Report

Analysis of an Esophageal cancer dataset

Gloria Lugoboni

13 June 2024

Contents

1 Introduction

1.1 The biological question

Esophageal cancer (ESCA) is one of the most common gastrointestinal tumors and is placed among the 10 highest mortality cancers worldwide [1], with a 5-year overall survival (OS) of circa 20% and high probability of developing metastasis [2] and resistance to chemo-radiotherapy [1]. ESCA can be differentiated into two main subtypes, esophageal squamous cell carcinoma (ESCC), which is the most prevalent form in China [3], and adenocarcinoma (AC), which is the most prevalent form in the United States [4].

1.2 The aim of the project

The aim of this study is to expand our knowledge of esophageal cancer, defining a set of variables that can help us distinguish cancer samples from healthy ones. Both supervised and unsupervised learning methods have been exploited here and the latter have been used to select a list of the best probes (the most important ones) to be given as input to 5 tools for functional enrichment analysis (Over-Representation Analysis and Network-based Analysis).

2 Methods

2.1 GSE199967

GSE199967 was retrieved from the Gene Expression Omnibus (GEO) database. The dataset, generated from Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version), contains a total of 42 specimens of tumor tissues and normal mucosal tissues that were collected from the Henan University of Chinese Medicine.

2.2 Principal Component Analysis - Unsupervised learning method

Principal component analysis (PCA) is a technique used to reduce the dimensionality of large datasets thanks to the identification of so-defined "principal components" [5]. This analysis was performed in R using ***prcomp*** from the ***stats*** package, giving as input the transposed expression data.

2.3 Clustering - Unsupervised learning methods

The term clustering is used to indicate a wide group of techniques that aim at identifying subgroups in a dataset. The idea behind this type of learning is to group similar objects, for this reason, it is important to define a criterion for similarity [6].

K-means clustering can be described as a technique based on an iterative relocation of data points into groups defining, therefore, clusters [7]. In the algorithm's initialization process, the number of the initial centroids k must be selected. This number can be derived based on previous empirical knowledge [8]. This analysis was performed in R using ***kmeans*** from the ***stats*** package. In this work $k = 2$ since we are dealing with a dataset containing samples of tumor tissues and normal mucosal tissues.

Hierarchical Clustering Hierarchical clustering is one of the possible techniques in which the resulting clustering is represented as an observation tree called dendrogram [9]. The analysis was performed using ***hclust*** from the ***stats*** package. In this work, the distance between observations was computed using the Euclidean distance (***method = "euclidean"***) while the cluster similarity was computed using both the average and complete linkage measure (***method = "ave"***, ***method = "complete"***). As, for k-means, the number of final clusters needs to be selected, and, for the same reason above, here we have $k = 2$. As a trial, $k = 6$ was selected, to understand if the method was able to extract better results with a higher number of clusters, accounting for the possibility of tumor heterogeneity in the tumor samples.

2.4 LIMMA - Feature Selection

The idea behind feature selection is to highlight a subset of relevant features obtained from a larger dataset. This process is essential to improve the accuracy and efficiency of several statistical methods and machine learning models [10]. Here, we use the R/Bioconductor software package ***limma***, which enables the analysis of gene expression data, both for microarray and RNA-sequencing data. Limma enables differential expression (DE) analysis by integrating several statistical principles and Bayes methods in a set of linear models to analyze entire gene expression experiments. The basic statistic test used to investigate significance is the moderated t-statistic, which is computed

for each probe and each contrast [11]. In this work, the design matrix was created defining separate coefficients for control and tumor samples. The ***lmFit()*** function from the ***limma*** package was used to test each gene for differential expression between the two groups (controls and tumors) using a linear model. After fitting our data to the linear model, we apply empirical Bayes smoothing with the ***eBayes()*** function. Because we are testing many different genes at once, we also performed multiple test corrections using the Benjamini-Hochberg method. The top genes were extracted using a log-fold change of 1.5.

2.5 Supervised Learning methods

With the term supervised learning, we indicate a set of methods able to define trained models starting from data associated with labels that provide information about a preexisting classification [12]. The input data for all the supervising methods were filtered by applying a row t-test (***rowttest*** function, part of the ***genefilter*** package). Specifically, the genes that resulted significant were maintained. The threshold for the p-value was set at 0.05. This procedure resulted in a dataset with 1988 variables (probe IDs), as opposed to the 41000 initial ones.

Each method has been tested through a 10-fold cross-validation using the ***train*** function from package ***caret***.

The models were compared in their performance based on their accuracy. For each model, the most important probes, also defined as features, were extracted, to compare them with the selected features obtained from LIMMA, this was possible using the function ***varImp*** of the ***caret*** package.

Random Forest is a supervised learning method in which multiple decision trees are combined to reach a single, most concordant, result. During this procedure, the input features are resampled in a "bootstrap" procedure creating from the extracted subset a decision tree. The predictions of the multiple trees are then merged calculating a mean value. From the final tree, it is possible to extract the most important feature (here genes) that guide the prediction/clusterization [13].
To train the model, the best parameters were chosen using the results obtained from the ***tuneRF*** function, part of the ***iRF*** package. The ***mtry*** linked to the lowest OOB error was equal to 11. The number of trees to be created by Random Forest was selected after several trials, and the final value for the parameter ***ntree*** is 500.

LDA is an approach used in supervised machine learning that, thanks to a process of dimensionality reduction, separates multiple classes resolving the multi-class classification problems. This is possible thanks to the creation of a linear decision surface that is used to correctly assign each observation to its class [14].
To train the model, it was necessary to select a prior probability of the classes present in the study. In the initial data, the classes were equally presented, therefore, the input for the ***train*** function was ***prior = c(0.5,0.5)***.

LASSO and Ridge are regression analysis methods that perform a process of parameter regularization by shrinking the regression coefficients and reducing some of them to zero. The non-zero coefficients are selected to be used in the model. The difference between Lasso and Ridge stands in the constraint applied [15].
To train the model, the best parameters were selected using a trial process via the ***TuneGrid*** parameter in the ***train*** function. After this process, the ***lambda*** value, which is needed to avoid overfitting and define the shrinking of the coefficients toward zero, was selected as equal to 0.02778940 for both models.

SCUDO (Signature-based Clustering for Diagnostic Purposes) is a rank-based method used for diagnostic and classification purposes, possible by the identification of sample-specific gene signatures that are used to build a graph and partition the data into homogeneous clusters on the basis of signature similarity [16].
The best parameters to train the model were selected using a trial process via the ***scudoModel*** function, part of the ***SCUDO*** package. The final values used for the model were ***nTop = 100, nBottom = 100, N = 0.25, maxDist = 1*** and ***beta = 1***.

2.6 Functional Enrichment Analysis

Functional enrichment analysis, also known as gene set analysis (GSA), is a well-known method used to analyze high-throughput data with the final aim of highlighting biological annotations that are over-represented in a list of genes with respect to a reference background[17].

Five different tools have been used to perform two different analyses:

- Over-Representation Analysis. This analysis is used to determine which a priori-defined gene sets are more present (over-represented) in a subset of "interesting" genes. In this study, ***gProfiler*** (in R) and ***DAVID*** (online) were used.

- Network-based Analysis. These methods are designed to investigate networks representing the interactions between the elements of a complex system with the final aim of extracting information on the connectivity and organization of the system. In this study, ***pathfindR*** (in R), ***enrichNet*** (online) and ***STRING*** (online) were used.

These analyses were performed on both the selected features obtained from LIMMA and the most important features extracted for each supervised model defined above. The results were then compared to understand if similar terms were identified by the two types of "future selection" methods: LIMMA vs supervised models.

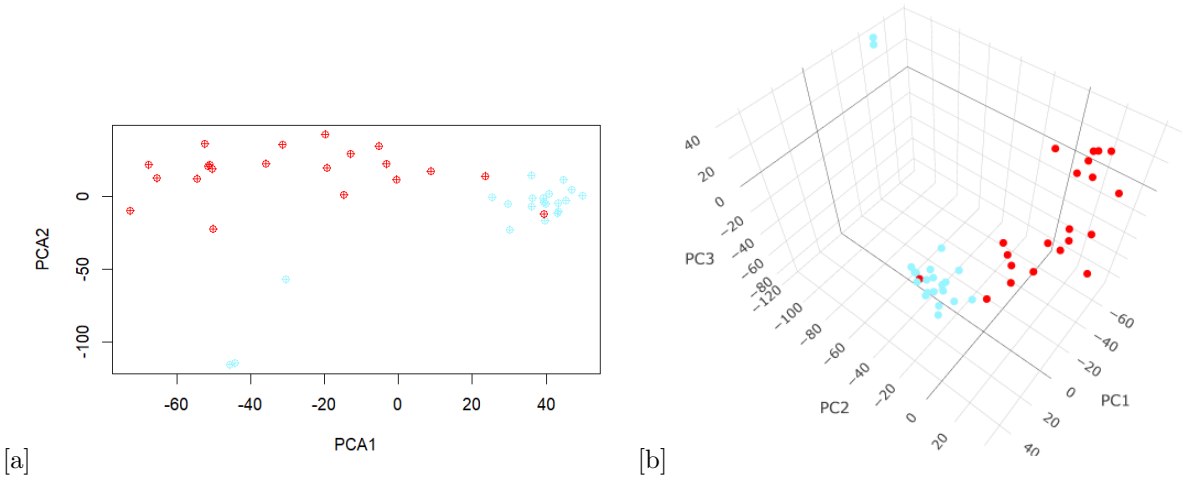
3 Results

3.1 Data visualization and Principal Component Analysis

No normalization was performed on the GSE199967 data since the boxplot obtained from the expression data showed a uniform distribution of the original data (Appendix fig. **??**).

A PCA analysis was effectuated on the complete gene expression datasets. We used the information retrieved from the metadata (tumor vs control) to investigate the ability of the PCA to stratify the data.

The results are presented in Figure 1, we can see that in both situations the identified two or three principal components can cluster and separate almost perfectly the Tumor samples from the Control ones and vice versa.



[a] [b]

Figure 1: **Principal Component Analysis**. (a) PCA plot in two dimensions, x-axes: PC1, y-axes: PC2, (b) PCA plot in three dimensions, x-axes: PC1, y-axes: PC2, z-axis: PC3. Controls are shown in light blue while tumors are shown in red.

3.2 Clustering - Unsupervised learning methods

K-means and hierarchical clustering were performed on the complete gene expression datasets with the objective of observing the ability of such unsupervised learning methods to identify the two main clusters present in the given dataset: control and tumor samples.

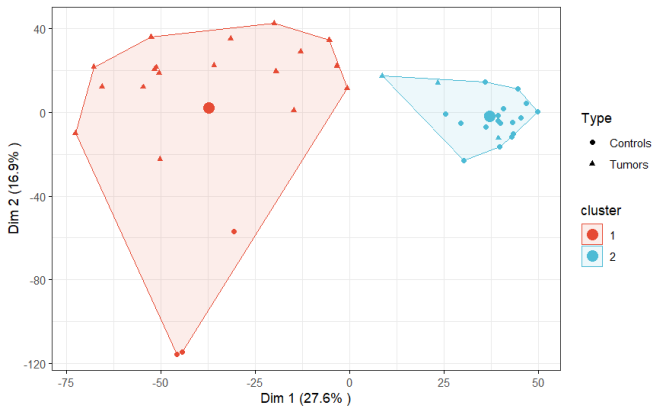
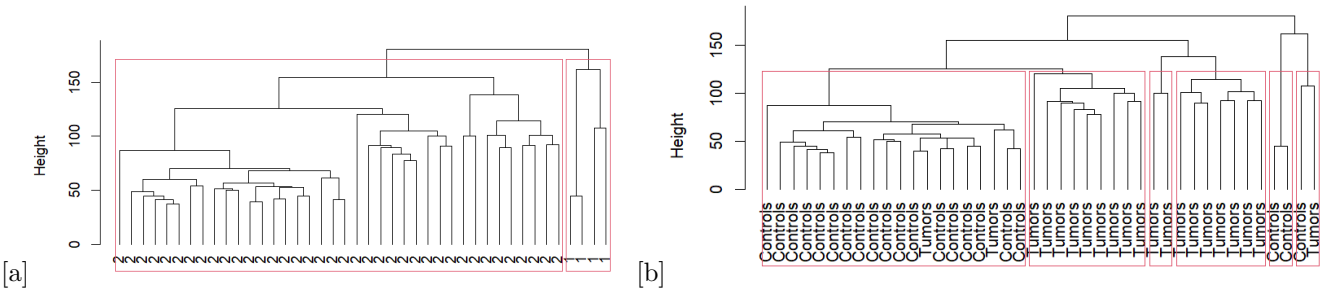


Figure 2: **Cluster analysis**. K-mean clustering, axes: Coordinates for the variables extracted from the first and the second PC

Starting with k-means, two clusters were identified, each containing 21 samples. From figure 2 it is possible to observe that the two clusters, tumors in red and controls in light blue, are well separated in space, as we observed in the PCA plots. Despite this, it is important to notice that the clusters are not able to separate perfectly the tumor and the control samples. Indeed, we can observe that three control samples were misclassified within the tumor one and a total of 3 tumor ones were misclassified as control samples. Nevertheless, these results are quite encouraging because K-means has been able to identify some patterns in the data.



[a] [b]

Figure 3: **Cluster analysis**. (a) Hierarchical clustering, x-axes: Identified clusters, y-axes: Height. k=2, (b) Hierarchical clustering, x-axes: Sample type (Tumor or Control), y-axes: Height. k=6.

Regarding the analysis operated with hierarchical clustering, here, only the results obtained by setting the linkage measure to complete are shown. From figure 3 [a] it is possible to observe that two clusters were identified, the first containing 39 samples and the second one containing 3 samples. It is clear that, with *k=2*, it is not possible to cluster the samples into the correct groups. By working on the k value it was possible to obtain a better clustering, as can be seen by figure 3 [b] where ***k=6***. It is possible to notice that almost all controls are clustered together, indeed only one control is misclassified with a tumor sample. The same can be said for the tumor samples, which are divided into 4 different clusters. One possible reason for this behavior could be connected to tumor heterogeneity, indeed, it is known that esophageal cancer is a complex malignancy characterized by a high level of intra-tumor heterogeneity, as cells in distinct microenvironmental niches acquire diverse phenotypic features [18]. This could reflect in the clustering process and generate 4 different groups of tumor subtypes characterized by different genetic expressions. Overall we can say that, thanks to hierarchical clustering, it is possible to obtain a good separation between tumor and control samples.

3.3 Feature selection - LIMMA

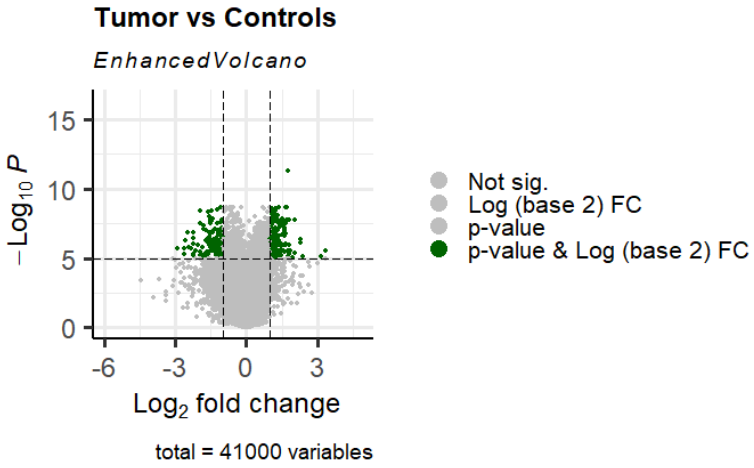


Figure 4: **Feature selection with LIMMA**. x-axis: Log2 Fold Change, y-axis: -Log10 p-value. The significative genes, highlighted in green, were selected based on thresholds on the p-value (0.05) and the log fold change (1.5).

The feature selection was applied to a total of 41000 genes. As it can be observed from figure 4, after the analysis, a total of 265 genes resulted as de-regulated in the tumor samples with respect to the control ones. Of these, 79 were found up-regulated and 186 down-regulated.

The differentially expressed genes were used later in this study to operate a functional enrichment analysis and understand if the highlighted terms were connected to esophageal cancer.

3.4 Supervised learning methods

As explained in the methods, a total of five models have been evaluated: Random Forest, Linear Discriminant Analysis, Ridge Regression, Lasso Regression and SCUDO (training and test networks in appendix fig. 11). Overall, all methods performed well with high accuracy, as can be seen from figure 5.

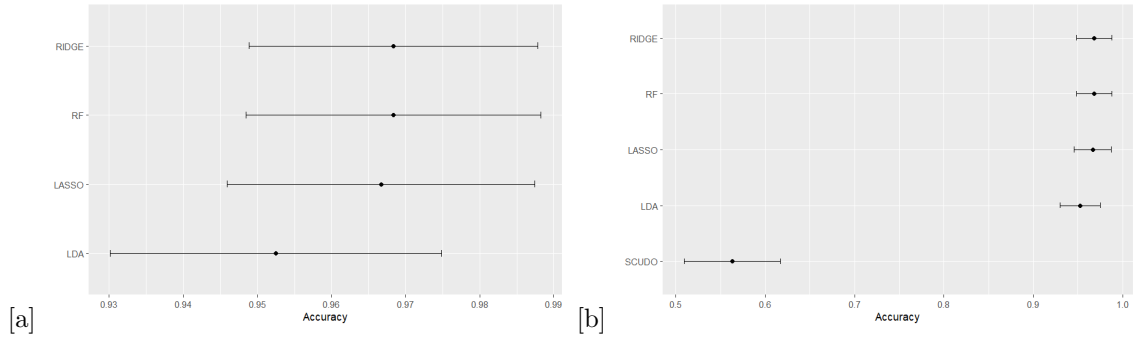


Figure 5: **Performance plot of the supervised models**.(a) All models except SCUDO. x-axis: accuracy, y-axis: model name. (b) All models. x-axis: accuracy, y-axis: model name. Some whiskers exceed 1.0. That might be explained by the low number of samples included in the analysis. SCUDO shows a high variance.

Among all, random forest showed the highest accuracy, equal to 0.9683333, followed directly by Ridge with an accuracy of 0.9658333. Figure 6 represents the OOB error rate against the number of trees used to train the model. The OOB error rate can be tough as the number of wrongly classified observations, therefore, the lower the OOB score, the most accurate our model is. Indeed, we can see that as the number of trees increases, the error rapidly decreases becoming stationary at a level < 0.1.

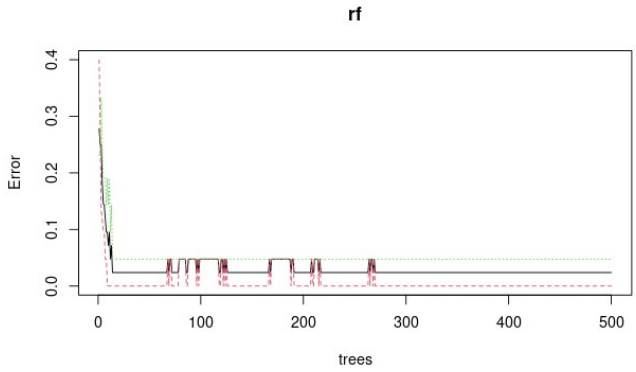


Figure 6: **Distribution of the out-of-bag (OOB) score over the total number of trees used to train the RF model.**

For each model, the most important variables were extracted (Appendix figure 12) using an arbitrary threshold set by looking at the distribution of the variable importance.

To take as an example Random Forest, the threshold was set at a value of importance equal to 10, since it is possible to observe from Fig. 7 that the majority of the variables sit between an importance value of 0 to 10. It was possible to observe that, for certain models, there is an agreement in the selection of the most important variable for the classification of the samples (Appendix Table 5). The same comparison was made with the genes extracted from LIMMA (Appendix Table 5).

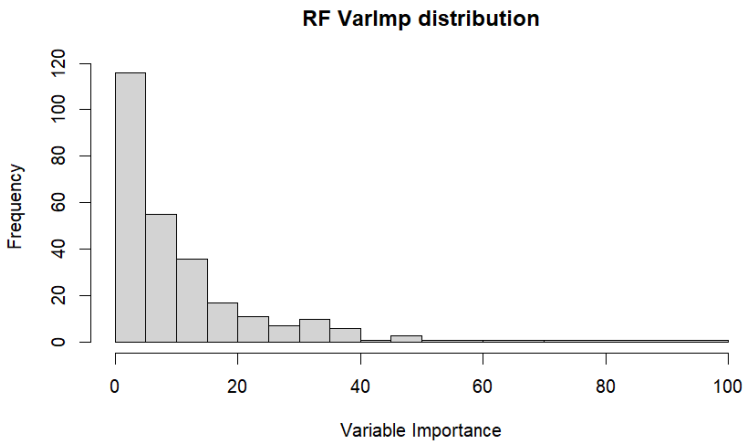


Figure 7: **Variable importance distribution in Random Forest.**

It is possible to observe that all the most important genes for random forest are also contained in the list of differentially expressed ones extracted from LIMMA.
It was possible to extract the following list of genes from the several models:

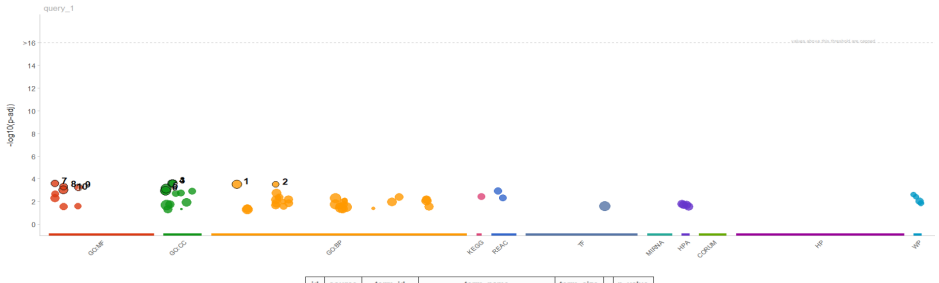
- 94 genes extracted as the most important features from the Random Forest model.
- 510 genes extracted as the most important features from the LDA model.
- 65 genes extracted as the most important features from the Ridge model.
- 293 genes extracted as the most important features from the LASSO model.
- 746 genes extracted as the most important features from the SCUDO model.

3.5 Functional Enrichment Analyses

The functional enrichment analyses have been performed on the set of genes extracted from LIMMA and extracted from the models as explained above. The lists were given to **gProfiler** and the most significative terms were considered.

An analysis of the terms highlighted that:

- The terms obtained from the random forest-selected genes mainly dealt with the extracellular matrix, serine-type peptidase activity and hydrolase activity (fig. 8). It has been shown in several studies that an enhanced activity of serine-type proteases and matrix metalloproteinases (MMPs) plays an important role in esophageal carcinogenesis [19].



id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0007155	cell adhesion	1512	3.1e-04
2	GO:BP	GO:0022617	extracellular matrix disassembly	61	3.2e-04
3	GO:CC	GO:0031012	extracellular matrix	555	2.5e-04
4	GO:CC	GO:0030312	external encapsulating structure	556	2.5e-04
5	GO:CC	GO:0005615	extracellular space	3303	7.3e-04
6	GO:CC	GO:0005576	extracellular region	4213	1.1e-03
7	GO:MF	GO:0004252	serine-type endopeptidase activity	180	2.7e-04
8	GO:MF	GO:0008236	serine-type peptidase activity	198	5.1e-04
9	GO:MF	GO:0017171	serine hydrolase activity	202	5.8e-04
10	GO:MF	GO:0008092	cytoskeletal protein binding	1002	9.0e-04

Figure 8: **gProfiler for RF most important variables**

- Regarding the genes selected as most important for LDA, the highlighter terms are mainly dealing with cell cycle and cell division (Appendix fig. 19). Indeed, it is known that aberrant cell proliferation is a hallmark of cancer [20].
- The genes extracted from the Ridge model had terms dealing with DNA replication (like Ctf18 RFC-like which is a subunit of a Replication Factor C (RFC)-like complex) and cell cycle (like NEDD8) (Appendix fig. 21). The same consideration done in the point above can be done for these terms.
- The terms obtained from the LASSO-selected genes mainly dealt with DNA replication, cell cycle, and cell organization (Appendix fig. 16). Indeed, these are the same terms observed above, that could explain to us that the models are using genes that are found aberrant in cancer with respect to controls to operate the classification.
- Finally, for the supervised methods, the terms highlighted starting from the genes obtained from SCUDO are, as expected from the observed trend, dealing with cell cycle and replication (Appendix fig. 22).
- Regarding the terms highlighted from the LIMMA selected genes, the majority dealt with the extracellular matrix organization and endonuclease activity (Appendix fig. 23). From the literature, it is possible to understand that the activity of such proteins is highly correlated with cancer cell invasion, migration, and metastasis [21].

These same lists were also given to **DAVID** which identified several terms with an adjusted p-value ranging between $E-0$ and $E-14$ (Appendix figures 13-18). The obtained terms showed high similarity with the one identified with gProfiler. Specifically, it is interesting to point out the term *palmoplantar keratoderma*, found from the list of genes retrieved with LIMMA. *Palmoplantar keratoderma* is a rare genetic disease that is associated with a very high lifetime risk of developing squamous cell carcinoma of the esophagus (OSCC), a subtype of esophageal cancer [22].

Both David and gProfiler found similar results, considering the terms with the highest p-values.

The same lists of genes have been used for STRING, enrichNet, and pathFinder (NbDA). Regarding **enrichNet**, significant results have been obtained for the gene lists derived from LDA, SCUDO, and LASSO. For LDA, statistically significant pathways identified with this analysis were DNA repair, with an XD score of 3.1157 and a p-value of 1.1e-08, and cell cycle, with an XD score of 2.1611 and a p-value of 5.3e-03. These results are similar to the one identified with the over-representation analysis, indeed, we can think that the identified genes have a strong biological significance. Regarding LASSO, the pathway of DNA replication was the only one identified as statistically significant, with an XD-score of 1.37490 and a p-value equal to 0.0015. Indeed, this is the same term that was highlighted in the previous analyses. Finally, regarding SCUDO, the identified pathways were cell cycle, with an XD-score of 1.5469 and a p-value of 9.9e-13, and DNA repair, with an XD-score of 3.1151 and a p-value of 1.2e-08. We can undoubtedly state that there is strong evidence of similarity between the current results and the precedent findings.

Regarding **STRING**, the only list of genes that were not characterized with more interactions than expected was the one obtained from Ridge. Table 3.5 reports the statistic and the functional enrichment of the obtained networks. It is possible to observe that there is an high similarity between these results and the one obtained from the previous analysis. For each network k-clustering was performed, and almost all the networks were characterized by the presence of one major cluster containing a vast portion of the nodes (fig. 9; Appendix fig. 24-27).

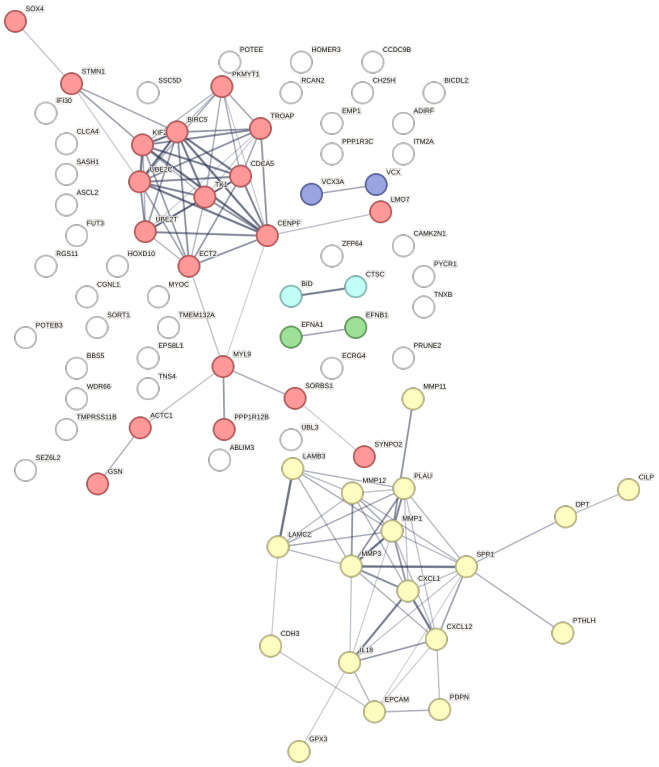


Figure 9: **Clusters obtained from RF STRING-network**. In red, cluster 1, with a total number of genes equal to 19 (genes regulating replication). In yellow, is cluster 2, with a total number of genes equal to 18 (genes regulating extracellular matrix degradation). Overall number of nodes: 94

STRING networks					
	Random Forest	LDA	Lasso	SCUDO	LIMMA
Expected number of edges	26	1354	192	2179	116
Number of interaction	103	4948	271	6635	449
PPI enrichment p-value	<1.0e-16	<1.0e-16	4.14e-08	<1.0e-16	<1.0e-16
Pathway	Degradation of the extracellular matrix	DNA replication	DNA replication	DNA replication	Degradation of the extracellular matrix
Biological Process	Serine-type endopeptidase activity; Regulation of replication	DNA replication, removal of RNA primer	Regulation of DNA replication	DNA replication preinitiation complex assembly	Extracellular matrix assembly; Positive regulation of exit from mitosis

pathFindR has been used on the most important features extracted from random forest selecting both KEGG and Reactome as the genesets. In the first case, terms mostly related to apoptosis, inflammation, and cell cycle are shown (Appendix fig.28). Some of these pathways were also identified previously, endorsing the results obtained in the functional enrichment analysis. Regarding the terms associated with apoptosis, it was possible to understand from the literature that during the early stages of esophageal carcinogenesis, the process of apoptosis starts and becomes more and more persistent. Indeed, hyperproliferative cells are more susceptible to apoptosis [23]. In the second case (fig. 10), the main terms concern extracellular matrix organization and metalloproteinase activity. These results highlight the concordance in the findings regarding the genes selected starting from the random forest model. Other terms highlighted deal with EPH-Ephrin signaling which correlates with an aberrant differentiation in esophageal cells when found hyperactivated [24].

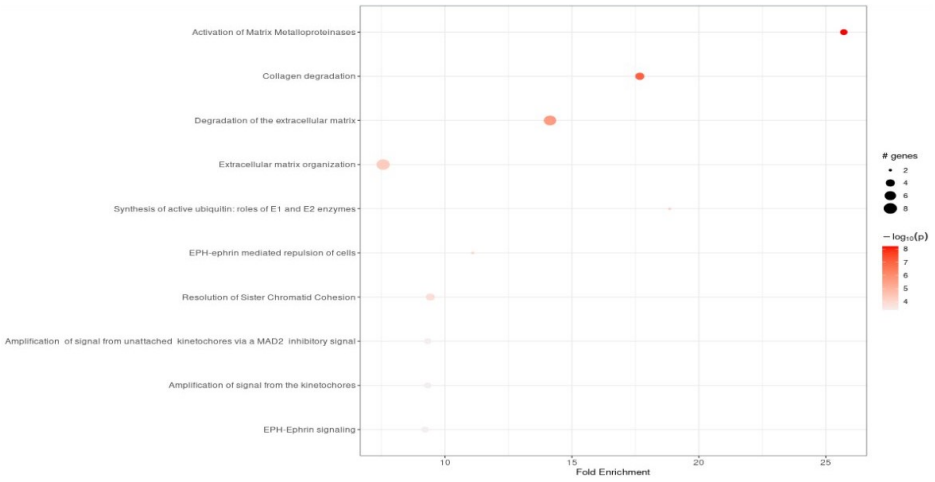


Figure 10: **PathFindR enrichment chart on important variables extracted from the Random Forest model.** Result obtained selecting Reactome as geneset.

4 Discussion

This analysis aims to find the set of variables that better distinguish esophageal cancer patients respect control ones. The original dataset contained 42 samples of which 21 were tumoral and 21 controls. Among the supervised learning methods, the best-performing one was random forest, followed by Ridge regression. A list of the most important variables for the several models was extracted and was used to make a comparison with the list of differentially expressed genes extracted from LIMMA. The idea behind this process is to understand if machine learning methods, like the one used in this analysis, define informative genes the ones that have functions connected to the de-regulated ones observed in the tumor samples and extracted from LIMMA. Indeed, it was possible to observe that for certain models, the most important variables represent genes that are differentially expressed (Appendix Table 5). Overall, the functional enrichment analysis highlighted terms and pathways related to matrix degradation and DNA replication. Records concerning matrix organization, metalloproteinase activity, and regulation of DNA replication and mitosis have been identified by David, gProfiler, enrichNet, STRING, and Pathfinder. Indeed, all the tools used for the functional analysis showed very concordant results.

These results are quite to be expected since it has been shown in several studies that an enhanced activity of matrix metalloproteinases (MMPs) plays an important role in esophageal carcinogenesis [19]. Plus, aberrant cell proliferation is a well-known hallmark of cancer [20].

Finally, thanks to this study, it was possible to observe a similarity in the functions associated with the list genes extracted using LIMMA, a well-known feature selection method used to analyze microarray data, and supervised learning methods, like random forest and LDA. We can therefore hypothesize that the extracted genes have a strong connection with esophageal cancer and could be used for further analysis to understand if the sets identified are specific for ESCA or share genes found in other cancers.

References

1. Et al., P. Z. Comprehensive analysis of a new immune-related prognostic signature for esophageal cancer and its correlation with infiltrating immune cells and target genes. *Ann Transl Med.* (2021).
2. Pennathur A, e. a. Oesophageal carcinoma. *The Lancet.* (2013).
3. Greathouse, K. e. a. Co-enrichment of cancer-associated bacterial taxa is correlated with immune cell infiltrates in esophageal tumor tissue. *Sci Rep 14, 2574* (2024).
4. Patel N, B. B. Incidence of Esophageal Cancer in the United States from 2001-2015: A United States Cancer Statistics Analysis of 50 States. *Cureus.* (2018).
5. T., J. I. & Jorge, C. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc* (2016).
6. Et. al., M. R. M. T. 1 - Analytics Defined, Information Security Analytics. *Syngress* (2015).
7. Morissette, L. & Chartier, S. The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology* (2013).
8. Et al, A. M. I. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences* (2023).
9. Hartigan, J. Statistical Clustering. *International Encyclopedia of the Social and Behavioral Sciences, Pergamon* (2001).
10. Et al, O. A. M.-L. Do feature selection methods for selecting environmental covariables enhance genomic prediction accuracy? *Front. Genet.* (2023).
11. Et al., M. E. R. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015).
12. A. Singh, N. T. & Sharma, A. A review of supervised machine learning algorithms. *3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (2016).
13. Mariana Belgiu, L. D. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* (2016).
14. E. I. G. Nassara, E. G.-M. & Kharouf, M. Linear Discriminant Analysis for Large-Scale Data: Application on Text and Image Data. *15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (2016).
15. L.E. Melkumova, S. S. Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering* (2017).
16. Lauria M Moyseos P, P. C. SCUDO: a tool for signature-based clustering of expression profiles. *Nucleic Acids Research.* (2015).
17. Garcia-Moreno A, L.-D. R. e. a. Functional Enrichment Analysis of Regulatory Elements. *Biomedicines.* (2022).
18. Junyu Li Lin Li, P. Y. e. a. Towards artificial intelligence to multi-omics characterization of tumor heterogeneity in esophageal cancer. *Seminar in Cancer Biology* (2023).
19. Et all., G. M. The role of matrix metalloproteinases (MMPs) and their inhibitors (TIMPs) in the development of esophageal cancer. *Folia Histochem Cytobiol.* (2012).
20. Gutschner T, D. S. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol.* (2012).
21. Qi Q Obianyo O, e. a. Blockade of Asparagine Endopeptidase Inhibits Cancer Metastasis. *J Med Chem.* (2017).
22. Anthony Ellis Janet M. Risk, e. a. Tylosis with oesophageal cancer: Diagnosis, management and molecular mechanisms. *Orphanet Journal of Rare Diseases volume* (2015).
23. Wang LD, e. a. Apoptosis and cell proliferation in esophageal precancerous and cancerous lesions: study of a high-risk population in northern China. *Anticancer Res.* (1999).
24. Et al., V. S. The Ephrin B2 Receptor Tyrosine Kinase Is a Regulator of Proto-oncogene MYC and Molecular Programs Central to Barrett's Neoplasia. *Gastroenterology* (2022).

5 Appendix

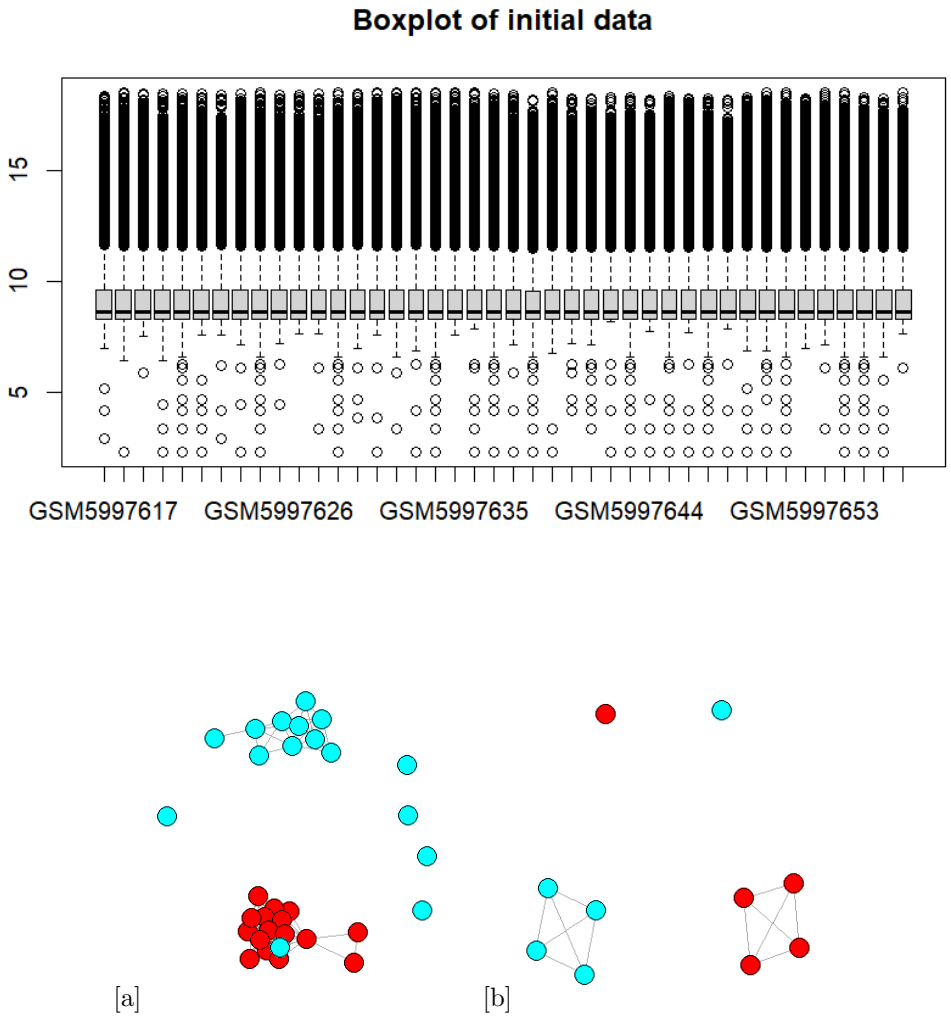


Figure 11: **SCUDO**. Training and test networks that were obtained for the SCUDO model.

Most important features shared between the models					
	Random Forest	LDA	Ridge	Lasso	SCUDO
Random Forest	94	32	10	2	42
LDA	32	510	27	121	489
Ridge	2	27	65	65	38
Lasso	10	121	65	293	174
SCUDO	42	489	38	174	746

LIMMA vs Supervised methods					
	Random Forest	LDA	Lasso	RIDGE	SCUDO
LIMMA	94/94	33/477	11/282	2/63	47/699

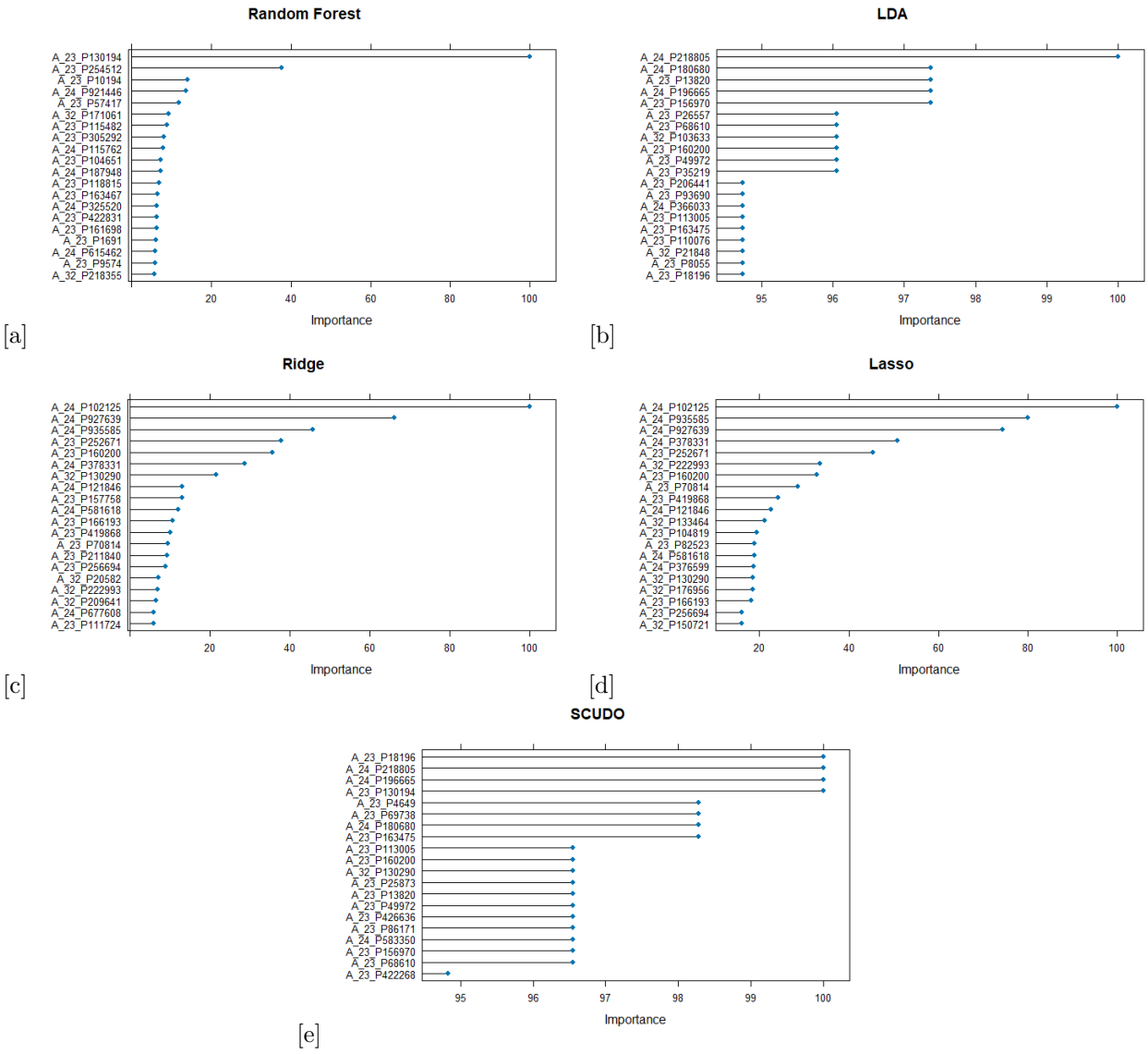
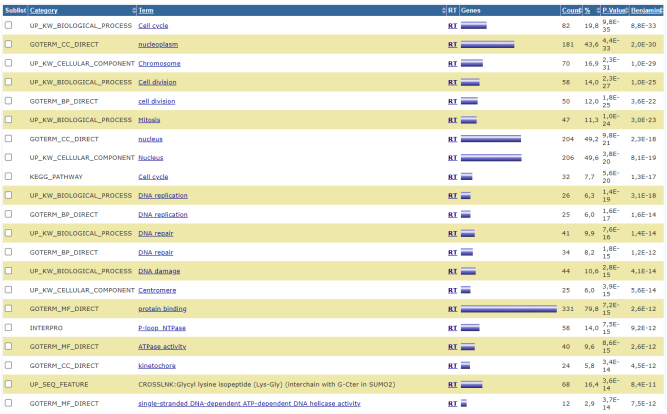


Figure 12: **Top 20 most important variables for the supervised models**. x-axis: Importance, y-axis: Probe ID

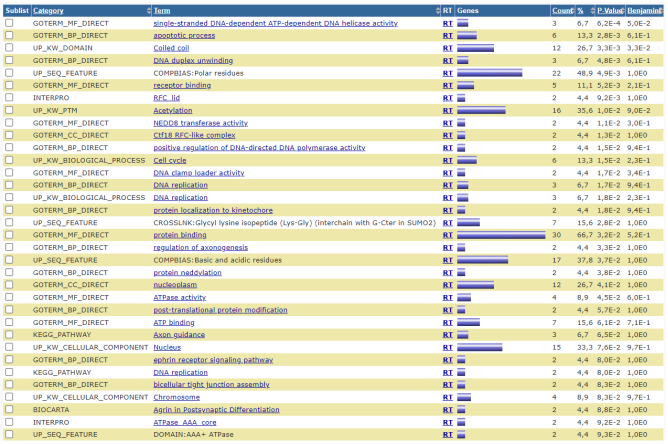
SubList	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	UP_KW_CELLULAR_COMPONENT	Extracellular matrix	RT		10	11,6	2,1E-6	5,7E-5
	INTERPRO	Pept_M10A_Zn_BS	RT		4	4,7	2,6E-5	3,9E-3
	INTERPRO	Hemopexin_CS	RT		4	4,7	3,8E-5	3,9E-3
	UP_KW_CELLULAR_COMPONENT	Secreted	RT		23	26,7	4,3E-5	5,8E-4
	GOTERM_CC_DIRECT	extracellular space	RT		21	24,4	6,6E-5	1,1E-2
	UP_SEQ_FEATURE	REPEAT:Hemopexin 3	RT		4	4,7	8,9E-5	1,4E-2
	UP_SEQ_FEATURE	REPEAT:Hemopexin 4	RT		4	4,7	8,9E-5	1,4E-2
	INTERPRO	M10A_MMP	RT		4	4,7	9,8E-5	3,9E-3
	INTERPRO	Hemopexin-like_repeat	RT		4	4,7	9,8E-5	3,9E-3
	INTERPRO	Pept_M10A	RT		4	4,7	9,8E-5	3,9E-3
	INTERPRO	Pept_M10_metallopeptidase	RT		4	4,7	9,8E-5	3,9E-3
	INTERPRO	Hemopexin-like_dom	RT		4	4,7	9,8E-5	3,9E-3
	UP_SEQ_FEATURE	REPEAT:Hemopexin 1	RT		4	4,7	1,0E-4	1,4E-2
	UP_SEQ_FEATURE	REPEAT:Hemopexin 2	RT		4	4,7	1,0E-4	1,4E-2
	GOTERM_MF_DIRECT	serine-type endopeptidase activity	RT		7	8,1	1,1E-4	1,7E-2
	INTERPRO	Hemopexin-like_dom_sf	RT		4	4,7	1,1E-4	3,9E-3
	SMART	HX	RT		4	4,7	1,9E-4	1,1E-2
	INTERPRO	Peptidase_Metallo	RT		4	4,7	2,0E-4	6,1E-3

Figure 13: **David chart for RF most important variables**



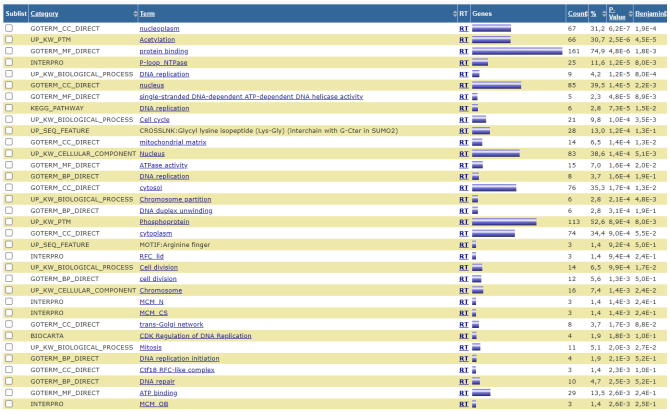
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	UP_KW_BIOLOGICAL_PROCESS	Cell cycle	RT		82	19,8	9,8E-35	8,8E-33
	GOTERM_CC_DIRECT	nucleoplasm	RT		181	43,6	4,4E-33	2,0E-30
	UP_KW_CELLULAR_COMPONENT	Chromosome	RT		70	16,9	2,3E-31	1,0E-29
	UP_KW_BIOLOGICAL_PROCESS	Cell division	RT		58	14,0	2,3E-27	1,0E-25
	GOTERM_BP_DIRECT	cell division	RT		50	12,0	1,8E-25	3,6E-22
	UP_KW_BIOLOGICAL_PROCESS	Mitosis	RT		47	11,3	1,0E-24	3,0E-23
	GOTERM_CC_DIRECT	nucleus	RT		204	49,2	9,8E-21	2,3E-18
	UP_KW_CELLULAR_COMPONENT	Nucleus	RT		206	49,6	3,8E-20	8,1E-19
	KEGG_PATHWAY	Cell cycle	RT		32	7,7	5,6E-20	1,3E-17
	UP_KW_BIOLOGICAL_PROCESS	DNA replication	RT		26	6,3	1,4E-19	3,1E-18
	GOTERM_BP_DIRECT	DNA replication	RT		25	6,0	1,6E-17	1,6E-14
	UP_KW_BIOLOGICAL_PROCESS	DNA repair	RT		41	9,9	7,6E-16	1,4E-14
	GOTERM_BP_DIRECT	DNA repair	RT		34	8,2	1,8E-15	1,2E-12
	UP_KW_BIOLOGICAL_PROCESS	DNA damage	RT		44	10,6	2,8E-15	4,1E-14
	UP_KW_CELLULAR_COMPONENT	Centromere	RT		25	6,0	3,9E-15	5,6E-14
	GOTERM_MF_DIRECT	protein binding	RT		331	79,8	7,2E-15	2,6E-12
	INTERPRO	P-loop_NTPase	RT		58	14,0	7,3E-15	9,2E-12
	GOTERM_MF_DIRECT	ATPase activity	RT		40	9,6	8,6E-15	2,6E-12
	GOTERM_CC_DIRECT	kinetochore	RT		24	5,8	3,4E-14	4,5E-12
	UP_SEQ_FEATURE	CROSSLNK:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2)	RT		68	16,4	3,6E-14	8,4E-11
	GOTERM_MF_DIRECT	single-stranded DNA-dependent ATP-dependent DNA helicase activity	RT		12	2,9	3,7E-14	7,5E-12

Figure 14: **David chart for LDA most important variables**



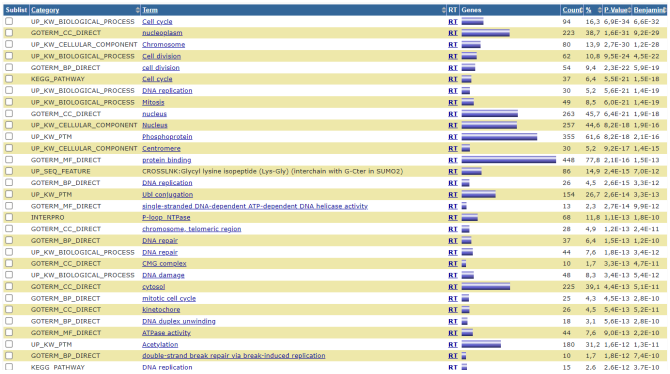
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_MF_DIRECT	single-stranded DNA-dependent ATP-dependent DNA helicase activity	RT		3	6,7	6,2E-4	5,0E-2
	GOTERM_BP_DIRECT	apoptotic process	RT		6	13,3	2,8E-3	6,1E-1
	UP_KW_DOMAIN	Coiled coil	RT		12	26,7	3,3E-3	3,3E-2
	GOTERM_BP_DIRECT	DNA duplex unwinding	RT		3	6,7	4,8E-3	6,1E-1
	UP_SEQ_FEATURE	COMPBIAS:Polar residues	RT		22	48,9	4,9E-3	1,0E0
	GOTERM_MF_DIRECT	receptor binding	RT		5	11,1	5,2E-3	2,1E-1
	INTERPRO	RFC_lid	RT		2	4,4	9,2E-3	1,0E0
	UP_KW_PTM	Acetylation	RT		16	35,6	1,0E-2	9,0E-2
	GOTERM_MF_DIRECT	NEDD8 transferase activity	RT		2	4,4	1,1E-2	3,0E-1
	GOTERM_CC_DIRECT	CHL18 RFC-like complex	RT		2	4,4	1,3E-2	1,0E0
	GOTERM_BP_DIRECT	positive regulation of DNA-directed DNA polymerase activity	RT		2	4,4	1,5E-2	9,4E-1
	UP_KW_BIOLOGICAL_PROCESS	Cell cycle	RT		6	13,3	1,5E-2	2,3E-1
	GOTERM_MF_DIRECT	DNA clamp loader activity	RT		2	4,4	1,7E-2	3,4E-1
	GOTERM_BP_DIRECT	DNA replication	RT		3	6,7	1,7E-2	9,4E-1
	UP_KW_BIOLOGICAL_PROCESS	DNA replication	RT		3	6,7	1,8E-2	2,3E-1
	GOTERM_BP_DIRECT	protein localization to kinetochore	RT		2	4,4	1,8E-2	9,4E-1
	UP_SEQ_FEATURE	CROSSLNK:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2)	RT		7	15,6	2,8E-2	1,0E0
	GOTERM_MF_DIRECT	protein binding	RT		30	66,7	3,2E-2	5,2E-1
	GOTERM_BP_DIRECT	regulation of axonogenesis	RT		2	4,4	3,3E-2	1,0E0
	UP_SEQ_FEATURE	COMPBIAS:Basic and acidic residues	RT		17	37,8	3,7E-2	1,0E0
	GOTERM_BP_DIRECT	protein neddylation	RT		2	4,4	3,8E-2	1,0E0
	GOTERM_CC_DIRECT	nucleoplasm	RT		12	26,7	4,1E-2	1,0E0
	GOTERM_MF_DIRECT	ATPase activity	RT		4	8,9	4,5E-2	6,0E-1
	GOTERM_BP_DIRECT	post-translational protein modification	RT		2	4,4	5,7E-2	1,0E0
	GOTERM_MF_DIRECT	ATP binding	RT		7	15,6	6,1E-2	7,1E-1
	KEGG_PATHWAY	Axon guidance	RT		3	6,7	6,5E-2	1,0E0
	UP_KW_CELLULAR_COMPONENT	Nucleus	RT		15	33,3	7,6E-2	9,7E-1
	GOTERM_BP_DIRECT	ephrin receptor signaling pathway	RT		2	4,4	8,0E-2	1,0E0
	KEGG_PATHWAY	DNA replication	RT		2	4,4	8,0E-2	1,0E0
	GOTERM_BP_DIRECT	bicellular tight junction assembly	RT		2	4,4	8,3E-2	1,0E0
	UP_KW_CELLULAR_COMPONENT	Chromosome	RT		4	8,9	8,3E-2	9,7E-1
	BIOCARTA	Agrin in Postsynaptic Differentiation	RT		2	4,4	8,8E-2	1,0E0
	INTERPRO	ATPase_AAA_core	RT		2	4,4	9,2E-2	1,0E0
	UP_SEQ_FEATURE	DOMAIN:AAA+ ATPase	RT		2	4,4	9,3E-2	1,0E0

Figure 15: **David chart for Ridge most important variables**



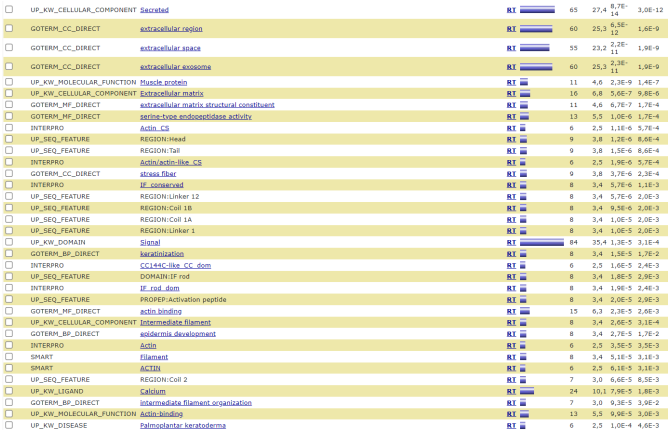
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_CC_DIRECT	nucleoplasm	RT		67	31,2	6,2E-7	1,9E-4
	UP_KW_PTM	Acetylation	RT		66	30,7	2,5E-6	4,5E-5
	GOTERM_MF_DIRECT	protein binding	RT		161	74,9	4,8E-6	1,8E-3
	INTERPRO	P-loop_NTPase	RT		25	11,6	1,2E-5	8,0E-3
	UP_KW_BIOLOGICAL_PROCESS	DNA replication	RT		9	4,2	1,2E-5	8,0E-4
	GOTERM_CC_DIRECT	nucleus	RT		85	39,5	1,4E-5	2,2E-3
	GOTERM_MF_DIRECT	single-stranded DNA-dependent ATP-dependent DNA helicase activity	RT		5	2,3	4,8E-5	8,9E-3
	KEGG_PATHWAY	DNA replication	RT		6	2,8	7,3E-5	1,5E-2
	UP_KW_BIOLOGICAL_PROCESS	Cell cycle	RT		21	9,8	1,0E-4	3,5E-3
	UP_SEQ_FEATURE	CROSSLNK:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2)	RT		28	13,0	1,2E-4	1,3E-1
	GOTERM_CC_DIRECT	mitochondrial matrix	RT		14	6,5	1,4E-4	1,3E-2
	UP_KW_CELLULAR_COMPONENT	Nucleus	RT		83	38,6	1,4E-4	5,1E-3
	GOTERM_MF_DIRECT	ATPase activity	RT		15	7,0	1,6E-4	2,0E-2
	GOTERM_BP_DIRECT	DNA replication	RT		8	3,7	1,6E-4	1,9E-1
	GOTERM_CC_DIRECT	cytosol	RT		76	35,3	1,7E-4	1,3E-2
	UP_KW_BIOLOGICAL_PROCESS	Chromosome partition	RT		6	2,8	2,1E-4	4,8E-3
	GOTERM_BP_DIRECT	DNA duplex unwinding	RT		6	2,8	3,1E-4	1,9E-1
	UP_KW_PTM	Phosphoprotein	RT		113	52,6	8,9E-4	8,0E-3
	GOTERM_CC_DIRECT	cytoplasm	RT		74	34,4	9,0E-4	5,5E-2
	UP_SEQ_FEATURE	MOTIF:Arginine finger	RT		3	1,4	9,2E-4	5,0E-1
	INTERPRO	RFC_lid	RT		3	1,4	9,4E-4	2,4E-1
	UP_KW_BIOLOGICAL_PROCESS	Cell division	RT		14	6,5	9,9E-4	1,7E-2
	GOTERM_BP_DIRECT	cell division	RT		12	5,6	1,3E-3	5,0E-1
	UP_KW_CELLULAR_COMPONENT	Chromosome	RT		16	7,4	1,4E-3	2,4E-2
	INTERPRO	MCM_N	RT		3	1,4	1,4E-3	2,4E-1
	INTERPRO	MCM_CS	RT		3	1,4	1,4E-3	2,4E-1
	GOTERM_CC_DIRECT	trans-Golgi network	RT		8	3,7	1,7E-3	8,8E-2
	BIOCARTA	CDK Regulation of DNA Replication	RT		4	1,9	1,8E-3	1,0E-1
	UP_KW_BIOLOGICAL_PROCESS	Mitosis	RT		11	5,1	2,0E-3	2,7E-2
	GOTERM_BP_DIRECT	DNA replication initiation	RT		4	1,9	2,1E-3	5,2E-1
	GOTERM_CC_DIRECT	CHL18 RFC-like complex	RT		3	1,4	2,2E-3	1,0E-1
	GOTERM_BP_DIRECT	DNA repair	RT		10	4,7	2,5E-3	5,2E-1
	GOTERM_MF_DIRECT	ATP binding	RT		29	13,5	2,6E-3	2,4E-1
	INTERPRO	MCM_OB	RT		3	1,4	2,6E-3	2,5E-1

Figure 16: **David chart for LASSO most important variables**



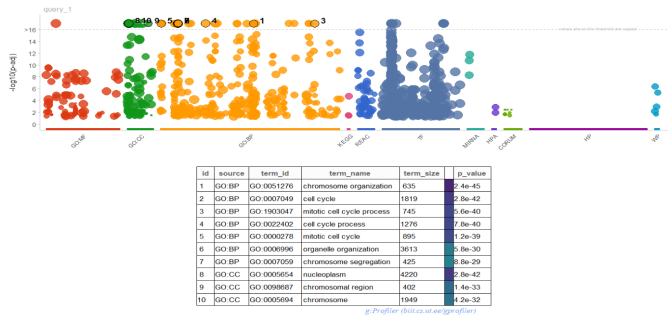
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	UP_KW_BIOLOGICAL_PROCESS	Cell cycle	RT		94	16,3	6,9E-34	6,6E-32
	GOTERM_CC_DIRECT	nucleoplasm	RT		223	38,7	1,6E-31	9,2E-29
	UP_KW_CELLULAR_COMPONENT	Chromosome	RT		80	13,9	2,7E-30	1,2E-28
	UP_KW_BIOLOGICAL_PROCESS	Cell division	RT		62	10,8	9,5E-24	4,5E-22
	GOTERM_BP_DIRECT	cell division	RT		54	9,4	2,3E-22	5,9E-19
	KEGG_PATHWAY	Cell cycle	RT		37	6,4	5,5E-21	1,5E-18
	UP_KW_BIOLOGICAL_PROCESS	DNA replication	RT		30	5,2	5,6E-21	1,4E-19
	UP_KW_BIOLOGICAL_PROCESS	Mitosis	RT		49	8,5	6,0E-21	1,4E-19
	GOTERM_CC_DIRECT	nucleus	RT		263	45,7	6,4E-21	1,9E-18
	UP_KW_CELLULAR_COMPONENT	Nucleus	RT		257	44,6	8,2E-18	1,9E-16
	UP_KW_PTM	Phosphoprotein	RT		355	61,6	8,2E-18	2,1E-16
	UP_KW_CELLULAR_COMPONENT	Centromere	RT		30	5,2	9,2E-17	1,4E-15
	GOTERM_MF_DIRECT	protein binding	RT		448	77,8	2,1E-16	1,5E-13
	UP_SEQ_FEATURE	CROSSLNK:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2)	RT		86	14,9	2,4E-15	7,0E-12
	GOTERM_BP_DIRECT	DNA replication	RT		26	4,5	2,6E-15	3,3E-12
	UP_KW_PTM	Ubl conjugation	RT		154	26,7	2,6E-14	3,3E-13
	GOTERM_MF_DIRECT	single-stranded DNA-dependent ATP-dependent DNA helicase activity	RT		13	2,3	2,7E-14	9,9E-12
	INTERPRO	P-loop_NTPase	RT		68	11,8	1,1E-13	1,8E-10
	GOTERM_CC_DIRECT	chromosome, telomeric region	RT		28	4,9	1,2E-13	2,4E-11
	GOTERM_BP_DIRECT	DNA repair	RT		37	6,4	1,5E-13	1,2E-10
	UP_KW_BIOLOGICAL_PROCESS	DNA repair	RT		44	7,6	1,8E-13	3,4E-12
	GOTERM_CC_DIRECT	CMG complex	RT		10	1,7	3,3E-13	4,7E-11
	UP_KW_BIOLOGICAL_PROCESS	DNA damage	RT		48	8,3	3,4E-13	5,4E-12
	GOTERM_CC_DIRECT	cytosol	RT		225	39,1	4,4E-13	5,1E-11
	GOTERM_BP_DIRECT	mitotic cell cycle	RT		25	4,3	4,5E-13	2,8E-10
	GOTERM_CC_DIRECT	kinetochore	RT		26	4,5	5,4E-13	5,2E-11
	GOTERM_BP_DIRECT	DNA duplex unwinding	RT		18	3,1	5,6E-13	2,8E-10
	GOTERM_MF_DIRECT	ATPase activity	RT		44	7,6	9,0E-13	2,2E-10
	UP_KW_PTM	Acetylation	RT		180	31,2	1,6E-12	1,3E-11
	GOTERM_BP_DIRECT	double-strand break repair via break-induced replication	RT		10	1,7	1,8E-12	7,4E-10
	KEGG_PATHWAY	DNA replication	RT		15	2,6	2,6E-12	3,7E-10

Figure 17: **David chart for SCUDO most important variables**



	UP_KW_CELLULAR_COMPONENT	Secreted	RT		65	27,4	8,7E-14	3,0E-12
	GOTERM_CC_DIRECT	extracellular region	RT		60	25,3	6,5E-12	1,6E-9
	GOTERM_CC_DIRECT	extracellular space	RT		55	23,2	2,2E-11	1,9E-9
	GOTERM_CC_DIRECT	extracellular exosome	RT		60	25,3	2,3E-11	1,9E-9
	UP_KW_MOLECULAR_FUNCTION	Muscle protein	RT		11	4,6	2,3E-9	1,4E-7
	UP_KW_CELLULAR_COMPONENT	Extracellular matrix	RT		16	6,8	5,6E-7	9,8E-6
	GOTERM_MF_DIRECT	extracellular matrix structural constituent	RT		11	4,6	6,7E-7	1,7E-4
	GOTERM_MF_DIRECT	serine-type endopeptidase activity	RT		13	5,5	1,0E-6	1,7E-4
	INTERPRO	Actin_CS	RT		6	2,5	1,1E-6	5,7E-4
	UP_SEQ_FEATURE	REGION:Head	RT		9	3,8	1,2E-6	8,6E-4
	UP_SEQ_FEATURE	REGION:Tail	RT		9	3,8	1,5E-6	8,6E-4
	INTERPRO	Actin/actin-like_CS	RT		6	2,5	1,9E-6	5,7E-4
	GOTERM_CC_DIRECT	stress fiber	RT		9	3,8	3,7E-6	2,3E-4
	INTERPRO	IF_conserved	RT		8	3,4	5,7E-6	1,1E-3
	UP_SEQ_FEATURE	REGION:Linker 12	RT		8	3,4	5,7E-6	2,0E-3
	UP_SEQ_FEATURE	REGION:Coil 1B	RT		8	3,4	9,5E-6	2,0E-3
	UP_SEQ_FEATURE	REGION:Coil 1A	RT		8	3,4	1,0E-5	2,0E-3
	UP_SEQ_FEATURE	REGION:Linker 1	RT		8	3,4	1,0E-5	2,0E-3
	UP_KW_DOMAIN	Signal	RT		84	35,4	1,3E-5	3,1E-4
	GOTERM_BP_DIRECT	keratinization	RT		8	3,4	1,5E-5	1,7E-2
	INTERPRO	CC144C-like_CC_dom	RT		6	2,5	1,6E-5	2,4E-3
	UP_SEQ_FEATURE	DOMAIN:IF rod	RT		8	3,4	1,8E-5	2,9E-3
	INTERPRO	IF_rod_dom	RT		8	3,4	1,9E-5	2,4E-3
	UP_SEQ_FEATURE	PROPEP:Activation peptide	RT		8	3,4	2,0E-5	2,9E-3
	GOTERM_MF_DIRECT	actin binding	RT		15	6,3	2,3E-5	2,6E-3
	UP_KW_CELLULAR_COMPONENT	Intermediate filament	RT		8	3,4	2,6E-5	3,1E-4
	GOTERM_BP_DIRECT	epidermis development	RT		8	3,4	2,7E-5	1,7E-2
	INTERPRO	Actin	RT		6	2,5	3,5E-5	3,5E-3
	SMART	Filament	RT		8	3,4	5,1E-5	3,1E-3
	SMART	ACTIN	RT		6	2,5	6,1E-5	3,1E-3
	UP_SEQ_FEATURE	REGION:Coil 2	RT		7	3,0	6,6E-5	8,5E-3
	UP_KW_LIGAND	Calcium	RT		24	10,1	7,9E-5	1,8E-3
	GOTERM_BP_DIRECT	intermediate filament organization	RT		7	3,0	9,3E-5	3,9E-2
	UP_KW_MOLECULAR_FUNCTION	Actin-binding	RT		13	5,5	9,9E-5	3,0E-3
	UP_KW_DISEASE	Palmoplantar keratoderma	RT		6	2,5	1,0E-4	4,6E-3

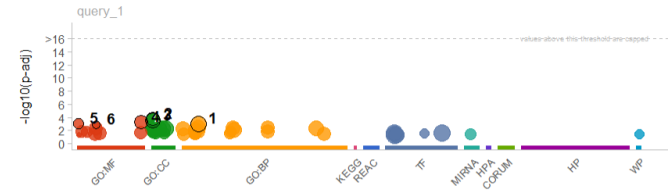
Figure 18: **David chart from selected genes using LIMMA**



id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0051276	chromosome organization	635	2.4e-45
2	GO:BP	GO:0007049	cell cycle	1819	2.8e-42
3	GO:BP	GO:1903047	mitotic cell cycle process	745	5.6e-40
4	GO:BP	GO:0022402	cell cycle process	1276	7.8e-40
5	GO:BP	GO:0000278	mitotic cell cycle	895	1.2e-39
6	GO:BP	GO:0006996	organelle organization	3613	5.8e-30
7	GO:BP	GO:0007059	chromosome segregation	425	8.8e-29
8	GO:CC	GO:0005654	nucleoplasm	4220	2.8e-42
9	GO:CC	GO:0098687	chromosomal region	402	1.4e-33
10	GO:CC	GO:0005694	chromosome	1949	4.2e-32

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

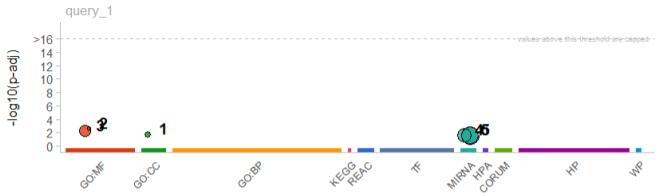
Figure 19: **gProfiler for LDA most important variables**



id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0006996	organelle organization	3613	1.0e-03
2	GO:CC	GO:0005759	mitochondrial matrix	496	2.0e-04
3	GO:CC	GO:0005694	chromosome	1949	3.2e-04
4	GO:MF	GO:0140657	ATP-dependent activity	580	6.0e-04
5	GO:MF	GO:0003678	DNA helicase activity	59	9.7e-04
6	GO:MF	GO:0017116	single-stranded DNA helicase activity	23	1.4e-03

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

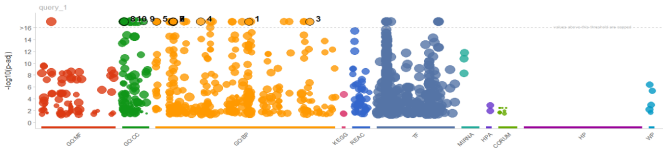
Figure 20: **gProfiler for LASSO most important variables**



id	source	term_id	term_name	term_size	p_value
1	GO:CC	GO:0031390	Ctf18 RFC-like complex	8	2.3e-02
2	GO:MF	GO:0019788	NEDD8 transferase activity	7	2.6e-03
3	GO:MF	GO:0017116	single-stranded DNA helicase activity	23	5.9e-03
4	MIRNA	MIRNA:hsa-miR-3169	hsa-miR-3169	47	2.8e-02
5	MIRNA	MIRNA:hsa-miR-519e-5p	hsa-miR-519e-5p	185	2.8e-02
6	MIRNA	MIRNA:hsa-miR-515-5p	hsa-miR-515-5p	188	3.1e-02

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

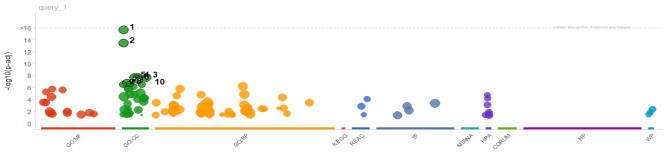
Figure 21: **gProfiler for Ridge most important variables**



id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0051276	chromosome organization	635	2.7e-45
2	GO:BP	GO:0007049	cell cycle	1819	3.3e-42
3	GO:BP	GO:1903047	mitotic cell cycle process	745	9.2e-40
4	GO:BP	GO:0022402	cell cycle process	1276	8.9e-40
5	GO:BP	GO:0000278	mitotic cell cycle	895	1.3e-39
6	GO:BP	GO:0006996	organelle organization	3613	3.0e-30
7	GO:BP	GO:0007059	chromosome segregation	425	9.4e-29
8	GO:CC	GO:0005654	nucleoplasm	4220	3.6e-42
9	GO:CC	GO:0098687	chromosomal region	402	1.5e-33
10	GO:CC	GO:0005694	chromosome	1949	4.9e-32

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

Figure 22: **gProfiler for SCUDO most important variables**



id	source	term_id	term_name	term_size	p_value
1	GO:CC	GO:0005576	extracellular region	4213	2.1e-16
2	GO:CC	GO:0005615	extracellular space	3303	3.3e-14
3	GO:CC	GO:1903561	extracellular vesicle	2133	2.0e-08
4	GO:CC	GO:0065010	extracellular membrane-bounded organelle	2134	2.0e-08
5	GO:CC	GO:0043230	extracellular organelle	2134	2.0e-08
6	GO:CC	GO:0070062	extracellular exosome	2109	4.6e-08
7	GO:CC	GO:0015629	actin cytoskeleton	515	1.4e-07
8	GO:CC	GO:0032432	actin filament bundle	95	2.3e-07
9	GO:CC	GO:0001725	stress fiber	86	3.3e-07
10	GO:CC	GO:0097517	contractile actin filament bundle	86	3.3e-07

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

Figure 23: **gProfiler for LIMMA most important variables**

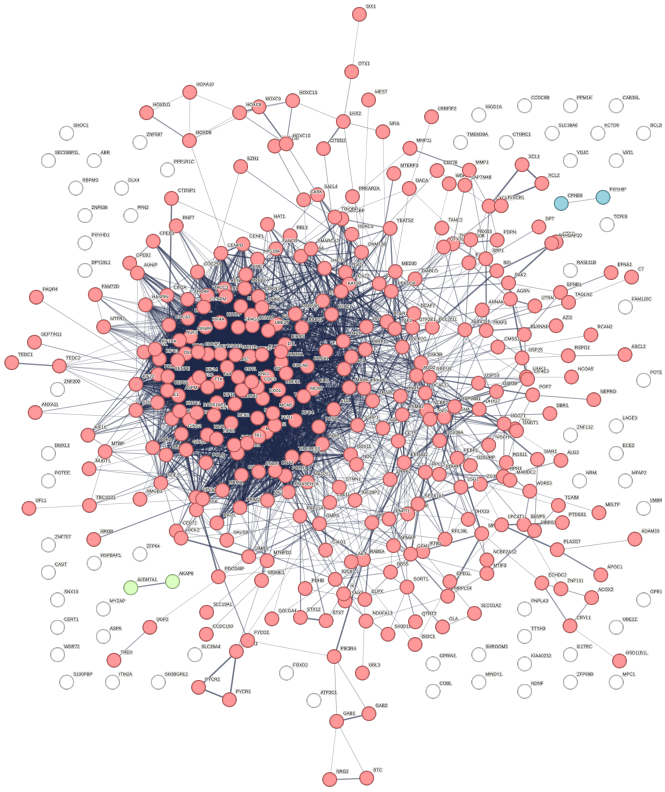


Figure 24: **Clusters obtained from LDA STRING-network**. In red, cluster 1, with a total number of genes equal to 338. Overall number of nodes: 405

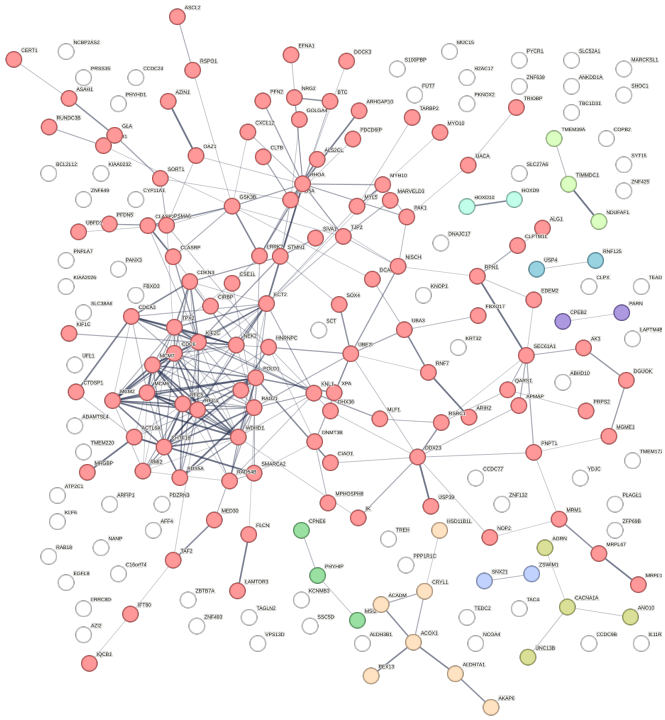


Figure 25: **Clusters obtained from LASSO STRING-network**. In red, cluster 1, with a total number of genes equal to 110. Overall number of nodes: 293

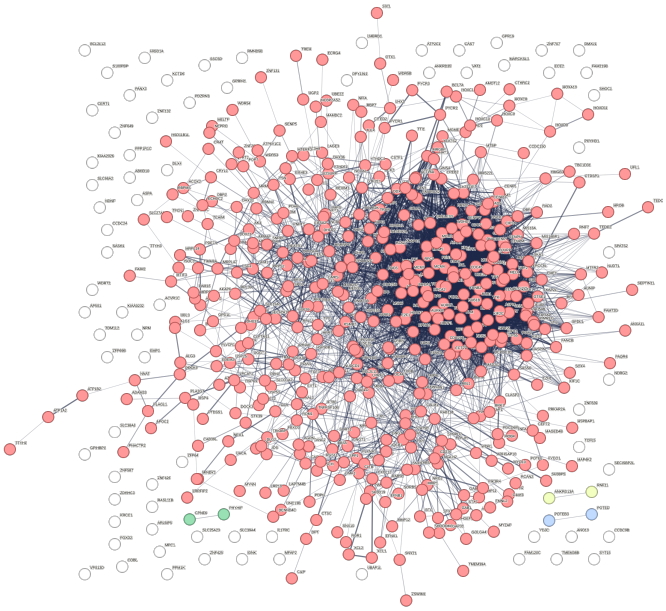


Figure 26: **Clusters obtained from SCUDO STRING-network**. In red, cluster 1, with a total number of genes equal to 455. Overall number of nodes: 561

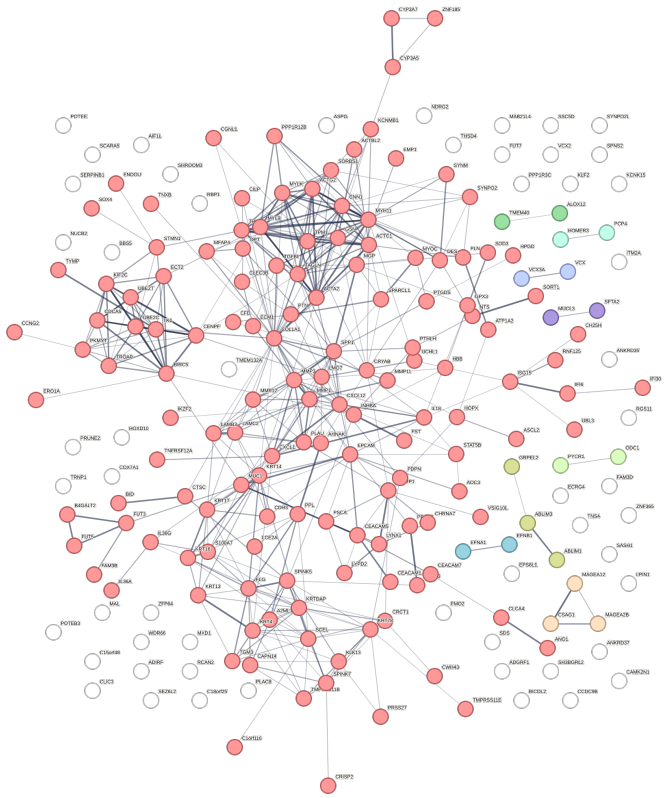


Figure 27: **Clusters obtained from LIMMA STRING-network**. In red, cluster 1, with a total number of genes equal to 141. Overall number of nodes: 265

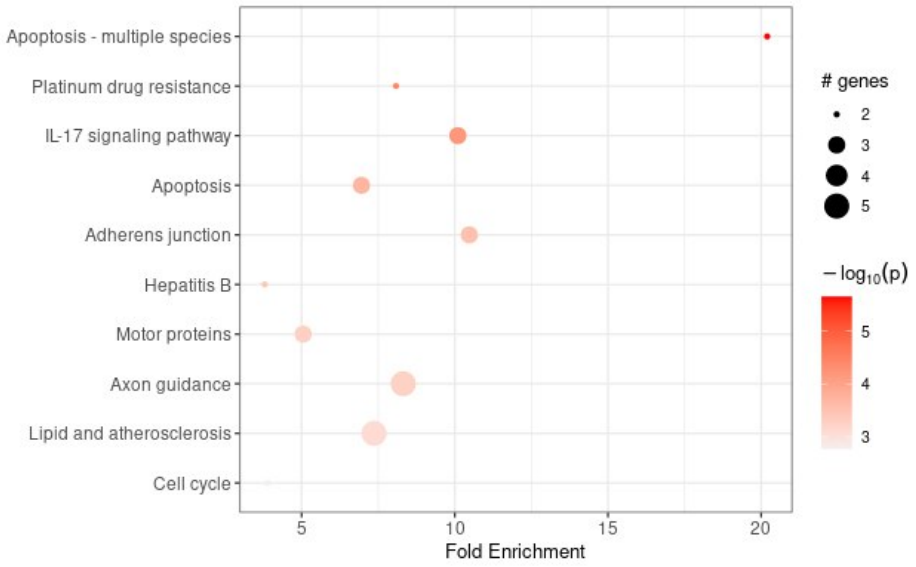


Figure 28: **PathFindR enrichment chart on important variables extracted from the Random Forest model.** Results obtained selecting KEGG as geneset.