Computational analysis of RNA expression among survival groups in amyotrophic lateral sclerosis

Sara Baldinelli[1]

[1] *Department of Cellular, Computational and Integrative Biology - CIBIO, University of Trento, Via Sommarive 14, 38123 Povo (TN), Italy*

Abstract

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease, familial in 10% of cases and targeting both upper and lower motor neurons, leading to a progressive loss of motor function. While the average life expectancy post-symptom onset ranges from 3 to 4 years, there exists a notable variability in survival times, with some individuals dying within months and others surviving for several years. To gain insight into the variability of survival time in ALS, we analysed RNA expression profiles of lymphoblastoid cell lines derived from 42 ALS patients (GEO accession GSE212131), of which 22 with short disease duration (less than 12 months) and 20 with long disease duration (over 6 years). Of the four different classifiers that we tested on a list of 691 differentially expressed genes (p-value < 0.01 using a t-test), namely Linear Discriminant Analysis [1], LASSO [2], Random Forest [3] and Ridge [4], Random Forest performed the best with a mean accuracy of 0.735. The functional analysis of the top 200 genes by importance according to the Random Forest revealed a number of pathways and reactions involved in the differential survival outcome, including spliceosome pathway activity (p-value = 1.0e-05) and metabolism of RNA (p-value = 1.3e-02). We are currently expanding the analysis by adding miRNA profiles for the same set of patients. These findings may help in the discovery and exploration of new potential biomarkers aimed at expanding our knowledge on ALS and at identifying potential therapeutic targets that may enhance the quality of life of people affected by this pathology.

Introduction

Amyotrophic lateral sclerosis (ALS) is a neural disorder characterized by the degeneration of motor neurons, loss of voluntary muscle control and premature mortality. While ALS is primarily considered a sporadic condition, approximately 10% of cases exhibit familial inheritance patterns, underlining the complex interplay of genetic and environmental factors in its pathogenesis [5]. One of the key aspects of ALS lies in its heterogeneity, evident in the spectrum of survival times observed among affected individuals. This not only makes it harder to find a prognosis but also underscores the need for a deeper understanding of the underlying molecular mechanisms driving disease progression and survival outcomes. A comprehensive analysis was performed on transcriptomic data from 42 ALS patients, aiming at getting insights in the gene expression patterns related with disease duration. This dataset, obtained from the Gene Expression Omnibus (GEO) under accession number GSE212131 [6], is splitted based on patients' disease duration (less than 12 months or more than 6 years). Among the five classifiers tested, Random Forest emerged as the optimal model, exhibiting a mean accuracy of approximately 73.5%. This model highlighted a set of top-ranking genes, which were then subjected to functional enrichment analysis. This analysis revealed significant biological pathways and molecular cascades implicated in disease progression and survival dynamics, with spliceosome activity, RNA metabolism, antigen processing and presentation, the IL-17 signaling pathway, and parathyroid hormone synthesis, secretion, and action, emerging as critical pathways in the molecular landscape of ALS duration variability.

1. Materials and Methods

1.1 Dataset

The dataset used in this analysis was taken from GEO database (accession number: GSE212131) [6], it contains transcriptomic data extracted from lymphoblastoid cell lines of 42 ALS patients, processed through Affymetrix microarray technology with a panel of 22011 variables (genes). The samples are splitted in two groups, one comprising 22 individuals with abbreviated disease duration ($<$ 12 months) and the other including patients exhibiting prolonged survival ($>$ 6 years).

1.2 ggplot2

ggplot2 (version 3.5.1) was utilized in order to visualize the accuracy scores for each algorithm employed and to get visual support in its choice. *ggplot2* [7] is an open-source data visualization package for creating graphics, based on the *Grammar of*

Graphics, a comprehensive framework for data visualization that breaks up graphs into semantic components, including scales and layers.

1.3 plotly
Through the *plotly* library (version 4.10.4) we were able to plot both the 2D and 3D interactive PCA using the `plot_ly()` function and coloring by group labels [8].

1.4 caret
The *caret* (Classification And REgression Training) package (version 6.0-94) is a set of functions that attempt to streamline the process for creating predictive models [9].

1.5 rScudo
SCUDO (Signature-based Clustering for Diagnostic Purposes) is a rank-based method for the analysis of gene expression profiles for diagnostic and classification purposes [10]. This package (version 1.18.1), based on the identification of sample-specific gene signatures composed of the most up- and down-regulated genes for that sample, was used to identify sample-specific gene signatures starting from gene expression data, and use them to build a graph of samples.

1.6 igraph
igraph [9] is a collection of network analysis tools with the emphasis on efficiency, portability and ease of use. This package (version 2.0.3) was utilized in our analysis to plot different network graphs.

1.7 gprofiler2
gprofiler2 [11] (version 0.2.3) is a bioinformatics tool that performs functional enrichment analysis of gene and pathway lists. We utilized it for the identification of enriched biological terms, pathways, and processes and visualized the results through the `ghostplot()` function.

1.8 pathfindeR
pathfindeR (version 2.3.1) was used to perform a functional analysis and the results were plotted with the `term_gene_graph()` and `enrichment_chart()` functions. Indeed, *pathfindeR* [12] is an R package designed for pathway analysis and visualization. It offers functionalities for identifying enriched pathways and biological processes based on gene expression data, facilitating the interpretation of molecular mechanisms of certain phenotypes or diseases.

1.9 Methods
In this study, we employed different methodologies to get meaningful insights from the transcriptomic data, taking advantadge of both supervised and unsupervised approaches. After an initial exploratory analysis, trough boxplots and principal component analysis, we employed diverse clustering algorithms, including K-means and hierarchical clustering, with a predetermined k-value of 2 due to the binary nature of our dataset: 'Short' labeling individuals with a more aggressive disease phenotype, and 'Long' representing those with a more indolent disease course. In particular, for what concerns the k-means analysis, we set a seed of 2900; specific parameters for the hierarchical clustering, instead, are the different methods that can be utilized, in this study 3 different ones were tested: the average, single and complete methods. After data preprocessing, we partitioned the dataset into training and test sets to facilitate robust model evaluation. This was done by sampling through the usage of the `sample()` and `setdiff()` functions. Comparing a set of machine learning algorithms, including Random Forest, Ridge regression, Lasso regression, Linear Discriminant Analysis, and Scudo, we tried to discover patterns predictive of survival outcomes in ALS patients. RF, LASSO, LDA and RIDGE classification were performed trough the `train()` function belonging to the *caret* package and, through the resampling of the respective results with the `resamples()` function, their accuracy was compared. The Scudo classification, instead, was performed thorugh the *rScudo* package, using the following functions: `scudoTrain()`, `scudoTest`, `scudoNetwork()` scudoPlot(trainNet, vertex.label = NA), and `scudoClassify()` in order to respectively perform the training, validate with the test and perform the classification. To identify clusters on map, `igraph::cluster_spinglass()` function was employed and the object obtained was plotted. Moreover, we conducted an enrichment analysis using *gprofiler2* to find enriched biological pathways underlying disease progression. To do so, the `gost()` function was used to perform a query starting from the gene list of most important gene obtained with the supervised analysis. Then, to visualize the results of the query in a Manhattan plot, we utilized the `gostplot()` function. These selected genes were subjected to gene-based network analysis using *PathfindeR*, getting insights into the molecular interactome governing ALS survival variability. To visualize the results of the network-based analysis, obtained through `run_pathfindR()`, we employed the `term_gene_graph()` and the `enrichment_chart()` functions. Finally, a literature search was performed in order to contextualize our findings and identify potential avenues for further exploration.

2. Results

2.1 Exploratory Analysis

The dataset consisted of mRNA expression profiles from ALS patients, divided into two groups: Short and Long. After importing the data, a boxplot was generated to visualize the distribution of gene expression levels across samples (**Figure 1**). It provided an overview of the data's spread and helped identify potential outliers.

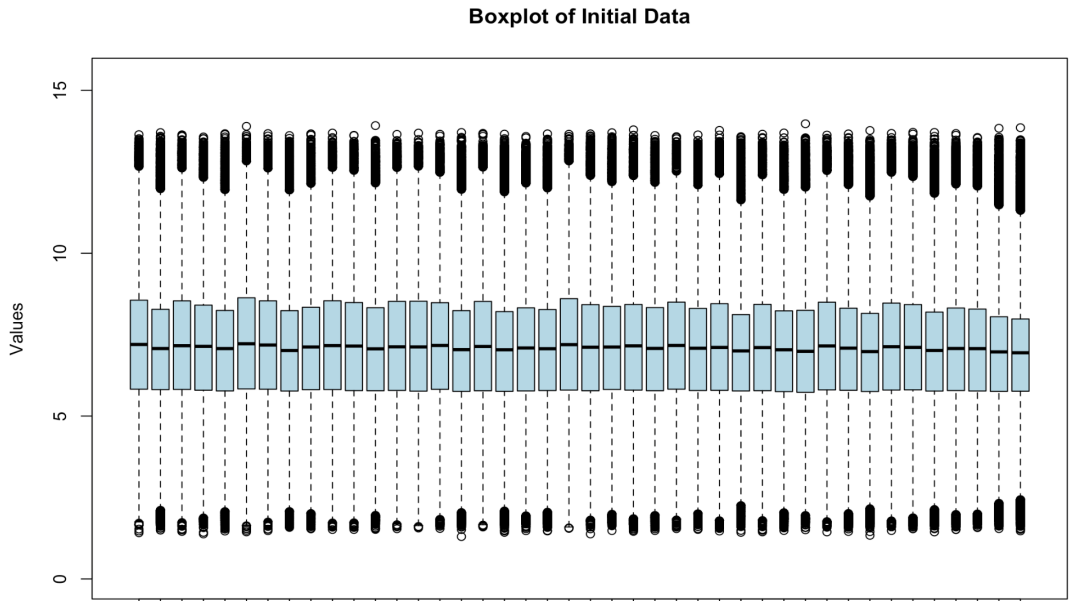


Figure 1. Boxplot of initial values per sample (x-axis): since median and variance looked well aligned, no further process of normalization has been done.

As an initial preprocessing step, we filtered out probes with insignificant differences between groups (p-value > 0.01), with a t-test obtaining a set of 691 genes. This filtering criteria should allow me to reduce significantly the amount of probes involved in the microarray without reducing too much the dataset. Since the next steps consists also in a selection of significant genes starting from this filtering, we avoided being to stringent on the selection. This is also the reason why we decided not to go for more advanced method nor multiple hypothesis correction because in this way the test would have produced highly selected results. PCA was then plotted to reduce the dimensionality of the dataset and visualize sample clustering patterns, but both 2D and 3D (**Figure 2**) PCA plots didn't manage in a nice and clear separation between the Short and Long groups.

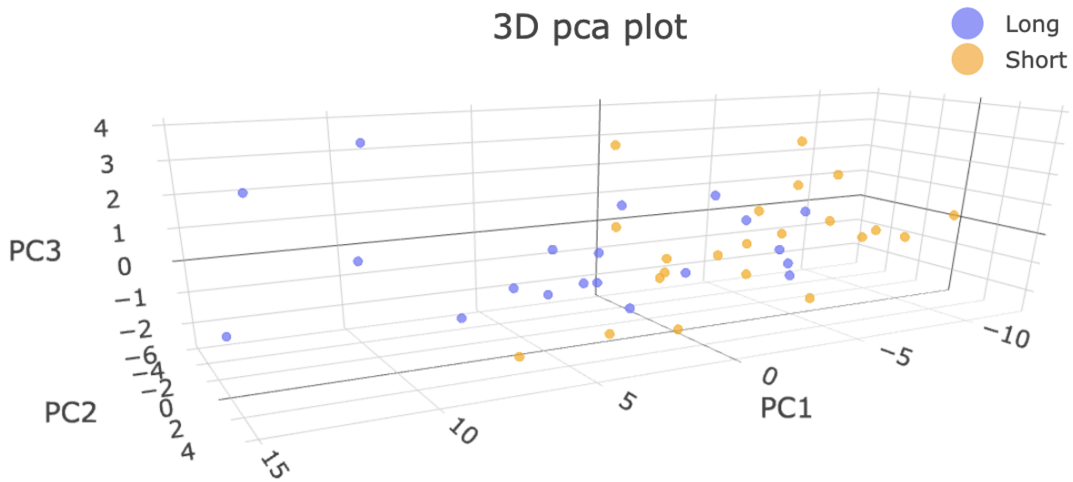


Figure 2. The plot represents the 3D PCA, apart from some samples which appear to be well separated, it is really hard to distinguish a pattern of division in the central region.

2.2 Clustering Analysis

Two clustering methods, k-means and hierarchical clustering, were utilized to identify natural groupings within the dataset. K-means clustering was employed to partition the samples into distinct groups based on their gene expression profiles. This unsupervised learning technique aims to identify clusters in the data by iteratively assigning each sample to the nearest centroid and updating the centroids until convergence. The analysis revealed the division of samples into short and long disease duration, but there was still an unclear distinction (**Figure 3**).

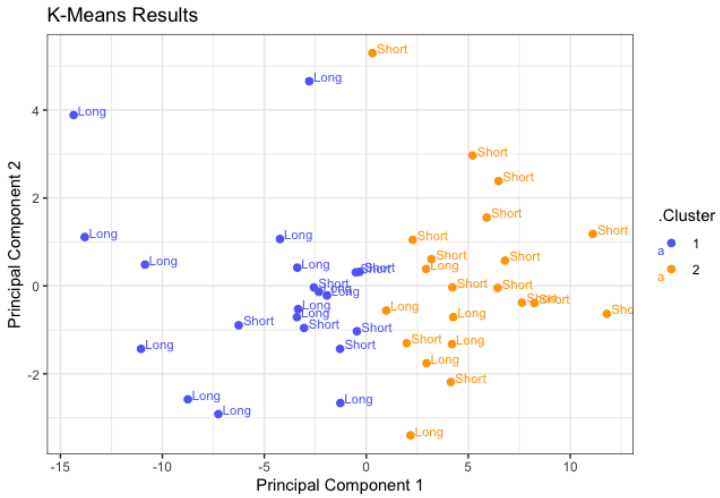


Figure 3. K-means clustering using seed = 2900, The labels assigned to each point in the 2D space correspond to the ground truth labels. This analysis revealed the division of some samples into the 2 groups, even though uncertainty in the separation can be noticed.

Hierarchical clustering, another unsupervised learning method, was utilized to further explore the structure of the data and assess the stability of clustering across different linkage methods, namely average, complete, and ward.D2. Each of them utilizes a different approach to measure the distance between clusters and merge them, resulting in varying cluster compositions and structures. Even though none of them managed to clearly distinguish the two clusters (the complete and ward.D2 methods are shown in **Figure S1** and **S2**, present in the supplementary (**Section 4**)), the average method helped in a separation of a 'Long' subcluster, as we can see from **Figure 4**. In particular, this method computes the average distance between all pairs of samples in two clusters and merges clusters with the smallest average distance.

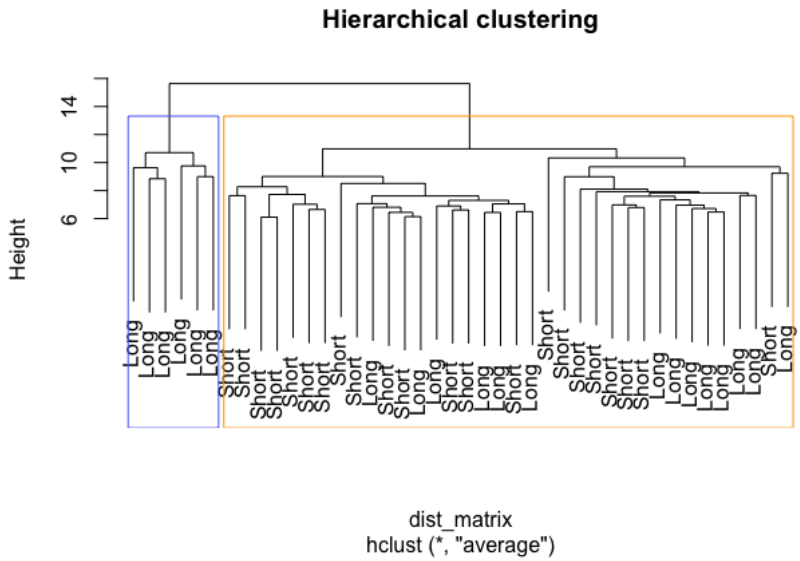


Figure 4. Hierarchical clustering exploiting average linkage. This method clusters the two groups after performing some initial branching. We can see a subcluster for the long survival group, but the other group stays highly noisy.

2.3 Supervised Analysis

With the aim of best classifying the dataset, several supervised machine learning algorithms, including Random Forest (RF), Linear Discriminant Analysis (LDA), LASSO and Ridge, were also employed. We decided to conduct a comparative analysis to evaluate their performance in distinguishing between ALS patient subgroups. Their accuracy was assessed using 10-fold cross-validation, providing the proportion of correctly classified samples out of the total samples. A graphical representation of the accuracy scores for each algorithm is presented in **Figure 5**, providing insights into their relative performance in distinguishing between ALS patient subgroups. The accuracies of the four algorithms are close (LASSO: 0.77, RF: 0.735, LDA: 0.715, RIDGE: 0.7), this is why our choice of picking RF (whose variable importance can be visualized in **Figure 6**) was based on the later performed enrichment analysis.

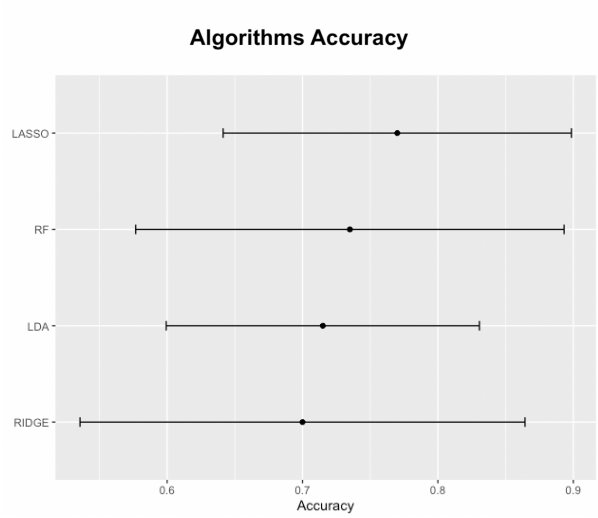


Figure 5. Accuracy differences for classification methods using cross validation available on caret. All the four methods have a similar accuracy with very high variability meaning that we cannot consider it as the only parameter to choose the best algorithm for classification. That is the reason why we went on with the analysis in order to look at the enrichment analysis with the different algorithms, arriving to the conclusion that the best option would be to choose RF.

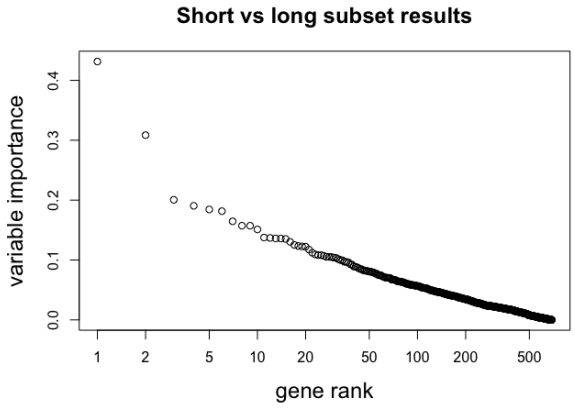


Figure 6. This plot shows in a decreasing order the gene rank for importance of random forest classification. As we already could have imagined by checking its accuracy, we can see that the performance is not the best, but still we decided to go on with the analysis to study just the most important genes, in particular the first 200.

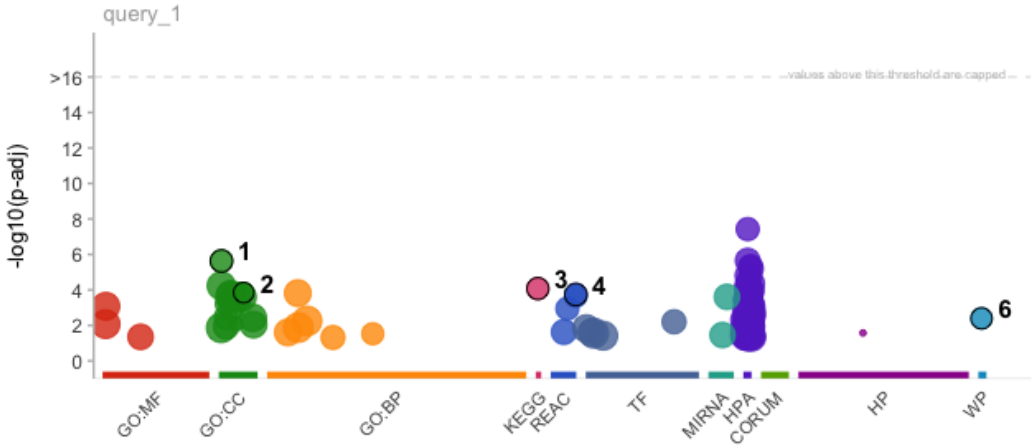
Beside these four algorithms also rScudo was also tested and the results scored around 0.62 of accuracy, reaching a result close to random with parameters nTop = nBottom = 25 and N = 0.4. Results of training and testing through rScudo are reported in the supplementary material (**Section 4, Figure S3 and S4**).

2.4 Enrichment Analysis

After having selected the best performing method for classification, we needed to convert the list of probe names mapping the genes to their corresponding HGNC symbols in order to perform the enrichment analysis. This was done through the use of the *biomaRt* library:

```
ensembl_Mart <- useMart("ensembl")
ensembl <- useEnsembl(biomart = "ensembl",
                      dataset = "hsapiens_gene_ensembl")
# look for conversion table for probes
ensembl_convertion_table <- getBM(attributes = c('affy_huex_1_0_st_v2', 'hgnc_symbol'),
      filters = 'affy_huex_1_0_st_v2',
      values = rf_importance_list,
      mart = ensembl)
```

Then, significant pathways and gene ontology terms related to ALS were identified and visualized through the usage of the *gprofiler2* package. As it can be seen from **Figure 7**, the most enriched terms (p-values ranging from 6.5e-03 to 5.1e-07) were all regarding the same biological process: mRNA splicing, suggesting its potential involvement in ALS pathology.



id	source	term_id	term_name	term_size	p_value
1	GO:CC	GO:0005681	spliceosomal complex	206	2.3e-06
2	GO:CC	GO:0071013	catalytic step 2 spliceosome	89	1.4e-04
3	KEGG	KEGG:03040	Spliceosome	174	8.4e-05
4	REAC	REAC:R-HSA-72163	mRNA Splicing - Major Pathway	202	1.8e-04
6	WP	WP:WP411	mRNA processing	126	4.0e-03

g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 7. gProfiler enrichment analysis results for the top 200 genes obtained from random forest for classification. We can see from this plot that we get many results, and the ones with the highest p-values are all about the splicing process as we can notice from the table.

2.5 Network Analysis

As last step we performed a network-based analysis through *pathfindR*, these tools allow us to discover new potential interactors or enriched pathways via including neighbours of the list of genes that we give them as input. This is done by considering links present in literature and performing the analysis on an enriched list. Since the set of genes will be expanded by the tools, we considered subsets of different length of the list (100, 200 and 300), still, the most relevant results for pathfindR have been obtained with the length of 200. For this reason we kept the same list for all approaches to increase coherence and allow a more fair comparison. For each selected gene, we computed the p-value using a two-sample t-test to assess the significance of gene expression differences between ALS patient subgroups. We then utilized the *pathfindR* package to perform pathway analysis based on the calculated p-values. Enriched terms and associated genes are visualized using network graphs (**Figure 8**) and the enrichment chart (**Figure 9**), allowing for the identification of relevant biological pathways and genes implicated in ALS heterogeneity.

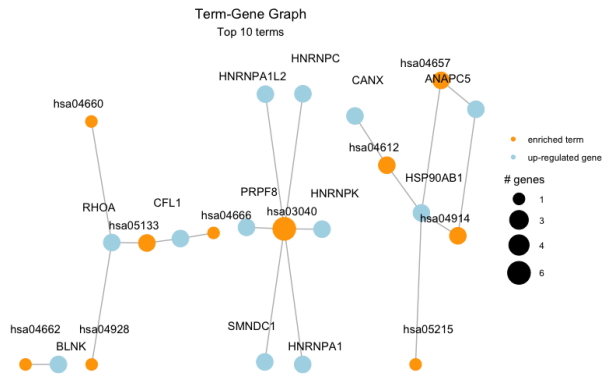


Figure 8. pathfindR network of interactions. This network have been achieved giving as input the list of the 200 most important genes for the RF classification.

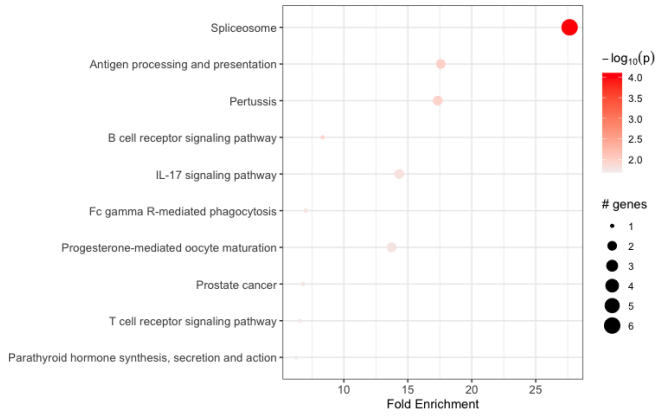


Figure 9. pathfindR enrichment chart obtained giving as input the pathfindR_results from the `run_pathfindR()` function.

Among all the enriched genes we can identify in **Figure 7**, we focused on the heterogeneous nuclear ribonucleoproteins (hnRNPs), a large family of RNA-binding proteins (RBPs) that contribute to multiple aspects of nucleic acid metabolism including alternative splicing, mRNA stabilization, and transcriptional and translational regulation [13]. At this point, through both *gprofiler2* (**Subsection 2.4**) enrichment and *pathfindR* network, we have two hints suggesting the potential involvement of splicing in ALS disease duration; but we can also confirm this result through the enrichment chart, (**Figure 9**): the most enriched term is, indeed, the *Spliceosome*. Beside the splicing mechanism, in the enrichment chart we can visualize other terms, such as *Antigen processing and presentations, Pertussis, IL-17 signaling pathway* and *Parathyroid hormone synthesis, secretion and action*, suggesting their involvement in the shorter duration of ALS disease.

3. Discussion

The computational analysis of RNA expression among survival groups in ALS provides insights into the molecular mechanisms involved in the variability in survival times observed in ALS patients. Through the integration of transcriptomic data and machine learning techniques, we aimed to unravel the gene expression patterns associated with disease duration and identify potential biomarkers and therapeutic targets for ALS. Our study utilized a dataset obtained from the GEO database containing transcriptomic data from lymphoblastoid cell lines of 42 ALS patients, divided into two groups based on disease duration: short ($<$ 12 months) and long ($>$ 6 years). After preprocessing the data and filtering out insignificant probes, we identified 691 differentially expressed genes using a t-test (p-value $<$ 0.01). Subsequently, we employed both supervised and unsupervised learning approaches to analyze the data.

Exploratory analysis revealed the distribution of gene expression levels across samples and provided an initial understanding of the dataset's characteristics. However, the PCA did not yield clear separation between short and long disease duration groups, indicating the complexity of ALS heterogeneity.

Clustering analysis using k-means and hierarchical clustering methods further explored the structure of the data. While k-means clustering showed some separation between short and long disease duration groups, hierarchical clustering using the average linkage method managed to identify a distinct subgroup within the long disease duration category.

Supervised machine learning algorithms, including Random Forest, LDA, LASSO, and Ridge regression, were employed to classify ALS patient subgroups based on gene expression profiles. Random Forest emerged as the optimal model, with a mean accuracy of 73.5%. The selection of Random Forest for further analysis was based on its performance and subsequent enrichment analysis results.

The enrichment analysis, revealed a strong association between ALS and mRNA splicing processes. The most enriched terms, identified using the *gprofiler2* package, consistently pointed to mRNA splicing (Figure 7). This finding is particularly noteworthy given the existing literature that implicates RNA metabolism disruptions in neurodegenerative diseases, including ALS and the p-values of the enriched terms, ranging from 6.5e-03 to 5.1e-07, underscore the statistical significance of these associations. Indeed, splicing is a tightly orchestrated process by which the brain produces protein diversity over time and space. While this process specializes and diversifies neurons, its deregulation may be responsible for their selective degeneration. In amyotrophic lateral sclerosis (ALS), splicing defects have been investigated at both singular gene level and higher-order level, involving the entire splicing machinery[14, 15]. In a study [15], the complete spectrum (396) of genes encoding splicing factors in the motor cortex (41) and spinal cord (40) samples from control and sporadic ALS (SALS) patients were analyzed. A substantial number of genes (184) displayed significant expression changes in tissue types or disease states, were implicated in distinct splicing complexes and showed different topological hierarchical roles based on protein–protein interactions.

Complementary to the enrichment analysis, the network analysis using *pathfindR* further elucidated the involvement of specific genes and pathways in ALS. By expanding the gene set through literature-based links and analyzing subsets of different lengths, we identified robust results particularly with a list of 200 genes. This consistent subset size across analyses ensured coherence and facilitated a fair comparison of findings. The two-sample t-tests conducted for each selected gene highlighted significant differences in gene expression between ALS patient subgroups, which were then utilized in the *pathfindR* pathway analysis. The visualization of enriched terms and associated genes through network graphs and enrichment charts (Figures 8 and 9) allowed us to pinpoint relevant biological pathways. Notably, the enrichment chart confirmed the involvement of the spliceosome, reinforcing the results obtained from *gprofiler2* and suggesting a critical role of RNA splicing deregulation in ALS pathogenesis.

Among the enriched genes, the hnRNPs emerged as significant. hnRNPs are a family of RNA-binding proteins involved in various aspects of nucleic acid metabolism, including alternative splicing and mRNA stabilization. Dysfunctions in the function of hnRNPs have been closely linked to neurodegenerative diseases, most prominently ALS and frontotemporal dementia (FTD), two diseases with significant genetic and pathological overlap [16].

Beyond splicing, the enrichment chart highlighted other significant pathways, such as antigen processing and presentation, the IL-17 signaling pathway, and parathyroid hormone synthesis, secretion, and action. Beyond splicing, the enrichment chart highlighted other significant pathways, such as antigen processing and presentation, the IL-17 signaling pathway, and parathyroid hormone synthesis, secretion, and action. These pathways' involvement suggests a broader spectrum of biological

processes contributing to the heterogeneity and progression of ALS.
Fow what concerns antigen processing and presentation, from a literature-based search, it was found how an increased activation of peripheral T cells in patients with sporadic ALS, triggered by IL-2 treatment, suggested a rise in antigen-experienced T cells in ALS blood. This was evidenced by in vitro culture with IL-2 for 14 days, pointing towards a potential weak autoantigen role for TDP-43 [17]. Furthermore, research on differentially expressed proteins in ALS has shown involvement in various pathways, including complement and coagulation cascades, NF-kappa B signaling, and ECM-receptor interactions, among others [18]. This broad activation of immune pathways underscores the complexity of the immune response in ALS.
The IL-17 signaling pathway, in paritcular, has been associated with inflammatory responses, which are increasingly recognized as important in neurodegenerative disease contexts. Interleukin 17A (IL-17A), the hallmark cytokine of Th17 cells, is part of a cytokine family that has been implicated in several inflammatory and autoimmune diseases, such as psoriasis, rheumatoid arthritis, and multiple sclerosis, as well as neurodegenerative diseases including ALS [19]. Notably, studies have observed fluctuating levels of IL-17A in ALS patients, with the highest concentrations appearing in the early stages of the disease [20]. These findings suggest that IL-17A might play a role in the early inflammatory processes of ALS.
Additionally, the connection between parathyroid hormone (PTH) synthesis, secretion, and action in ALS is complex and somewhat controversial. Some studies have shown improvements in muscular endurance in ALS patients with primary hyperparathyroidism (PHP) following parathyroid adenoma resection, suggesting a potential therapeutic angle. However, other research has not found significant pathogenic associations between parathyroid dysfunction and ALS progression. Despite these mixed findings, certain similarities in the progression patterns of PHP and ALS, such as muscle weakness and atrophy, hint at an intriguing, but not fully understood, link between these conditions [21].

Togehter, these enriched pathways indicate a multifaceted interaction between immune response, inflammation, and hormonal regulation in ALS, underscoring the disease's complexity and the need for a comprehnsive approach to its study and treatment.

4. Conclusions and Future Perspectives

In this study, we conducted a comprehensive computational analysis of RNA expression profiles among survival groups in amyotrophic lateral sclerosis (ALS), uncovering significant molecular pathways and genes associated with the disease's progression and heterogeneity. By leveraging transcriptomic data from 42 ALS patients and employing both unsupervised and supervised machine learning approaches, we identified potential genetic signatures differentiating short and long disease duration groups. Our findings highlighted the critical role of RNA splicing in ALS, as evidenced by the consistent enrichment of mRNA splicing-related terms in both *gprofiler2* and *pathfindR* analyses. This suggests an involvement of splicing deregulation in ALS pathogenesis, corresponding also with existing literature on the subject. Moreover, the identification of hnRNPs as significant players in this process may offer potential therapeutic targets. Beside RNA splicing, our analysis revealed additional pathways implicated in ALS, such as antigen processing and presentation, IL-17 signaling, and parathyroid hormone synthesis and action, suggesting a complex interplay between immune response, inflammation, and hormonal regulation in ALS, and emphasizing the need for a multifaceted approach to understanding and treating the disease.

Future studies should focus on validating our findings through experimental approaches, such as quantitative PCR and protein expression analysis, to confirm the involvement of identified genes and pathways in ALS. But also the integration of other omics data, such as miRNA profiles, could provide a more comprehensive view on ALS disease duration. Based on our findings, future research could explore the development of targeted therapies aimed at modulating RNA splicing and immune response pathways. Implementing personalized medicine approaches by stratifying ALS patients based on their molecular profiles could enable tailored therapeutic interventions, potentially improving outcomes and quality of life for individuals with ALS.

Abbreviations

ALS: Amyotrophic lateral sclerosis; **GEO:** Gene Expression Omnibus; **PCA:** Principal Component Analysis; **LDA:** Linear Discriminant Analysis; **GO:** Gene Ontology;**hnRNP:** heterogeneous nuclear ribonucleoprotein;**RBP:** RNA-binding protein; **FTD:** frontotemporal dementia; **miRNA:** microRNA.

Data Availability

The data utilized in this study can be downloaded from GEO database, with the following accession number: GSE212131. The code utilized for the analysis can be retrieved from the public repository on GitHub at the following link: `https://github.com/Sara-Baldinelli/NBDA-project`.

References

[1] Petros Xanthopoulos, Panos M Pardalos, Theodore B Trafalis, Petros Xanthopoulos, Panos M Pardalos, and Theodore B Trafalis. Linear discriminant analysis. *Robust data mining*, pages 27–33, 2013.

[2] Jonas Ranstam and Jonathan A Cook. Lasso regression. *Journal of British Surgery*, 105(10):1348–1348, 2018.

[3] Steven J Rigatti. Random forest. *Journal of Insurance Medicine*, 47(1):31–39, 2017.

[4] Gary C McDonald. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100, 2009.

[5] Masaya Oda, Yuishin Izumi, and Ryuji Kaji. Gene mutations in familial amyotrophic lateral sclerosis. *Brain and Nerve= Shinkei Kenkyu no Shinpo*, 63(2):165–170, 2011.

[6] Rachel Waller, Joanna J Bury, Charlie Appleby-Mallinder, Matthew Wyles, George Loxley, Aditi Babel, Saleh Shekari, Mbombe Kazoka, Helen Wollff, Ammar Al-Chalabi, et al. Establishing mrna and microrna interactions driving disease heterogeneity in amyotrophic lateral sclerosis patient survival. *Brain Communications*, 6(1):fcad331, 2024.

[7] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[8] Carson Sievert. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020.

[9] Max Kuhn. Caret: classification and regression training. *Astrophysics Source Code Library*, pages ascl–1505, 2015.

[10] Matteo Ciciani, Thomas Cantore, and Mario Lauria. rscudo: an r package for classification of molecular profiles using rank-based signatures. *Bioinformatics*, 36(13):4095–4096, 2020.

[11] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic acids research*, 51(W1):W207–W212, 2023.

[12] Ege Ulgen, Ozan Ozisik, and Osman Ugur Sezerman. pathfindr: an r package for comprehensive identification of enriched pathways in omics data through active subnetworks. *Frontiers in genetics*, 10:425394, 2019.

[13] Thomas Geuens, Delphine Bouhy, and Vincent Timmerman. The hnrnp family: insights into their role in health and disease. *Human genetics*, 135:851–867, 2016.

[14] Hitomi Tsuiji, Yohei Iguchi, Asako Furuya, Ayane Kataoka, Hiroyuki Hatsuta, Naoki Atsuta, Fumiaki Tanaka, Yoshio Hashizume, Hiroyasu Akatsu, Shigeo Murayama, et al. Spliceosome integrity is defective in the motor neuron diseases als and sma. *EMBO molecular medicine*, 5(2):221–234, 2013.

[15] Valentina La Cognata, Giulia Gentile, Eleonora Aronica, and Sebastiano Cavallaro. Splicing players are differently expressed in sporadic amyotrophic lateral sclerosis molecular clusters and brain regions. *Cells*, 9(1):159, 2020.

[16] Maria D Purice and J Paul Taylor. Linking hnrnp function to als and ftd pathology. *Frontiers in neuroscience*, 12:351731, 2018.

[17] Swetha Ramachandran, Veselin Grozdanov, Bianca Leins, Katharina Kandler, Simon Witzel, Medhanie Mulaw, Albert C Ludolph, Jochen H Weishaupt, and Karin M Danzer. Low t-cell reactivity to tdp-43 peptides in als. *Frontiers in immunology*, 14:1193507, 2023.

[18] Lin Chen, Ningyuan Wang, Yingzhen Zhang, Dongxiao Li, Caili He, Zhongzhong Li, Jian Zhang, and Yansu Guo. Proteomics analysis indicates the involvement of immunity and inflammation in the onset stage of sod1-g93a mouse model of als. *Journal of proteomics*, 272:104776, 2023.

[19] Junjue Chen, Xiaohong Liu, and Yisheng Zhong. Interleukin-17a: the key cytokine in neurodegenerative diseases. *Frontiers in aging neuroscience*, 12:566922, 2020.

[20] Milan Fiala, Madhuri Chattopadhay, Antonio La Cava, Eric Tse, Guanghao Liu, Elaine Lourenco, Ascia Eskin, Philip T Liu, Larry Magpantay, Stephen Tse, et al. Il-17a is increased in the serum and in spinal cord cd8 and mast cells of als patients. *Journal of neuroinflammation*, 7:1–14, 2010.

[21] Alexios-Fotios A Mentis, Anastasia M Bougea, and George P Chrousos. Amyotrophic lateral sclerosis (als) and the endocrine system: Are there any further ties to be explored? *Aging Brain*, 1:100024, 2021.

Supplementary Material

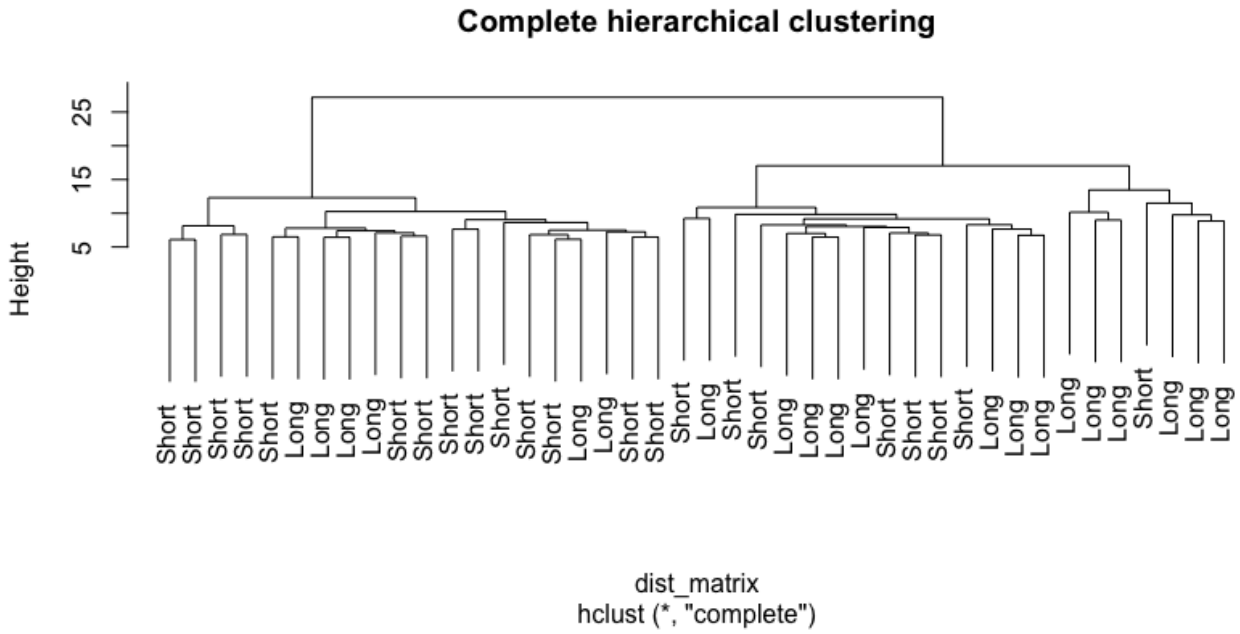


Figure S1. Hierarchical clustering exploiting complete linkage.

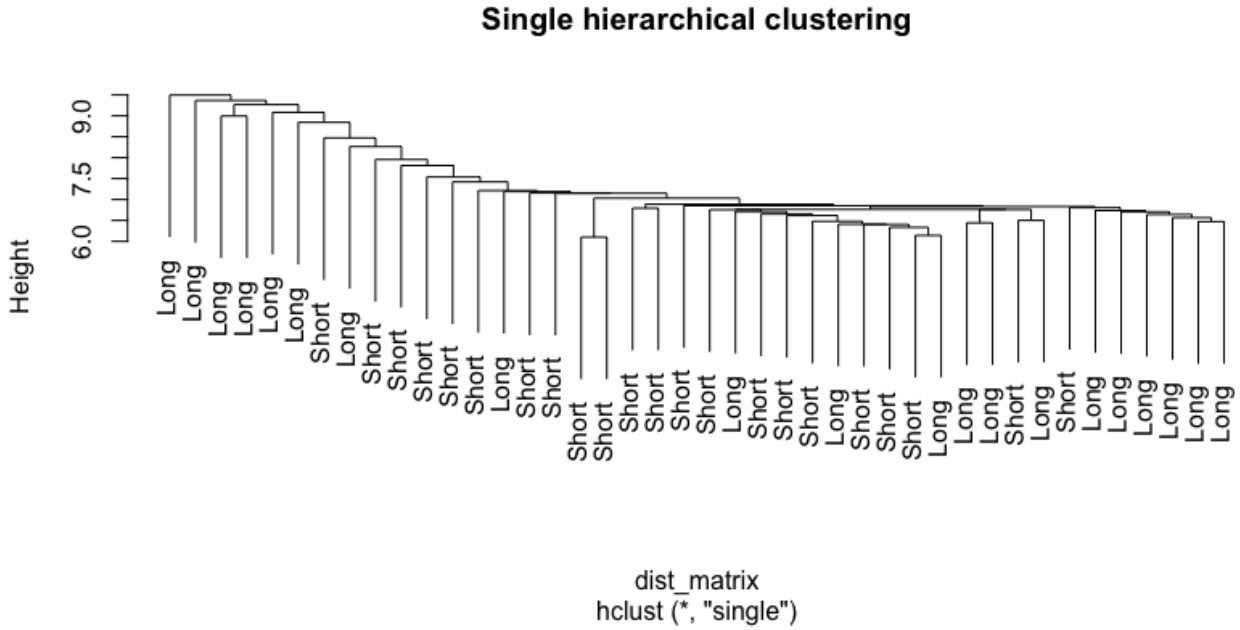


Figure S2. Hierarchical clustering exploiting single linkage.

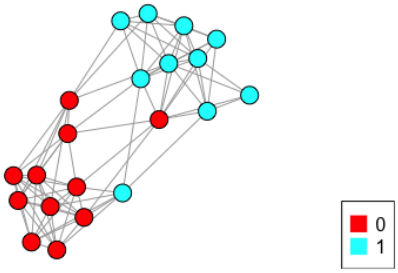


Figure S3. Training network obtained from Scudo.

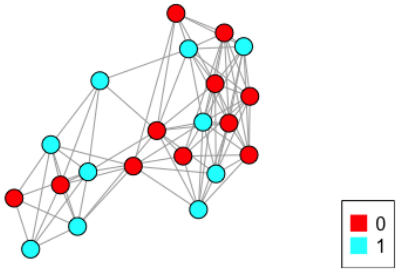


Figure S4. Test network obtained from Scudo. Even if samples in training have been perfectly separated, test looks not clearly classified. Accuracy is around 0.62.

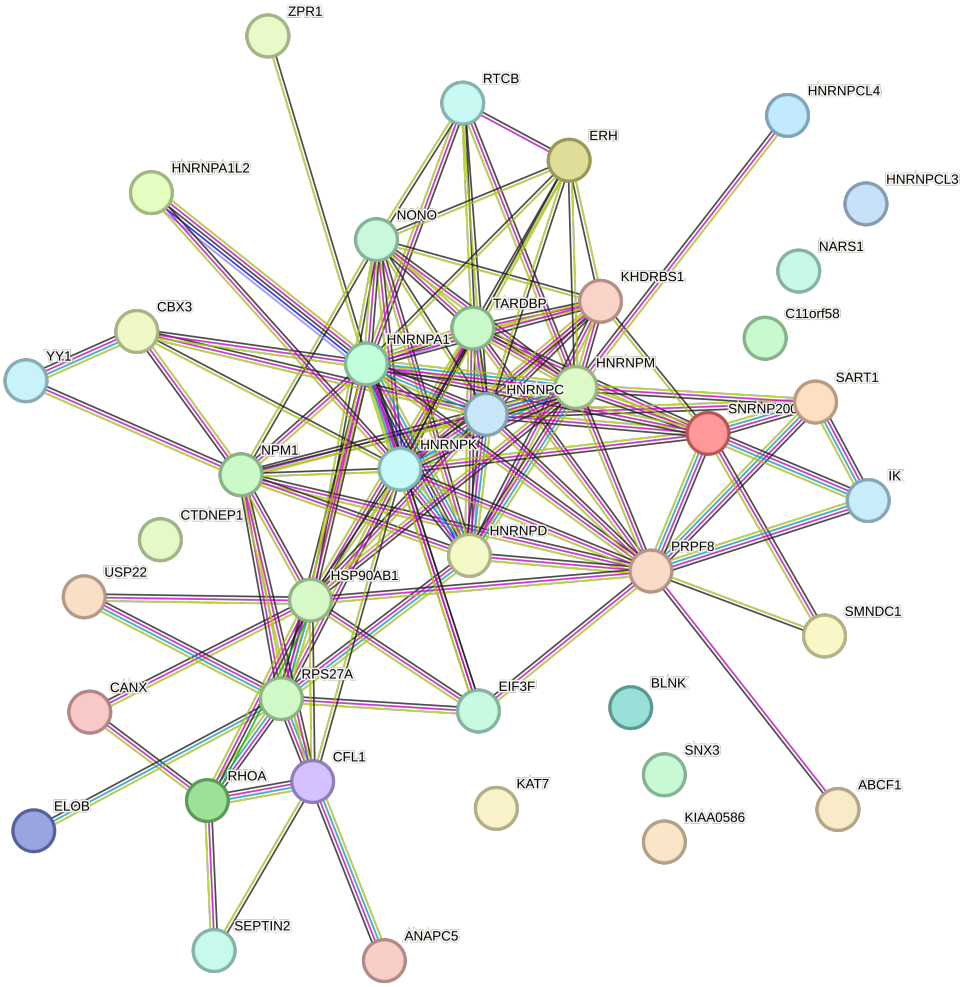


Figure S5. STRING network of interactions, the links with a wider stroke means a higher confidence level. This network have been achieved giving as input the list of the 200 most important genes for the RF classification. We can clearly see a cluster of nodes in the middle, representing mostly heterogeneous nuclear ribonucleoproteins related to RNA splicing and metabolism.