrichment for the discovery of three distinct molecular subtypes in the postmortem cortex and spinal cord. Methodological bias is partially addressed through the cell deconvolution analysis. Utilizing bootstrapping, the expression of hybrid subtypes is demonstrated and concordance of intra-patient subtype presentation between the cortex and spinal cord is shown to be statistically significant.

In Chapter 3, survival analysis and Cox regression are leveraged in an effort to link ALS subtypes to clinical variability, identifying weak associations with disease duration and age at onset. Supporting findings from univariate and multivariate survival, WGCNA uncovers subtype-specific gene subsets significantly correlated with survival and age of onset. Results broadly agree between the postmortem cortex and spinal cord.

In Chapter 4, application of differential expression analysis identifies seven marker genes consistently upregulated in the postmortem cortex and spinal cord of one of the subtypes. Marker genes are then utilized to develop numerous supervised classifiers –

generally showing clinically useful stratification accuracies in three different holdout cohorts.

In Chapter 5, the findings and implications from Chapters 2–4 are discussed and future directions are proposed.

Text and figures in this PhD dissertation are adapted from a previously published article and one currently *in preparation*:

- Chapters 2, 3, and 4: Reprinted with permission from: **Eshima J**, O'Connor SA, Marschall E, NYGC ALS Consortium, Bowser R, Plaisier CL, Smith BS. Molecular subtypes of ALS are associated with differences in patient prognosis. *Nature Communications*, **14**, 95 (2023).
- Chapters 2, 3, and 4: **Eshima J**, Pennington TR, Choudhury R, Garcia JM, Fricks J, Smith BS. (2024). Elevated expression of *B4GALT6*, *GABRA1*, *GAD2*, *GLRA3*, *HTR2A*, *PCSK1*, and *SLC17A6* are postmortem markers for the ALS-Ox subtype. [*in preparation*].

Chapter 2

UNSUPERVISED CLUSTERING IDENTIFIES THREE ALS SUBTYPES IN THE POSTMORTEM CORTEX AND SPINAL CORD

2.1 Introduction

Amyotrophic Lateral Sclerosis is a neurodegenerative disease with poorly understood clinical heterogeneity, underscored by significant differences in patient age at onset, symptom progression, therapeutic response, disease duration, and comorbidity presentation. The effects of this gap in understanding can be seen in long clinical diagnostic timelines, hampered by an absence of disease specific biomarkers, subjective scoring metrics, and presentation of symptoms that overlap with other motor neuron disorders early in the disease course, often leading to misdiagnosis[40,233,254]. The lack of diagnostic and prognostic biomarkers has led to the utilization of a patient classification system based on the site of onset and symptom presentation, which poorly predicts differences in patient pathology, survival, treatment responsiveness, and symptom progression[55,339]. As a consequence, challenges experienced in the clinic – including in the design of clinical trials and subsequent lack of effective ALS treatments – are directly linked to the underlying disease heterogeneity.

More broadly, recent efforts have been directed towards identifying the phenotypes and mechanisms driving clinical heterogeneity in neurodegeneration. In Alzheimer's patients, neuroimaging-derived subtypes demonstrated differences in clinical presentation, survival, age of onset, rate of progression, and age of death, providing critical new insight into disease heterogeneity[346]. Further, associations between age of onset, rate of progression, and symptom presentation in Alzheimer's disease has led to the implementation of a recently developed clinical classification system (early vs late onset with additional subvariants) that better predicts the individuals' disease course[230]. More directly, in the context of ALS, one group has recently developed a predictive model to stratify patients and inform prognosis, using patient-derived clinical information[343]. Collectively, these efforts demonstrate advantages of stratifying patients within these broad neurodegenerative disease spectrums and reflect a growing clinical need to enable this more personalized perspective.

Strategies to assess the molecular foundation of ALS heterogeneity have primarily applied '-omic' methodologies in combination with unsupervised clustering for disease subtype discovery[8,237,317]. Tam *et al*. established an important foundation for this chapter, using frontal and motor postmortem cortex transcriptomics to stratify a cohort of 77 ALS patients into three distinct subtypes[317]. They further demonstrate the direct interplay between TDP-43 and transposable elements using eCLIP-seq, providing key insight into the pathological role of transposable elements in ALS, given the near ubiquitous nature of TDP-43 cellular inclusions (~97%)[202,266,317].

In the second chapter of this dissertation, the publicly available ALS cohort derived from the NYGC ALS Consortium and Target ALS groups was stratified, leading to the identification of three molecular subtypes from > 200 individuals with ALS. Features distinguishing each subtype were subject to enrichment, offering a wholistic framework to better understand pathological differences between the ALS subtypes. Stratification is performed on the postmortem cortex and spinal cord separately, but converge toward similar pathological themes regardless of the central nervous system region considered. Intra-patient presentation of subtypes and the related coherence (concordance) is considered for all patient samples. Lastly, limitations on the interpretation of the results are demonstrated through cellular deconvolution of bulk tissue RNA-seq data.

2.2 Materials and Methods

2.2.1 Data Sources

Within the NCBI Gene Expression Omnibus (GEO) data repository, the study with accession GSE153960 contains RNA-seq data from 1659 tissue samples, spanning 11 regions of the CNS, from 439 patients with ALS, frontotemporal lobar degeneration, or comorbidities for ALS-Alzheimer’s (ALS/AD) or ALS-FTLD. These 1659 tissue samples were filtered such that only the individuals belonging to the groups ALS-TDP, ALS/FTLD, ALS/AD, and ALS-SOD1 were considered. Cortex samples were derived from the frontal, medial motor, lateral motor, and ‘unspecified’ motor cortex regions, while spinal cord samples were derived from the cervical, thoracic, and lumbar regions. From the postmortem cortex, 451 unique tissue transcriptomes, corresponding to 208 individuals with ALS, passed the filtering criteria. An additional 135 transcriptomes were derived from non-neurological controls (*n*=93 samples; *n*=58 patients) and patients with FTLD exclusively (*n*=42 samples and patients). Similarly, in the postmortem spinal cord, 428 tissue transcriptomes are considered from 206 individuals with ALS, >85% of which are included in the postmortem cortex cohort. An additional 91 samples, corresponding to 56 individuals, are derived from non-neurological controls exclusively. Files were transferred using Globus[111].

2.2.2 Study Approval

The NYGC ALS Consortium samples presented in this work were acquired through various IRB protocols from member sites and the Target ALS postmortem tissue core and transferred to the NYGC in accordance with all applicable foreign, domestic, federal, state, and local laws and regulations for processing, sequencing, and analyses.[266]

Postmortem brain tissues from patients with FTLD-TDP or PSP and from cognitively normal individuals were obtained from the Mayo Clinic Florida Brain Bank. Diagnosis was independently ascertained by trained neurologists and neuropathologists upon neurological and pathological examinations, respectively. Written informed consent was given by all participants or authorized family members, and all protocols were approved by the IRB and ethics committee of the Mayo Clinic.[266]

2.2.3 Alignment and Quantification

Quantification of gene expression was performed using RSEM[196], as detailed by Prudencio *et al*.[266]. The processed gene count matrix was accessed directly from the GEO Accession ([https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE153960]) and counts were rounded to integers as recommended by the authors of RSEM.

SQuIRE[359] was selected for transposable element quantification, as this alignment pipeline provides locus-specific TE counts, allowing for a deeper analysis beyond TE subfamilies. Similar to RSEM, SQuIRE applies the Expectation Maximization (EM) algorithm to optimize the allocation of multi-mapped reads – an important step when sequence similarity between transcripts is high. SQuIRE's Fetch**,** Clean**,** Map**,** and Count functions were utilized to align and quantify locus-specific transposable elements. The EM 'tot_counts' values were selected as the estimate for sequencing reads attributed to the transposable elements. The hg38 build was used during mapping, with default trim and EM parameters, and a read length of 100 or 125 base pairs depending on the sequencing platform specified. A scoring threshold of ≥ 99 was used to restrict the number of false positive TEs (1%), with few uniquely mapping reads. Quantification was performed separately in the postmortem cortex and spinal cord. Only the locus-specific TEs with at least one count for all ALS samples were included in downstream analysis, resulting in 1474 unique TE features in the postmortem cortex and 475 in the postmortem spinal cord. The naming scheme for the locus-specific transposable elements is presented in SQuIRE[359], however in brief, TE feature names included the mapping chromosome, start and stop base pairs, transposable element subfamily, family and superfamily identifiers, base mismatches in parts per thousand, and sense or antisense stand annotation. TE lists for each cohort are available online in supplementary datasets from related publications[98,99]. The SQuIRE software pipeline was executed on ASU's Agave high performance computing cluster.

2.2.4 Addressing Technical and Biological Covariates

As discussed by Prudencio *et al.*[266], the large ALS cohort size required the utilization of two different sequencing platforms (HiSeq 2500 and NovaSeq 6000, Illumina, San Diego, CA) to complete the analysis. Exploratory differential expression in the cortex considering sequencing platforms as the design equation factor revealed strong batch effects in gene expression, evident by more than one-third of all genes falling below the Benjamini-Hochberg[26] corrected p-value threshold (37.2%, 22478/60403; including TEs). To correct for these batch effects, the approach outlined by Prudencio *et al*. was followed and resulted in splitting the ALS cohort based on sequencing platform. In the cortex cohort, the NovaSeq subset contained 255 patient transcriptomes (*n*=106 female, *n*=149 male), while the HiSeq subset contained 196 (*n*=97 female, *n*=99 male). In the spinal cord cohort, the NovaSeq subset contained 273 patient transcriptomes (*n*=126 female, *n*=147 male), while the HiSeq subset contained 155 (*n*=75 female, *n*=80 male).

DESeq2[208] was initially applied to perform a preliminary differential expression analysis on gene and TE counts to screen for covariate-dependent expression. In the postmortem cortex cohort, differential expression was utilized to guide the removal of sex-dependent genes prior to clustering. As described by Prudencio *et al*., sex was determined using XIST and UTY expression. Default parameters were used for DESeq2 differential expression, with male specified as the reference level and the ‘betaPrior’ argument in the DESeq() function set to true. A Benjamini-Hochberg corrected p-value ≤ 0.05 was selected as the threshold for removal of sex-dependent genes. In the postmortem spinal cord, additional covariates were individually screened for dependent gene expression including: site of collection (NYGC versus Target ALS), RIN, and tissue region. Further, given previous work[151], it was understood that cell type composition strongly influences bulk tissue expression in the spinal cord and used marker genes defined by the same study to remove these tissue-dependent features. Glial marker genes were obtained from Table S3 in *ref. 18*.

Following the removal of covariate-dependent genes using differential expression, the raw count matrix was subject to a variance stabilizing transformation (VST) to address heteroskedasticity in gene counts[208]. The VST counts were then subject to rank ordering by median absolute deviation (MAD) and the top 5,000 and 10,000 features were retained for unsupervised clustering analysis by non-negative matrix factorization (NMF), in the spinal cord and cortex cohorts, respectively[192,255]. Fewer features were used in the spinal cord cohort due to lower gene expression levels. This process was completed independently for all sequencing platform subsets.

2.2.5 Rank Estimation

Factorization rank was estimated in R, Version 4.0.3 (The R Foundation for Statistical Computing, Vienna, Austria) using the NMF package[116]. A rank of 3 for clustering analysis, based on the plots of the cophenetic correlation coefficient for ranks spanning 2 to 6. Quality measures were estimated using 50 iterations at each rank and the default seeding method. The nsNMF (non-smooth non-negative matrix factorization) method variant was utilized for all NMF clustering[255].

2.2.6 Non-negative Matrix Factorization

Non-negative matrix factorization was performed in SAKE, a convenient tool for RNA-seq sample pre-processing, filtering, clustering and visualization[148] (Version 0.4.0). In the cortex cohort, the top 10,000 MAD genes, after a variance stabilizing transformation, were utilized as the input into SAKE, while the top 5,000 were used in the spinal cord cohort due to more stringent filtering of covariate-dependent genes. In SAKE, no samples were removed during the quality control step, and further transformations in the filtering step were not necessary. During non-negative matrix factorization, selected parameters include factorization rank = 3, iterations = 200 (cortex) or 100 (spinal cord), and NMF method set to nsNMF.

To robustly assign ALS sample subtypes, 10 independent rounds of NMF clustering were performed in SAKE. For each patient sample, the ALS subtype with a simple majority was assigned. For a small number of edge cases, an eleventh round of NMF clustering was used as a tiebreaker to reach the simple majority threshold. This process was completed for both sequencing platform groups in each tissue cohort.

2.2.7 Feature Selection

After each replicate of NMF clustering, gene and TE feature scores[175] were calculated for all transcripts from each cohort and sequencing platform subgroup. Feature scores were averaged across all clustering replicates and reordered. In the postmortem cortex cohort, the top 1000 features from both sequencing platform cohorts were combined, and after the removal of duplicates, 1,681 transcripts remained for enrichment, corresponding to 891 gene symbols. Feedback from peer-reviewers[99] guided the decision to increase the number of genes considered in the spinal cord cohort. Therefore, all features from both sequencing platform subgroups were combined, and after the removal of duplicates, 8,163 transcripts remained for enrichment, corresponding to 5438 gene symbols.

2.2.8 Enrichment

In the postmortem cortex cohort, the 891 gene symbols were then enriched using two independent approaches, GSEA[310] (Version 4.1.0, Broad Institute, Boston, MA) and Enricher[185]. Healthy control donors were selected as the reference phenotype during enrichment. Transcripts without a corresponding gene symbol (HGNC) were excluded from the enrichment analysis, including TEs. The minimum gene set size was adjusted to 5, and all other parameters were maintained as the default. For the enrichment, the canonical pathways contained in the Reactome database[159], a custom gene set containing markers of disease-associated microglia[68,171], and curated gene sets for Alzheimer's, Parkinson's, and ALS[31,165] were used. Pathway heatmaps reflecting gene enrichment by phenotype were built using the 'Rank Metric Score' tabulated during GSEA. Enrichr[185] was performed to support subtype-specific pathway expression observed during GSEA, utilizing the Fisher's exact test with Benjamini-Hochberg multiple hypothesis test correction[26]. Hypergeometric enrichment analysis was considered in the context of the Reactome 2016 database. Upregulation and downregulation of pathways was determined using subtype-specific differential expression, with each feature assigned to two of the three subtypes based on the maximum and minimum median expression on the DESeq2 median-of-ratios scale.

To perform GSEA in the postmortem spinal cord cohort, transcript expression was normalized to the DESeq2 median-of-ratios scale for enrichment after removing covariate dependent genes. Default parameters were maintained, aside from lowering the minimum gene set size to 5 and maximum to 150. Canonical pathways contained in the Reactome database[159] were leveraged and pathway normalized enrichment scores are presented for each ALS subtype. Non-neurological controls were designated as the reference level. To further support subtype-specific pathway enrichment observed in GSEA, hypergeometric enrichment analysis was performed using Enrichr[185], the Reactome 2022 database, and the subtype feature assignment approach detailed above[99]. Enrichment *p*-values are determined by Fisher's exact test, and presented as –$\log_{10}$ transformed values after FDR adjustment. The *p*-value heatmap is color-coded to indicate upregulation or downregulation relative to the other subtypes, and blank cells indicate an FDR adjusted *p*-value > 0.05.

2.2.9 Bootstrapping and Hybrid Subtypes

Given that previously established predictor gene sets for ALS subtype were not available, ALS-Glia, ALS-Ox, and ALS-TD predictor gene sets in the cortex cohort were derived from the gold, navy, and maroon eigengenes, respectively and defined in a later chapter. Transcript counts were considered on the DESeq2 median-of-ratios scale, adjusted for RIN, site of collection, and sequencing platform covariates. Difficulty identifying ALS-Glia and ALS-TD specific eigengenes in the postmortem spinal cord cohort restricted this analysis to the cortex cohort exclusively.

Subtype scores were defined as the average expression of subtype-specific predictor genes minus the average expression of all features from the cortex or spinal cord cohorts. Scores were calculated for 100 different sets of predictors (per subtype) and used to define a 5% cutoff for the expected subtype score[175]. Each sampled predictor gene set contained the same number of features as the original eigengene, and were generated by randomly sampling the eigengenes with replacement. For example, the expected subtype score for ALS-Glia patients was determined by first generating 100 predictor sets by randomly sampling features comprising the gold eigengene. Then, the average (sample-wise) ALS-Glia expression was determined for each of the 100 predictor sets and subtracted from the average (sample-wise) ALS-Glia expression of all 1,681 classification genes.

After repeating this analysis for the ALS-Ox and ALS-TD subtypes, using their respective eigengenes, 100 subtype scores were generated for all cortex samples (*n*=203 female, *n*=248 male). A 5% cutoff for the expected subtype score was then established, per sample, and final subtype classification thresholds were determined by weighting expected subtype scores according to the observed proportion of patient samples in each subtype (obtained from clustering). Bootstrapping was then applied, involving the sampling of predictor gene sets (with replacement) and calculation of subtype scores for 1000 iterations.

Patient samples were initially placed at the origin, and moved in the direction of the subtype vertex after passing the corresponding subtype threshold. Therefore, the x, y, and z axis vertices reflect the expression of a single subtype, while the other three vertices capture a combination of two subtypes. Individual points that passed a given subtype threshold in >50% of bootstrap iterations were filled with their respective subtype colors. Samples were considered to express a hybrid subtype state if one subtype threshold was passed >50% of the time and simultaneously passed a second subtype threshold >40% of the time.

2.2.10 Cellular Deconvolution

In the postmortem cortex cohort, cell deconvolution was performed using CIBERSORTx[244] with reference single cell RNA-sequencing expression from the developing human brain available from Nowakowski *et al.*[247] (http://bit.ly/cortexSingleCell). Raw data were filtered and normalized to the DESeq2 median-of-ratios scale. Cell types were grouped into 10 major cell types: neuronal progenitor, excitatory neuron, inhibitory neuron, glial progenitor, astrocyte, microglia, endothelial, mural, choroid, and unknown. Marker genes for each major cell type were identified using Seurat's[289] function FindAllMarkers() (Version 4.0.3). Marker genes were used to generate medioids (i.e., cell type signatures) to use as the reference for cell deconvolution. The ALS cohort was normalized using DESeq2 with count values on the median-of-ratios scale. All overlapping MAD transcripts between the NovaSeq and HiSeq cohorts were used, totaling 7372 transcripts, to ensure a sufficient number of transcripts were available for deconvolution. Transcripts without a mapped gene symbol and transposable elements were removed from the analysis which led to 4912 total genes. Lastly, transcripts not shared between ALS and control cohorts (n=586; n=267 female, n=319 male) and Nowakowski cell type signatures were removed. 1881 transcripts remained and were used as input into CIBERSORTx. Quantile normalization was disabled in CIBERSORTx, which is recommended for RNA-seq data, and 500 permutations were used for significance analysis. Significant differences in cell type fractions were assessed using the nonparametric Wilcoxon rank sum test[204] with Bonferroni[92] correction. In the postmortem spinal cord cohort, cell type proportions from bulk expression deconvolution were obtained from a previous study[151].

2.2.11 Assignment of Patient-Level Subtype

For most ALS patients considered, multiple postmortem tissue samples from the cortex and spinal cord were subject to RNA-sequencing. Patients were assigned a subtype label only if there was a majority consensus among their frontal and motor cortex samples, or if there was a single sample characterized. The same approach was taken to assign patient-level subtypes in the spinal cord cohort. This strategy is termed the majority agreement approach herein. ALS patients which failed to reach a majority consensus in the cortex or spinal cord (independently) were labeled 'Discordant'. A 'global' or full consideration of all available postmortem samples using the majority agreement approach is presented in *Chapter 3* for the 192 patients with observations from both the cortex and spinal cord.

2.2.12 Intra-Patient Concordance Analysis

Postmortem cortex subtype labels were used to assess concordance with the molecular phenotype presented in the spinal cord of the same patients, given that most patients had observations from both regions of the CNS. Agreement is considered at the tissue-level rather than CNS level (i.e. cortex, spinal cord), to avoid sample dependence concerns with the majority agreement approach described above. Subtype discordance between the cortex and spinal is color-coded using the same scheme presented in the hybrid subtype analysis to inform which discordant subtype was more common, given the postmortem cortex observation. The *p*-values were estimated using bootstrapping, where distributions for the expected number of concordant patient samples were generated from 10,000 iterations, after adjusting sampling probabilities to reflect observations from the cortex and spinal cord. Subtype re-sampling probabilities were adjusted separately for the postmortem cortex and spinal cord cohorts, with probabilities in the cortex set equal to 239/451, 84/451, and 128/451 for Ox, Glia, and TD subtypes respectively. The spinal cord probabilities were 139/428, 106/428, and 183/428, for Ox, Glia, and TD respectively. The same approach was used when considering concordance in the NovaSeq and HiSeq platforms independently, with all values presented in Figure 2.11. True concordant values were compared against the derived distribution for estimation of *p*-values assuming a one-tailed binomial distribution.

2.3 Results

To assess whether ALS patient heterogeneity is reflected in postmortem gene expression, unsupervised clustering analysis was first performed in SAKE[148] using ALS transcriptomes from the postmortem cortex (*n*=451) and spinal cord (*n*=428) as presented in Figure 2.1. As shown in Figure 2.1D, over 85% (*n*=192) of ALS patients were shared between the cortex and spinal cord cohorts. SQuIRE[359] was implemented to quantify transposable element expression with chromosomal locus specificity. TE features were filtered to ensure the retained transcripts had unique mapping reads and counts in all ALS patient samples per cohort. In response to preliminary differential expression analyses and relevant work[266], both postmortem cohorts were split according to sequencing platform (NovaSeq 6000 and HiSeq 2500, Illumina, San Diego, CA, USA) and clustered independently. Following the assignment of sample subtypes, gene enrichment and cellular deconvolution are leverage to provide additional insight into subtype-specific pathologies and the influence of bulk tissue RNA-sequencing on the observed phenotypes. Prior to clustering, a variance stabilizing transformation was applied following size factor estimation, as shown in Figure 2.2, and the removal of covariate-dependent genes was performed using DESeq2 differential expression[208]. In the postmortem cortex cohort, sex was the only covariate considered prior to clustering, while RIN, tissue, and site of collection were additional covariates screened for dependent gene expression in the postmortem spinal cord. Estimation of factorization rank was then performed in R, and a rank of 3 was chosen given the quality metrics presented in Figure 2.3. No specific filter or expression cut-off was applied to remove lowly expressed genes, however most features were expressed with a mean of normalized counts > 10, as shown in Figure 2.4.

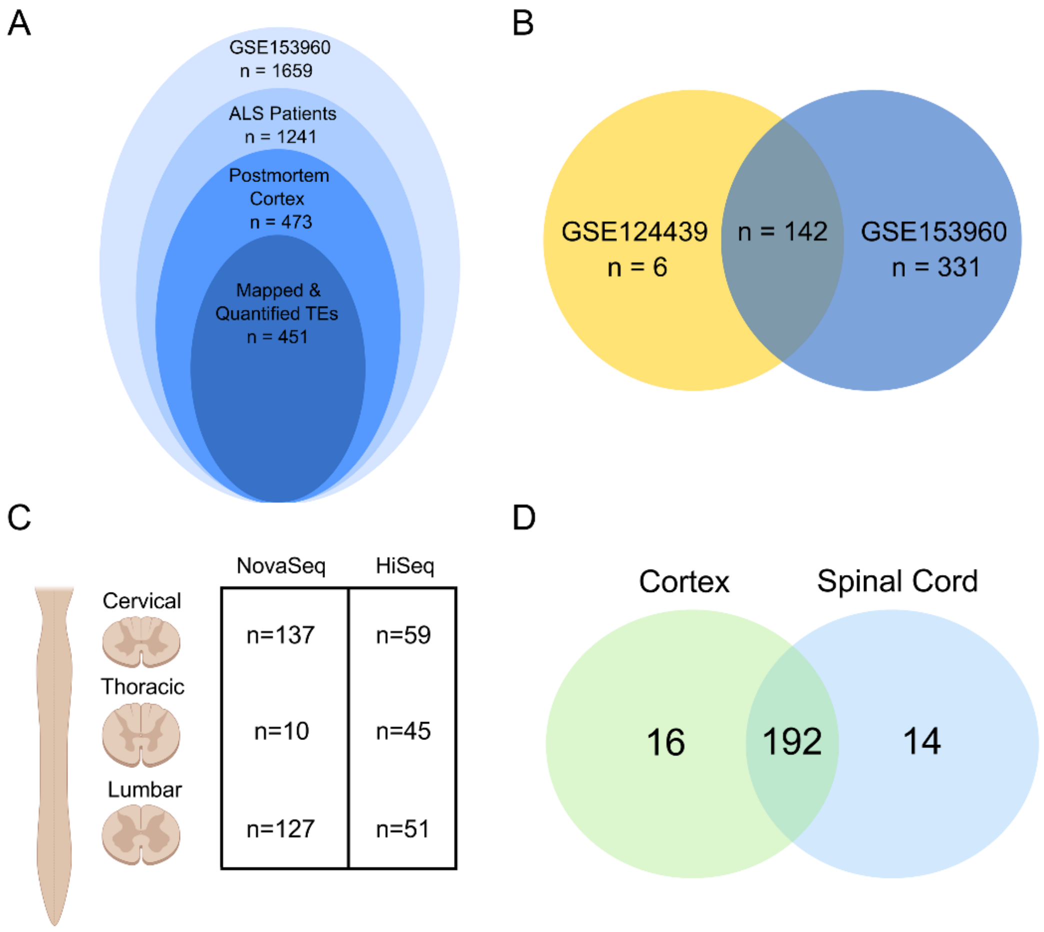


Figure 2.1: The postmortem cortex and spinal cord cohorts. (A) Selection of the postmortem cortex ALS cohort from the GSE153960 repository. Transcriptomes associated with the frontal and motor cortex were the only tissue sites considered in this cohort. Control samples are not shown, but included 93 transcriptomes from healthy control donors and 42 from frontotemporal lobar degeneration patients. (B) Comparison of common ALS patient postmortem cortex samples between the foundational study from Tam *et al.*[317] (GSE124439) and the repository utilized in this analysis. (C) Overview of sample numbers from the spinal cord region, separated by sequencing platform. (D) Venn diagram showing patient overlap between the postmortem cortex and spinal cord cohorts considered in this dissertation. Over 85% of patients were included in both cohorts, allowing assessment of intra-patient concordance.

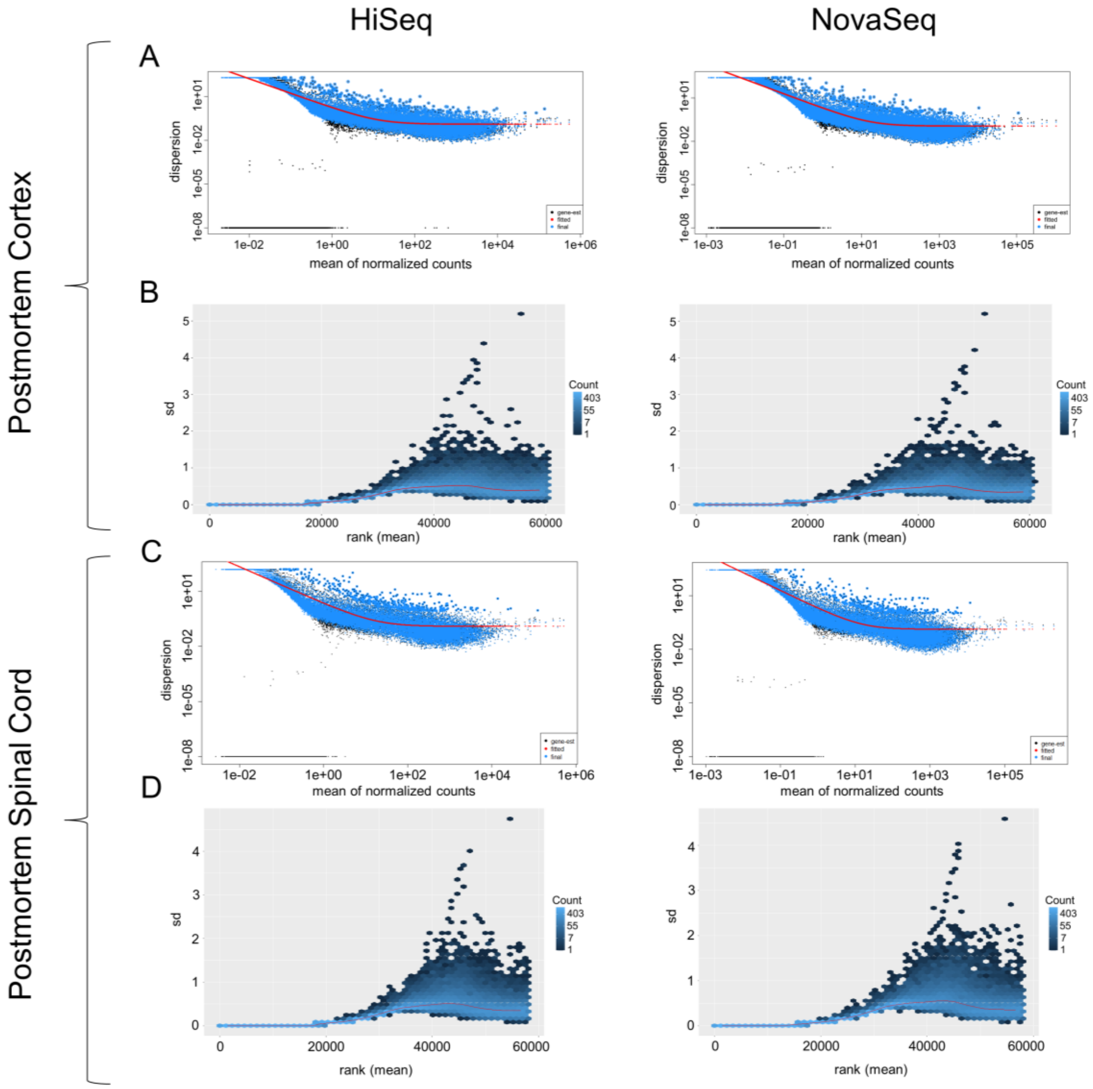


Figure 2.2: Diagnostic plots following DESeq2-based normalization. In the postmortem cortex, (A) estimation of size factors are performed first and is followed by a (B) variance stabilizing transformation to address heteroskedasticity concerns, ensuring the most variable genes are not strictly the most highly expressed. Equivalently, in the postmortem spinal cord cohort, (C) estimation of size factors is followed by (D) a variance stabilizing transformation. Diagnostic plots are relatively consistent throughout all cohorts and sequencing platform subgroups, indicating that the effects due to normalization are similar.

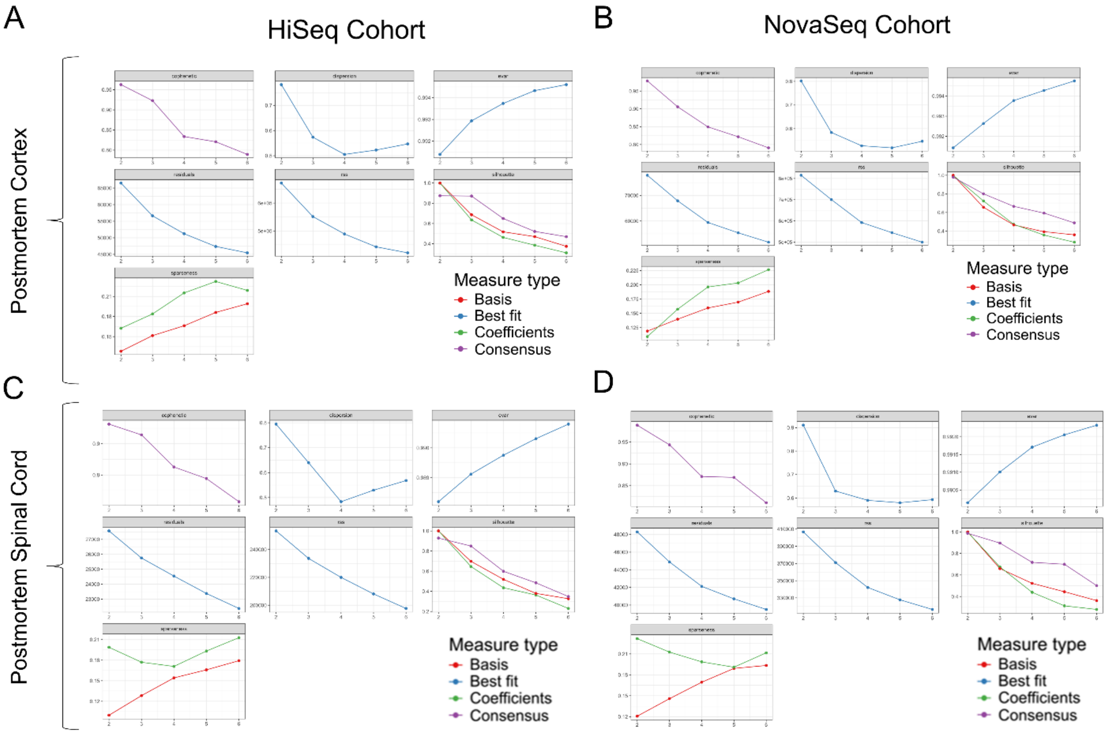


Figure 2.3: Estimation of rank using the NMF package in R. Ranks 2-6 were considered and quality measures were estimated using 50 iterations at each rank with the default seeding method. Cluster number estimation (rank) was performed independently in the postmortem cortex (A) HiSeq subgroup and (B) NovaSeq subgroup, and the postmortem spinal cord (C) HiSeq subgroup and (D) NovaSeq subgroup.

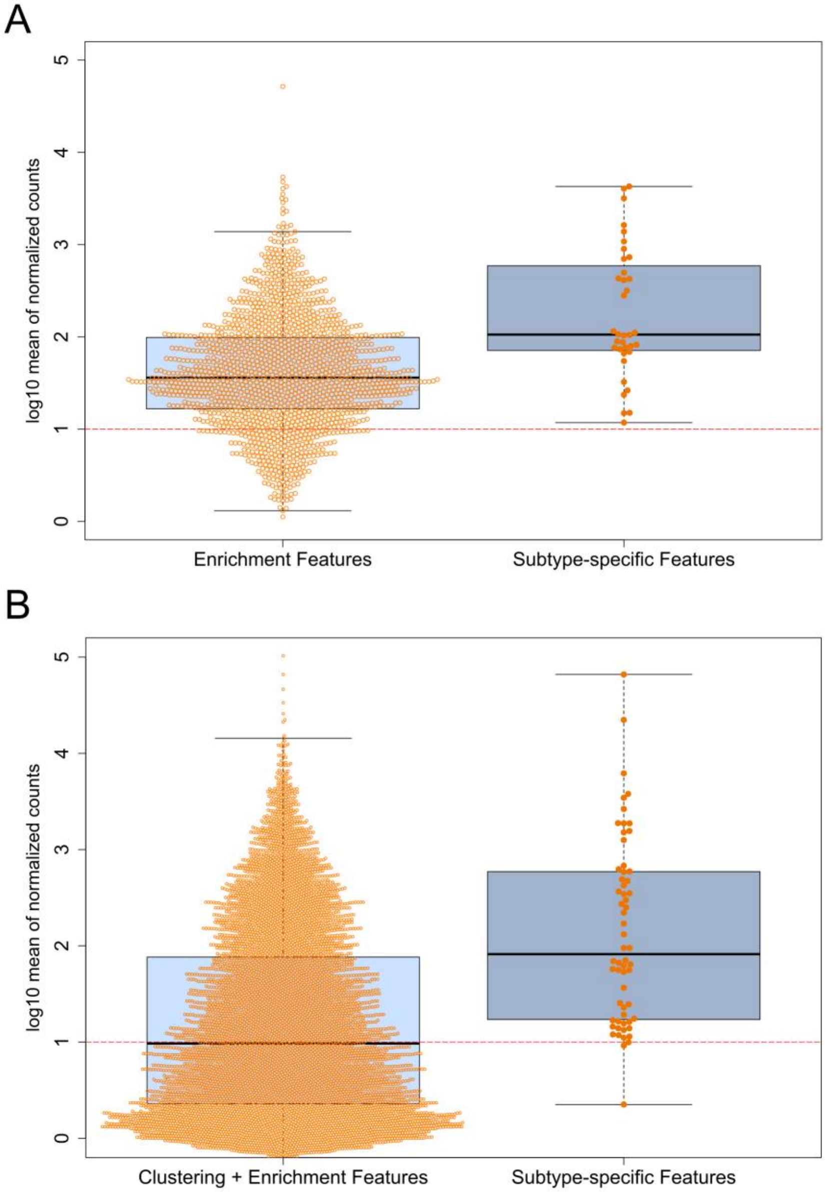


Figure 2.4: Feature expression in the postmortem cortex and spinal cord cohorts following covariate-dependent gene removal and DESeq2 normalization. (A) Mean of normalized counts adjusted for RIN, site of collection and sequencing platform covariates, presented on the $\log_{10}$ scale, showing the 1,681 genes used in enrichment, WGCNA, and differential expression ("Enrichment Features"). All cortex subtype-specific features (n=36) show mean expression above 10 normalized counts (red dashed line). (B) Mean of normalized counts adjusted for RIN, site of collection, tissue region, and sequencing platform covariates, presented on the $\log_{10}$ scale, showing ~8100 genes used in clustering and enrichment. Two of the 60 subtype-specific features had mean normalized expression below 10 counts (red dashed line).

2.3.1 Unsupervised Clustering in the ALS Postmortem Cortex

After filtering for the top 10,000 most variably expressed transcripts by median absolute deviation, non-smooth non-negative matrix factorization (nsNMF)[255] was applied in SAKE[148] to identify subgroups of ALS patients based on gene expression in the postmortem cortex. Eleven independent rounds of clustering were performed, and soft clustering probabilities were used to assign single subtypes to all samples with representative results depicted in Figure 2.5. Three distinct patterns of gene expression were identified in both the NovaSeq and HiSeq ALS cohorts as shown in Figure 2.6A and 2.6F. Principal component analysis (PCA) further demonstrated the ability to separate the putative ALS subtypes into three distinct clusters when considering the first and second principal components as seen in Figure 2.6B and 2.6G. To illustrate subtype-specific expression, two transcripts associated with each subtype were considered in the principal component space on the VST scale, as depicted in Figure 2.6C-E and Figure 2.6H-J. Taken together, these results support the existence of three distinct patterns of gene and TE expression within the ALS postmortem cortex transcriptome.

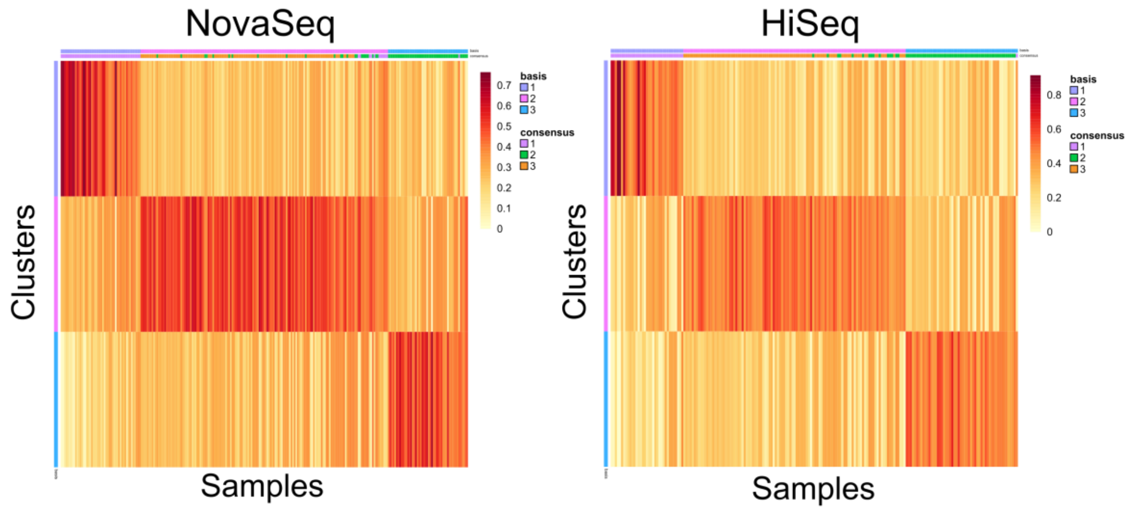


Figure 2.5: Representative soft clustering probabilities for patient samples from the postmortem cortex, separated by sequencing platform. Unsupervised clustering was performed using the non-smooth variant of the NMF algorithm[255]. Basis groupings were used for assignment of sample subtype.

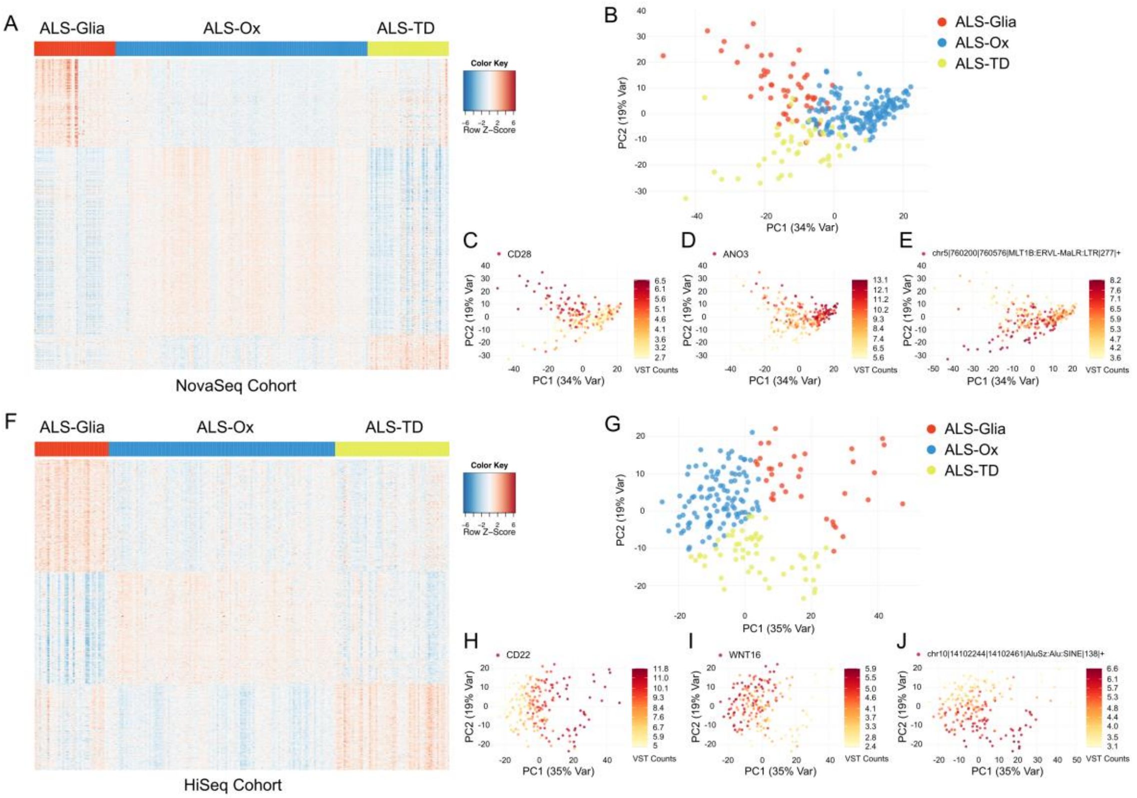


Figure 2.6: (A) Heatmap of 741 genes and transposable elements selected by SAKE[148] shows transcript overexpression in a subtype-specific fashion for the NovaSeq cohort (*n*=255 biologically independent samples). Transcript counts are z-score normalized. (B) Principal component analysis shows three distinct clusters, when considering the first two principal components. (C) Sample expression of *CD28* transcripts was plotted in the same PCA space, with elevated counts seen for the ALS-Glia subtype. A darker color corresponds to higher feature expression. (D) Expression of the *ANO3* gene shows specificity for the oxidative stress and altered synaptic signaling subtype. (E) The ALS-TD subtype shows specific upregulation of transposable element chr5|760200|760576|MLT1B:ERVL-MaLR:LTR|277|+ compared to the other two subtypes. (F) Heatmap of 618 genes and TEs shows subtype-specific expression in the HiSeq cohort (*n*=196 biologically independent samples). (G) PCA considering the HiSeq cohort shows three distinct clusters of ALS patient transcriptomes. (H) Elevated expression of *CD22* is seen in the activated glia subtype. (I) Subtype-specific expression of *WNT16* in the ALS-Ox subtype. (J) chr10|14102244|14102461|AluSz:Alu:SINE|138|+ is overexpressed in the ALS-TD subtype.

2.3.2 Unsupervised Clustering in the ALS Postmortem Spinal Cord

Eleven independent rounds of unsupervised clustering were performed on 428 transcriptomes derived from cervical, thoracic and lumbar postmortem tissue, corresponding to 206 unique ALS patients. To help address bulk tissue effects and differences in cell type composition between the cortex and spinal cord, oligodendrocyte, microglia, astrocyte, and endothelial cell marker genes were removed (n = 1282), presented by Humphrey *et al*.[151], given their findings show differential gene expression in the ALS spinal cord is partially driven by cell type composition. The majority of these features showed gene expression dependent on one of the four covariates previously addressed for both the NovaSeq (1061/1282) and HiSeq (855/1282) cohorts. Following the removal of covariate-dependent and glial marker genes, the top 5,000 most variable transcripts by median absolute deviation were selected. The number of features considered was reduced relative to the cortex cohort to limit the consideration of noisy transcripts. Mirroring the cortex cohort, clustering probabilities were used to assign a single subtype to each patient sample, with representative results shown in Figure 2.7. Results capture three distinct expression profiles in both the NovaSeq and HiSeq cohorts as seen in Figure 2.8A-B, and yield three clusters following application of principal component analysis seen in Figure 2.8C-D.

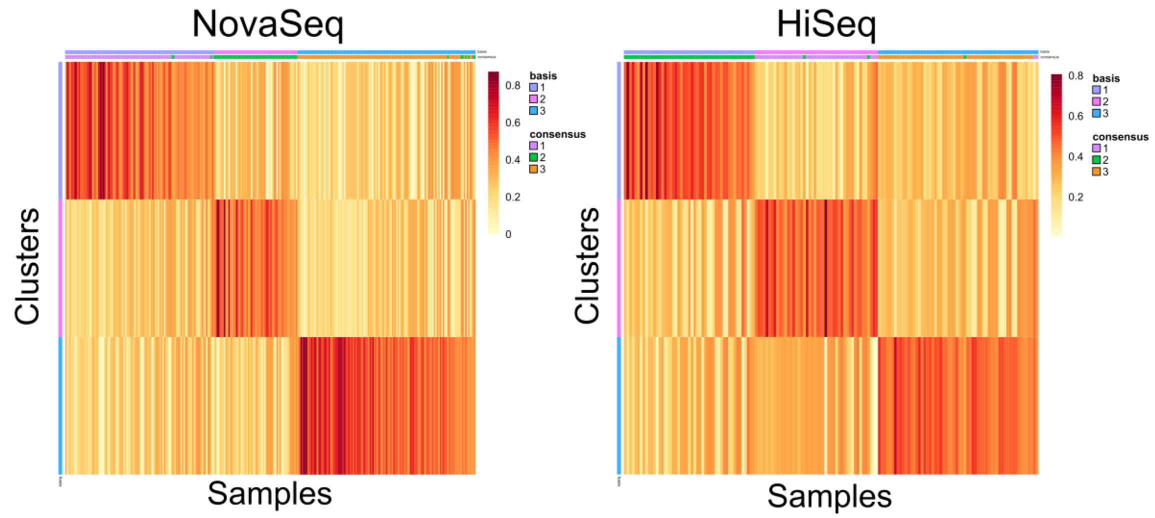


Figure 2.7: Representative soft clustering probabilities for patient samples from the postmortem spinal cord, separated by sequencing platform. Unsupervised clustering was performed using the non-smooth variant of the NMF algorithm[255]. Basis groupings were used for assignment of sample subtype.

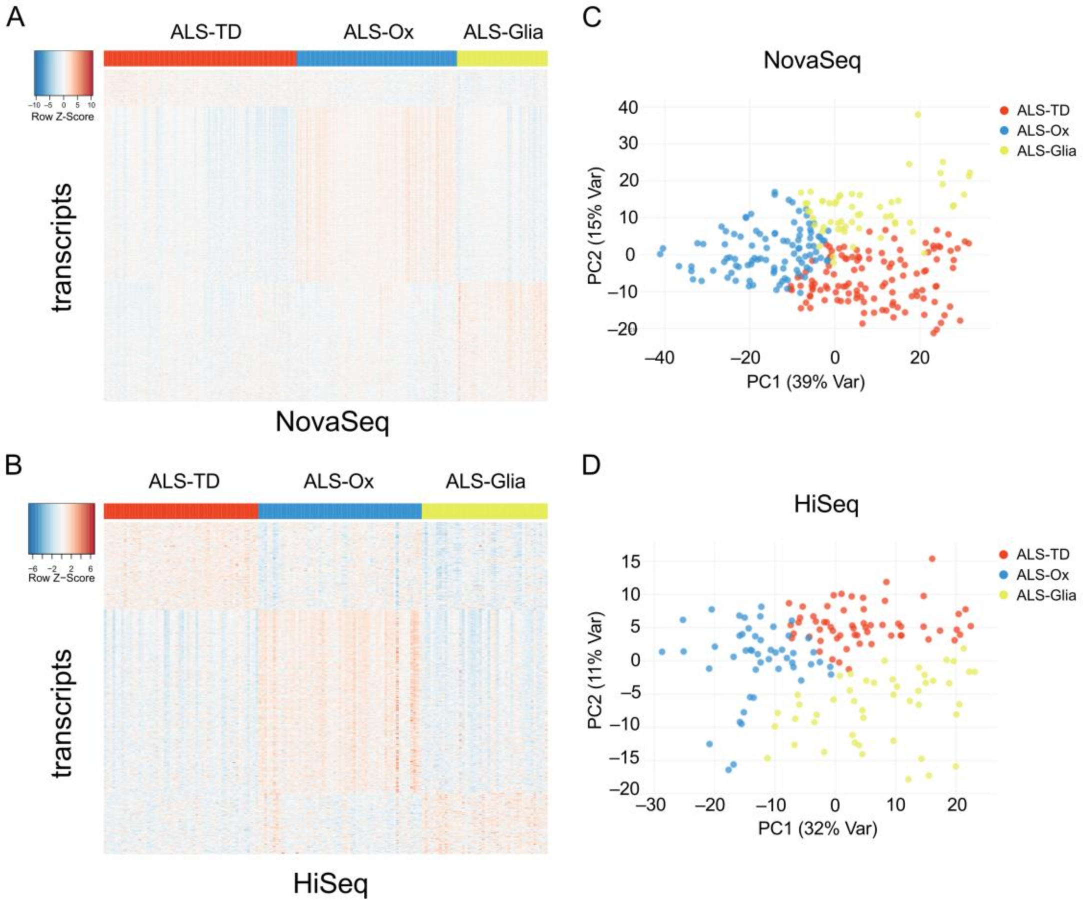


Figure 2.8: (A) Heatmap showing subtype specific gene expression in the NovaSeq cohort, comprised of 273 biologically independent tissue samples (*columns*) and 763 transcripts (*rows*). (B) Gene expression in the HiSeq cohort, with 155 biologically independent tissue samples (*columns*) and 567 transcripts (*rows*). Both heatmaps are presented after z-score normalization with features selected by SAKE[148]. Principal component analysis shows three distinct clusters when considering the first two principal components, in both the (C) NovaSeq and (D) HiSeq subgroups.

2.3.3 Enrichment in the ALS Postmortem Cortex

Gene scores from NMF clustering[175] were used to select the top 1000 features most uniquely associated with a single subtype from each sequencing platform (1,681 total), done in an effort to better isolate phenotypes unique to each ALS cluster. Gene enrichment was performed using two independent approaches, Gene Set Enrichment Analysis (GSEA)[310] and hypergeometric enrichment analysis[185] (Fisher's exact test). Subtype-specific pathway enrichment was observed for each ALS subtype, as shown in Figure 2.9A-D. In the first subtype, termed ALS-Glia[317], enrichment for immunological signaling and activation, genes implicated in a pro-neuroinflammatory microglia state in Alzheimer's (Disease-Associated Microglia, DAM)[171], and markers of neural cell death were observed as seen in Figure 2.9A, 2.9B, and 2.9G. Transposable element expression was greatly reduced in ALS-Glia samples compared to the other two subtypes, as shown in Figure 2.9E and Figure 2.10A and 2.10C.

Enrichment of the remaining two subtypes, termed ALS-TD and ALS-Ox[317], suggest some overlapping disease mechanisms, such as altered ECM maintenance and the influence of post translational modification machinery, as observed in Figure 2.9C, 2.9D, 2.9I, and 2.9J. Furthermore, while the ALS-Ox subtype had the strongest expression of the locus-specific TEs, as seen in Figure 2.9E, the ALS-TD subtype showed elevated TE expression more often than the control groups and ALS-Glia subtype as shown in Figure 2.10A. To distinguish the ALS-TD subtype from ALS-Ox, the unique downregulation of RNA polymerase II transcriptional genes is observed for the TD subtype as seen in Figure

2.9H. This enrichment evidence, along with univariate features considered later, were used to define this ALS subgroup by transcriptional dysregulation (TD), rather than TE expression, as in previous work[317].

In the ALS-Ox subtype the distinct enrichment of Alzheimer's associated genes is observed, but not genes previously associated with ALS or Parkinson's disease, which may reflect stringent filtering during NMF score-based feature selection. Additionally, negative enrichment for genes involved in oxidative phosphorylation are seen in Figure 2.9D, and weak positive enrichment for synaptic signaling shown in Figure 2.9D and 2.9K, when compared to the control cohort. It is worth noting that subtype enrichment results generally agree with the findings reported by Tam *et al*.[317], important given the overlapping patient cohort depicted in Figure 2.1B, although custom TE enrichment was not performed and some key differences are observed for the ALS-Ox group. Given these results, ALS subtype naming conventions presented by Tam *et al*.[317] were maintained, where appropriate.

In the NovaSeq cohort there was roughly a ratio of 3:1.4:1 observed for the ALS-Ox, ALS-TD, and ALS-Glia subtypes, respectively. The HiSeq cohort showed a similar proportion of ALS subtypes, with an approximate 3:1.9:1 ratio observed for the ALS-Ox, ALS-TD, and ALS-Glia subtypes, respectively. Similar subtype proportions in both sequencing platform subgroups are seen as evidence for the limited influence of instrumentation on the expression of ALS phenotypes postmortem, strengthening the overall utility of patient stratification.

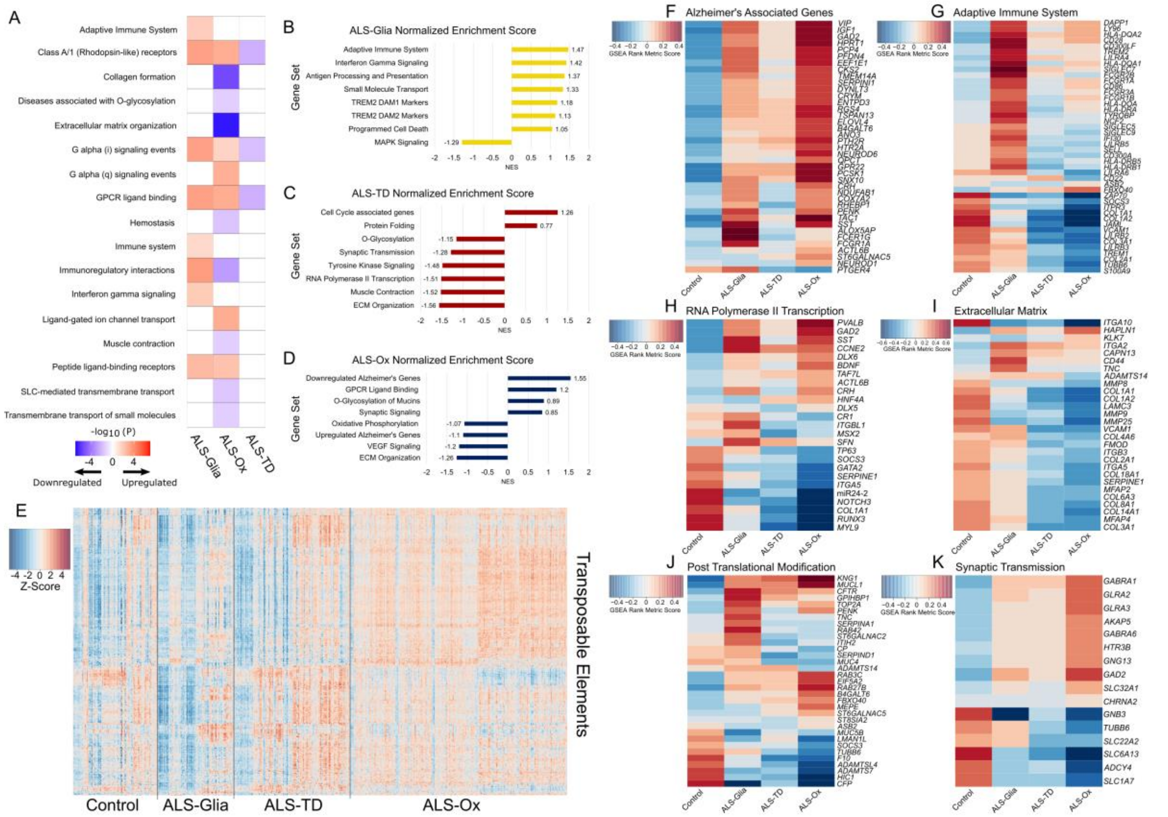


Figure 2.9: (A) Benjamini-Hochberg adjusted *p*-values, derived from a Fisher's exact test, are presented on the –$\log_{10}$ scale. All presented pathways are significantly enriched in at least one subtype. Negative enrichment is encoded as the negative magnitude of the –$\log_{10}$(adjusted *p*-value). *P*, Fisher's exact test, one-tailed, Benjamini-Hochberg method for multiple hypothesis test correction. (B-D) Gene sets enriched in each ALS subtype are presented along the Y-axis, with GSEA normalized enrichment score (NES) presented along the X-axis. (E) Heatmap of transposable element expression, with 426 unique TEs and 544 biologically independent transcriptomes. Patient samples were plotted by subgroup, with the thin black lines denoting sample separation by subtype. TE count values were subject to VST, followed by z-score normalization, with red indicating elevated expression. (F-K) Pathways enriched specifically for one or more subtypes were generated using GSEA rank metric scores. Genes comprising each functional pathway are included, with subtype-specific gene enrichment scores encoded on a red-blue scale.

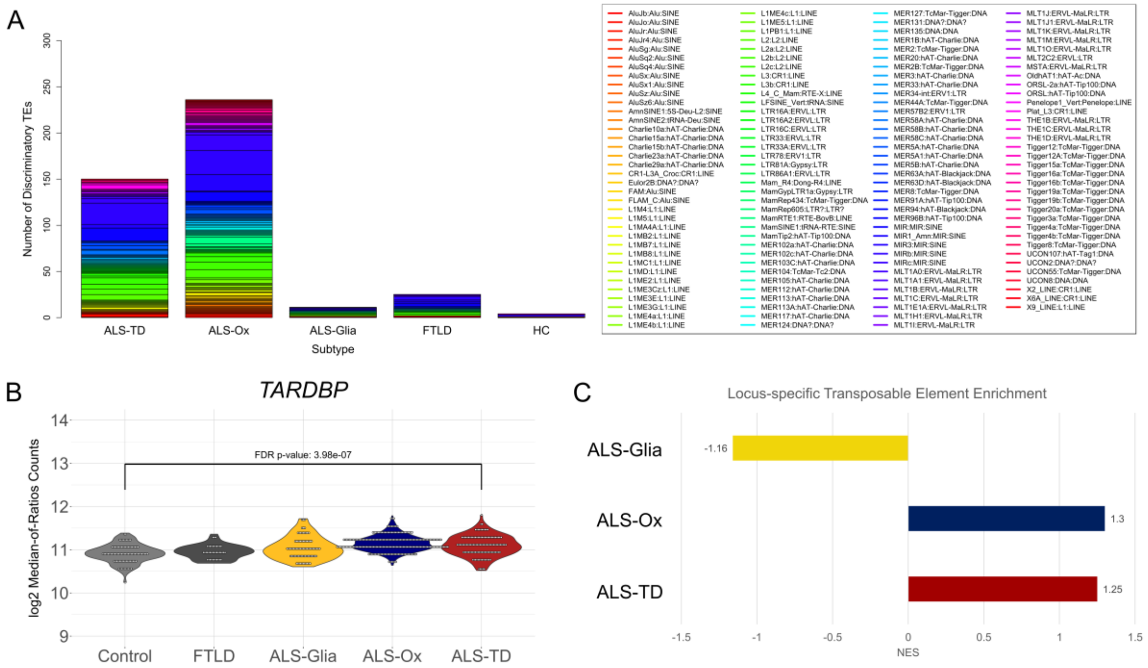


Figure 2.10: (A) Cortex transposable elements with gene locus resolution were assigned to the group which demonstrated the largest average expression on the median-of-ratios scale, and reveals characteristic expression in both the TD and Ox subtypes. (B) Cortex *TARDBP* expression, encoding TDP-43, is shown for healthy controls, patients with frontotemporal degeneration, and each ALS subtype on the DESeq2 median-of-ratios scale. *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction. Previous works have demonstrated direct interactions between TDP-43 and transposable elements (TEs) and implicated TE subfamilies as subtype specific features[206,317]. However, normalized expression is relatively consistent across ALS subtypes and significant differences in expression are not observed, suggesting *TARDBP* expression is not a defining characteristic of a single subtype. (C) A locus-specific transposable element feature set, derived from SQuIRE[359], was utilized to perform TE enrichment using GSEA. Normalized enrichment scores for each of the three subtypes are plotted, with healthy controls specified as the reference. Results indicate that elevated expression of TEs are characteristic of both the ALS-Ox and ALS-TD subtypes.

2.3.4 Enrichment in the ALS Postmortem Spinal Cord

Similar to the approach taken in the cortex cohort, enrichment was performed using all 5000 features from the NovaSeq and HiSeq cohorts, and combined them to yield 8,163 non-duplicate transcripts (corresponding to 5438 gene symbols). The resulting feature set was used to perform hypergeometric enrichment analysis with Enrichr[185] and the Reactome[159] pathway database, according to our previously described approach[99], as shown in Figure 2.11A. GSEA[310] was leveraged to enrich each stratified group against a non-neurological control cohort comprised of 91 donor samples from the cervical, thoracic, and lumbar regions of the spinal cord as shown in Figure 2.11B. In agreement with the phenotypes identified in the ALS postmortem cortex[99], significant enrichment for neuroinflammatory signatures is observed in ALS-Glia patients when compared to the other two subtypes. Interestingly, normalized enrichment scores indicate negative enrichment for neuroinflammatory pathways in the ALS spinal cord, relative to controls, whereas positive enrichment is observed in the cortex[99] – findings which may be linked to cell type composition[151] and stringent filtering of glial marker genes. In ALS-Ox patients, statistically significant positive enrichment for genes associated with synaptic signaling is observed, mirroring the phenotype observed in the postmortem cortex. Pathways associated with RNA metabolism and processing were the most strongly enriched in the ALS-TD subtype when compared to controls using GSEA, although significant associations were not observed at an adjusted *p*-value < 0.05 by either enrichment approach. Collectively, findings demonstrate that unsupervised clustering of the ALS spinal cord recapitulates many of the same subtype signatures observed in the postmortem cortex. Conversely, phenotypic differences between the groups are less pronounced than in the cortex, evident in the magnitude of GSEA normalized enrichment scores. These findings likely reflect differences in cell type composition between the two regions of the central nervous system and the fact that most glial marker genes demonstrated covariate-dependent expression and were subsequently filtered.

Allocation to each of the three subtypes in the NovaSeq cohort more closely agrees with findings from the frontal and motor cortex, with ALS-Glia representing the rarest subtype (21.9% of spinal transcriptomes compared to 19.2% in the cortex[99]) corresponding to a Glia:Ox:TD ratio of 1:1.5:2, as shown in Figure 2.12. In the HiSeq cohort, a Glia:Ox:TD subtype ratio of 1:1:1.4 was observed, shown in Figure 2.12, corresponding to ~30% of patients classified as ALS-Glia, indicating the selected sequencing platform influences the detectable subtype expression signature. In both cases, the transcriptional dysregulation (TD) subtype was the most commonly assigned – as compared to the oxidative stress subtype in the postmortem cortex – potentially a consequence of some dependency on the removed covariates as seen in Figure 2.13A-D, weak neuronal expression associated with cell type composition in the spinal cord as observed in Figure 2.13E, and stringent filtering of covariate-dependent genes.

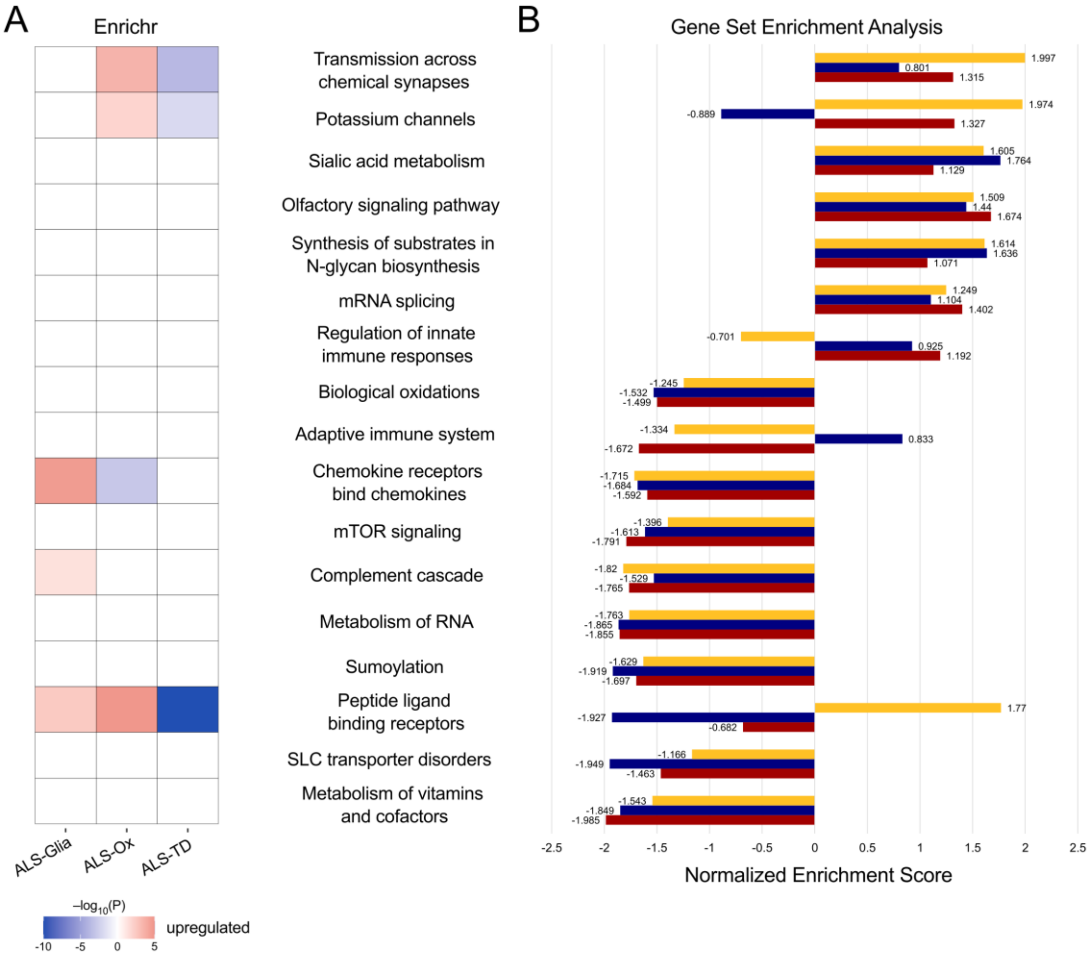


Figure 2.11: Enrichment analysis using (A) Fisher exact tests performed with Enrichr[185] with FDR[26] adjusted *p*-values presented on the $-\log_{10}$ scale. Genes were assigned to patient clusters depending on median expression and downregulated pathways are indicated with the negative magnitude of the $-\log_{10}$ transformed *p*-value. (B) GSEA[310] with non-neurological controls specified as the reference level for calculation of all normalized enrichment scores.

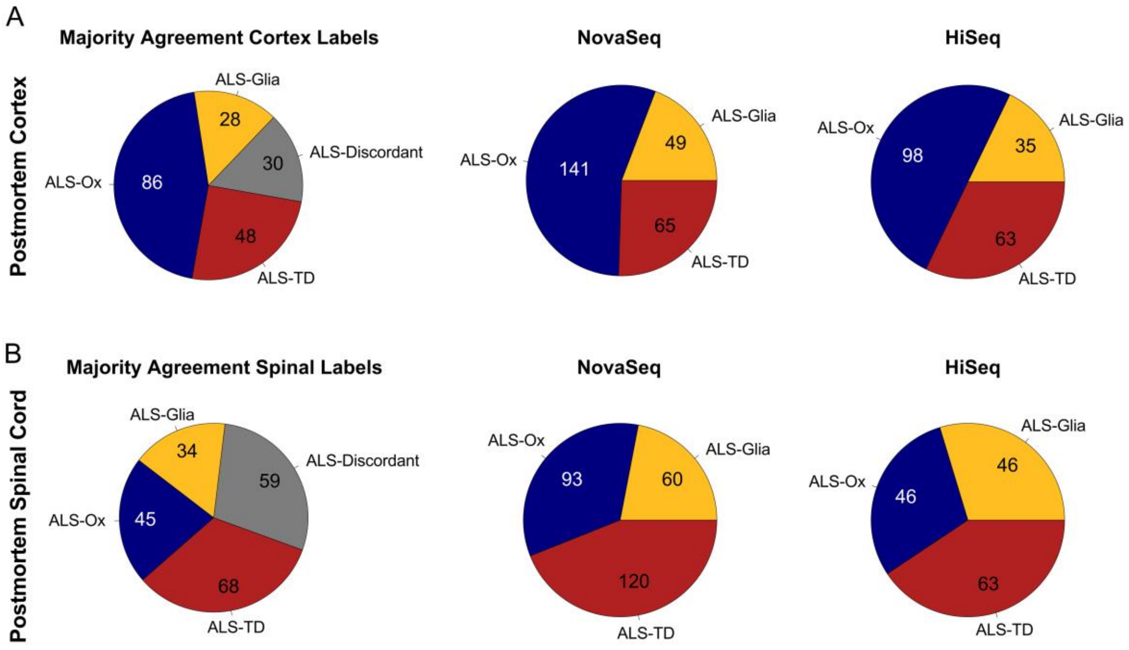


Figure 2.12: Generalized comparison of subtypes assigned to the ALS postmortem cortex and spinal cord. Pie charts showing patient level subtypes using the majority agreement approach described previously[99,317] in the (A) postmortem cortex and (B) the postmortem spinal cord. A larger fraction of patients was found to be "Discordant" or "ALS-TD" in the spinal cord as compared to the cortex, potentially reflecting cell type composition differences in each tissue region. When accounting for sequencing platform in the assignment of patient subtype at the individual sample level, the NovaSeq cohort more closely mirrors the subtype proportion in the postmortem cortex. Yet, in both cases, the ALS-Ox subtype was not the most common, pointing to cell composition effects given work from Humphrey *et al*.[151]

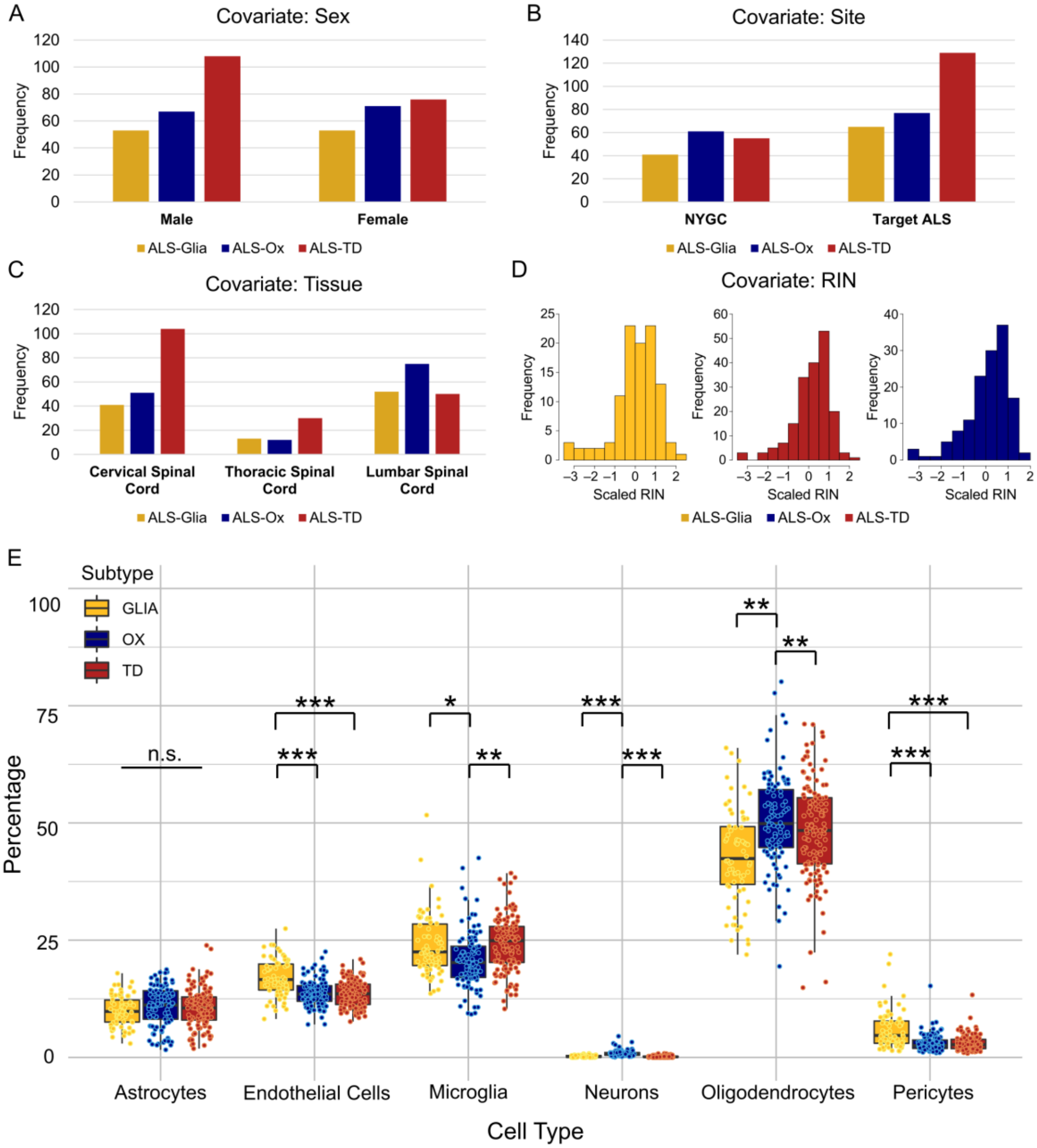


Figure 2.13: Diagnostic plots show spinal cord subtype is partially influenced by multiple covariates after dependent gene removal. Covariate diagnostic plots following assignment of sample subtype using non-smooth non-negative matrix factorization showing (A) sex, (B) site of collection, (C) tissue region, and (D) RIN. While scaled RIN does not appear to have a strong effect on the assigned subtype, the remaining covariates are seen to influence subtype despite removal of covariate-dependent genes using differential expression. Strictly considering this spinal cord cohort, the NYGC site of collection/processing and the lumbar tissue region appear more robust to covariate-dependent gene expression. (E)

Estimated cell type fractions in the postmortem spinal cord (n = 293 unique transcriptomes, 137 cervical, 36 thoracic, 120 lumbar), considered in the context of the ALS subtypes. Estimates were previously calculated by Humphrey *et al*.[151] using the MuSiC algorithm[342] with reference single-nucleus RNA-seq data from Mathys *et al*.[218] Significant differences in cell type percentages were assessed using a two-sided Wilcoxon rank sum test with Bonferroni *p*-value adjustment. Adjusted *p*-values are denoted using the following scheme: *** $p < 1\text{E-}5$; ** $p < 0.001$; * $p < 0.05$. Pairwise comparisons not depicted have an adjusted *p*-value > 0.05.

2.3.5 Cellular Deconvolution

During bulk tissue RNA-sequencing, mRNA is extracted and purified from many cells, a process that is usually controlled using the initial weight of the tissue sample. As single cell sequencing technologies have shown, the individual cellular composition of most tissue is heterogeneous, varying spatially, temporally, and between organisms of the same species. Therefore, measured gene expression in bulk tissue RNA-sequencing is generally influenced, to some extent, by the varying cellular composition of the tissue, rather than reflect some underlying disease state or process[151,259]. Consequentially, the interpretation of the stratified subtypes is limited without an improved understanding for cell type composition in the bulk tissue transcriptomes. As previously discussed, cellular deconvolution is a matrix decomposition strategy which leverages reference single-cell RNA-seq expression data and the original bulk tissue transcriptome to identify the contribution of individual cell types to the overall expression signature.

In an effort to address potential biases in the cortex cohort during bulk tissue sequencing, cell deconvolution was performed using CIBERSORTx[244], with DESeq2 normalized count values (additional details in Methods section) and reference single cell expression from Nowakowski *et al*.[247] as seen in Figure 2.14. Ten cell type signatures (including "Unknown") were generated from the single cell expression and used to estimate cell percentages in the bulk expression data. Significant differences between prefrontal and motor cortices are observed in microglial, glial progenitor, vascular cell, and inhibitory neuron fractions as seen in Figure 2.14A. Weak significant differences are observed in the excitatory neurons, and no significant differences are seen in the astrocytes. Taken together, these findings indicate cell percentages in the frontal and motor cortex may partially explain subtype-specific expression, although tissue region in the CNS does not strongly influence the assignment of ALS subtype, as seen in Figure 2.15A. When considering cell type percentages in each subtype, some significant differences were observed, as seen in Figure 2.14B. The ALS-Ox subtype had a greater average percentage of excitatory neurons as compared to the ALS-Glia subtype (Bonferroni-adjusted *p*-value < 1E-5). Similarly, the ALS-Ox subtype demonstrated a significantly greater percentage of inhibitory neurons as compared to the other two subtypes. These results suggest that the ALS-Ox phenotype is partially driven by bulk tissue cell fractions, yet these differences are small in the case of ALS-Ox versus ALS-TD patients, supporting neuronal stress and altered inhibition as hallmarks of the ALS-Ox subtype. The percentage of endothelial and mural cells in ALS-Ox postmortem cortices suggests expression implicating blood-brain barrier dysfunction may be driven by bulk tissue biases. Some significant differences in microglial fraction are observed between the Glia and Ox subtypes (Bonferroni-adjusted *p*-value < 1E-9) and Glia and TD subtypes (Bonferroni-adjusted *p*-value < 1E-7), suggesting that differences in cell type fractions may, in part, explain elevated expression of microglial marker genes in ALS-Glia patients. However, it is important to emphasize that no significant differences in astrocyte fraction were observed between the ALS-Glia subtype and the other two subtypes, indicating upregulated neuroinflammatory signaling in ALS-Glia patients remains a defining characteristic. Cell deconvolution was also performed on the healthy control and FTLD patients, with results presented in Figure 2.15B.

In the postmortem spinal cord, cell fractions were previously estimated by Humphrey *et al*.[151] using the MuSiC algorithm[342] with reference single-nucleus RNA-seq data from Mathys *et al*.[218] A total of 293 patient samples are shown in Figure 2.13E, 137 from the cervical region of the spinal cord, 36 from thoracic, and 120 from the lumbar region. Similar to cell deconvolution results from the postmortem cortex cohort, no significant differences in astrocyte percentages were observed between subtypes in the spinal cord cohort. Also mirroring the cortex results, significant differences between subtypes were observed in neuronal, endothelial, pericyte, and microglial cell fractions. Collectively, results from cellular deconvolution demonstrate that the findings of this work are partially influenced by bulk tissue RNA-sequencing limitations, however unique astrocyte gene expression in the ALS-Glia subtype suggests that the observation of distinct disease states during stratification cannot be attributed to analytical bias alone. The stratification of ALS patients using single-cell RNA-sequencing data is likely to provide significant new insight into this challenge.

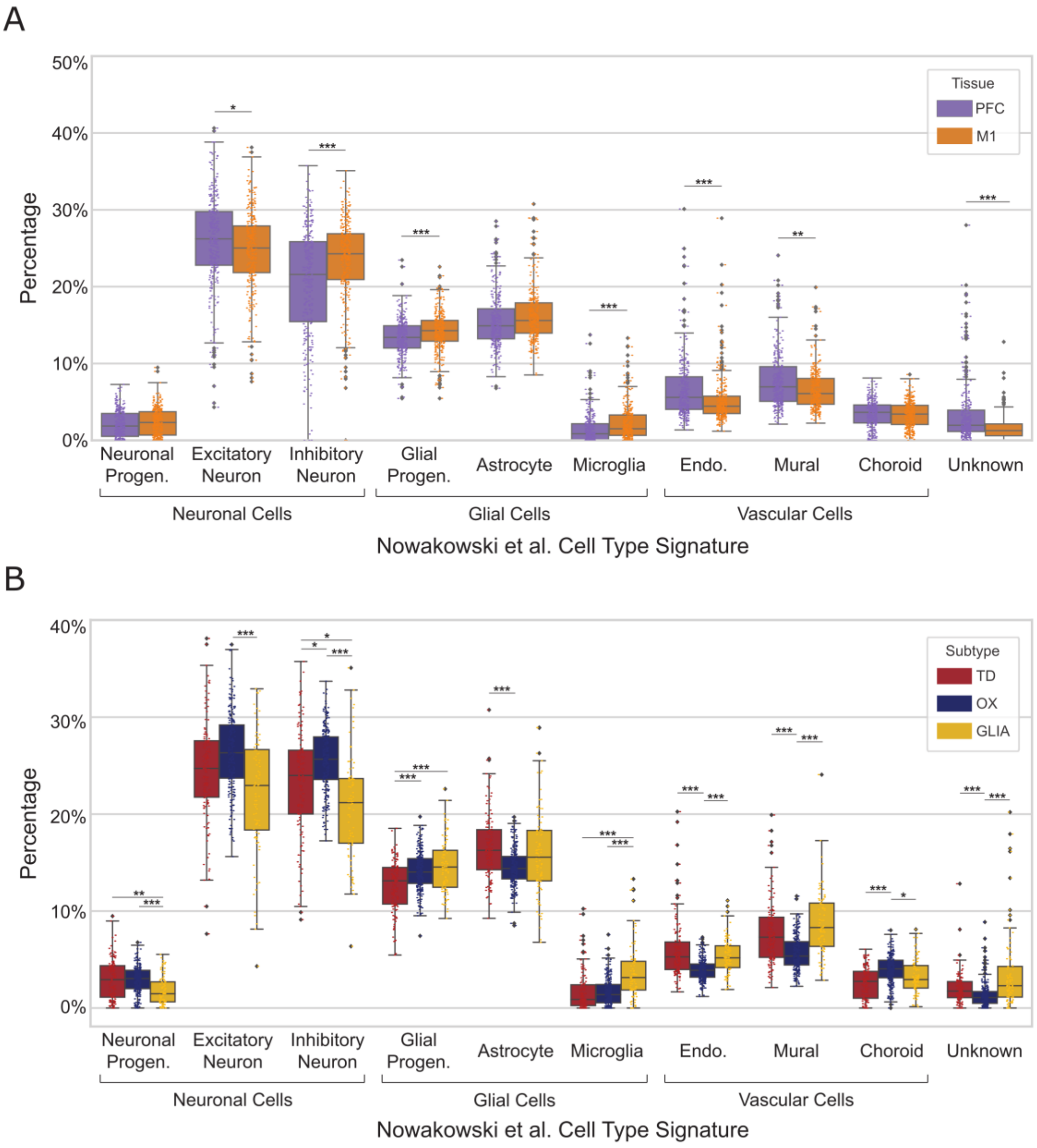


Figure 2.14: Bulk tissue cell deconvolution in ALS subtypes. (A) Cell type percentages in the prefrontal and motor cortices for all patient samples considered in this study. (B) Fractions of cell types in the frontal and motor postmortem cortex, considered in the context of the ALS subtypes. Significant differences in cell type percentages were assessed using a two-sided Wilcoxon rank sum test with Bonferroni *p*-value adjustment. Adjusted *p*-values are denoted using the following scheme: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. The median is indicated by the solid black line, and first and third quartiles are captured by the bounds of the box. Boxplot whiskers are defined as the first and third quartiles –/+ interquartile range times 1.5, respectively, and outliers are denoted as solid black points. Minimum and maximum values are captured by the lowermost and uppermost points, respectively, or whisker bound if no outliers are shown.

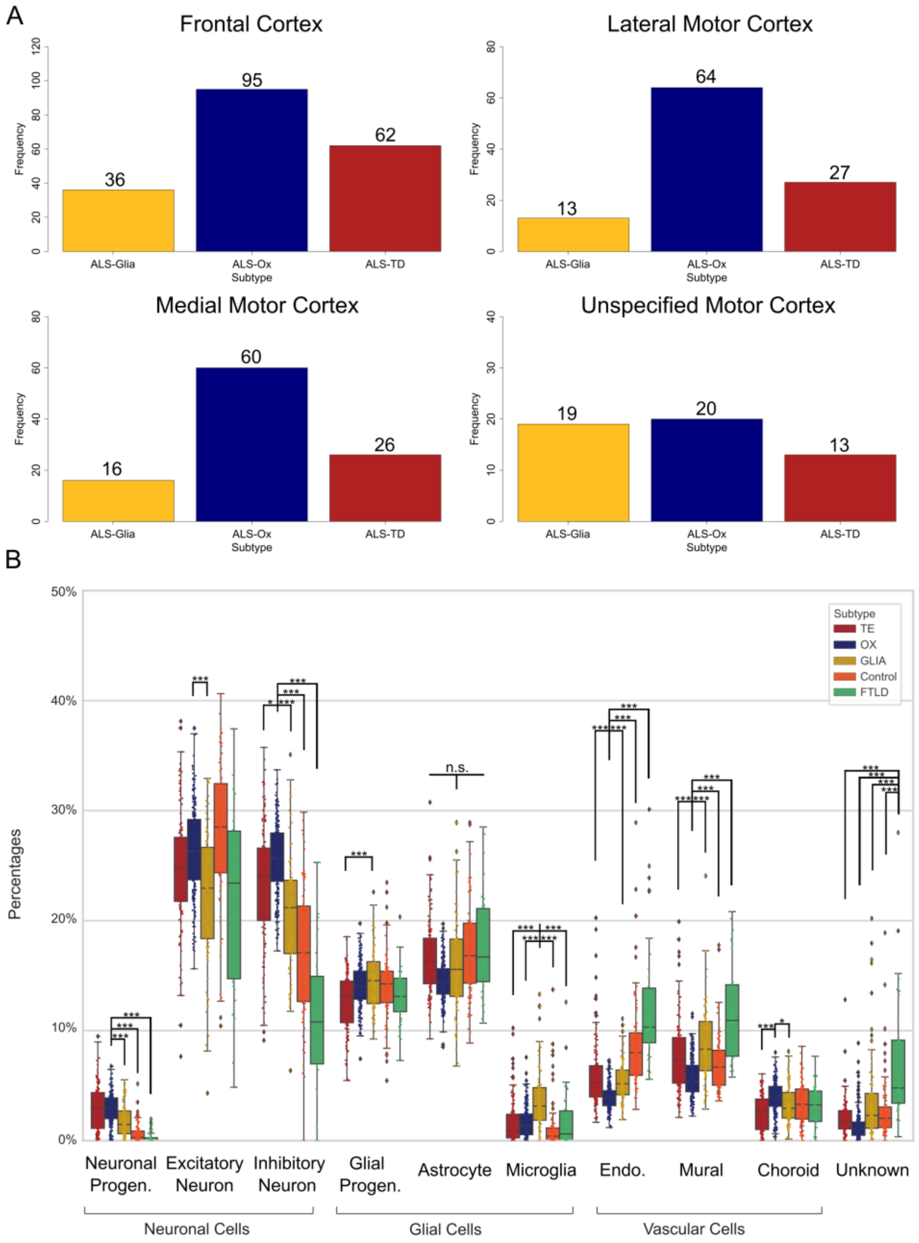


Figure 2.15: (A) Patient subtypes in each tissue region considered in this study. Approximately the same ratio of Glia, Ox and TD patients are observed in the frontal and specified motor cortices, roughly matching the ratio observed during unsupervised clustering, indicating brain region is not a confounding factor with subtype. (B) Cell type percentages in the frontal and motor cortex, considered in the context of ALS patients and controls (*n*=585). Cell deconvolution was performed using CIBERSORT[244], with DESeq2

median-of-ratio counts and references expression from Nowakowski *et al.*[247]. Significant differences in cell populations were assessed using a Wilcoxon rank sum test (two-sided) with Bonferroni adjustment. Adjusted p-values are denoted using the following scheme: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. n.s. – not significant. The median is indicated by the solid black line, and first and third quartiles are captured by the bounds of the box. Boxplot whiskers are defined as the first and third quartiles –/+ interquartile range times 1.5, respectively, and outliers are denoted as solid black points. Minimum and maximum values are captured by the lowermost and uppermost points, respectively, or whisker bound if no outliers are shown.

2.3.6 Subtype Concordance in the Postmortem Cortex and Spinal Cord

Having performed the independent clustering and enrichment of the postmortem cortex and spinal cord cohorts and arriving at similar conclusions in both, intra-patient agreement (concordance) was then assessed. As shown in Figure 2.1D, the large number of shared patients between the two postmortem cohorts allowed for the comparison of subtypes presented throughout the ALS central nervous system. Subtype labels from the frontal, medial motor, lateral motor, and ‘unspecified’ motor cortex were combined with labels from the cervical, thoracic, and lumbar regions of the spinal cord, where available. The average number of postmortem tissue transcriptomes available for each patient was approximately four.

Concordance was considered at the sample level, to ensure independence, and presented as a matrix of pie charts, with spinal cord region along the rows and cortical region along the columns, as seen in Figure 2.16. Excluding the unspecified motor cortex samples, the highest concordance was found between the frontal cortex and lumbar region of the spinal cord for the ALS-Glia subtype (46.2%), the medial motor cortex and lumbar region for the ALS-Ox subtype (46.4%), and the medial motor cortex and cervical spinal cord for the ALS-TD subtype (69.6%). The higher concordance observed in the ALS-TD subtype, and generally lower overall agreement, likely stems from bias towards the ALS-TD subtype during unsupervised clustering, despite removal of covariate dependent genes as seen in Figure 2.13A-D, and differences in cell type composition as compared to the cortex as seen in Figure 2.13E. Similarly, the lumbar region is seen to more closely reflect the proportion of subtypes observed in the cortex[99], as seen in Figure 2.13C, generally corresponding to a noticeable improvement in concordance between the cortex and lumbar spinal cord relative to other spinal cord regions for ALS-Ox and ALS-Glia patients, as shown in Figure 2.16. More generally, concordance between the lumbar spinal cord and each region of the cortex, excluding the lateral motor, is statistically significant with agreement being higher than would be expected by random chance. Using bootstrapping, distributions for the expected number of concordant patient samples were generated from 10,000 iterations, after adjusting sampling probabilities to reflect subtype observations in the cortex and spinal cord. True concordant values were compared against the derived distribution for estimation of *p*-values, assuming a one-tailed binomial distribution, which were found to be 0.012 in the frontal cortex, 0.031 in the medial motor cortex, 0.160 in the lateral motor cortex, and 0.021 in the ‘unspecified’ motor cortex. Statistically significant concordance was also observed between the cervical region of the spinal cord and the frontal cortex ($p = 0.026$).

The concordance analysis was extended by further separating patients by sequencing platform to assess dependence on instrumentation, as shown in Figures 2.17, 2.18, and 2.19. The NovaSeq 6000 sequencing platform was found to outperform the HiSeq 2500 sequencing platform in the cervical (42.3% vs 35.4%) and lumbar spinal cord (47.7% vs 35.4%) but not the thoracic spinal cord (32.0% vs 37.9%). In the NovaSeq cohort, the highest concordance for the ALS-Glia subtype remains the same tissue pairing at 48.0%, and for ALS-Ox the highest agreement was seen between the lateral motor cortex and lumbar spinal cord at 48.4% (excluding the unspecified motor cortex and pairings with a single observation) – although lower sample numbers may partially explain these differences. In the NovaSeq subset, intra-patient concordance between the postmortem cortex and lumbar region of the spinal cord remains statistically significance in the frontal cortex (p = 0.024), medial motor cortex (p = 0.041), lateral motor cortex (p = 0.037), and 'unspecified' motor cortex (p = 0.010). Additionally, concordance was significant between the cervical region of the spinal cord and the frontal cortex (p = 0.024) and 'unspecified' motor cortex (p = 0.050). No tissue pairings were found to be statistically significance in the HiSeq subset. Collectively the analysis shows weak to moderate agreement in subtype presented throughout the ALS postmortem cortex and spinal cord, despite differences in cell type composition, after removal of covariate-dependent genes. Concordance between the cortex and spinal cord phenotype was demonstrated to be dependent on tissue and sequencing platform, and find the lumbar region of the spinal cord shows the highest overall concordance with the cortical phenotype in this cohort. Bootstrapping shows that the proportion of concordant patients in the NovaSeq subset consistently exceeds the number expected by random chance when comparing any region of the cortex with the lumbar region of the spinal cord, improving the likelihood that patient subtypes are expressed throughout the individual’s central nervous system and influence the disease course.

Patient concordance was further considered by screening for patients assigned the same subtype in every sample considered in this study and the previous[99]. A total of 45 patients were found to pass this criterion (45/222; 20.3%) and stratify these patients further to find five ALS-Glia patients, 12 ALS-Ox patients, and 19 ALS-TD patients coherently assigned a single subtype in both the cortex and spinal cord, as shown in Figure 2.20. While the lower patient number limits the extrapolation of these findings, a surprising association with sex and a stark difference in disease duration in this concordant patient subset is observed, as observed in Figure 2.20B-E.

Despite the statistical limitations (sample independence) of assigning patient subtype using the majority agreement approach, concordance between the majority postmortem cortex subtype[99] and majority spinal cord subtype are considered using this method. As shown in Figure 2.21, more than half of the patients were found to be discordant (68.2%), which likely reflects differences in cell type composition between the cortex and spinal cord and highlights the challenges of linking patient phenotype between these two regions.

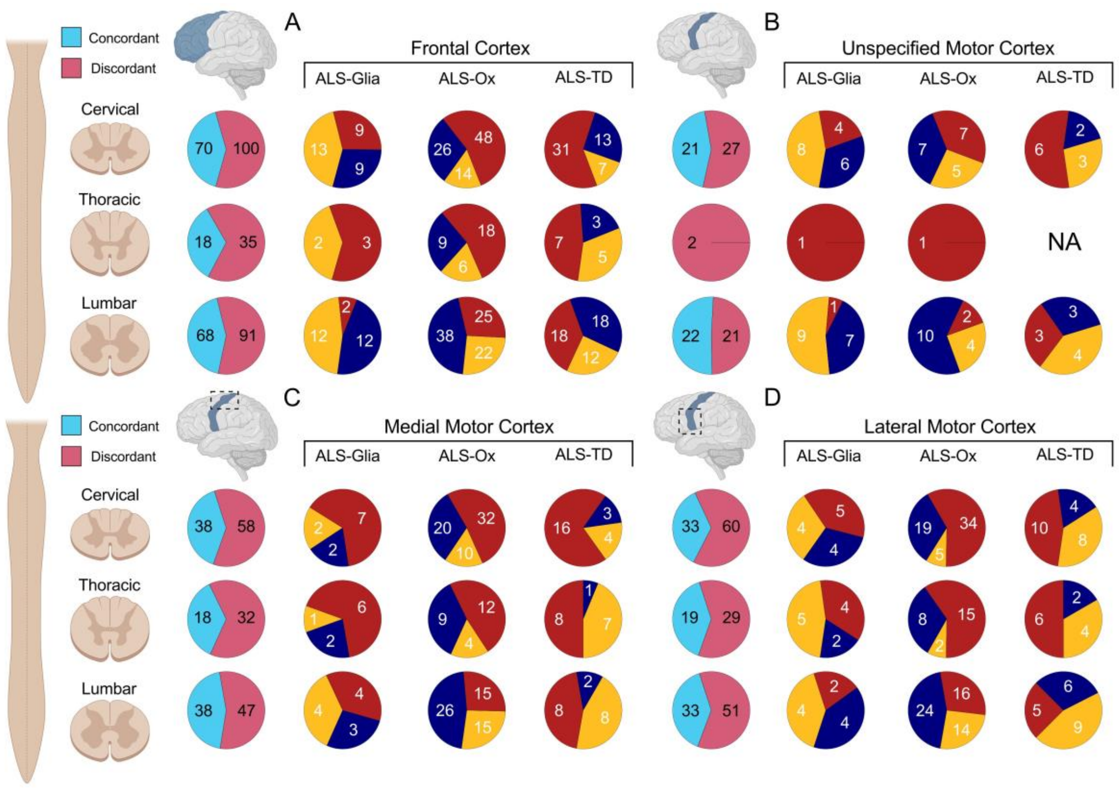


Figure 2.16: Agreement between the subtype assigned to the cervical, thoracic, and lumbar regions (*rows*) and (A) the frontal cortex, (B) the unspecified motor cortex, (C) the medial motor cortex, and (D) the lateral motor cortex – for all available samples. Pie charts are first presented as an aggregate of all paired tissue samples (light blue and pink) and in a subtype-specific manner. Concordance at the subtype level (*columns*) has been color coded to indicate agreement (gold, navy, and maroon) or disagreement (orange, green, and purple) between the two tissue regions compared. No patients assigned ALS-TD in the unspecified motor cortex had a corresponding thoracic spinal cord sample in this cohort. Created with BioRender.com.

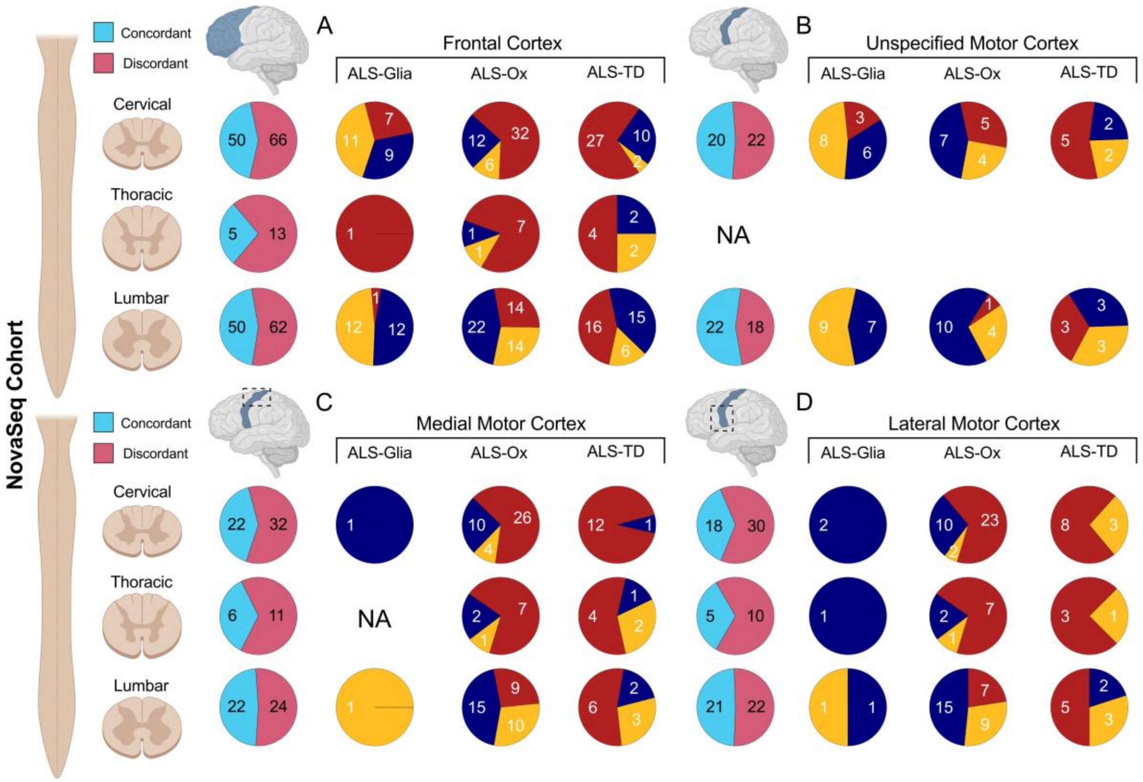


Figure 2.17: Tissue specific concordance between the postmortem cortex and spinal cord in the NovaSeq cohort. Agreement between the subtype assigned to the cervical, thoracic, and lumbar regions (*rows*) and (A) the frontal cortex, (B) the unspecified motor cortex, (C) the medial motor cortex, and (D) the lateral motor cortex – for all available NovaSeq spinal cord samples. Pie charts are first presented as an aggregate of all paired tissue samples (light blue and pink) and in a subtype-specific manner. Concordance at the subtype level (*columns*) has been color coded to indicate agreement (gold, navy, and maroon) or disagreement (orange, green, and purple) between the two tissue regions compared. No patients assigned ALS-TD in the medial motor cortex had a corresponding thoracic spinal cord sample in this cohort. There were no intra-patient tissue pairings between the thoracic spinal cord and unspecified motor cortex in this cohort. Created with BioRender.com.

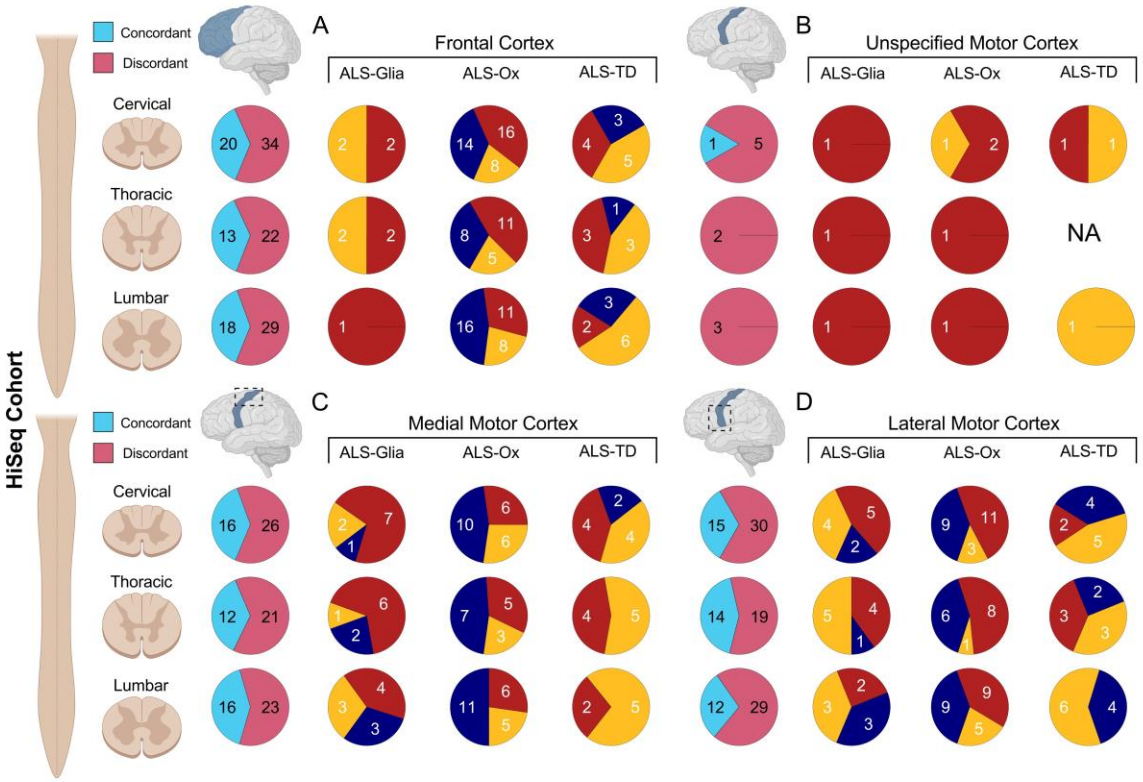


Figure 2.18: Tissue specific concordance between the postmortem cortex and spinal cord in the HiSeq cohort. Agreement between the subtype assigned to the cervical, thoracic, and lumbar regions (*rows*) and (A) the frontal cortex, (B) the unspecified motor cortex, (C) the medial motor cortex, and (D) the lateral motor cortex – for all available HiSeq spinal cord samples. Pie charts are first presented as an aggregate of all paired tissue samples (light blue and pink) and in a subtype-specific manner. Concordance at the subtype level (*columns*) has been color coded to indicate agreement (gold, navy, and maroon) or disagreement (orange, green, and purple) between the two tissue regions compared. No patients assigned ALS-TD in the unspecified motor cortex had a corresponding thoracic spinal cord sample in this cohort. Created with BioRender.com.

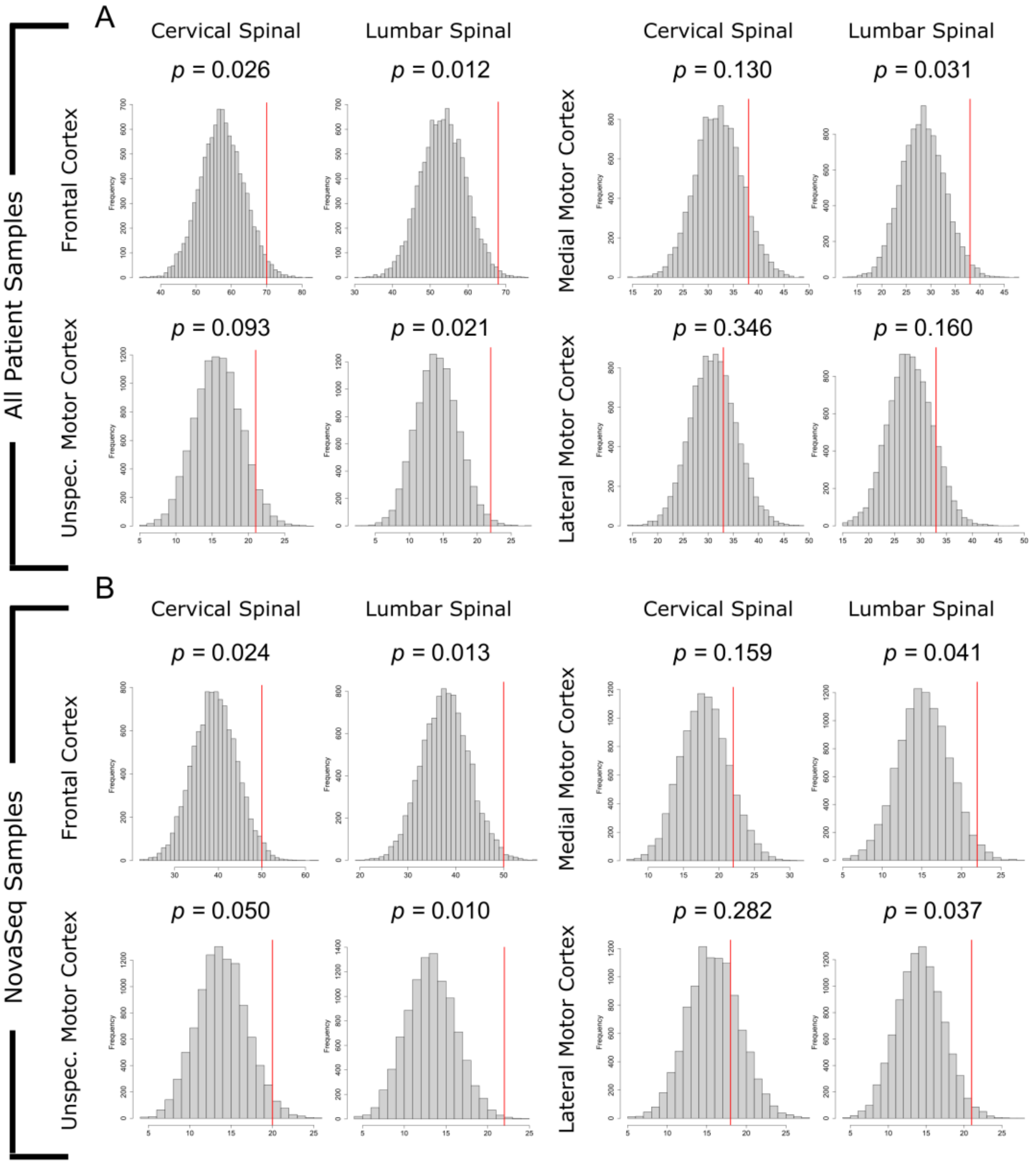


Figure 2.19: Intra-patient concordance *p*-values, estimated from 10,000 rounds of bootstrapping after independently adjusting for the proportion of patient subtypes observed in the cortex and spinal cord cohorts. The red line indicates the true number of concordant samples observed for a given cortex – spinal tissue pairing. *P*-value estimates are provided for (A) all patient samples and (B) the NovaSeq subset exclusively**.** The NovaSeq platform generally outperforms the HiSeq, with no tissue pairings having statistical significance in the HiSeq subset. No significant agreement was observed in the thoracic spinal cord in any cohort.

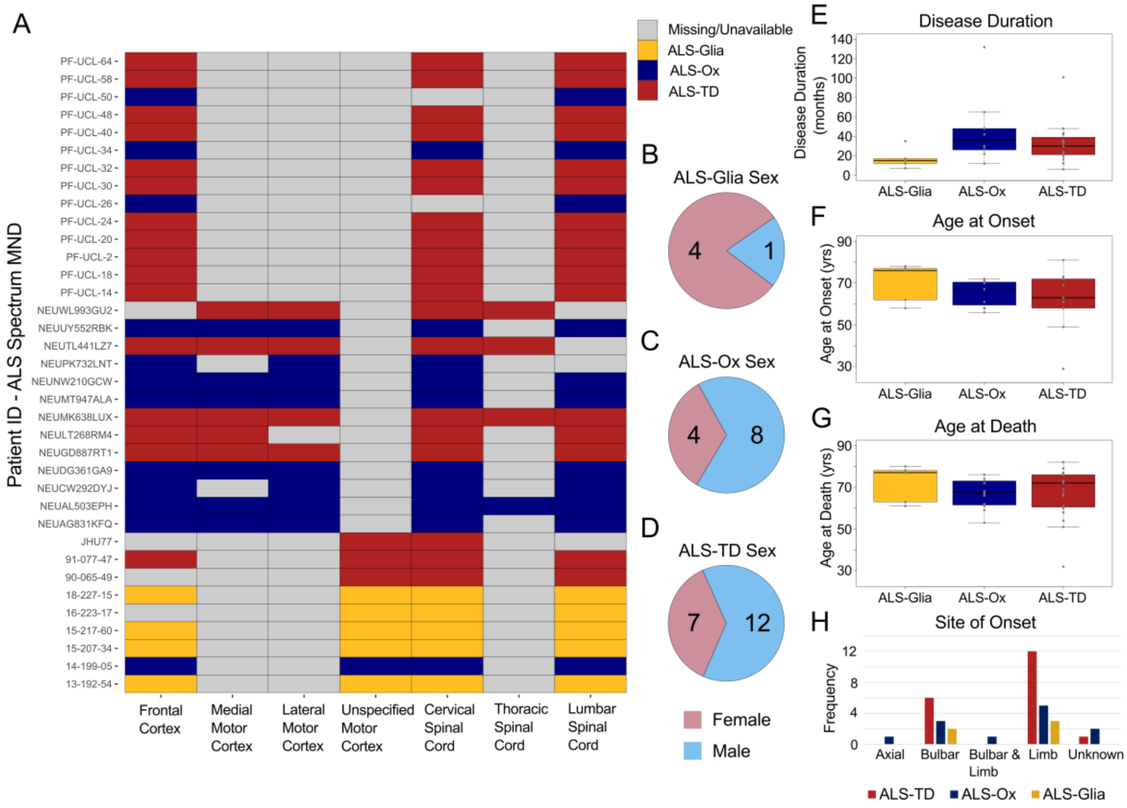


Figure 2.20: ALS patients with subtype concordance throughout the central nervous system. Multiple patients demonstrated perfect concordance across all available tissue transcriptomes and are presented using publicly available de-identified IDs. (A) Heatmap showing subtype assignment to each region of the cortex and spinal cord considered in this study. Gray cells correspond to unavailable or not applicable tissue transcriptomes. Patients without observations from both the postmortem cortex and spinal cord are excluded from this plot, but are available in Supplementary Dataset 4. Patient sex is presented for (B–D) each ALS subtype. Interestingly, patients concordant for the ALS-Glia subtype are primarily female, potentially indicating sex-dependent differences in the presentation of disease phenotype. Clinical parameters for concordant patients are plotted as boxplots, and show (E) disease duration, (F) age at onset, (G) age at death, and (H) site of symptom onset. Statistical tests were not performed due to limited patient number.

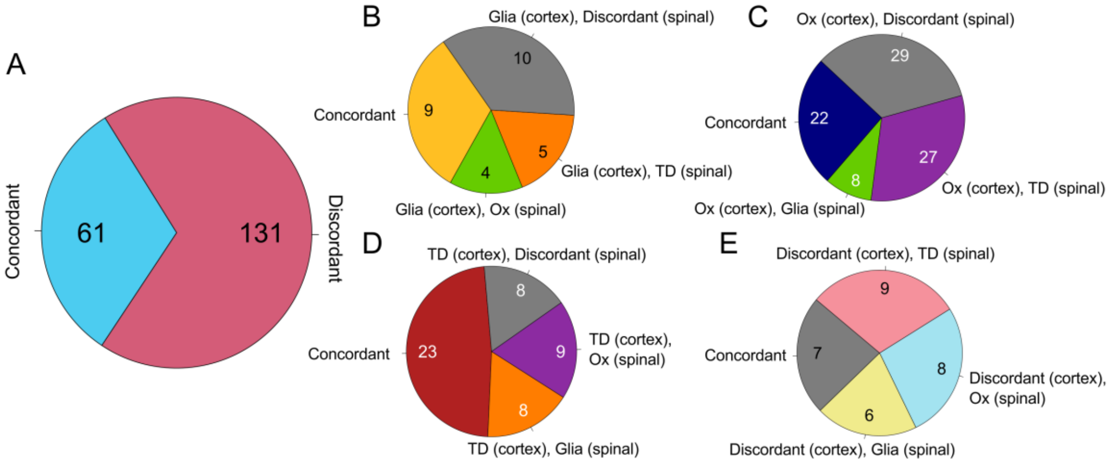
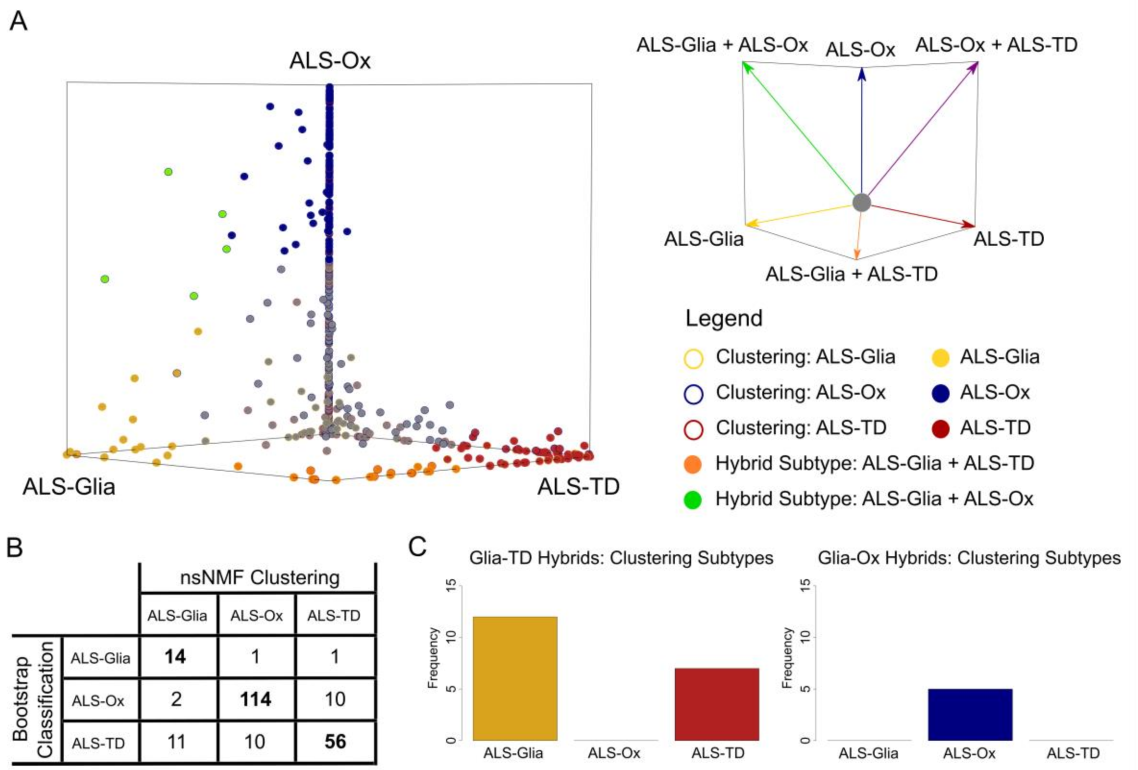


Figure 2.21: Concordance between the postmortem cortex and spinal cord using the majority agreement approach. A meta level analysis comparing subtype assigned in the cortex with the spinal cord using the majority agreement approach – in which patients were assigned a subtype if a single sample was available or by majority if two or more samples were available. Patient subtype was assigned in the cortex and spinal cord independently. (A) The majority (68.2%) of patient samples did not show concordance between the postmortem cortex and spinal cord when using the majority agreement approach – indicating this method is not the optimal way to manage repeat patient sampling. (B) For patients that presented as ALS-Glia in the cortex, ~32% of individuals were assigned the same subtype in their spinal cord by majority agreement. Discordant was the most common subtype assigned, likely reflecting limitations due to cell type composition of the spinal cord. (C) Patients with the oxidative stress phenotype in the cortex demonstrated similar concordance, with ~25% of patients assigned the same subtype in their spinal cord. Encouragingly, very few patients classified as ALS-Ox in the cortex were assigned ALS-Glia in their spinal cord, suggesting cell type composition does not act as a confounding factor in the expression of ALS-Ox marker genes in the spinal cord but weakens the detectable signal. (D) Patients presenting as ALS-TD in their cortex showed the highest concordance with the spinal cord (~48%), but likely reflects bias towards this subtype in the spinal cord transcriptomes. Interestingly, this bias towards ALS-TD in the spinal cord does not appear to be due to RIN, given that the mean RIN was 6.14 for all cortex samples yet 6.50 for all samples from the spinal cord. (E) Patients initially showing discordance for their postmortem cortex subtype are reassigned to ALS subtypes with roughly the same frequency. The concordance label in this figure indicates both the postmortem cortex and spinal cord presented as “Discordant” by the majority agreement method.

2.3.7 Presentation of Hybrid Subtypes

As shown in Figures 2.5, 2.6, 2.7, 2.8, 2.9, and 2.11, evidence for the co-presentation of ALS phenotypes within the ALS cortex cohort is observed. Therefore, to better understand the transcriptional landscape of these ALS molecular subtypes, the classification approach outlined by Patel *et al*.[259] was leveraged. This classification strategy leverages subtype-specific and co-expressed gene subsets (“eigengenes”) obtained from the WGCNA analysis discussed in Chapter 3. As detailed in the methods section above, subtype scores were calculated using predictor gene sets derived from the ALS-Glia (gold), ALS-Ox (navy), and ALS-TD (maroon) eigengenes. Bootstrapping was applied to define 95% confidence intervals on the expected subtype scores and assign patient samples to one or more subtypes if they passed the 5% threshold. As shown in Figure 2.22, the majority of classified patient samples demonstrated gene expression characteristic of a single subtype (220/244). However, for a subset of patients, hybrid gene expression characteristic of both the ALS-Glia and ALS-TD subtypes (n=19), as well as the ALS-Glia and ALS-Ox subtypes (n=5) was observed. Interestingly, despite shared disease themes between the ALS-Ox and ALS-TD groups as seen in Figure 2.8C-D, the bootstrap-based classification approach shows that these two subtypes are generally expressed independently. Furthermore, no patient samples were seen to express all three subtypes simultaneously, evident by the fact that all samples fall along one of the three faces of the hexagonal plot in Figure 2.22A. Sample subtypes obtained from the unsupervised clustering analysis are encoded as border colors, and generally show agreement between the two approaches, shown in Figure 2.22B. All patient samples shown to express a hybrid ALS phenotype were initially clustered into one of the two subtypes comprising the hybrid state, as shown in Figure 2.22C, further supporting the interpretation of this analysis. Taken together, the results capture the heterogeneous spectrum of ALS disease phenotypes in this cohort and reveal that a small subset of ALS postmortem cortex transcriptomes show evidence for the co-presentation of subtype states.

Stemming from the challenges encountered during stratification and similar phenotype presentation in ALS-Glia and ALS-TD patients when using spinal cord gene expression, hybrid subtyping analysis was not performed in the spinal cord cohort. As shown in Figure 2.11, 2.13 and 2.14, the higher proportion of glia cells in the spinal cord, relative to the cortex, likely drives some of the shared pathways and mechanisms found to be enriched in the two subtypes. These effects propagated forward into the WGCNA analysis performed in *Chapter 3.3.5*, and made definition of ALS-Glia and ALS-TD eigengenes difficult. Thus, consideration of hybrid subtype presentation in the spinal cord cohort would likely show overestimation of Glia-TD hybrids and reflect a severe dependence on cell type composition.



Bootstrap Classification	nsNMF Clustering: ALS-Glia	ALS-Ox	ALS-TD
ALS-Glia	**14**	1	1
ALS-Ox	2	**114**	10
ALS-TD	11	10	**56**

Figure 2.22: Score-based classification uncovers hybrid subtype states in the ALS cortex cohort. (A) Subtype scoring was implemented with bootstrapping to assess the spectrum of disease phenotypes presented in ALS. Each point corresponds to a single transcriptome derived from the frontal or motor postmortem cortex, *n*=451 biologically independent samples. Patient samples were initially placed at the origin, moved in the direction of the subtype axis for each round of bootstrapping that passed the subtype score threshold, and could only reach the vertex if the patient sample passed the threshold in all rounds of bootstrapping. Data points are filled according to the bootstrap-based subtype assignment and borders are included to denote the patient subtype obtained from unsupervised clustering. Transcriptomes which approached the vertices shared by two subtypes are considered to express a hybrid subtype state. Patient samples are color coded gray if they failed to pass the subtype score thresholds in ≥ 50% of bootstrap iterations. (B) Confusion matrix showing unsupervised clustering results in each classification subtype. (C) Clustering results in Glia-TD and Glia-Ox hybrids.

2.4 Discussion and Conclusion

Limited understanding for variable ALS onset and progression has limited clinical trial success and slowed the development of effective therapeutics. Collectively, unsupervised clustering and enrichment results from the postmortem cortex and spinal cord cohorts converge on similar phenotypes, lending strength for the contribution of independent, yet occasionally concurrent, molecular subtypes to ALS disease heterogeneity. In both postmortem cohorts, ALS patient transcriptomes[266] were stratified into three subtypes defined by distinct gene expression phenotypes, termed ALS-Glia[317], ALS-Ox[317], and ALS-TD. Guided by enrichment analyses, gene expression associated with activated glia and neuroinflammatory signaling are observed in the ALS-Glia subtype. Although enrichment scores take opposite magnitudes in the ALS-Glia cortex and spinal cord relative to non-neurological controls, results continue to implicate neuroimmune pathologies relative to the other two subtypes, and likely reflect more stringent gene filtering applied to spinal cord transcriptomes to address cell compositional differences. Despite these differences, the ALS-Ox subtype was found to be highly conserved in the ALS central nervous system. Gene expression primarily implicates altered synaptic signaling in both the postmortem cortex and spinal cord with oxidative and proteotoxic stress also elevated in the cortex – more typical neurodegenerative themes. Consideration of locus-specific transposable elements via SQuIRE revealed that both the ALS-TD and ALS-Ox subtypes strongly overexpressed TEs compared to healthy control donors and ALS-Glia patients in the cortex. In ALS-TD cortex samples, the unique expression of transcription and translation associated genes was observed, including transcription factors, regulatory microRNAs, mRNA traditionally marked for nonsense mediated decay, pseudogenes, antisense, intronic, and long non-coding RNAs. These findings led us to define the final subtype by transcriptional dysregulation, rather than transposable element expression. Features generally indicating transcriptional dysregulation in the ALS-TD spinal cord continued to be observed, including pseudogenes, antisense, read-through transcripts, and regulatory RNA.

Findings from this Chapter offer an important postmortem foundation for future ALS patient stratification in the clinic and prior to enrollment in clinical trials. Clustering and enrichment analyses, in agreement with prior studies[8,237,317], jointly indicate separate pathological mechanisms in each subtype, offering one explanation for variable and generally weak patient response to existing therapies like Riluzole and Edaravone. Further, enrichment analyses offer new perspectives on subtype-specific therapeutic targets, providing a pharmacological foundation for future drug discovery. Choice of sequencing platform was seen to influence the allocation of patient samples in the spinal cord cohort but not in the cortex cohort after carrying out the analyses separately for samples derived from NovaSeq 6000 and HiSeq 2500 instruments. Further, bulk tissue RNA-sequencing and variability in cell composition is seen to influence the overall expression profile and restricts the interpretation and utility of these findings to some extent. Finally, considering agreement in subtype presentation at the patient level, modest but statistically significant agreement is observed between the subtype(s) expressed in the postmortem cortex and spinal cord, likely influenced by both technical and biological confounding factors. Results show that the lumbar region of the spinal cord generally demonstrates the highest concordance with all regions of the cortex, regardless of the individuals' subtype. In addition, the NovaSeq platform appears to outperform the HiSeq platform when using intra-patient concordance as the measure. Importantly, concordance between the lumbar spinal cord and frontal, medial motor, lateral motor, and 'unspecified' motor cortex regions in the NovaSeq subset is higher than would be expected by random chance (p = 0.003–0.027), after adjusting for subtype proportions observed in each cohort, as shown in Figure 2.18, strengthening claims for the existence and pathological relevance of the neurodegenerative subtypes in ALS patients.

Chapter 3

PATIENT SUBTYPES CAPTURE SURVIVAL DIFFERENCES AND INDICATE ASSOCIATIONS WITH AGE AT DISEASE ONSET

3.1 Introduction

ALS patient variability in disease onset, age, presentation of symptoms, presence of comorbidities like Alzheimer's disease or FTLD, rates of functional decline, and overall duration have long been reported to vary, even in relatively small cohorts with similar underlying characteristics or in cohorts with different genetic mutations on the same protein[67,81,170,284,315]. As previously discussed in *Chapter 2*, the use of a patient stratification system based on the site of symptom onset poorly predicts disease progression and patient outcomes[55,339]. Further, patient heterogeneity represents a major complication in the design of clinical trials[315] and the identification of reliable and quantifiable molecular features to aid patient stratification and improve success rate remains a potent need. As noted in *Chapters 1.1* and *1.2.1*, genetic models of the disease have offered evidence for a spectrum of mechanistic disease pathways in ALS neurodegeneration. Linking mechanistic heterogeneity to clinical variability would provide an important foundation for improving clinical trial success and guide development of more personalized therapeutics.

Offering key insight into the extent of mechanistic heterogeneity in patients, groups report incomplete recovery of all phenotypic aspects of ALS using patient-derived stem cell models, and further show dependency on the gene mutation used as the basis of the

ALS disease model[60,176,223,338]. Further capturing mechanistic heterogeneity, groups leveraging the C9orf72 hexanucleotide repeat expansion as the genetic ALS model frequently report the presence of aggregated dipeptide repeats of varying amino acid composition, cellular effects, toxicity, and localization[15,113,214]. Similar conclusions are found when leveraging animal models of ALS, where genetic background of the mouse has been shown to influence microglial activation and heterogeneity in pathology and progression[181]. Further, heterogeneous disease outcomes have been reported in mouse models when using the same mutant protein, TDP-43, but with different mutation loci (M337V and Q331K)[7]. Finally, in human cohorts, clear evidence for continued mechanistic heterogeneity is demonstrated in hereditary patients with the mutations to the *SOD1* protein. Despite the fact that the same protein is implicated in the pathology, significant variability in disease onset and progression are observed when considering common mutations like A4V, H46R, and D90A[170,270]. Extensive work has linked mutation variability to toxic gain of function, with proposed mechanisms implicating altered substrate affinity, metal-free protein stability, and the propensity for monomer aggregation, rather than loss of dismutase (enzymatic) activity[23,47,199,267,321,351]. Collectively, these works and others show that a multitude of factors likely contribute to the heterogeneity observed in patients and emphasize the challenges associated with the stratification of patients to improve clinical trial outcomes.

While the characterization of the ALS neurodegenerative spectrum provides important insight into disease variability in patients, associations with clinical parameters have not yet been established, calling into question the biological relevance of the proposed subtypes. In the third chapter of this dissertation, subtype labels from *Chapter 2* are leveraged to perform univariate and multivariate survival analyses, weakly linking the immunological subtype, ALS-Glia, to a shorter disease duration and oxidative stress subtype to a longer duration. In line with findings from survival analysis, results from WGCNA converge on the same interpretation, as the ALS-Glia eigengene is significantly negatively correlated with disease duration in the postmortem cortex. Similar gene expression profiles between ALS-Glia and ALS-TD subtypes in the spinal cord – likely a consequence of cell type composition[151] – makes assignment of subtype eigengenes challenging. However, one immunological eigengene continues to agree with findings from survival, as a significantly negative correlation is seen between eigengene ‘expression’ and disease duration – indicating a negative effect on rate of disease progression – although the eigengene shows specificity for the ALS-TD subtype.

The association of subtype with survival variability represents an important step towards the beneficial stratification of patients, but as Cox Regression shows, only remains part of the picture as other factors like sex and age are seen to contribute more towards differences in patient hazard. Further, disease duration does not capture important clinical considerations like rate of functional decline, and future works linking expression-based subtypes to alternative parameters like ALSFRS-R, metabolic readings, or neuroimaging data may present further insight into the contribution of the expressed subtypes on clinical and mechanistic heterogeneity.

3.2 Methods

3.2.1 Assignment of Patient-Level Subtype for Clinical Analysis

To allow for the consideration of patient clinical parameters, the majority subtype approach outlined in *Chapter 2.2.11* is leveraged i) in the cortex and spinal cord cohorts separately and ii) 'globally' across all available samples from the cortex and spinal cord. In brief, patients were assigned a subtype label only if there was a majority consensus. ALS patients which failed to reach a majority consensus were labeled 'Discordant'. When applying the majority agreement approach in the later case, the patient cohort was filtered to ensure observations were available from both the postmortem cortex and spinal cord, totaling 192 individuals.

3.2.2 Assessment of Clinical Parameters in ALS Subtypes

Differences in ALS survival by subtype was assessed using the Kaplan-Meier analysis[80,167] with application of the log-rank statistical test. Left censoring was applied in the postmortem cortex cohort. Subtype-specific differences in age of symptom onset and age at death were analyzed using ANOVA tests. *Post hoc* analysis used a two-sided t-test with FDR *p*-value adjustment. A Chi-squared test of independence was applied to assess subtype dependency for site of symptom onset, genetic mutation, or FTLD and Alzheimer's comorbidity.

3.2.3 Cox Regression for Multivariate Survival

To address sample dependence due to repeat patient measures, Cox proportional hazard models were constructed in R to assess multivariate contribution to patient survival[80,169,178,324,325]. Sex, subtype, age at symptom onset, and disease group covariates are included as fixed effects and obtain hazard ratios from the exponential of the β coefficient. Regression diagnostics are separated by tissue region, and first show residuals plotted by covariate followed by plots of scaled Schoenfeld residuals for each covariate level. By running regression in each tissue region independently, the need to incorporate patient-specific random effects into the proportional hazard model is bypassed. Model term *p*-values were calculated from the coefficient z-scores, while testing of the proportional hazard assumption at the covariate level was performed using the score test, with $p < 0.05$ indicating time-dependent hazard and assumption violation[80]. All covariates in all models are observed to meet the assumption of having proportional hazards over the survival duration, excluding the disease group covariate in the lateral motor cortex model ($p = 0.04$).

3.2.4 Weighted Gene Co-Expression Network Analysis (WGCNA)

Co-expressed gene sets associated with disease duration, age of symptom onset, and age at death were assessed using the Weighted Gene Co-Expression Network Analysis (WGCNA) package in R[188] (Version 1.70-3, University of California, Los Angeles). The minimum module size was set to 25 and a soft power of 13 was selected in the cortex, while a power of six was selected in the spinal cord, to ensure the assumptions of scale-free topology are met[188]. Gene expression was maintained on the variance stabilizing transformation (VST) scale. Eigengenes were enriched for the biological process (BP) gene ontology. Transcripts without a corresponding gene symbol (ENSG and TEs) were included in WGCNA to ensure features relevant for the ALS-TD subtype were still considered.

In the cortex cohort, the top 1,000 most relevant transcripts from each sequencing platform subgroup were considered during construction of the eigengene heatmap, using variance-stabilizing transformation count values. The same scale was used in the spinal cord cohort, however the 5,000 most variable transcripts from each sequencing platform were combined, yielding a total of 8,163 transcripts. Eigengenes were assessed for upregulation or downregulation in each subtype using dummy regression, with subtype as the predictor and sample-wise eigengene expression as the response variable. For each eigengene, a linear regression model was constructed, setting one of the three subtypes to a value of 1, and the other two to a value of 0. The sign and magnitude of the β coefficient from the non-zero term reflects subtype-specific eigengene expression. Correlation and regression p-values are $-\log_{10}$ transformed prior to plotting.

3.3 Results

3.3.1 Subtype Clinical Outcomes in the Cortex Cohort

Following the assignment of cortex sample subtype from unsupervised clustering, patient clinical parameters were considered in the context of their respective subtypes. A survival analysis[167] was performed to determine whether the three molecular subtypes of ALS capture some of the clinical heterogeneity seen in patient disease duration. ALS patients (n=208) were only assigned a subtype if there was a majority consensus among frontal and motor cortex samples or a single tissue sample was characterized for a given patient (additional details in Methods section). Importantly, multiple tissue samples from the same donor are classified as the same subtype (80.8%; 126/156), lending support to the patient-level subtype assignment methodology.

Notably, the results show significant differences in patient survival, with the ALS-Glia subtype associated with the shortest disease duration and a median survival of 28 months as seen in Figure 3.1A. Pairwise comparisons using the log-rank test showed significant differences in survival between ALS-Glia and ALS-Ox subtypes ($p = 0.015$) and ALS-Glia and ALS-TD subtypes ($p = 0.0043$) but not between the ALS-Ox and ALS-TD subtypes ($p = 0.30$) after (left) censoring patients with an unknown age of onset but recorded disease duration. Consideration of patient age of symptom onset showed a nonsignificant trend toward the latest disease onset for the ALS-Glia subtype (63.2 ± 1.83 years; presented as mean ± standard error) and earliest disease onset for the ALS-Ox subtype (60.4 ± 1.16 years) as seen in Figure 3.1B. The oldest median age at death was observed for the ALS-TD subtype (66.7 ± 1.33 years) and youngest median age at death for the ALS-Ox subtype (64.0 ± 1.05 years), which likely reflects some dependency on the age of symptom onset, as observed in Figure 3.1C. As shown in Figure 3.1D, site of symptom onset shows roughly the same proportion of patients with bulbar and limb onset across the three subtypes. Subtype comorbidity for FTLD was analyzed using a Chi-square test of independence, although subtype dependency in the co-presentation of ALS and FTLD was not observed ($p = 0.59$).

This analysis was also performed with ALS patients that did not reach a majority agreement for subtype presentation in available tissue samples, termed ALS-Discordant[317]. Among the 208 unique patients in the cortex cohort (n=95 female, n=113 male), 30 were found to be discordant (n=17 female, n=13 male), as summarized in Table 3.1. As seen in Figure 3.2, similar results are observed, with significant differences in patient survival ($p < 0.05$) and the latest age of onset maintained for the ALS-Glia subtype (nonsignificant). Patient clinical parameters are also considered in the context of the hybrid subtypes identified in *Chapter 2.3.7,* with results presented in Figure 3.3. In addition, given the large number of patient transcriptomes shared between this cohort and important foundational work from Tam *et al*.[317], agreement of subtype labels for the 140 samples in common were considered, as presented in Figure 2.1. 85% agreement (119/140) in sample classification is reported, despite differences in the features used for patient stratification, with results presented in Table 3.2. Finally, given known genetic associations with ALS, *C9orf72* and *SOD1* mutation frequency is considered in the context of the identified subtypes, presented in Figure 3.4.

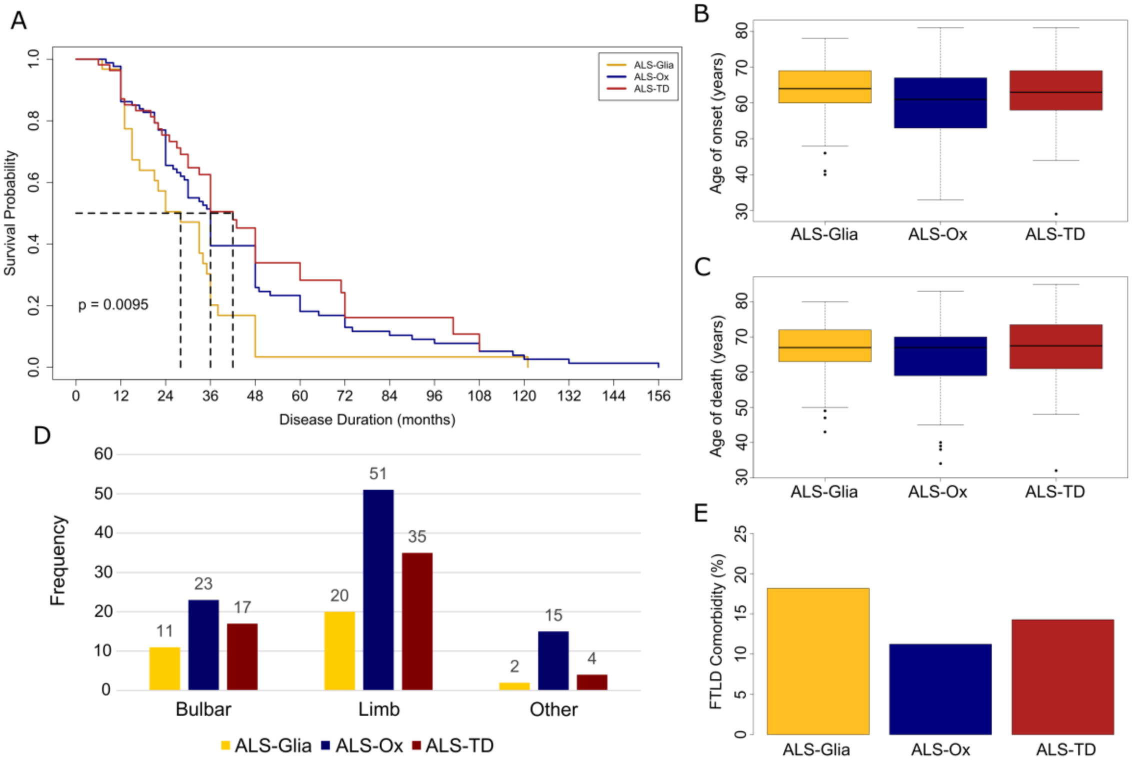


Figure 3.1: Assessment of ALS patient clinical parameters in the context of cortex subtypes. (A) Kaplan-Meier survival for the three identified ALS subtypes, with *n*=150 patients. Patients without an available age of onset or disease duration were excluded from this analysis. The ALS-Glia subtype is significantly associated with a shorter survival duration ($p < 0.01$, log-rank test). The ALS-Ox subtype had a median survival duration of 36 months, while the ALS-TD group had the longest median survival (42 months). (B) Age of disease onset plotted as boxplots for the three ALS subtypes, with *n*=151 patients. No significant differences are observed in age of onset by subtype. The median is indicated by the solid black line, and first and third quartiles are captured by the bounds of the box. Boxplot whiskers are defined as the first and third quartiles –/+ interquartile range times 1.5, respectively, and outliers are denoted as solid black points. Minimum and maximum values are captured by the lowermost and uppermost points, respectively, or whisker bound if no outliers are shown. (C) Age at death plotted as boxplots for the ALS-Glia, ALS-Ox, and ALS-TD subtypes, with *n*=178 patients. Again, no significant differences are observed. (D) ALS subtype site of symptom onset, with the ‘Other’ category comprising axial (4), axial-limb (2), bulbar-limb (4), axial-bulbar (1), generalized (1), and unknown (9) sites of onset. (E) FTLD comorbidity was converted to a percentage and plotted as a bar graph. A Chi-square test of independence was used to assess whether ALS subtype and FTLD comorbidity were associated ($p = 0.59$, one-tailed).

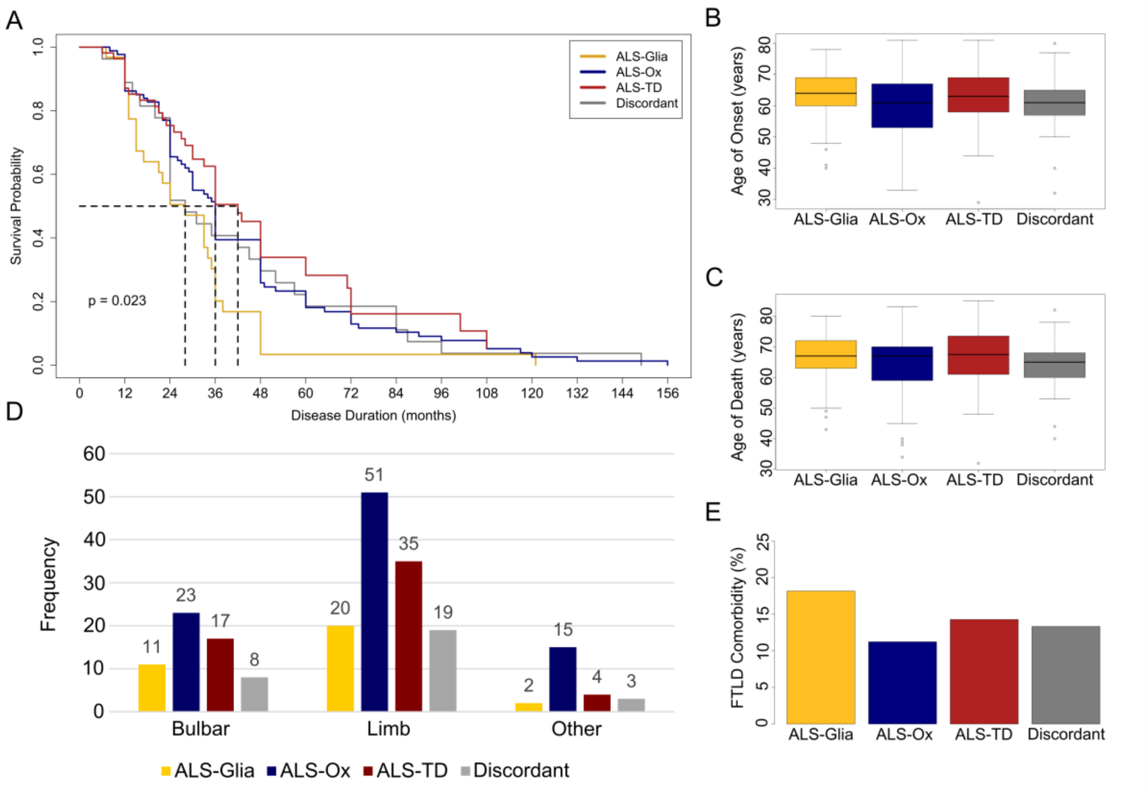


Figure 3.2: Assessment of ALS patient clinical parameters including discordant patients. (A) Kaplan-Meier survival analysis including the three ALS subtypes and ‘discordant’ patients (*n*=177). Pairwise comparisons showed significant differences in survival between the ALS-Glia and ALS-Ox subtypes (*p* = 0.015) and ALS-Glia and ALS-TD subtypes (*p* = 0.0043). *P*, log-rank test. (B) Ages of ALS symptom onset are plotted as boxplots, separated by disease group (*n*=180). The ALS-Glia subtype shows a nonsignificant trend towards the latest symptom onset. The median is indicated by the solid black line, and first and third quartiles are captured by the bounds of the box. Boxplot whiskers are defined as the first and third quartiles –/+ interquartile range times 1.5, respectively, and outliers are denoted as solid black points. Minimum and maximum values are captured by the lowermost and uppermost points, respectively, or whisker bound if no outliers are shown. (C) Age at death is shown for the three ALS subtypes and discordant patients (*n*=208). (D) Site of symptom onset are shown for all ALS patients included in this analysis, and a chi-square test of independence suggests site of symptom onset and subtype are not strongly associated. The ‘other’ category is comprised of axial (4), axial-limb (2), bulbar-limb (4), axial-bulbar (2), generalized (1), and unknown (11) sites of onset. (E) Frontotemporal lobar degeneration comorbidity is shown as a percentage, for all ALS patient groups considered in this analysis. A chi-square test of independence again suggests FTLD comorbidity and ALS subtype are not strongly associated.

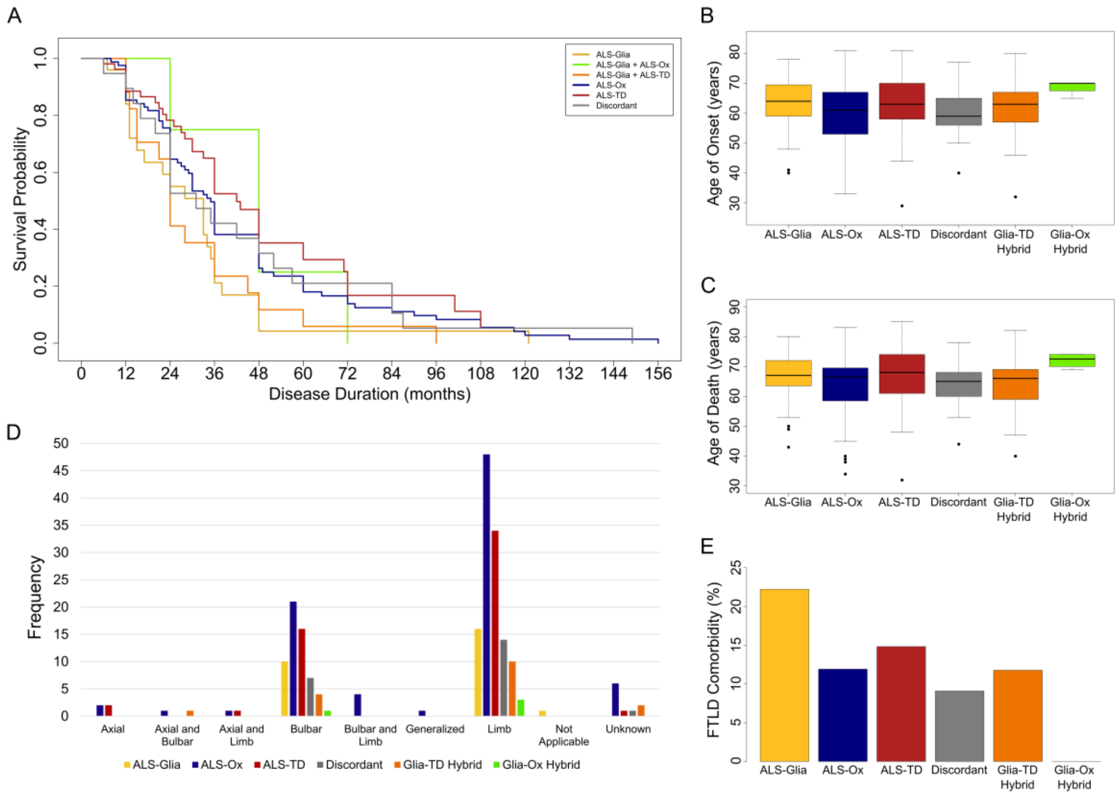


Figure 3.3: Clinical parameters in patients with hybrid subtype samples. Patients were assigned to a hybrid subtype if one or more tissue samples passed the thresholds detailed in *Chapter 2.2.9*. (A) Kaplan-Meier survival analysis including the three ALS subtypes, hybrids, and 'discordant' patients (*n*=177). Interestingly, survival in Glia-TD hybrids mirrors survival in the ALS-Glia subtype, with significant differences observed when compared to the ALS-TD subtype (p = 0.007), and survival differences trending towards significance when compared to the ALS-Ox subtype (p = 0.085). *P*, log-rank test. Findings suggest the elevated inflammatory phenotype seen in ALS-Glia patients is sufficient to drive fast progression in ALS, irrespective of co-expressed phenotypes, although additional work is needed to assess the consistency of hybrid subtype expression in other cohorts. (B) Age of symptom onset, plotted as boxplots, and separated by subtype (*n*=180). No significant differences are observed between the Glia-TD hybrids and other subtypes. The median is indicated by the solid black line, and first and third quartiles are captured by the bounds of the box. Boxplot whiskers are defined as the first and third quartiles –/+ interquartile range times 1.5, respectively, and outliers are denoted as solid black points. Minimum and maximum values are captured by the lowermost and uppermost points, respectively, or whisker bound if no outliers are shown. (C) Age of death, separated by subtype (*n*=208). (D) Site of symptom onset for all disease subtypes. (E) FTLD comorbidity in each disease subtype, presented as a percentage. The small number of Glia-Ox hybrids limits the interpretation of differences observed in survival, age of onset ($p <$ 0.05 for all pairwise comparisons), age of death ($p <$ 0.05 for all pairwise comparisons), and FTLD comorbidity.

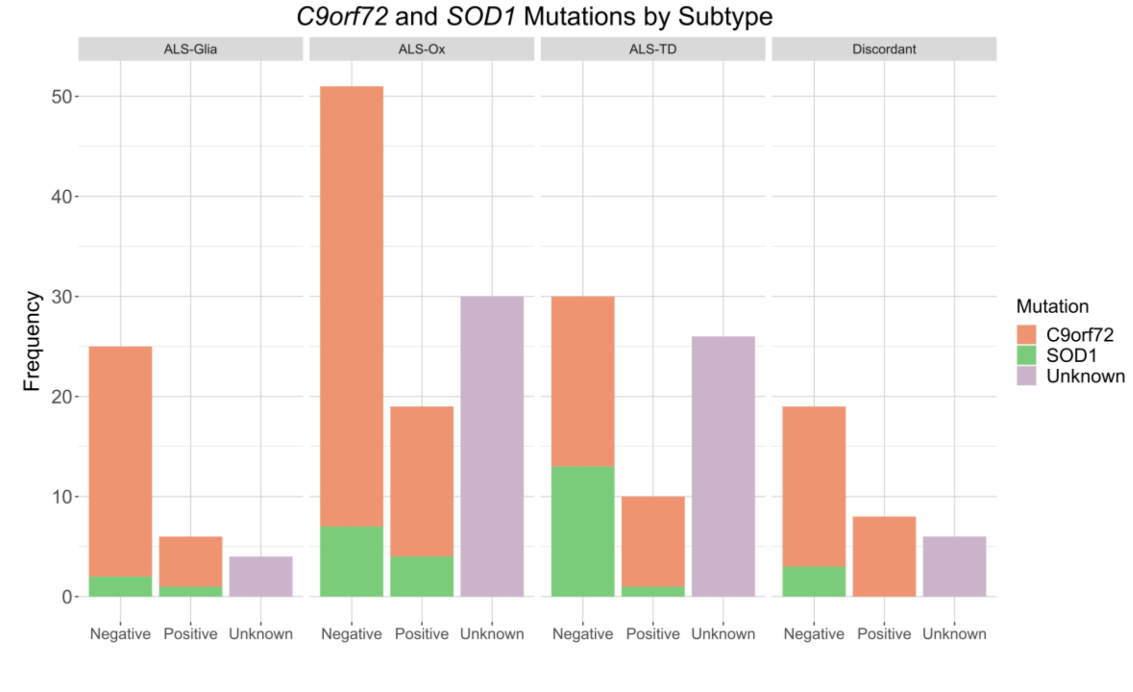


Figure 3.4: Mutation frequency in the cortex cohort. Stacked bar chart showing *C9orf72* and *SOD1* mutation frequency in the ALS cohort. A chi-squared test of independence was performed to assess mutation dependency on subtype. After removal of the "unknown" categorical variable, the null hypothesis (no association between ALS subtype and common genetic drivers) was accepted for both *C9orf72* (p = 0.47, one-tailed) and *SOD1* (p = 0.21, one-tailed). It is important to note that the limited number of observations for *SOD1* may drive inaccurate estimation of the chi-squared test statistic.

Table 3.1. Cortex cohort demographics, with clinical parameters separated by ALS subtype. Disease Duration, Age of Onset, and Age of Death metrics presented as mean ± standard error.

	Cohort Demographics		
A-L: Axial and Limb Onset B-L: Bulbar and Limb Onset A-B: Axial and Bulbar Onset	ALS Spectrum	FTLD	Healthy Control Donors
	(n = 208)	(n = 42)	(n = 58)
Sex			
Female	95 (45.7%)	18 (42.9%)	28 (48.3%)
Male	113 (54.3%)	24 (57.1%)	30 (51.7%)
Tissue Site	n = 451	n = 42	n = 93
Frontal Cortex	193 (42.8%)	42 (100%)	56 (60.2%)
Lateral Motor Cortex	104 (23.1%)	0	18 (19.4%)
Medial Motor Cortex	102 (22.6%)	0	19 (20.4%)
Motor Cortex Unspecified	52 (11.5%)	0	0
ALS Subtype		NA	NA
ALS-Glia	33 (15.9%)	–	–
ALS-TD	56 (26.9%)	–	–
ALS-Ox	89 (42.8%)	–	–
ALS-Discordant	30 (14.4%)	–	–
Disease Duration (months)		Not Available	NA
ALS-Glia	29.1 ± 3.81	–	–
ALS-TD	38.1 ± 3.40	–	–
ALS-Ox	41.8 ± 3.18	–	–
ALS-Discordant	42.4 ± 6.35	–	–
Age of Onset (years)		Not Available	NA
ALS-Glia	63.2 ± 1.83	–	–
ALS-TD	62.7 ± 1.68	–	–
ALS-Ox	60.4 ± 1.16	–	–
ALS-Discordant	60.9 ± 1.87	–	–
Site of Onset		NA	NA
ALS-Glia	Bulbar: 11; Limb: 20; Unknown: 2	–	–
ALS-TD	Bulbar: 17; Limb: 35; Axial: 2; A-L: 1; Unknown: 1	–	–
ALS-Ox	Bulbar: 23; Limb: 51; Axial: 2; A-B: 1; A-L: 1; B-L: 4; Generalized: 1; Unknown: 6	–	–
ALS-Discordant	Bulbar: 8; Limb: 19; Unknown: 2; A-B: 1	–	–
Age of Death (years)		66.5 ± 9.5	64.8 ± 15.7 *
ALS-Glia	66.1 ±1.60	–	–
ALS-TD	66.7 ± 1.33	–	–
ALS-Ox	64.0 ± 1.05	–	–
ALS-Discordant	64.0 ± 1.65	–	–
FTLD Comorbidity	27/208 (13.0%)	42/42 (100%)	NA
ALS-Glia	6/33 (18.2%)	–	–
ALS-TD	8/56 (14.3%)	–	–
ALS-Ox	10/89 (11.2%)	–	–
ALS-Discordant	4/30 (13.3%)	–	–

Table 3.2. Subtype concordance matrix highlights the strong agreement of subtype labels (85%) between this analysis and the foundational work from Tam *et al*.[317] for the 140 samples in common.

Study Concordance Matrix		Eshima *et al*.[99]		
		ALS-TD	ALS-Ox	ALS-Glia
Tam *et al*.[317]	ALS-TE	21	6	0
	ALS-Ox	9	79	1
	ALS-Glia	1	4	19

3.3.2 Subtype Clinical Outcomes in the Spinal Cord Cohort

After stratification of the spinal cord cohort, patient clinical parameters were examined to determine if subtype level differences in survival are maintained. A Kaplan-Meier survival analysis was performed after assigning patient-level subtype using majority agreement between all available regions of the spinal cord or if a single tissue sample was characterized for a given patient (26/206; 12.6%). Similar to the postmortem cortex, a significantly shorter survival duration is observed in the ALS-Glia subtype when compared to ALS-Ox (p = 0.032) and Discordant (p = 0.023) groups but not the ALS-TD subtype (p = 0.27) as seen in Figure 3.5A. As shown in Figure 2.13 and 2.14, the higher proportion of glial cells comprising spinal cord tissue likely drives weaker phenotypic differences between ALS-Glia and ALS-TD subtypes, relative to the cortex, which may partially explain the similar survival curves observed in Figure 3.5A. The mean survival duration for ALS-Glia patients was found to be 31.3 ± 3.97 months (mean ± SE), while ALS-Ox patients were found to have the longest mean survival duration at 45.6 ± 4.87 months, with all values presented in Table 3.3. Interestingly, a significant difference in the age of disease onset between ALS-Glia and discordant patients is seen after FDR[26] correction (FDR p = 0.018), following the trend seen in the postmortem cortex, as seen in Figure 3.5B. Differences in age at death were not significant after FDR correction, shown in Figure 3.5C. No significant relationships were found between disease comorbidity and ALS subtype in the spinal cord using Chi-squared tests of independence, as shown in Figure 3.5D and 3.5E. *C9orf72* and *SOD1* mutation frequency are again plotted in the context of the identified spinal cord subtypes, presented in Figure 3.6. Collectively, findings generally agree with results from the postmortem cortex, but also reflect some of the challenges associated with stratifying patients using spinal cord gene expression.

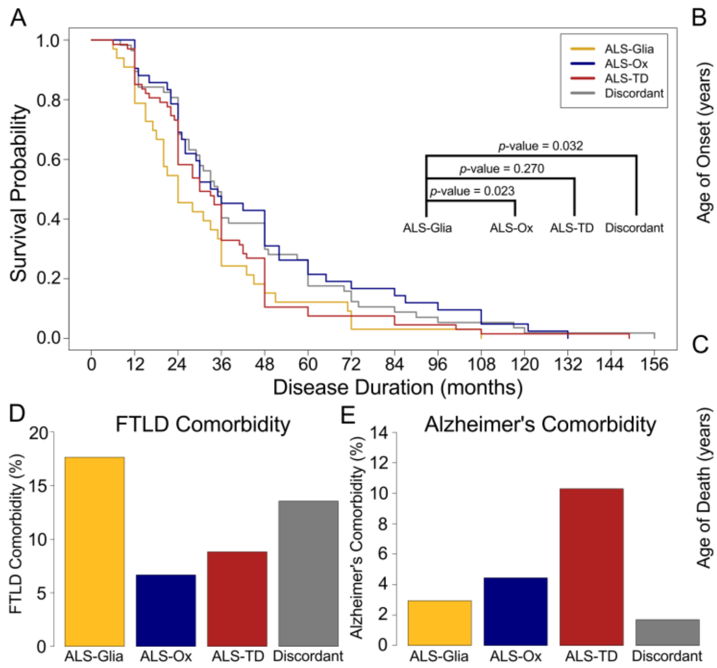


Figure 3.5: (A) Kaplan-Meier survival analysis[167] using patient subtypes (*n*=206) defined by spinal cord transcriptomes. Subtypes were assigned if the majority of available tissue regions were concordant, otherwise the patients were assigned to the 'Discordant' group. No left censoring was applied. The ALS-Glia subtype is observed to have a significantly shorter survival duration when compared to the ALS-Ox and Discordant groups. (B) Age of onset (*n*=206) and (C) age at death (*n*=206) are presented as boxplots for each subtype. T-tests with a false discovery rate correction were applied, and the Glia subtype was seen to have a significantly later age of onset as compared to the Discordant group. Comorbidity for (D) FTLD and (E) Alzheimer's disease are presented as bar plots. Chi-squared tests of independence were performed and neither FTLD (p = 0.38) nor Alzheimer's (p = 0.15) comorbidity was seen to be associated with ALS subtype.

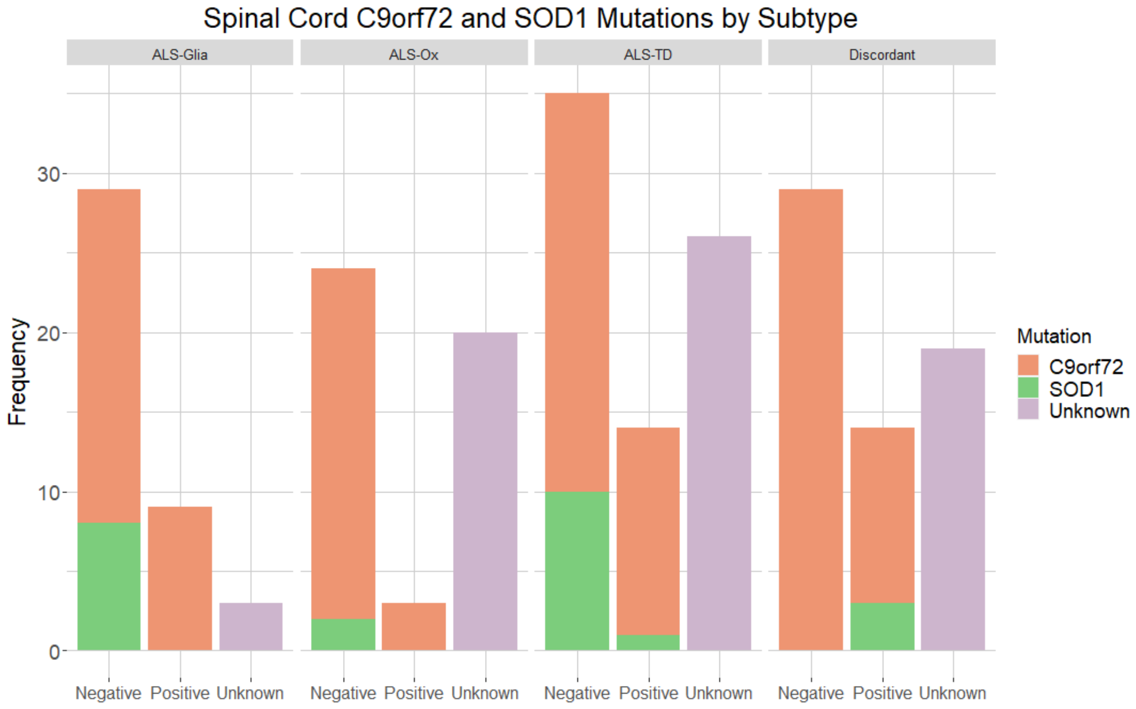


Figure 3.6: Mutation frequency in the spinal cord cohort. Stacked bar chart showing *C9orf72* and *SOD1* mutation frequency in the ALS cohort. A chi-squared test of independence was performed to assess mutation dependency on subtype. After removal of the "unknown" categorical variable, the null hypothesis (no association between ALS subtype and common genetic drivers) was accepted for *C9orf72* (*p* = 0.26, one-tailed) but rejected for *SOD1* (*p* < 0.001, one-tailed). It is important to note that the limited number of observations for *SOD1* may drive inaccurate estimation of the chi-squared test statistic.

Table 3.3. ALS patient and healthy control demographics from the spinal cord cohort. Disease duration, age of onset, and age of death statistics are provided as mean ± standard error. *Two non-neurological control donors had an age of death listed as "90 or Older". An estimate of 90 years was used for all samples listed as such.

Cohort Demographics		
A-L: Axial and Limb Onset B-L: Bulbar and Limb Onset A-B: Axial and Bulbar Onset	ALS Spectrum (n = 206)	Healthy Control Donors (n = 56)
Sex		
Female	97 (47.1%)	31 (55.4%)
Male	109 (52.9%)	25 (44.6%)
Tissue Site	n = 428	n = 91
Cervical Spinal Cord	195 (45.6%)	40 (44.0%)
Thoracic Spinal Cord	55 (12.9%)	9 (9.9%)
Lumbar Spinal Cord	178 (41.6%)	42 (46.2%)
ALS Subtype		NA
ALS-Glia	34 (16.5%)	–
ALS-TD	68 (33.0%)	–
ALS-Ox	45 (21.8%)	–
ALS-Discordant	59 (28.6%)	–
Disease Duration (months)		NA
ALS-Glia	31.3 ± 3.97	–
ALS-TD	36.0 ± 3.01	–
ALS-Ox	45.6 ± 4.87	–
ALS-Discordant	43.2 ± 3.98	–
Age of Onset (years)		NA
ALS-Glia	64.7 ± 1.50	–
ALS-TD	61.6 ± 1.58	–
ALS-Ox	62.6 ± 1.63	–
ALS-Discordant	58.4 ± 1.41	–
Site of Onset		NA
ALS-Glia	A-B: 1; A-L: 1; Bulbar: 8; B-L: 1 Limb: 23	–
ALS-TD	A-L: 1; Bulbar: 21; B-L: 1; Limb: 37; Unknown: 8	–
ALS-Ox	Axial: 2; Bulbar: 12; B-L: 2; Limb: 26; Generalized: 1; Unknown: 2	–
ALS-Discordant	Axial: 1; A-B: 1; Bulbar: 16; B-L: 1 Limb: 39; Unknown: 1	–
Age of Death (years)		63.8 ± 2.16 *
ALS-Glia	67.4 ± 1.48	–
ALS-TD	65.5 ± 1.36	–
ALS-Ox	66.1 ± 1.15	–
ALS-Discordant	62.7 ± 1.29	–
FTLD Comorbidity	23/206 (11.2%)	NA
ALS-Glia	6/34 (17.6%)	–
ALS-TD	6/68 (8.8%)	–
ALS-Ox	3/45 (6.7%)	–
ALS-Discordant	8/59 (13.6%)	–

3.3.3 Subtype Clinical Outcomes in both Cohorts

Extending the consideration of clinical parameters in the cortex and spinal cord cohorts independently, the majority agreement approach is leveraged to assign a single, patient-level subtype from all available samples. Patients without observations from both the cortex and spinal cord were excluded from this analysis, totaling 192 unique individuals. Providing support for the majority agreement approach and overall methodology, subtype proportions derived from all samples mirrors the ratios observed in the postmortem cortex alone, demonstrating that phenotypic agreement broadly occurs throughout the cortex and spinal cord – shown in Figure 3.7. In line with previous univariate survival analyses in the cortex and spinal cord, more defined differences between the subtypes in the context of disease duration and age at onset are observed, as shown in Figure 3.8. Survival differences are greater between the ALS-Glia patients and ALS-Ox and discordant patients, seen in Figure 3.8A. Shown in Figure 3.8B, a significantly later age of onset is seen in ALS-Glia patients, relative to the other two ALS subtypes, after correcting for multiple hypothesis testing. Significant differences in age at death are also seen but are confounded with age of onset. FTLD and Alzheimer's comorbidity was not seen to be associated with ALS subtype, nor was site of symptom onset, as shown in Figure 3.8D, 3.8E, and 3.8F. The combined cohort demographics are provided in Table 3.4. Different results are driven by the re-classification of patient subtype, accounting for added observations from either the postmortem cortex or spinal cord. Worth mentioning, slow progressors in the ALS-Glia subtype are no longer observed, suggesting misclassification, and point to the potential benefit of resampling patient tissue transcriptomes to obtain a robust CNS-level subtype. Collectively, the consideration of patient outcomes in the context of a ‘global’ ALS subtype continues to indicate phenotype-dependent differences in disease duration and age at onset.

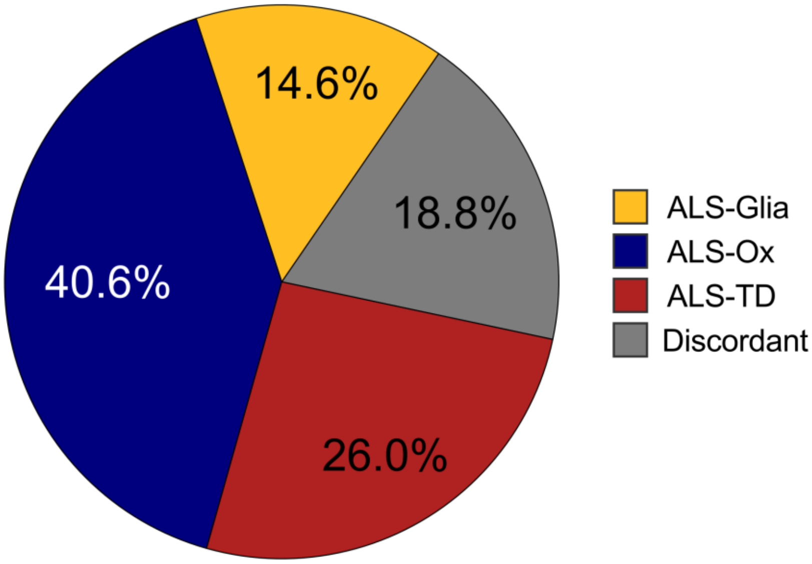


Figure 3.7: Assignment of patient subtype using the majority agreement approach. All patient samples from the postmortem cortex and spinal cord were considered, where available.

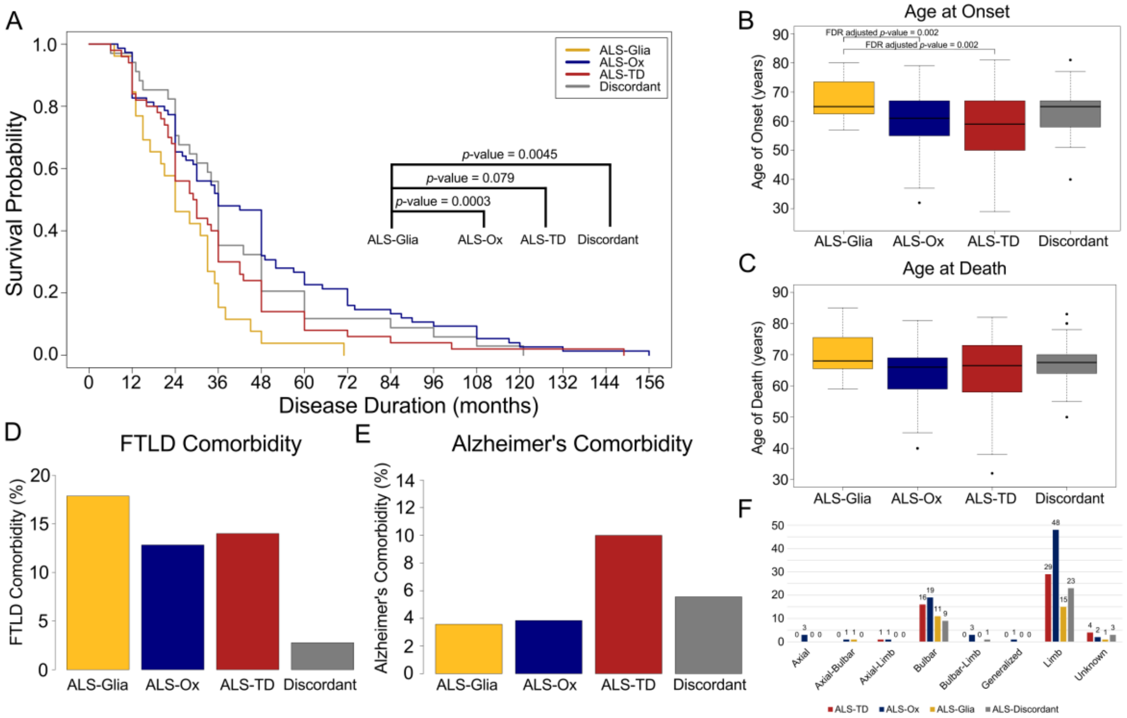


Figure 3.8: Survival and clinical parameters analysis using postmortem cortex and spinal cord observations. (A) Kaplan-Meier survival analysis using patient subtypes (*n*=192) defined by all available cortex and spinal cord transcriptomes. Patients without postmortem observations from both the cortex and spinal cord were excluded. Subtypes were assigned if the majority of available tissue regions were concordant, otherwise the patients were assigned to the 'Discordant' group. The ALS-Glia subtype is observed to have a significantly shorter survival duration when compared to the ALS-Ox and Discordant groups. (B) Age of onset (*n*=192) and (C) age at death (*n*=192) are presented as boxplots for each subtype. T-tests with a false discovery rate correction were applied, and the Glia subtype was seen to have a significantly later age of onset as compared to the other two subtypes (FDR *p* = 0.002). Comorbidity for (D) FTLD and (E) Alzheimer's disease are presented as bar plots. Chi-squared tests of independence were performed and neither FTLD (*p* = 0.256) nor Alzheimer's (*p* = 0.486) comorbidity was seen to be associated with ALS subtype. (F) Site of symptom onset is presented for each subtype, and no significant associations were observed using a Chi-squared test of independence (*p* = 0.564).

Table 3.4. Cohort demographics for all available ALS samples considered in this work. Patients without observations from both the cortex and spinal cord were excluded. Disease duration, age of onset, and age of death statistics are provided as mean ± standard error. *Two non-neurological control donors had an age of death listed as "90 or Older". An estimate of 90 years was used for all samples listed as such. Site of onset abbreviations: A-L: Axial-Limb, B-L: Bulbar-Limb, A-B: Axial-Bulbar

Cohort Demographics	ALS Spectrum (n = 192)
Sex	
Female	89 (46.4%)
Male	103 (53.6%)
Tissue Site	n = 832
Frontal Cortex	180 (21.6%)
Medial Motor Cortex	101 (12.1%)
Lateral Motor Cortex	100 (12.0%)
Unspecified Motor Cortex	48 (5.8%)
Cervical Spinal Cord	181 (21.8%)
Thoracic Spinal Cord	55 (6.6%)
Lumbar Spinal Cord	167 (20.1%)
ALS Subtype	
ALS-Glia	28 (14.6%)
ALS-TD	50 (26.0%)
ALS-Ox	78 (40.6%)
ALS-Discordant	36 (18.8%)
Disease Duration (months)	39.5 ± 2.1
ALS-Glia	26.8 ± 2.8
ALS-TD	35.2 ± 3.6
ALS-Ox	46.0 ± 3.8
ALS-Discordant	41.3 ± 4.6
Age of Onset (years)	61.9 ± 0.8
ALS-Glia	67.5 ± 1.4
ALS-TD	58.3 ± 2.1
ALS-Ox	60.7 ± 1.2
ALS-Discordant	63.6 ± 1.4
Site of Onset	
ALS-Glia	A-B: 1; Bulbar: 11; Limb: 15; Unknown: 1
ALS-TD	A-L: 1; Bulbar: 16; Limb: 29; Unknown: 4
ALS-Ox	Axial: 3; A-B: 1; A-L: 1; Bulbar: 19; B-L: 3; Generalized: 1; Limb: 48; Unknown: 2
ALS-Discordant	Bulbar: 9; B-L: 1 Limb: 23; Unknown: 3
Age of Death (years)	65.6 ± 0.7
ALS-Glia	69.9 ± 1.3
ALS-TD	64.1 ± 1.7
ALS-Ox	64.3 ± 1.0
ALS-Discordant	67.1 ± 1.2
FTLD Comorbidity	23/192 (12.0%)
ALS-Glia	5/28 (17.9%)
ALS-TD	7/50 (14.0%)
ALS-Ox	10/78 (12.8%)
ALS-Discordant	1/36 (2.8%)
Alzheimer’s Comorbidity	11/192 (5.7%)
ALS-Glia	1/28 (3.6%)
ALS-TD	5/50 (10.0%)
ALS-Ox	3/78 (3.8%)
ALS-Discordant	2/36 (5.6%)

3.3.4 Cox Regression for Multivariate Survival

Stemming from the repeat measure of patient tissue transcriptomes postmortem, the assignment of patient level subtype using the majority agreement approach suffers from sample dependence concerns when performing survival analysis. Further, standard survival analyses consider the influence of a single variable on differences in disease duration, making it challenging to address the effects of other important biological covariates like sex and age at disease onset. Tissue-dependent cox regression provides a framework to address both shortcomings of the majority agreement survival analysis shown in Figures 3.1, 3.2, 3.3, 3.5 and 3.8. In addition to informing covariate-dependent survival differences, the exponential of the fitted β coefficient provides the hazard ratio which is useful for informing how much better or worse a given state or level is relative to the specified reference level. While powerful and informative, the undertaking of this analysis proved more challenging than initially anticipated due to the unique structure of the data – given the repetition of patient measures was performed at a single point in time, rather than across time or multiple visits, as is more common.

To better understand how repeat patient measures and the majority agreement approach[99,317] influences previously observed survival differences in the cortex[99], multivariate survival analyses were performed independently in each tissue region using Cox proportional hazard regression[12,80,169,324,325]. As shown in Figure 3.9, survival analysis in each region of the CNS shows weaker differences in disease duration dependent on subtype, revealing that sex and disease group (ALS-FTLD, ALS/Alzheimer's, ALS-SOD1, and ALS-TDP) contribute more to differences in disease duration. Model diagnostics presented in Figure 3.10 show the proportional hazard assumption is met for all covariates. Reference levels are set as males, ALS-TDP disease group, and the ALS-TD subtype. Subtype effect on survival was not statistically significant in the six regions presented, although the ALS-Ox subtype trended towards significance in the cervical (p = 0.095) and thoracic (p = 0.064) tissues with hazard ratios between 0.5 and 0.9. The ALS-Glia covariate is less consistent, with hazards ranging from 0.77 to 1.4, but were more generally equivalent or slightly higher than the ALS-TD reference level. Large confidence intervals seen in the disease group covariate likely reflect low sample number, although it is notable to see the greatest difference in hazard ratio in this covariate and shows disease comorbidity negatively affects risk of death. Interestingly, sex-dependent differences in hazard ratio are seen, with females showing an increased risk for death in the range of 14–80% for the six regions presented. Furthermore, the same model is considered excluding the disease group covariate to account for possible collinearity with subtype. Similar results are seen, with the ALS-Ox hazard ratio between 0.58 and 0.93, ALS-Glia hazard ratio between 0.82 and 1.57, and elevated hazard in females (1.10 – 1.81). Taken together, the ALS-Ox subtype may be weakly associated with a better patient prognosis, however the inclusion sex, age, and disease group covariates while maintaining observational independent demonstrates limitations with the majority agreement approach and that the effects due to these factors contributes more to survival differences than subtype.

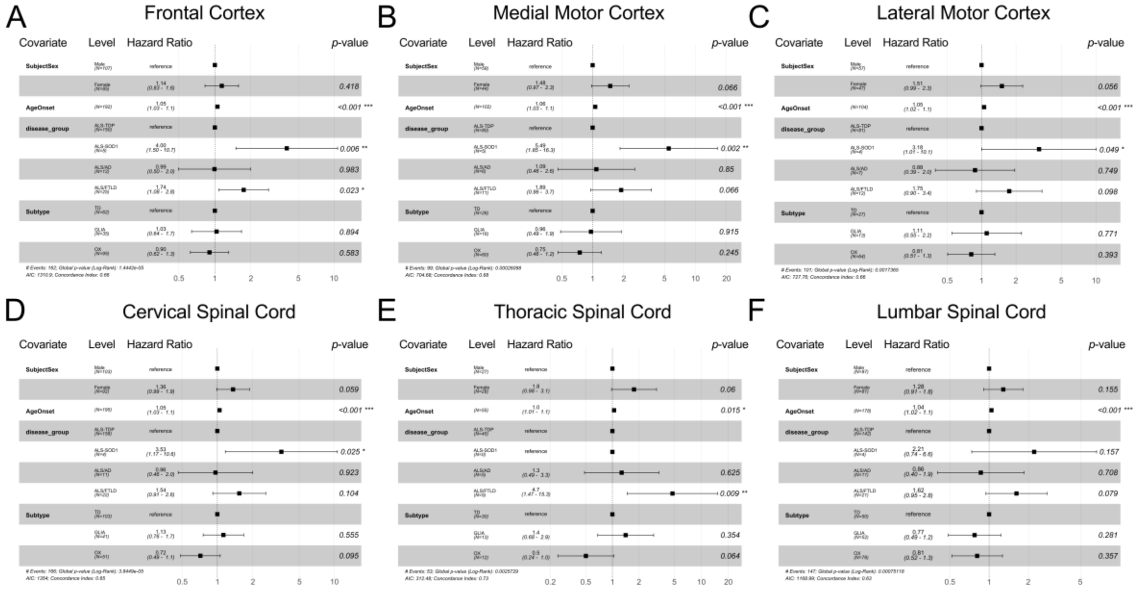


Figure 3.9: Independent and Identically Distributed (IID) Survival Analysis. Tissue region specific survival analyses for the ALS (A) frontal cortex (n = 193), (B) medial motor cortex (n = 102), (C) lateral motor cortex (n = 104), (D) cervical spinal cord (n = 195), (E) thoracic spinal cord (n = 55), and (F) lumbar spinal cord (n = 178). The "unspecified motor cortex" (n = 52) was not considered. The effects due to sex and disease group can be seen to contribute more to survival differences among patients however, across most tissue regions, there is a nonsignificant but consistent trend toward a lower hazard associated with the ALS-Ox subtype. Model terms are presented as hazard ratios with the 95% confidence interval shown. Terms are separated by covariate and subgroup, with reference levels indicated.

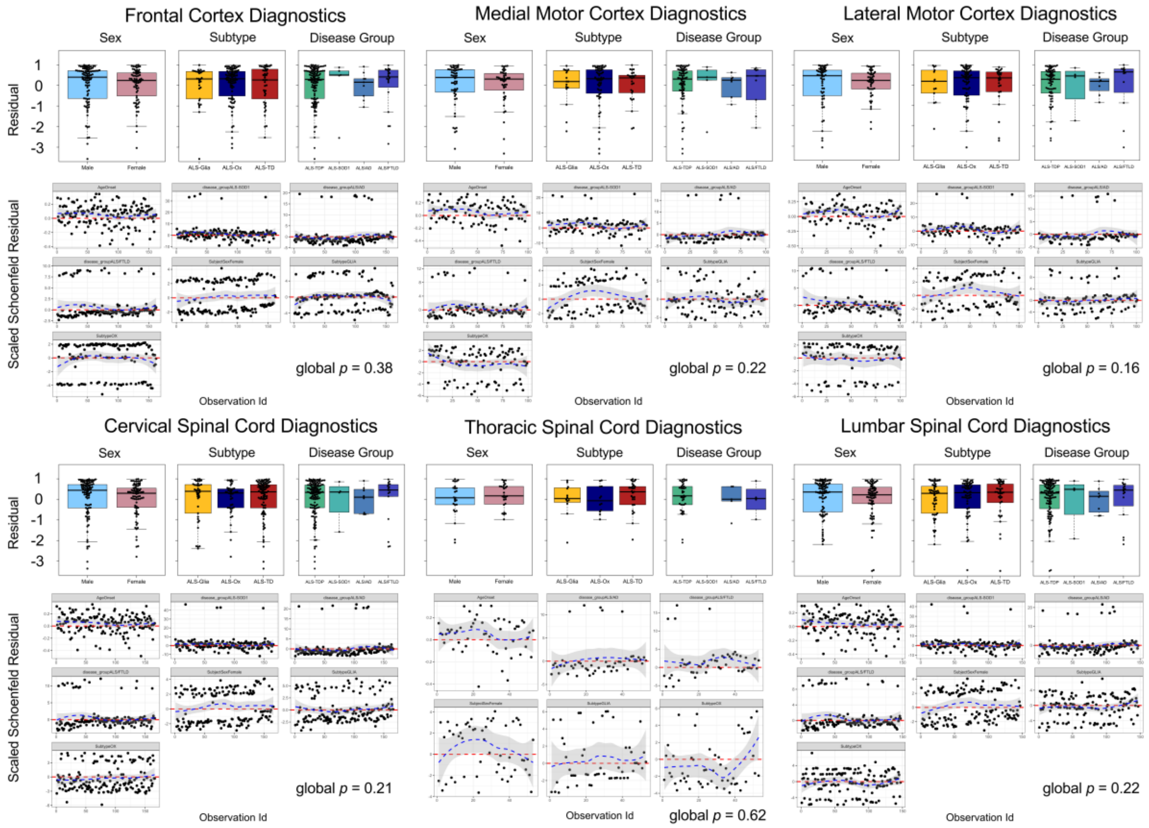


Figure 3.10: Cox proportional hazard model diagnostics. Sample level observations from the postmortem cortex and spinal cord were separated by tissue region and utilized to construct Cox proportional hazard regression models. Sex, disease group, age of onset, and subtype were included as model covariates, yielding a total of 728 observations without missing data in the six Cox models shown. For each model constructed, residuals are plotted by covariate, and generally show weak or null dependency on the variable level. To assess adherence to the proportional hazard model assumption, scaled Schoenfeld residual plots are shown for each covariate level, with score test *p*-values[325] determined using the 'km' time transformation. All covariates are seen to meet the assumption of having proportional hazards over the survival duration, excluding the disease group covariate in the lateral motor cortex model ($p = 0.04$).

3.3.5 Weighted Gene Co-Expression Network Analysis

In an effort to strengthen associations between ALS gene expression subtypes and differences in patient clinical parameters, a weighted gene co-expression (correlation) network analysis[188] was performed in the cortex and spinal cord cohorts independently. In both cases, a soft threshold (power) of the signed network was selected to ensure the assumption of scale free topology is roughly met, as shown in Figure 3.11.

In the cortex cohort, the identification of eigengene clusters based on the correlation of gene expression is shown as a dendrogram and heatmap plot in Figure 3.12. Results indicate the maroon and gold eigengenes are significantly correlated with ALS clinical parameters, as shown in Figure 3.13. Expression of the maroon eigengene is seen to be negatively correlated with age of symptom onset and age at death. Conversely, the gold eigengene is seen to be positively correlated with age of onset and death, yet negatively correlated with disease duration as seen in Figure 3.13A. The observed relationship between the gold eigengene and patient clinical parameters indicates that elevated expression drives a later disease onset but a shorter survival duration. Transcripts comprising subtype-specific eigengenes are presented in Table 3.5.

Cortex eigengenes were enriched for gene ontology, and the gold eigengene was seen to be strongly linked to the immune system ($p < 5\times10^{-16}$, Fisher exact test, one-tailed, Bonferroni-corrected) as shown in Figure 3.13A. Importantly, ALS-Glia specific overexpression was observed for the majority of features included in the gold eigengene, as observed in Figure 3.13B. The maroon eigengene – primarily composed of transposable elements, long non-coding RNA, pseudogenes, and poorly characterized transcripts (Ensembl IDs) – was not significantly linked to any gene ontologies, although a general association with transcription was given. ALS-TD specific expression was observed for many of the features comprising the maroon eigengene as seen in Figure 3.13C. Four representative transcripts from the navy (ALS-Ox) eigengene are shown in Figure 3.14. Subtype-specificity for eigengene expression was further assessed using the β coefficient from dummy regressions considering subtype as the binary predictor and sample-wise eigengene expression as the response, as shown in Figure 3.13A.

Mirroring the approach taken in the cortex cohort, WGCNA was applied to the spinal cord cohort and leveraged the top 5,000 most variable genes form the NovaSeq and HiSeq subgroups, calculated by median absolute deviation. The identification of eigengenes based on correlation of gene expression is shown as a dendrogram and heatmap plot in Figure 3.15. Despite identifying some similarly enriched eigengenes in the cortex and spinal cord cohorts, they were intentionally *not* redefined as gold, maroon, and navy given that the transcripts comprising subtypes-specific eigengenes are not equivalent. Results show the weighted correlation analysis in the spinal cord uncovers an ALS-Ox specific eigengene (“turquoise”) that shares many features with the navy eigengene from the cortex, as presented in Figures 3.16 and 3.17. As shown in Figure 3.16, the turquoise ALS-Ox eigengene was significantly enriched for the neuropeptide signaling gene ontology ($p < 1\times10^{-3}$, Fisher exact test, one-tailed, Bonferroni-corrected) but not significantly associated with disease duration and age of onset (p = 0.08 and 0.09, respectively). Likely reflecting the challenges stratifying ALS-Glia and ALS-TD patients using spinal cord expression, identification of Glia- and TD-specific eigengenes was difficult. Guided by dummy regression coefficients, the green eigengene was assigned to the ALS-Glia subtype while the salmon eigengene demonstrated the strongest specificity for ALS-TD as seen in Figure 3.16. The green ALS-Glia eigengene was not found to be significant enriched for gene ontology and weakly associated with age of onset ($p = 0.008$). A negative correlation with age of onset was observed, opposite to the effects seen in the gold ALS-Glia eigengene from the cortex. The salmon eigengene was found to be significantly enriched for regulation of type I interferon production gene ontology ($p < 0.05$, Fisher exact test, one-tailed, Bonferroni-corrected) and was significantly negatively correlated with disease duration ($p < 1 \times 10^{-11}$). Although the negative association between elevated immunological eigengene expression and disease duration persists in the spinal cord, this phenotype is difficult to distinguish in ALS-Glia and ALS-TD patients and is assigned to the TD subtype instead of Glia (by intuition), again implicating cell compositional challenges and clustering limitations in the spinal cord.

It is relevant to consider a similar WGCNA analysis, performed by Humphrey *et al.*[151], with many of the same spinal cord patient samples considered in this work. Differences in results likely reflect the method used to select input features into WGCNA and the inclusion of transposable element transcripts in this dissertation. Humphrey *et al.* use all differentially expressed genes relative to controls and consider each region of the spinal cord independently during WGCNA analysis[151]. This work uses the top 5,000 most variably expressed transcripts (by median absolute deviation) within the ALS cohort exclusively, and considers each sequencing platform independently in the selection of features, facilitating the discovery of subtype-specific eigengenes.

The clinical parameter analysis in *Chapter 3.3.1*, *3.3.2*, and *3.3.3* is further supported by WGCNA results in the cortex, given the ALS-Glia subtype shows the oldest median age of onset and a significantly shorter disease duration – as captured by the gold eigengene shown in Figure 3.13. Similarly, in the spinal cord, expression of the immune-related salmon eigengene is seen to be significantly negatively correlated with the duration of the disease as shown in Figure 3.16 – although specificity was found for the ALS-TD subtype rather than ALS-Glia, based on the β coefficients from dummy regression. Despite challenges associated with cell composition differences, the cortex and spinal cord WGCNA results converge on similar interpretations, lending additional support for the proposed subtype-associated clinical heterogeneity in ALS.

Table 3.5. Features comprising subtype-specific eigengenes from the postmortem cortex.

ALS-Glia	ALS-Ox	ALS-TD
FCGR3A	NPR3	ENSG00000228741
HLA-DRB1	IMPG1	PSORS1C1
SCIN	OLFM4	HPN-AS1
HLA-DQA1	LINC01361	chr5|137489774|137490129|L2c:L2:LINE|341|+
APOC2	GPR26	KRT8P42
ENSG00000261795	VIP	C10orf62
CD300A	ENSG00000257501	ENSG00000261710
TREM2	LINC01140	chr18|76982063|76982196|MIRb:MIR:SINE|271|-
ALOX5AP	FNDC9	DUOX1
SERPINA1	HTR3B	chr1|205529732|205529939|MIR:MIR:SINE|294|+
CD68	ZNF702P	
SLC7A7	NME5	GAS6-AS1
SIGLEC7	ENSG00000205562	chr11|67418109|67418229|MIR3:MIR:SINE|292|-

AIF1

TLR7

LY86

CX3CR1

HLA-DRB5

CCR5

CD44

SLAMF8

HLA-DRA

SERPINA3

GPR34

MSR1

SNX20

HLA-DOA

CD86

APOBR

FCER1G

TYROBP

ENSG00000273259

SIGLEC9

HAMP

SCIMP

IGSF6

APOC1

RAB42

OLR1

FCGR1A

CD300LF

CHI3L2

CLEC7A

P2RY13

IFI30

ANKRD22

LILRA4

RNASE2

P2RY12

TYMP

RGS1

CP

RNASE6

FCGR1B

ENSG00000286069

PPP1R17

NPY2R

GGH

SMPX

ENSG00000283025

C2orf80

LINC00643

CASQ1

CCDC68

ENSG00000260878

AKAP5

LRRC53

ENSG00000223812

SYTL5

ENSG00000260401

FBXO40

IGF1

COX7A2

SLC17A8

ENTPD3

KITLG

GLRA2

MCHR2

LAMP5

TPH2

FAM19A2

PCP4L1

BEX5

GAD2

GJD2

TMEM126A

VSTM2A-OT1

NPY5R

MAL2

PCP4

ENSG00000270111

ENSG00000260412

TMEM155

CDKN3

DAW1

ENSG00000286220

chr17|74248710|74248902|MIR:MIR:SINE|219|-

LINC00862

SLC4A9

ENSG00000279495

chr17|82658820|82659186|MLT1C:ERVL-MaLR:LTR|238|+

chr11|113245172|113245450|AluJb:Alu:SINE|166|+

ENSG00000262188

FAM95C

chr1|204965807|204966058|MIRb:MIR:SINE|282|+

CD22

chr18|76992147|76992483|MER1B:hAT-Charlie:DNA|141|-

chr10|22568490|22568589|L2b:L2:LINE|263|-

chr9|19550846|19551149|L1MA4A:L1:LINE|136|-

PACRG-AS3

chr13|24298816|24299096|L3:CR1:LINE|314|-

HSD17B3

OR6W1P

SNX18P3

PRR26

TRPV5

ENSG00000215068

DUOXA1

ENSG00000259807

ENSG00000280206

ENSG00000240265

ENSG00000283486

LINC00639

TRPV6

chr1|225866202|225866280|MIR3:MIR:SINE|295|-

ENSG00000279803

ENSG00000248710

TNRC6C-AS1

ENSG00000231840

KEL

ENSG00000279360

ENSG00000241218

TWF1P1

ENSG00000279996

TLR8
CD69
NCF2
CSTA
IL21R
PTPN7
CCL2
LSP1
FPR1
IL1B
MS4A4A
RGS18
FPR3
SELL
FCGR1CP
ODF3B
GPR183
S100A9
NAPSB
GYPC
LILRB3
LY96
DAPP1
LILRA6
FPR2
SOCS3
BCL2A1

ENSG00000278727
WNT2
PRKAG2-AS1
ZRSR1
ENSG00000273301
ENSG00000261037
LY86-AS1
STYK1
HTR2C
SERTM1
CNTN6
LINC01202
DACH2
TYRP1
ABCC12
SYT4
NDUFAB1
SLC26A4-AS1
IL1RAPL2
STAT4
NEK2
TMEM196
GABRA1
ST6GALNAC5
NWD2
CLGN
KLHL14
TDO2
KRT222
HAPLN1
FAM19A1
ENSG00000236841
ENSG00000246363
HS6ST2
USMG5
ENSG00000261728
INHBA-AS1
OLFM3
ENSG00000230852
PCSK1
C3orf80

NANOGP4
ENSG00000285269
chr6|39795360|39795610|MIRb:MIR:SINE|332|+
AGPAT4-IT1
ENSG00000261121
ENSG00000283914
DCLK3
chr10|22599150|22599430|L1MB8:L1:LINE|204|-
KRT8P13
PACRG-AS1
ENSG00000286159
ENSG00000281969
ENSG00000254491
ENSG00000279161
ENSG00000228510
DNM1P47
ENSG00000229492
ENSG00000280571
ENSG00000255176
ENSG00000239828
ENSG00000250072
ENSG00000274184
C9orf139
LINC00940
MAGEC3
ENSG00000261329
ENSG00000248015
ENSG00000268518
ENSG00000225877
GPR83
ADAMTS14
ENSG00000240687
ENSG00000273151
ENSG00000266844
FFAR1
LINC00877
HIST1H1T

LINC01511
RTKN2
MIR133A1HG
VSTM2A
NXPH2
TESPA1
TAC1
FLRT3
GPR22
KCNB2
GPR149
PFDN4
FAM3C2
QPCT
MAS1
HTR5A.AS1
CALB1
C7orf61
ENSG00000272321
PTH2R
ESRP1
HTR2A
C10orf105
MZT1
FAM162B
GLRA3
PLS1
ASAH2B
ENSG00000278962
LINC00507
RGS4
NEUROD1
MAD2L1
NEUROD6
ANO3
GDA
SYNPR
ENSG00000280105
KCNIP4
HPRT1
RAB27B

TMEM200A

OTOGL

ENSG00000261292

SLC27A2

RFK

LINC00460

ENSG00000286084

ELOVL4

FAM3C

ENSG00000282033

DYDC2

SDR16C5

ENSG00000267034

ENSG00000260838

ENSG00000233508

TRBC2

ENSG00000229425

TSPAN13

C1D

ENSG00000279981

SYT10

NUDT4P1

B4GALT6

HSP90AB4P

EIF5A2

LRRC2

ENSG00000279013

ENSG00000265579

LIN28B

PWAR5

SNX10

CERKL

EPHX4

RAB3C

chr12|54950114|54950247|MIR1_Amn:MIR:SINE|372|-

OXGR1

ENSG00000228971

SERPINI1

PTGER3

DYNLT3

ARL6
SOSTDC1
ENSG00000249436
ENSG00000267160
LINC01378
NSRP1P1
KCNV1
WNT16
FGF9
CTXN2
PVALB
LINC01616
ABRACL
ANKRD34C
TMEM14A
GULP1
ENSG00000254187
BMP3
SMIM18
MEPE
RASL11B
PKIB
ENSG00000261542
SPTSSB
CTXN3
ENSG00000248115
ENSG00000279675
OSTN
LANCL3
ENSG00000257522
PTPN3
ENSG00000214265
EEF1E1
SMIM17
SLC17A6
ENSG00000259834

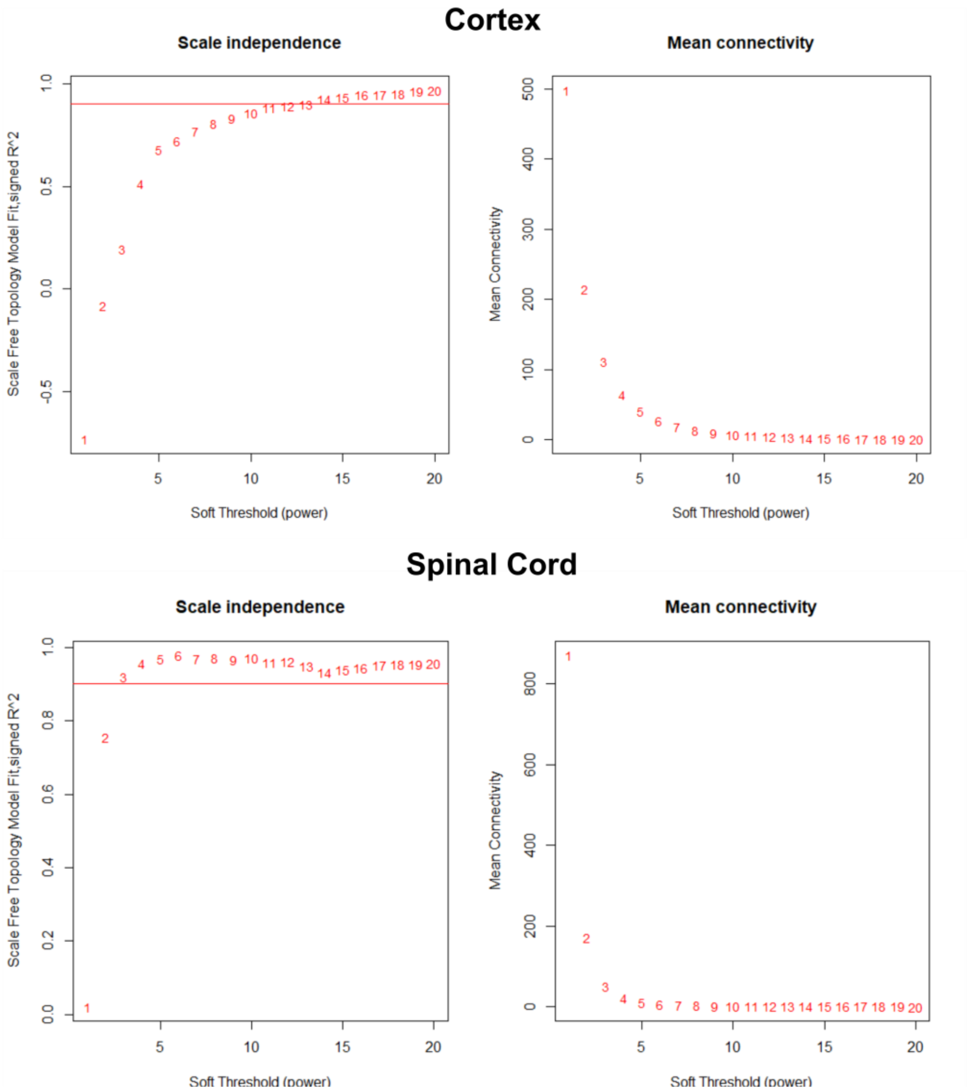


Figure 3.11: WGCNA scale free topology in the postmortem cortex and spinal cord cohorts. A soft threshold of 13 was selected for the cortex cohort, while a soft threshold of 4 was selected for the spinal cohort. Although the assumption of scale free topology (red line) is roughly met for both cases, it is recognized that the power of 13 does not fully ‘saturate’ the R^2 value in the cortex and thus a value of 17 may have been a better choice.

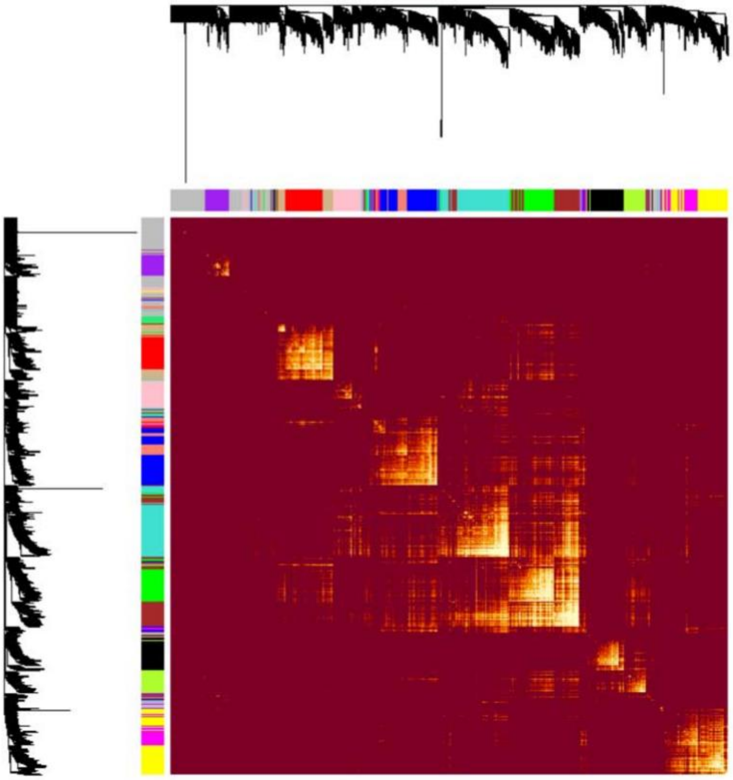


Figure 3.12: Eigengene correlation heatmap and clustering dendrogram identifies transcript sets that are co-expressed in the ALS postmortem cortex. The turquoise, purple, and magenta eigengenes were redefined as the navy, gold, and maroon eigengenes, respectively. All 1,681 transcripts are shown.

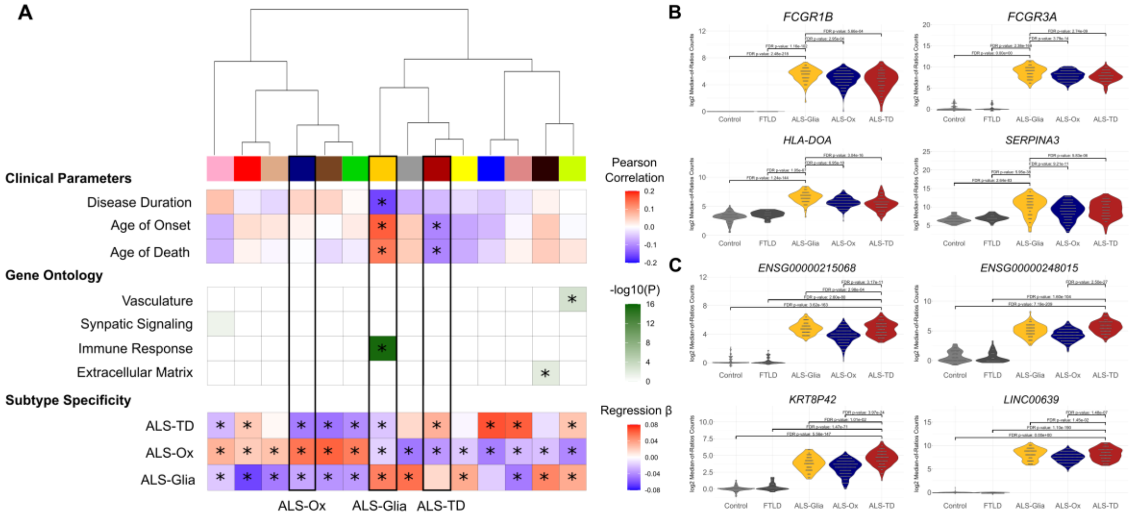


Figure 3.13: WGCNA elucidates subtype-specific disease pathways and eigengenes associated with ALS patient clinical outcomes. (A) Heatmap depicting eigengenes significantly correlated with ALS patient age of disease onset, age of death, and disease duration (univariate regression, two-tailed). Eigengene labels, moving left to right in the dendrogram, are: pink, red, tan, navy (ALS-Ox), brown, green, gold (ALS-Glia), gray, maroon (ALS-TD), yellow, blue, salmon, black, and green-yellow. Eigengenes were enriched for gene ontology and Bonferroni-adjusted *p*-values are shown (Fisher's exact test, one-sided). Subtype-specific expression of eigengenes was determined using dummy regression (two-tailed), with the β coefficient presented as a heatmap. A positive β coefficient denotes subtype upregulation of transcripts comprising the particular eigengene. Bonferroni-adjusted *p*-values less than 0.05 are denoted with *. (B) Univariate plots showing gene expression levels of four representative features (*FCGR1B*, *FCGR3A*, *HLA-DOA*, *SERPINA3*) in the gold eigengene – with evidence for ALS-Glia specificity. *P*, DESeq2[208] differential expression using the negative binomial distribution, two-tailed, false discovery rate (FDR) method for multiple hypothesis test correction. (C) ALS-TD specific expression of four representative features (*ENSG00000215068*, *ENSG00000248015*, *KRT8P42*, *LINC00639*) in the maroon eigengene. *P*, same as B.

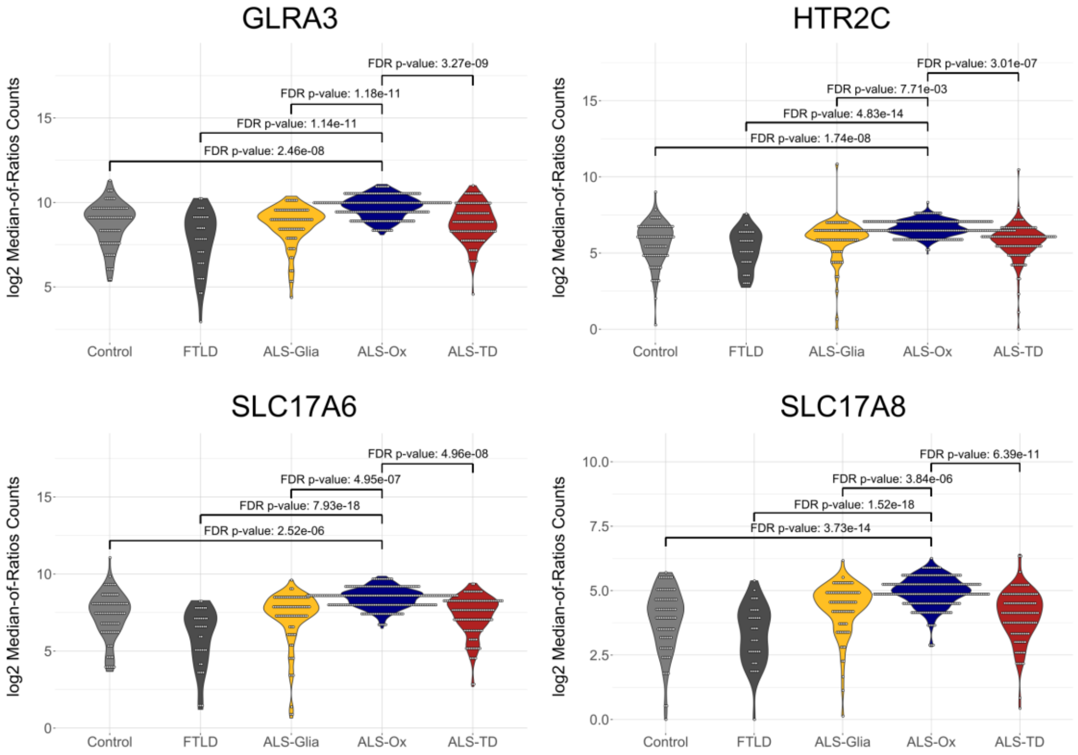


Figure 3.14: Univariate violin plots showing gene expression levels of four representative features (*GLRA3*, *HTR2C*, *SLC17A6*, *SLC17A8*) in the navy eigengene – with evidence for ALS-Ox specificity.

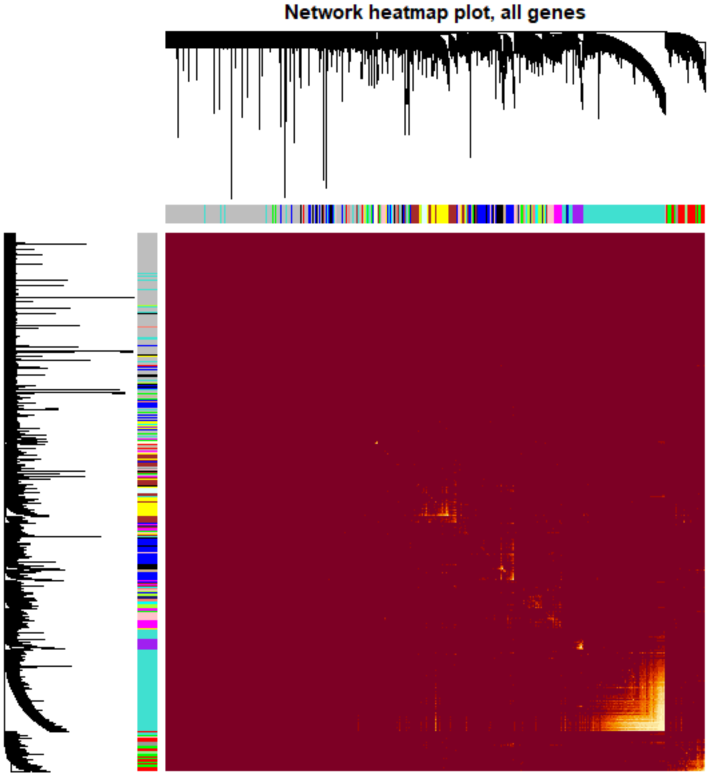


Figure 3.15: Eigengene correlation heatmap and clustering dendrogram identifies transcript sets that are co-expressed in the ALS postmortem spinal cord. All 8,163 transcripts are shown.

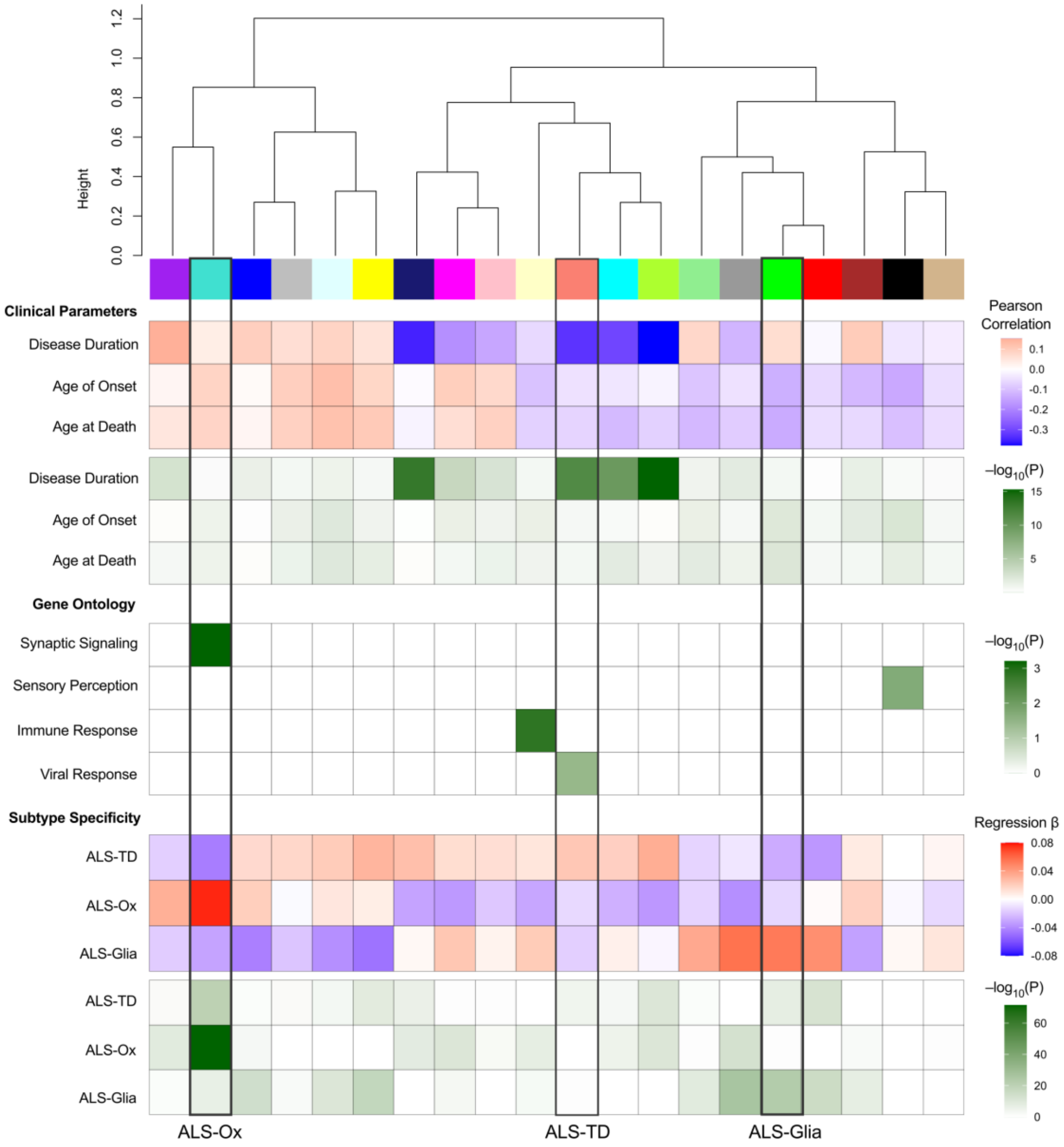


Figure 3.16: Heatmap depicting spinal cord eigengenes significantly correlated with ALS patient age of disease onset, age of death, and disease duration (univariate regression, two-tailed). Eigengene labels, moving left to right in the dendrogram, are: purple, turquoise (ALS-Ox), blue, grey, light cyan, yellow, midnight blue, magenta, pink, light yellow, salmon (ALS-TD), cyan, green-yellow, light green, grey60, green (ALS-Glia), red, brown, black, and tan. Eigengenes were enriched for gene ontology and Bonferroni-adjusted *p*-values are shown (Fisher's exact test, one-sided). Subtype-specific expression of eigengenes was determined using dummy regression (two-tailed), with the β coefficient presented as a heatmap. A positive β coefficient denotes subtype upregulation of transcripts comprising the particular eigengene.

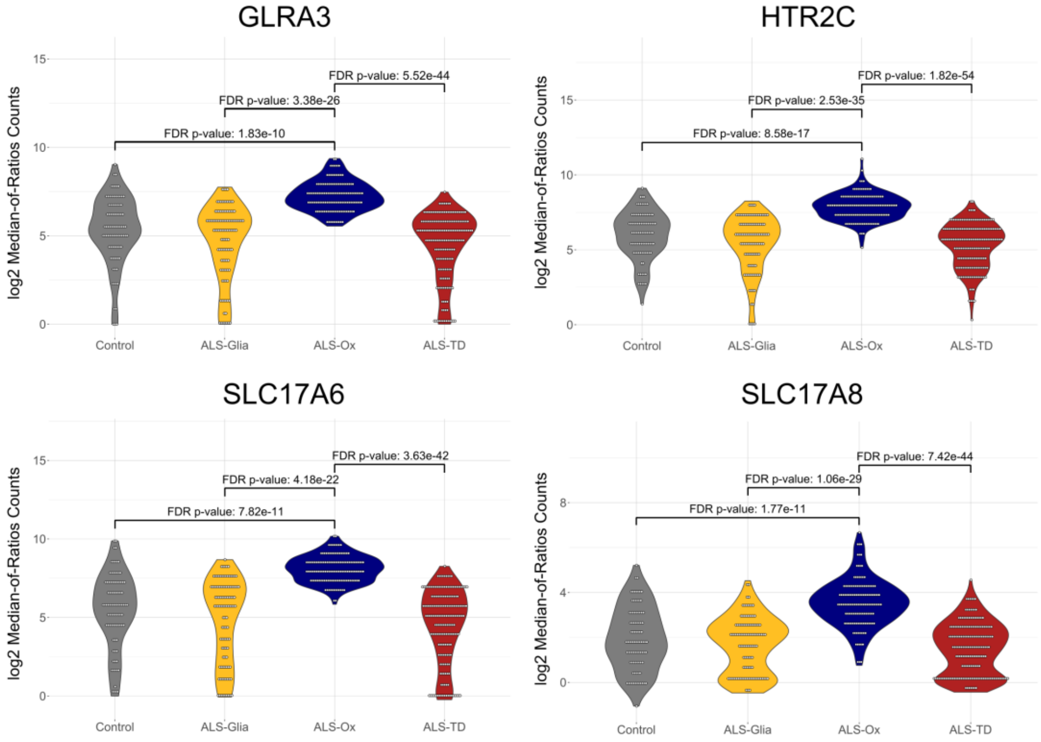


Figure 3.17: Transcripts comprising the turquoise eigengene are elevated in the ALS-Ox subtype, with *GLRA3*, *HTR2C*, *SLC17A6*, and *SLC17A8* shown – also found in the cortex ALS-Ox eigengene.

3.4 Discussion and Conclusion

While previous works arrive at similar findings related to the presentation of molecular phenotypes in ALS patients, associations with clinical parameters had not yet been established[8,237,317]. Collectively, survival analysis and WGCNA weakly link subtype presentation in the postmortem cortex and spinal cord to variability in disease duration and age at onset. Univariate survival analysis leveraging the majority agreement approach arrives at the same conclusion in the cortex and spinal cord cohorts when performed independently and together. The ALS-Glia subtype is seen to have the shortest disease duration while ALS-Ox typically shows the longest. After correcting for other relevant covariates like age, sex, and disease group using the Cox proportional hazard survival framework, the effects due to subtype are diminished. In most cases, sex and disease group were seen to contribute more to differences in patient hazard, yet results continued to indicate a lower risk of death associated with the ALS-Ox subtype, relative to ALS-TD patients. Using a more indirect approach, WGCNA provides the ability to link the expression of gene subsets to clinical measures through correlation. In the cortex cohort, maroon (ALS-TD) and gold (ALS-Glia) eigengenes were significantly correlated with ALS clinical parameters. The Glia eigengene was negatively correlated with disease duration, arriving at a similar conclusion as the univariate survival analysis. In the spinal cord cohort, clustering challenges linked to cell compositional differences made the stratification of Glia and TD subtypes difficult, which was subsequently observed in the WGCNA analysis. The ALS-Ox eigengene is recaptured, however assignment of Glia and TD eigengenes limited broad agreement between cortex and spinal cord results.

Importantly, the identified relationship between elevated inflammatory gene expression and shorter disease duration, in ALS-Glia patients, is well supported by previous works[24,34,356]. Using ALS mouse models expressing mutant *SOD1*, Beers *et al*.[24] and Biollée *et al*.[34] both show that microglia become activated and accelerate disease progression, while Yamanaka *et al*.[356] leveraged Cre-mediated gene excision to demonstrate astrocytes also modulate progression through microglial activation. While these studies do not find associations between glial activation and age at onset, the cortex WGCNA analysis captures a statistically significant positive correlation between inflammatory gene expression and age of onset – potentially a consequence of differences in sample size between these works and our own. Lending additional support to these findings, Humphreys *et al*.[151] consider spinal cord samples from the same cohort[266] and identify activated microglia modules (eigengenes) negatively correlated with disease duration. Similar associations with disease duration are identified in the spinal cord WGCNA using the top 5,000 most variable features, but contrary to results from the cortex, report weak negative correlations with age of onset in the two eigengenes enriched for immunological responses. As previously discussed, different associations with age of onset between the cortex and spinal cord cohorts likely reflects challenges associated with stratifying ALS-Glia and ALS-TD patients using spinal cord expression.

Worth discussing briefly, the majority agreement approach using both the cortex and spinal cord subtypes shows surprising intra-patient agreement, with < 20% of patients classified as discordant. Further, survival and age of onset differences between subtypes become more pronounced and suggest that there may be real benefit in the repeat characterization of patient samples to obtain a more robust assessment of subtype. However, the benefit from repeat sampling may present an unnecessary burden to living patients if successfully translated beyond bulk tissue expression, but may aid ongoing postmortem research.

Findings from this Chapter provide an important link between clinical variability and phenotypic heterogeneity in the cortex and spinal cord of ALS patients. Univariate survival consistently shows a shorter survival duration in patients that present as ALS-Glia in a majority of the tissue regions characterized. After adjusting for other important covariates like age, sex, and disease group (ALS-FTLD, ALS/Alzheimer's, ALS-SOD1, and ALS-TDP), differences in survival dependent on subtype become considerably weaker and non-significant – although the oxidative stress phenotype is generally associated with a lower hazard (0.5–0.9) compared to the other two subtypes. Converging with findings from survival analyses, WGCNA in both the cortex and spinal cord identifies immunological co-expressed gene subsets significantly and negatively correlated with the duration of the disease. In the cortex, the eigengene showed specificity for the ALS-Glia subtype through dummy regression but was assigned to the ALS-TD subtype in the spinal cord. Conflicting findings are attributed to cell compositional differences between the cortex and spinal cord, as shown in Figures 2.13E and 2.14, and suggest the cortex eigengenes may be less influenced by bulk tissue RNA-sequencing bias. Further, cortex immune eigengene shows an association with age of onset, findings that are observed in the ALS-Glia subtype following a direct t-test comparison (with FDR correction) of subtype age of onsets, defined by the majority agreement approach in all available tissues – shown in Figure 3.8B.

This effect persists when considering patients that present as a single subtype in all tissue samples considered, shown in Figure 2.20F, but sample number limits the interpretation to some extent. While associations with survival duration and age of onset are important for establishing a role for phenotypic heterogeneity in ALS, additional work is needed to determine if and how the molecular subtypes contribute to site of symptom onset, rate of progression, disease comorbidity, or genetic risk. Lending some support to this statement, genetic mutations to the *SOD1* protein were found to be dependent on ALS subtype in the spinal cord cohort, although stratification bias towards the ALS-TD subtype and chi-squared tests in the cortex indicate the effect is weak, if at all present, as seen in Figures 2.12, 3.4, 3.6. Broadly, findings from this Chapter provide an important foundation for the design of more effective clinical trials by demonstrating phenotypic heterogeneity captures some of the variability observed in patient survival.

Chapter 4

DIFFERENTIAL EXPRESSION IDENTIFIES MARKER GENES ELEVATED IN THE CORTEX AND SPINAL CORD OF ALS-OX PATIENTS

4.1 Introduction

By demonstrating that the ALS subtypes capture some of the clinical heterogeneity observed in patients, a need arises to enable the stratification of patients using quantifiable markers with direct benefits in clinical trial design. Working to address this need, one group leveraged patient derived clinical measures in a multivariate statistical model, but report limited predictive success[343], indicating accuracy may be improved by including molecular measures like gene or protein expression. More recently, another group applied proteomics to identify a set of 59 proteins in patient CSF that was able to stratify fast progressors (>1 unit decrease in ALSFRS-R per month) from slow progressors (<0.5 unit decrease in ALSFRS-R per month) and validated their predictive performance in an independent patient cohort[337]. Broadly, both groups demonstrate the feasibility of stratifying ALS patients using clinical and molecular measures, while the latter study hints at the advantages of utilizing a systems biology framework to address this need. Building on the results from *Chapters 2* and *3*, analyses are performed to identify transcripts that achieve clinically useful stratification accuracy.

In the fourth chapter of this dissertation, differential expression analysis was applied in an effort to identify subtype-specific transcript expression and establish marker genes. Elevated expression of marker genes *B4GALT6*, *GABRA1*, *GAD2*, *GLRA3*, *HTR2A*, *PCSK1*, and *SLC17A6* was observed in both the postmortem cortex and spinal cord of ALS-Ox patients and shown to be independent of the count normalization approach used. Similarly, overexpression of genes *MYL9, ST6GALNAC2*, and *TAGLN* was observed in the cortex and spinal cord of ALS-Glia patients, although stratification challenges with the immunological subtype in the spinal cord and similar FTLD expression in the cortex restricts the utility to some extent. Leveraging ALS-Ox marker genes, supervised learning shows these features are sufficient to stratify the oxidative stress subtype from all other patient samples, including FTLD and non-neurological controls, with appreciable sensitivity and specificity. Three holdout cohorts were tested, and strong classifier performance was maintained in each, demonstrating technical and methodological batch effects are weaker than the phenotype signal used for stratification. Results from this chapter provide a promising foundation for the translation of ALS-Ox marker genes. While limitations are considered in detail in *Chapter 5.2*, it is important to emphasize that the findings from this dissertation cannot be extended to living patients, and additional confirmatory studies should be carried out to examine marker genes expression premortem and longitudinally throughout the disease course.

4.2 Methods

4.2.1 Differential Expression

Differential expression analysis was performed separately in the cortex and spinal cord cohorts. In the cortex cohort, transcript counts were normalized to the median-of-ratios scale using DESeq2 size factor estimation to better allow comparison between patient samples[208]. Subtype-specific differential expression of transcripts was determined using a multifactor design equation, accounting for sequencing platform, RIN, and site of sample collection covariates. One ALS-TD sample (CGND-HRA-01732) did not have a RIN value available and was subsequently excluded from the analysis due to an incomplete design equation. Pairwise analysis was performed using the constrast() argument, for all combinations. Genes and TEs with an FDR adjusted p-value ≤ 0.05 were considered to be significant. All patient samples from the cortex, including controls ($n = 586$; $n = 267$ female, $n = 319$ male), were considered during normalization. Counts on the median-of-ratios scale were $\log_2$ transformed before plotting. For heatmap presentation, z-scores were calculated using ALS patients to establish gene-wise mean expression and deviation, with expression values on the $\log_2$ median-of-ratios scale. FDR adjusted p-values, derived from DESeq2 differential expression were $-\log_{10}$ transformed prior to plotting. A few additional genes not included in the 1,681 features used for classification, enrichment, and networking, were also considered during the univariate analysis out of disease relevance and include *TARDBP*, *OXR1*, *BECN1*, *BECN2*, *SOD1*, *UBQLN1*, *UBQLN2*, *UCP2*, and *TXN*. Many of these added genes were used during unsupervised clustering as some of the top 10,000 most variable features calculated by median absolute deviation.

In the spinal cord cohort differential transcript expression between ALS subtypes was considered using DESeq2[208] with counts presented on the $\log_2$ scale, following size factor normalization (median-of-ratios). All patient samples were used to estimate size factors for normalization (n = 519 samples). A multifactor design equation was implemented, which included platform, site, RIN, tissue, and subtype covariates. Pairwise comparisons were performed using the contrast() argument, and FDR-adjusted p-values < 0.05 were considered to be significant. For presentation as a heatmap, transcript expression was z-score normalized, and observations that fell outside four standard deviations were adjusted to ±4 for plotting purposes only. FDR-adjusted p-values were $-\log_{10}$ transformed prior to plotting. Differential expression analysis from the cortex was determined previously[99] and reused in the presentation of subtype marker genes in *Chapter 4.3.3*.

4.2.2 Supervised Classification

Supervised classification was applied to two separate scenarios: (1) in the cortex cohort exclusively using 299 subtype-specific transcripts shared between sequencing platform subgroups and (2) in both cohorts using ALS-Ox marker genes. In the first scenario, machine learning classifiers were developed in Python (Version 3.8.8, Python Software Foundation, Wilmington, DE) using the Scikit-learn framework[260] (Version 0.24.1). Supervised classifiers were constructed using training and testing datasets generated from a 70% / 30% split of the ALS NovaSeq cohort (*n* = 255 transcriptomes). 100-fold cross validation was applied to assess performance in the testing cohort. The ALS HiSeq cohort (*n* = 196 transcriptomes) was designated as the holdout dataset to assess performance metrics when classifying new patient samples. Transcript counts on the VST scale were utilized during classifier development. Classifier recall, precision, and F1 scores were calculated for all ALS subtypes after each round of cross validation. Four different models were considered, k-nearest neighbors (KNN), linear support vector classification (Linear SVC), multilayer perceptron (MLP), and random forest (RF). To limit the inclusion of platform-dependent genes, the top 1000 features (by gene score[175]) were further filtered so that only genes and TEs shared between the two sequencing platform cohorts were retained, totaling 299. The k-nearest neighbor classifier was built with k neighbors = 5, distance calculated using the Manhattan metric, weights = distance, and all other parameters as default. The linear SVC classifier was constructed using class weights defined by the proportion of subtypes in the NovaSeq cohort, max iterations = 100,000 and default for all other parameters. The multilayer perceptron neural network was built using three hidden layers (five total), with 100 ‘neurons’ comprising each hidden layer, learning rate = 0.0001, hyperbolic tangent activation function, random state = 1, max iterations = 10,000 and default settings for all remaining parameters. Finally, the random forest was developed using n estimators = 1000, oob score = True, class weights defined by the proportion of subtypes in the NovaSeq cohort, and default for all other parameters. All models were constructed using the ‘one-vs-rest’ multi-class strategy.

In the second scenario, the multi-class problem from the first scenario was simplified to the binary case of predicting “ALS-Ox” or “not ALS-Ox”. Classifiers to stratify ALS-Ox from all other patients (‘NotOx’) – including FTLD (*n*=42 cortex samples[99]) and non-neurological controls (*n*=93 cortex samples[99] and *n*=91 spinal samples) – were developed using an 80% / 20% train/test split and three unique holdout cohorts comprised of (i) all cortical transcriptomes, (ii) all spinal cord transcriptomes, and (iii) all samples analyzed by HiSeq. 100-fold cross-validation was used to estimate F1 scores, with predictions made using the max distance metric and the first component. PLS-DA was performed using the ‘Mixomics’ library in R[279]. Using the same train/test split, cross-validation, and holdout cohorts, additional classifiers were developed in Python (Version 3.9.10, Python Software Foundation, Wilmington, DE) using the scikit-learn framework[260] (Version 1.3.0). Five different models were considered, which included k-nearest neighbors (KNN), linear discriminant analysis (LDA), multilayer perceptron (MLP), random forest (RF), and support vector machine classification (SVM). Default parameters were maintained unless otherwise noted. For the k-nearest neighbor classifier the number of neighbors was set to 8. For the SVM, a linear kernel was used with the regularization parameter, ‘C’ set to 0.025. Finally, the multilayer perceptron classifier was built using three hidden layers, with 100 ‘neurons’ comprising each hidden layer. The learning rate was set to 0.0001, while alpha was set equal to 1E-5.

4.2.3 Quantification of Truncated *STMN2*

Given relevant work from Tam *et al*.[317] and Prudencio *et al*.[266] – which consider *TARDBP*, truncated stathmin-2, and transposable element expression – these features are reexamined in the context of the stratified ALS cohort. Quantification of the normal length and truncated form of *STMN2* were determined previously by Prudencio *et al*. and provided by the NYGC ALS Consortium[266]. In brief, the relative abundance (as a percentage of total expression) of truncated *STMN2* was determined by parsing splice junction tables from STAR and dividing transcripts uniquely mapping to exon 1 through exon 2a by all reads coming from exon 1 on the fragments per kilobase of exon per million mapped fragments (FPKM) scale[266]. Computational methods for the quantification of transposable elements with gene locus resolution are detailed in *Chapter 2.2.3*.

4.3 Results

4.3.1 Differential Expression in the Cortex Cohort

To provide additional insight into subtype-specific gene expression, differential expression analysis was performed, considering the 1,681 features used in classification, enrichment, and WGCNA. Transcript counts were normalized using DESeq2 size factor estimation[208] and $\log_2$ transformed (additional details in Methods section). Shown in Figures 4.1, 4.2, 4.3, and 4.4, the heatmap and violin plots reflect ALS-Glia, ALS-Ox, and

ALS-TD specific transcript expression. Out of the 36 highly differentially expressed transcripts selected to support the characterization of the three ALS phenotypes, 33 were observed to have distinctive expression in a single subtype following differential expression, independent of the RNA-seq platform used for analysis. Feature assignment by gene score generally supports these findings, with 24 of 33 clustering transcripts assigned to the same subtype regardless of the platform used for sequencing, shown in Table 4.1. A few interesting features show rather large differences in normalized expression between controls and ALS subtypes, which may suggest simple thresholding could be used to distinguish the two cohorts, presented in Figure 4.5. To lend additional strength to these findings, a differential expression analysis was performed considering FTLD controls and ALS-FTLD patients exclusively. As seen in Figure 4.6, despite shared pathological mechanisms in these two patient cohorts, ALS-FTLD patients maintain distinct expression of features presented in Figure 4.5. Importantly, some of these genes and transcripts have not been previously associated with ALS neurodegeneration, offering additional insight into disease pathologies and potential targets for diagnostic or therapeutic development.

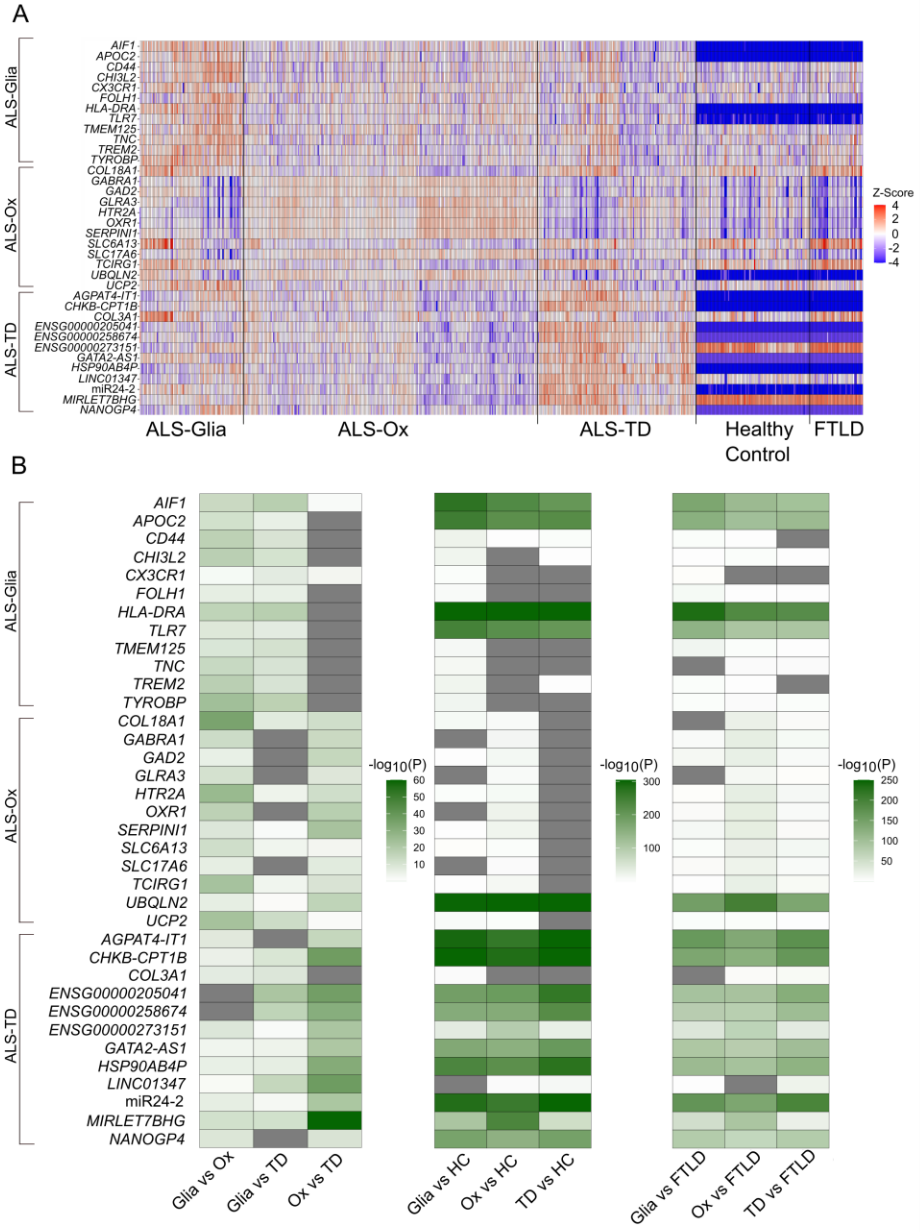


Figure 4.1: Subtype-specific transcript expression. (A) Heatmap showing expression of 36 subtype-specific transcripts for all patient samples considered in this study. Count values are adjusted for RIN, site of sample preparation, and sequencing platform covariates. Expression is z-score normalized, using ALS patient expression to define the mean and standard deviation. Control samples with a z-score < –4 are adjusted to –4 for plotting purposes. **(B)** Presentation of FDR-adjusted *p*-values following pairwise differential expression analysis. *P*-values are $-\log_{10}$ transformed prior to plotting. Gray colored entries indicate an adjusted *p*-value > 0.05. *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

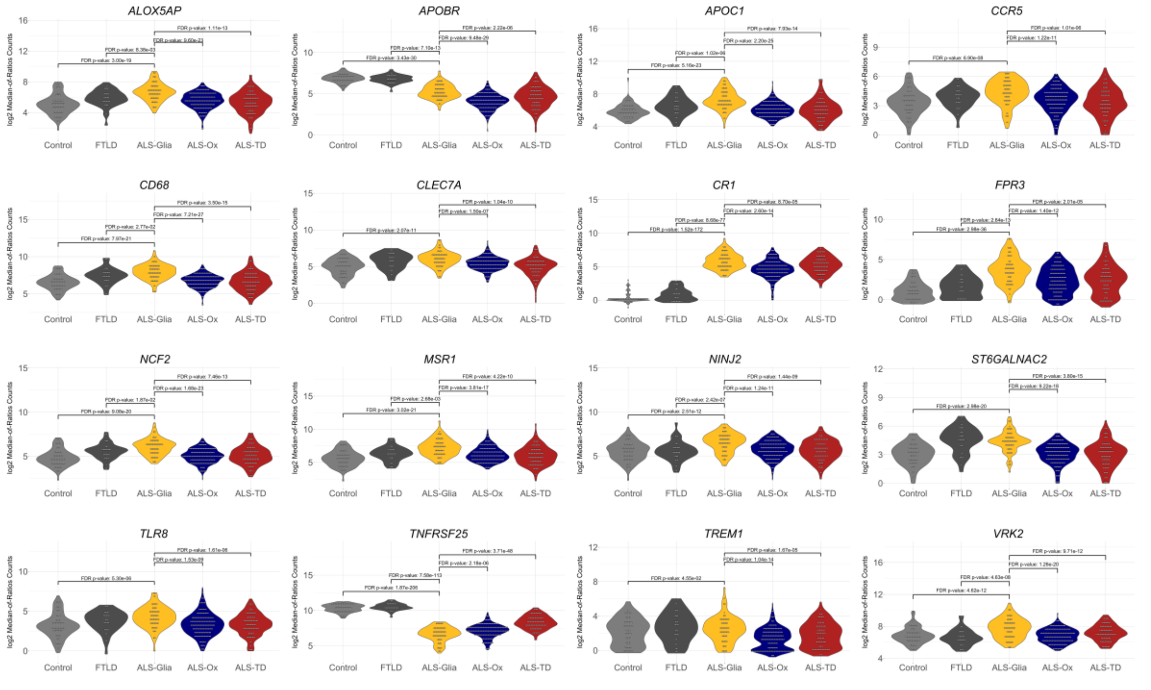


Figure 4.2: Supplemental transcripts for the ALS-Glia subtype. Violin plots show ALS-Glia specific expression for 16 supporting genes: (*top left*) *ALOX5AP*, *APOBR*, *APOC1*, *CCR5*, *CD68*, *CLEC7A*, *CR1*, *FPR3*, *MSR1*, *NCF2*, *NINJ2*, *ST6GALNAC2*, *TLR8*, *TNFRSF25*, *TREM1*, and *VRK2*. Genes are generally associated with glial activation, neuroinflammation, and a pro-apoptotic phenotype. *p*-values have been adjusted for RIN, site of collection, and sequencing platform covariates. *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

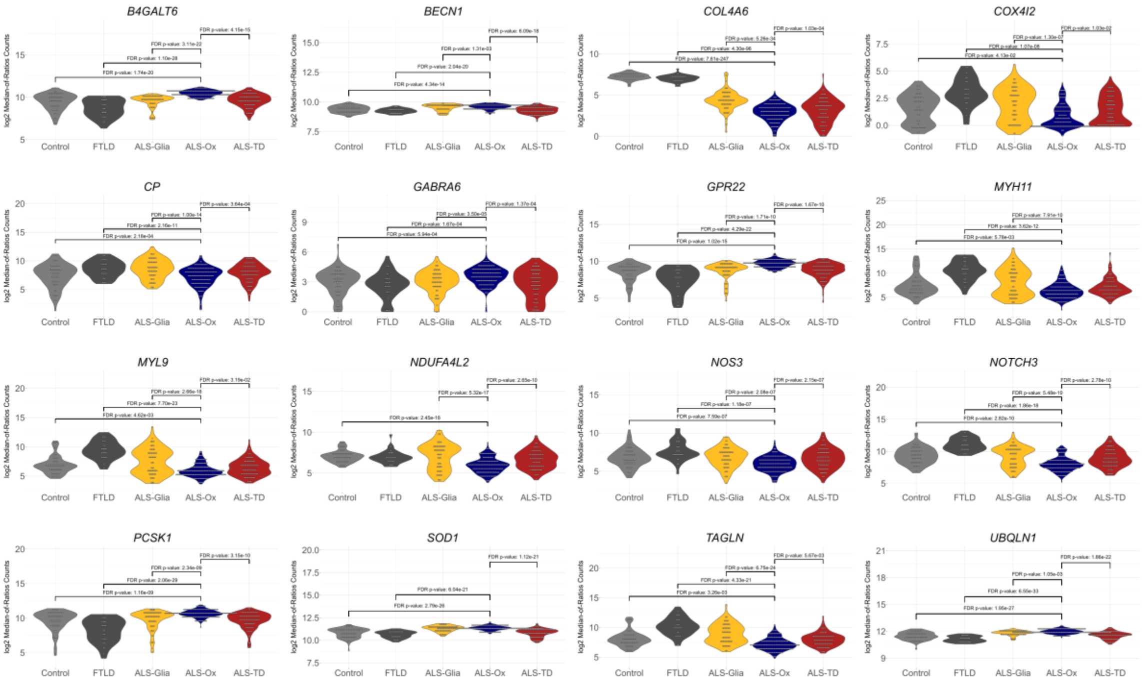


Figure 4.3: Supplemental transcripts for the ALS-Ox subtype. ALS-Ox specific gene expression is shown as violin plots, and include: (*top left*) *B4GALT6*, *BECN1*, *COL4A6*, *COX4I2*, *CP*, *GABRA6*, *GPR22*, *MYH11*, *MYL9*, *NDUFA4L2*, *NOS3*, *NOTCH3*, *PCSK1*, *SOD1*, *TAGLN*, and *UBQLN1*. Supporting genes are generally associated with synaptic signaling, blood-brain barrier integrity, oxidative stress, and proteotoxic stress. *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

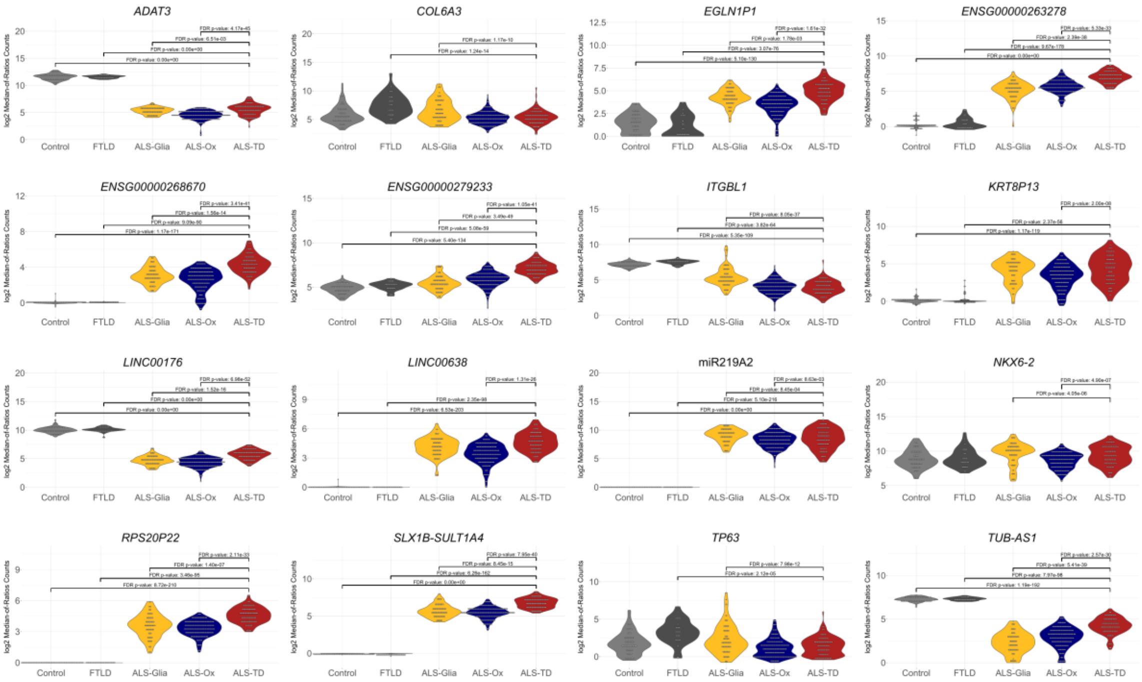


Figure 4.4: Supplemental transcripts for the ALS-TD subtype. ALS-TD specific feature expression is shown as violin plots, and include: (*top left*) *ADAT3*, *COL6A3*, *EGLN1P1*, *ENSG00000263278*, *ENSG00000268670*, *ENSG00000279233*, *ITGBL1*, *KRT8P13*, *LINC00176*, *LINC00638*, *MIR219A2*, *NKX6-2*, *RPS20P22*, *SLX1B-SULT1A4*, *TP63*, and *TUB-AS1*. Supporting genes are generally associated with transcriptional regulation. *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

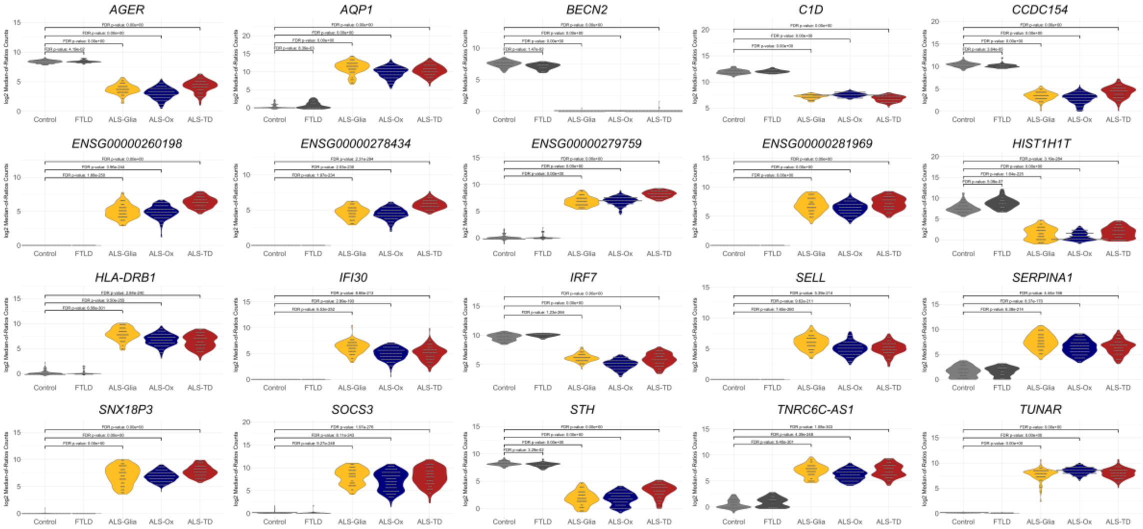


Figure 4.5: Characteristic gene expression distinguishes ALS patients from controls. Genes strongly differentially expressed between ALS patients and controls. Violin plots indicate simple thresholding could be utilized to distinguish ALS patients from controls and some genes further show subtype-specific upregulation or downregulation. Of notable interest, elevated expression of *STH* in the brain is known to serve as a marker for Parkinson’s and other neurodegenerative diseases, including FTLD, and is observed to be strongly downregulated in all ALS patients. These findings offer a potential marker for the stratification of FTLD patients and ALS patients with FTLD comorbidity. *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

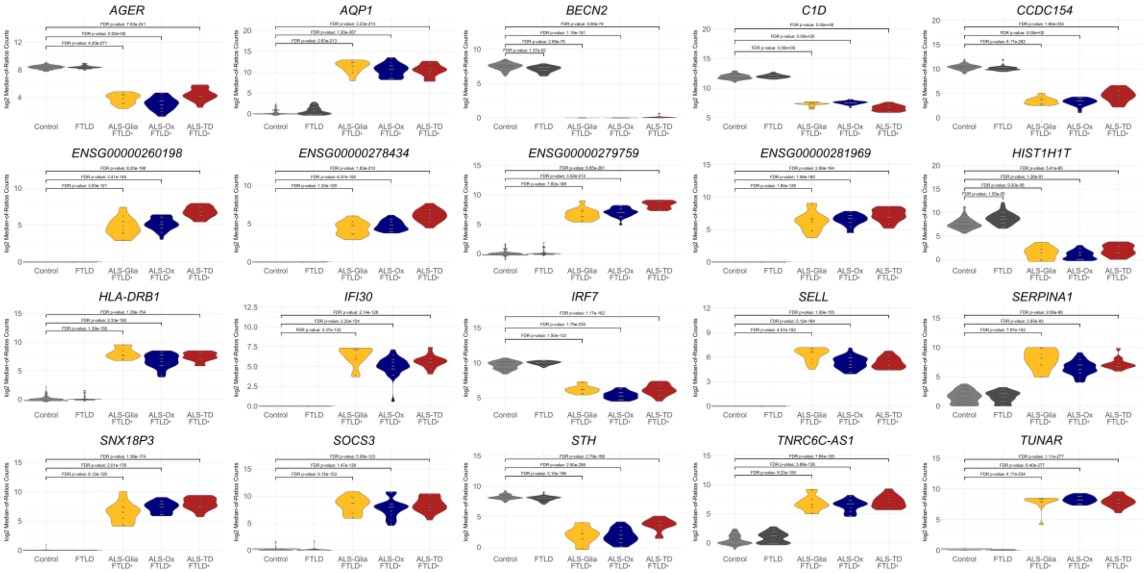


Figure 4.6: Differential expression analysis considering features distinguishing ALS-FTLD from FTLD. Features presented in Figure 4.5 are reconsidered, excluding FTLD⁻ ALS patients. Differential expression between ALS-FTLD and FTLD patients is maintained, further suggesting these features are specific to ALS pathology. *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

Table 4.1: Subtype specificity for transcript expression determined by NMF gene score[175]. Transcript gene scores are calculated in the NovaSeq and HiSeq subsets independently, and generally show strong agreement between the two Illumina sequencing platforms. Genes that were considered in the differential expression analysis exclusively are marked ‘N/A’.

Transcript	NovaSeq Subtype	HiSeq Subtype
AIF1	ALS-Glia	ALS-Ox
APOC2	ALS-Glia	ALS-Glia
CD44	ALS-Glia	ALS-Glia
CHI3L2	ALS-Glia	ALS-Glia
CX3CR1	ALS-Glia	ALS-Ox
FOLH1	ALS-TD	ALS-Glia
HLA-DRA	ALS-Glia	ALS-Glia
TLR7	ALS-Glia	ALS-Glia
TMEM125	ALS-Glia	ALS-Glia
TNC	ALS-Glia	ALS-Glia
TREM2	ALS-Glia	ALS-Glia
TYROBP	ALS-Glia	ALS-Glia
COL18A1	ALS-Glia	ALS-Glia
GABRA1	ALS-Ox	ALS-Ox
GAD2	ALS-Ox	ALS-Ox
GLRA3	ALS-Ox	ALS-Ox
HTR2A	ALS-Ox	ALS-Ox
OXR1	N/A	N/A
SERPINI1	ALS-Ox	ALS-Ox
SLC6A13	ALS-Glia	ALS-TD
SLC17A6	ALS-Ox	ALS-Ox
TCIRG1	ALS-Glia	ALS-Glia
UBQLN2	N/A	N/A
UCP2	N/A	N/A
AGPAT4-IT1	ALS-TD	ALS-Glia
CHKB-CPT1B	ALS-TD	ALS-TD
COL3A1	ALS-Glia	ALS-Ox
ENSG00000205041	ALS-TD	ALS-TD
ENSG00000258674	ALS-TD	ALS-TD
ENSG00000273151	ALS-TD	ALS-Glia
GATA2-AS1	ALS-TD	ALS-TD
HSP90AB4P	ALS-TD	ALS-TD
LINC01347	ALS-TD	ALS-TD
miR24-2	ALS-TD	ALS-Glia
MIRLET7BHG	ALS-TD	ALS-TD
NANOGP4	ALS-TD	ALS-Glia

4.3.2 Differential Expression in the Spinal Cord Cohort

Differential expression was applied to identify subtype-specific transcript expression in the spinal cord cohort. After adjusting for sex, site of collection, RIN, tissue, and sequencing platform covariates, transcript expression that uniquely defines each subtype is recovered regardless of analytical platform, as shown in Figure 4.7A. Differential expression *p*-values, after FDR adjustment and $-\log_{10}$ transformation, are presented in Figure 4.7B as heatmaps, using pairwise comparisons for all group combinations. Expression of transcripts stratifying ALS-Glia and ALS-TD subtypes in the spinal cord is weaker, evident in the heatmap and the differential expression *p*-values relative to the cortex[99], and likely reflects differences in cell type composition in the spinal cord as seen in Figures 2.13 and 2.14.

Spinal cord differential expression analysis is extended by considering other relevant transcripts, including those found to stratify this cohort when considering postmortem cortex transcriptomes[99]. Neuroinflammatory genes *AIF1*, *CD68, HLA-DRA*, *TREM2*, and *TYROBP* were among the most elevated transcripts in the cortex of ALS-Glia patients[99] but not the spinal cord, likely reflecting regional differences in cell type populations, shown in Figure 4.8. Further these transcripts were included in the 1282 glial marker genes removed prior to clustering – which may partially explain the similar expression of these transcripts in ALS-Glia and ALS-TD subtypes. Oxidative and proteotoxic stress genes *BECN1*, *OXR1*, *SERPINI1*, *SOD1*, and *UBQLN2* generally show weaker differences in spinal cord expression when compared to the other two subtypes and again implicate cell compositional differences. Interestingly, transcriptional regulators miR24-2 and *NKX6-2* show specificity for the postmortem cortex, and *NKX6-2* expression is most elevated in the spinal cord of ALS-TD patients, relative to the other two subtypes.

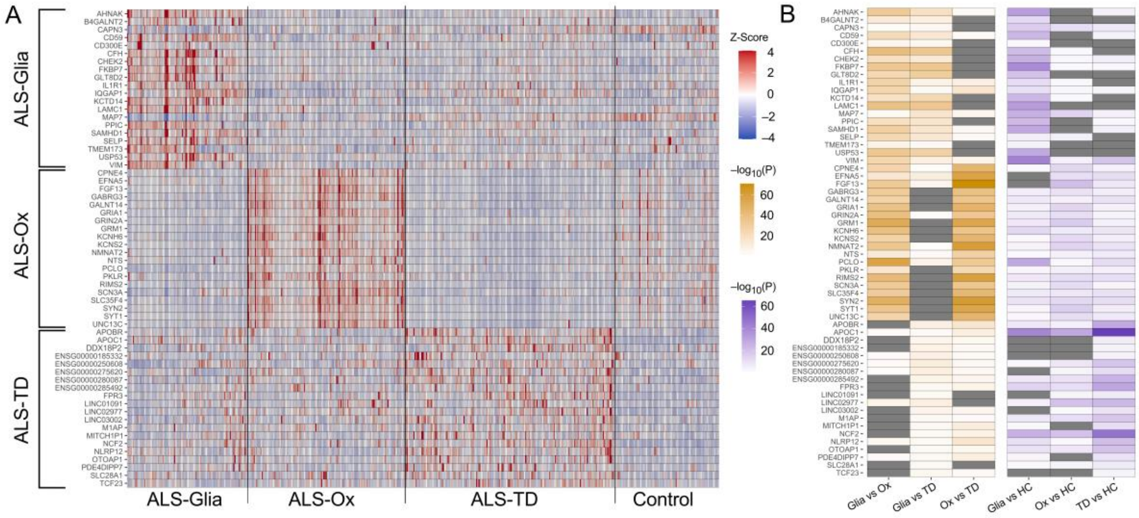


Figure 4.7: Differential expression identifies transcripts in the spinal cord that stratify subtypes and ALS-Ox markers shared between the cortex and spinal cord. (A) Heatmap showing z-score normalized expression following transformation to the median-of-ratios scale for each subtype. For plotting, z-scores < –4 or > 4 are adjusted to –4 and 4, respectively. All presented genes have mean raw counts > 10 and are expressed uniquely in a single ALS subtype. A total of 519 spinal cord samples are shown along the columns, grouped by subtype. (B) Heatmaps showing $–\log_{10}$ transformed differential expression FDR-adjusted *p*-values using pairwise comparisons. Gray cells indicated an adjusted *p*-value > 0.05.

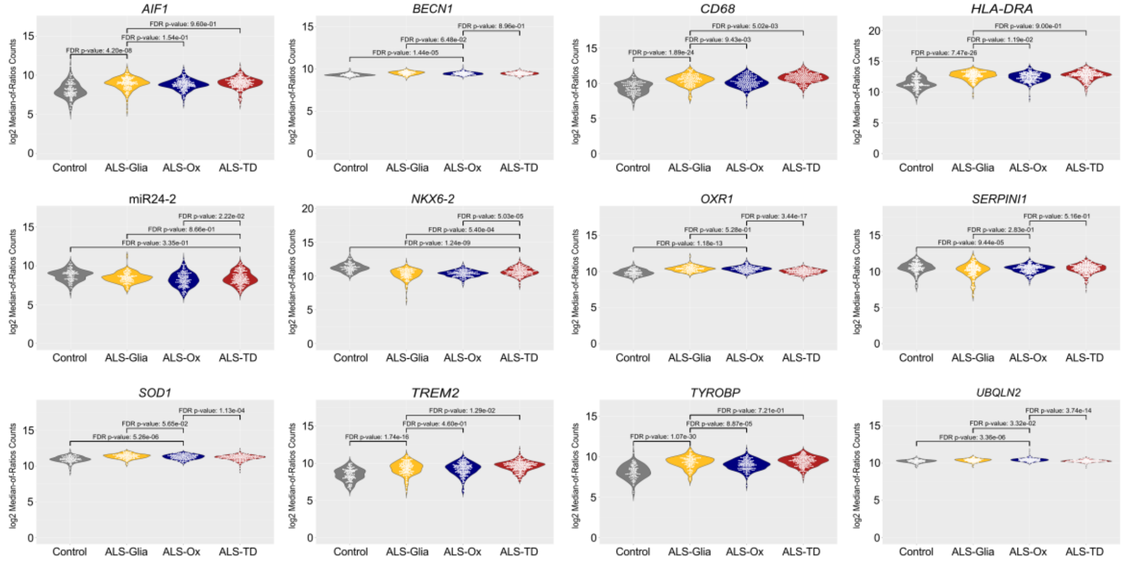


Figure 4.8: The neuroinflammatory subtype (ALS-Glia) is obscured in the ALS spinal cord. Transcripts found to stratify this ALS cohort using the postmortem cortex[99] are reconsidered in the spinal cord. ALS-Glia cortex transcripts *AIF1*, *CD68*, *HLA-DRA*, *TREM2*, and *TYROBP* show weak or non-significant differences in expression compared to the other two subtypes in the spinal cord. Genes associated with proteotoxic and oxidative stress are elevated in the cortex of ALS-Ox patients but not in the spinal cord, seen in the expression of *BECN1*, *OXR1*, *SERPINI1*, *SOD1*, and *UBQLN2* yet tissue composition at the cellular level may partially explain these differences. *NKX6-2* but not miR24-2, both associated with the regulation of transcription, showed weak but consistent upregulation in the cortex and spinal cord of ALS-TD patients.

4.3.3 ALS-Ox Marker Genes

Most notably, a total of ten transcripts are identified with consistently elevated expression, irrespective of tissue region in the postmortem ALS central nervous system, relative to the other subtypes and non-neurological controls in this cohort. Seven of these transcripts were specific for the ALS-Ox subtype, presented in Figure 4.9, while the remaining three were specific for ALS-Glia shown in Figure 4.10. A total of 1,104 unique tissue transcriptomes were considered, from 5 distinct regions of the central nervous system, corresponding to 222 ALS patients, 88 non-neurological controls, and 42 frontotemporal dementia (FTLD) patients. ALS-Ox marker genes *GABRA1*, *GAD2*, *GLRA3*, *HTR2A*, *PCSK1*, and *SLC17A6* collectively implicate changes to synaptic signaling, with elevation of inhibitory receptors and enzymes involved in the biosynthesis of inhibitory neurotransmitters. Upregulation of *ST6GALNAC2* in ALS-Glia samples and *B4GALT6* in ALS-Ox suggests protein glycosylation may play a surprisingly central role in the presentation of ALS subtype. As may be expected, expression of the subtype marker genes was generally different in the cortex and spinal cord regions. Notably, ALS-Ox marker genes were found to better stratify this patient cohort when considering spinal cord expression, evident in the FDR-adjusted *p*-values, which offers promise for clinical translation. Difficulty stratifying ALS-Glia and ALS-TD patients in the spinal cord may limit the practicality of Glia marker transcripts. Further ALS-Glia marker genes were non-significantly or weakly upregulated in the cortex when compared to FTLD patients, suggesting disease mimics may not be easily distinguishable from the Glia subtype.

To demonstrate differential gene expression is not dependent on median-of-ratios transcript normalization, marker gene expression is considered following FPKM normalization in a refined set of ALS patients with observations available from both the postmortem cortex and spinal cord (*n* = 192 ALS patients, 88 non-neurological controls) – shown in Figure 4.11. ALS patient samples were binned into one of three categories in an effort to capture the spectrum of phenotypes typically observed in most patients, which include: "Concordant ALS-Ox" (100% of tissue samples are ALS-Ox), "At least 50% ALS-Ox", and "Generally not ALS-Ox" (<50% of tissue samples are ALS-Ox). All ALS-Ox marker genes show a decreasing trend in the median expression as intra-patient concordance for the ALS-Ox subtype decreases. Further, patients that generally do not present as ALS-Ox maintain marker gene expression at a level similar to non-neurological controls. Recognizing that the transcripts per million (TPM) count unit is better suited for cross-sample comparison, future works should examine ALS-Ox marker gene expression using TPM counts derived from alignment procedures. Collectively, these findings show the ALS-Ox marker genes defined in this work provide a foundation to stratify ALS patients and account for the moderate intra-patient concordance observed between the cortex and spinal cord CNS regions.

Finally, to provide additional insight into the subset of patients that demonstrated perfect subtype concordance throughout the postmortem cortex and spinal cord (Figure 2.20), marker gene expression, genetic risk, and truncated *STMN2* abundance are considered in this subset ($n_{\text{ALS-Glia}}$ = 5, $n_{\text{ALS-Ox}}$ = 12, $n_{\text{ALS-TD}}$ = 19) and compared to all other patients (n=177). A total of six patients were 100% concordant in the spinal cord (2 or more observations) but did not have observations from the cortex available and were subsequently excluded from the analysis. These excluded de-identified IDs include: 15-215-48, 97-126-29, PF-UCL-12, PF-UCL-28, 97-125-35, PF-UCL-62. Similarly, in the cortex, three patients were 100% concordant (2 or more observations) but did not have observations from the spinal cord available and were also excluded from this analysis. De-identified IDs subject to exclusion include: 13-191-47, NEUVZ387WGH, NEUZV622ZHF. As seen in Figure 4.12, median ALS-Ox marker gene expression was elevated in the perfectly concordant patient subset, compared to all other groups including FTLD mimics in the cortex. For most genes in either region, expression in the "Not 100% Concordant" category spans the entire count range indicating that thresholding is insufficient to accurately stratify patients. FTLD patients show similar expression of ALS-Glia marker genes in the cortex, but tissue region dependent differences are clearly observed for *MYL9*, *TAGLN*, and *ST6GALNAC2* suggesting additional work could demonstrate specificity for the ALS-Glia spinal cord. As presented in Figure 4.13, elevated expression of truncated *STMN2* was less distinct in the ALS-Ox spinal cord and cortex relative to the other two subtypes and patients without full concordance. Interestingly, genetic mutation showed a significant association with perfectly concordant patient subtype (p = 0.0124). Excluding 13 concordant ALS-TD patients with an unknown hexanucleotide repeat expansion length, five of the six remaining individuals were positive for the *C9orf72* mutation as seen in Figure 4.13. Collectively, marker gene expression in the perfectly concordant patient subset lends additional strength to their association with ALS molecular subtype and implicates a genetic component in ALS-TD concordance.

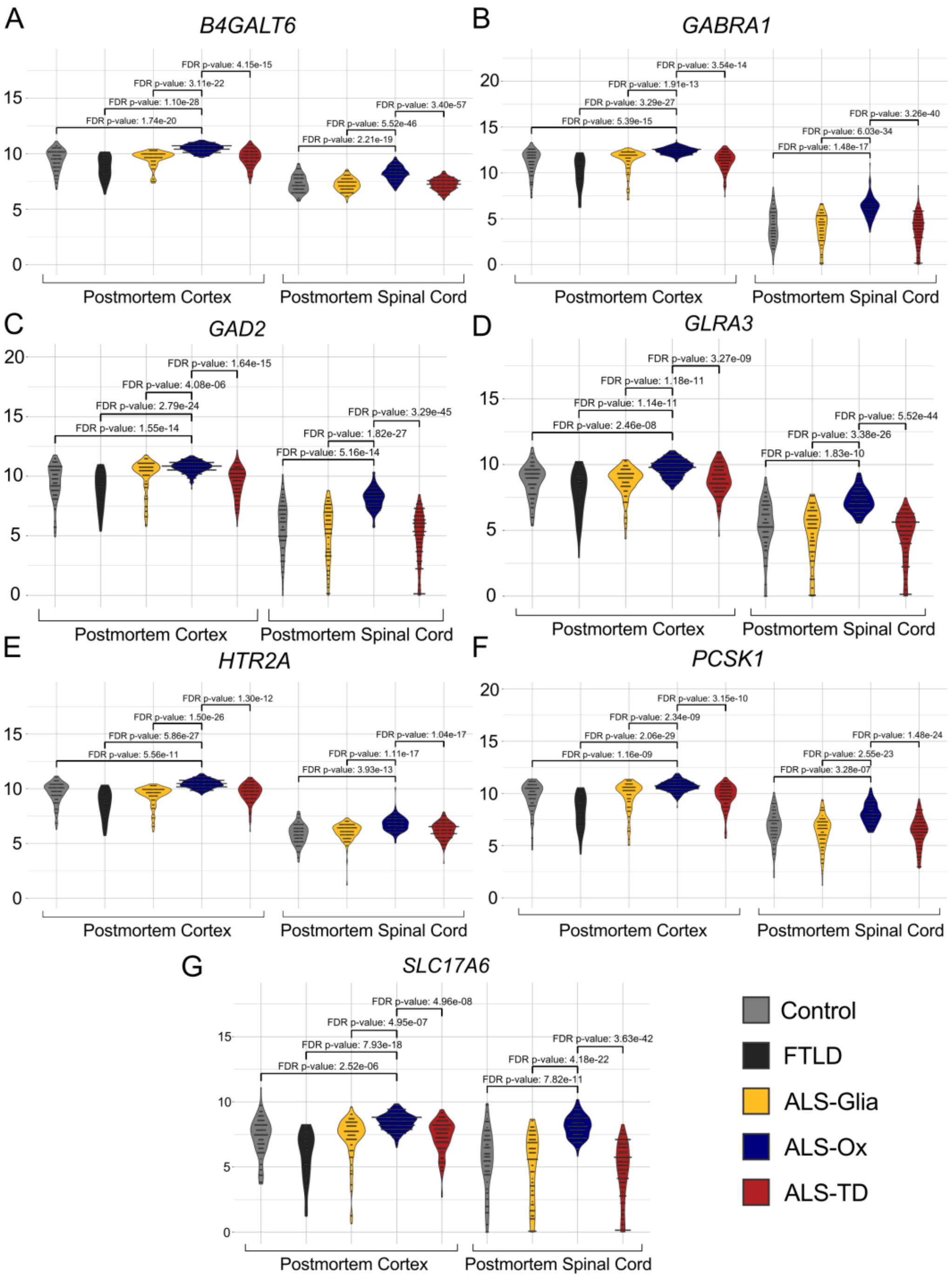


Figure 4.9: ALS-Ox marker genes in the postmortem cortex and spinal cord. Marker genes show coherent elevated expression throughout the central nervous system. Expression is separated both by subtype and CNS region for (A) *B4GALT6*, (B) *GABRA1*, (C) *GAD2*, (D) *GLRA3*, (E) *HTR2A*, (F) *PCSK1*, and (G) *SLC17A6*. All counts are presented on the $\log_2$ transformed median-of-ratio scale. All differential expression *p*-values are FDR adjusted.

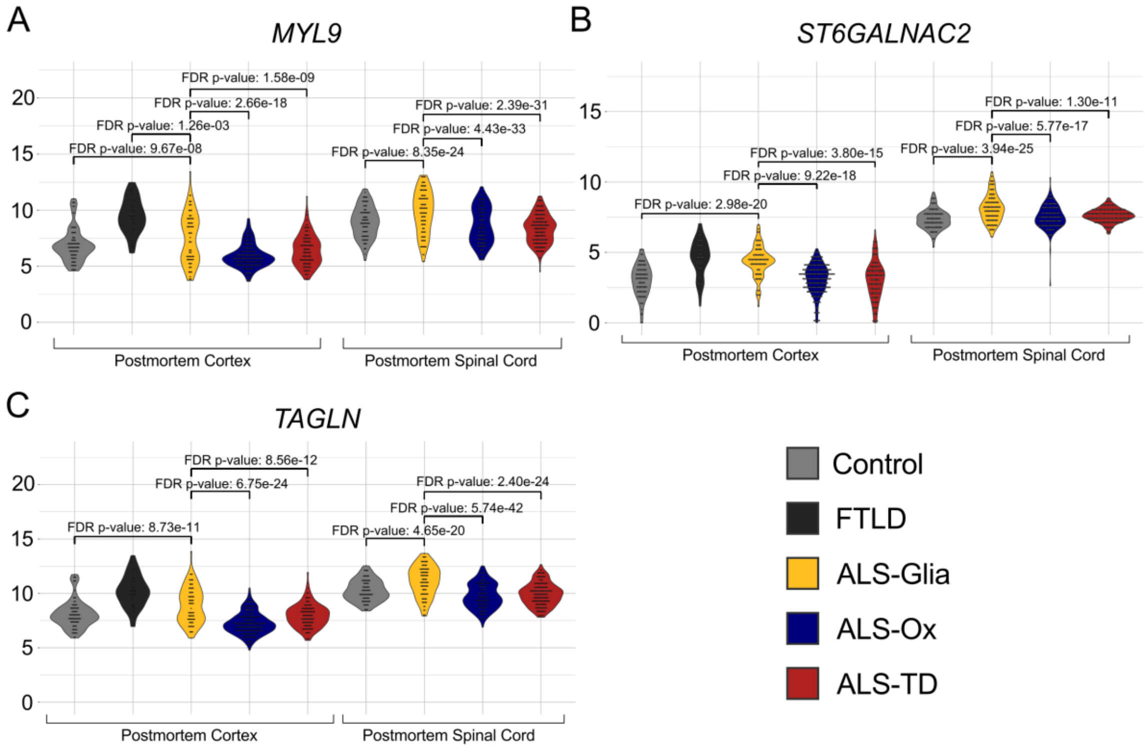


Figure 4.10: ALS-Glia marker genes in the postmortem cortex and spinal cord. Expression is separated both by subtype and CNS region for (A) *MYL9*, (B) *ST6GALNAC2*, and (C) *TAGLN*. All counts are presented on the $\log_2$ transformed median-of-ratio scale. All differential expression *p*-values are FDR adjusted.

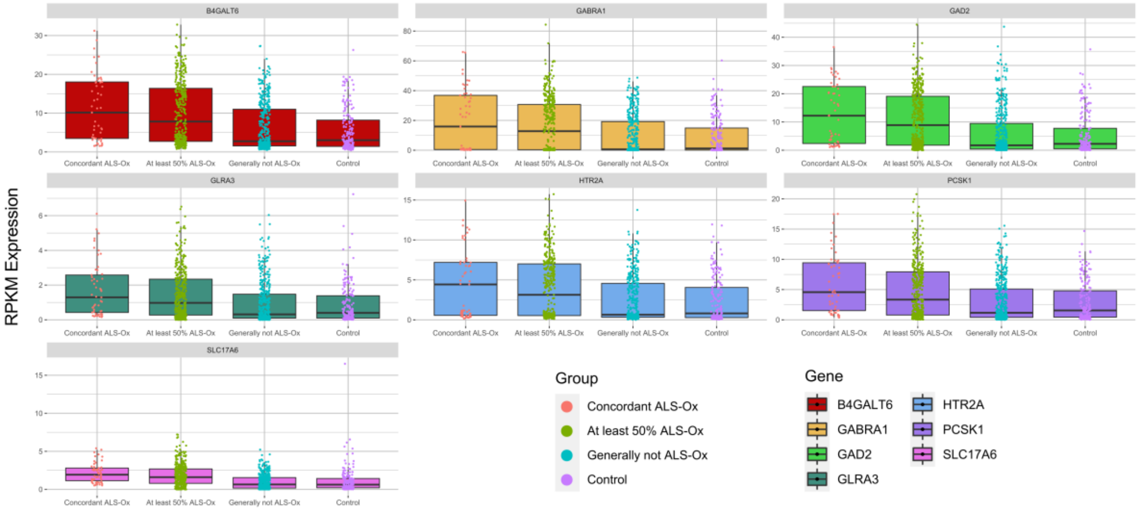


Figure 4.11: ALS-Ox marker gene expression on the FPKM count scale. Sample-level expression of ALS-Ox marker genes after grouping by ALS-Ox percentage, calculated by taking the number of intra-patient samples defined as ALS-Ox divided by the total number of samples from the patient. Transcript expression is normalized by library size to the FPKM scale. Concordant ALS-Ox patients were defined as ALS-Ox in all available tissue samples, while the ‘generally not ALS-Ox’ category is defined as less than 50% of samples classified as ALS-Ox. A total of 53 unique samples were included in the ‘Concordant ALS-Ox’ category, 393 samples in the ‘at least 50% ALS-Ox’, 410 samples in ‘generally not ALS-Ox’, and 184 control samples – corresponding to 206 ALS patients and 88 non-neurological controls.

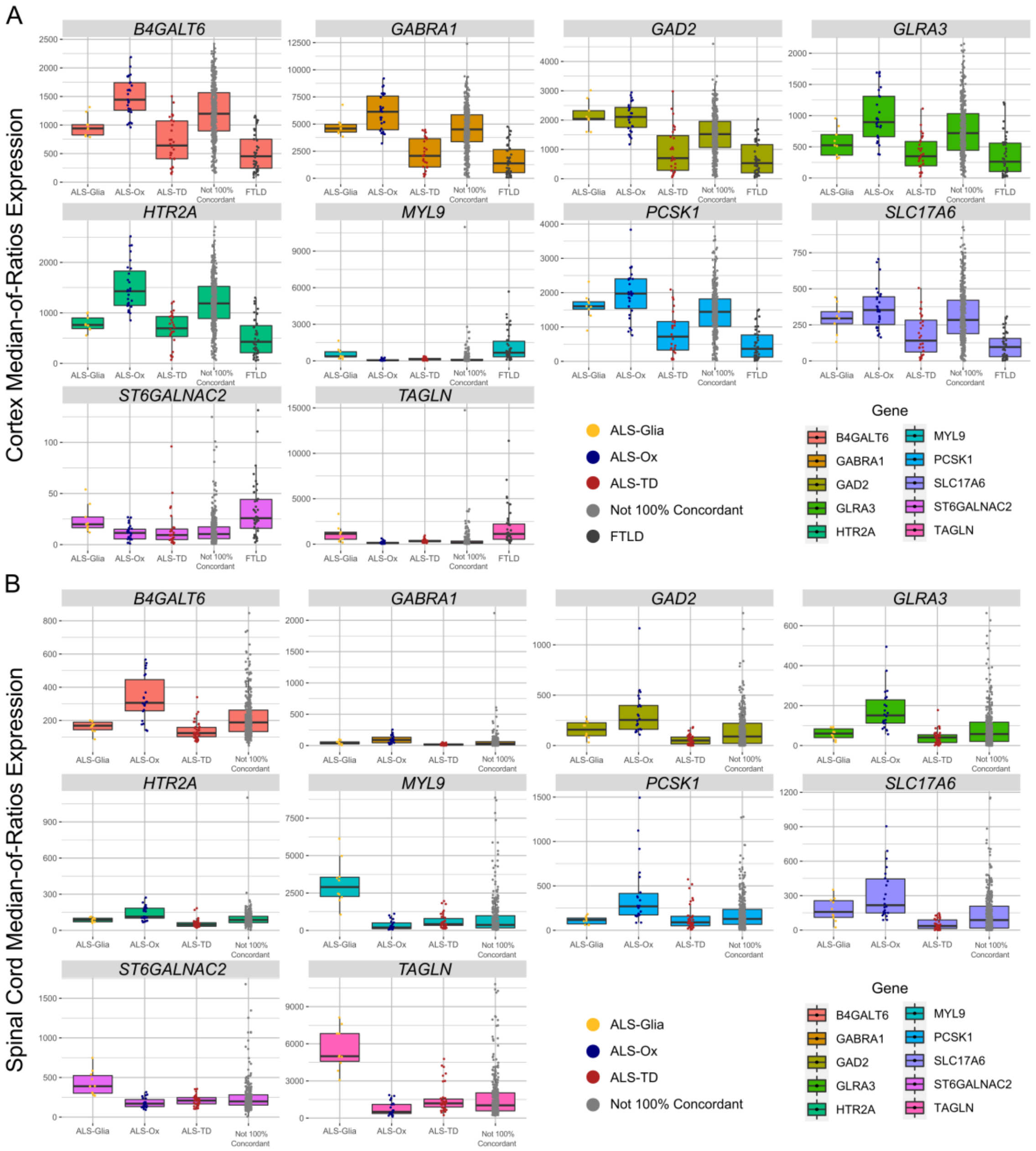


Figure 4.12: Marker gene expression in the perfectly concordant patient subset presented in Figure 2.20. Marker gene expression is presented on the DESeq2 median-of-ratios scale for samples originating from the (A) postmortem cortex or (B) spinal cord. In the cortex, a total of 9, 27, and 28 unique tissue samples are included in the ALS-Glia, ALS-Ox, and ALS-TD categories, respectively. 42 FTLD mimics are included and the remaining 380 transcriptomes are included in the “Not 100% Concordant” category. In the spinal cord, a total of 10 ALS-Glia, 22 ALS-Ox, and 38 ALS-TD transcriptomes comprise the perfectly concordant categories, with the remaining 346 samples as “Not 100% Concordant”.

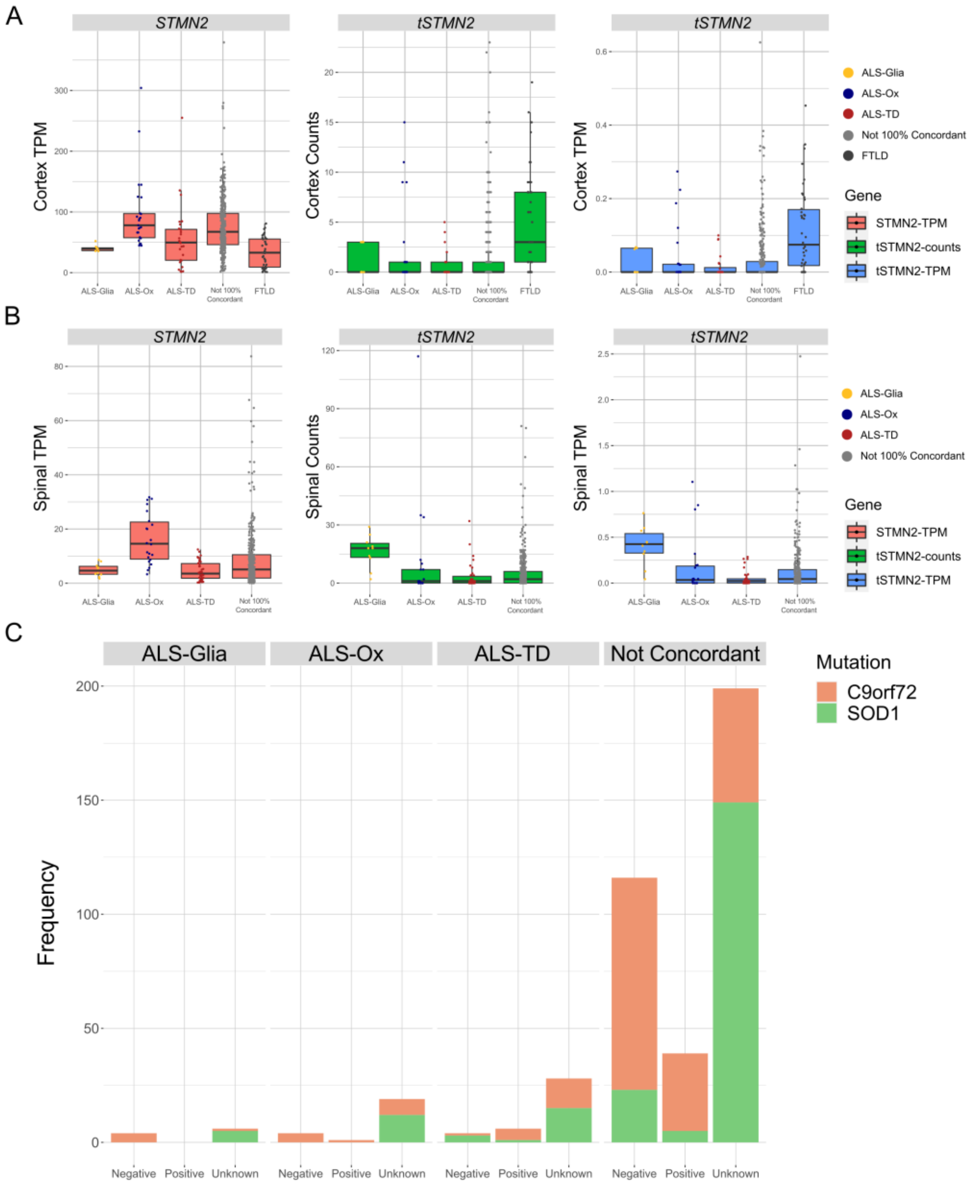


Figure 4.13: Truncated Stathmin-2 expression and genetic risk in the perfectly concordant patient subset in Figure 2.20. Full length and truncated Stathmin-2 expression presented on the transcript per million (TPM) or raw count scales in the (A) postmortem cortex and (B) spinal cord. In the cortex, a total of 9, 27, and 28 unique tissue samples are included in the ALS-Glia, ALS-Ox, and ALS-TD categories, respectively. 42 FTLD mimics are included and the remaining 380 transcriptomes are included in the "Not 100% Concordant" category. In the spinal cord, a total of 10 ALS-Glia, 22 ALS-Ox, and 38 ALS-TD transcriptomes comprise the perfectly concordant categories, with the remaining 346 samples as "Not 100% Concordant". (C) Genetic mutations in the perfectly concordant subset, with chi-squared test of independence indicating a significant association between (concordant) subtype and *C9orf72* mutation with ($p = 0.0009$) and without ($p = 0.0124$) the 'unknown' category.

4.3.4 Supervised Classification

In an effort to demonstrate the feasibility of patient stratification using postmortem cortex gene expression and facilitate clinical translation, supervised learning algorithms were leveraged to construct four subtype classifiers. To limit sequencing platform batch effects, only the features shared between the top 1,000 NovaSeq and HiSeq genes (by gene score) were included, totaling 299 transcripts. An independent validation cohort was not publicly available at the time of analysis, so it was decided to "holdout" all patient samples analyzed by HiSeq in an effort to reduce overestimation of classifier performance when applied to new cohorts. Predictive performance for each subtype was considered using the 'one-vs-rest' multi-class classification framework. As may be expected given the presentation of hybrid subtypes seen in Figure 2.22 and differential expression results presented in Figure 4.1, sensitivity and specificity metrics in the holdout cohort were relatively poor for all classifiers constructed, as shown in Figure 4.14. High F1 scores in the training and testing cohorts are expected and classification results in the cortex broadly highlight the challenges in using large gene panels to stratify ALS patients.

Building on the promising identification of seven ALS-Ox marker genes consistently upregulated in the cortex and spinal cord (Figure 4.9), a second round of classifiers were constructed using FPKM normalized expression of these features exclusively. In each case, classifier performance was assessed using FPKM normalized expression, an 80/20 train-test split, 100-fold cross validation, two classes ("Ox" and "NotOx"), and three different holdout (validation) cohorts comprised of all (i) postmortem spinal cord samples (ii) postmortem cortex samples and (iii) HiSeq samples. The first holdout cohort estimates predictive accuracy when assigning spinal cord subtype using cortex expression, while the opposite is true in the second holdout cohort. The final holdout cohort is designed to better estimate predictive accuracy when applied to new patient cohorts accounting for instrument-dependent expression. While it may be reasonable to assume that predicting the cortex phenotype using spinal cord expression is more clinically useful as it limits diagnostic invasiveness, models were constructed for both cases to demonstrate the capability of the subtype-specific transcripts to stratify the cohort regardless of region-dependent expression differences.

With the aim of reducing clinical diagnostic burden, classifiers were constructed from all three-gene combinations of ALS-Ox marker genes and screened for predictive power using partial least squares discriminate analysis[279] (PLS-DA). After training and testing the classifiers, the three-gene combination of *GAD2*, *GLRA3,* and *SLC17A6* was found to slightly outperform other gene combinations when predicting subtype in the spinal cord validation cohort (AUC = 0.927), shown in Figure 4.15A. Conversely, when training on the spinal cord cohort, *HTR2A, SLC17A6,* and *B4GALT6* gene set showed the highest predictive accuracy after application to the cortex validation cohort (AUC = 0.881), shown in Figure 4.15B. In the HiSeq validation cohort, *B4GALT6, GLRA3, SLC17A6,* and demonstrated the highest predictive accuracy (AUC = 0.831), suggesting these transcripts may be more invariant to differences in sample preparation and instrumentation, as seen in Figure 4.15C. Furthermore, the same gene combination of *B4GALT6, GLRA3,* and *SLC17A6* demonstrated the highest average AUC across all three validation cohorts (AUC

= 0.873), indicating these genes may be most robust for assigning ALS-Ox patient subtype. When compared to the PLS-DA classifier using all seven ALS-Ox marker genes, a decrease in predictive power is observed in the spinal cord holdout (AUC = 0.922), the cortex holdout (AUC = 0.861), and the HiSeq holdout (AUC = 0.809).

Leveraging results from PLS-DA, the classification analysis was extended by performing supervised machine learning using k-nearest neighbor (KNN), linear discriminant analysis (LDA), random forest (RF), support vector machine classifier (SVM), and multilayer perceptron (MLP) classification frameworks[260]. Classifiers were constructed using FPKM normalized expression of (i) the best three gene combination from PLS-DA (*B4GALT6*, *GLRA3*, *SLC17A6*), shown in Figure 4.16 and (ii) all seven ALS-Ox marker genes, shown in Figure 4.17. Using the top three discriminatory genes, the SVM classifier demonstrates the highest overall predictive accuracy when stratifying ALS-Ox and “not ALS-Ox” with median F1 scores from the test cohort ranging between 0.62–0.73 for ALS-Ox and 0.83–0.90 for ‘not ALS-Ox’, and holdout cohort AUCs ranging from 0.86–0.89. Similar performance is observed in the MLP classifier. In agreement with results observed during PLS-DA, the seven gene classifier generally demonstrated worse predictive accuracy in the cortex (AUCs = 0.84–0.86) and spinal cord (AUCs = 0.76–0.86) holdout cohorts, as seen in Figure 4.17A and 4.17B. However, improved predictive accuracy was seen when the seven-gene classifiers were applied to the HiSeq holdout cohort, with AUCs ranging from 0.86–0.91, shown in Figure 4.17C, suggesting the seven-gene classifier may outperform the three-gene as the ‘strength’ of batch effects and confounding covariates increases. Collectively, classification results demonstrate that the set of ALS-Ox marker genes established in this work can achieve appreciable stratification accuracy when predicting patient phenotype between regions of the central nervous system – with different cell type composition – or when using different instrumentation for quantification of gene expression.

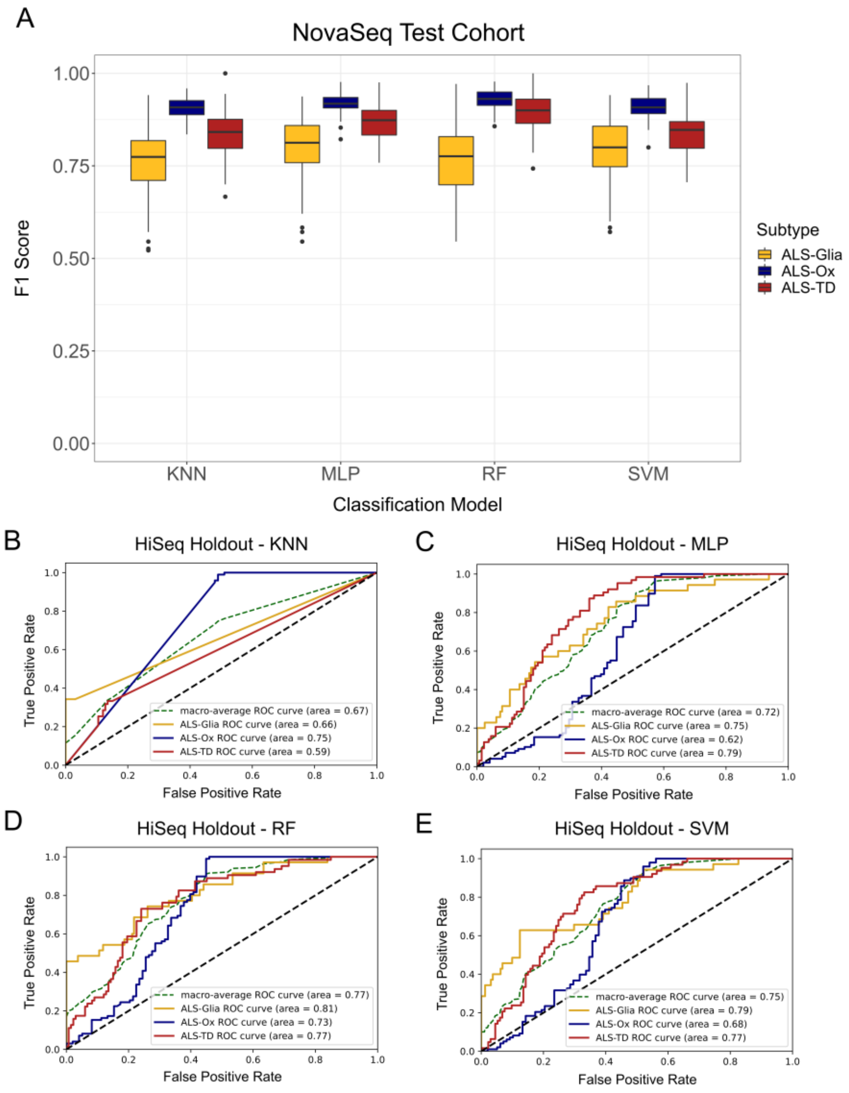


Figure 4.14: Supervised classification in the postmortem cortex cohort exclusively using 299 features. (A) F1 scores from 100-fold cross validation with the NovaSeq cohort are shown as boxplots, with *n*=208 patients in the training cohort and *n*=89 patients in the test

cohort. Four classification methods were considered (KNN, MLP, RF, and linear SVC) and predictive metrics are separated by subtype label. The MLP classifier demonstrated the highest average F1 score for the ALS-Glia subtype (0.80), while the RF classifier showed the best performance when predicting the ALS-Ox (0.93) and ALS-TD subtypes (0.90). The median is indicated by the solid black line, and first and third quartiles are captured by the bounds of the box. Boxplot whiskers are defined as the first and third quartiles –/+ interquartile range times 1.5, respectively, and outliers are denoted as solid black points. Minimum and maximum values are captured by the lowermost and uppermost points, respectively, or whisker bound if no outliers are shown. (B) ROC plot showing false positive rate (1-specificity) versus the true positive rate (sensitivity) for the KNN classifier when applied to the holdout (HiSeq) cohort. Given the multi-class nature of this analysis, three classifiers were constructed accounting for each binary case, using a 'one-versus-rest' approach. (C) ROC plot showing predictive metrics for the MLP classifier. (D) Sensitivity and specificity metrics for the random forest classifier when applied to the holdout cohort. (E) ROC plot for the linear SVM classifier show similar performance to the RF and MLP models. Using net reclassification improvement and integrated discrimination improvement methodology no single classifier was observed to outperform the others in the case of Glia vs rest. The SVM classifier was determined to outperform all other classifiers for the Ox vs rest case, and both the MLP and SVM classifiers were superior when compared to the RF model in the TD vs rest case.

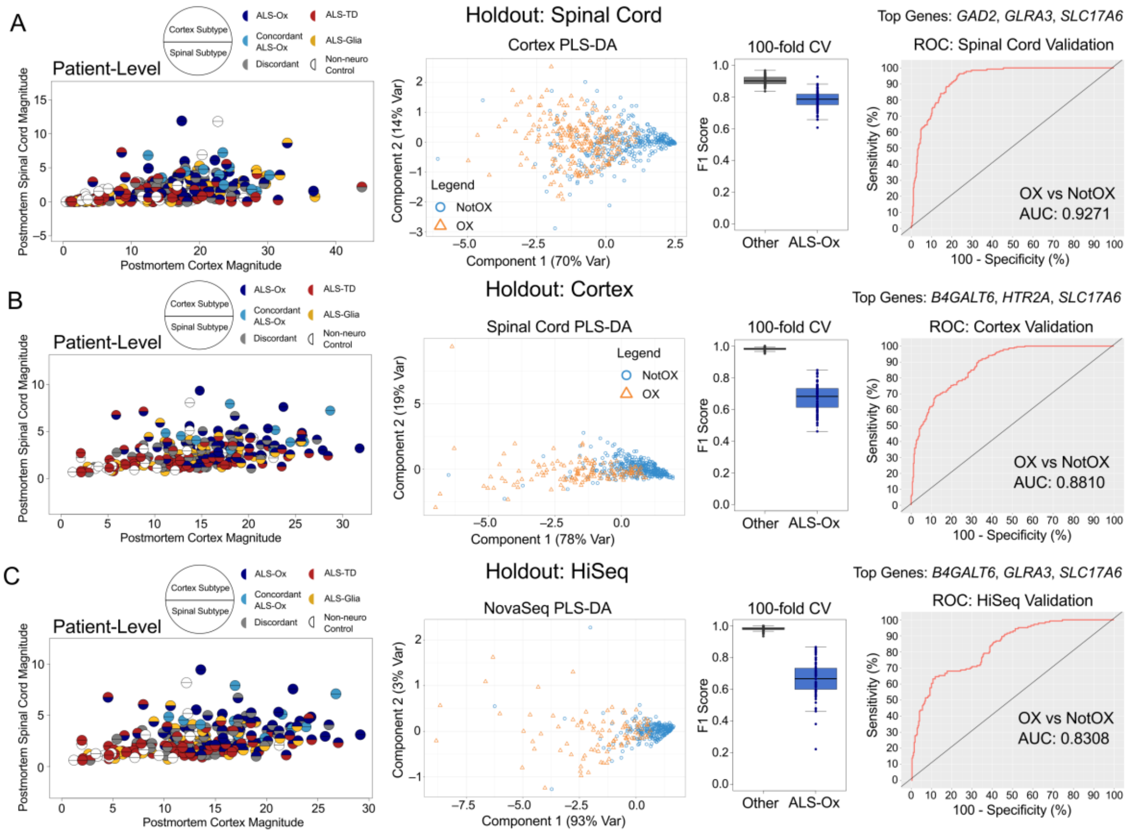


Figure 4.15: Three-gene PLS-DA classifiers for ALS-Ox patients. Partial least squares discriminate analysis with expression of transcripts normalized to the FPKM scale using library size and transcript length estimates from GRCh38.p12. In each case, visualization of patients is first performed using ALS-Ox marker genes, taking the mean FPKM expression magnitude – for the three-gene combination – from each available tissue sample. Majority assigned subtype is color-coded with the postmortem cortex and spinal cord presented in the upper and lower half circles, respectively. The PLS-DA classifier was then trained and tested using an 80/20 split of (A) all postmortem cortex samples, and validated using all spinal cord samples, (B) all postmortem spinal cord samples, and validated on the cortex holdout, and (C) all NovaSeq samples, and validated on the HiSeq holdout. Following PLS-DA, the training cohort is plotted using the first two components. Test cohort F1 metrics are presented as boxplots for 100 rounds of cross validation predicting ALS-Ox against all other samples, including non-neurological controls (‘Other’). Lastly, ROC plots showing application of the PLS-DA classifier to each of the three holdout cohorts. The top gene combination is provided for each holdout cohort, and the *B4GALT6*, *GLRA3*, *SLC17A6* trio was found to have the highest average AUC across all three holdout datasets.

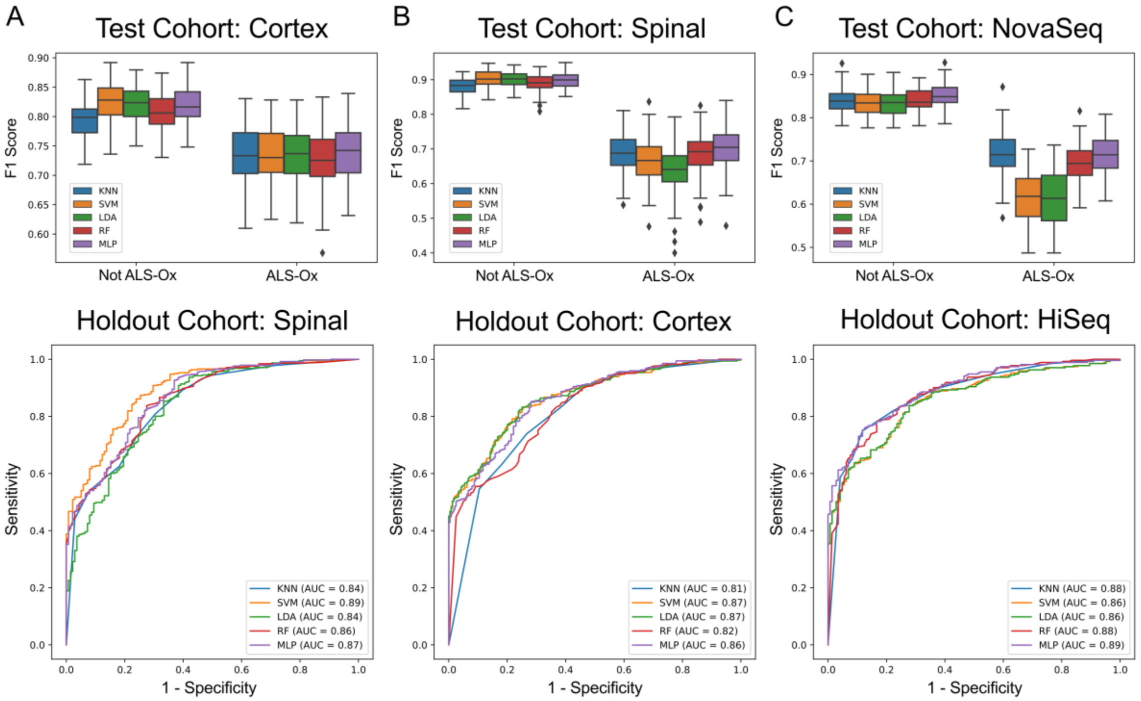


Figure 4.16: Supervised classification of ALS-Ox samples using expression of *B4GALT6*, *GLRA3*, and *SLC17A6* marker genes. Five different classification algorithms were considered. F1 scores obtained from 100-fold cross validation in the test cohort are presented first and separated by classification level (‘Ox’ vs ‘NotOx’). A combined total of 1,104 ALS and control samples are considered, with *n*=377 (~34%) assigned the ALS-Ox label. The five classifiers were constructed and applied to three different holdout/validation cohorts comprised of (A) all postmortem spinal cord samples (*n*=519), (B) all postmortem cortex samples (*n*=585), and (C) all samples analyzed by HiSeq (*n*=415). ROC plots are presented second, for each classifier, and show sensitivity and 1-specificity metrics when applied to the specified holdout cohort.

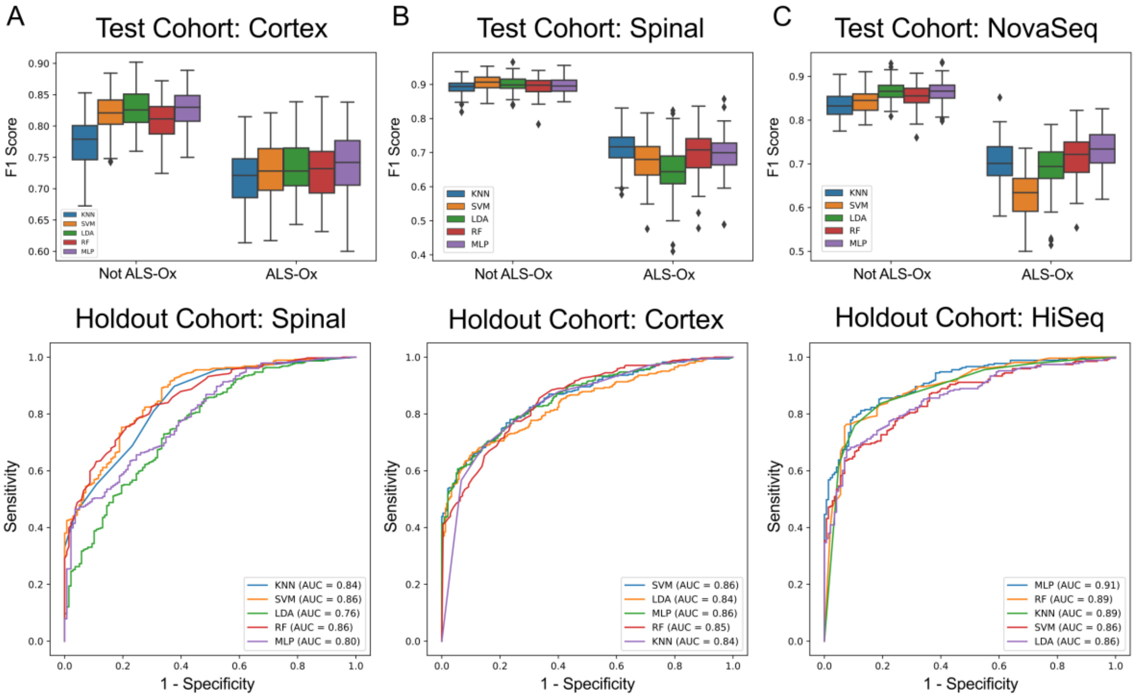


Figure 4.17: Supervised machine learning classifiers using all seven ALS-Ox marker genes. Five different classifiers were constructed, using FPKM normalized expression. F1 scores from 100 rounds of cross validation are presented as boxplots, for each classifier considered. F1 scores are separated by class level (ALS-Ox and 'Not ALS-Ox'). A combined total of 1,104 ALS and control samples are considered, with *n*=377 (~34%) assigned the ALS-Ox label. Classifiers were constructed and applied to three different holdout cohorts comprised of (A) all postmortem spinal cord samples (*n*=519), (B) all postmortem cortex samples (*n*=585), and (C) all samples analyzed by HiSeq (*n*=415). ROC plots are presented second, for each classifier, and show sensitivity vs 1-specificity metrics when applied to the specified holdout cohort.

Table 4.2. Confusion matrix for the three gene classifiers presented in Figure 4.16A, which uses the cortex transcriptomes for training and the spinal cord for validation.

Three Gene Confusion Matrix			Predicted	
Classifier			ALS-Ox	Not ALS-Ox
KNN	Observed	ALS-Ox	105	33
		Not ALS-Ox	104	277
LDA	Observed	ALS-Ox	92	46
		Not ALS-Ox	64	317
MLP	Observed	ALS-Ox	101	37
		Not ALS-Ox	76	305
RF	Observed	ALS-Ox	100	38
		Not ALS-Ox	76	305
SVM	Observed	ALS-Ox	97	41
		Not ALS-Ox	42	339

Table 4.3. Confusion matrix for the three gene classifiers presented in Figure 4.16B, which uses the spinal cord transcriptomes for training and the cortex for validation.

Three Gene Confusion Matrix			Predicted	
Classifier			ALS-Ox	Not ALS-Ox
KNN	Observed	ALS-Ox	148	91
		Not ALS-Ox	59	287
LDA	Observed	ALS-Ox	103	136
		Not ALS-Ox	20	326
MLP	Observed	ALS-Ox	131	108
		Not ALS-Ox	30	316
RF	Observed	ALS-Ox	115	124
		Not ALS-Ox	26	320
SVM	Observed	ALS-Ox	113	126
		Not ALS-Ox	24	322

Table 4.4. Confusion matrix for the three gene classifiers presented in Figure 4.16C, which uses the NovaSeq transcriptomes for training and the HiSeq for validation.

Three Gene Confusion Matrix Classifier			Predicted	
			ALS-Ox	Not ALS-Ox
KNN	Observed	ALS-Ox	112	32
		Not ALS-Ox	49	222
LDA	Observed	ALS-Ox	77	67
		Not ALS-Ox	28	243
MLP	Observed	ALS-Ox	97	47
		Not ALS-Ox	31	240
RF	Observed	ALS-Ox	98	46
		Not ALS-Ox	36	235
SVM	Observed	ALS-Ox	80	64
		Not ALS-Ox	28	243

Table 4.5. Confusion matrix for the classifiers developed using all seven ALS-Ox marker genes presented in Figure 4.17A. Cortex transcriptomes are used for model training and the spinal cord is used to validate predictive performance.

Seven Gene Confusion Matrix Classifier			Predicted	
			ALS-Ox	Not ALS-Ox
KNN	Observed	ALS-Ox	106	32
		Not ALS-Ox	119	262
LDA	Observed	ALS-Ox	85	53
		Not ALS-Ox	92	289
MLP	Observed	ALS-Ox	89	49
		Not ALS-Ox	84	297
RF	Observed	ALS-Ox	92	46
		Not ALS-Ox	65	316
SVM	Observed	ALS-Ox	92	46
		Not ALS-Ox	38	333

Table 4.6. Confusion matrix for the classifiers developed using all seven ALS-Ox marker genes presented in Figure 4.17B. Spinal cord transcriptomes are used for model training and the cortex is used to validate predictive performance.

Seven Gene Confusion Matrix			Predicted	
Classifier			ALS-Ox	Not ALS-Ox
KNN	Observed	ALS-Ox	166	73
		Not ALS-Ox	62	284
LDA	Observed	ALS-Ox	100	139
		Not ALS-Ox	36	310
MLP	Observed	ALS-Ox	148	91
		Not ALS-Ox	48	298
RF	Observed	ALS-Ox	144	95
		Not ALS-Ox	47	299
SVM	Observed	ALS-Ox	120	119
		Not ALS-Ox	36	310

Table 4.7. Confusion matrix for the classifiers developed using all seven ALS-Ox marker genes presented in Figure 4.17C. NovaSeq transcriptomes are used for model training and the HiSeq is used to validate predictive performance.

Seven Gene Confusion Matrix			Predicted	
Classifier			ALS-Ox	Not ALS-Ox
KNN	Observed	ALS-Ox	116	28
		Not ALS-Ox	44	227
LDA	Observed	ALS-Ox	85	59
		Not ALS-Ox	35	236
MLP	Observed	ALS-Ox	107	37
		Not ALS-Ox	36	235
RF	Observed	ALS-Ox	103	41
		Not ALS-Ox	32	239
SVM	Observed	ALS-Ox	80	64
		Not ALS-Ox	26	245

4.3.5 TDP-43 Associated Pathology and Truncated Stathmin-2

Given the near ubiquitous nature of TDP-43 cellular inclusions in ALS patients, expression of the transcript encoding the protein, *TARDBP*, was examined in both the cortex and spinal cord cohorts with all ALS patients and controls. No significant differences in *TARDBP* expression were observed between any of the ALS subtypes, as seen in Figure 4.18. These findings are generally supported by the foundational study from Neumann *et al*.[243], which identify extensive post translation modification (hyperphosphorylation) to the TDP-43 protein, but conflict with findings from Tam *et al*. in a subset of the ALS cohort considered in this dissertation[317]. Larger sample size and different normalization procedures may explain some of the discrepancy in subtype *TARDBP* expression.

TDP-43 is known to participate in the transcriptional regulation of the *STMN2* gene, where cytoplasmic mislocalization of TDP-43 is associated with cryptic exon splicing and the truncation of the normal length *STMN2* transcript, resulting in a non-functional protein and impaired axonal regeneration and neuromuscular junction maintenance[20,152,179,201,229]. Prudencio *et al*.[266] previously quantified expression of truncated *STMN2* in this cohort, therefore an extension to this analysis was performed by considering truncated and normal length *STMN2* in the context of the identified cortex subtypes, presented in Figure 4.19. Although no subtype was seen to characteristically express truncated *STMN2* in the cortex, ALS-Ox samples had significantly upregulated expression of the full length *STMN2* transcript. In the ALS-Ox spinal cord, statistically significant upregulation of the *STMN2* transcript and the truncated pathological form associated with TDP-43 cryptic exon splicing is observed when compared to the ALS-TD subtype, shown in Figure 4.20. These findings are unexpected, given that a hyperactive TDP-43 pathology is reported in ALS-TE patients by Tam *et al*.[317] and our own enrichment and differential expression results continue to implicate changes to transcriptional regulation in the ALS-TD subtype. Intuitively, overexpression of the *STMN2* cryptic exon is to be expected[20], yet to the contrary cryptic exon expression is *lowest* in the ALS-TD subtype. Thus, results indicate that additional work is needed to clarify the role of the TDP-43 protein and its association with cryptic exon splicing in ALS-TD patients. One possible explanation could be phenotype-dependent clearance of truncated transcripts with ALS-TD patients best "primed" for degradation, but quantification or methodological bias cannot be ruled out. Lending support to the proposed subtype-dependent variability in cryptic exon expression, a TDP-43 repressed exon in the *UNC13A* gene is only observed in a subset of patient samples (< 40%)[46,212].

As demonstrated by Tam *et al*. and others[197,206,317], the TDP-43 protein serves as an extensive transcriptional regulator of non-coding RNA, binding both DNA and RNA to modulate expression[64]. TDP-43 interacts with a diversity of non-coding RNA classes including intronic, antisense, intergenic, long and short non-coding, 3` and 5` untranslated regions, and LINE, SINE, and LTR retrotransposons[317]. As a consequence of the demonstrated link between TDP-43 and retrotransposon expression, TEs were quantified in both the cortex and spinal cord cohorts with gene locus resolution using SQuIRE[359]. Worth noting, TE features considered in this work differs from that of Tam *et al*.[317] given their TE quantification pipeline[160] sums gene locus expression from the same family, a step that was found to introduce undesirable batch effects following hierarchical clustering and co-expression analyses[99]. TE expression was considered in both the cortex and spinal cord cohorts using the quality control and expression thresholds detailed in *Chapter 2.2.3*. Only TEs that passed median absolute deviation and sequencing platform filtering steps are considered herein, totaling 426 in the cortex and 86 in the spinal cord. After adjusting for chromosome length (in Mbp), no chromosomes were found to be overrepresented in the cortex but interesting differences are observed in the spinal cord, as seen in Figure 4.21 – although lower feature number may partially explain the results. Examining retrotransposon expression at the family level, analogous to the TE features considered by Tam *et al*.[317], LINE, SINE, and SINE-VTNR-*Alu* elements are seen to contribute most to the observed TE expression profile, in agreement with other works[97,239] – presented in Figure 4.22. TE expression was not a defining characteristic of a single subtype, as seen in Figures 2.9E and 2.10, and differentially expressed retrotransposons are observed in each cortex subtype, presented in Figures 4.23, 4.24, and 4.25. In the spinal cord cohort, fewer globally expressed TEs are recovered, yet results continue to show a lack of specificity for retrotransposon expression in a single subtype as seen in Figure 4.26. Interestingly, ALS-Glia spinal cord samples generally continue to show lower expression of TEs relative to the other two subtypes. Taken together, TE expression is altered in ALS when compared to controls and broad differences are observed in expression levels between subtypes (Figure 2.9, 2.10, and 4.26A) – although these differences don’t associate strongly with TE sequence similarity (family). Continued work is needed to clarify the role of TE expression in the presentation of ALS subtypes and studies considering genome-wide methylation in tandem may provide significant new insight.

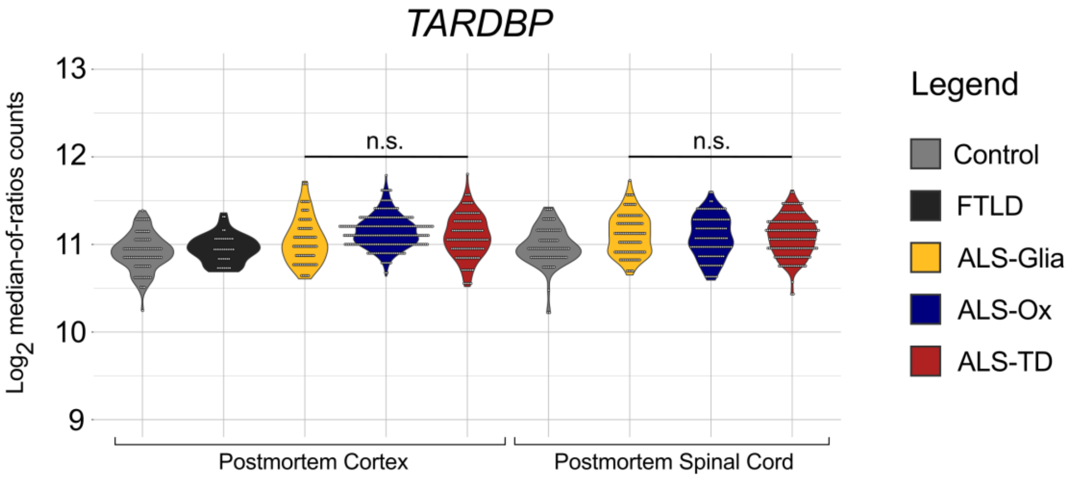


Figure 4.18: Expression of transcript *TARDBP*, encoding ALS disease-associated protein TDP-43, is presented to show expression differences between subtypes are not observed in the postmortem spinal cord, as well as the cortex. These findings support TDP-43 pathology occurring at the protein level rather than transcript level.

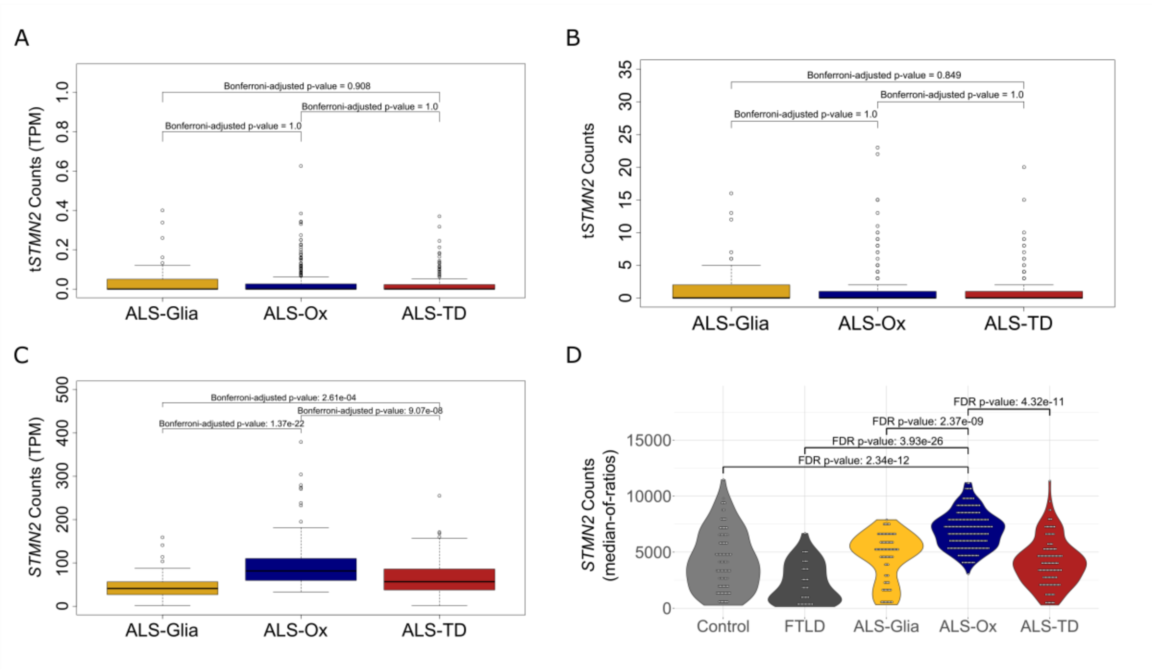


Figure 4.19: Truncated and normal length Stathmin-2 in the postmortem cortex. A Mann-Whitney U test (two-sided) was used to assess statistical significance in tSTMN-2 expression on both the (A) TPM scale and (B) raw count scale. After adjusting p-values for multiple hypothesis testing using the Bonferroni method, truncated *STMN2* expression was not observed to have any association with ALS subtype. (C) Full length transcript *STMN2* counts on TPM scale, evaluated using the Mann-Whitney U test (two-sided), with Bonferroni-adjusted p-values shown. (D) Full length transcript *STMN2* counts on the DESeq2 median-of-ratios scale. Healthy control donors and FTLD patients are included, in an effort to improve the estimation of size factors for normalization. *P*, DESeq2[208] differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

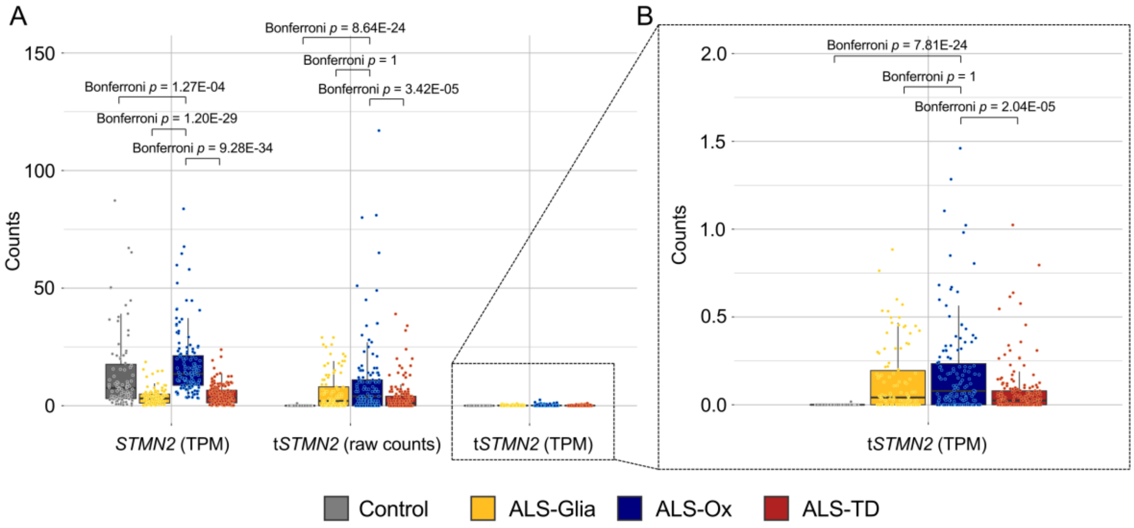


Figure 4.20: Expression of full length and truncated *STMN2* in the spinal cord. (A) A two-sided Mann-Whitney U test was used to assess statistical significance in *STMN2* and truncated *STMN2* expression on both TPM scale and raw count scale. After adjusting p-values for multiple hypothesis testing using the Bonferroni method[36], truncated *STMN2* expression was elevated in the postmortem spinal cord of ALS-Ox patients when compared to the ALS-TD subtype on both count scales, further supporting phenotypic differences between ALS-Ox and ALS-TD patients. These finds are somewhat surprising, given ALS-TD pathology appears more closely linked to transcription as compared to ALS-Ox. (B) Truncated *STMN2* counts on the TPM scale are replotted for visual clarity. No statistically significant differences in the expression of *STMN2* or truncated *STMN2* are observed between ALS-Glia and ALS-TD subtypes.

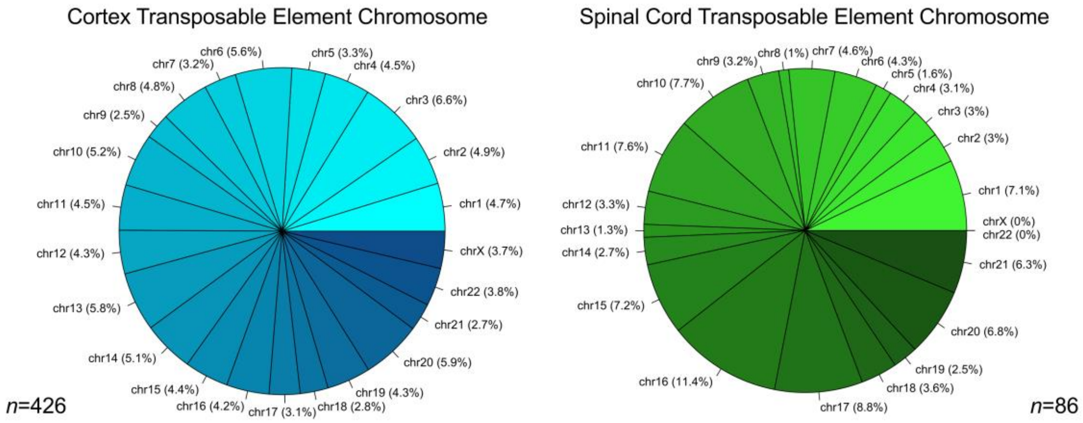


Figure 4.21: Assessment of chromosome origin for TEs identified in the cortex (*left*) and spinal cord (*right*). Only the TEs included in the top genes selected by gene score and shared by both sequencing platforms were considered in the analysis, totaling 426 non-redundant transcripts in the cortex and 86 in the spinal cord. Frequency was adjusted for chromosome length in Mbp.

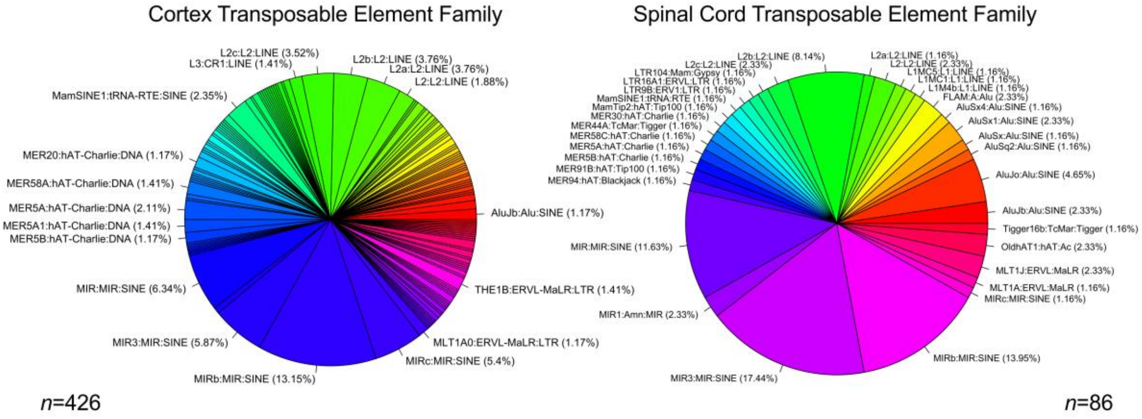


Figure 4.22: Assessment of transposon family for TEs identified in the cortex (*left*) and spinal cord (*right*). For the cortex, family names with <1% frequency were not printed but pie chart colors match the legend presented in Figure 2.10A.

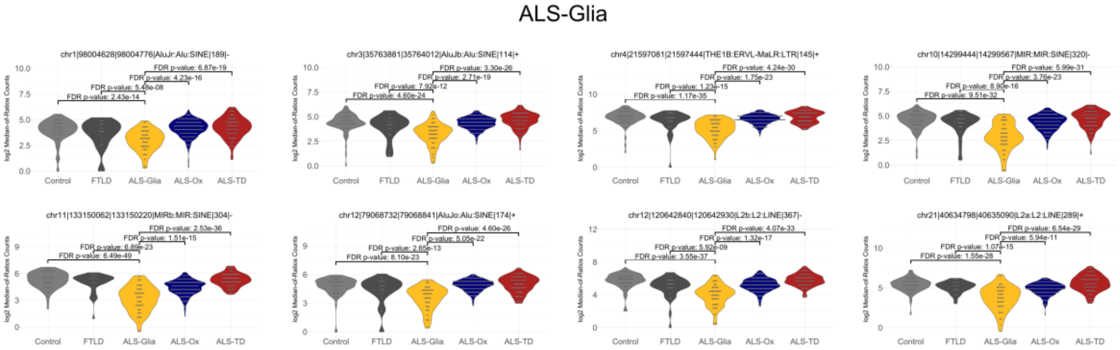


Figure 4.23: ALS-Glia specific transposable element expression in the cortex, with eight representative transcripts shown. The ALS-Glia subtype was defined by downregulated expression of TEs, as compared to other ALS subtypes and controls. *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

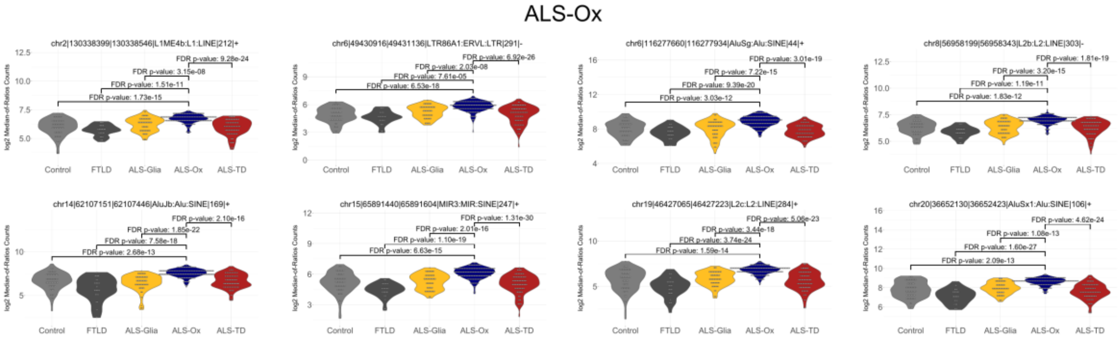


Figure 4.24: ALS-Ox specific transposable element expression in the cortex, with eight representative transcripts shown. The ALS-Ox subtype was defined by upregulated expression of long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs). *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

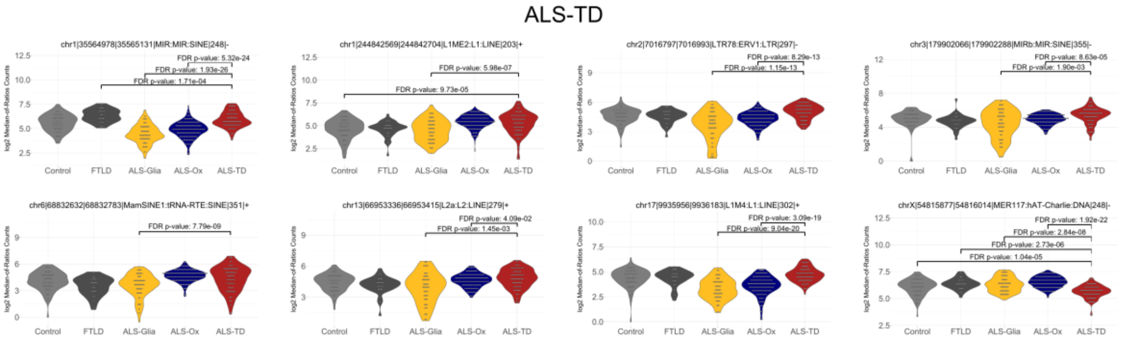


Figure 4.25: ALS-TD specific transposable element expression in the cortex, with eight representative transcripts shown. The ALS-Ox subtype was defined by upregulated expression of long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs). *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

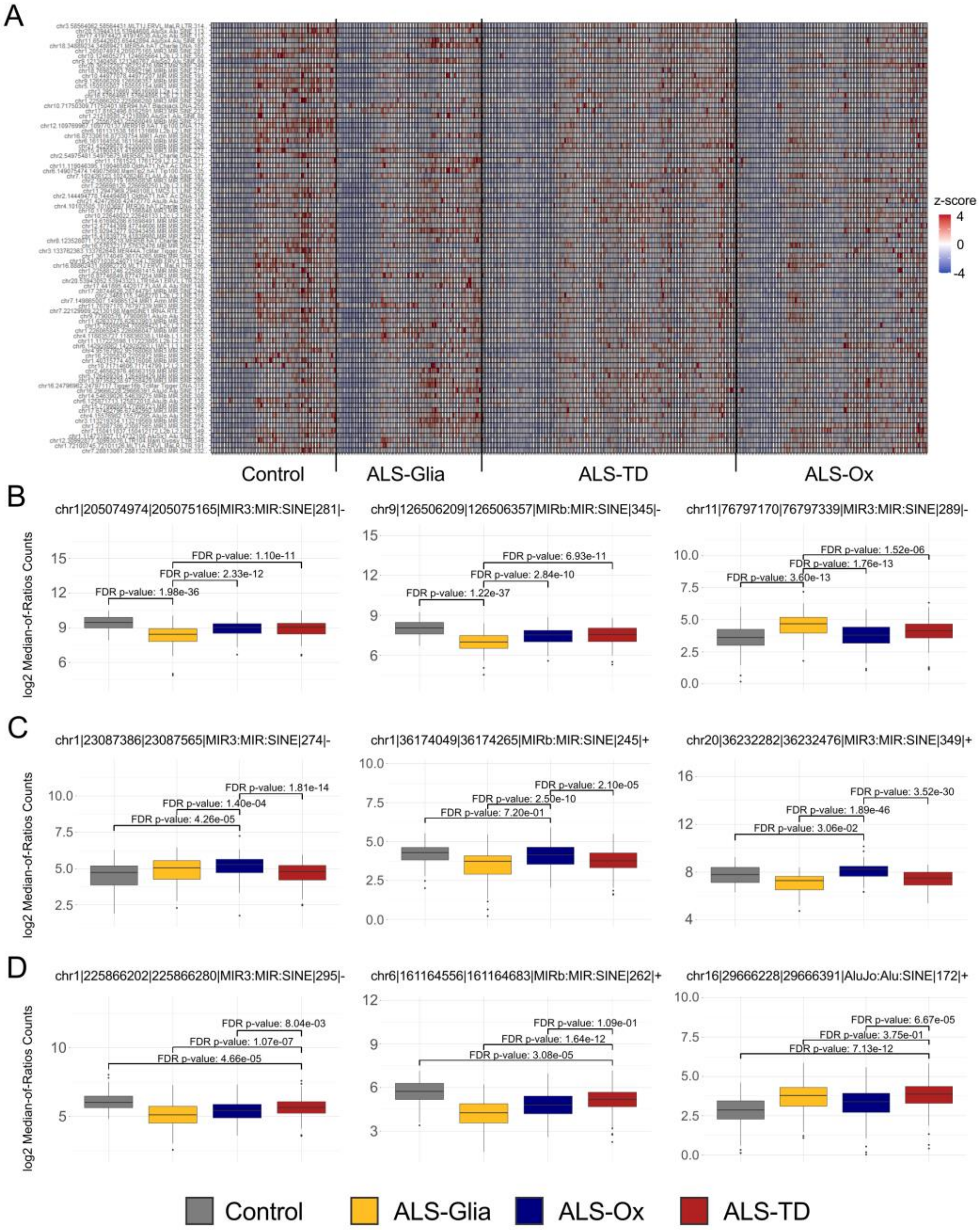


Figure 4.26: Retrotransposon expression in the ALS spinal cord. TEs were only considered if they passed filtering by median absolute deviation and were shared between sequencing platform subgroups, totaling 86 non-redundant transcripts. (A) A heatmap showing spinal cord TE expression in 519 samples (*n*=428 ALS and 91 controls) with count values normalized by DESeq2 and z-score transformed before plotting. Three differentially expressed TEs are shown for (B) ALS-Glia samples, (C) ALS-Ox samples, and (D) ALS-TD samples. Expression values are normalized to the DESeq2 median-of-ratios scale and $\log_2$ transformed prior to plotting. *P*, DESeq2 differential expression using the negative binomial distribution, two-tailed, FDR method for multiple hypothesis test correction.

4.4 Discussion

The implications and relevance of differentially expressed transcripts are considered in the context of each ALS subtype, and is followed by a more general discussion of the results obtained in this Chapter.

4.4.1 ALS-Glia

In the ALS-Glia subtype from the cortex cohort, significantly elevated expression of microglia, astrocyte, and oligodendrocyte marker genes is observed (*AIF1*[317], *CCR5*[109], *CD44*[219], *CD68*[32], *CHI3L2*[285], *CR1*[88], *CX3CR1*[56], *HLA-DRA*[141], *MSR1*[172], *TLR7*[51], *TMEM125*[53], *TNC*[53], *TREM2*[171], and *TYROBP*[171,366]). ALS-Glia upregulation of *CHI3L2*, *CX3CR1*, *FOLH1*, *HLA-DRA*, *ALOX5AP*, *CCR5*, *CR1*, *FPR3*, *NCF2*, *TLR8*, and *TNFRSF25* generally indicates a pro-neuroinflammatory and pro-apoptotic disease phenotype[29,88,114,119,177,190,205,213,224,226,285]. ALS-Glia negative enrichment for PI3K/AKT signaling further supports a pro-apoptotic disease phenotype[144]. Elevated expression of *TREM2*, *TYROBP*, and *CLEC7A* may suggest a compensatory neuroprotective mechanism, where the activated (DAM) microglia state enhances phagocytic clearance and slows neurodegeneration[171,316]. The DAM phenotype is also known to promote ROS generation and neuroinflammation[298], obscuring the relationship between disease-associated microglia and ALS-Glia pathogenesis. Alterations to lipid metabolism in the ALS-Glia subtype are evidenced by *APOBR*, *APOC1*, and *APOC2* overexpression compared to ALS-Ox and ALS-TD patients, and may further reflect the elevated *APOE* and *LPL* expression seen in disease-associated microglia[161,171]. Interestingly, upregulated expression of transcripts *CX3CR1*, *TYROBP* and *TREM2* in this subtype possibly suggests dysregulation or competition between homeostatic and activated microglia phenotypes[171]. Relatedly, increased expression disease associated astrocyte[135] (DAA) marker genes is observed in the ALS-Glia subtype, including *ITIH3*, *KCNIP4*, *PDGFD*, *ST6GALNAC5*, and *TNC*. Interestingly, ALS-Glia expression of DAA genes suggests the astrocyte population in these patients captures both disease-associated and homeostatic phenotypes, when compared to healthy control donors.

Consistent with the cortex ALS-Glia subtype, characteristic expression of many Fc-gamma receptors and MHC Class II molecules are seen. Heightened *VRK2* expression suggests some anti-apoptotic regulation occurs in ALS-Glia patients[234]. Overexpression of *FOLH1* may provide evidence for glutamate excitotoxicity susceptibility in the ALS-Glia subtype[119]. Elevated transcription of *ST6GALNAC2* suggests alterations to post-translational protein O-glycosylation, while *NINJ2* expression may support the proclivity for neuronal damage and death. Although additional work is needed to better understand the consequences of the apparently dichotomous microglial phenotypes in the ALS-Glia frontal and motor cortex, these results clearly demonstrate that a subset of ALS patients are defined by glial activation and elevated inflammatory signaling. More generally, activated microglia and astrocytes are known promote cytotoxicity in motor neurons[198,369], providing a direct framework linking the neuroinflammatory phenotype in ALS-Glia patients to more rapid disease progression.

In the spinal cord, ALS-Glia specific expression was primarily recovered from genes associated with neuroinflammation and included: *CD59*, *CD300E*, *CFH*, *IL1R1*, *SAMHD1*, *SELP*, and *TMEM173* (*STING1*). Outside of the expected neuroinflammatory phenotype and more closely resembling the ALS-Ox cortex phenotype, elevation of *AHNAK* implicates blood-spinal cord barrier and calcium homeostasis disruptions while expression of *IQGAP1*, *LAMC1*, *MAP7*, and *VIM* suggests changes to cytoskeletal and extracellular matrix organization. In line with findings from the cortex, glycosylation gene *ST6GALNAC2* remains upregulated in the spinal cord of ALS-Glia samples, while elevated *B4GALNT2* and *GLT8D2* lends additional support for the relevance of this post-translational modification in the presentation of ALS subtype. Additionally, *MYL9* and *TAGLN* genes remain overexpression in the ALS-Glia spinal cord when compared to the other two subtypes. Broad downregulation of *CAPN3* autocatalytic protease relative to non-neurological controls suggests proteotoxic stress is a conserved phenotype in the spinal cord of all ALS patients. Of the transcripts associated with the ALS-Glia subtype in the spinal cord, *AHNAK, CAPN3, CD59, CFH, CHEK2, IL1R1*, *LAMC1*, *MAP7, SAMHD1*, *STING1*, and *VIM* have been previously linked to CNS injury, neurodegeneration, or ALS specifically[14,41,112,130,135,209,261,273,293,318,331,333]. Interestingly, *SAMHD1* has been shown to regulate L1 retrotransposon expression[318] and overexpression in ALS-Glia patients continues to support a role for TEs in ALS neurodegeneration. While not directly associated with neurodegeneration, P-selectin (*SELP*) has been shown to mediate the microglial phenotype in glioblastoma, with elevated P-selectin protein levels associated with a more aggressive tumor proliferation and invasion[361], in good agreement with the

ALS-Glia phenotype and established clinical associations. Elevated expression of *STING1* in the spinal cord of ALS-Glia patients implicates the recently elucidated cGAS-STING signaling pathway in age- and neurodegenerative-related inflammation and microglia activation[130]. Finally, in agreement with the DAA phenotype observed in the cortex, elevated expression of DAA marker gene *VIM* continues to implicate reactive astrocytes in the ALS-Glia phenotype[135].

4.4.2 ALS-Ox

The cortex ALS-Ox subtype is defined by oxidative stress, evidenced by upregulated expression of *OXR1* and *SOD1* and downregulation of *CP* (ceruloplasmin), *UCP2*, and oxidative phosphorylation genes *NDUFA4L2, TCIRG1*, and *COX4I2*[6,165,207,220,250]. *NDUFA4L2* and *BECN1* expression further implicate impaired autophagy in ALS-Ox pathology[242,354]. Subtype-specific expression of many synaptic signaling associated genes are observed, including: *GABRA1* (GABA receptor), *GABRA6*, *GAD2* (catalyzes production of GABA), *GLRA2* (glycine receptor), *GLRA3, HTR2A* (serotonin receptor), *KCNV1* (voltage-gated ion channel), *KCNMB1*, *PCSK1*[9], *SLC6A13* (GABA transporter), *SLC17A6* (glutamate transporter), *SLC17A8* (glutamate transporter), and *TCIRG1* (proton transporter associated with synaptic vesicle formation[165]). Together, the upregulated transcription of *GABRA1, GABRA6, GAD2, GLRA2,* and *GLRA3* and downregulation of *SLC6A13* strongly suggest increased inhibition in the ALS-Ox frontal and motor cortex. Increased expression of *SLC17A6* and *SLC17A8* is hypothesized to reflect a neuronal process to alleviate reduced excitability. Elevated transcription of

BECN1, *PFDN4*, *SERPINI1* (neuroserpin), *UBQLN1*, and *UBQLN2* suggests proteotoxic stress is also a defining characteristic of this ALS subtype[9,103,146,147,317].

Downregulation of *NOS3*, *NOTCH3*, *MYH11*, *MYL9*, and *TAGLN* in the cortex may implicate pericyte and vascular smooth muscle cell dysfunction and alterations to the blood-brain barrier in ALS-Ox patients[158,313,314]. Similar to the ALS-Glia subtype, *B4GALT6* overexpression suggests changes to the O-glycosylated proteome. Evidence for alterations to the extracellular matrix, in the frontal and motor cortex of ALS-Ox patients, is observed in the downregulated expression of *ADAMTSL4, ADAMTS7, ADAMTS14, COL1A1, COL1A2, COL2A1, COL3A1, COL4A6, COL6A3*, *COL8A1, COL14A1, COL18A1,* and *TAGLN*. Interestingly, Collins *et al*. demonstrate that alterations to the extracellular matrix persist at the protein level[71]. Importantly, upregulated transcription of marker genes *GABRA1*, *GAD2*, *HTR2A*, and *PCSK1* is observed in ALS-Ox patients, which have been previously reported to be downregulated in Alzheimer's patients[319], suggesting distinct synaptic signaling pathological mechanisms. Taken together, these results generally suggest ALS-Ox patients reflect more traditional neurodegenerative themes, such as oxidative and proteotoxic stress, impaired blood-brain barrier function, and alterations to synaptic signaling.

In the ALS-Ox spinal cord, elevated expression of genes *B4GALT6*, *GABRA1*, *GAD2*, *GLRA3*, *HTR2A*, *PCSK1*, and *SLC17A6* persists, leading us to define these seven transcripts as marker genes. Expression levels are dependent on tissue region but remain consistently overexpressed in ALS-Ox patients relative to FTLD and non-neurological controls. In agreement with findings from the cortex, ALS-Ox specific transcript expression in the spinal cord was generally associated with neuronal signaling and included: *GABRG3*, *GRIA1*, *GRIN2A*, *GRM1*, *KCNH6*, *KCNS2*, *NTS*, *PCLO*, *RIMS2*, *SCN3A*, *SLC35F4*, *SYN2*, *SYT1*, and *UNC13C*. Oxidative and proteotoxic stress phenotypes are weaker in the ALS-Ox spinal cord as compared to the cortex, seen in the expression of genes *BECN1*, *OXR1*, *SOD1*, and *UBQLN2*. Similar to the Glia phenotype, upregulation of glycosylation genes *B4GALT6* and *GALNT14* suggests this PTM plays a role in disease heterogeneity. Interestingly, *CPNE4* has been previously linked to ALS as a SNP-associated risk gene[353] and spatially associated with type 1 excitatory dorsal neurons in the spinal cord of adults[355]. Elevated expression of *EFNA5* and *NMNAT2* suggests a neuroprotective mechanism in the spinal cord of ALS-Ox patients[122,283] – in good agreement with ALS-Ox survival. Expression of *UNC13C* has not been previously associated with ALS, however TDP-43 mediated cryptic splicing of paralog *UNC13A* (and *UNC13B*[46]) is detected in a subset of ALS patients in two separate studies[46,212] and uniquely elevated expression in ALS-Ox patients may implicate a related pathological mechanism.

4.4.3 ALS-TD

The defining characteristic of ALS-TD patients in the cortex is the dysregulation of transcription, evident by the overexpression of pseudogenes (*EGLN1P1, ENSG00000213197, HSP90AB4P, KRT8P13, NANOGP4, RPS20P22)*, intronic and antisense transcripts (*AGPAT4-IT1, GATA2-AS1, TUB-AS1, ENSG00000205041, ENSG00000263278, ENSG00000268670,* and *ENSG00000273151*), long non-coding RNA (*LINC00176, LINC00638, LINC01347*), and nonsense-mediated decay mRNA

(*ARHGAP19-SLIT1*, *C1QTNF3-AMACR*, *CHKB-CPT1B*, and *SLX1B-SULT1A4*). Upregulated expression of microRNAs miR24-2, miR219A2, miR3648-1, and *MIRLET7BHG*, relative to the other ALS cortex subtypes, provides additional support for transcriptional and translational dysregulation in ALS-TD patients. miR24-2 has been previously shown to participate in many diseases, including neurodegeneration, serving to regulate cellular proliferation, differentiation, and apoptosis[65]. miR219A2 is known to modulate oligodendrocyte differentiation and remyelination and has been previously reported to be downregulated in the brains of Alzheimer's patients[287,340]. *MIRLET7BHG* (*LET-7B* host gene) is also known to regulate gene expression and has been shown to interact with glial receptor *TLR7* to promote neurodegeneration[193]. Therefore, downregulation of *TLR7* in the ALS-TD subtype may reflect a neuroprotective state. Altered expression of transcription factors *NKX6-2* and *RUNX3*, relative to controls, further emphasizes transcription as a central pathological mechanism in ALS-TD patients.

Similar to the cortex ALS-Ox subtype, downregulation of transcripts encoding extracellular matrix proteins and characteristic expression of some transposable elements is observed. Surprisingly, *TARDBP* (encoding TDP-43) transcription was not a defining feature of ALS-TD patients and expression was relatively conserved across ALS subtypes, with only moderate upregulation observed compared to healthy controls. Transcription of *ADAT3* in ALS-TD patients suggests that the pathological dysregulation of transcription and translation extends to tRNAs[291]. Consistent with the ALS-TD cortex phenotype, elevated expression of many novel mRNA transcripts was observed, with some examples being *ENSG00000258674*, *ENSG00000279233*, *ENSG00000279712*, *ENSG00000228434*, *ENSG00000234913*, and *ENSG00000250397*. Downregulation of *TP63* suggests alterations to *TP53* signaling and an anti-apoptotic phenotypic state in the ALS-TD subtype[357]. This interpretation is further supported by the survival analyses, given ALS-TD patients generally demonstrated a longer median disease duration and lower hazard relative to ALS-Glia patients. Taken together, these results suggest poor control of gene transcription in ALS-TD frontal and motor cortices and provide additional insight into the role of TEs in this subtype.

Extending findings from the cortex, the ALS-TD subtype in the spinal cord similarly implicates transcription, seen in the upregulation of transcripts *DDX18P2*, *ENSG00000185332* (*TMEM105* lncRNA), *ENSG00000250608* (*NUDT16-DT*), *ENSG00000275620* (*FLJ16779* lncRNA), *ENSG00000280087*, *ENSG00000285492*, *LINC01091*, *LINC02977*, *LINC03002*, *MITCH1P1*, *OTOAP1*, *PDE4DIPP7*, *SLC28A1* (pyrimidine nucleoside importer), and *TCF23*. Differential expression was generally weaker in the ALS-TD spinal cord, and likely reflects technical and cell compositional bias associated with stratification of patients using bulk spinal cord transcriptomes. Illustrating these challenges, *APOBR*, *APOC1*, and *FPR3* were found to be upregulated in the ALS-Glia cortex but were most elevated in the ALS-TD spinal cord, suggesting an unclear but possibly relevant role for these genes in mechanistic heterogeneity. Few ALS-TD transcripts show previous links to neurodegeneration, however *NCF2* is reportedly upregulated in ALS in response to neuroinflammation and clearly illustrates a phenotypic spectrum by providing a direct link to increased oxidative stress via ROS production[335]. Although not directly related to neurodegeneration, *NLRP12* has been shown to be an important inhibitor of inflammation and NF-κB signaling[10,245], and ALS-TD upregulation begins to suggest a separate (possibly neuroprotective) inflammatory response when considered in the context of characteristic ALS-Glia gene expression.

4.4.4 Concluding Thoughts

Through differential expression analysis, consistent elevation of marker genes *B4GALT6*, *GABRA1*, *GAD2*, *GLRA3*, *HTR2A*, *PCSK1*, and *SLC17A6* is seen in the postmortem cortex and spinal cord of ALS-Ox patients. These marker genes offer an important foundation for future patient stratification, as it is demonstrated that their association with the ALS-Ox phenotype is not dependent on normalization scale and supervised learning differentiates this phenotype from all others – including FTLD controls – regardless of tissue region or sequencing platform used for characterization. An additional three genes, *MYL9*, *ST6GALNAC2*, *TAGLN* show consistent upregulation in the ALS-Glia cortex and spinal cord relative to the other two subtypes, but are similarly expressed in FTLD controls, suggesting they may have less utility for patient stratification. Moreover, while additional work is needed to link the subtypes to clinical heterogeneity – including rate of progression, genetic risk, exposure, and lifestyle – a clear trend emerges in phenotype-dependent differences in survival with the inflammatory subtypes typically most aggressive and the ALS-Ox subtype less so. Survival differences between ALS-Ox and ALS-TD need further clarification, however marker genes provide a preliminary but important foundation for improved clinical trial design by enabling the stratification of ALS-Ox patients from others, including FTLD mimics.

Among the inflammatory genes associated with ALS, the chitinases (*CHIT1*, *CHI3L1*) have been considered extensively, with many groups demonstrating that elevated expression is linked to ALS progression and disease duration[120,156,307,326,336]. Consistent with these studies, and others[285], elevated expression of another member of the chitinase family, *CHI3L2,* is shown to be uniquely upregulated in ALS-Glia frontal and motor cortices. Similarly, mutations to the *GLT8D1* gene have been previously linked to fALS[74], and elevated expression of the related gene *GLT8D2* was found in the ALS-Glia spinal cord. Cyclophilin A (*PPIA*) has been previously linked to TDP-43 pathology and neurodegeneration in mice[256] and elevated in the CSF of sporadic ALS patients[257]. Elevated expression of Cyclophilin C (*PPIC*) in the spinal cord of ALS-Glia patients lends additional support for the participation of these molecular chaperones in disease pathology.

Briefly considering the results from a similar stratification analysis performed at the protein level using patient CSF[337], seven of the 59 features reliably differentiating fast and slow progressors at four different pairwise comparisons of collection timepoints are also differentially expressed between ALS subtypes at the postmortem transcript level (*CP*, *CFB*, *CD300A*, *FAM19A2*, *KNG1*, *RBP4*, *SERPIND1*). An additional 19 transcripts mirrored findings at the protein level performed by Vu *et al*.[337], which reported a total of 198 proteins stratifying fast and slow progressors for one or more timepoint comparisons, and included: *APOC2*, read through transcript *C1QTNF3-AMACR*, *CBLN4*, *FAM19A1*, *FCGR3A*, *FRZB*, *GDA*, *IGF2*, *ITIH2*, *ITIH3*, *NGFR*, ALS-Ox marker gene *PCSK1*, *QPCT*, *SELL*, *SERPINA3*, *SERPINA5*, *SERPINI1*, *SST*, and *VSTM2A*. Shared features at the protein and transcript level offer additional support for the relevance of these features in ALS patient heterogeneity and indicate additional insight should be gained by considering both system levels in a multi-omics framework. Moreover, elevated abundance of the protein encoded by ALS-Ox marker gene *PCSK1* in the CSF of slow ALS progressors matches findings at the transcript level in the postmortem cortex and spinal cord, and indicates translation of subtype markers genes to living patients should be readily feasible.

Chapter 5

CONCLUSIONS AND FUTURE WORK

This chapter outlines the major significance and contributions of each of the specific aims that were described in the dissertation abstract and introduction. Also included are publications and conference presentations that contributed to the production of this dissertation, and others that occurred concurrently. Challenges and limitations with this work are discussed, and followed by possible future directions for clinical translation. A supplemental analysis is provided in the future directions section which considers the expression of marker genes and subtype-specific transcripts from the cortex in patient-derived induced pluripotent stem cells (iPSCs).

5.1 Significance and Contributions

5.1.1 *Chapter 2*

Clinical and mechanistic heterogeneity observed in ALS has traditionally been poorly understood and has contributed to limited success in clinical trials. To address this gap in understanding, a publicly available ALS cohort (GSE153960) was stratified using postmortem bulk gene expression from the cortex and spinal cord. Three distinct molecular subtypes are recovered, and enrichment defines these subtypes by elevated inflammatory signaling (ALS-Glia), altered synaptic signaling and oxidative stress (ALS-Ox), and dysregulated transcription (ALS-TD). While stratification in the spinal cord was hindered by technical and bulk tissue bias, intra-patient concordance is statistically significant and strengthens the notion that subtype presentation is generally conserved throughout the patient's central nervous system and contributes to patient heterogeneity. Further, concordance analysis shows the lumbar region of the spinal cord is most concordant with the phenotype presented in the frontal and motor cortex. Through bootstrapping, it is demonstrated that ALS patient samples can present in hybrid phenotype states, but show that this interpretation, and overall phenotype presentation, may be driven in part by cell composition bias encountered by bulk tissue RNA-sequencing. Outcomes from this work led to one peer-reviewed publication[99] and a second *in preparation*[98]. Collectively in this chapter, results show that patients can be grouped into molecular subtypes with similar postmortem gene expression patterns that recapitulate many mechanistic phenotypes identified by genetic ALS models[321]. Findings expand upon previous gene expression-based stratification studies[8,237,317], linking additional genes to ALS phenotypic variability and providing additional insight into the role of retrotransposon expression leading to the redefinition of one subtype (ALS-TD).

5.1.2 *Chapter 3*

Previous works show no association between molecular subtype and clinical heterogeneity but are generally limited by sample size[8,237,317]. Our consideration of more than 200 ALS patients provided increased statistical power and enabled the detection of a weak but persistent subtype association with survival. Using the previously described majority agreement approach, ALS-Glia patients repeatedly show a significantly shorter disease duration compared to ALS-Ox patients, with the general trend non-significantly maintained after correcting for other relevant covariates like age, sex, and disease comorbidity through Cox regression. A similar trend towards a later age of onset is observed in ALS-Glia patients and good agreement with survival and age of onset associations in the results from the cortex WGCNA analysis is found. Additional support for findings in this chapter are seen in a recent study which identifies 59 primarily inflammatory proteins which differentiate fast progressors from slow progressors using the ALSFRS-R rate of decline[337]. Outcomes from this work led to one peer-reviewed publication[99] and a second *in preparation*[98]. Collectively in this chapter, an important link is established between clinical variability (survival, age of onset) and ALS mechanistic heterogeneity observed through postmortem gene expression. Moreover, this link implies that the stratification of patients can lead to improved clinical trial outcomes by grouping similarly risked patients into a single cohort and excluding those likely to have an inherently lower or higher risk of death.

5.1.3 *Chapter 4*

The lack of prognostic biomarkers capable of stratifying patients has limited clinical trial success and the design of more personalized and effective therapeutics. To address this need, differential expression was used to identify the consistent elevation of marker genes *B4GALT6*, *GABRA1*, *GAD2*, *GLRA3*, *HTR2A*, *PCSK1*, and *SLC17A6* in the postmortem cortex and spinal cord of ALS-Ox patients. These marker genes offer an important foundation for future patient stratification, as their association with the ALS-Ox phenotype is shown to be independent of the normalization scale. Furthermore, supervised learning differentiates this phenotype from all others, including FTLD controls, regardless of tissue region or sequencing platform used for characterization. An additional three genes, *MYL9*, *ST6GALNAC2*, *TAGLN* show consistent upregulation in the ALS-Glia cortex and spinal cord relative to the other two subtypes, but are similarly expressed in FTLD controls, suggesting they may have less utility for patient stratification. Outcomes from this work led to one peer-reviewed publication[99] and a second *in preparation*[98]. The marker genes established in this chapter provide a preliminary but important foundation for improved clinical trial design by enabling the stratification of ALS-Ox patients from others, including FTLD mimics.

5.1.4 *Contributions*

The following lists peer-reviewed publications and conference oral and poster presentations that contributed to this dissertation. Items donated with * indicate the contribution is unrelated to this dissertation but concurrent.

<u>List of Publications</u>

- **Jarrett Eshima**, Taylor Renee Pennington, Raiyan Choudhury, Jordan M. Garcia, John Fricks, Barbara S. Smith. (2024). Elevated expression of *B4GALT6*, *GABRA1*, *GAD2*, *GLRA3*, *HTR2A*, *PCSK1*, and *SLC17A6* are postmortem markers for the ALS-Ox subtype. [*in preparation*]. Preprint available at: https://www.medrxiv.org/content/10.1101/2024.03.21.24304538v1
- Taylor Renee Pennington, **Jarrett Eshima**, Barbara S. Smith. (2023). Identification of volatile metabolites produced from gut microbial levodopa metabolism using an untargeted metabolomics approach. *BMC Microbiology*, [*in review*].*
- **Jarrett Eshima**, Taylor Pennington, Youssef Abdellatif, Joel F. Lusk, Angela Ponce Olea, Benjamin D. Ambrose, Ethan Marschall, Christopher Miranda, Paula Phan, Christina Aridi, Barbara S. Smith. (2023). An engineered culture vessel and flow system to improve the *in vitro* analysis of volatile organic compounds. *Nature Communications Engineering*, [*in review*].*
- **Jarrett Eshima**, Samantha A. O'Connor, Ethan Marschall, NYGC ALS Consortium, Robert Bowser, Christopher L. Plaisier, Barbara S. Smith. (2023). Molecular subtypes of ALS are associated with differences in patient prognosis. *Nature Communications*, *14*(1), 95.
- Christopher Miranda, Madeleine Howell, Joel Lusk, Ethan Marschall, **Jarrett Eshima**, Trent Anderson, and Barbara S. Smith. (2021). Automated Microscope-Independent Fluorescence Guided Micropipette. *Biomedical Optics Express*, *12*(8), 4689-4699.*

- **Jarrett Eshima**, Trenton J. Davis, Heather D. Bean, John Fricks, Barbara S. Smith. (2020). A Metabolomic Approach for Predicting Diurnal Changes in Cortisol. *Metabolites, 10*(5), 194.*
- **Jarrett Eshima**, Stephanie Ong, Trenton J. Davis, Christopher Miranda, Devika Krishnamurthy, Abigael Nachtsheim, John Stufken, Christopher Plaisier, John Fricks, Heather Bean, and Barbara S. Smith. (2019). Monitoring changes in the healthy female metabolome across the menstrual cycle using GC×GC-TOFMS. *Journal of Chromatography B*, *1121*, 48-57.*
- Joel Lusk, Christopher Miranda, Madeleine Howell, Matthew Chrest, **Jarrett Eshima**, and Barbara S. Smith. (2019). Photoacoustic Flow System for the Detection of Ovarian Circulating Tumor Cells Utilizing Copper Sulfide Nanoparticles. *ACS Biomaterials Science & Engineering*, *5*(3), 1553-1560.*

List of Poster Presentations

- **Jarrett Eshima**, Samantha A. O’Connor, Ethan Marschall, NYGC ALS Consortium, Robert Bowser, Christopher L. Plaisier, Barbara S. Smith “Transcriptomic-based stratification identifies molecular subtypes of ALS with differences in prognosis”, Poster Presentation, Society for Neuroscience. San Diego, California (2022).
- **Jarrett Eshima**, Samantha A. O’Connor, Ethan Marschall, NYGC ALS Consortium, Robert Bowser, Christopher L. Plaisier, Barbara S. Smith “ALS patient stratification identifies pathological subtypes in postmortem cortex

transcriptomes", Poster Presentation, St. Jude Future Fellow Research Conference. Memphis, Tennessee (2022).

- **Jarrett B. Eshima**, Joel F. Lusk, Benjamin D. Ambrose, Taylor Pennington, Paula Phan, Youssef Abdellatif, Barbara S. Smith "Identification of Volatile Biomarkers in a Controlled Microenvironment", Poster Presentation, Arizona Biomedical Research Commission. Phoenix, AZ (2022).*
- **Jarrett B. Eshima**, Aris Mosely, Mireya Herrera, Gilbert Ramos, Lora Nordstrom, Gwenn Levit, and Barbara S. Smith "Applying Multi-omics for the Identification of Early Psychosis Diagnostic Biomarkers", Virtual Poster Presentation, Arizona Biomedical Research Commission. Phoenix, AZ (2021).*
- **Jarrett B. Eshima**, Joel F. Lusk, Benjamin D. Ambrose, Ethan B. Marschall, Esther Sim, Yuka Sugamura, Alison Haymaker, Barbara S. Smith "Biodome: Identification of Ovarian Cancer Volatile Biomarkers in a Controlled Microenvironment", Virtual Poster Presentation, Arizona Biomedical Research Commission. Phoenix, AZ (2021).*
- **Jarrett Eshima**, Esther Sim, Joel F. Lusk, Ethan B. Marshall, Barbara S. Smith "Monitoring Ovarian Cancer Progression Through Metabolic Biomarkers", Poster Presentation in Cancer Technologies Track, Biomedical Engineering Society. Philadelphia, PA (2019).*
- **Jarrett Eshima**, Trenton J. Davis, Heather D. Bean, John Fricks, Barbara S. Smith "A Metabolomic Approach to Non-Invasively Track Changes in Cortisol

for Mental Health Applications”, Poster Presentation in Bioinformatics, Computational and Systems Biology Track, Biomedical Engineering Society. Philadelphia, PA (2019).*

- Stephanie Ong**, Jarrett Eshima**, Christopher Miranda, Trenton Davis, Heather Bean, and Barbara S. Smith “Monitoring Women's Fertility Through Volatile Biomarkers”, Poster Presentation in Diagnostics and Imaging Tract, MCTB Symposium, Tempe, AZ (2017).*
- Vi Nguyen, **Jarrett Eshima**, Samantha Brenna, and Barbara S. Smith “Identifying Volatile Hormone Signatures for Monitoring Female Reproductive Health”, Poster Presentation in Molecular and Cellular Engineering Functional Materials and Sensors Track, Biomedical Engineering Society. Phoenix, AZ (2017).*
- Stephanie Ong**, Jarrett Eshima**, Christopher Miranda, Trenton Davis, Heather Bean, and Barbara S. Smith “Monitoring Women's Fertility Through Volatile Biomarkers”, Poster Presentation in Cellular and Molecular Bioengineering Track, Biomedical Engineering Society. Phoenix, AZ (2017).*
- **Jarrett Eshima**, Stephanie Ong, and Barbara S. Smith “Identification of Volatile Metabolic Biomarkers Correlated to Changes in Hormone Levels”, Poster Presentation in Translational Biomedical Engineering Track for Undergraduates, Biomedical Engineering Society. Phoenix, AZ (2017).*

List of Oral Presentations

- **Jarrett Eshima**, Samantha A. O'Connor, Ethan Marschall, NYGC ALS Consortium, Robert Bowser, Christopher L. Plaisier, Barbara S. Smith "A stratification of ALS patient transcriptomes idnetifiese molecular subtypes associated with differences in patient prognosis", Invited Oral Presentation, ABRC-Flinn Foundation 8th Annual Research Conference. Phoenix, AZ (2023).
- **Jarrett Eshima**, Trenton J. Davis, Heather D. Bean, John Fricks, Barbara S. Smith "Predicting Diurnal Changes in Cortisol using a Metabolic Approach for Mental Health Applications", Invited Oral Presentation, Arizona Biomedical Research Commission. Phoenix, AZ (2020).*

Provisional and Non-Provisional U.S. Patents

- Provisional U.S. Patent Application **No. 62/912,868**, "Minimally Invasive Metal Detector"*
- Non-Provisional U.S. Patent Application **No. 17/715,352**, "Devices and Systems for Non-Destructive Collection and Monitoring of Biological Volatiles"*
- Non-Provisional U.S. Patent Application **No. 18/621,914**, "Biomarkers for Amyotrophic Lateral Sclerosis Stratification"

5.2 Challenges and Limitations

One important limitation of this dissertation stems from the strict consideration of postmortem gene expression from ALS patients. Additional work is necessary before the patient stratification marker genes can be truly leveraged to improve clinical trials and enable the design of more personalized therapeutics, most importantly the translation to living patients. The detection of elevated mRNA or translated protein in CSF and plasma presents challenges due to rapid clearance, generally low concentrations, and strict regulation through the blood brain and blood spinal cord barriers. Additionally, longitudinal studies would be necessary to demonstrate the biomarker tracks with the disease course before early patient stratification can be realized. Offering promise for eventual translation, seven transcripts differentially expressed between ALS-Glia and ALS-Ox samples were among 59 CSF-derived proteins found to stratify fast and slow ALS progressors[337], although *MYL9*, *ST6GALNAC2*, and *TAGLN* were not among the reported features. Furthermore, significant differences in the expression of truncated *STMN2* between ALS-Ox and ALS-TD patients suggests sensitive assays targeting established pathological transcripts may offer secondary use in the stratification of patients, in addition to diagnosis.

Another limitation of these findings stems from the dependency of the bulk tissue expression profile on cell compositional differences, both in the region of the CNS characterized and between the same region in different patients. The consequences of this dependency are discussed in *Chapter 1.3.4* and throughout this dissertation, however looking forward, significant insight should be gained from single cell RNA-sequencing approaches and help clarify the contribution of individual cell types to the overall ALS subtype spectrum.

Significant challenges were encountered during the clustering of spinal cord transcriptomes and required the removal of covariate dependent genes and glial cell marker genes and restricting the number of genes to 5,000. Technical bias remained but the concordance analysis and discovery of marker genes lends strength to the overall validity of the subtypes identified in this dissertation. Importantly, enrichment differences between subtypes were generally weaker in the spinal cord compared to the cortex, although the ALS-Ox phenotype remains clearly distinguishable. Additionally, difficulties were encountered during the development of the Cox multivariate survival models. Through much trial and error, it was determined that the safest use was to avoid the consideration of all tissue regions simultaneously, which would imply the implementation of a patient specific random effect term. This application is supported by the coxme R package, but the repeat measure at death, rather than across the disease course than is the typical implementation, led to the eventual application presented in this dissertation. More experienced survival statisticians may consider using Supplemental Dataset 2 in the second publication associated with this work to replicate findings and address the knowledge limitations encountered here. All scripts used in this dissertation have also been made publicly available at: https://github.com/BSmithLab. More specific to the results, Cox regression shows the effect due to subtype is generally weaker than sex and disease comorbidity and suggests that more work is needed to clarify the influence of molecular variability on survival – although other works generally arrive at similar conclusions relating neuroinflammation and shorter disease duration[24,34,114,198,316,356]. Finally, no subtype marker genes were available for the hybrid subtyping analysis performed in

Chapter 2 and the analysis required the use of subtype-specific eigengenes defined in the WGCNA analysis from Chapter 3. The original analysis from Patel *et al*.[259] had validated subtype marker genes from TCGA and single cell resolution, and our approach leaves room for improvement and additional study.

With the use of publicly available transcriptomic data, a few additional limitations were encountered. Certain influential covariates, like postmortem interval, agonal stage, and hospitals where postmortem samples were obtained, were not available for consideration. Similarly, the use of cerebellum expression as quality controls to address technical and biological covariates were unavailable for all patients. The ability to associate subtype with unavailable patient measures, like rate of functional decline through ALSFRS-R measures, neuroimaging data, or physiological recordings presents additional opportunities to link molecular heterogeneity to clinical variability. Disease mimics other than FTLD were not available for consideration, and marker gene specificity should be validated against mechanistically similar diseases like AD, Parkinson's, SMA, PLS, Progressive supranuclear palsy, and PMA. Lastly, independent verification of RNA expression levels in tissue by RT-qPCR or immunostaining could not be performed, representing an important next step in successful translation.

5.3 Future Directions

Most promisingly, stratification of fast and slow ALS progressors using quantitative proteomics and decline in the ALSFRS-R by Vu *et al*.[337] found ALS-Ox marker gene PCSK1 is elevated in slow progressors when comparing first and last CSF collection timepoints. Agreement between study findings at the transcript and protein level offer a highly favorable foundation for successful translation of stratification (prognostic) biomarkers. Conversely, cell compositional differences throughout the spinal cord may introduce unwanted variability in the quantified protein abundance of stratification biomarkers like PCSK1, although the concordance analysis in *Chapter 2* provides an initial recommendation to leverage lumbar-derived CSF, where feasible – coincidentally advantageous given current health practices. Assuming the expression of proteins encoded by the subtype marker genes remain good predictors of patient variability across the wider sporadic population, developing minimally invasive assays to quantify protein expression in patient CSF should enable the stratification of living patients for improved clinical trial design. To establish concrete guidelines and robust predictive models using CSF expression levels, longitudinal measures should be retrospectively validated against subtype obtained from the postmortem cortex transcriptome. Moreover, other established measures, both molecular (phosphorylated neurofilaments, IL-8[184], MCP-1[184], Transthyretin[269], Cystatin C[269], SERPINA4[337], among others[337]) and clinical (site of symptom onset, genetic mutations, diagnostic delay, etc.), should be considered in the same model in an effort to achieve higher stratification accuracies[343]. Along the same line and discussed previously, future works should examine associations between patient subtype presentation and other responsive clinical measures like ALSFRS-R, neuroimaging, forced vital capacity, hand held dynamometry, motor unit number estimation, and electrical impedance myography. Large scale efforts like Target ALS and Answer ALS[21] provide a crucial resource foundation for these recommended future directions. Putative prognostic biomarkers should be integrated with other relevant measures into a single predictive model with special care given to the analytical and normalization procedures implemented to maximize assay accessibility and minimize batch effects and variability. The predictive models can then be leveraged by clinicians, patient organizations, governments, and the pharmaceutical industry to guide patient enrollment into clinical trials and accelerate the development of phenotype-specific therapeutics or repurposing of existing drugs.

Stemming from cell compositional differences and the difficulty encountered during the stratification of ALS-Glia and ALS-TD subtypes using spinal cord transcriptomes, the use of protein abundance from lumbar-derived CSF to separate these two groups is anticipated to encounter challenges. To help circumvent this, one possible strategy may be to link the ALS-Glia phenotype in the cortex to variable expression profiles in isolated peripheral blood mononuclear cells (PBMCs) from patient blood. Given blood-brain barrier disruption is a common hallmark of neurodegenerative diseases and the migration of these cells to the brain have been observed, these cells may reflect underlying differences in patient phenotype, especially in the context of ALS-Glia neuroinflammation.

Given the mechanisms implicated in the ALS subtypes and known genetic associations with ALS, an opportunity was noticed to consider marker gene expression in patient-derived iPSCs to determine the feasibility in leveraging these cell models for mechanistic research, drug discovery, and drug repurposing. To accomplish this, publicly available RNA-sequencing data was utilized from NCBI Gene Expression Omnibus Accession: GSE210969[82]. This work performs bulk cell line RNA-seq on eight different iPSC lines derived from 3 “healthy donors”, 3 ALS patients with mutations to TDP-43, and 2 ALS patients with mutations to the SOD1 protein. Low sample number restrict this analysis, however differential expression analysis only weakly demonstrates an association between genetic mutation and marker gene expression with TDP-43 iPSCs more closely resembling the ALS-TD subtype.

As seen in Figure 5.1, ALS-Ox marker gene *SLC17A6* is elevated in SOD1 iPSCs compared to TDP43 iPSCs and are similarly expressed in patient frontal and motor cortex tissue. The remaining ALS-Ox marker genes show limited differences in expression, although the expected trend is weakly seen in *B4GALT6* and *GABRA1*. ALS-TD cortex genes *CHKB-CPT1B*, *ENSG00000205041*, *LINC01347*, and miR24-2 are weakly elevated in TDP-43 iPSCs compared to SOD1and match expression levels observed in patient cortical tissue. ALS-Glia marker genes and DAM transcript *TREM2* were generally not uniquely expressed in the motor neurons differentiated from SOD1- and TDP-43 ALS patient iPSCs and suggest more complicated cell models, like organoids, are required to successfully recapitulate the disease phenotype. More broadly given these findings, SOD1 iPSCs are hypothesized to serve as a more appropriate model system for the identification of ALS-Ox specific drug targets and therapies, while iPSCs harboring pathological mutations to the TDP-43 protein will better serve the discovery of ALS-TD drug targets and therapies. Follow-up analyses with larger patient numbers and single cell resolution should clarify the preliminary associations observed here, however other groups have arrived at similar conclusions[223].

Outside of the development of more phenotype specific cell models, future works may also look to consider ALS subtype presentation in other CNS tissue regions like the choroid plexus and cerebellum to better understand associations with the blood brain barrier and better address expression covariates. While disease duration was used as the clinical measure to stratify the cohorts, consideration of other relevant clinical measures like ALSFRS-R may provide additional insight and strengthen existing associations. Independent validation of elevated marker gene expression should be performed in other patient cohorts to demonstrate that the effects observed in this work are not dependent on the myriad of covariates known to influence postmortem bulk tissue gene expression. To the contrary and lending strength to the findings in this work, 85% agreement is reported in the subtypes assigned to 140 shared postmortem cortex samples independently considered by Tam *et al*.[317]. Eventually, successful translation of marker genes and stratification of living patient cohorts will enable the design of more effective clinical trials, guide selection of maximal response measures in the clinic, spur the development of more personalized therapeutics, and better inform the patient and entire caretaking team of probable event timelines.

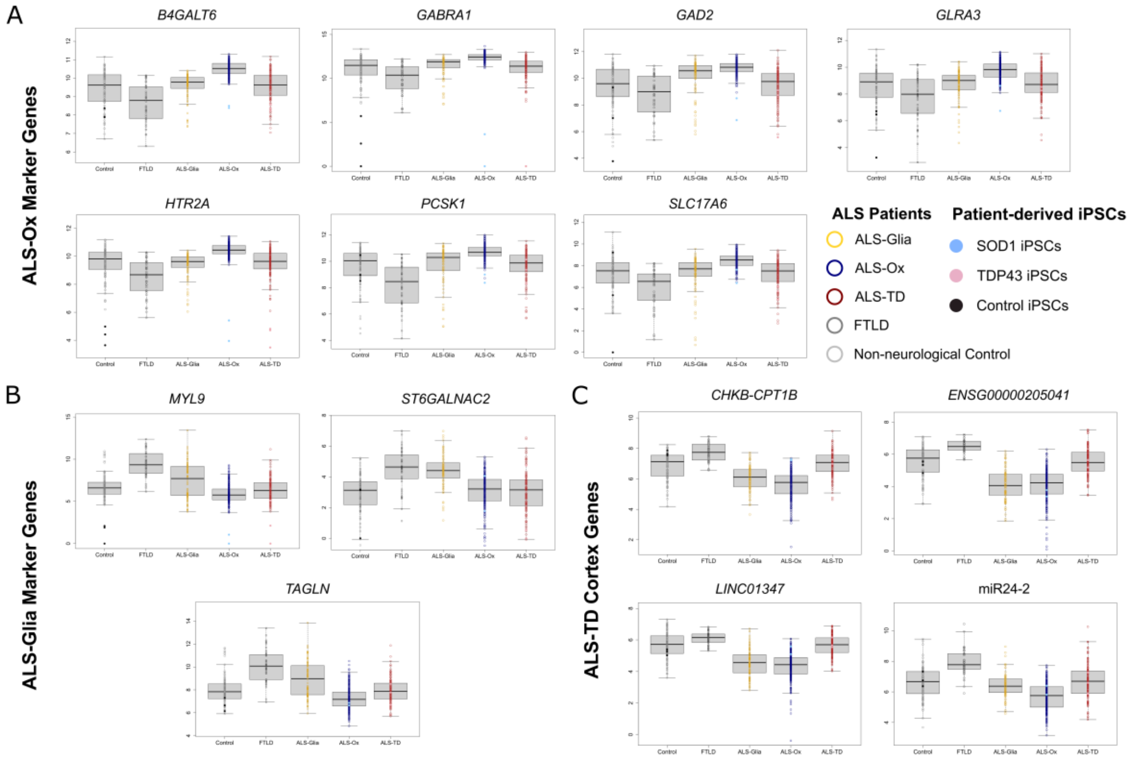


Figure 5.1: ALS patient-derived iPSC expression of subtype marker genes. ALS patient expression from the cortex are plotted as open points and iPSC expression is overlaid as solid points. SOD1 iPSCs are plotted in the ALS-Ox group while TDP-43 iPSCs are plotted in the ALS-TD group. (A) ALS-Ox marker genes *B4GALT6*, *GABRA1*, *GAD2*, *GLRA3*, *HTR2A*, *PCSK1*, and *SLC17A6*. (B) ALS-Glia marker genes *MYL9*, *ST6GALNAC2*, and *TAGLN* and disease associated microglia marker *TREM2*. (C) Four representative transcripts, *CHKB-CPT1B*, *ENSG00000205041*, *LINC01347*, and miR24-2 differentially expressed in the ALS-TD postmortem cortex. All expression is presented on the DESeq2 normalized median-of-ratios scale following a $\log_2$ transformation.

REFERENCES

[1] Abadi, M. et al. {TensorFlow}: a system for {Large-Scale} machine learning. In12th USENIX symposium on operating systems design and implementation (OSDI 16) 265-283 (2016).

[2] Abel, O., Powell, J.F., Andersen, P.M., Al-Chalabi, A., ALSoD: A user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. *Human mutation* **33**, 1345-1351 (2012).

[3] Amaral, P.P., Mattick, J.S., Noncoding RNA in development. *Mammalian genome* **19**, 454-492 (2008).

[4] Anders, S., Huber, W., Differential expression analysis for sequence count data. *Nature Precedings*, 1-1 (2010).

[5] Andersen, P.K., Gill, R.D., Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100-1120 (1982).

[6] Andrews, Z.B., Diano, S., Horvath, T.L., Mitochondrial uncoupling proteins in the CNS: in support of function and survival. *Nature Reviews Neuroscience* **6**, 829-840 (2005).

[7] Arnold, E.S. et al. ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *Proceedings of the National Academy of Sciences* **110**, E736-E745 (2013).

[8] Aronica, E. et al. Molecular classification of amyotrophic lateral sclerosis by unsupervised clustering of gene expression in motor cortex. *Neurobiology of disease* **74**, 359-376 (2015).

[9] Artenstein, A.W., Opal, S.M., Proprotein convertases in health and disease. *New England Journal of Medicine* **365**, 2507-2518 (2011).

[10] Arthur, J.C., Lich, J.D., Aziz, R.K., Kotb, M., Ting, J.P., Heat shock protein 90 associates with monarch-1 and regulates its ability to promote degradation of NF-κB-inducing kinase. *The Journal of Immunology* **179**, 6291-6296 (2007).

[11] Ash, P.E. et al. Unconventional translation of C9ORF72 GGGGCC expansion generates insoluble polypeptides specific to c9FTD/ALS. *Neuron* **77**, 639-646 (2013).

[12] Austin, P.C., A tutorial on multilevel survival analysis: methods, models and applications. *International Statistical Review* **85**, 185-203 (2017).

[13] Babaian, A., Mager, D.L., Endogenous retroviral promoter exaptation in human cancer. *Mobile DNA* **7**, 1-21 (2016).

[14] Bahia, E.l. et al. Complement activation at the motor end-plates in amyotrophic lateral sclerosis. *Journal of neuroinflammation* **13**, 1-12 (2016).

[15] Balendra, R., Isaacs, A.M., C9orf72-mediated ALS and FTD: multiple pathways to disease. *Nature Reviews Neurology* **14**, 544-558 (2018).

[16] Bao, W., Kojima, K.K., Kohany, O., Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna* **6**, 1-6 (2015).

[17] Barabási, A.L., Bonabeau, E., Scale-free networks. *Scientific american* **288**, 60-69 (2003).

[18] Barry, G. et al. The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. *Molecular psychiatry* **19**, 486-494 (2014).

[19] Baruzzo, G. et al. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature methods* **14**, 135-139 (2017).

[20] Baughn, M.W. et al. Mechanism of STMN2 cryptic splice-polyadenylation and its correction for TDP-43 proteinopathies. *Science* **379**, 1140-1149 (2023).

[21] Baxi, E.G. et al. Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical and multi-omics data from induced pluripotent cell lines. *Nature neuroscience* **25**, 226-237 (2022).

[22] Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* **37**, 38-44 (2019).

[23] Beckman, J.S., Carson, M., Smith, C.D., Koppenol, W.H., ALS, SOD and peroxynitrite. *Nature* **364**, 584-584 (1993).

[24] Beers, D.R. et al. Wild-type microglia extend survival in PU. 1 knockout mice with familial amyotrophic lateral sclerosis. *Proceedings of the National Academy of Sciences* **103**, 16021-16026 (2006).

[25] Bellera, C.A. et al. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC medical research methodology* **10**, 1-12 (2010).

[26] Benjamini, Y., Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289-300 (1995).

[27] Benson, D.A. et al. GenBank. *Nucleic acids research* **41**, D36-D42 (2012).

[28] Bhan, A., Soleimani, M., Mandal, S.S., Long noncoding RNA and cancer: a new paradigm. *Cancer research* **77**, 3965-3981 (2017).

[29] Bido, S. et al. Microglia-specific overexpression of α-synuclein leads to severe dopaminergic neurodegeneration by phagocytic exhaustion and oxidative toxicity. *Nature communications* **12**, 6237 (2021).

[30] Black, D.L., Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103**, 367-370 (2000).

[31] Blalock, E.M. et al. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences* **101**, 2173-2178 (2004).

[32] Boche, D., Perry, V.H., Nicoll, J.A., Activation patterns of microglia and their identification in the human brain. *Neuropathology and applied neurobiology* **39**, 3-18 (2013).

[33] Boeke, J.D., Garfinkel, D.J., Styles, C.A., Fink, G.R., Ty elements transpose through an RNA intermediate. *Cell* **40**, 491-500 (1985).

[34] Boillée, S. et al. Onset and progression in inherited ALS determined by motor neurons and microglia. *Science* **312**, 1389-1392 (2006).

[35] Bolger, A.M., Lohse, M., Usadel, B., Trimmomatic: a flexible trimmer for Illumina sequence data. B*ioinformatics* **30**, 2114-2120 (2014).

[36] Bonferroni, C., Statistical class theory and probability calculus. *Publications of the Higher Institute of Economic and Commercial Sciences of Florence* **8**, 3-62 (1936).

[37] Bonifacino, T. et al. Nearly 30 years of animal models to study amyotrophic lateral sclerosis: a historical overview and future perspectives. *International journal of molecular sciences* **22**, 12236 (2021).

[38] Boratyn, G.M. et al. BLAST: a more efficient report with usability improvements. *Nucleic acids research* **41**, W29-W33 (2013).

[39] Bourque, G. et al. Ten things you should know about transposable elements. *Genome biology* **19**, 1-12 (2018).

[40] Bowser, R., Turner, M.R., Shefner, J., Biomarkers in amyotrophic lateral sclerosis: opportunities and limitations. *Nature Reviews Neurology* **7**, 631 (2011).

[41] Bretheau, F. et al. The alarmin interleukin-1α triggers secondary degeneration through reactive astrocytes and endothelium after spinal cord injury. *Nature Communications* **13**, 5786 (2022).

[42] Brettschneider, J. et al. TDP-43 pathology and neuronal loss in amyotrophic lateral sclerosis spinal cord. *Acta neuropathologica* **128**, 423-437 (2014).

[43] Brettschneider, J., Petzold, A., Süssmuth, S.D., Ludolph, A.C., Tumani, H., Axonal damage markers in cerebrospinal fluid are increased in ALS. *Neurology* **66**, 852-856 (2006).

[44] Bright, F. et al. Neuroinflammation in frontotemporal dementia. *Nature Reviews Neurology* **15**, 540-555 (2019).

[45] Brooks, B.R., Miller, R.G., Swash, M., Munsat, T.L., El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotrophic lateral sclerosis and other motor neuron disorders* **1**, 293-299 (2000).

[46] Brown, A.L. et al. TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature* **603**, 131-137 (2022).

[47] Bruijn, L.I. et al. Aggregation and motor neuron toxicity of an ALS-linked SOD1 mutant independent from wild-type SOD1. *Science* **281**, 1851-1854 (1998).

[48] Bruijn, L.I. et al. ALS-linked SOD1 mutant G85R mediates damage to astrocytes and promotes rapidly progressive disease with SOD1-containing inclusions. *Neuron* **18**, 327-338 (1997).

[49] Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P., Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* **101**, 4164-4169 (2004).

[50] Burrell, J.R., Kiernan, M.C., Vucic, S., Hodges, J.R., Motor neuron dysfunction in frontotemporal dementia. *Brain* **134**, 2582-2594 (2011).

[51] Butchi, N.B., Du, M., Peterson, K.E., Interactions between TLR7 and TLR9 agonists and receptors regulate innate immune responses by astrocytes and microglia. *Glia* **58**, 650-664 (2010).

[52] Byron, S.A., Van Keuren-Jensen, K.R., Engelthaler, D.M., Carpten, J.D., Craig, D.W., Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics* **17**, 257-271 (2016).

[53] Cahoy, J.D. et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *Journal of Neuroscience* **28**, 264-278 (2008).

[54] Cairns, N.J. et al. TDP-43 in familial and sporadic frontotemporal lobar degeneration with ubiquitin inclusions. *The American journal of pathology* **171**, 227-240 (2007).

[55] Calvo, A. et al. Factors predicting survival in ALS: a multicenter Italian study. *Journal of neurology* **264**, 54-63 (2017).

[56] Cardona, A.E. et al. Control of microglial neurotoxicity by the fractalkine receptor. *Nature neuroscience* **9**, 917-924 (2006).

[57] Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research* **44**, D471-D480 (2016).

[58] Cedarbaum, J.M. et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the neurological sciences* **169**, 13-21 (1999).

[59] Chen, E.Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics* **14**, 1-14 (2013).

[60] Chen, H. et al. Modeling ALS with iPSCs reveals that mutant SOD1 misregulates neurofilament balance in motor neurons. *Cell stem cell* **14**, 796-809 (2014).

[61] Chen, Y., Kim, J.K., Hirning, A.J., Josić, K., Bennett, M.R., Emergent genetic oscillations in a synthetic microbial consortium. *Science* **349**, 986-989 (2015).

[62] Chen, Y.Z. et al. DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *The American Journal of Human Genetics* **74**, 1128-1135 (2004).

[63] Cheng, J. et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149-1154 (2005).

[64] Chen-Plotkin, A.S., Lee, V.M., Trojanowski, J.Q., TAR DNA-binding protein 43 in neurodegenerative disease. *Nature Reviews Neurology* **6**, 211-220 (2010).

[65] Chhabra, R., Dubey, R., Saini, N., Cooperative and individualistic functions of the microRNAs in the miR-23a~ 27a~ 24-2 cluster and its implication in human diseases. *Molecular cancer* **9**, 1-16 (2010).

[66] Chiò, A. et al. ALS phenotype is influenced by age, sex, and genetics: a population-based study. *Neurology* **94**, e802-e810 (2020).

[67] Chiò, A., Calvo, A., Moglia, C., Mazzini, L., Mora, G., Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study. *Journal of Neurology, Neurosurgery & Psychiatry* **82**, 740-746 (2011).

[68] Chiu, I.M. et al. A neurodegeneration-specific gene-expression signature of acutely isolated microglia from an amyotrophic lateral sclerosis mouse model. *Cell reports* **4**, 385-401 (2013).

[69] Chuong, E.B., Elde, N.C., Feschotte, C., Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics* **18**, 71-86 (2017).

[70] Cleveland, D.W., Laing, N., Hurse, P.V., Brown Jr., R.H., Toxic mutants in Charcot's sclerosis. *Nature* **378**, 342-343 (1995).

[71] Collins, M.A., An, J., Hood, B.L., Conrads, T.P., Bowser, R.P., Label-free LC–MS/MS proteomic analysis of cerebrospinal fluid identifies protein/pathway alterations and candidate biomarkers for amyotrophic lateral sclerosis. *Journal of proteome research* **14**, 4486-4501 (2015).

[72] Collisson, E.A., Bailey, P., Chang, D.K., Biankin, A.V., Molecular subtypes of pancreatic cancer. *Nature reviews Gastroenterology & hepatology* **16**, 207-220 (2019).

[73] Cooper-Knock, J. et al. Clinico-pathological features in amyotrophic lateral sclerosis with expansions in C9ORF72. *Brain* **135**, 751-764 (2012).

[74] Cooper-Knock, J. et al. Mutations in the glycosyltransferase domain of GLT8D1 are associated with familial amyotrophic lateral sclerosis. *Cell reports* **26**, 2298-2306 (2019).

[75] Cooper-Knock, J. et al. Sequestration of multiple RNA recognition motif-containing proteins by C9orf72 repeat expansions. *Brain* **137**, 2040-2051 (2014).

[76] Corrà, F., Agnoletto, C., Minotti, L., Baldassari, F., Volinia, S., The network of non-coding RNAs in cancer drug resistance. *Frontiers in oncology* **8**, 327 (2018).

[77] Costa, V., Aprile, M., Esposito, R., Ciccodicola, A., RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *European Journal of Human Genetics* **21**, 134-142 (2013).

[78] Costa-Silva, J., Domingues, D., Lopes, F.M., RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one* **12**, e0190152 (2017).

[79] Cowling, V.H., Regulation of mRNA cap methylation. *Biochemical Journal* **425**, 295-302 (2010).

[80] Cox, D.R., Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187-202 (1972).

[81] Cudkowicz, M.E., McKenna-Yasek, D., Chen, C., Hedley-Whyte, E.T., Brown Jr., R.H., Limited corticospinal tract involvement in amyotrophic lateral sclerosis subjects with the A4V mutation in the copper/zinc superoxide dismutase gene. *Annals of neurology* **43**, 703-710 (1998).

[82] Dash, B.P., Freischmidt, A., Weishaupt, J.H., Hermann, A., Downstream effects of mutations in SOD1 and TARDBP converge on gene expression impairment in patient-derived motor neurons. *International Journal of Molecular Sciences* **23**, 9652 (2022).

[83] Deczkowska, A. et al. Disease-associated microglia: a universal immune sensor of neurodegeneration. *Cell* **173**, 1073-1081 (2018).

[84] Devenney, E., Vucic, S., Hodges, J.R., Kiernan, M.C., Motor neuron disease-frontotemporal dementia: a clinical continuum. *Expert review of neurotherapeutics* **15**, 509-522 (2015).

[85] Dillies, M.A. et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics* **14**, 671-683 (2013).

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

[87] Dobrowolny, G. et al. A longitudinal study defined circulating microRNAs as reliable biomarkers for disease prognosis and progression in ALS human patients. *Cell death discovery* **7**, 4 (2021).

[88] Doens, D., Fernández, P.L., Microglia receptors and their implications in the response to amyloid β for Alzheimer's disease pathogenesis. *Journal of neuroinflammation* **11**, 1-14 (2014).

[89] Doi, Y., Atsuta, N., Sobue, G., Morita, M., Nakano, I., Prevalence and incidence of amyotrophic lateral sclerosis in Japan. *Journal of epidemiology* **24**, 494-499 (2014).

[90] Donnelly, C.J. et al. RNA toxicity from the ALS/FTD C9ORF72 expansion is mitigated by antisense intervention. *Neuron* **80**, 415-428 (2013).

[91] Dubnau, D., Losick, R., Bistability in bacteria. *Molecular microbiology* **61**, 564-572 (2006).

[92] Dunn, O.J., Multiple comparisons among means. *Journal of the American statistical association* **56**, 52-64 (1961).

[93] Elamin, M. et al. Executive dysfunction is a negative prognostic indicator in patients with ALS without dementia. *Neurology* **76**, 1263-1269 (2011).

[94] Elbarbary, R.A., Lucas, B.A., Maquat, L.E., Retrotransposons as regulators of gene expression. *Science* **351**, aac7247 (2016).

[95] ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).

[96] Engström, P.G. et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods* **10**, 1185-1191 (2013).

[97] Erwin, J.A., Marchetto, M.C., Gage, F.H., Mobile DNA elements in the generation of diversity and complexity in the brain. *Nature Reviews Neuroscience* **15**, 497-506 (2014).

[98] Eshima, J. et al. Elevated expression of B4GALT6, GABRA1, GAD2, GLRA3, HTR2A, PCSK1, and SLC17A6 are postmortem markers for the ALS-Ox subtype. (2024). Preprint available at: https://doi.org/10.1101/2024.03.21.24304538

[99] Eshima, J. et al. Molecular subtypes of ALS are associated with differences in patient prognosis. *Nature communications* **14**, 95 (2023).

[100] Estes, P.S. et al. Motor neurons and glia exhibit specific individualized responses to TDP-43 expression in a Drosophila model of amyotrophic lateral sclerosis. *Disease models & mechanisms* **6**, 721-733 (2013).

[101] Estes, P.S. et al. Wild-type and A315T mutant TDP-43 exert differential neurotoxicity in a Drosophila model of ALS. *Human molecular genetics* **20**, 2308-2321 (2011).

[102] Ewing, A.D., Kazazian, H.H., High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome research* **20**, 1262-1270 (2010).

[103] Fabbro, S., Seeds, N.W., Plasminogen activator activity is inhibited while neuroserpin is up-regulated in the Alzheimer disease brain. *Journal of neurochemistry* **109**, 303-315 (2009).

[104] Faghihi, M.A. et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. *Nature medicine* **14**, 723-730 (2008).

[105] Feneberg, E., Gray, E., Ansorge, O., Talbot, K., Turner, M.R., Towards a TDP-43-based biomarker for ALS and FTLD. *Molecular neurobiology* **55**, 7789-7801 (2018).

[106] Ferrante, R.J. et al. Evidence of increased oxidative damage in both sporadic and familial amyotrophic lateral sclerosis. *Journal of neurochemistry* **69**, 2064-2074 (1997).

[107] Ferreira, P.G. et al. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nature communications* **9**, 490 (2018).

[108] Ferrer, J., Dimitrova, N., Transcription regulation by long non-coding RNAs: mechanisms and disease relevance. *Nature Reviews Molecular Cell Biology*, 1-20 (2024).

[109] Flynn, G., Maru, S., Loughlin, J., Romero, I.A., Male, D., Regulation of chemokine receptor expression in human microglia and astrocytes. *Journal of neuroimmunology* **136**, 84-93 (2003).

[110] Forgy, E.W., Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics* **21**, 768-769 (1965).

[111] Foster, I., Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing* **15**, 70-73 (2011).

[112] Freibaum, B.D., Chitta, R.K., High, A.A., Taylor, J.P., Global analysis of TDP-43 interacting proteins reveals strong association with RNA splicing and translation machinery. *Journal of proteome research* **9**, 1104-1120 (2010).

[113] Frottin, F., Pérez-Berlanga, M., Hartl, F.U., Hipp, M.S., Multiple pathways of toxicity induced by C9orf72 dipeptide repeat aggregates and G4C2 RNA in a cellular model. *Elife* **10**, e62718 (2021).

[114] Fuhrmann, M. et al. Microglial Cx3cr1 knockout prevents neuron loss in a mouse model of Alzheimer's disease. *Nature neuroscience* **13**, 411-413 (2010).

[115] Ganesalingam, J. et al. Combination of neurofilament heavy chain and complement C3 as CSF biomarkers for ALS. *Journal of neurochemistry* **117**, 528-537 (2011).

[116] Gaujoux, R., Seoighe, C., A flexible R package for nonnegative matrix factorization. *BMC bioinformatics* **11**, 1-9 (2010).

[117] Gendron, T.F. et al. Antisense transcripts of the expanded C9ORF72 hexanucleotide repeat form nuclear RNA foci and undergo repeat-associated non-ATG translation in c9FTD/ALS. *Acta neuropathologica* **126**, 829-844 (2013).

[118] Gene Ontology Consortium, Gene ontology consortium: going forward. *Nucleic acids research* **43**, D1049-D1056 (2015).

[119] Ghadge, G.D. et al. Glutamate carboxypeptidase II inhibition protects motor neurons from death in familial amyotrophic lateral sclerosis models. *Proceedings of the National Academy of Sciences* **100**, 9554-9559 (2003).

[120] Gille, B. et al. Inflammatory markers in cerebrospinal fluid: independent prognostic biomarkers in amyotrophic lateral sclerosis?. *Journal of Neurology, Neurosurgery & Psychiatry* **90**, 1338-1346 (2019).

[121] Gillespie, M. et al. The reactome pathway knowledgebase 2022. *Nucleic acids research* **50**, D687-D692 (2022).

[122] Gilley, J., Coleman, M.P., Endogenous Nmnat2 is an essential survival factor for maintenance of healthy axons. *PLoS biology* **8**, e1000300 (2010).

[123] Glass, K., Girvan, M., Annotation enrichment analysis: an alternative method for evaluating the functional properties of gene sets. *Scientific reports* **4**, 4191 (2014).

[124] Goel, M.K., Khanna, P., Kishore, J., Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research* **1**, 274 (2010).

[125] Gorbunova, V. et al. The role of retrotransposable elements in ageing and age-associated diseases. *Nature* **596**, 43-53 (2021).

[126] Grabowski, P.J., Black, D.L., Alternative RNA splicing in the nervous system. *Progress in neurobiology* **65**, 289-308 (2001).

[127] Grambsch, P.M., Therneau, T.M., Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515-526 (1994).

[128] GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).

[129] Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nature medicine* **21**, 1350-1356 (2015).

[130] Gulen, M.F. et al. cGAS–STING drives ageing-related inflammation and neurodegeneration. *Nature* **620**, 374-380 (2023).

[131] Guo, C., Sun, L., Chen, X., Zhang, D., Oxidative stress, mitochondrial damage and neurodegenerative diseases. *Neural regeneration research* **8**, 2003 (2013).

[132] Guo, Q. et al. In situ structure of neuronal C9orf72 poly-GA aggregates reveals proteasome recruitment. *Cell* **172**, 696-705 (2018).

[133] Gurney, M.E. et al. Motor neuron degeneration in mice that express a human Cu, Zn superoxide dismutase mutation. *Science* **264**, 1772-1775 (1994).

[134] Gurney, M.E., Fleck, T.J., Himes, C.S., Hall, E.D., Riluzole preserves motor function in a transgenic model of familial amyotrophic lateral sclerosis. *Neurology* **50**, 62-66 (1998).

[135] Habib, N. et al. Disease-associated astrocytes in Alzheimer's disease and aging. *Nature neuroscience* **23**, 701-706 (2020).

[136] Han, X. et al. The lncRNA Hand2os1/Uph locus orchestrates heart development through regulation of precise expression of Hand2. *Development* **146**, dev176198 (2019).

[137] Hanson, K.A., Kim, S.H., Wassarman, D.A., Tibbetts, R.S., Ubiquilin modifies TDP-43 toxicity in a Drosophila model of amyotrophic lateral sclerosis (ALS). *Journal of Biological Chemistry* **285**, 11068-11072 (2010).

[138] Hardiman, O., Van Den Berg, L.H., Kiernan, M.C., Clinical diagnosis and management of amyotrophic lateral sclerosis. *Nature reviews neurology* **7**, 639-649 (2011).

[139] Harrison, P.W. et al. Ensembl 2024. *Nucleic Acids Research* **52**, D891-D899 (2024).

[140] Hayashi, Y., Homma, K., Ichijo, H., SOD1 in neurotoxicity and its controversial roles in SOD1 mutation-negative ALS. *Advances in biological regulation* **60**, 95-104 (2016).

[141] Hayes, G.M., Woodroofe, M.N., Cuzner, M.L., Microglia are the major cell type expressing MHC class II in human white matter. *Journal of the neurological sciences* **80**, 25-37 (1987).

[142] Hayes, L.R., Duan, L., Bowen, K., Kalab, P., Rothstein, J.D., C9orf72 arginine-rich dipeptide repeat proteins disrupt karyopherin-mediated nuclear import. *Elife* **9**, e51685 (2020).

[143] Hedges, D.J., Deininger, P., Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **616**, 46-59 (2007).

[144] Heras-Sandoval, D., Pérez-Rojas, J.M., Hernández-Damián, J., Pedraza-Chaverri, J., The role of PI3K/AKT/mTOR pathway in the modulation of autophagy and the clearance of protein aggregates in neurodegeneration. *Cellular signalling* **26**, 2694-2701 (2014).

[145] Hickman, A.B., Dyda, F., Mechanisms of DNA transposition. *Mobile DNA III*, 529-553 (2015).

[146] Hiltunen, M. et al. Ubiquilin 1 modulates amyloid precursor protein trafficking and Aβ secretion. *Journal of Biological Chemistry* **281**, 32240-32253 (2006).

[147] Hjerpe, R. et al. UBQLN2 mediates autophagy-independent protein aggregate clearance by the proteasome. *Cell* **166**, 935-949 (2016).

[148] Ho, Y.J. et al. Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. *Genome research* **28**, 1353-1363 (2018).

[149] Hochberg, Y., A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800-802 (1988).

[150] Holm, S., A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70 (1979).

[151] Humphrey, J. et al. Integrative transcriptomic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes. *Nature Neuroscience* **26**, 150-162 (2023).

[152] Humphrey, J., Emmett, W., Fratta, P., Isaacs, A.M., Plagnol, V., Quantitative analysis of cryptic splicing associated with TDP-43 depletion. *BMC medical genomics* **10**, 1-17 (2017).

[153] Igaz, L.M. et al. Expression of TDP-43 C-terminal fragments in vitro recapitulates pathological features of TDP-43 proteinopathies. *Journal of Biological Chemistry* **284**, 8516-8524 (2009).

[154] Iguchi, Y. et al. Exosome secretion is a key pathway for clearance of pathological TDP-43. *Brain* **139**, 3187-3201 (2016).

[155] Ilieva, H., Polymenidou, M., Cleveland, D.W., Non–cell autonomous toxicity in neurodegenerative disorders: ALS and beyond. *Journal of Cell Biology* **187**, 761-772 (2009).

[156] Illán-Gala, I. et al. CSF sAPPβ, YKL-40, and NfL along the ALS-FTD spectrum. *Neurology* **91**, e1619-e1628 (2018).

[157] Ito, H. et al. Treatment with edaravone, initiated at symptom onset, slows motor decline and decreases SOD1 deposition in ALS mice. *Experimental neurology* **213**, 448-455 (2008).

[158] Iwase, K. et al. Induction of endothelial nitric-oxide synthase in rat brain astrocytes by systemic lipopolysaccharide treatment. *Journal of Biological Chemistry* **275**, 11929-11933 (2000).

[159] Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic acids research* **48**, D498-D503 (2020).

[160] Jin, Y., Tam, O.H., Paniagua, E., Hammell, M., TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593-3599 (2015).

[161] Jong, M.C., Hofker, M.H., Havekes, L.M., Role of ApoCs in lipoprotein metabolism: functional differences between ApoC1, ApoC2, and ApoC3. *Arteriosclerosis, thrombosis, and vascular biology* **19**, 472-484 (1999).

[162] Jurka, J., Walichiewicz, J., Milosavljevic, A., Prototypic sequences for human repetitive DNA. *Journal of molecular evolution* **35**, 286-291 (1992).

[163] Kaikkonen, M.U., Lam, M.T., Glass, C.K., Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovascular research* **90**, 430-440 (2011).

[164] Kamath, P.S. et al. A model to predict survival in patients with end–stage liver disease. *Hepatology* **33**, 464-470 (2001).

[165] Kanehisa, M., Goto, S., KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).

[166] Kanekura, K. et al. Poly-dipeptides encoded by the C9ORF72 repeats block global protein translation. *Human molecular genetics* **25**, 1803-1813 (2016).

[167] Kaplan, E.L., Meier, P., Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53**, 457-481 (1958).

[168] Karolchik, D. et al. The UCSC genome browser database. *Nucleic acids research* **31**, 51-54 (2003).

[169] Kassambara, A., Kosinski, M., Biecek, P., Fabian, S., survminer: Drawing Survival Curves using ‘ggplot2’. (2017).

[170] Kaur, S.J., McKeown, S.R., Rashid, S., Mutant SOD1 mediated pathogenesis of amyotrophic lateral sclerosis. *Gene* **577**, 109-118 (2016).

[171] Keren-Shaul, H. et al. A unique microglia type associated with restricting development of Alzheimer’s disease. *Cell* **169**, 1276-1290 (2017).

[172] Khoury, J.E. et al. Scavenger receptor-mediated adhesion of microglia to β-amyloid fibrils. *Nature* **382**, 716-719 (1996).

[173] Kiernan, M.C. et al. Amyotrophic lateral sclerosis. *The lancet* **377**, 942-955 (2011).

[174] Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, 1-13 (2013).

[175] Kim, H., Park, H., Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **23**, 1495-1502 (2007).

[176] Kiskinis, E. et al. Pathways disrupted in human ALS motor neurons identified through genetic correction of mutant SOD1. *Cell stem cell* **14**, 781-795 (2014).

[177] Klegeris, A., McGeer, P.L., Cyclooxygenase and 5-lipoxygenase inhibitors protect against mononuclear phagocyte neurotoxicity. *Neurobiology of aging* **23**, 787-794 (2002).

[178] Kleinbaum, D.G., Klein, M., Survival analysis a self-learning text. (Springer, New York, 1996).

[179] Klim, J.R. et al. ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. *Nature neuroscience* **22**, 167-179 (2019).

[180] Kodama, Y. et al. DNA data bank of Japan: 30th anniversary. *Nucleic acids research* **46**, D30-D35 (2018).

[181] Komine, O. et al. Genetic background variation impacts microglial heterogeneity and disease progression in amyotrophic lateral sclerosis model mice. *Iscience* **27**, (2024).

[182] Kristensen, V.N. et al. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer* **14**, 299-313 (2014).

[183] Krug, L. et al. Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. *PLoS genetics* **13**, e1006635 (2017).

[184] Kuhle, J. et al. Increased levels of inflammatory chemokines in amyotrophic lateral sclerosis. *European journal of neurology* **16**, 771-774 (2009).

[185] Kuleshov, M.V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90-W97 (2016).

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

[187] Kwong, L.K. et al. Novel monoclonal antibodies to normal and pathologically altered human TDP-43 proteins. *Acta neuropathologica communications* **2**, 1-9 (2014).

[188] Langfelder, P., Horvath, S., WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 1-13 (2008).

[189] Langmead, B., Salzberg, S.L., Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).

[190] Le, Y., Murphy, P.M., Wang, J.M., Formyl-peptide receptors revisited. *Trends in immunology* **23**, 541-548 (2002).

[191] Lederer, C.W., Torrisi, A., Pantelidou, M., Santama, N., Cavallaro, S., Pathways and genes differentially expressed in the motor cortex of patients with sporadic amyotrophic lateral sclerosis. *BMC genomics* **8**, 1-26 (2007).

[192] Lee, D.D., Seung, H.S., Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791 (1999).

[193] Lehmann, S.M. et al. An unconventional role for miRNA: let-7 activates Toll-like receptor 7 and causes neurodegeneration. *Nature neuroscience* **15**, 827-835 (2012).

[194] Leinonen, R. et al. The European nucleotide archive. *Nucleic acids research* **39**, D28-D31 (2010).

[195] Leung, K.M., Elashoff, R.M., Afifi, A.A., Censoring issues in survival analysis. *Annual review of public health* **18**, 83-104 (1997).

[196] Li, B., Dewey, C.N., RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 1-16 (2011).

[197] Li, W., Jin, Y., Prazak, L., Hammell, M., Dubnau, J., Transposable elements in TDP-43-mediated neurodegenerative disorders. *PLoS one* **7**, e44099 (2012).

[198] Liddelow, S.A. et al. Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481-487 (2017).

[199] Lindberg, M.J., Tibell, L., Oliveberg, M., Common denominator of Cu/Zn superoxide dismutase mutants associated with amyotrophic lateral sclerosis: decreased stability of the apo state. *Proceedings of the National Academy of Sciences* **99**, 16607-16612 (2002).

[200] Lindquist, S.G. et al. Corticobasal and ataxia syndromes widen the spectrum of C9ORF72 hexanucleotide expansion disease. *Clinical genetics* **83**, 279-283 (2013).

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

[202] Ling, S.C., Polymenidou, M., Cleveland, D.W., Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. *Neuron* **79**, 416-438 (2013).

[203] Lippmann, R., An introduction to computing with neural nets. *ACM SIGARCH Computer Architecture News* **16**, 7-25 (1988).

[204] Litchfield, J.J., Wilcoxon, F., A simplified method of evaluating dose-effect experiments. *Journal of pharmacology and experimental therapeutics* **96**, 99-113 (1949).

[205] Liu, C., Cui, G., Zhu, M., Kang, X., Guo, H., Neuroinflammation in Alzheimer's disease: chemokines produced by astrocytes and chemokine receptors. *International journal of clinical and experimental pathology* **7**, 8342 (2014).

[206] Liu, E.Y. et al. Loss of nuclear TDP-43 is associated with decondensation of LINE retrotransposons. *Cell reports* **27**, 1409-1421 (2019).

[207] Loeffler, D.A. et al. Increased regional brain concentrations of ceruloplasmin in neurodegenerative disorders. *Brain research* **738**, 265-274 (1996).

[208] Love, M.I., Huber, W., Anders, S., Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 1-21 (2014).

[209] Lukiw, W.J., Surjyadipta, B., Dua, P., Alexandrov, P.N., Common micro RNAs (miRNAs) target complement factor H (CFH) regulation in Alzheimer's disease (AD) and in age-related macular degeneration (AMD). *International journal of biochemistry and molecular biology* **3**, 105 (2012).

[210] Lutz, C., Mouse models of ALS: Past, present and future. *Brain Research* **1693**, 1-10 (2018).

[211] Ma, F., Pellegrini, M., ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinformatics* **36**, 533-538 (2020).

[212] Ma, X.R. et al. TDP-43 represses cryptic exon inclusion in the FTD–ALS gene UNC13A. *Nature* **603**, 124-130 (2022).

[213] Ma, Y., Haynes, R.L., Sidman, R.L., Vartanian, T., TLR8: an innate immune receptor in brain, neurons and axons. *Cell Cycle* **6**, 2859-2868 (2007).

[214] Mackenzie, I.R. et al. Quantitative analysis and clinico-pathological correlations of different dipeptide repeat protein pathologies in C9ORF72 mutation carriers. *Acta neuropathologica* **130**, 845-861 (2015).

[215] Malaspina, A., Kaushik, N., De Belleroche, J., Differential expression of 14 genes in amyotrophic lateral sclerosis spinal cord detected using gridded cDNA arrays. *Journal of neurochemistry* **77**, 132-145 (2001).

[216] Mastrangelo, A. et al. Amyloid-Beta Co-Pathology Is a Major Determinant of the Elevated Plasma GFAP Values in Amyotrophic Lateral Sclerosis. *International Journal of Molecular Sciences* **24**, 13976 (2023).

[217] Mathis, S., Goizet, C., Soulages, A., Vallat, J.M., Le Masson, G., Genetics of amyotrophic lateral sclerosis: A review. *Journal of the Neurological Sciences* **399**, 217-226 (2019).

[218] Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332-337 (2019).

[219] Matsumoto, T. et al. CD44 expression in astrocytes and microglia is associated with ALS progression in a mouse model. *Neuroscience letters* **520**, 115-120 (2012).

[220] Mattiazzi, M. et al. Mutated human SOD1 causes dysfunction of oxidative phosphorylation in mitochondria of transgenic mice. *Journal of biological chemistry* **277**, 29626-29633 (2002).

[221] Mattick, J.S. et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology* **24**, 430-447 (2023).

[222] Mattick, J.S., Makunin, I.V., Non-coding RNA. *Human molecular genetics* **15**, R17-R29 (2006).

[223] Matus, S., Medinas, D.B., Hetz, C., Common ground: stem cell approaches find shared pathways underlying ALS. *Cell stem cell* **14**, 697-699 (2014).

[224] Mc Guire, C., Beyaert, R., van Loo, G., Death receptor signalling in central nervous system inflammation and demyelination. *Trends in neurosciences* **34**, 619-628 (2011).

[225] McClintock, B., Chromosome organization and genic expression. In *Cold Spring Harbor symposia on quantitative biology* **16**, 13-47 (1951).

[226] McGeer, P.L., McGeer, E.G., Glial reactions in Parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society* **23**, 474-483 (2008).

[227] McGoldrick, P., Joyce, P.I., Fisher, E.M., Greensmith, L., Rodent models of amyotrophic lateral sclerosis. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1832**, 1421-1436 (2013).

[228] McWilliam, H. et al. Analysis tool web services from the EMBL-EBI. *Nucleic acids research* **41**, W597-W600 (2013).

[229] Melamed, Z.E. et al. Premature polyadenylation-mediated loss of stathmin-2 is a hallmark of TDP-43-dependent neurodegeneration. *Nature neuroscience* **22**, 180-190 (2019).

[230] Mendez, M.F., Early-onset Alzheimer disease and its variants. *Continuum (Minneapolis, Minn.)* **25**, 34 (2019).

[231] Millecamps, S. et al. Phenotype difference between ALS patients with expanded repeats in C9ORF72 and patients with mutations in other ALS-related genes. *Journal of medical genetics* **49**, 258-263 (2012).

[232] Miller, R.G., Moore, D.H., Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). *Cochrane database of systematic reviews* **3**, (2012).

[233] Mitchell, J.D. et al. Timelines in the diagnostic evaluation of people with suspected amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND)–a 20-year review: can we do better?. *Amyotrophic Lateral Sclerosis* **11**, 537-541 (2010).

[234] Monsalve, D.M. et al. Human VRK2 modulates apoptosis by interaction with Bcl-xL and regulation of BAX gene expression. *Cell death & disease* **4**, e513-e513 (2013).

[235] Monti, S., Tamayo, P., Mesirov, J., Golub, T., Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* **52**, 91-118 (2003).

[236] Moore, J.E. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).

[237] Morello, G. et al. Integrative multi-omic analysis identifies new drivers and pathways in molecularly distinct subtypes of ALS. *Scientific reports* **9**, 9968 (2019).

[238] Morita, S. et al. Combined survival analysis of prospective clinical trials of gefitinib for non–small cell lung cancer with EGFR mutations. *Clinical Cancer Research* **15**, 4493-4498 (2009).

[239] Muotri, A.R. et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903-910 (2005).

[240] Murtagh, F., Contreras, P., Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 86-97 (2012).

[241] Murtagh, F., Legendre, P., Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. *Journal of classification* **31**, 274-295 (2014).

[242] Nassif, M. et al. Pathogenic role of BECN1/Beclin 1 in the development of amyotrophic lateral sclerosis. *Autophagy* **10**, 1256-1271 (2014).

[243] Neumann, M. et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* **314**, 130-133 (2006).

[244] Newman, A.M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12**, 453-457 (2015).

[245] Normand, S. et al. Proteasomal degradation of NOD2 by NLRP12 in monocytes promotes bacterial tolerance and colonization by enteropathogens. *Nature communications* **9**, 5338 (2018).

[246] Noto, Y.I. et al. Elevated CSF TDP-43 levels in amyotrophic lateral sclerosis: specificity, sensitivity, and a possible prognostic value. *Amyotrophic Lateral Sclerosis* **12**, 140-143 (2011).

[247] Nowakowski, T.J. et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318-1323 (2017).

[248] O'Connor, S.A. et al. Neural G0: a quiescent-like state found in neuroepithelial-derived cells and glioma. *Molecular systems biology* **17**, e9522 (2021).

[249] O'Leary, N.A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733-D745 (2016).

[250] Oliver, P.L. et al. Oxr1 is essential for protection against oxidative stress-induced neurodegeneration. *PLoS genetics* **7**, e1002338 (2011).

[251] Orlacchio, A. et al. SPATACSIN mutations cause autosomal recessive juvenile amyotrophic lateral sclerosis. *Brain* **133**, 591-598 (2010).

[252] Paatero, P., Tapper, U., Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111-126 (1994).

[253] Pace, J.K., Feschotte, C., The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome research* **17**, 422-432 (2007).

[254] Paganoni, S. et al. Diagnostic timelines and delays in diagnosing amyotrophic lateral sclerosis (ALS). *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* **15**, 453-456 (2014).

[255] Pascual-Montano, A., Carazo, J.M., Kochi, K., Lehmann, D., Pascual-Marqui, R.D., Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE transactions on pattern analysis and machine intelligence* **28**, 403-415 (2006).

[256] Pasetto, L. et al. Defective cyclophilin A induces TDP-43 proteinopathy: implications for amyotrophic lateral sclerosis and frontotemporal dementia. *Brain* **144**, 3710-3726 (2021).

[257] Pasetto, L. et al. Targeting extracellular cyclophilin A reduces neuroinflammation and extends survival in a mouse model of amyotrophic lateral sclerosis. *Journal of Neuroscience* **37**, 1413-1427 (2017).

[258] Passmore, L.A., Coller, J., Roles of mRNA poly (A) tails in regulation of eukaryotic gene expression. *Nature Reviews Molecular Cell Biology* **23**, 93-106 (2022).

[259] Patel, A.P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).

[260] Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).

[261] Phatnani, H. et al. An integrated multi-omic analysis of iPSC-derived motor neurons from C9ORF72 ALS patients. *Iscience* **24**, (2021).

[262] Philips, T., Rothstein, J.D., Rodent models of amyotrophic lateral sclerosis. *Current protocols in pharmacology* **69**, 5-67 (2015).

[263] Picher-Martel, V., Valdmanis, P.N., Gould, P.V., Julien, J.P., Dupré, N., From animal models to human disease: a genetic approach for personalized medicine in ALS. *Acta neuropathologica communications* **4**, 1-29 (2016).

[264] Platt, R.N., Vandewege, M.W., Ray, D.A., Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research* **26**, 25-43 (2018).

[265] Potvin-Trottier, L., Lord, N.D., Vinnicombe, G., Paulsson, J., Synchronous long-term oscillations in a synthetic gene circuit. *Nature* **538**, 514-517 (2016).

[266] Prudencio, M. et al. Truncated stathmin-2 is a marker of TDP-43 pathology in frontotemporal dementia. *The Journal of clinical investigation* **130**, e139741 (2020).

[267] Prudencio, M., Hart, P.J., Borchelt, D.R., Andersen, P.M., Variation in aggregation propensities among ALS-associated variants of SOD1: correlation to human disease. *Human molecular genetics* **18**, 3217-3226 (2009).

[268] Rahmani, E. et al. BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. *Genome biology* **19**, 1-18 (2018).

[269] Ranganathan, S. et al. Proteomic profiling of cerebrospinal fluid identifies biomarkers for amyotrophic lateral sclerosis. *Journal of neurochemistry* **95**, 1461-1471 (2005).

[270] Ratovitski, T. et al. Variation in the biochemical/biophysical properties of mutant superoxide dismutase 1 enzymes and the rate of disease progression in familial amyotrophic lateral sclerosis kindreds. *Human molecular genetics* **8**, 1451-1460 (1999).

[271] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.L., Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-1555 (2002).

[272] Reijn, T.S., Abdo, W.F., Schelhaas, H.J., Verbeek, M.M., CSF neurofilament protein analysis in the differential diagnosis of ALS. *Journal of neurology* **256**, 615-619 (2009).

[273] Reinhardt, P. et al. Genetic correction of a LRRK2 mutation in human iPSCs links parkinsonian neurodegeneration to ERK-dependent changes in gene expression. *Cell stem cell* **12**, 354-367 (2013).

[274] Renton, A.E. et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257-268 (2011).

[275] Renton, A.E., Chiò, A., Traynor, B.J., State of play in amyotrophic lateral sclerosis genetics. *Nature neuroscience* **17**, 17-23 (2014).

[276] Ripps, M.E., Huntley, G.W., Hof, P.R., Morrison, J.H., Gordon, J.W., Transgenic mice expressing an altered murine superoxide dismutase gene provide an animal model of amyotrophic lateral sclerosis. *Proceedings of the National Academy of Sciences* **92**, 689-693 (1995).

[277] Rivals, I., Personnaz, L., Taing, L., Potier, M.C., Enrichment or depletion of a GO category within a class of genes: which test?. *Bioinformatics* **23**, 401-407 (2007).

[278] Robinson, M.D., McCarthy, D.J., Smyth, G.K., edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

[279] Rohart, F., Gautier, B., Singh, A., Lê Cao, K.A., mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS computational biology* **13**, e1005752 (2017).

[280] Romano, G., Veneziano, D., Acunzo, M., Croce, C.M., Small non-coding RNA and cancer. *Carcinogenesis* **38**, 485-491 (2017).

[281] Rosen, D.R. et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362**, 59-62 (1993).

[282] Rothstein, J.D., Van Kammen, M., Levey, A.I., Martin, L.J., Kuncl, R.W., Selective loss of glial glutamate transporter GLT-1 in amyotrophic lateral sclerosis. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* **38**, 73-84 (1995).

[283] Rué, L. et al. Reduction of ephrin-A5 aggravates disease progression in amyotrophic lateral sclerosis. *Acta neuropathologica communications* **7**, 1-15 (2019).

[284] Sabatelli, M., Conte, A., Zollino, M., Clinical and genetic heterogeneity of amyotrophic lateral sclerosis. *Clinical genetics* **83**, 408-416 (2013).

[285] Sanfilippo, C. et al. CHI3L1 and CHI3L2 overexpression in motor cortex and spinal cord of sALS patients. *Molecular and Cellular Neuroscience* **85**, 162-169 (2017).

[286] Sanger, F., Nicklen, S., Coulson, A.R., DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463-5467 (1977).

[287] Santa-Maria, I. et al. Dysregulation of microRNA-219 promotes neurodegeneration through post-transcriptional regulation of tau. *The Journal of clinical investigation* **125**, 681-686 (2015).

[288] Sasaki, S., Iwata, M., Mitochondrial alterations in the spinal cord of patients with sporadic amyotrophic lateral sclerosis. *Journal of Neuropathology & Experimental Neurology* **66**, 10-16 (2007).

[289] Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A., Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **33**, 495-502 (2015).

[290] Schaefer, M., Plant, T.D., Stresow, N., Albrecht, N., Schultz, G., Functional differences between TRPC4 splice variants. *Journal of Biological Chemistry* **277**, 3752-3759 (2002).

[291] Schaub, M., Keller, W., RNA editing by adenosine deaminases generates RNA and protein diversity. *Biochimie* **84**, 791-803 (2002).

[292] Schroeder, A. et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC molecular biology* **7**, 1-14 (2006).

[293] Schwab, N., Tator, C., Hazrati, L.N., DNA damage as a marker of brain damage in individuals with history of concussions. *Laboratory Investigation* **99**, 1008-1018 (2019).

[294] Shahidullah, M. et al. Defects in synapse structure and function precede motor neuron degeneration in Drosophila models of FUS-related ALS. *Journal of Neuroscience* **33**, 19590-19598 (2013).

[295] Šidák, Z., Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* **62**, 626-633 (1967).

[296] Siddique, T. et al. Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity. *New England Journal of Medicine* **324**, 1381-1384 (1991).

[297] Simillion, C., Liechti, R., Lischer, H.E., Ioannidis, V., Bruggmann, R., Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC bioinformatics* **18**, 1-14 (2017).

[298] Simpson, D.S., Oliver, P.L., ROS generation in microglia: understanding oxidative stress and inflammation in neurodegenerative disease. *Antioxidants* **9**, 743 (2020).

[299] Smid, M. et al. Subtypes of breast cancer show preferential site of relapse. *Cancer research* **68**, 3108-3114 (2008).

[300] Smit, A.F.A., Riggs, A.D., MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic acids research* **23**, 98-102 (1995).

[301] Smit, A.F.A., Tóth, G., Riggs, A.D., Jurka, J., Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *Journal of molecular biology* **246**, 401-417 (1995).

[302] Smyth, G.K., Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, 397-420 (2005).

[303] Snowden, J.S. et al. Distinct clinical and pathological characteristics of frontotemporal dementia associated with C 9ORF72 mutations. *Brain* **135**, 693-708 (2012).

[304] Sonntag, K.C. et al. Limited predictability of postmortem human brain tissue quality by RNA integrity numbers. *Journal of neurochemistry* **138**, 53-59 (2016).

[305] Stark, R., Grzelak, M., Hadfield, J., RNA sequencing: the teenage years. *Nature Reviews Genetics* **20**, 631-656 (2019).

[306] Stathakis, D., How many hidden layers and nodes?. *International Journal of Remote Sensing* **30**, 2133-2147 (2009).

[307] Steinacker, P. et al. Chitotriosidase (CHIT1) is increased in microglia and macrophages in spinal cord of amyotrophic lateral sclerosis and cerebrospinal fluid levels correlate with disease severity and progression. *Journal of Neurology, Neurosurgery & Psychiatry* **89**, 239-247 (2018).

[308] Steinacker, P. et al. TDP-43 in cerebrospinal fluid of patients with frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Archives of neurology* **65**, 1481-1487 (2008).

[309] Storkebaum, E. et al. Treatment of motoneuron degeneration by intracerebroventricular delivery of VEGF in a rat model of ALS. *Nature neuroscience* **8**, 85-92 (2005).

[310] Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545-15550 (2005).

[311] Sud, M. et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic acids research* **44**, D463-D470 (2016).

[312] Süssmuth, S.D. et al. CSF glial markers correlate with survival in amyotrophic lateral sclerosis. *Neurology* **74**, 982-987 (2010).

[313] Sweeney, M.D., Ayyadurai, S., Zlokovic, B.V., Pericytes of the neurovascular unit: key functions and signaling pathways. *Nature neuroscience* **19**, 771-783 (2016).

[314] Sweeney, M.D., Kisler, K., Montagne, A., Toga, A.W., Zlokovic, B.V., The role of brain vasculature in neurodegenerative disorders. *Nature neuroscience* **21**, 1318-1331 (2018).

[315] Swinnen, B., Robberecht, W., The phenotypic variability of amyotrophic lateral sclerosis. *Nature Reviews Neurology* **10**, 661-670 (2014).

[316] Takahashi, K., Rochford, C.D., Neumann, H., Clearance of apoptotic neurons without inflammation by microglial triggering receptor expressed on myeloid cells-2. *The Journal of experimental medicine* **201**, 647-657 (2005).

[317] Tam, O.H. et al. Postmortem cortex samples identify distinct molecular subtypes of ALS: retrotransposon activation, oxidative stress, and activated glia. *Cell reports* **29**, 1164-1177 (2019).

[318] Tam, O.H., Ostrow, L.W., Hammell, M.G., Diseases of the nERVous system: retrotransposon activity in neurodegenerative disease. *Mobile DNA* **10**, 1-14 (2019).

[319] Tan, M.G. et al. Genome wide profiling of altered gene expression in the neocortex of Alzheimer's disease. *Journal of neuroscience research* **88**, 1157-1169 (2010).

[320] Tano, K., Akimitsu, N., Long non-coding RNAs in cancer progression. *Frontiers in genetics* **3**, 219 (2012).

[321] Taylor, J.P., Brown Jr., R.H., Cleveland, D.W., Decoding ALS: from genes to mechanism. *Nature* **539**, 197-206 (2016).

[322] The Cancer Genome Atlas Research Network, Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).

[323] Therneau, T., Mixed effects Cox models. *CRAN repository* (2015).

[324] Therneau, T.M., Grambsch P.M., Modeling Survival Data: Extending the Cox Model. (Springer, New York, 2000).

[325] Therneau, T.M., Lumley, T., Package ‘survival’. *R Top Doc* **128**, 28-33 (2015).

[326] Thompson, A.G. et al. CSF chitinase proteins in amyotrophic lateral sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry* **90**, 1215-1220 (2019).

[327] Tijms, B.M. et al. Cerebrospinal fluid proteomics in patients with Alzheimer’s disease reveals five molecular subtypes with distinct genetic risk profiles. *Nature aging*, 1-15 (2024).

[328] Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515 (2010).

[329] Turner, M.R., Parton, M.J., Shaw, C.E., Leigh, P.N., Al-Chalabi, A., Prolonged survival in motor neuron disease: a descriptive study of the King’s database 1990–2002. *Journal of Neurology, Neurosurgery & Psychiatry* **74**, 995-997 (2003).

[330] Van der Maaten, L., Hinton, G., Visualizing data using t-SNE. *Journal of machine learning research* **9**, (2008).

[331] Vassileff, N. et al. Revealing the proteome of motor cortex derived extracellular vesicles isolated from amyotrophic lateral sclerosis human postmortem tissues. *Cells* **9**, 1709 (2020).

[332] Verstraete, E. et al. TDP-43 plasma levels are higher in amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis* **13**, 446-451 (2012).

[333] Vissing, J. et al. A heterozygous 21-bp deletion in CAPN3 causes dominantly inherited limb girdle muscular dystrophy. *Brain* **139**, 2154-2163 (2016).

[334] Vizcaíno, J.A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology* **32**, 223-226 (2014).

[335] Volonté, C., Apolloni, S., Parisi, C., Amadio, S., Purinergic contribution to amyotrophic lateral sclerosis. *Neuropharmacology* **104**, 180-193 (2016).

[336] Vu, L. et al. Cross-sectional and longitudinal measures of chitinase proteins in amyotrophic lateral sclerosis and expression of CHI3L1 in activated astrocytes. *Journal of Neurology, Neurosurgery & Psychiatry* **91**, 350-358 (2020).

[337] Vu, L. et al. Proteomics and mathematical modeling of longitudinal CSF differentiates fast versus slow ALS progression. *Annals of clinical and translational neurology* **10**, 2025-2042 (2023).

[338] Wainger, B.J. et al. Intrinsic membrane hyperexcitability of amyotrophic lateral sclerosis patient-derived motor neurons. *Cell reports* **7**, 1-11 (2014).

[339] Walhout, R., Verstraete, E., Van Den Heuvel, M.P., Veldink, J.H., Van Den Berg, L.H., Patterns of symptom development in patients with motor neuron disease. *Amyotrophic lateral sclerosis and frontotemporal degeneration* **19**, 21-28 (2018).

[340] Wang, H. et al. miR-219 cooperates with miR-338 in myelination and promotes myelin repair in the CNS. *Developmental cell* **40**, 566-582 (2017).

[341] Wang, N. et al. UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics* **31**, 137-139 (2015).

[342] Wang, X., Park, J., Susztak, K., Zhang, N.R., Li, M., Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications* **10**, 380 (2019).

[343] Westeneng, H.J. et al. Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. *The Lancet Neurology* **17**, 423-433 (2018).

[344] White, E.S., Baralle, F.E., Muro, A.F., New insights into form and function of fibronectin splice variants. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* **216**, 1-14 (2008).

[345] White, K. et al. Effect of postmortem interval and years in storage on RNA quality of tissue at a repository of the NIH NeuroBioBank. *Biopreservation and biobanking* **16**, 148-157 (2018).

[346] Whitwell, J.L. et al. Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: a case-control study. *The Lancet Neurology* **11**, 868-877 (2012).

[347] Wilkerson, M.D., Hayes, D.N., ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573 (2010).

[348] Wilkinson, M.E., Charenton, C., Nagai, K., RNA splicing by the spliceosome. *Annual review of biochemistry* **89**, 359-388 (2020).

[349] Wishart, D.S. et al. HMDB 5. 0: the human metabolome database for 2022. *Nucleic acids research* **50**, D622-D631 (2022).

[350] Wold, S., Esbensen, K., Geladi, P., Principal component analysis. *Chemometrics and intelligent laboratory systems* **2**, 37-52 (1987).

[351] Wong, P.C. et al. An adverse property of a familial ALS-linked SOD1 mutation causes motor neuron disease characterized by vacuolar degeneration of mitochondria. *Neuron* **14**, 1105-1116 (1995).

[352] Wu, T.D., Nacu, S., Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-881 (2010).

[353] Xiao, L. et al. Multiple-tissue integrative transcriptome-wide association studies discovered new genes associated with amyotrophic lateral sclerosis. *Frontiers in genetics* **11**, 587243 (2020).

[354] Xu, W.N. et al. Mitochondrial NDUFA4L2 attenuates the apoptosis of nucleus pulposus cells induced by oxidative stress via the inhibition of mitophagy. *Experimental & molecular medicine* **51**, 1-16 (2019).

[355] Yadav, A. et al. A cellular taxonomy of the adult human spinal cord. *Neuron* **111**, 328-344 (2023).

[356] Yamanaka, K. et al. Astrocytes as determinants of disease progression in inherited amyotrophic lateral sclerosis. *Nature neuroscience* **11**, 251-253 (2008).

[357] Yang, A. et al. p63, a p53 homolog at 3q27–29, encodes multiple products with transactivating, death-inducing, and dominant-negative activities. *Molecular cell* **2**, 305-316 (1998).

[358] Yang, S., Lim, K.H., Kim, S.H., Joo, J.Y., Molecular landscape of long noncoding RNAs in brain disorders. *Molecular Psychiatry* **26**, 1060-1074 (2021).

[359] Yang, W.R., Ardeljan, D., Pacyna, C.N., Payer, L.M., Burns, K.H., SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic acids research* **47**, e27-e27 (2019).

[360] Yao, R.W., Wang, Y., Chen, L.L., Cellular functions of long noncoding RNAs. *Nature cell biology* **21**, 542-551 (2019).

[361] Yeini, E. et al. P-selectin axis plays a key role in microglia immunophenotype and glioblastoma progression. *Nature communications* **12**, 1912 (2021).

[362] Yeo, G., Holste, D., Kreiman, G., Burge, C.B., Variation in alternative splicing across human tissues. *Genome biology* **5**, 1-15 (2004).

[363] Yugi, K., Kubota, H., Hatano, A., Kuroda, S., Trans-omics: how to reconstruct biochemical networks across multiple ‘omic’layers. *Trends in biotechnology* **34**, 276-290 (2016).

[364] Zemmour, D., Pratama, A., Loughhead, S.M., Mathis, D., Benoist, C., Flicr, a long noncoding RNA, modulates Foxp3 expression and autoimmunity. *Proceedings of the National Academy of Sciences* **114**, E3472-E3480 (2017).

[365] Zeng, T. et al. BACE1-AS prevents BACE1 mRNA degradation through the sequestration of BACE1-targeting miRNAs. *Journal of chemical neuroanatomy* **98**, 87-96 (2019).

[366] Zhang, B. et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer’s disease. *Cell* **153**, 707-720 (2013).

[367] Zhang, P., Wu, W., Chen, Q., Chen, M., Non-coding RNAs and their integrated networks. *Journal of integrative bioinformatics* **16**, 20190027 (2019).

[368] Zhang, R. et al. Topology-dependent interference of synthetic gene circuit function by growth feedback. *Nature chemical biology* **16**, 695-701 (2020).

[369] Zhao, W. et al. Activated microglia initiate motor neuron injury by a nitric oxide and glutamate-mediated mechanism. *Journal of Neuropathology & Experimental Neurology* **63**, 964-977 (2004).

[370] Zhao, Y. et al. TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. *Journal of translational medicine* **19**, 1-15 (2021).

[371] Zou, Z.Y. et al. Genetic epidemiology of amyotrophic lateral sclerosis: a systematic review and meta-analysis. *Journal of Neurology, Neurosurgery & Psychiatry* **88**, 540-549 (2017).

APPENDIX A

CONTINUATION OF ACKNOWLEDGEMENTS

Through the Smith Lab, I have been introduced to some of the most genuine and talented individuals that I hope to call lifelong friends – all of whom deserve a specific shout out. Christopher Miranda was the first PhD student in the Smith Lab and was someone that I looked up to throughout. Chris is a scientific beast, and it would be impossible to find a day when he was not the first to arrive and last to leave. He showed me firsthand the value of persistence, hard work, and passion for research – I could not have had a better scientific role model. Likewise, Joel Lusk was a role model and mentor for me, and essentially all other students in the lab. His patience, attentiveness, creativity, willingness to help, and breadth of knowledge are truly superpowers. Also, we all owe you a huge thank you for creating a digital copy of the chemical inventory and pretty much single-handedly keeping us compliant with the ever-changing rules of lab work. Mostly outside of the lab, I'd like to thank you both, and Ethan Marschall, for your friendship and lifelong memories including happy hours, national conferences, attending my first oral presentation, an expensive sword and watermelon, custom lightsaber, bold financial bets, house parties, and doing cool science demonstrations for the public. I'd like to thank colleagues: (i) Ethan Marschall for always being willing to chat with me, take a coffee break, handling purchasing, or listen to me vent about oddly specific things, (ii) Benjamin Ambrose for your long-term commitment to our project, enthusiasm, and creativity, and (iii) the trio of remarkable students, Youssef Abdellatif, Paula Phan, and Taylor Pennington, and technician Angela Ponce Olea that helped get the Biodome project to the finish line. I'd like to thank Ethan Marschall, Taylor Pennington, Jordan Garcia, and Paula Phan for their help with figure preparation throughout my PhD. I'd also like to thank all current and former Smith Lab members for their mentorship, friendship, insight, conversation, positivity, and good memories: Stephanie Ong, Matthew Chrest, Blake Browning, Justin Mieth, Madeleine Howell, Vi Nguyen, Yuka Sugamura, Raiyan Choudhury, Esther Sim, Jorge Favian Rios, Joshua Burgett, Derek Smetanick, Jamie Akbari-Carpenter, and Jordan Garcia.

Outside of the Smith Lab, I am thankful for the opportunity to work alongside or in collaboration with a number of talented individuals. To my proteomic mentors, Ronghu Wu and Suttipong “Jay” Suttapitugsakul, thank you for your endless knowledge, crucial guidance, and for showing me an extended amount of patience during my headfirst dive into the deep end of biochemistry. The skillsets I learned during my REU helped to expand my systems biology perspective and allowed me to draw important parallels across mass spectrometry-based omics. To my long-time colleague in the area of volatile metabolomics, Dr. Trenton Davis, thank you for entertaining my stupid questions, quashing paranoia, helping with maintenance, method development, and data analysis and for always responding to late-night texts generally along the lines of “help, the instrument isn’t working”. To Vlad Voziyanov, Fadi Gerges, Livia De Mesquita Teixeira, Michael Daeger, Megan Donnay, Ethan Marschall, Sophia, Joe Guimond, and Taylor Pennington, thank you all for the much-needed happy hours. To our collaborators at TGen, thank you for showing us plenty of patience and providing significant insight and support. To my scientific partner in crime, Taylor Pennington, thank you for the extra support and motivation when things were difficult, coffee breaks, late-night writing sessions, scientific enthusiasm and persistence, custom glove bag attempt, and unforgettable moments like cleaning open wounds with ethanol, reviving the mass spectrometer, gel electrophoresis pains, introducing me to Alani energy drinks, having me look into a CDC important license for a very specific bacteria strain, covering the RNAi protocol when I was unwell, and helping me locate lipofectamine on a Sunday night (thank you Benjamin Bartelle). To the SBHSE support staff, thank you for supporting the day-to-day and helping enable student success.

To Dahn Troung, Ali Navaei, Harpinder Saini, Jamie Veldhuizen, Alejandra Patino-Guerrero, Arati Sridharan, Jonathan Duncan, Vlad Voziyanov, Fadi Gerges, Livia De Mesquita Teixeira, Michael Daeger, Elizabeth Jitendran, Swathy Kumar, Samantha O’Connor, Sierra Wilferd, Marisa Masles, Yi Ren, Christina Forbes, Joe Holland, Hamid Esmaeili, Kalpana Ravi, Twinkle Manoharan, Parisa Taklifi, Oscar Perez, Ronald Nelson Jr., Megan Donnay, Alex Morales, Kaila Gemenes, Gradi Bamfonga, Joe Guimond, Adam Kindelin, Ben Gonzalez, Christine Roeger, Solo Pyon, Laura Hawes, Megan Maurer, and Patricia Stepp, thank you all for your past and present friendships, conversation, happy hours, collaboration, knowledge, perspective, and support. You all have helped make the day-to-day and general PhD experience a highly positive one. I’d like to thank Samantha O’Connor for her collaboration and cell deconvolution analysis, and Sierra Wilferd for her collaboration and discussion on a cool grant idea. Further, I’d like to give a sincere thank you the following influential faculty that contributed to my growth and continued to support me outside of their classroom: Michael Sobrado, Jit Muthuswamy, Mehdi Nikkhah, Christopher Buneo, Madeline Andrews, Xiaojun Tian, Jerry Coursen, and Vincent Pizziconi.

Without the following funding sources, this dissertation would not have been possible: National Science Foundation, Flinn Foundation, and Arizona State University. Likewise, without the crucial initiatives by the New York Genome Center ALS Consortium and Target ALS – and support from the patients and family members – new insight gained from this dissertation would have been limited. By extension, the ALS Association and Tow Foundation are thanked for supporting the generation of RNA-sequencing data used in this work. Lastly, insight gained from this dissertation is attributed to the mountain of important research on which it stands.