Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 02 July 2026
Revised 06 Aug 2026
Accepted 19 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (17s): 879–886

KEYWORDS:
Amyotrophic lateral sclerosis, alternative splicing, RNA-Seq, deep learning, CNN, RNN, unidirectional LSTM, BiLSTM and StringTie.

Investigating Alternative Splicing Events in Amyotrophic lateral Sclerosis with RNA-Sequence Analysis and Deep Learning.

A. Sathya[1], Harshita.M[2], LanceLot Lawra LI[2], Patan Sabiha[2], J. Jino Blessy[2*]

[1]Department of Artificial Intelligence and Data Analytics, Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Chennai 60016, India.
[2,*]Department of Bioinformatics, Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Chennai 60016, India.
Corresponding Author's Email ID: jinoblessy@sriramachandra.edu.in

ABSTRACT: Amyotrophic lateral sclerosis (ALS) is a progressive chronic neurodegenerative disorder caused by the degeneration of motor neurons and associated with dysregulated RNA metabolism, specifically alternative splicing. Investigation into AS abnormalities may help illuminate the molecular mechanisms of the disease and identify new transcriptomic biomarkers. RNA-Seq data from ALS and normal individuals were compared using a computational workflow that included HISAT2 for alignment, FeatureCounts for read counting, and DESeq2 for differential gene expression analysis, and StringTie with BEDTools for transcript assembly and analysis. Extracted AS associated transcript sequences were classified using the deep learning models CNN, RNN, unidirectional LSTM, and BiLSTM. Differential expression analysis revealed 349 statistically significant dysregulated genes. The performance of the models was assessed and found to be strongest for BiLSTM which achieved an accuracy and F1-score of 81% in differentiating ALS samples from normal individuals on the basis of transcript sequences, outperforming the CNN, RNN, and unidirectional LSTM architectures. This suggests the BiLSTM model can effectively account for sequence dependencies and distinguish between features that identify AS associated transcript sequences. A deep learning approach using a BiLSTM framework offers a scalable method to identify alternatively spliced transcripts and potential transcriptomic biomarkers in ALS. The model requires external validation on large independent datasets and combination with multi-omics data to further improve prediction accuracy and biological insight.

1. INTRODUCTION

ALS is a fatal neurodegenerative disease, characterized by a selective loss of motor neurons in the brain and spinal cord, which leads to the gradual development of progressive muscle weakness, atrophy and ultimately paralysis [1]. The prevalence of ALS is between 2-3 per 100,000 population per year, slightly higher in men than women [2]. The 90-95% of cases are sporadic (SALS), while 5-10% are familial and are generally caused by mutations in genes such as SOD1, TARDBP, FUS and C9orf72 [3]. Age of onset is between 40 and 70, and because the diagnosis is variable and the presentation may overlap with many other neuromuscular diseases the mean age of diagnosis is usually delayed. The common features of presentations include weakness of the limbs, dysarthria or bulbar palsy, and death typically occurs within 3 to 5 years of onset as a result of respiratory failure [4]. The cause of ALS is not yet fully understood but involves mechanisms such as glutamate excitotoxicity, oxidative stress, mitochondrial dysfunction, defective axonal transport, neuroinflammation and altered RNA metabolism [5]. Modern NGS

technologies, RNA-Seq and single-cell transcriptomics have recently elucidated gene expression profiles, alternate splicing events and molecular networks that cause motor neuron death in ALS [6-7].

Particularly, misregulation of splicing factors and RNA binding proteins such as FUS and TDP-43 is pathologically important due to their critical role in processing neuronal transcripts necessary for cell survival and function [8]. The approved drugs Riluzole and Edaravone only have modest benefits in decelerating disease progression, and cannot cure ALS patients [9]. New therapeutic approaches such as antisense oligonucleotide treatments, gene therapy, and immunotherapy are aiming at correcting genetic defects or targeting signalling pathways that drive the diseases [10]. The new integrative technologies, transcriptomics and proteomics can lead to the discovery of DEGs, PPIs and signalling pathways related to ALS [11].

More recent studies also suggest the ubiquitin-proteasome system, autophagy and immune response pathways to be associated with ALS neurodegeneration and hence potential therapeutic targets. The PI3K/AKT/mTOR signalling pathway, a central regulator of cell survival and metabolism, has been considered to be a major hub for ALS neurodegeneration and studies are focused on identifying its contribution to motor neuron degeneration [12]. Through the use of high-throughput RNA-seq data and different computational tools, we intended to identify significant DEGs, splicing patterns and associated pathways, which can serve as diagnosis markers.

2. MATERIALS AND METHODS

2.1 Data Collection

Two RNA-seq data sets were downloaded from the Gene Expression Omnibus (GEO): GSE268090 and GSE217625. The two datasets are from ALS and control group respectively, and all samples were performed with high-depth paired-end sequencing. A total of 14 samples (7 ALS, 7 control) were selected (Table 1).

Table 1. Next generation RNA sequencing samples for ALS control and case from GEO database.

GEO DATASET	SAMPLE TYPE
SRR29124587	Control
SRR29124588	Control
SRR29124589	Control
SRR29124591	Control
SRR22262474	Control
SRR22262484	Control
SRR22262486	Control
SRR29124599	case
SRR29124601	case
SRR29124603	case
SRR29124604	case
SRR22262476	case
SRR22262480	case
SRR22262498	case

2.2 Data Pre-processing

The raw RNA-seq reads were subjected to a number of pre-processing steps to check the quality and ready to be followed analysis. Quality Assessment: A fastqc was employed to examine the quality of the raw FASTQlines, the parameters like bases quality score, GC percentage, sequence duplication position and appendix contamination were inspected. Trimming: A trimmomatic was employed to remove the poor quality reads and appendix sequences, and parameters were tuned to ensure the read length longer than 36 bp on average quality score beyond Q20. Alignment: Clean reads were aligned to human genome (GRCh38) using HISAT2 was chosen based on its speed and ability to splice-align.

2.3 Discovery of alternative Splicing Events

ASEs was discovered using both statistical and visualization tools paraphrase assembly and quantification StringTie was used to convert aligned reads to transcripts and quantify gene expression. StringTie generates GTF files with paraphrases, which are used for the discovery of splice events. Differential Expression Analysis Feature Counts was used to generate count matrices and DESeq2 was used to determine differentially expressed genes and transcripts between control and ALS groups. Genes that had an adjusted ($|\log2FC|$ 1.5, $p < 0.05$) were considered significant. Visualization NanoPlot provided an overview and representation of the distribution and quality of read length and enabled to check transcriptome integrity. Splicing event discovery GFF files of StringTie was parsed with custom scripts and bedtools commands. Utility such as bedtools intersect, merge, flankBed, and getfasta were used to extract exonic, intronic and splice site information. Typical splice events like exon skipping, alternative donor/acceptor, and intron retention were then identified.

2.4 Point Engineering and Data Preparation

The features that were pulled out of splice events were then formalized in preparation for model training. Strand data, paraphrase ID and splice junction position were then added. The latter was split into training and test sets with appropriate normalization and one-hot garbling applied for categorical. Sequences were padded or cropped to equal length for compliance with neural network input requirements.

2.5 Model Development

To classify and predict critical splicing events, four deep literacy infrastructures were designed. These are: LSTM: LSTM armature was utilized because it can remember long-term dependency in succession biological information. LSTM networks re-read sequences of nucleotides and extract temporal patterns in splicing junctions. CNN: CNNs are utilized to extract localized motifs and position dependencies in sequences. This helped to identify the conserved patterns in splice sites. BiLSTM is an advanced variant of RNN which processes input in both forward and backward direction. BiLSTM is an extension where the sequence is read in both directions (forward and backward). This captures context from both past and future states in a sequence. This results in improved performance for many NLP tasks like language modelling and sequence classification. The cross-entropy loss function and the Adam optimizer were utilized to train all the models. All the models were trained on 70:30 train-test split. Performance metrics used for evaluation of these models include accuracy, precision, recall, F1-score, and support.

3. RESULTS AND DISCUSSION

Data Preprocessing and Alignment

All 14 RNA-Seq samples (7 treatments and 7 controls) were obtained from the NCBI GEO database. FastQC was performed to evaluate base quality and GC content for RNA-Seq samples. The quality assessment of the sequencing data. The key sequence metrics such as the duplication rate, GC content and length distribution for reads. The forward reads have a duplication rate of 55.0% and a GC content of 51.0% with a median read length of 151 bp and a sequence amount of 13.0 million reads. The reverse reads have a duplication rate of 34.3% and GC content of 50.0% with a median read length of 59 bp and a total of 21.6 million reads. The similarity in the GC content of forward and reverse reads (~50%) indicates that the overall sequences had no significant nucleotide bias and good GC content. However, the high duplication level in the forward reads indicates possible over-sequencing or PCR duplicates. The per-sequence GC content plot of the FastQC results appears to show normal distributions for both samples and good overlap between the samples with small deviation of peaks between samples. Reads were trimmed using Trimmomatic. Finally, reads were mapped to the human genome reference using HISAT2 with mapping efficiency >90% for almost all the samples.

Natural Resources for Human Health
eISSN: 2583-1194

Gene Expression and Splicing Analysis

Differential gene expression analysis was performed using DESeq2 and revealed a plethora of significantly dysregulated genes between treatment and control samples. MA plot clearly illustrates both up and down regulated groups of genes (Figure 1). The samples were clearly distinguishable between treatment and control groups in a PCA plot. PCA of the RNA-seq data shows that the majority of the samples can be divided clearly according to the first principal component (PC1) that explains 82% of total variation and the second principal component (PC2) explain another 8%. The bulk of the samples separates clearly according to PC1 so we conclude that there is considerable variation in the transcriptional output of the samples. There is a second distinct cluster. Two clear groups are visible; one grouping on the negative region of the PC1 spectrum, while the other on the positive. Case and controls appear in both clusters so we hypothesize that the major sources of variance are probably not solely due to the experimental condition. Within each group samples are very tightly clustered together, which is testament to the excellent repeatability and small intra-group variance. One interpretation of a second cluster, however, is that batch effects, technical variation or undetected biological variation could be present (Figure 1).

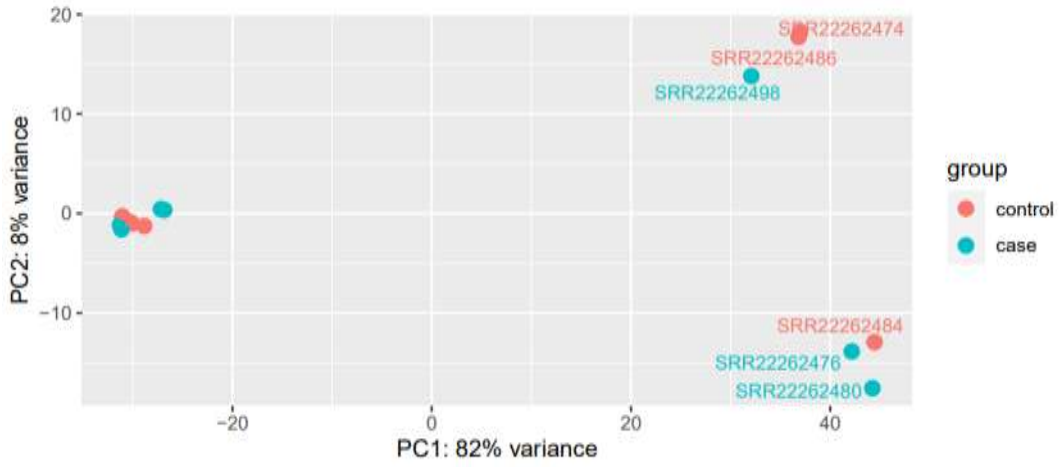


Figure 1: A group of variance for ALS sample.

The sample-to-sample distance heatmap shows the similarity between all RNA-Seq samples, based on their gene expression values. The colour of the boxes represent the pairwise distances between two samples, with the darkest blue indicating similarity and light colours representing distances between samples. The hierarchical clustering dendrogram on the top side shows a visual representation of grouping the samples according to their distance. Each group of samples is highly similar to each other, as displayed in the dark blue squares along the main diagonal, which represents the consistency of the within-group expression profile. Between the two groups of samples, the distances are significantly higher, thus indicating that these two groups of samples are very different in expression profile, likely resulting from either the biological variability, experimental conditions or possible batch effect, as suggested by two distinct groups in the PCA plot and dendrogram (Figure 2). Alternative splicing events were determined using stringtie, the following types of AS events were detected: Exon skipping (SE), Intron retention (IR) and Alternative 5'/3' splice sites (A5SS/A3SS).These alternative splicing events were significantly enriched in neural development, synaptic signalling and cytoskeleton reorganization pathways.

Natural Resources for Human Health
eISSN: 2583-1194

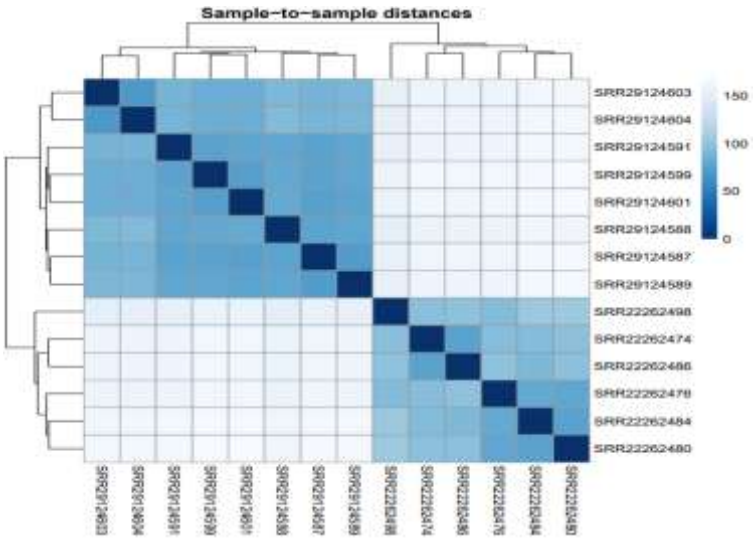


Figure 2: PCA Shows Sample to sample distances for control and cases of ALS disease.

Deep Learning-Based Classification of Splicing Events

The one-hot encoded FASTA sequences of the determined splicing regions were subsequently provided to four deep learning models, whose performance were measured with conventional metrics (accuracy, precision, recall and F1-score). Different deep learning models were evaluated for the prediction of ALS-related AS events, measured by accuracy, precision, recall, and F1-score (Figure 3). Out of the five models evaluated, the BiLSTM model showed best results of 81% accuracy, 82% precision, 81% recall and 81% F1-score (with support = 37 samples). The CNN model's results were fair at 78% accuracy, 78% precision, 78% recall and 78% F1-score (with support = 68 samples). Slightly lower was the LSTM at 73% in all measured metrics (accuracy, precision, recall and F1-score) (with support = 68 samples). The hybrid RNN + BiLSTM model resulted in worst prediction among the tested models, and showed an accuracy of 70%, 71% precision, 70% recall and 69% F1-score (with support = 68 samples). From these findings it appears the BiLSTM model has the best potential in capturing the sequence dependencies and contextual data required to predict ALS related alternative splicing events, due to its architecture, compared to the other methods implemented. To further understand the performance of the best model that is the BiLSTM model, we consider its confusion matrix (Figure 4). In this figure it shows that the model has correctly classified 15 cases of Class 0 and 13 cases of Class 1. The model has predicted 5 false positive for class 0 and 4 false negative for Class 1. The accuracy calculated is approximately 75.7 %. The confusion matrix states the model has a good classification with 15 false positive cases and 4 false negative cases and good overall prediction for Class 0 and Class 1 (Figure 4).

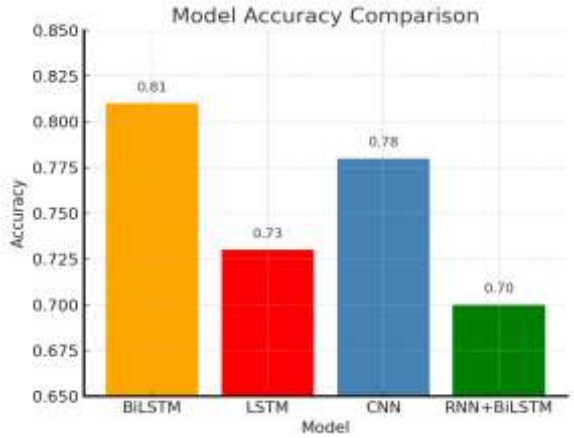


Figure 3: Bar chart shows models accuracy comparison BiLSTM, LSTM, CNN and RNN+BiLSTM for ALS related alternative splicing events.

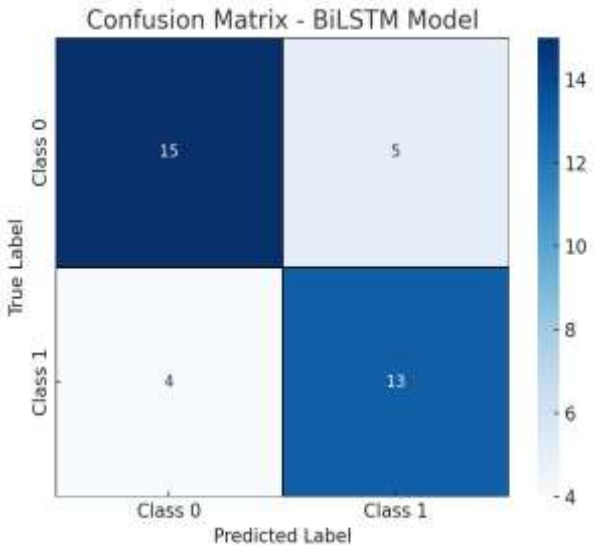


Figure 4: Confusion matrix using BiLSTM

3.4 Discussion

We performed integrative RNA-Seq data processing, differential gene expression analysis, alternative splicing detection, and deep learning-based classification of alternative splicing to analyse the transcriptomic changes in ALS. We found important discoveries in disease progression molecular mechanism and alternative splicing prediction using computational models. The data quality was first examined. High base quality scores and regular GC content distribution (50-55%) of both RNA-Seq samples were confirmed. High alignment rates (>90%) to reference genome were also reported [13].

Thousands of dysregulated genes were identified between treated and control group, which represents significant transcriptomic remodelling due to ALS. MA plot shows the distribution of upregulated and downregulated genes which is expected due to alternation of cell functions. Furthermore, treated group clusters distinctly separate from control group as shown by PCA plot which implies that these differences are biological and not artifacts. This is consistent with findings from previous studies where dysregulated gene expression was commonly observed in neurodegenerative diseases, including ALS [14-15].

One of the most important steps of this study was to discover alternative splicing events in ALS patients. There were various categories of alternative splicing events identified using StringTie, including SE, IR, and alternative splice site usage (A5SS and A3SS). Exon skipping events were the most abundant among all the alternative splicing events discovered which is a well-known characteristic of the human genome [16]. A large portion of alternative splicing events occurred in genes that are implicated in neural development, synaptic function and cytoskeletal organization. Those biological processes are essential for neuronal function and cell integrity which are commonly affected in ALS. Specifically, disturbance of cytoskeletal dynamics and neuronal signalling are known hallmarks of motor neuron diseases [17-18].

To investigate the potential of sequence information for predicting ALS related alternative splicing, we developed various deep learning models to predict the splicing events from RNA sequence. Among all models used BiLSTM gained the best accuracy, precision, recall and F1 score as 81% accuracy, 82% precision, 81% recall and 81% F1 score. The reason may attribute to the nature of BiLSTM that can capture information from both forward and reverse direction and learn sequences which is important since there may be sequence dependency among different positions in the splicing site. CNN is also performing well by capturing local motifs, however, it still lower than BiLSTM may mean that local information is not sufficient to predict splicing event and bidirectional information are more informative for alternative splicing prediction. LSTM is also worse than BiLSTM in prediction, it is likely because of limited information it learns from single direction sequence. The hybrid RNN + BiLSTM model gives the worse result possibly because of increased complexity. Previous RNA-seq big data can serve as a knowledge resource of splicing regulation through deep learning, to help researchers to understand alternative splicing using RNA-seq data more precisely with less coverage [19].

Overall, comparing the prediction results of various deep learning models indicate that exploiting the sequence context is highly efficient in predicting the alternative splicing events of ALS patients, especially the BiLSTM model. However, the present study has some limitations. The accuracy of the BiLSTM model suggests that it can adequately model sequences in the dataset as both classes are classified somewhat evenly. Through the use of a confusion matrix we can conclude that the model performs reasonably with an accuracy of 75.7% which is correctly classifying the bulk of Class 0 and Class 1, with relatively few false positives or negatives. Some of these false positives and negatives arise as the characteristics of Class 0 and Class 1 are somewhat overlapping due to shared similarities and potentially biological variation. The model has a tendency to predict with a good level of consistency though and has reasonable overall predictive accuracy. Moreover, the detection of biological informative alternative splicing events and the performance of BiLSTM suggests that computational approaches have potential to assist further discovery in neurodegenerative disease mechanisms.

4. CONCLUSION

This study provided a comprehensive transcriptomic analysis in order to study the alternative splicing in ALS. We have combined the RNA-seq data analysis, differential gene expression analysis, and the deep learning based classification approach. A high quality sequencing data with uniform GC content and the high alignment rate makes downstream analysis to be reliable. There were the obvious transcriptional differences between the treated group and the control group in terms of the differential gene expression and the result from PCA plot, indicating the changes in terms of molecular pathway related to the disease. In the analysis of alternative splicing, we have found that several event types were overrepresented in genes related to neural development, synaptic transmission and cytoskeleton organization, which include exon skipping, intron retention and alternative splice site usage. These indicate that splicing dysregulation was highly important in the development of ALS. Also, when apply deep learning models to classify the samples based on the splicing events, BiLSTM model performed significantly better than the other models to predict the alternative splicing related to ALS. The result indicate that it is very important to capture both directions sequence dependency, and the CNN and LSTM model had a moderate performance. And it shows the contribution that computational models can make in identifying the molecular mechanisms involved in ALS neurodegenerative diseases.

ACKNOWLEDGEMENTS

We thank the Department of Bioinformatics, Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research for their computational facilities.

AUTHOR CONTRIBUTIONS

Harshita: Methodology, Curation, Formal Analysis & Laboratory Analysis.

J. Jino Blessy: Conceptualization, Investigation, Supervision, Writing—Review & Editing.

A. Sathya: Writing—Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

All data generated and used during the current study are from publically available resources, i.e., NCBI Gene Expression Omnibus (GEO) with accession numbers cited in the manuscript.

REFERENCES

[1] Brown, R.H., Al-Chalabi, A., 2017. Amyotrophic lateral sclerosis. New England Journal of Medicine 377(2), 162–172.

[2] Ohi, T., Mitsui, J., Shatunov, A., Jones, A.R., Iacoangeli, A., Al-Chalabi, A., et al., 2013. Genetic epidemiology of amyotrophic lateral sclerosis: A study of 149 cases. Brain 136(6), 1693–1701.

[3] Renton, A.E., Chio, A., Traynor, B.J., 2014. State of play in amyotrophic lateral sclerosis genetics. Nature Neuroscience 17(1), 17–23.

[4] Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E.M., Logroscino, G., Robberecht, W., et al., 2017. Amyotrophic lateral sclerosis. Nature Reviews Disease Primers 3, 17071.

[5] Taylor, J.P., Brown, R.H. Jr., Cleveland, D.W., 2016. Decoding ALS: From genes to mechanism. Nature 539(7628), 197–206.

[6] Mehta, A.R., Chataway, T., Collin, L., Bilsland, L.G., Houlden, H., Sibley, C.R., 2021. Molecular mechanisms of neurodegeneration in amyotrophic lateral sclerosis: Insights from genomics and transcriptomics. Nature Reviews Neurology 17(7), 433–448.

[7] Prudencio, M., Humphrey, J., Pickles, S., Brown, A.L., Hill, S.E., Kachergus, J.M., et al., 2020. Truncated STMN2 is a marker of TDP-43 pathology in frontotemporal dementia. Journal of Clinical Investigation 130(11), 6080–6092.

[8] Ling, S.C., Polymenidou, M., Cleveland, D.W., 2013. Converging mechanisms in ALS and FTD: Disrupted RNA and protein homeostasis. Neuron 79(3), 416–438.

[9] Jaiswal, M.K., 2019. Riluzole and edaravone: A tale of two amyotrophic lateral sclerosis drugs. Medicinal Research Reviews 39(2), 733–748.

[10] Miller, T.M., Pestronk, A., David, W., Rothstein, J.D., Simpson, E., Appel, S.H., et al., 2022. An antisense oligonucleotide against SOD1 delivered intrathecally for patients with SOD1 familial ALS: A phase 1, randomized, first-in-human study. Lancet Neurology 21(3), 231–240.

[11] Tam, O.H., Rozhkov, N.V., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., et al., 2019. Postmortem cortex samples identify distinct molecular subtypes of ALS: Implications for precision medicine. Neuron 102(6), 1050–1062.

[12] Perlson, E., Hanz, S., Ben-Yaakov, K., Segal-Ruder, Y., Seger, R., Fainzilber, M., 2005. Vimentin-dependent spatial translocation of an activated MAP kinase in injured nerve. Neuron 45(5), 715–726.

[13] Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: A fast spliced aligner with low memory requirements. Nature Methods 12(4), 357–360.

[14] Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(12), 550.

[15] Prudencio, M., Belzil, V.V., Batra, R., Ross, C.A., Gendron, T.F., Pregent, L.J., et al., 2015. Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. Nature Neuroscience 18(8), 1175–1182.

[16] Pan, Q., Shai, O., Lee, L.J., Frey, B.J., Blencowe, B.J., 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics 40(12), 1413–1415.

[17] Ling, J.P., Pletnikova, O., Troncoso, J.C., Wong, P.C., 2015. TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. Science 349(6248), 650–655.

[18] Tollervey, J.R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A.L., Župunski, V., Patani, R., 2011. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nature Neuroscience 14(4), 452–458.

[19] Zhang, Z., Pan, Z., Ying, Y., Xie, Z., Adhikari, S., Phillips, J., Carstens, R.P., Black, D.L., Wu, Y., Xing, Y., 2019. Deep-learning augmented RNA-seq analysis of transcript splicing. Nature Methods 16(4), 307–310.