Analysis of RNA expression on different amyotrophic lateral sclerosis subgroups

Filippo Alberto Mirolo, 239906

September 8, 2023

1 Abstract

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease caused by death of motor neurons in the brain and in the spinal cord. With the progression of the disease the loss of neurons involves muscular paralysis and even death. ALS is a complex disease which can arise in a specific location, for then spreads in the organism. The uprising of ALS is due to a large number genetic factors, poorly mapped and understood at this date [MP20]. In addition ALS can be classified depending on his phenotype and how much severe the symptoms are. As sad so, in this project I exploit data from the GEO (Gene Expression Omnibus) database (accession number GSE212131), which contains gene expression profile of patients with severe and less uncompromising phenotype. Literature [Car23, Sin20] correlates ALS with pro inflammatory environment and harmful immune response; according to these result my analysis found the two macro terms as differential expressed in the two groups.

2 Introduction

ALS is a very fast progressive and severe neurological disease, for which there are only few and not really performing treatments. The main reason behind the difficulty in finding good and reliable method to tackle the pathogenesis, is due to the really high genetic heterogeneity of ALS. Nowadays the scientific community is trying to find more and more genetic features to characterize the disease with the aim to develop better strategies to treat patients. Right now there are a bunch of gene recognized to be impactful on arising the ALS, the mains ones are: chromosome 9 open reading frame 72 (C9orf72), superoxide dismutase 1 (SOD1), TAR DNA-binding protein 43 (TARDBP), fused in sarcoma (FUS) and TANK-binding kinase 1 (TBK1). In this project, the data analysed are composed by two groups: a short (more severe) duration of the disease (less than twelve months), a long (more feeble) duration of the disease (greater than 6 years). The aim is to compare the transcription profile and distinguish some key features to differentiate these two phenotype. The analysis is composed of three main steps, an exploratory analysis, comprehensive of principal component analysis (PCA), and feature selection. Second, were used few classification methods (logistic regression, ridge regression, lasso regression, support vector machines, etc) to extract the most predictive features (genes) of short and long subgroups. Lastly, a more deep analysis was performed on the list of gene extract from the most performing model, a comparison with the current knowledge was also exploit. All the script are available at this GitHub link.

3 Materials and methods

3.1 Data set

As sad before the data used are taken from the GEO database (accession number GSE212131). This data set contain 42 samples (patients) and 22011 variables (genes), with a tiny part as control variables; coming from Affymetrix microarray technology. The 42 samples are spilt in:

- 22 patients who suffered a short (less then 12 months) disease duration
- 20 patients who suffered a long (greater then 6 years) disease duration

Transcriptomics profile were extract for both groups from lymphoblastoid cell lines (LCLs).

3.2 Methods

This project rest on a variety of advance methods to extract information from data, as supervised and unsupervised methods. In this section will be briefly presented them. Starting with an exploratory analysis of the data I begin with a box plot to check the data distribution and eventually to find out some anomalies. Then were performed PCAs and several clustering methods, the PCAs were conducted before and after the feature selection (done with a T test). The comparison of the two PCAs highlights a major difference before and after the restriction of the variables number. K-means and hierarchical clustering were also performed with a k value of 2, because the data set is compose by two classes. For all the project the classes will be defined as Short for the more severe phenotype and Long for the more feeble one. After that I continue testing some methods, but before I had to split the samples in training and test set and keep them for all the models. Random Forrest was the first applied and then was performed Ridge regression, Lasso regression, Support Vector Machines and Scudo. Using the best performing method I extract, from the sub set of gene selected with the T test, 200 genes. These were used to perform Over Representation Analysis with *GProfiler* then the same 200 genes were also used to assert gene base network analysis with *STRING* and *PathfindeR*. Lastly I compare my result with the information available in the literature. For further information check the script at this link.

4 Results

4.1 Principal Component Analysis

As first approach of the exploratory analysis I performed a PCA with all the variables (genes) contained in the data set. From this plot (figure 1) is not possible to recognize a pattern of division. After I have performed a feature selection through a T test with a 0.01 p-value, I kept all the variables with a p-value below the threshold (0.01) and construct a new data set with 575 features. I have again performed a PCA (figure 1), in which is now clear a better clustering of the samples. This procedure arrange a second data set containing only those gene with a different distribution between the two samples groups (Short and Long). The main reason in doing so is to reduce the number of probes to reduce noise, which can decrease the accuracy later methods. That's why from now on, all the analysis and methods will be using this reduced data set.

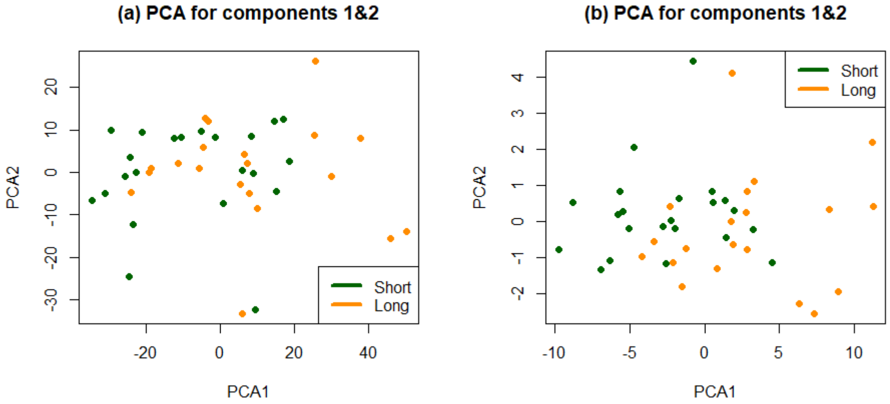

Figure 1: The plot *a* represent the PCA before the feature selection, where is really hard to distinguish a pattern of division. The plot *b* was performed after the feature selection, and even if there is no a clear division.

4.2 Clustering methods

In this section the project continue with the exploratory analysis going more in depth with some clustering methods, as K-means and Hierarchical clustering. From the two PCA and also from the conformation of the data (the lack of a control group), I don't expect a nice clustering. The reason behind is that the patients, even if have different phenotype, will likely have a considerable number of features in common.

K-means

The results of the K-means algorithm are shown in the figure 2, from where we can see the similarities with the second PCA. The cluster division is far from perfect, but is possible to observe a slightly division between the groups, even though there are mis classifications.

Hierarchical Clustering

An other clustering method applied was hierarchical clustering (figure 2), after few trials with different linkage (complete, centroid, single and average), the best one turn out to be complete linkage, which as for the K-means show a division but also a considerable number of mis classifications.

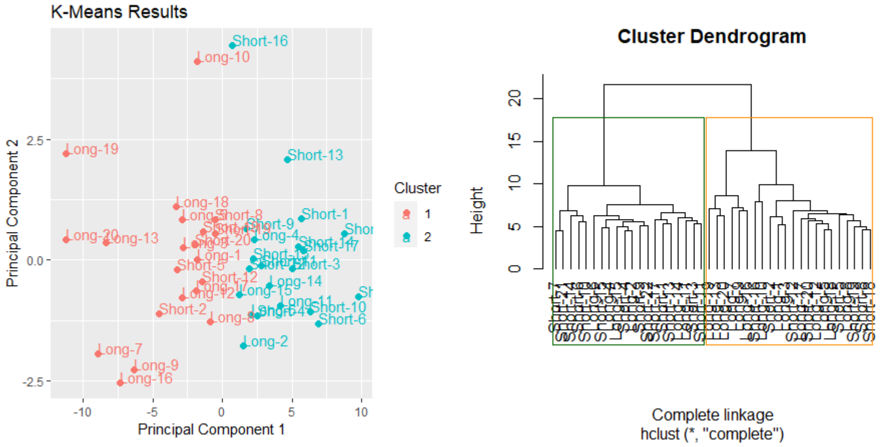

Figure 2: Are represented the result of the K-means tool in the left and the hierarchical clustering at the right. Both are not able to cluster very well the data.

4.3 Supervised methods

The project proceed with the evaluation of different predictions methods to extrapolate the more impactful genes of the reduced data set, the decision will be based on the accuracy of the prediction on the test set.

Random Forrest

The first supervised method tested was the Random Forrest; it was tried out with increasing number of trees to further increase accuracy, at the end the most perfoming was with the parameter ntree = 5000. The performance of the Random Forrest model is represented in figure 3.

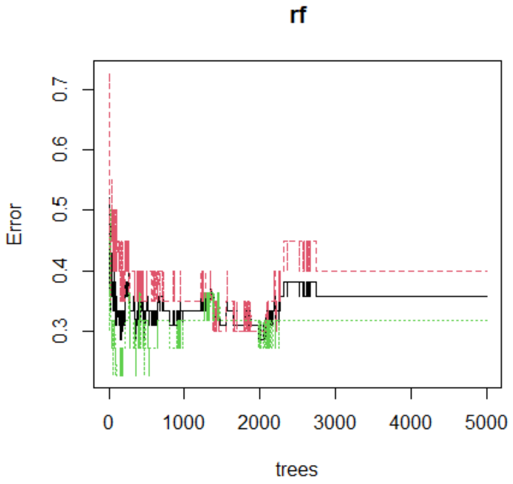

Figure 3: Random forest classification error rare for increasing number of trees. This approach reach a platou after 3000 trees, but the error is quite high.

Support Vector Machines

A Support Vector Machines (SVM) was also trained with the same spilt of the data of Random Forrest, for this model was trained two SVMs one with a linear kernel and a second with a polynomial kernel. Both perform quite well on the data, with an accuracy around 0.8, but out of the two was chosen the one with the polynomial kernel, which turn out to be the best model out of all.

Classification methods

Several prediction methods have been evaluated with the same split (of train and test set) as the Random Forrest and SVMs methods to maintain stability in the analysis. These methods are: LDA, Lasso, Ridge, Logistic regression and rScudo (which had an accuracy of circa 0.76, not shown in figure 4). All of these methods turn out to be not so good in predicting the test set, especially in comparison with the two SVMs, with an accuracy around 0.7. As regard rScudo the accuracy is better and is the one nearest the SVM with the polynomial kernel. The rScudo methods is been used with parameters nTop = 12, nBottom = 12 and N=0.25. Results of the rScudo network are shown in attachment section (Figure 9). The overall accuracies evaluations are shown in figure 4.

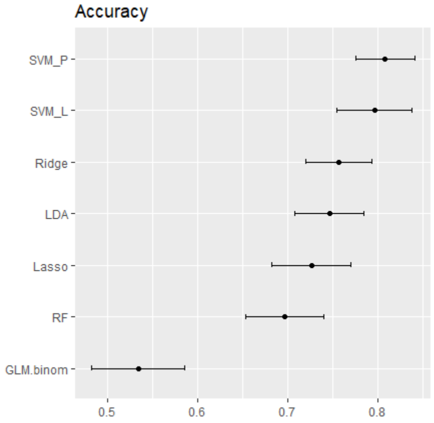

Figure 4: Plot of all the accuracies of the models tested.

4.4 Network Base Analysis

From the SVM with the polynomial kernel was extracted a list of 200 genes,the most important features according to the model and used for the analysis downstream.

Over Representation Analysis

Over Representation Analysis can be performed with tools as Gprofiler and David. Out of these two I used only GProfiler and not David, because David left out from the analysis too many genes. Given as input all the 200 features retrived from the model, the output of Gprofiler can be seen in figure 10 in attachment. All the point lie in the region of transcription factors, this mean that the majority of the gene are related with the switching of the transcription pattern of cells. Although all the point shown can't reach the p-value threshold (setted by the tools) is remarkable the redundancy of provenience from transcription factor database, and make also sense that a possible division between the groups is formed by transcription factors. Is also to mention that a small part of the features selected were not been found from the tools, some of them are Open Reading Frame (ORF) protein or related factors.

Network base analysis

At last was performed network analysis to search for connections between features and highlight important pathways to differentiate the two groups. This procedure was firstly done with the pathfindR library, which results are visible in figure 5 and sequentially with STRING. The mosts enriched pathways are (as shown in figure 5) the complement cascade, G alpha signaling events and the RHOA family. The complement cascade is related to an inflammatory environment and clear microbes and damaged cells. G alpha signaling evens take place during the transduction of signaling through the GTP-GDP and cyclic GMP switch. The RHOA family instead is composed by GTPases proteins, which are responsable mainly for the transduction of signaling and more in specific to the remodelling of the cythoskeleton, actin stress fibers formation and actomyosin contraction.

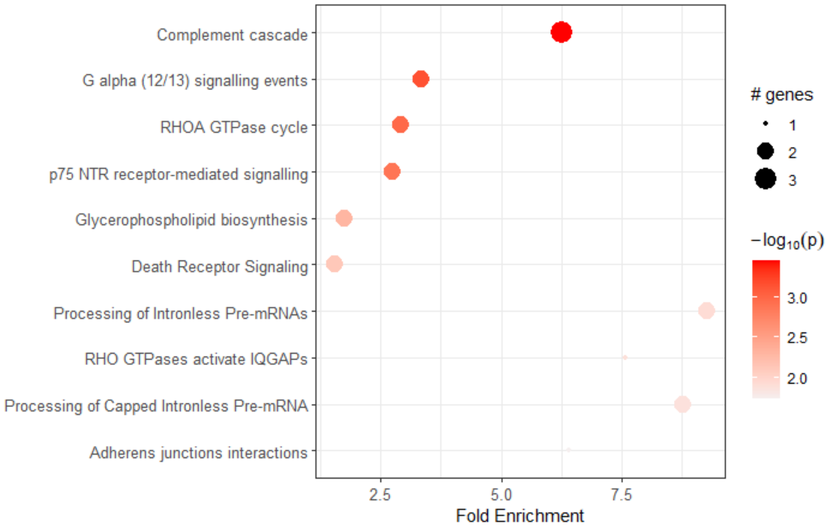

Figure 5: PathfindR plot showing the enriched terms with the respective p-value and number of genes involved.

As regard the STRING tool, all the 200 genes was gave as input at the tools, and as before some genes were not recognised. STRING is a database of protein protein interaction, so the output is a connected network; the links between proteins can be known interactions (data coming from literature) and predicted interactions. Out of the all genes given in input, was kept only those which created the biggest sub cluster, very few proteins connections were cut out. In the figure 6 is represent the resulting sub cluster, in which are highlight with different colors smaller groups of protein more deep link.

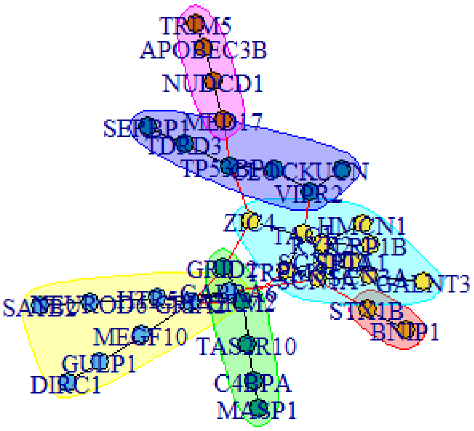

Figure 6: STRING output without the non interconnected proteins. The cluster are composed by: 1) TRIM5, APOBEC3B, NUDCD1, MED17; 2) STX1B, BNIP1; 3)DIRC1, GULP1, HTR5A, NEUROD6, SATB2, GABRA6, GRIA2, MEGF10; 4)C4BPA, GRM2, TAS2R10, GRID2, MASP1; 5)UCN, VIPR2, CLOCK, TP53BP1, DTRD3, SERBP1; 6) GALNT3, HMCN1, LRPIB, SCN3A, SPTA1, RYR2, SCN10A, TAC1, SCN1A, TRPM8, ZIC4

5 Discussion

This project has the aim to find differences in the gene expression between the Long disease phenotype and the Short phenotype. The 200 final genes were selected thanks to the SMVs, and even if cross validation was used to evaluated all the models, the number of samples (too small) could be a limit of this project. I want to underline that at the end, no control variables are kept in the final list of features, which is evidence of the effectiveness of the analysis. The ORA analysis tells that there are a discreet number of genes belongs to the transcription factor family, which is reasonable for the splitting of the two phenotypes. Meanwhile the Network base analysis with pathfindR, shows as enriched terms related to inflammation (Complement cascade), the RHOA family (associated with cytoskeleton regulation) and transduction of signaling. Lastly STRING return a network with all the 200 genes, which was cleaned removing all the unconnected genes, remaining with only a connected sub cluster. These genes was again divided in more connected and smaller clusters. These cluster involve genes related to pro-inflammatory environment, immune response activation, cell division and growth, neuro trasmitters, complement activation, chromatin remodelling and apoptotic reaction. All of these are tightly bind to the mechanism and the compartments of the ALS. These gene are, according to this analysis, candidate to differentiate the two phenotypes, of course the two group have lots of gene (that could be linked to ALS) in common in term of expression because they are different phenotypes of the same disease. Anyway is also reasonable that these different expressed genes are redundant with the ones strongly related to the disease in general; this implies that the different phenotypes depend on the same causes (pathways), but the more dangerous one is due to a grater activation of the same pathways or a wider (in number of genes) activation. An other thing honorable of mention is that all this tools cannot use some of the genes, which are cut out of the analysis; results like C2orf73, KIAA1949 and few others. Probably some of them are relevant (Open reading frames are impactful in this disease [CS23]), that's why is needed a deeper analysis. Since these results come from gene expression my suggestion is to make further analysis with more samples and with data coming from miRNA; which in my opinion are relevant due to the disease mechanism. To sum up, based on these finding the main differences between the Long and Short disease duration lies on the more inflammatory environment, the more outrageous immune response, that probably influence the differentiation and secretory pattern of motor neurons and associated cells.

References

[Car23] M.; Di Giulio S.; Mariano S.; Panzarini E. Carata, E.; Muci. Looking to the future of the role of macrophages and extracellular vesicles in neuroinflammation in als. *Int. J. Mol. Sci.*, 24:11251, 2023.

[CS23] Menkes Daniel L. Chong, Zhao Zhong and Nizar. Souayah. Pathogenesis underlying hexanucleotide repeat expansions in c9orf72 gene in amyotrophic lateral sclerosis. *Neurosciences*, 2023.

[MP20] Van Damme P. Masrori P. Amyotrophic lateral sclerosis: a clinical review. *Eur J Neurol.*, 2020.

[Sin20] Devasahayam G. Singh, E. Neurodegeneration by oxidative stress: a review on prospective use of small molecules for neuroprotection. *Mol Biol Rep*, 47:3133–3140, 2020.

Attachment

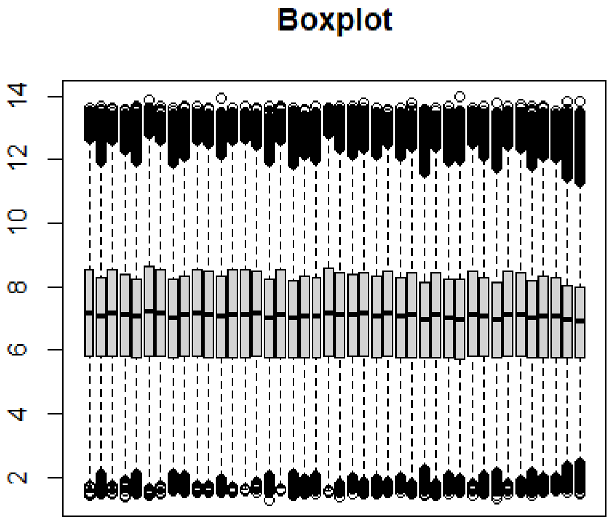

Figure 7: Box plot of initial values per sample (x-axis), without further processing. Since median and variance looked well aligned, no further process of normalization have been done.

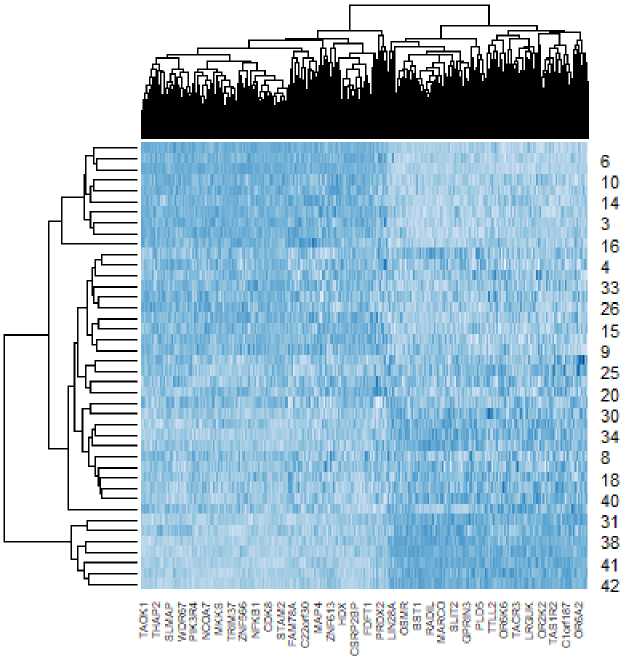

Figure 8: Heat map after the feature selection

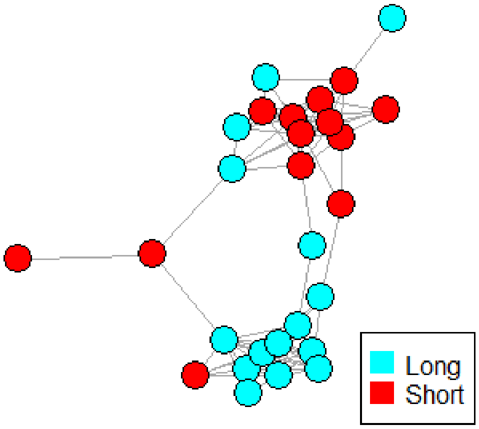

Figure 9: RScudo training performance

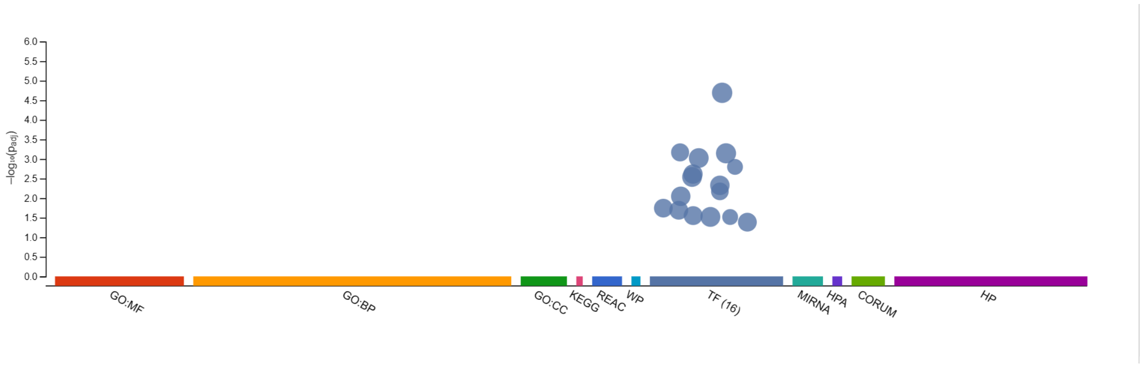

Figure 10: Gprofiler found reference only in Transcription Factors data bases

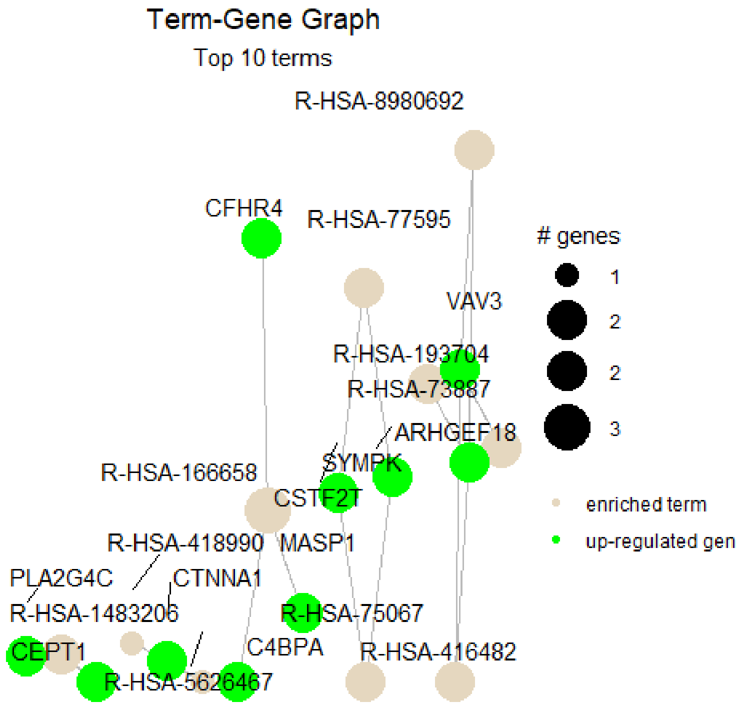

Figure 11: pathfindR network of interactions. This network have been achieved giving as input the list of the 200 most important genes.