University Paris-Saclay

UFR Sciences

Bioinformatics Project Report

Topic:

Post-mortem Gene Analysis: Identification of Links with ALS

Made by:

- Nourchen MARZOUKI
- Inès DUFLOS

Table of contents

Introduction

Amyotrophic Lateral Sclerosis (ALS), also known as Lou Gehrig's disease, is a progressive neurodegenerative disorder characterised by the degeneration of motor neurons in the brain and spinal cord. Individuals affected by ALS experience muscle weakness, paralysis, and ultimately respiratory failure. Unfortunately, there is currently no cure for this devastating condition, and the average survival from symptom onset is only 3 to 5 years.

The urgent need for effective ALS treatments has led to numerous clinical trials, but disappointingly, over 40 drugs have failed to achieve their primary endpoints. Understanding the underlying molecular mechanisms and identifying reliable biomarkers are critical steps toward developing targeted therapies and improving patient outcomes.

Objective

The primary objective of our project is to identify biomarkers associated with ALS. To achieve this, we will leverage RNA-Seq sequencing data obtained from post-mortem brain cortex biopsies. These samples come from both ALS-diagnosed and non-diagnosed individuals. By analysing gene expression profiles, we aim to uncover specific molecular signatures that correlate with ALS pathology.

Step 1 – Data Pre-processing

In this part, the goal was to prepare our data so we could use it for the rest of the project.

Gather RNA Counts

First, we had to extract the data from the files we had and put it in a DataFrame. The DataFrame named 'data_matrix' contains all the genes samples of our data, each line corresponds to a patient and each column correspond to a gene.

Gather Sample Annotations

Once we had our data in 'data_matrix', we wanted to have more information concerning each patient samples. So, we created a DataFrame named 'data_annotation' where we included various information related to the samples that could be useful, such as the region from which the sample was taken or whether the patient was affected by ALS or not. Additionally, we linked this DataFrame with 'data_matrix' by including corresponding row numbers in the 'Num_ligne_matrix' column.
We wanted to make sure that our data in 'data_annotation' was correct so we manually created tests to see if the sample were correctly linked to their information's.

Step 2 – Descriptive Analysis

The objective in this part was to better understand the data that we were manipulating in order to gain a better understanding and facilitate their subsequent utilization.

Sample Description

Visualisation of data

As shown in Figure 1, the ALS Spectrum MND group exhibits a notably larger sample size compared to other groups, with over 140 samples. This substantial dataset is pivotal for identifying molecular signatures associated with ALS pathology.

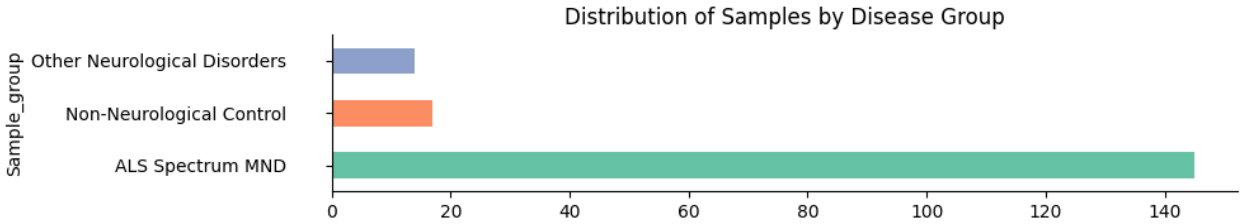


Figure 1: Distribution of Samples by Disease Group

As shown in Figure 2, the stacked bars represent various CNS subregions within each disease group, with colour coding for clear differentiation. The ALS Spectrum MND group exhibits a higher number of samples, particularly from the Motor Cortex.

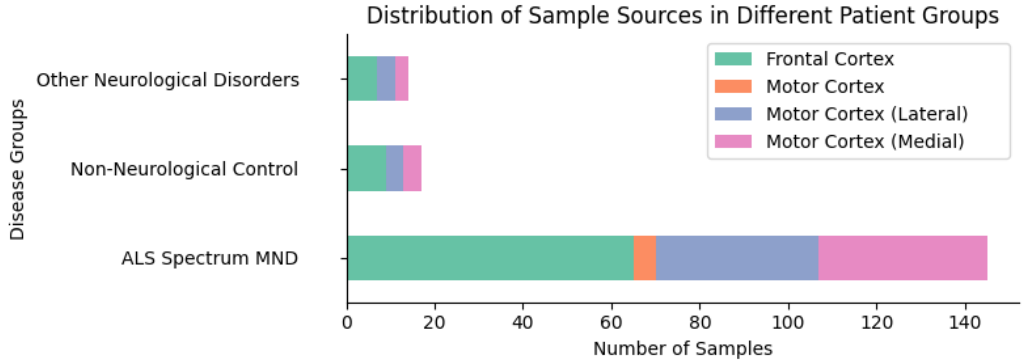


Figure 2: Distribution of Sample Sources in Different Patient Groups

As shown in Figure 3, the means of each sample are dispersed between around 500 and 1900. Most of them are between 700 and 1100, we can see that few samples mean are far away on the right on the histogram with values over 1200 and a few sample means are under the value of 600.

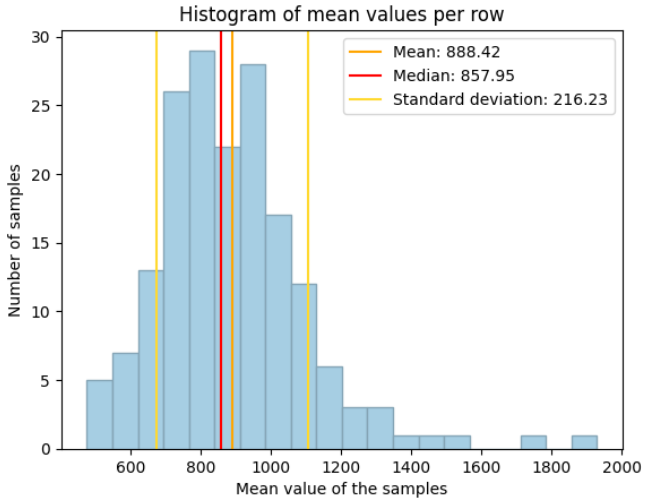


Figure 3: Histogram of mean values per sample

Samples that have a mean of over 1200

We decided to look closer to the sample that have a mean value over 1200 to see if there could be a corelation between the mean value being high and the patient having ALS.
There are 11 samples that have a mean of over 1200 and all of them are from patients having ALS, so there could be a corelation, but we can't be sure at this moment of the process.

Samples that have a mean of below 600

We decided to look closer to the sample that have a mean value below 600 to see if there could be a corelation between the mean value being low and the patient having ALS.
There are 9 samples that have a mean of below 600 and 8 of them are from patients having ALS, so there could be a corelation, but we can't be sure at this moment of the process.

RNA Counts Description

In this part, we did the mean of each gene to see if a gene with a high or a low mean could be a gene responsible for ALS.

Top 10 genes with the highest mean

The 10 genes with the highest mean are: MIRb:MIR:SINE, MALAT1, L2a:L2:LINE, MIR:MIR:SINE, L2c:L2:LINE, AluJb:Alu:SINE, MIRc:MIR:SINE, L2b:L2:LINE, MIR3:MIR:SINE, AluSx:Alu:SINE.
Most of them could be relating to ALS since when we look at the patient who have those genes at a high level, we can see that for almost every patients, the patient has ALS.

Top 10 genes with the smallest mean

To find the top 10 genes with the smallest mean we looked at the genes that have a mean over 0.056818181818 because it corresponds to 10/176. We chose the limit value of 10/176 because any value below that is not very relevant since it means that only 5 or 6 sample have that gene.
So, the 10 genes with the lowest mean are: AADACL4, ALKBH3-AS1, BANCR, BRD7P3, CD300LD, CDH16, CEACAM7, CFHR2, CMA1, CPLX4.
Some of them could be relating to ALS since when we look at the patient who have those genes at a low level, we can see that for most patients, the patient has ALS.

Step 3 – PCA

For the PCA we chose to use only PC1, PC2 and PC3 because they have variance of over 0.05 as show on Figure 4.

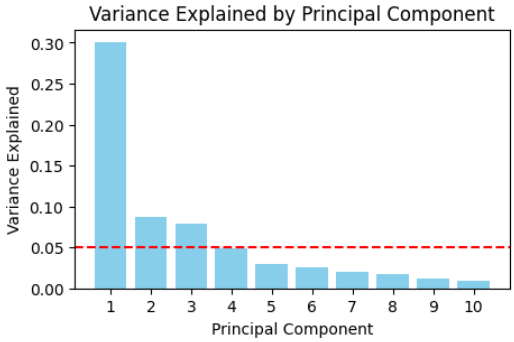
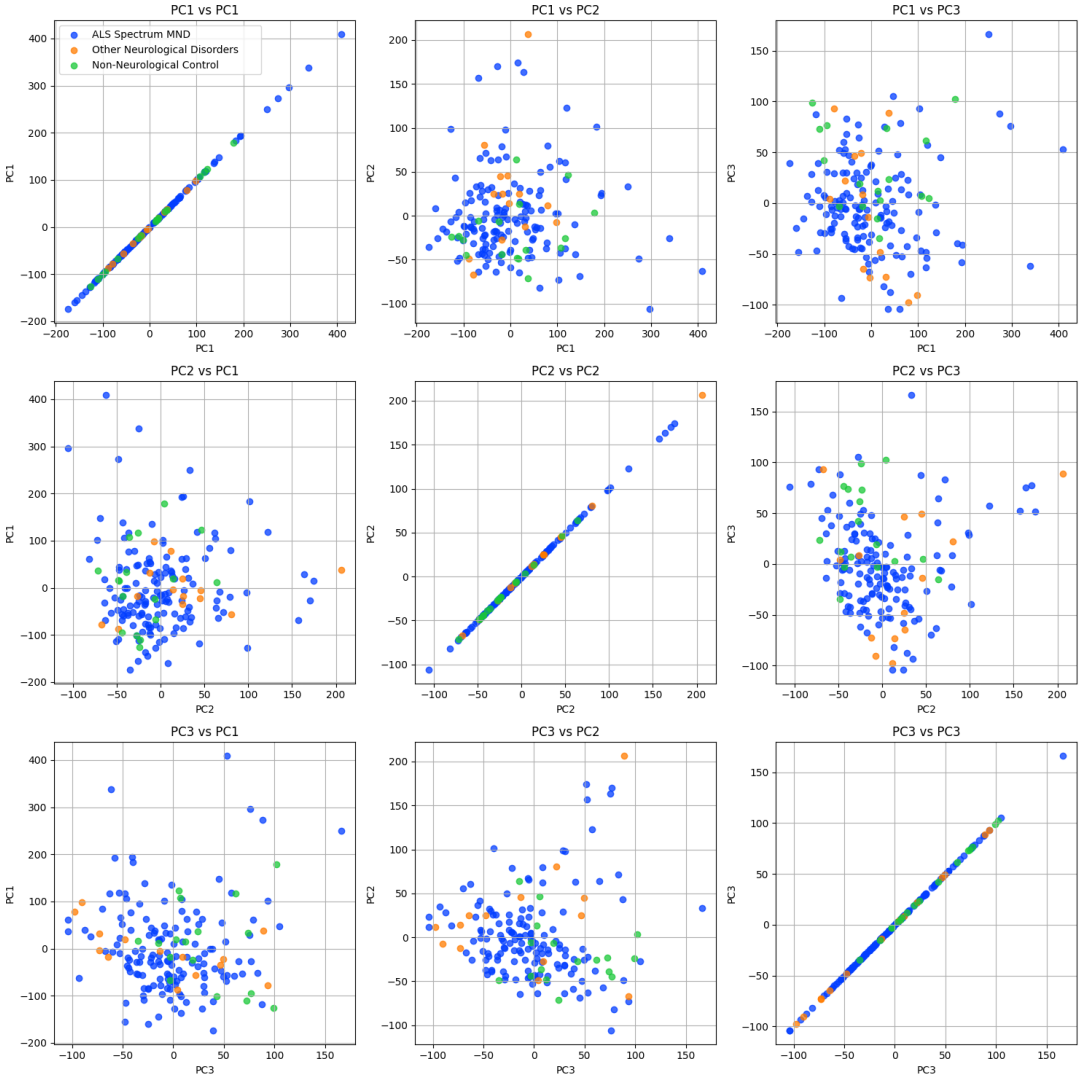


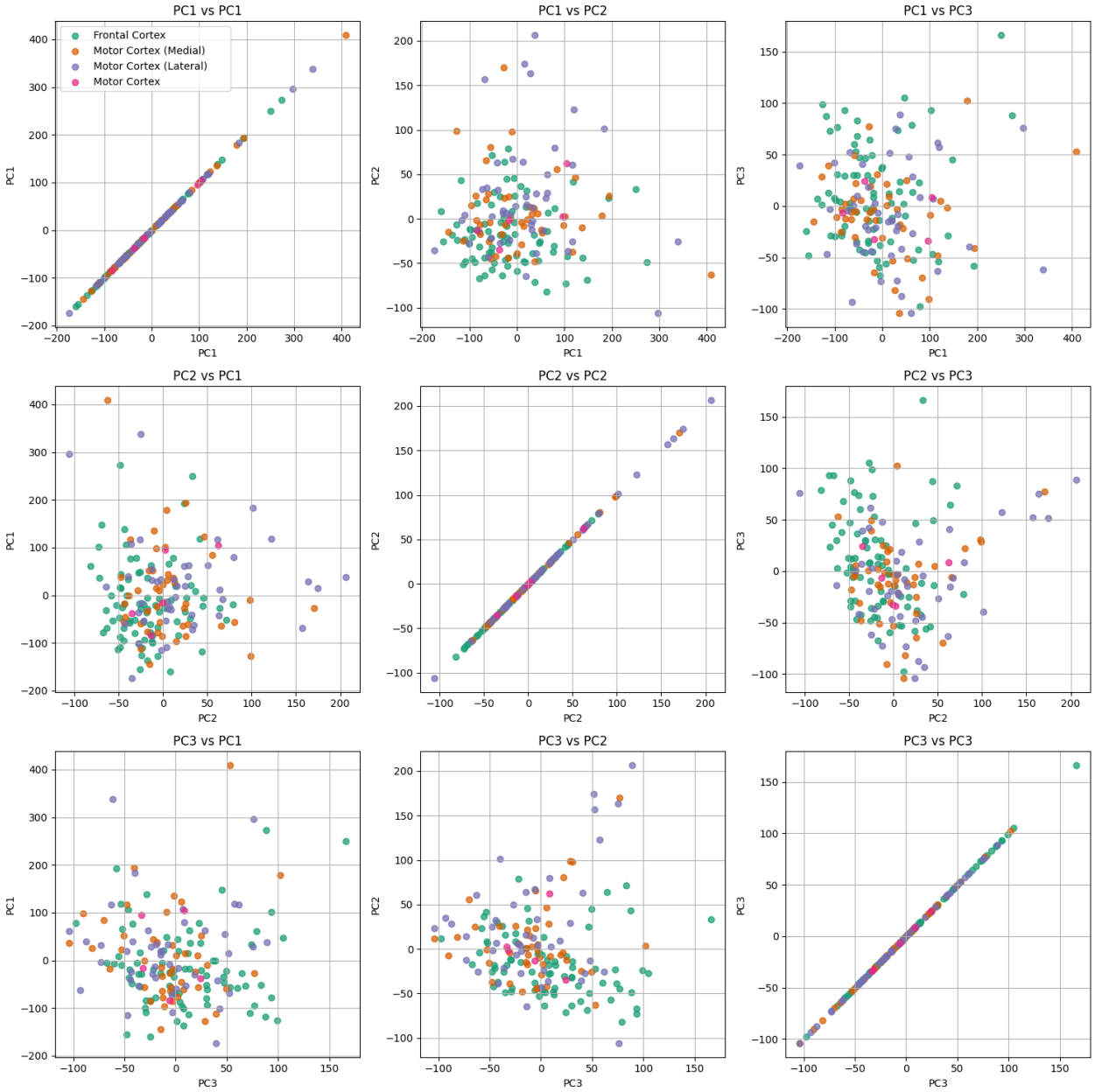
Figure 4: Variance Explained by Principal Component

We first did the PCA to see if it could show groups based on sample groups, but as shown in Figure 5 no groups were visible after doing the PCA.



<u>Figure 5:</u> PCA on Sample Groups

Then, we did the PCA to see if it could show groups based on CNS Subregion, but as shown in Figure 5 no groups were visible after doing this PCA.



<u>Figure 6:</u> PCA on CNS Subregion

Step 5 – Univariate Analysis

In this part, we used pydeseq2 to learn more about our data.
On the Figure 7, we can see that most of the genes that are corresponding have positive values. We can also see that they are divided into 2 groups, the first one very grouped on the right and the other more dispersed. So, there could be 2 “types” of ALS.

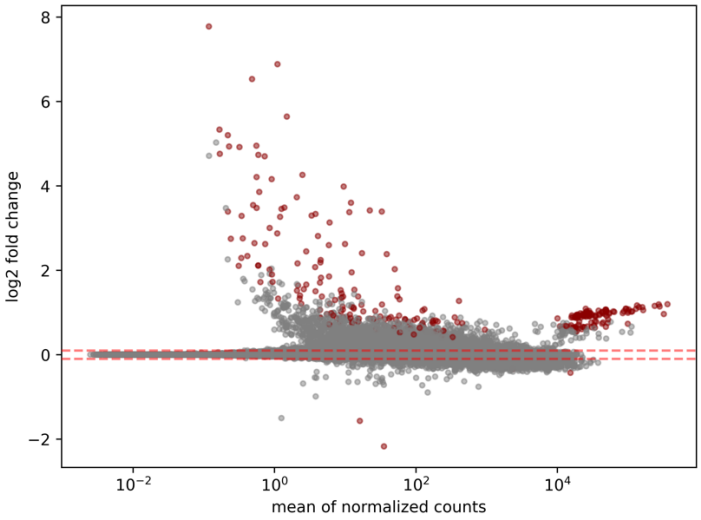


Figure 7: Volcano plot using pydeseq2

Step 6 – Multivariate Analysis

In this part, we used Elastic-Net to be able to extract the first 100 candidate genes that could be at the origin of ALS. The 100 candidates that we found are:

1) RPPH1
2) GFAP
3) L1PA5:L1:LINE
4) SLC1A2
5) L1PA4:L1:LINE
6) KIF5A
7) L1PA3:L1:LINE
8) L1PA7:L1:LINE
9) SYT1
10) PLP1
11) SCD
12) XIST
13) CALM1
14) ZNF483
15) DST
16) L1M5:L1:LINE
17) AluJb:Alu:SINE
18) MBP
19) MIR3648-1
20) SPP1
21) L1PA2:L1:LINE
22) ACTB
23) ATRNL1
24) MAP2
25) SORT1
26) MEF2C
27) PSD3
28) L1PA6:L1:LINE
29) TF
30) AluJr:Alu:SINE
31) L1ME4a:L1:LINE
32) APC

33) L2b:L2:LINE
34) NORAD
35) IDS
36) L1PB1:L1:LINE
37) ENC1
38) MIR3648-2
39) VCAN
40) NEAT1
41) AHNAK
42) SLC1A3
43) TPPP
44) L2a:L2:LINE
45) CAMK2A
46) NEFH
47) AluJo:Alu:SINE
48) HSP90AA1
49) CADM2
50) SYT11
51) OIP5-AS1
52) RMRP
53) RTN3
54) NECAB1
55) SCN2A
56) SERPINA3
57) MIR663AHG
58) APP
59) ATP1A2
60) DPYSL2
61) PRKACB
62) GAS7
63) L3:CR1:LINE
64) CLSTN1
65) ABCA2
66) MIRb:MIR:SINE
67) PPP3CA
68) MOBP
69) MAP4K4
70) PGM2L1
71) CPE
72) HSPA8
73) GRIN2B
74) MTSS1L
75) PREPL
76) CNP
77) ATP8A1
78) L1ME1:L1:LINE
79) ATP2B1
80) TSC22D1
81) ADGRV1
82) SPOCK1
83) FAIM2
84) AQP4
85) YWHAH
86) PTPRZ1
87) TMOD2
88) YWHAG

89) NDRG2
90) NTRK2
91) SPARC
92) SRRM2
93) PWAR5
94) SPTBN1
95) GAPDH
96) CNTN2
97) SYT2
98) L1MC4a:L1:LINE
99) TSPYL1
100) LAMP2

Some of the genes in this candidate list were revealed to be likely to cause ALS in step 2 when we looked at the genes with the highest and lowest means. This list therefore confirms certain suspicions that we had.

Conclusion

During this project, we had to analyse data from patient gene sample and find genes that could be responsible for ALS. By looking at our data, learning more about it and using different python tools we were able to extract 100 genes from our data that could be responsible for ALS.

This project was an excellent opportunity for us to learn how to work on data about genes using python. During this project we encountered a few difficulties, we had a lot of other projects to do and exams to prepare so we had to manage our time correctly to do this project.

Finally, this project helped us to learn more about both python and bio informatics.