PARKINSON'S DISEASE PATIENTS AND CONTROL SUBJECTS TRANSCRIPTOMIC ANALYSIS

Letizia De Pietri[1]

[1]University of Trento

Abstract

Parkinson's disease (PD) is the most prevalent movement disorder in elderly adults. This study concerns a transcriptomic profiling of peripheral blood mononuclear cells from patients affected by PD and control subjects. Both unsupervised and supervised methods were exploited to detect analogies and differences in the gene expression profile of the two groups. Although none of the employed unsupervised methods was able to detect patterns in the data for efficient splitting in the two expected groups, some supervised methods reached quite satisfactory levels of classification accuracy. Moreover, the involvement of some genes with PD found in literature is reflected in their expression pattern in the two groups.

Keywords: Parkinson's disease, transcriptomic analysis, machine learning, data analysis

1 INTRODUCTION

The study concerns an in depth analysis of transcriptome data coming from patients with Parkinson's disease (PD) and control subjects (Series GSE49126, Mutez et al. (2014)). PD is the most prevalent movement disorder in elderly adults but there is a big limitation in studying this disease that is the difficulty of obtaining brain tissues for performing analyses. To overcome this limitation, RNA was extracted from peripheral blood mononuclear cells (PBMC), these cells indeed constitute an easily accessible window onto the multi-organ transcriptome. More specifically, 20 elderly healthy controls constitute the control group while samples belonging to PD patients amount to a total of 30. To measure expression levels, RNA was hybridized on 4x44k Agilent expression microarrays which covered 41k unique genes and transcripts.
In the current analysis, a first attempt with unsupervised methods was performed to explore the possibility that RNA expression data differ between cases and controls groups; following, also supervised methods were employed to perform and evaluate samples classification and feature selection. More precisely, the utilized methods are Principal Component Analysis (PCA), K-means clustering, hierarchical clustering, Random Forest (RF), Linear Discriminant Analysis (LDA), logistic regression with Lasso regularization and a SCUDO based analysis. Finally, functional annotation was performed to integrate results together with a network analysis. All the mentioned analyses were carried out with the usage of R software environment version 4.3.3 (R Core Team (2021)) and its packages.

2 MATERIALS AND METHODS

2.1 Data pre-processing

Data were retrieved using the *getGEO* function, provided by the *GEOquery* library (Davis and Meltzer (2007)). After having checked that no NA values were present, a first inspection was conducted

through a boxplot (Figures 1 and 2 in supplementary material). To correct for the asymmetric data distribution, a logarithmic transformation was applied together with a median centering normalization that allowed to stabilize variance across samples.

2.2 PCA

Following, a PCA was performed. The *prcomp* function from the *stats* library (R Core Team et al. (2013)) requires genes on the columns and samples on the rows. A scree plot showing the variance explained by each component was given as a first output (Figure 3 in supplementary material).
Besides the standard two dimensional plot, using the *plot3d* function provided by the *rgl* library (R Core Team et al. (2013)), a 3D plot was also produced with the aim of increasing the total variance explained by the components included in the plot.

2.3 K-means clustering and hierarchical clustering

To try with another unsupervised approach, k-means clustering algorithm was applied to our data exploiting the *kmeans* function from the *stats* library (R Core Team et al. (2013)). The *K* parameter was set to 2 and to visualize the clustering a PCA was performed and represented using the *ggplot2* library (Villanueva and Chen (2019)).
Another unsupervised clustering approach was attempted by performing a hierarchical clustering. The *hclust* function was applied on a distance matrix built on the transpose of the expression matrix; four different agglomerative methods were used (complete-linkage, average-linkage, single-linkage and centroid-linkage).

2.4 Random Forest

Using the *randomForest* function from the homonym library (Liaw and Wiener (2002)), a RF classifier was built on a total of 1000 trees and the 200 genes that were the most important for classification were extracted. Finally, the matrix of expression values for the top 25 genes was obtained and a heatmap was generated on it.

2.5 Linear Discriminant Analysis

To perform LDA, t-tests were performed to compare gene expression between the two groups and genes with a p-value of less than 0.1 were selected. A subset of the data containing 20 controls and 20 affected was selected and it was split into training and test set: the training set included 15 control and 15 affected samples while the test set included the remaining 5 control and 5 affected samples. After that, an LDA model was trained on the training set using the *lda* function from the *MASS* library (Venables and Ripley (2002)) and it was used to predict class labels for the test set. Groups projections on the LDA axis were plotted together with a ROC curve to evaluate model performance.

2.6 Lasso

The *glmnet* package (Friedman et al. (2010)) was exploited to perform logistic regression with Lasso regularization. After having selected 20 controls and 20 cases, 15 controls and 15 affected samples were randomly chosen for training. The remaining 5 controls and 5 affected samples were used for testing. A logistic regression (*family = 'binomial'*) with Lasso regularization was then fitted on the properly prepared training data using the *glmnet* function. Then, cross validation was performed on the training set to select the opitmal lambda using the *cv.glmnet* function and the selected lambda value was used to predict class labels and probabilities (*predict* function) on the test data. After that, predicted probabilities were used to plot the ROC curve and to calculate the AUC to evaluate model performance using *ROCR* library (Sing et al. (2005)).

2.7 Models comparison with 10 fold cross validation

The *CARET* library (Kuhn and Max (2008)) was used to run 10 fold cross validation to compare the performance of RF, LDA analysis and Lasso models. After the set up of the 10 fold cross validation, all the three models were indeed trained specifying the *trControl* parameter. The *resamples* function was used to combine the results of the models for comparison, and the accuracy of the three models was plotted using *ggplot* (Villanueva and Chen (2019)).

2.8 rScudo

The *createDataPartition* function from the *CARET* library (Kuhn and Max (2008)) was used to split the data (20 controls and 20 affected samples) into training and testing sets. Then, the *rScudo* library (Ciciani et al. (2020)) was employed; SCUDO is a rank-based method for the analysis of gene expression profiles. The function *scudoTrain* was exploited to compute gene signatures for each sample; 25 top and bottom features were specified together with a significance level (alpha) of 0.05. After viewing the top 5 up regulated consensus signatures in the two groups, the model was validated on test data using the *scudoTest* function. A network for the test results was then built with the *scudoNetwork* function and plotted with the *scudoPlot* function (Figure 6 in supplementary material). A supervised classification on the test set based on the previously trained network was performed thanks to the *scudoClassify* function. Statistics were calculated with the *confusionMatrix* function from *CARET* (Kuhn and Max (2008)).

2.9 Functional annotation

g:Profiler (Reimand et al. (2007)) is the tool that was employed for functional annotation. The lists that were annotated were the 25 genes up regulated in the control or in the affected groups obtained through the *consensusUpSignatures* function. Gene symbols were extracted from the dataset and given in input to the annotation tool. Results were then studied through a literature search to extrapolate biological meaning in the context of Parkinson's diasease.

2.10 Interactions Network Analysis

A first attempt of network analysis was performed using the *pathfindR* library and the *run_pathfindR* function. First, a list of differentially expressed genes was generated through the *limma* (Ritchie et al. (2015)) library that is specifically designed for performing differential expression on microarray and RNA-seq data. Then, the list of genes with associated p-value was given in input to *pathfindR*.
A second attempt was performed with STRING (Szklarczyk et al. (2023)). The already cited lists of 25 genes up regulated in the control or in the affected groups were expanded using the STRING database. STRING is a repository of protein-protein interactions which allowed to find interactions between the query proteins and other proteins known to interact either from curated databases or because experimentally determined. Note that since many microarray terms are actually transcripts without a known function and name, the two lists amount to less than 25 actual extracted genes/proteins (21 for controls and 15 for affected). Enriched ontologies outputted by STRING were analyzed.

3 Results and Discussion

3.1 Data pre-processing

Figure 1 depicts the boxplot obtained after logarithmic transformation and median centering normalization. Expression data consist of 43376 measurements for each sample.

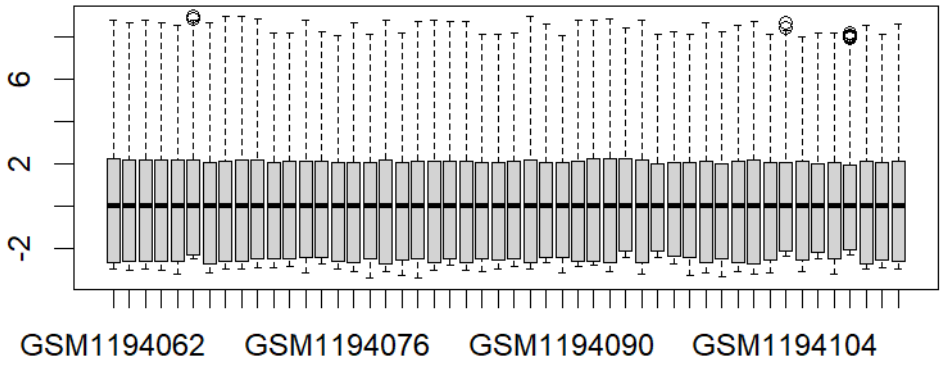


Figure 1. Boxplot resulting from logarithmic transformation and median centering normalization.

3.2 PCA

Figure 2 refers to the output of the PCA: we can observe that the two components are not able to split individuals in two groups corresponding to cases (magenta) and controls (blue). This suggests that PCA did not detect underlying differences in the expression profiles of the groups.

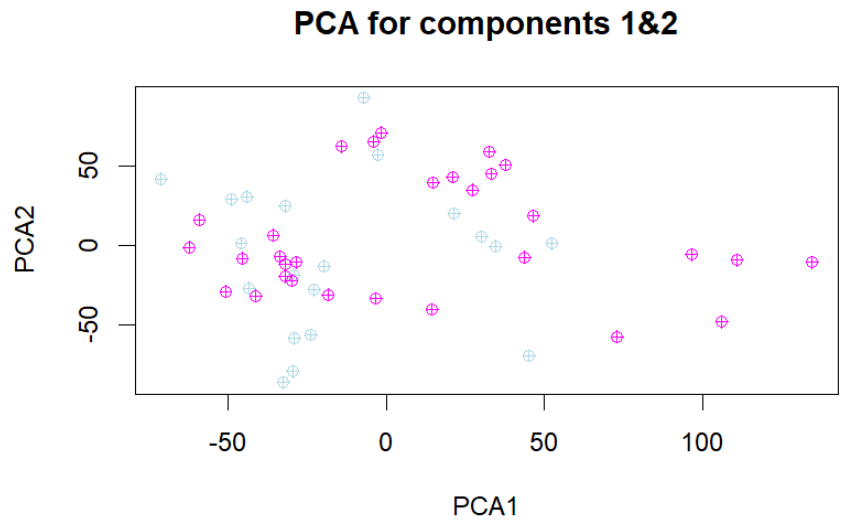


Figure 2. 2D PCA plot: magenta samples are cases and blue samples are controls.

A 3D plot **Figure 3** was also produced to see if including the third component, hence increasing the explained variance, helped in splitting data in a meaningful way. The grouping remains ambiguous as cases and controls do not cluster in an apparent informative way.

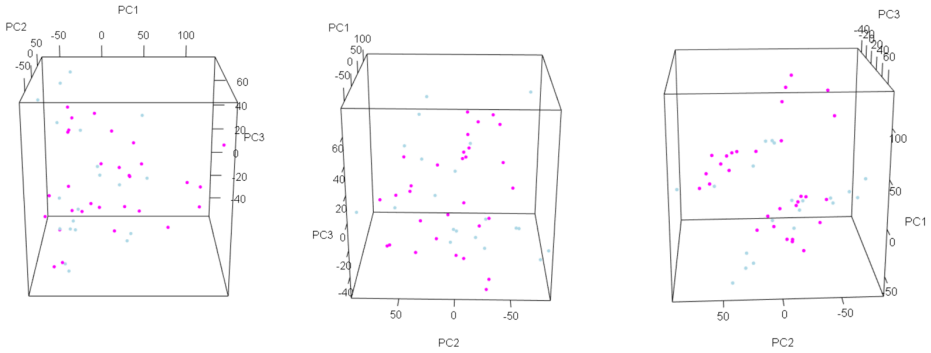


Figure 3. 3D PCA plots from three different perspectives: magenta samples are cases and blue samples are controls.

3.3 K-means clustering and hierarchical clustering

Figure 4 shows the visualization through dimensionality reduction of the k-means clustering. It can be observed that the two groups resulting from the clustering do not correspond to the cases and controls groups.

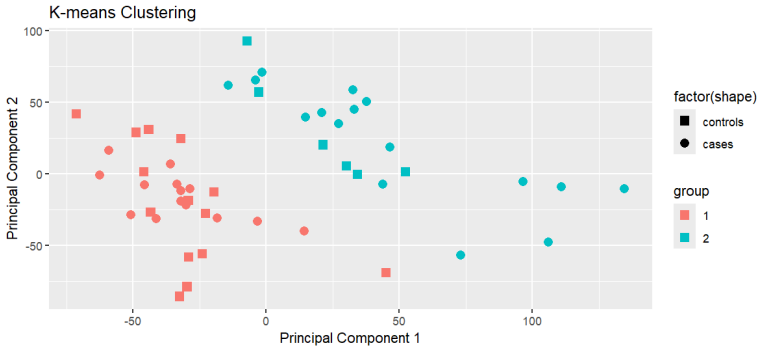


Figure 4. K-means clustering: the analysis does not split data into apparently meaningful groups. Control samples are represented as squares, cases samples are represented as circles.

The result of hierarchical clustering obtained with the complete-linkage method (supplementary material, Figure 4) gave the same result as the one obtained with the average-linkage method. The grouping obtained did not reflect the cases-controls splitting. The other two methods (centroid-linkage and single-linkage) outputted even worse results as they produced two unbalanced clusters, one of them composed of one sample only (supplementary material, Figure 5).

3.4 Random Forest

Figure 5 displays the heatmap built on the expression values of the 25 genes that were found to be the most 'important' ones by the RF classifier. We can observe that some samples belonging to the same group are clustering together and some genes have similar expression patterns across samples; for example genes 43812 and 11461 have low expression values in controls and higher values in cases; the opposite happens for genes 36488 and 31423. These genes correspond, respectively, to 'LOC647121' (also known as 'EMBP1'), 'ZADH2' (also known as 'PTGR3'), 'EPB41' and 'SLC4A1'. The EPB41 gene encodes for the Erythrocyte Membrane Protein Band that constitutes the red cell membrane cytoskeletal network; at the same time, the composition of erythrocytes membrane has been found to be a biomarker for Parkinson's disease (Tian et al. (2019)). The expression pattern of the EPB41 gene in our samples might reflect these findings. It is interesting to note that the SLC4A1 gene might reflect these findings as well, it is indeed expressed in the erythrocyte plasma membrane and it presents the same expression pattern as the EPB41 gene.

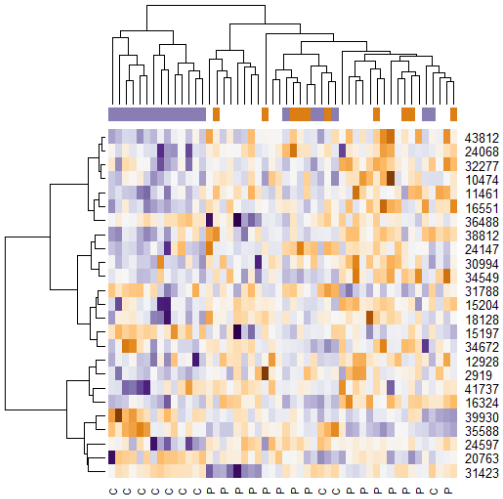


Figure 5. Heatmap built on the expression values of the top 25 genes. On the columns we have hierarchical clustering of samples and on the rows clustering of genes. Gene expression values range from purple to orange. Labels are shown both for genes (rows) and samples (C for control, P for patient).

3.5 Linear Discriminant Analysis

In **Figure 6**, which depicts the projection on the LDA axis of the two groups, we can observe that 4 samples are misclassified according to this projection. When evaluating the performance of this model on the test set, a specificity of 1 and an accuracy of 0.8 were reached; an affected sample was indeed classified as control. Overall, this LDA model reached satisfactory results in classifying the two groups even though the evaluation was performed on 10 samples only.

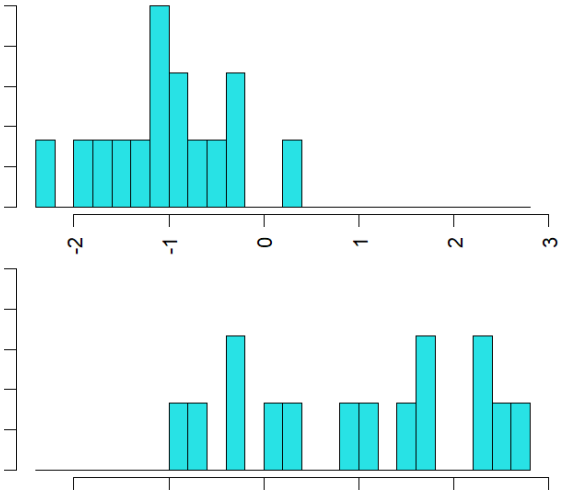


Figure 6. Projection of the two groups on the LDA axis: 4 samples are misclassfied.

3.6 Lasso

By performing cross validation, a value of lambda of 0.4214898 was selected as this is the one that minimized the cross validated error. Also, model performance evaluation outputted an AUC of 0.5, meaning that the model performed no better than random.

3.7 Models comparison with 10 fold cross validation

Figure 7 helps in comparing the performances of LDA, RF models and Lasso in terms of accuracy. This figure shows that LDA and RF reached a higher accuracy (around 0.68) with respect to Lasso (0.6). Given that, LDA and RF might be the best methods to classify our samples or new samples in the same biological context.
Although the original accuracy value computed without repeated cross validation for LDA classification was 0.8, after conducting 10 fold cross validation the accuracy value decreased to less than 0.7 as previously specified. This value is probably a more precise and realistic one which, even though worse, could still represent a satisfactory classification potential.

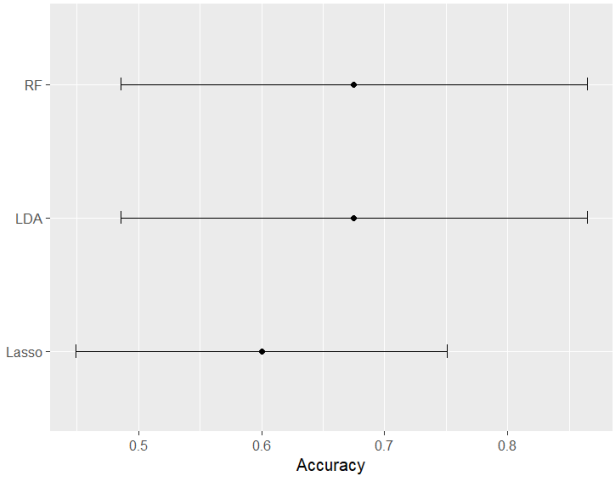


Figure 7. CARET plot showing accuracy values for RF, LDA and Lasso calculated through 10 fold cross validation.

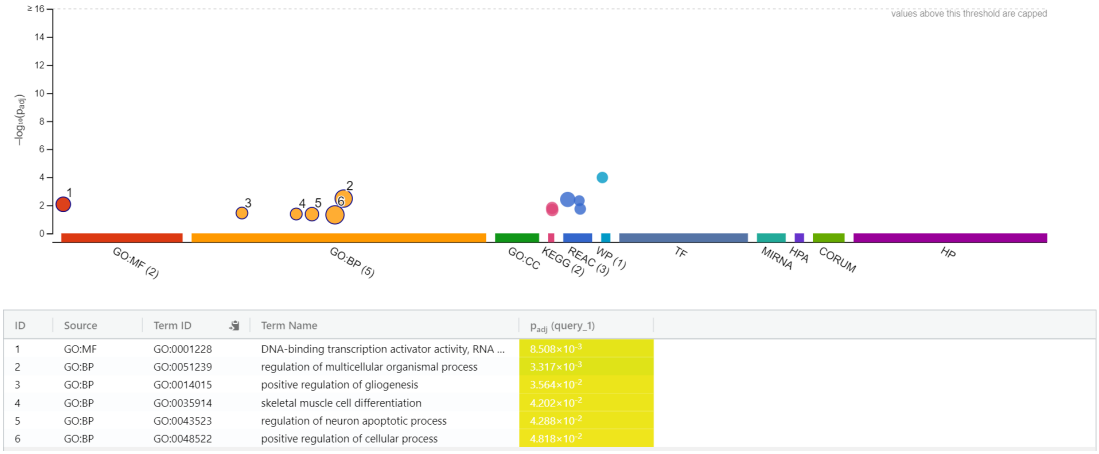
3.8 rScudo

The signatures that were found to be up regulated in the control group are: IRF5, EGR1, PVALB, NR4A2 and HBEGF; while the ones up regulated in the affected group are: LOC729983, C9orf140 (also known as SAPCD2), PRG4 and CLIC5. The genes up regulated in the controls are involved in the modulation of cell growth and differentiation, transcription regulation, calcium ion binding and epidermal growth factor receptor signaling pathway. It is interesting to note that decreased levels of the NR4A2 gene product is associated with PD (Le et al. (2003)); since this gene is found to be more expressed in controls, its expression pattern between the two groups in our dataset is in accordance with NR4A2 low expression levels association with Parkinson's. The genes up regulated in the affected group are involved in cell proliferation, boundary lubrication within articulating joints, myoblast proliferation and endothelial cell maintenance. These functions do not have an apparent biological connection to the context of PD.
In the network created by SCUDO (supplementary material, Figure 6) it was not possible to detect a clear splitting of the two groups, this means that there is an overlapping of gene expression patterns of the two groups based on the top 25 up and down regulated genes.
Regarding performance evaluation, *CARET* statistics revealed an accuracy of 0.5 which marks a random classification. These results suggest that the model is not able to efficiently distinguish the two groups.

3.9 Functional annotation

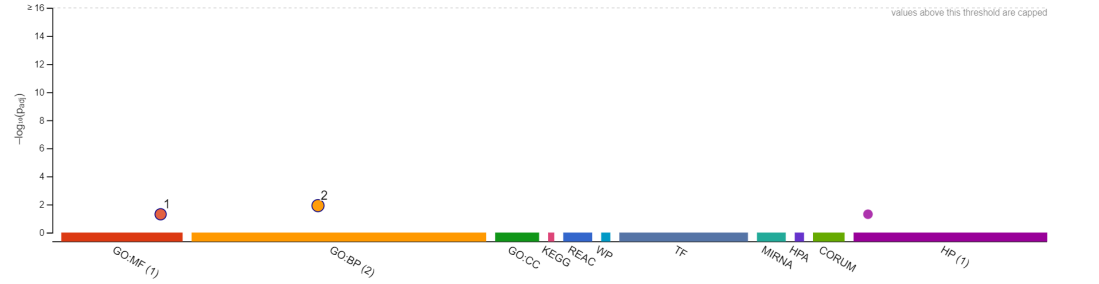
The output of *g:Profiler* for the genes up regulated in the control group is found in **Figure 8**. The fact that gliogenesis is positively regulated in control patients is in accordance with the possible impairement of this process in PD patients (Lee et al. (2023)). Moreover, some mechanisms that were

found to be up regulated in controls, and so relatively down regulated in cases, such as 'positive regulation of cellular process', 'citokine signaling in immune system' and 'oncostatin M signaling pathway' participate in higher level processes that were found to be disrupted in PD patients (cell survival, inflammation and immune processes) (Mutez et al. (2014)). Apparently, no other biologically meaningful up regulated pathways, either in cases (shown in **Figure 9**) or in controls, are found in the *g:Profiler* output; actually, 'cell-cell adhesion mediator activity' is up regulated in patients even if literature highlights the fact that it was detected to be down regulated in the PD context (Chapman (2014)).



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0001228	DNA-binding transcription activator activity, RNA ...	8.508×10^{-3}
2	GO:BP	GO:0051239	regulation of multicellular organismal process	3.317×10^{-3}
3	GO:BP	GO:0014015	positive regulation of gliogenesis	3.564×10^{-2}
4	GO:BP	GO:0035914	skeletal muscle cell differentiation	4.202×10^{-2}
5	GO:BP	GO:0043523	regulation of neuron apoptotic process	4.288×10^{-2}
6	GO:BP	GO:0048522	positive regulation of cellular process	4.818×10^{-2}

Figure 8. g:Profiler output of the list of genes up-regulated in control samples.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0098632	cell-cell adhesion mediator activity	4.988×10^{-2}
2	GO:BP	GO:0045104	intermediate filament cytoskeleton organization	1.208×10^{-2}

Figure 9. g:Profiler output of the list of genes up-regulated in affected samples.

3.10 Interactions Network Analysis

The network analysis with *pathfindR* did not succeed as one gene only (SSH1, a phosphatase that regulates actin filament dynamics) was found to be significantly differentially expressed between the two groups (p-values in Figure 7, supplementary material). Another method for differential expression analysis should be attempted.

After having performed STRING (Szklarczyk et al. (2023)) expansion, the controls network reached a total of 31 interacting proteins starting from 21, while the cases network reached a total of 25 terms from the initial 15.

A first enrichment result highlights that the control up-regulated genes are involved in regulation of neuron death. This might represent a controlled and regulated neuron death which is in accordance with the unaffected condition of the brain tissue in this group. Also, transcription activity molecular functions are enriched in the network. The average node degree of the network (**Figure 9A**) is 5.29

while the average local clustering coefficient is 0.67; this underlines many mutual proteins interactions. It is worth highlighting that one of the hub nodes, NR4A2, has been found to be associated with PD if mutated. The over-expression of this gene in controls (as it is present in the original list of up regulated genes), and hence its efficient translation probably in absence of mutations, highlights its importance in determining the healthy condition of the group.
Concerning the cases network (**Figure 9B**), the mainly enriched biological processes are DNA replication activities and the other ontology terms are enriched in the same context of DNA replication. No apparent connection exists between PD and increased levels of DNA replication activity. Also, the lower average local clustering coefficient indicates a less connected network of proteins. No relevant informations were extrapolated from this network.

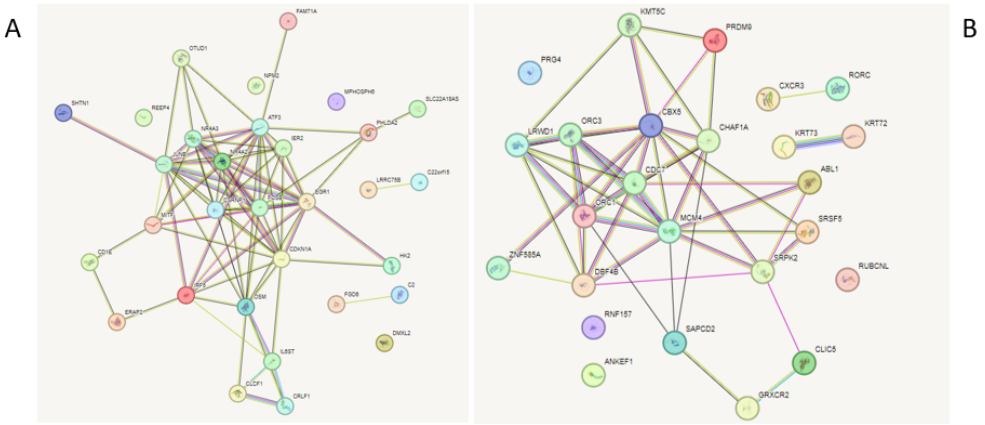


Figure 10. STRING produced interaction networks: Figure A controls network, Figure B cases network.

4 Conclusion

In conclusion, none of the unsupervised methods that were employed was able to perform a meaningful splitting of the data, the two groups found by clustering methods were not corresponding to the cases and controls groups and the PCA did not catch any informative pattern in the data. For what concerns supervised methods, both LDA and RF reached an accuracy of almost 0.7 in classifying samples when trained on a proportion of our dataset, none of the other supervised methods reached such accuracy value. Although a value of 0.7 is not optimal, it might still reflect an acceptable performance in classification of new samples. This is an important insight into the diagnosis potential of these methods. Regarding the network analysis, STRING network detected an important hub node (NR4A2) that might play an important role in determining the disease status. Also, some correspondences were found between the expression pattern of some genes in our samples and their association with PD reported in literature, supporting their relevance as possible PD markers. Lastly, a concordance was detected between the pathways found to be enriched in patients in this analysis and the ones found to be associated with PD in the dataset reference study (Mutez et al. (2014)).

Acknowledgment

A special thank you goes to professor Mario Lauria for his precious help throughout the development of the project.

References

Chapman, M. A. (2014). Interactions between cell adhesion and the synaptic vesicle cycle in parkinson's disease. *Medical hypotheses*, 83(2):203–207.

Ciciani, M., Cantore, T., and Lauria, M. (2020). rscudo: an r package for classification of molecular profiles using rank-based signatures. *Bioinformatics*, 36(13):4095–4096.

Davis, S. and Meltzer, P. S. (2007). Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847.

Friedman, J., Tibshirani, R., and Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Kuhn and Max (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Le, W.-d., Xu, P., Jankovic, J., Jiang, H., Appel, S. H., Smith, R. G., and Vassilatis, D. K. (2003). Mutations in nr4a2 associated with familial parkinson disease. *Nature genetics*, 33(1):85–89.

Lee, A. J., Kim, C., Park, S., Joo, J., Choi, B., Yang, D., Jun, K., Eom, J., Lee, S.-J., Chung, S. J., et al. (2023). Characterization of altered molecular mechanisms in parkinson's disease through cell type–resolved multiomics analyses. *Science Advances*, 9(15):eabo2467.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Mutez, E., Nkiliza, A., Belarbi, K., de Broucker, A., Vanbesien-Mailliot, C., Bleuse, S., Duflot, A., Comptdaer, T., Semaille, P., Blervaque, R., et al. (2014). Involvement of the immune system, endocytosis and eif2 signaling in both genetically determined and sporadic forms of parkinson's disease. *Neurobiology of disease*, 63:165–170.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team, R. et al. (2013). R: A language and environment for statistical computing.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, 35(suppl_2):W193–W200.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881.

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., et al. (2023). The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646.

Tian, C., Liu, G., Gao, L., Soltys, D., Pan, C., Stewart, T., Shi, M., Xie, Z., Liu, N., Feng, T., et al. (2019). Erythrocytic α-synuclein as a potential biomarker for parkinson's disease. *Translational neurodegeneration*, 8:1–12.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Villanueva, R. A. M. and Chen, Z. J. (2019). ggplot2: elegant graphics for data analysis.