Position Analysis of Low Complexity and Compositionally Biased Regions in ALS-Associated Proteins with Liquid-Liquid Phase Separation Properties

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disease characterized by the progressive degeneration of motor neurons, yielding several muscle and respiratory defects ultimately leading to death. Recent experimental data suggest a direct relationship between proteins forming Liquid Liquid Phase Separated condensates and ALS occurrence. Here, we leverage computational methods along with pre-existing online tools to address statistically significant differences between ALS/LLPS related proteins and non-ALS/LLPS related proteins. In this study, we employ computational methods and existing online tools to investigate the differences between ALS-associated LLPS proteins and non-ALS LLPS proteins. By analyzing these features, we aim to uncover insights into the molecular mechanisms underlying ALS and identify potential biomarkers for the disease.

Key Words: ALS, LLPS, LCRs, CBRs, proteins

Introduction

Amyotrophic Lateral Sclerosis (ALS) is a neuromuscular disease that affects the nerve cells in the brain and spinal cord and leads to gradual degradation of the motor neuron. When nerve cells break down, muscle weakness occurs, and more often than not, paralysis and breathing failure follows (Brotman et al., 2024). Despite decades of research, no clear cause of ALS has emerged. It is thought to result from a mix of genetic predisposition, environmental influences and cellular dysfunctions. Despite great progress in understanding the molecular pathways, there are still no curative therapies and care is therefore essentially palliative.

Epidemiology of ALS

The incidence of ALS ranges from 1.5 to 2.5 per 100,000 across global regions (Morgan et al., 2016). This disease primarily affects individuals in the 50–70-year age range, with an approximate 1.5 to 1 male-to-female ratio. About 90% of all cases are considered sporadic, in other words, they arise or have no known familial history, while some (10%) cases are familiar linked to genetic mutations like C9orf72 (This mutation is been associated with the dysregulation of RNA, abnormal protein aggregation and the dysregulation of the nucleocytoplasmic transport through the autophagy-proteostasis pathway), SOD1 (has been linked to oxidative stress, mitochondrial dysfunction and has a role in the pathway of PI3K-Akt signaling pathway), TARDBP (involved in RNA processing, TDP-43 protein aggregation, and disruption of cytoskeletal pathways that lead to axonal transport issues), and FUS (affecting RNA metabolism, transport, and dynamics of stress granules, which impacts the autophagy-proteostasis pathway) Rizea et al., 2024.

Mechanisms of ALS Pathogenesis

Several related systems contribute to the progression of ALS through multiple complex mechanisms. The defining feature of ALS includes protein aggregation among other factors. TDP-43 and FUS along with other abnormal misfolded proteins accumulate excessively within cells which disrupt cellular processes and break the relationships between autophagy-proteostasis and RNA metabolism (Masrori & Van Damme, 2020). The processing of RNA continues to serve as a crucial ALS factor because alterations in TARDBP and FUS genes disrupt RNA functions which subsequently causes motor neuron damage (Marchal-Crespo et al., 2019). The synaptic cleft chemical glutamate creating excessive levels triggers excitotoxicity which leads to neuronal cell death and oxidative stress (Yonehara & Roska, 2017). The PI3K-Akt pathway drives neuronal apoptosis by enhancing oxidative stress and energy deficits which results from mitochondrial dysfunction leading to undermined neuronal integrity (Foust et al., 2009). Research indicates that activated microglia and astrocytes release pro-inflammatory cytokines which speed up neuronal death (Landegger et al., 2017).

Pathways Involved in ALS

Study findings have identified the dysregulation of the PI3K-Akt pathway as a main mechanism in ALS since this signaling pathway controls neuronal survival and apoptosis in patients (Deverman et al., 2016). The autophagy-proteostasis pathway functions as a critical mechanism for protein homeostasis yet its failure causes toxic aggregates to accumulate that affect motor neurons together with glial cells (Cronin et al., 2014). The harm done to motor neuron cells by disturbances of the cytoskeleton becomes a primary source leading to motor neuron disease pathology (Tervo et al., 2016).

Liquid-Liquid Phase Separation (LLPS): Formation and Functions

Liquid-Liquid Phase Separation (LLPS) is the fundamental mechanism underlying the formation of membraneless organelles in eukaryotic cells. Formation of such condensates occurs when macromolecules, mostly proteins along with nucleic acids, demix from the solution and form dense, liquid-like droplets that are membranes yet distinguishable from the surrounding cellular content. Such organization of macromolecules enables compartmentalization of cellular materials without the need of a lipid layer, thus enabling a spatiotemporal and dynamic control mechanism of biochemical processes (Zhang et al., 2020).

Several functionalities have been attributed to LLPS formation. Inside the cell nucleus LLPS facilitates fundamental processes such as transcriptional regulation by forming transcriptional factories, RNA synthesis and organization in nucleoli, chromatin segregation in active and inactive parts and others . For example in mouse cells large heterochromatic regions spanning pericentromeric chromosomal regions are marked with H3K9me3 histone modification and organized along with Heterochromatin Protein 1 (HP1) in large heterochromatic foci in the form of nuclear condensates, while ribosomal proteins using

rRNAs as scaffolds are organized into the most profound nuclear condensates, the nucleoli(Schoelz and Riddle 2022).

Role of Proteins with Intrinsically Disordered Regions (IDRs) in LLPS Formation

A major driver of LLPS formation is proteins with intrinsically disordered regions (IDRs). IDRs are protein regions of lower complexity that do not adopt any specific secondary or tertiary structure. Often these regions correspond to highly flexible protein regions with biased amino acid frequency that allow for different conformations, thus increasing the interaction space with other IDRs. Interestingly, it has been demonstrated that such disordered protein regions often cary post translational modification sites, subsequently allowing for dynamic regulation of the phase separation process (Zganget al., 2020, Mittag et al., 2021)

Relationship Between ALS and LLPS

Recent data support the presence of direct association of ALS and LLPS formation. Interestingly FUS and TDP-43, two RNA binding proteins have been recognized as key players in ALS. Mutations on the protein sequence of the two proteins lead to improper phase separation while also favoring the formation of protein aggregates found in patients with ALS.
FUS, which has IDR and RNA-binding domains, phase-separates into liquid-like condensate in the physiological conditions. Mutations of FUS causing ALS, however, disrupt its native phase behavior and triggers liquid-like condensate to solid-like aggregate transition typical of neurodegeneration. It is enhanced by FUS interaction with G-quadruplex RNAs that enhance its phase separation and resulting aggregation (Ishiguro et al., 2021). Mutations of TDP-43 similarly eliminate its RNA-binding function and trigger pathological LLPS and formation of toxic inclusion bodies in neurons.

Materials and Methods

In this study, we constructed a comprehensive dataset of ALS-associated proteins with phase separation characteristics by integrating data from diverse sources.
The initial dataset was downloaded from DisGeNET, a database of gene-disease associations, where 126 ALS-associated genes were downloaded. These were intersected with DrLLPS, a database of proteins known to undergo liquid-liquid phase separation (LLPS). The intersection resulted in 48 genes that are both ALS-associated and undergo LLPS.

To create a control dataset for analysis, we first subtracted the 48 ALS-related genes from the whole DrLLPS database. We then selected 48 other genes that are related to LLPS but not ALS randomly. We finally combined these two dataset (diseased and control), creating a final dataset of 96 genes.

The protein sequences for these 96 genes were retrieved using BioPython, a comprehensive set of Python tools for computational biology. Sequence handling capabilities of BioPython were utilized to create FASTA format files containing the complete protein sequences for each gene entry.

The resultant FASTA files were run through PlatoLoco, which is a specialist bioinformatics tool for protein sequence analysis with particular focus on the detection and characterization of Low Complexity Regions (LCRs) and Compositionally Biased Regions (CBRs) in proteins. Results were provided as JSON output for analysis.

Low Complexity Regions (LCRs) are segments of protein sequences that have a biased amino acid composition, which is frequently expressed in the form of repetitive elements or low amino acid diversity. LCRs can be significant for protein function, stability, and interactions, and particularly for phase separation, where they can facilitate the formation of biomolecular condensates.

Compositional Biased Regions (CBRs) are the areas wherein certain amino acids are over-represented compared to their typical distribution in proteins. CBRs can influence protein folding, stability, and interactions, and are often implicated in protein aggregate and condensate formation, which is relevant in a variety of biological processes and diseases.

Results

Several tests were performed in order to investigate the Role of Low Complexity Regions and Compositional Bias Regions in ALS disease.

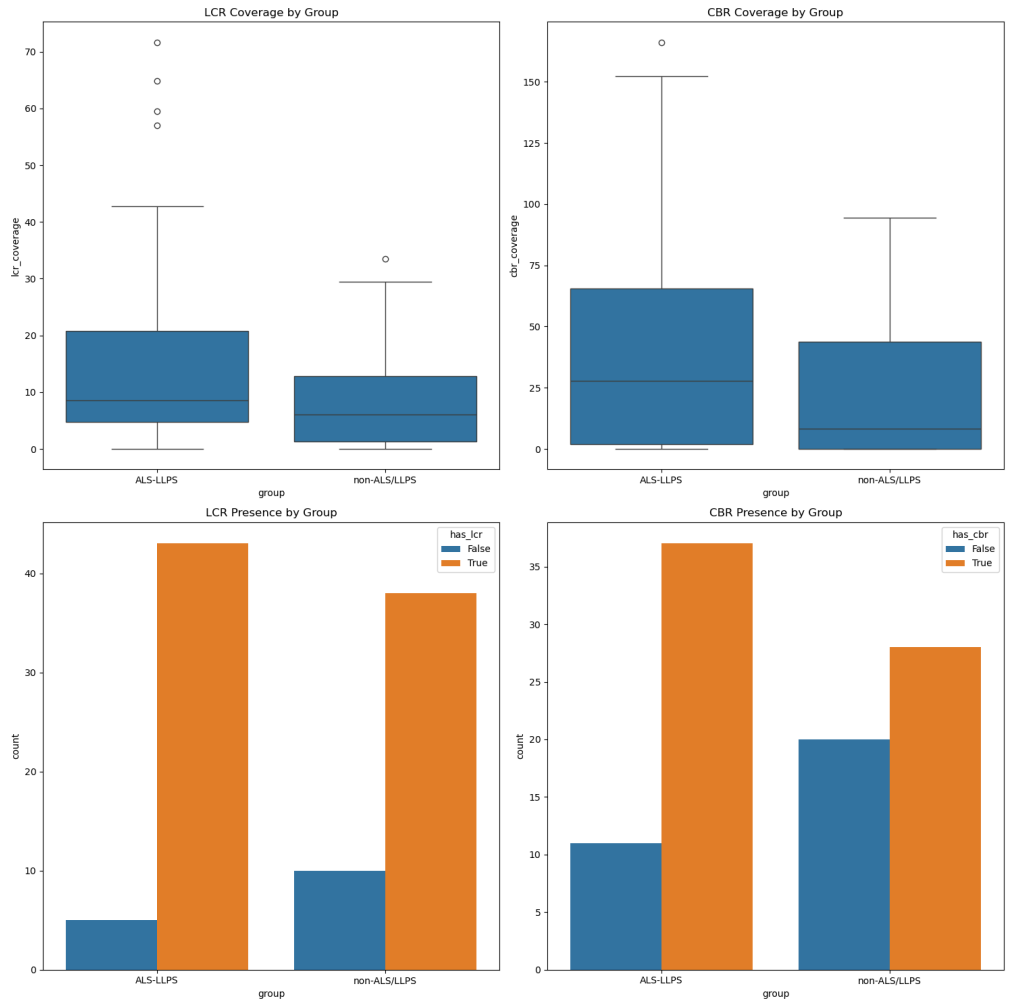
Statistical Analysis

We developed a t-test to assess differences in the number and coverage of low-complexity regions (LCRs) and charged-biased regions (CBRs) between ALS-LLPS proteins and a control group. Analysis of the number of LCRs revealed no significant difference (p-value: 0.052) but suggests a trend toward ALS-LLPS proteins having more LCRs compared to the control group. We observed the same trend for the number of CBRs with a non-statistically significant result (p-value: 0.089).

In terms of coverage, ALS-LLPS proteins cover a significantly higher proportion of their sequence occupied by LCRs compared to controls ($t = 2.604$, $p = 0.011$). Similarly, CBR coverage was also significantly greater in ALS-LLPS proteins ($t = 2.327$, $p = 0.022$). These results indicate that a significantly large part of the proteins contain LCRs and CBRs.

We also used a Chi-square test to check the importance of the presence or absence of LCRs and CBRs associated with the disease in LLPS proteins. The null hypothesis (H_0) in both cases states that the presence of LCRs or CBRs is independent of ALS-LLPS, meaning there is no meaningful association between these features and the disease-related protein group. First, for the LCR presence, Chi-square has shown that 89.6% (43) of ALS-LLPS proteins contained LCRs, compared to 79.2% (38) of non-ALS/LLPS proteins. Still, there is no statistically significant association between

LCR presence and ALS-LLPS status ($\chi^2 = 1.264$, $p = 0.261$). Chi-square has shown that for the CBR presence, 77.1% (37) of ALS-LLPS proteins contain CBRs, compared to 58.3% (28) of non-ALS/LLPS proteins. The result for this case is the same as above, as there is no statistically significant association between CBRs and the ALS-LLPS. The results of the chi-square suggest a potential trend toward a higher proportion of CBRs in ALS-LLPS proteins.



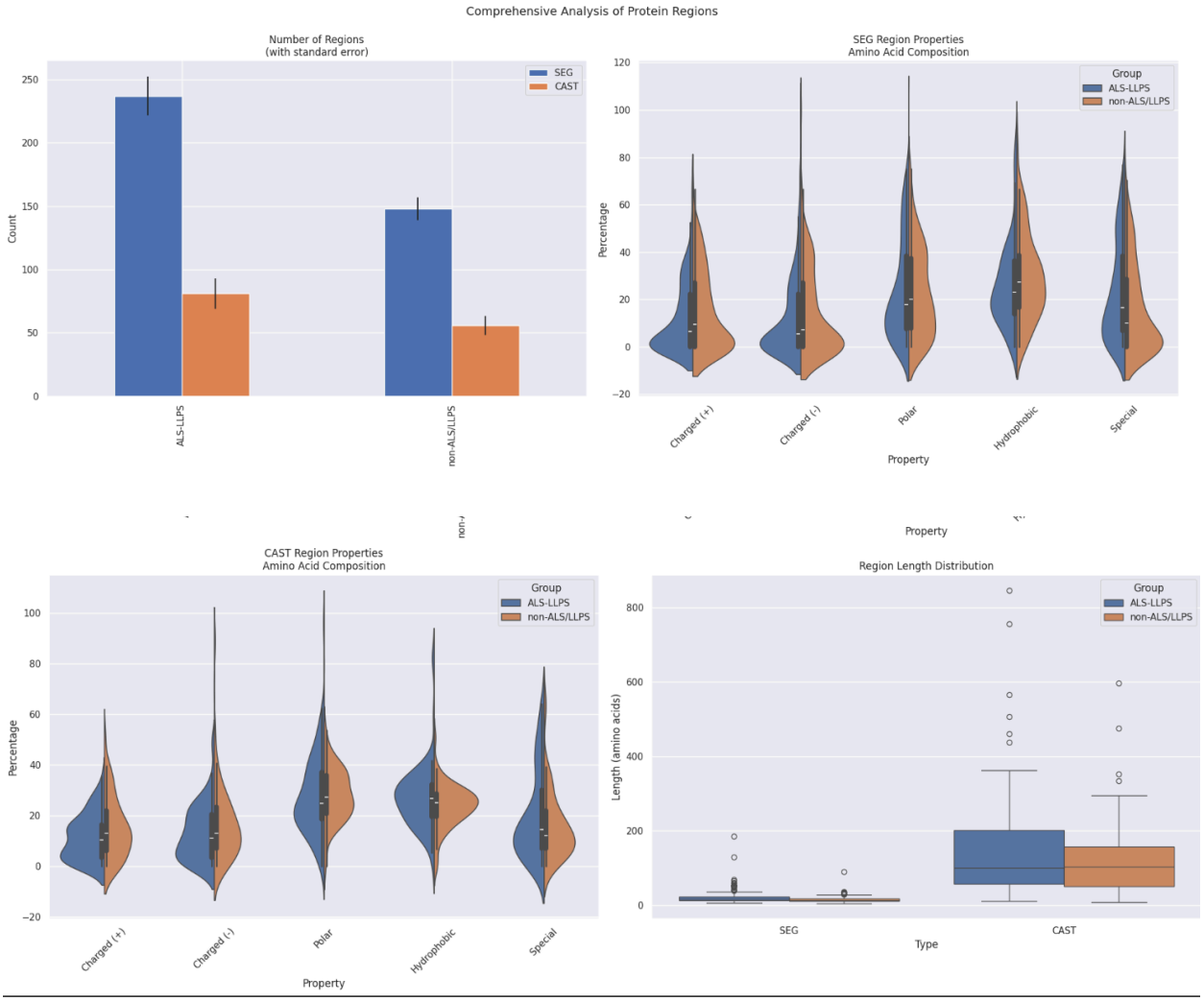
Sequence Composition Analysis

Then, we executed a sequence composition analysis using SEG, a tool for identification of low-complexity regions, and CAST for detecting compositionally biased regions. We applied these tools to the proteins associated with the disease and the control group to analyze differences in sequence composition. On the one hand, the analysis conducted with SEG showed that ALS-LLPS genes have more SEG regions (237) than non-ALS/LLPS genes (148) and that SEG regions in ALS-LLPS genes are significantly longer, with a statistically significant difference (Mean: 21.1 vs. 15.9, $p = 0.001$). On the other hand, the analysis performed with CAST showed that ALS-LLPS genes have more CAST regions (81) than non-ALS/LLPS genes (56). Although there is no significant difference in CAST region length ($p=0.213$).

Amino acid composition analysis revealed that SEG-identified regions in ALS-LLPS proteins were significantly enriched (p=0.004) in special residues such as glycine (23.6%) and proline (16.9%), which are known to take part in protein structural properties. For the same analysis, CAST-identified regions had a significantly lower percentage of positively charged residues in ALS-LLPS genes (11.5% vs. 15.2%, p = 0.024).

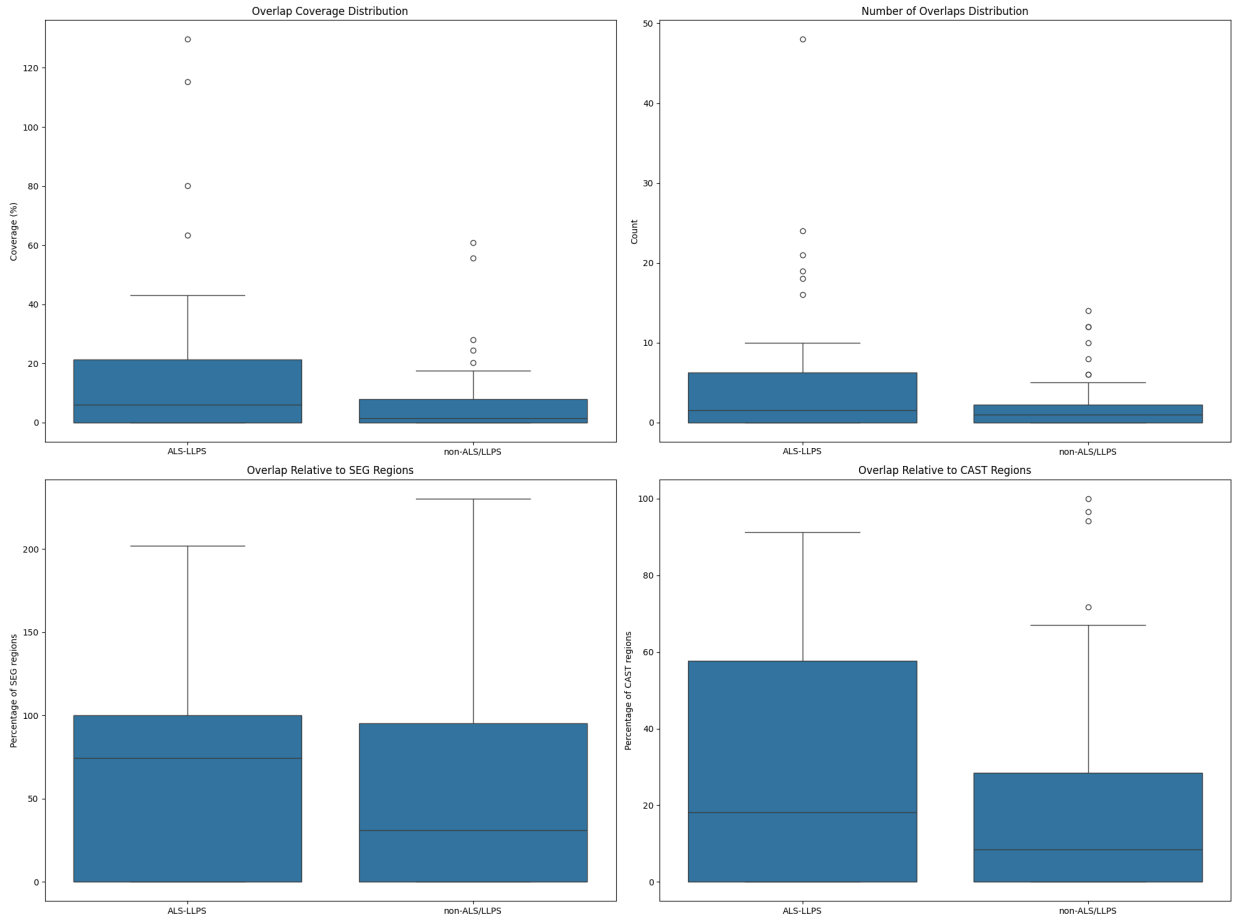
Region Distribution Analysis

We compared the ALS-LLPS and non-ALS/LLPS proteins in terms of the frequency and coverage of SEG- and CAST-identified areas. ALS-LLPS proteins showed about double the coverage (15.69% vs. 8.24%) and had over 60% more SEG-identified regions per protein (4.94 vs. 3.08). Also, there is higher variability in ALS/LLPS (SD: 17.72% vs. 8.41%), suggesting patterns of heterogeneous distribution. Similar results were seen for CAST-identified areas, with ALS-LLPS proteins showing considerably greater coverage (39.75% vs. 22.59%) and 44% more regions per protein (1.69 vs. 1.17). Additionally, these proteins showed higher variability (SD: 41.84% vs. 28.37%), indicating variations in the distribution of CAST-defined compositionally biased areas within ALS-LLPS proteins.



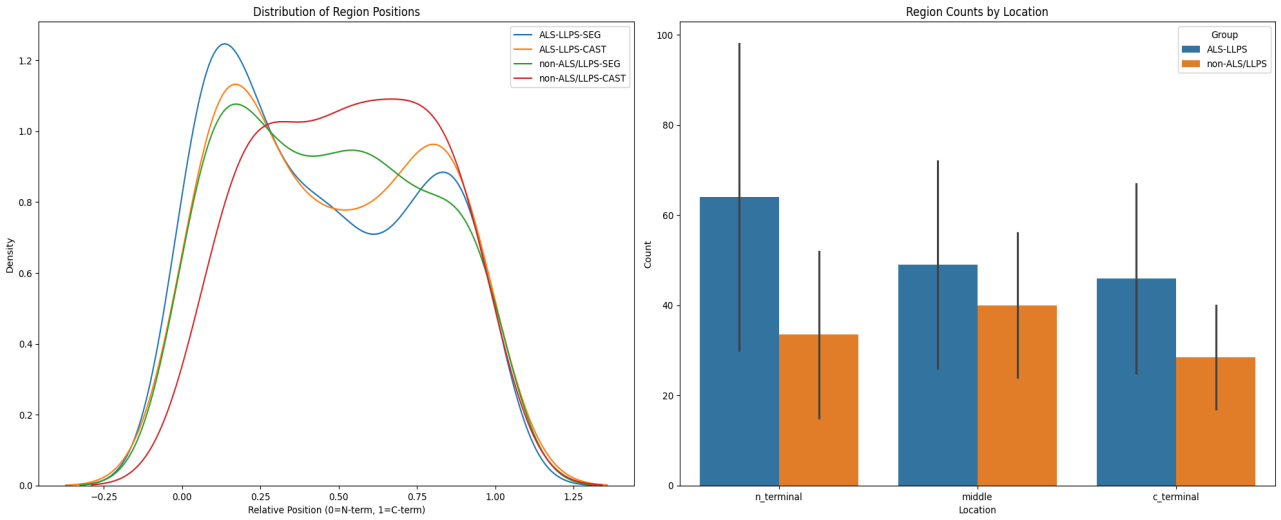
SEG-CAST Overlap Analysis

Next, we analyzed the overlap between SEG and CAST identified regions. Overlap refers to sequence regions that are detected by both tools. Our analysis shows significantly higher overlap in ALS-LLPS proteins, with SEG-CAST overlap covering a larger proportion of the sequence (16.86% vs. 6.87%, p = 0.029) and occurring more frequently (5.15 overlaps vs. 2.21 overlaps, p = 0.034). These results indicate that ALS-LLPS are more likely to contain regions of low complexity and compositionally biased sequence features. However, the overlap to SEG ratio is not statistically significant (p=0.138). This ratio measures what percentage of SEG regions are involved in overlaps with CAST regions, implying that while ALS-LLPS proteins may have a greater tendency toward overlap, the effect is variable across proteins. Similarly, the proportion of CAST-identified regions that overlap with SEG-identified regions, referred to as the overlap-to-CAST ratio, was analyzed. This metric assesses how much of the compositionally biased sequence space is also classified as low complexity. ALS-LLPS proteins showed a higher mean overlap-to-CAST ratio (30.56% vs. 19.41%, Median: 18.06% vs. 8.50%), suggesting a stronger involvement of CAST-identified regions in overlapping segments. While this trend suggests greater integration of CAST-identified regions within low-complexity sequences in ALS-LLPS proteins, the statistical significance was not important (p = 0.069).



Position Analysis

We analyzed the positional distribution of SEG- and CAST-identified regions using a normalized scale ranging from 0 (N-terminus) to 1 (C-terminus). The SEG-identified areas in non-ALS LLPS proteins were closer to uniform distribution (Mean: 0.471, Median: 0.460) compared to those in ALS-LLPS proteins that showed a small N-terminal bias (Mean: 0.443, Median: 0.419). This difference was not statistically significant (p=0.401).



Discussion

The above analysis showed distinctly different features between ALS-LLPS proteins and the control group. T-tests, displayed that ALS-LLPS proteins contain a higher number and greater coverage of both low-complexity regions (LCRs) and compositionally biased regions (CBRs), suggesting these features are enriched in disease-associated proteins. On the other hand, chi-square, showed that the presence or absence of these regions without checking anything else, was not statistically significant between the two groups. Sequence composition analysis demonstrated that ALS-LLPS proteins contain longer SEG-identified regions, a higher frequency of specific amino acid residues in these regions, and a reduced proportion of charged residues in CAST-identified regions. FInally, the overlap analysis further highlighted that ALS-LLPS proteins have significantly greater overlap coverage and frequency between SEG- and CAST-identified regions, suggesting an exchange between low-complexity and compositional bias.

References

- StatPearls. "Amyotrophic Lateral Sclerosis." ncbi.nlm.nih.gov

- Oxford Academic. "Pathogenesis of Amyotrophic Lateral Sclerosis." academic.oup.com
- MDPI. "Understanding Amyotrophic Lateral Sclerosis: Pathophysiology." mdpi.com
- Frontiers in Molecular Neuroscience. "Key Disease Mechanisms Linked to Amyotrophic Lateral Sclerosis." frontiersin.org
- PMC. "Amyotrophic Lateral Sclerosis: A Clinical Review." pmc.ncbi.nlm.nih.gov
- Frontiers in Neuroscience. "RNA Dysregulation in ALS." frontiersin.org
- Nature Neuroscience. "Excitotoxicity and Glutamate Transporters in ALS." nature.com
- Cell Reports. "Mitochondrial Dysfunction and ALS Progression." cell.com
- Journal of Neuroinflammation. "Neuroinflammation and ALS." jneuroinflammation.com
- Molecular Neurobiology. "The PI3K-Akt Pathway in ALS." springer.com
- Journal of Cell Science. "Autophagy and Proteostasis in ALS." jcs.biologists.org
- Brain Research. "Cytoskeletal Disruptions in ALS." sciencedirect.com