Machine learning for neurodegenerative disease diagnosis and biomarker extraction

Alessandro Dipalma, Francesco Iannelli & Marco Petix

May 16 2022

Abstract

Gene expression data provide invaluable insights into disease mechanisms and there is a huge unmet need for effective treatments of neurodegenerative diseases as therapeutic options are still limited. The complexity of the molecular mechanisms underlying neuronal degeneration and the heterogeneity of the patient population present massive challenges to the development of early diagnostic tools. In this work we present a machine learning oriented approach for the automated extraction of insightful biomarkers and we show how such tools can aid clinicians in patient diagnosis.

1 Method

In this central section we will detail the setup of our experiment and the methodology, ranging from a description of the dataset of our choice to the steps we followed throughout our analysis.

1.1 Data source

We opted to use GSE26927 Durrenberger et al. [2012] for our case study. The dataset, which is publicly available on the Gene Expression Omnibus, is an expression profiling by array and it contains a total of 118 samples selected post-mortem on the basis of a confirmed diagnosis of 6 different neurodegenerative diseases: Alzheimer's disease (AD), Huntington's disease (HD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS) and schizophrenia (SCHIZ). The gene expression analysis was performed on the RNA samples using the Illumina whole genome. All the labelling and hybridisation of the samples was carried out in a single experiment by the Imperial College group. The exact patients' subdivision and tissues for each disease are depicted in Table 1. It was common opinion among us that GSE26927 was very suitable for our task, as not only the analysis of whole-genome expression patterns across these major disorders could offer an outstanding opportunity both to look into exclusive disease-specific changes and, more importantly, to look for potential common molecular pathogenic mechanisms, but it also allowed us not to face the tiresome and difficult quest of integrating multiple datasets. Indeed, we attempted the latter at the beginning of our study but it proved to be quite difficult due to differences in sensitivity between heterogeneous microarray platforms and pre-processing techniques, which were different in every gene expression experiment we analyzed.

1.2 Statistical Analysis

We first performed an in-depth exploratory analysis to properly assess the quality and the nature of the data at hand. The patients had 20589 gene expression measurements each.

Name	Disease	Control	Tissue
AD	11	7	Entorhinal cortex
HD	10	10	Ventral head of the caudate nucleus
PD	12	8	Substantia nigra
ALS	10	10	Cervical spinal cord
MS	10	10	Frontal gyri
SCHIZ	10	10	Superior temporal cortex

Table 1: Tissues and patients for each disease.

Age The youngest patient had died at 18 years old and the oldest at 104. As depicted in Figure 1, the age distribution was bimodal, with the modes being at 50 years old and 75 years old respectively. We spotted a patient whose age was missing and we imputed this datum with the median of the ages in the same sex and disease group.

Sex The patients sex distribution was skewed towards males, as there were 76 males versus 42 females, with the control group of the ALS disease being the most skewed with only male patients, as can be seen from Figure 2. As Figure 3 shows, there was not relevant skewness in the age distribution for the control and disease groups of any of the 6 pathologies except for AD, where the patients in the disease group were significantly older than the control ones.

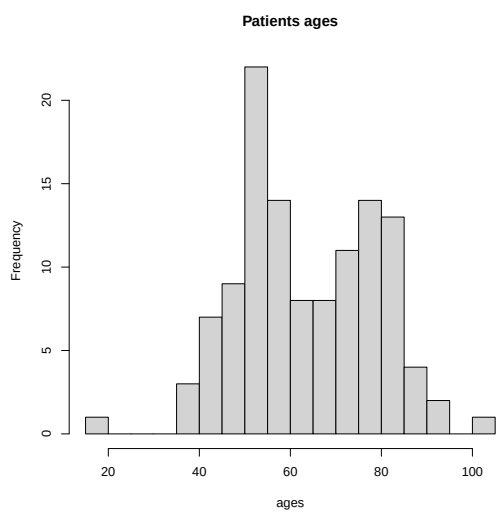


Figure 1: Patients ages distribution.

Gene Expression Matrix From the distribution of the gene expression matrix entries we noted few negative values close to 0, approximately the 0.008% of the total, with 0 itself being the median. The largest negative entry was -10^{-6}, the smallest was -35 and -2.04 accounted for their mean. Moreover, the positive entries were not as tightly distributed and spanned from a small region to the right of 0 to $O(10^4)$. Such statistics are commonly found in gene expression data processed via background normalization Sui et al. [2009], where 0 represent absence of expression and few negative entries are to be expected due to background noise; for this reason we chose to substitute all negative values with 0.0.

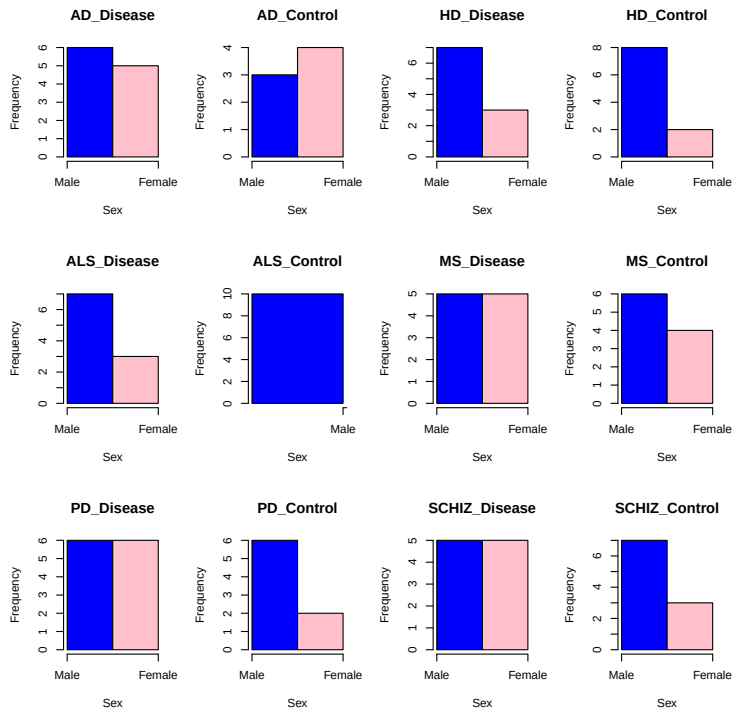


Figure 2: Patients sex distribution across disease and control groups.

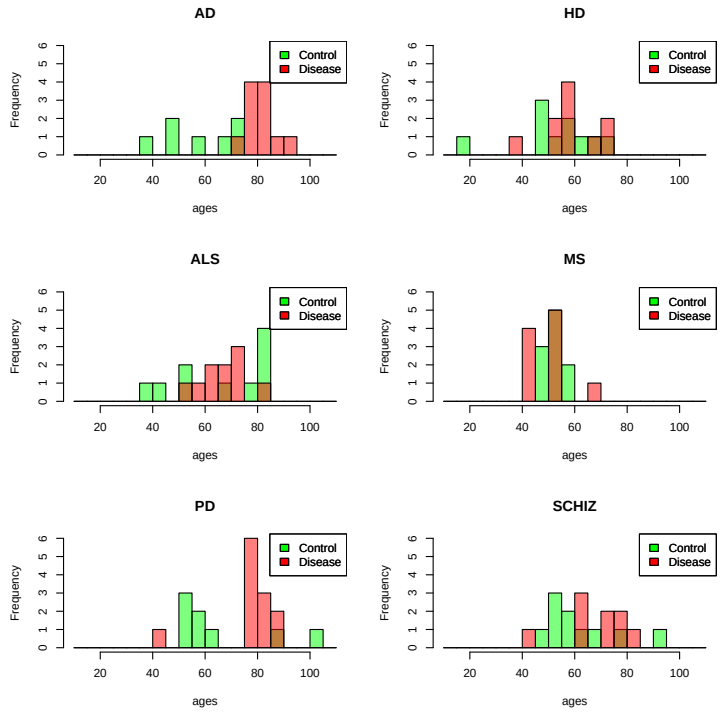


Figure 3: Patients age distribution across disease and control groups.

1.3 Data Split

As we were well aware that performing the differential gene expression and the predictive analyses on the whole dataset was not only methodologically fallacious but could also lead to excessively optimistic results, we proceeded by splitting each control and disease group in two disjoint sets with a 70-30% ratio, one for training and validation and the other for testing, for a total of 12 sets, 2 for each disease. We paid great attention so that each split presented the same statistics, and most importantly, the same disease-control proportions of the whole sample. Fortunately for us the `caret` Khun [2008] R package offered many utilities to perform such operations.

1.4 Differential Gene Expression Analysis (DGEA)

We performed two experiments using the `limma` Ritchie et al. [2015] R package: one aimed at measuring different expressions between the control and the disease groups for each of the six pathologies in our dataset whilst the other between every possible disease-disease pair. Results are pictured in Figure 14 and Figure 15 in the Appendix. In both experiments we filtered the differences using a p-value threshold of 0.05, then, for sake of simplicity, we inspected only the first 100, sorting them by ascending p-value. We measured statistically significant differences only for 3 out of the 6 disease and control populations: more than 100 genes for HD, 32 for PD and 34 for ALS. On the contrary, the second experiment highlighted many more differences in gene expression between the populations affected by the 6 diseases, in fact, we spotted statistically significant differences for 12 out of the possible 15 pairs, with the exception of AD-SCHIZ, AD-MS and MS-SCHIZ.

1.5 Predictive Analysis and Feature Selection (FS)

Having extracted a consistent pool of candidates biomarkers, we designed a predictive pipeline using the `scikit-learn` Buitinck et al. [2013] Python library. Our objective was twofold: to estimate the predictive power of the genes from the differential expression analysis, further refining them, and to extract some for AD, MS and SCHIZ since which we still had none.

Model While building the pipeline we followed Occam's razor principle and we kept its structure as simple as possible. After a preliminary model selection we opted to stick with a 3 stage pipeline structured as follows: first a z-score normalizer followed by a feature selector in the form of a logistic regression with L1 penalty and, at last, a grid-search CV with a random forest tree as the classifier. Do note that we also attempted classification with a decision tree, as it was far more interpretable than the random forest, but the latter achieved better results on the validation set.

Experiments In our first batch of experiments we fed in input to our pipeline the whole set of 20589 genes, merging the train and validation splits for each disease, effectively creating 2 monolitic train and validation sets. In such way we used 84 patients for training and validation and 34 for testing. This first attempt at a multi-disease classifier failed to produce any meaningful results, as the performances were not satisfactory and the logistic regressor failed to converge even for very high number of iterations. We then proceeded to train 6 binary pipelines, having the same structure as before, each discriminating among one disease and control group. We fed in input to these models the genes from the differential expression analysis, if available, or all the 20589 genes if not. This attempt produced far better results, which are depicted trough Table 5 to Table 10 in the Appendix, and we were able to achieve perfect a classification score at testing time on HD, ALS and SCHIZ. It was thus notable that genes isolated by the differential expression analysis for HD and ALS presented a remarkable predictive power.

Feature Sets We extracted 2 set of features from each of the 6 binary pipelines, one from the logistic regression and one from the random forest, for a total of 12. We then trained 2 multi-disease classifier: one on the set-union of the 6 sets extracted from the 6 logistic regressions and the other one on the set-union of the 6 from the random forests.

Multi-class Predictor The set union of the features selected by the 6 different logistic regressions was coarser and counted roughly 5K genes while the one resulting from merging the features extracted from the 6 different random forest classifiers counted only 21 genes. We built a new pipeline made of 2 stages: z-score normalization and random forest classifier with grid-search CV. In the end, we were able to achieve satisfactory scores for the pipeline trained on the second, smaller, set of 21 genes, as shown in Table 2. Hyperparameters configurations for both single-disease and multi-disease pipelines can be found in Table 11 in the Appendix. The list of the 21 genes together with the importance scores from the multi-disease classifier can be found in Table 4.

Class	Precision	Recall	F1 Score	Support
AD	1.0	1.0	1.0	3
HD	1.0	0.67	0.80	3
PD	1.0	1.0	1.0	3
ALS	1.0	1.0	1.0	3
MS	1.0	1.0	1.0	3
SCHIZ	1.0	0.67	0.80	3
Control	0.89	1.0	0.94	16

Table 2: Test set classification report for the final multi-disease classifier.

We also tried experimenting with the genes extracted from the disease-disease pairs during the differential gene expression analysis but with no avail whatsoever.

1.6 Clustering

We continued our study by performing several clustering runs. Emerging patterns in the data, both in terms of patients and genes, could indicate similarities between groups of individuals affected by the 6 pathologies and could hint to us meaningful findings relevant to their diagnosis.

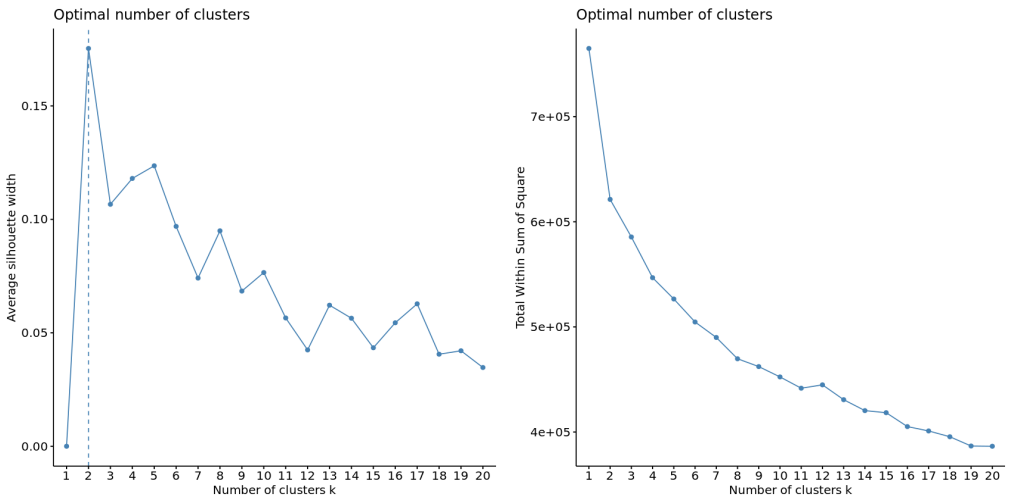
We designed a clustering pipeline using the `factoextra` Kassambara et al. [2020] R package. Such library provided many algorithms, including prototype-based approaches, such as k-means, and hierarchical ones. The application of the former was preceded by a preliminary analysis of the optimal number of centroids by inspecting silhouette plots and elbow plots.

As pre-processing steps, we applied the natural logarithm to the whole gene expression matrix, substituting all null entries with 1 for numerical reasons, and a z-score normalisation.

To further investigate the relevance of clusters, experiments performed over the entire expression matrix were always paired with experiments performed on the matrix masked over the 21 genes previously extracted from the multi-disease classifier and shown in Table 4.

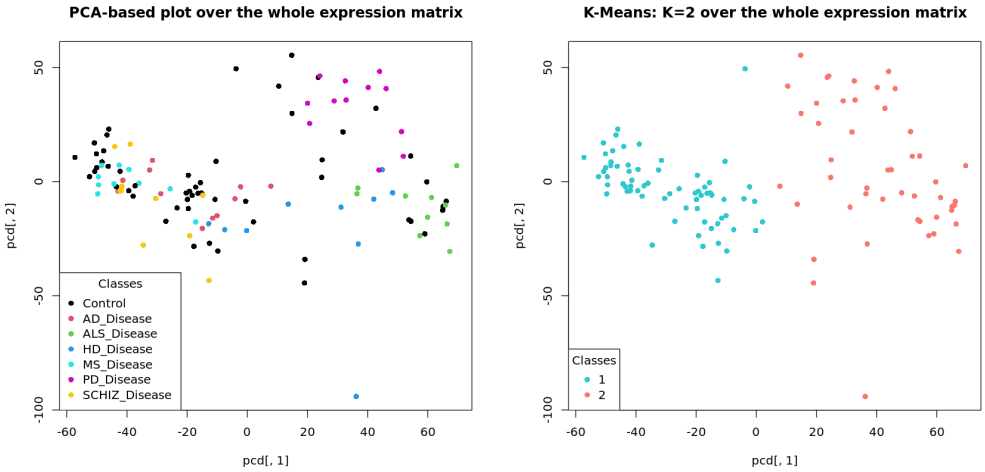
In each experiment, cluster labels were obtained for each set of patients and genes of interest and such labeling was subsequently used to highlight the samples in the space spanned by the first 2 principal components of the matching subset of rows, for genes, or columns, for patients, from the expression matrix. Furthermore, we performed several experiments by leveraging the t-SNE algorithm with the aim of highlighting potentially interesting structures in the data but such endeavour did not ultimately lead to any meaningful results.

Patients Clustering As already mentioned, fairly common techniques, such as studying variation of silhouettes and the elbow method, were used to determine the optimal number of clusters. Figure 4 displays the result of this preliminary analysis.



(a) (b)

Figure 4: (a) Silhouette. (b) Elbow-method for k-means on the whole gene set.



(a) (b)

Figure 5: (a) PCA-based plot of the patients on whole gene set. (b) k-means clustering of said patients for k = 2.

The silhouette peaks shown in Figure 4 were used as candidates for the optimal value of k. This survey ultimately produced many clusterings, which we inspected extensively. The most significant result we observed during this study was the frequent joint clustering of patients affected by AD, MS and SCHIZ. As a matter of fact, the differential gene expression analysis failed to produce statistically meaningful differences for the following pairs of diseases: AD-SCHIZ, AD-MS and MS-SCHIZ, as can be seen from Figure 15.

An example of such a grouping can be seen in Figure 5. This result, although promising, could have been caused by the high number of unfiltered genes, for this reason we carried out a further exploration, which we described in the following section. Another, but perhaps weaker, similarity was noted between patients affected by HD and ALS. Furthermore, patients with PD were often well isolated from the others in many of our experiments.

We continued our clustering analysis on the patients population by shifting to hierarchical methods. Such effort led to conclusions very much in line with what already observed for the prototype-based clustering. Figure 6 displays the dendrogram resulting from the application of said technique to the entire expression matrix. As a final remark, the application of both hierarchical and prototype-based clustering techniques to the 21 genes isolated by means of the multi-disease classifier did not offer any other meaningful insight, nevertheless, we observed a neater separation of the patients affected by different diseases on such set of genes.

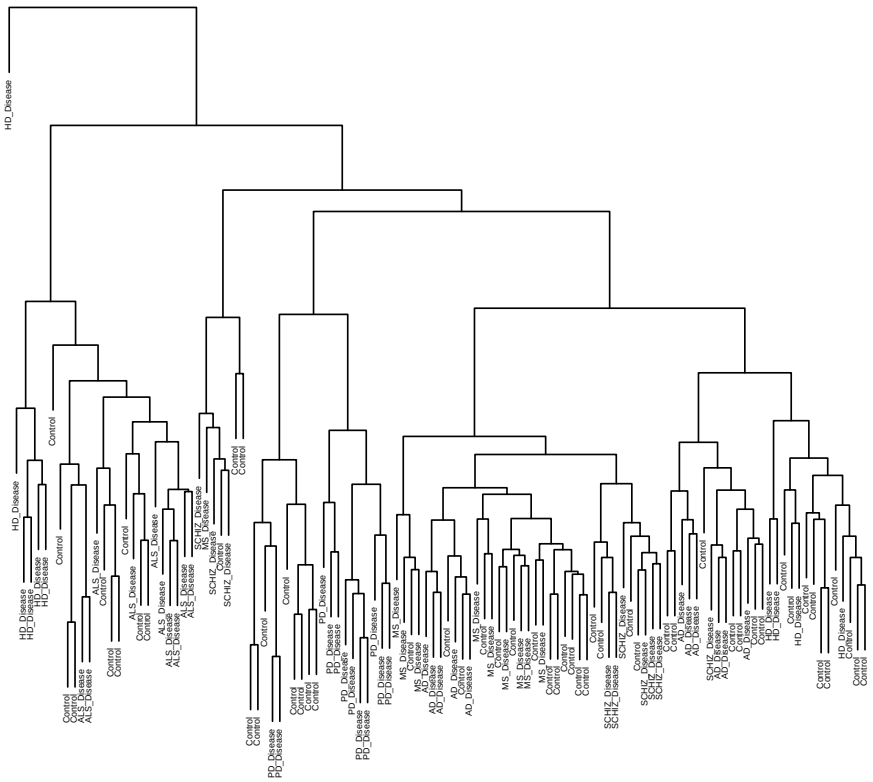
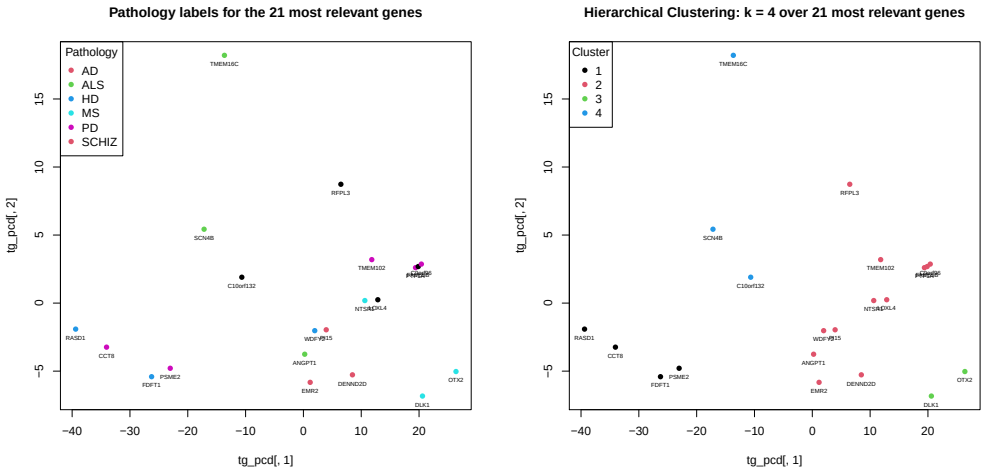


Figure 6: Dendrogram resulting from the hierarchical clustering of the patients represented by the expression matrix.

Genes Clustering Due to the prohibitive complexity of applying the clustering pipeline just described to the entire set, counting well over 20K genes, we narrowed the analysis to the 21 most relevant ones extracted from the multi-disease classifier.

We fed in input the corresponding rows of the pre-processed gene expression matrix to both the k-means and the hierarchical clustering algorithms, hoping to gain more insights on the previous results. We derived a series of labels, mapping each of the 21 genes to a disease, by associating to each gene the disease for which the binary classifier that isolated it was originally trained on. These labels greatly helped in characterizing clusters, as shown in Figure 7, and in identifying similarities between genes involved in the diagnostic process of SCHIZ and MS such as: RASD1, CCT8, FDFT1 and PSME2.



(a) (b)

Figure 7: (a) PCA-based plot of the 21 most relevant genes according to the classifiers and (b) hierarchical clustering of said genes for k = 4.

None of the 21 labels accounted for a gene being simultaneously relevant for the diagnosis of multiple diseases. Nevertheless, we performed pairwise intersections between the 6 sets of genes isolated by the logistic regressors of the 6 binary pipelines and we were able to identify several hundreds common genes suggesting synergies between the 6 pathologies, with many co-occurrences for AD, MS and SCHIZ. In order to refine this results and to further quantify the predictive power of our multi-disease classifier, we projected the intersection of these common genes on the set from Table 4. The results are shown in Table 3.

Pathologies	Diagnostically Relevant Genes
AD, MS	NTSR1
AD, PD	NTSR1
AD, SCHIZ	PTF1A
ALS, SCHIZ	DENND2D
MS, PD	NTSR1
MS, SCHIZ	WDFY2, TMEM102

Table 3: List of genes diagnostically relevant for more than one pathology.

1.7 Pathway (PW) Enrichment Analysis

As the final step of our study we performed a pathway enrichment analysis. We had gathered some interesting results during the previous phases and the next challenge was to provide further context and an explanation to both the predictive power shown by the multi-disease classifier and to the patterns emerged during the clustering analysis.

Protein to Protein Interaction We first focused on single diseases, carrying out a protein to protein interaction analysis using the genes highlighted via the DGEA in Section 1.4 for HD, PD and ALS, and those selected by the logistic regressors in Section 1.5 for AD, ALS and MS. It's important to note that the former group was smaller than the latter, comprising roughly 150 genes versus 5000.

The network was built as follows: we computed a Pearson correlation matrix using the expression value of each gene, then, we used it to build an expression graph. Nodes were connected by an edge if the positive correlation score between the corresponding genes was greater than 0.9 or the negative correlation was smaller than -0.6. This representation had richer expressive power with respect to a simple heatmap and enabled us to easily visualize groups of correlated genes. The so obtained gene networks are depicted in Figure 8. Such networks were very big and densely connected, for this reason we decided that a in-depth analysis was out of the scope of our work.

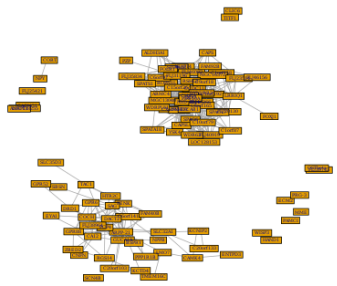
Enrichment Map An enrichment map was built as in Reimand et al. [2019], by querying via `g:Profiler` Raudvere et al. [2019] the genes for each disease. We used the DGEA selection from Section 1.4 for HD, ALS and PD and the FS selection from Section 1.5 for AD, MS and SCHIZ. The setup for the 6 queries was the following:

1. Query run as ordered one.
2. Only annotated genes for statistical domain scope.
3. g:SCS threshold for significance threshold.
4. p-value ≤ 0.01 for user threshold.
5. GO biological process, Reactome, KEGG for data sources.
6. Filter pathways with less than 5 genes or more than 2000.

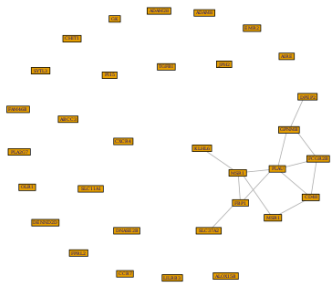
We observed an unbalance in the number pathways retrieved for DGEA and FS groups due to their different sizes. Nevertheless, not only we were able to link the sub-networks but we also leveraged the smaller communities resulting from HD, PD and ALS as a starting point to study their structures.

After loading the pathway list in the Cytoscape using the EnrichmentMap addon, we used AutoAnnotate to perform clustering over the pathway graph, greatly improving the network interpretation, as we were able to separate communities and spotted crucial pathways linking diseases sub-networks.

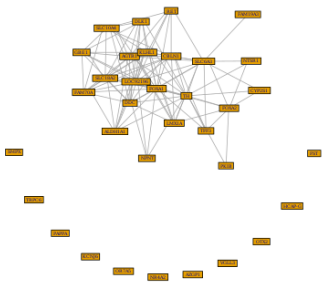
The final enrichment map had 403 nodes, 4472 edges and a total of 16 connected components. The map is shown in its integrity in Figure 9. We highlighted in red the edges between pathways comprising at least one the 10 genes with highest feature rank from Table 4. Relevant pathways were well localized in tightly connected clusters. Moreover, pathways including the 21 genes constituted the 45% of the enrichment map.



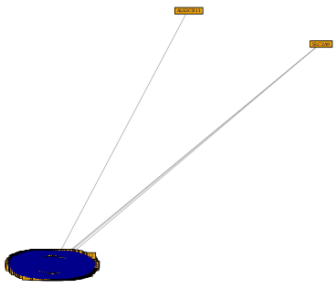
(a) HD



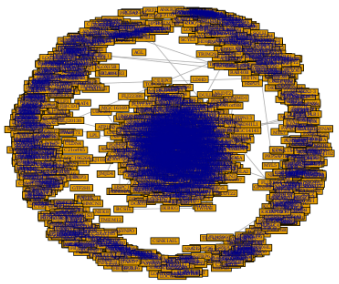
(b) ALS



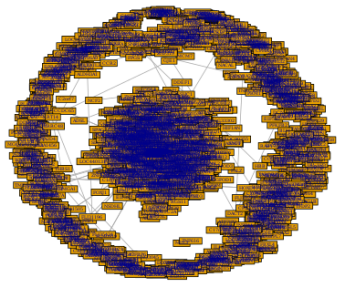
(c) PD



(d) AD



(e) MS



(f) SCHIZ

Figure 8: P2P interaction networks.

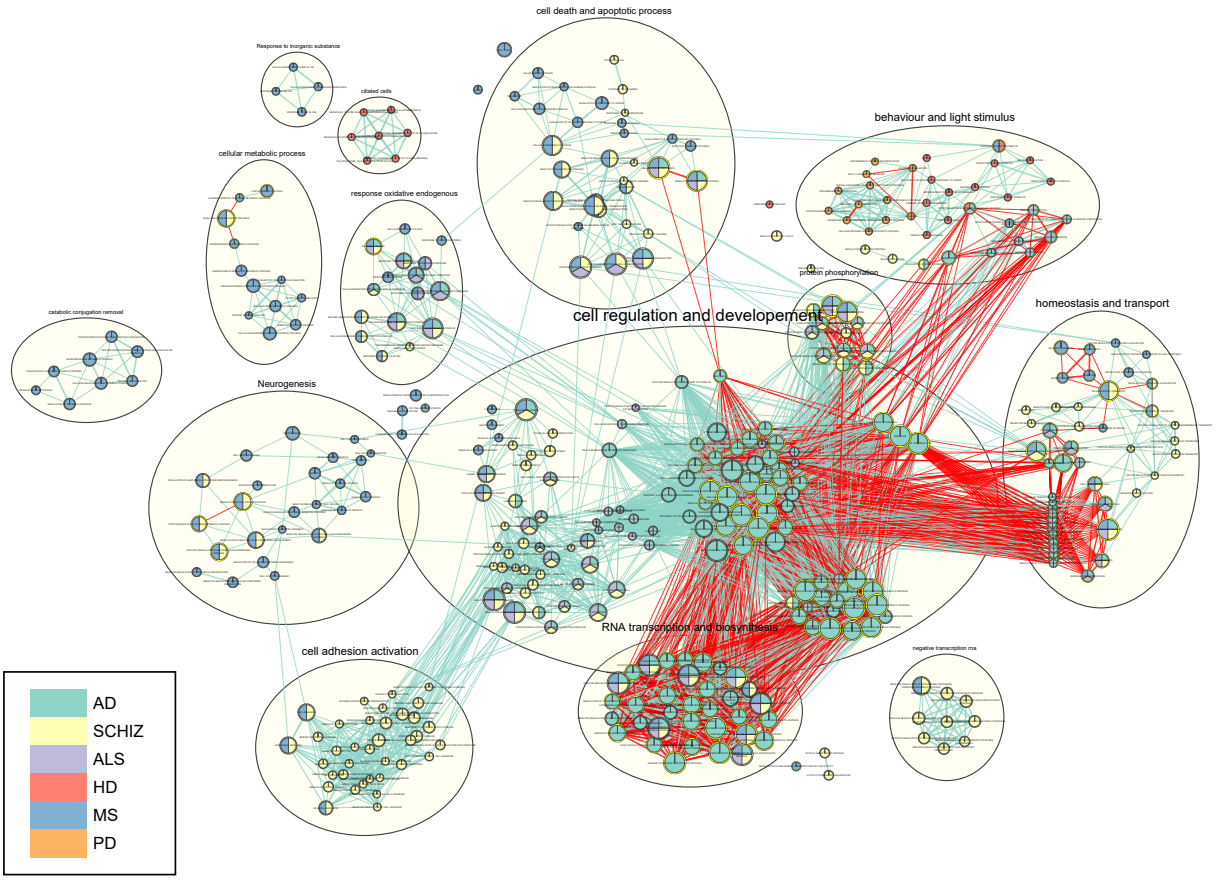


Figure 9: Enrichment map. Nodes are pathways: they are colored accordingly to the disease that genes partaking in them belonged to and their size is proportional to the number of such genes. An edge links two pathways if they share at least a gene and its thickness is proportional to the number of shared genes.

Enrichment Map Analysis As Figure 9 shows, pathways retrieved for MS and SCHIZ were always in the same communities, except for those in the *response to inorganic substances* and *catabolic processes* clusters that were retrieved only for MS. The ALS pathway set was a subset of AD with strong coverage, as only 5 pathways were missing from the former to completely overlap the latter. Do note that, as a consequence, since the DGEA was able to isolate genes for ALS, the same such genes could have offered insights on AD.

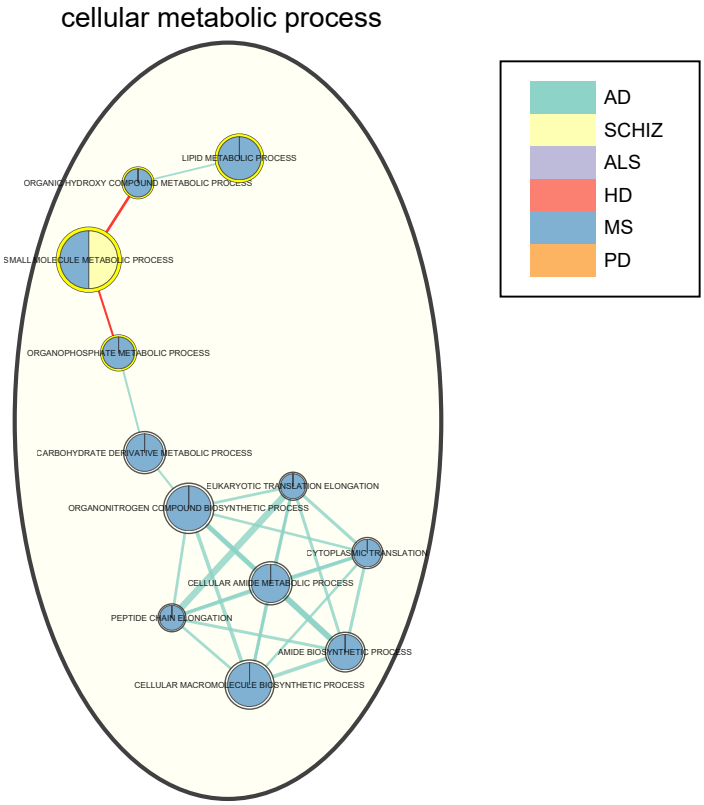


Figure 10: Pathways involving FDFT1 gene.

In the following we will focus on some note-worthy genes; their ranking, according to the multi-disease classifier, is reported in Table 4. The gene with the highest rank is FDFT1, which is involved in small molecule metabolic process and lipid metabolic process pathways, shown in Figure 10. This gene is related to cholesterol synthesis, Stefan et al. [2021]. As pictured in Figure 11 *G protein-coupled receptor signaling pathway* (GPCRs) links pathways retrieved for HD, ALS and AD. GPCRs regulate the majority of known physiological processes and have a prominent position in the clinic, moreover, defects in the GPCRs can be found in various dysfunctions and diseases [1]. Furthermore, we found a study Azam et al. [2020] proposing a therapy targeting GPCRs for neurological diseases such as PD. The HD pathways in Figure 11 link the ones retrieved for PD through the *behaviour* and *locomotory behaviour* pathways; behaviour and locomotion are compromised by the two diseases. Furthermore, the *decreased and abnormal facial expression* pathways share the DLK1 gene, which is also used by the multi-disease classifier. Another relevant pathway present in Figure 11 is the *dopaminergic neuron differentiation* and is the only one in the enrichment map that includes the OTX2 gene, 4th in Table 4. Moreover, the Neurotensin receptor 1 (NTSR1) is involved in 60.6 % of relevant pathways in our network. As stated in NCBI [2022] the gene belongs to the large super-family of G-protein coupled receptors and mediates the multiple functions of neurotensin, such as hypotension, hyperglycemia, hypothermia, antinociception, and regulation of intestinal motility and secretion. Pathways involving the WDFY2 gene, 10th in Table 4, are condensed in a small connected component related to protein phosphorilation, as shown in Figure 12; WDFY2 is involved in cognitive decline in older adults Kamboh et al. [2019].

[1] https://www.cusabio.com/pathway/G-protein-coupled-receptor-signaling-pathway.html

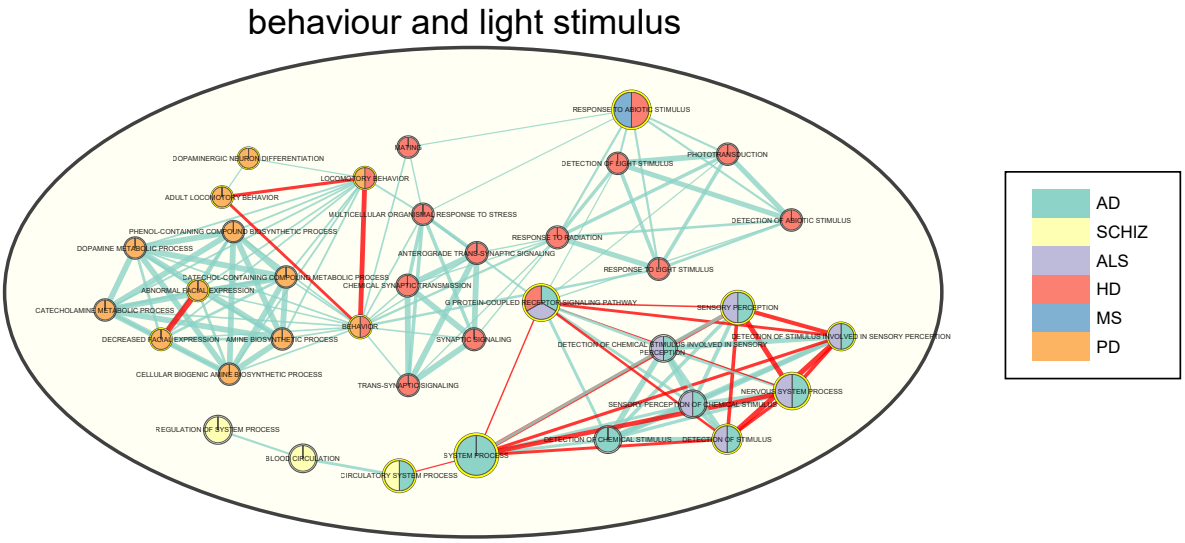


Figure 11: Community surrounding GPCRs pathway.

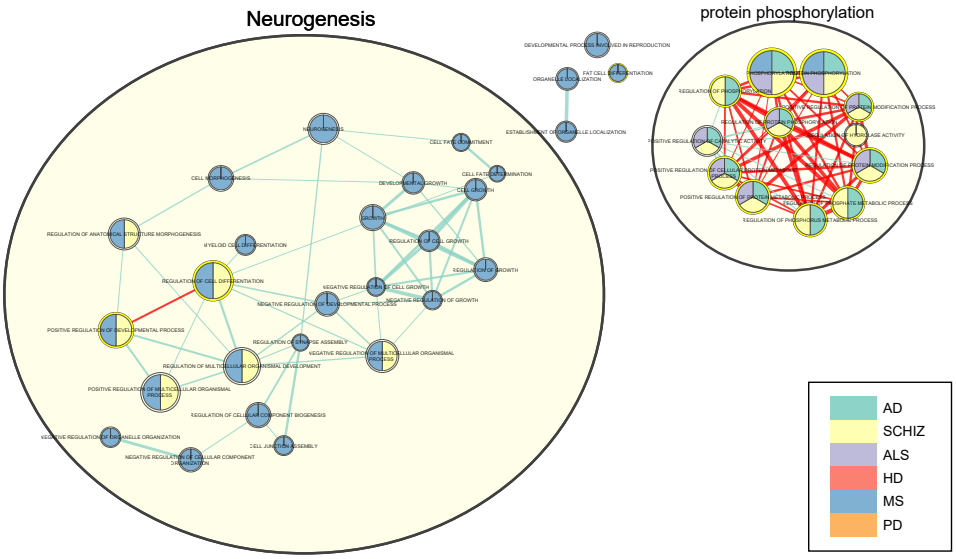


Figure 12: pathways involving WDFY2 gene.

TMEM102 is involved in regulation of mitochondrial outer membrane permeabilization and in apoptotic signaling pathway; response to cytokine, and signal transduction, acts upstream of or within positive regulation of T cell migration and positive regulation of cell adhesion NCBI [2022].

RFPL3 and PTF1A are active in almost every overlapping pathways. The union of the two sets is depicted in Figure 13. PTF1A is thought to be involved in the maintenance of exocrine pancreas-specific gene expression including elastase 1 and amylase. Mutations in this gene cause cerebellar agenesis and loss of expression is seen in ductal type pancreas cancers Sellick et al. [2004].

As a final consideration, 8 out of the 20 genes used by the classifier the query on g:Profiler did not return any related pathway.

Rank	Gene	Importance	retrieved pathways
1	FDFT1	0.1014	small molecule metabolic process, lipid metabolic process
2	EMR2	0.0940	
3	DENND2D	0.0831	
4	OTX2	0.0747	dopaminergic neuron differentiation
5	LOXL4	0.0696	localization, cellular protein modification process, cellular macromolecule metabolic process, transport, protein metabolic process, organonitrogen compound metabolic process, macromolecule modification, protein modification process, cellular protein metabolic process [...]
6	C10orf132	0.0584	
7	SCN4B	0.0537	
8	NTSR1	0.0524	positive regulation of biological process, cell surface receptor signaling pathway, regulation of nitrogen compound metabolic process, organophosphate metabolic process, cellular ion homeostasis, regulation of primary metabolic process, positive regulation of apoptotic process, transmembrane transport [...]
9	DLK1	0.0514	decreased facial expression, abnormal facial expression
10	WDFY2	0.0441	protein phosphorylation, regulation of protein phosphorylation, Factor: ER71; motif: ACCGGAARYN; match class: 1, regulation of protein modification process, fat cell differentiation, phosphorylation, positive regulation of cellular protein metabolic process, regulation of phosphorus metabolic process [...]
11	PSME2	0.0440	antigen processing and presentation, positive regulation of cellular protein metabolic process, positive regulation of molecular function, positive regulation of catalytic activity, regulation of hydrolase activity, positive regulation of protein metabolic process, regulation of cell cycle [...]
12	ANGPT1	0.0434	
13	TMEM102	0.0370	programmed cell death, regulation of cell motility, regulation of protein modification process, biological adhesion, apoptotic process, cell migration, apoptotic signaling pathway, cell chemotaxis, regulation of apoptotic process, positive regulation of cell motility, regulation of programmed cell death, regulation of locomotion [...]
14	RASD1	0.0356	negative regulation of nucleobase-containing compound metabolic process [...]
15	CCT8	0.0349	positive regulation of nucleobase-containing compound metabolic process, positive regulation of biosynthetic process, positive regulation of macromolecule biosynthetic process, positive regulation of cellular biosynthetic process [...]
16	RFPL3	0.0331	positive regulation of biological process, regulation of macromolecule metabolic process, cellular nitrogen compound biosynthetic process, regulation of nitrogen compound metabolic process, regulation of primary metabolic process, positive regulation of cellular biosynthetic process, positive regulation of cellular process [...]
17	TMEM16C	0.0279	
18	PI15	0.0215	
19	PTF1A	0.0209	positive regulation of biological process, regulation of nitrogen compound metabolic process, regulation of primary metabolic process, regulation of nucleic acid-templated transcription, cellular biosynthetic process, developmental process, positive regulation of nitrogen compound metabolic process, aromatic compound biosynthetic process [...]
20	FAM55B	0.01804	
21	C9orf96	0.0	

Table 4: Genes used by the multi-disease classifier together with importance score and common pathway memberships.

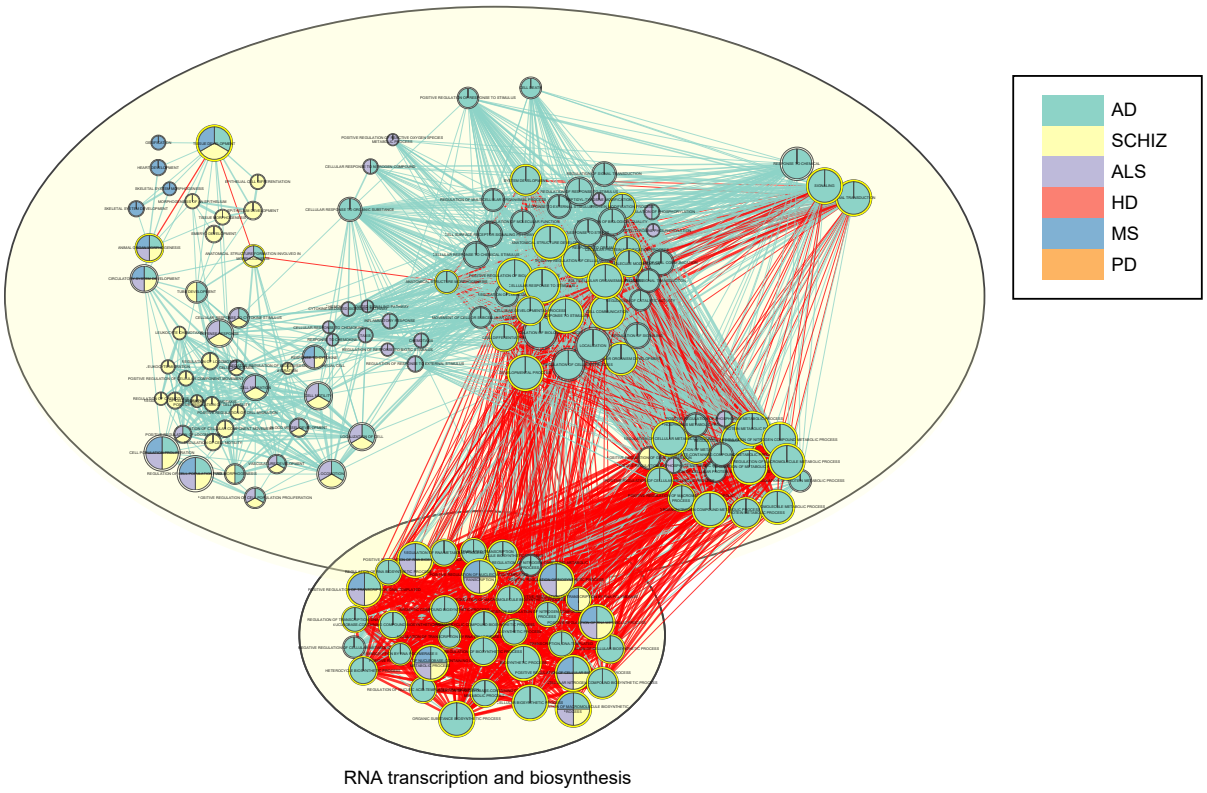


Figure 13: Pathways involving RFPL3 and PTF1A genes.

2 Discussion and Conclusion

In this work we leveraged machine learning techniques as a core tool for extracting relevant bio-markers in a population of patients affected by 6 neurodegenerative diseases. We succeeded in extracting relevant bio-markers for all of the diseases and validated their predictive power using simple learning algorithms. We further demonstrated how such tools are invaluable helpers not only for scientists but also for clinicians, since our final model achieved close to optimal results in diagnosing patients. Combining differential gene expression and learning algorithms was of paramount importance to get achieve working models and, most importantly, satisfactory performances.

In the final stages of our case study we focused on providing an explanation to the behavior of the classifier and to the patterns found in the clustering process. We tried to isolate the biological processes involving the genes with highest relevance score according to our model, although a proper interpretation of the results of our study would require trained biologists and clinicians. However, we noted that most of the genes used by our predictive model have well known relations with biological processes of the nervous system.

References

Durrenberger P.F., Fernando F.S., Magliozzi R., Kashefi S.N. et al. *Selection of novel reference genes for use in the human central nervous system: a BrainNet Europe Study.* Acta Neuropathol 2012 Dec;124(6):893-903.

Sui Y, Zhao X, Speed TP, Wu Z. *Background adjustment for DNA microarrays using a database of microarray experiments.* J Comput Biol. 2009;16(11):1501-1515. doi:10.1089/cmb.2009.0063

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.

Kuhn, M. (2008). *Building Predictive Models in R Using the caret Package.* Journal of Statistical Software, 28(5), 1 - 26. doi:http://dx.doi.org/10.18637/jss.v028.i05

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, Fabian, Mueller, A., Grisel, O. and G. Varoquaux. (2013). *API design for machine learning software: experiences from the scikit-learn project.* In ECML PKDD Workshop: Languages for Data Mining and Machine Learning (pp. 108–122).

U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, J. Vilo. *g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update).* Nucleic Acids Research 2019; doi:10.1093/nar/gkz369.

Azam, S.; Haque, M.E.; Jakaria, M.; Jo, S.-H.; Kim, I.-S.; Choi, D.-K. *G-Protein-Coupled Receptors in CNS: A Potential Therapeutic Target for Intervention in Neurodegenerative Disorders and Associated Cognitive Deficits.* Cells 2020, 9, 506. https://doi.org/10.3390/cells9020506

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D., and Bader, G. D. (2019). *Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap.* Nature protocols, 14(2), 482–517. https://doi.org/10.1038/s41596-018-0103-9

Kassambara, A. and Mundt, F. (2020) *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses.*

M. Ilyas Kamboh, Kang-Hsien Fan, Qi Yan, Joanne C. Beer, Beth E. Snitz, Xingbin Wang, Chung-Chou H. Chang, F. Yesim Demirci, Eleanor Feingold, Mary Ganguli (2019).*Population-based genome-wide association study of cognitive decline in older adults free of dementia: identification of a novel locus for the attention domain* Neurobiology of Aging, 84, 239.e15-239.e24.https://doi.org/10.1016/j.neurobiolaging.2019.02.024

Sellick, Gabrielle S, Barker, Karen T, Stolte-Dijkstra, Irene, Fleischmann, Christina, J Coleman, Richard, Garrett, Christine, Gloyn, Anna L, Edghill, Emma L, Hattersley, Andrew T, Wellauer, Peter K, Goodwin, Graham, Houlston, Richard S (2004). *Mutations in PTF1A cause pancreatic and cerebellar agenesis* Nature Genetics, 36, 1301-1305. https://doi.org/10.1038/ng1475

Stefan A. Berghoff, Lena Spieth, Ting Sun, Leon Hosang, Constanze Depp, Andrew O. Sasmita, Martina H. Vasileva, Patricia Scholz, Yu Zhao, Dilja Krueger-Burg, Sven Wichert, Euan R. Brown, Kyriakos Michail, Klaus-Armin Nave, Stefan Bonn, Francesca Odoardi, Moritz Rossner, Till Ischebeck, Julia M. Edgar, Gesine Saher (2021). *Neuronal cholesterol synthesis is essential for repair of chronically demyelinated lesions in mice*, Cell Reports, Volume 37, Issue 4, 2021, 109889, ISSN 2211-1247. https://doi.org/10.1016/j.celrep.2021.109889.

National Center for Biotechnology Information. *PubChem Gene Summary for Gene 4923, NTSR1 - neurotensin receptor 1 (human).* https://pubchem.ncbi.nlm.nih.gov/gene/NTSR1/human. Accessed May 16, 2022.

National Center for Biotechnology Information. *PubChem Gene Summary for Gene 284114, TMEM102 - transmembrane protein 102.* https://www.ncbi.nlm.nih.gov/gtr/genes/284114/. Accessed May 16, 2022.

Kaliszewski, M., Knott, A. and Bossy-Wetzel, E. Primary cilia and autophagic dysfunction in Huntington's disease. Cell Death Differ 22, 1413–1424 (2015). https://doi.org/10.1038/cdd.2015.80

Appendix

Class	Precision	Recall	F1 Score	Support
Control	0.67	1.0	0.80	2
Disease	1.0	0.67	0.80	3

Table 5: Test set classification report for the AD classifier.

Class	Precision	Recall	F1 Score	Support
Control	1.0	1.0	1.0	3
Disease	1.0	1.0	1.0	3

Table 6: Test set classification report for the HD classifier.

Class	Precision	Recall	F1 Score	Support
Control	1.0	0.67	0.80	3
Disease	0.75	1.0	0.86	3

Table 7: Test set classification report for the multiple sclerosis disease classifier.

Class	Precision	Recall	F1 Score	Support
Control	1.0	0.5	0.67	2
Disease	0.75	1.0	0.86	3

Table 8: Test set classification report for the PD classifier.

Class	Precision	Recall	F1 Score	Support
Control	1.0	1.0	1.0	3
Disease	1.0	1.0	1.0	3

Table 9: Test set classification report for the amyotrophic lateral sclerosis disease classifier.

Class	Precision	Recall	F1 Score	Support
Control	1.0	1.0	1.0	3
Disease	1.0	1.0	1.0	3

Table 10: Test set classification report for the schizophrenia disease classifier.

Model	Trees (RF)	Alpha (RF)	Max Iters (LR)	Solver (LR)	Folds (CV)
Single Disease	[2, 3, 4, 5]	[$1e^{-5}$, $1e^{-4}$]	[1000, 2000, 3000]	`saga`	2
Many Disease	[4, 5, 7, 8]	[$1e^{-4}$, $5e^{-4}$, $1e^{-3}$, $5e^{-3}$]	3000	`saga`	[3, 5]

Table 11: Hyperparameters configurations used in the model selection phase. RF stands for random forest, LR for logistic regression and CV for cross validation.

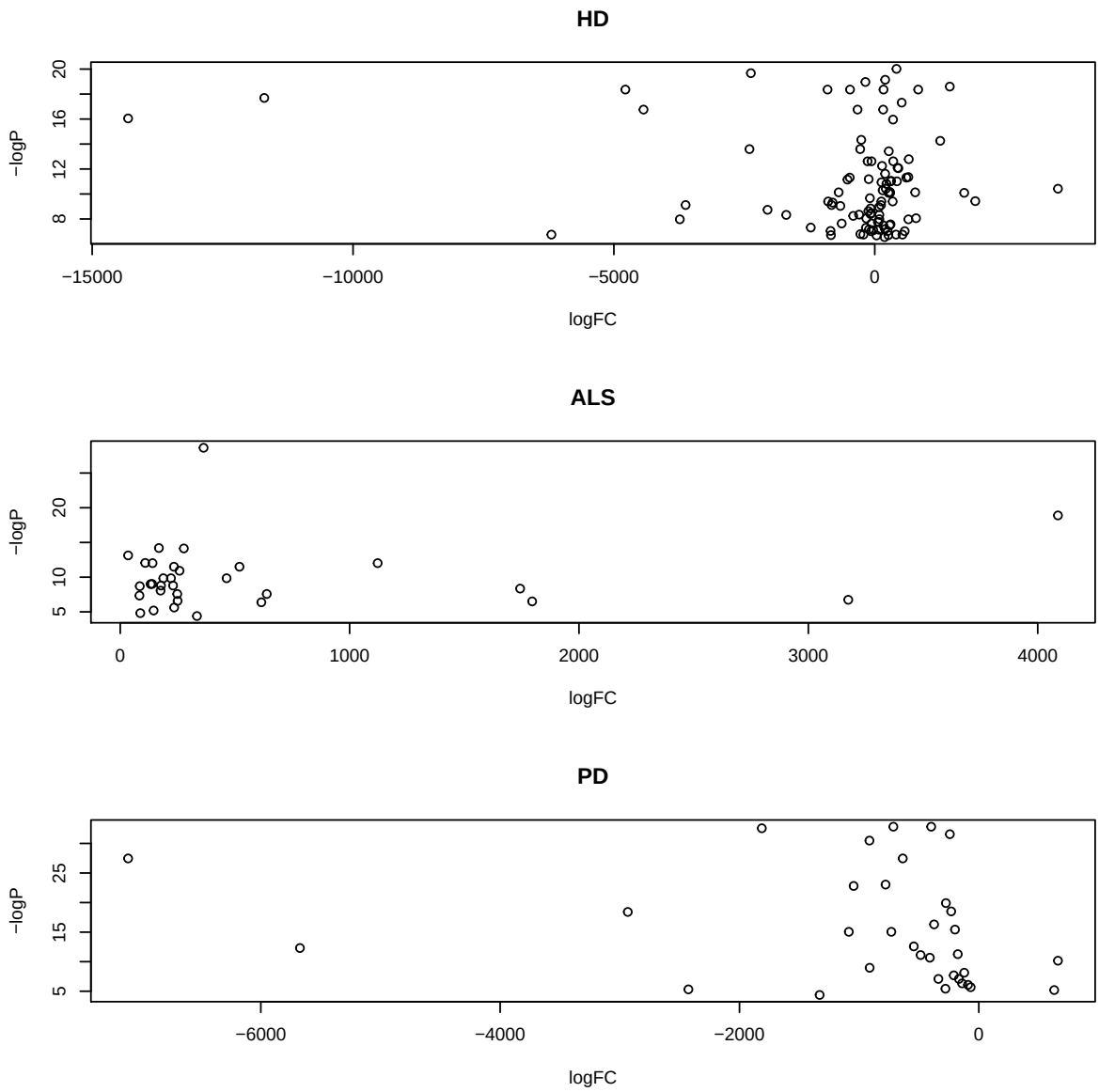


Figure 14: Volcano plot with the results from the differential gene expression analysis conducted on the 6 disease-control groups.

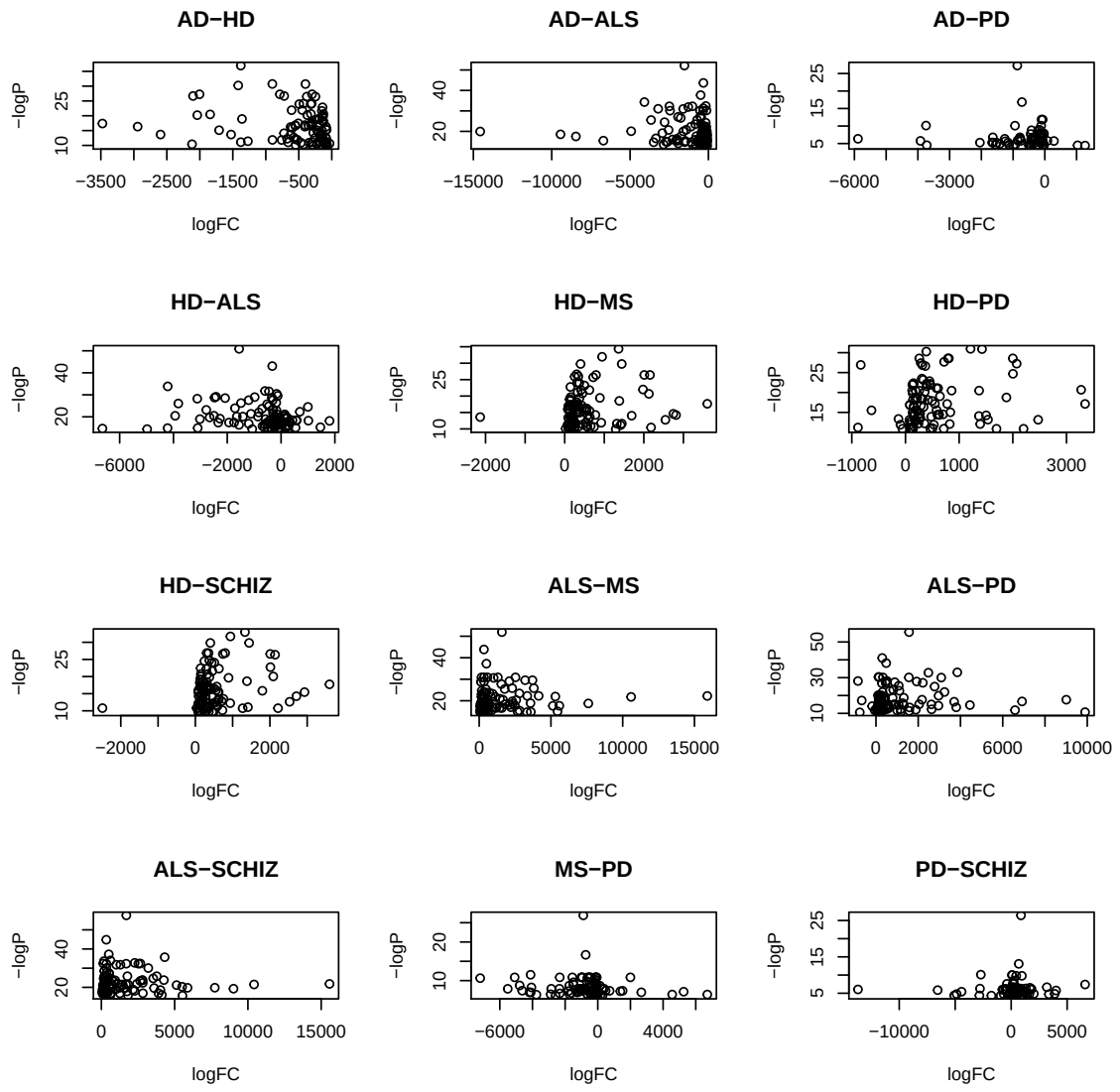


Figure 15: Volcano plot with the result from the differential gene expression analysis conducted on all the possible disease-disease pairs.