Analysis of Colorectal cancer from Microarray data

Andrea Tonina (ID:239615)

Supervisor: Prof. Mario Lauria

3 September 2024

Contents

Chapter 1

Introduction

Colorectal cancer (CRC), which includes both colon and rectal cancer, has become a significant health concern. It ranks as the third most commonly diagnosed cancer and the second leading cause of cancer-related deaths, with approximately 9.4% of cancer-related deaths in 2020 attributed to CRC [1].
Colorectal cancer (CRC) exhibits significant heterogeneity, with three primary histological subtypes: adenocarcinoma (AC), mucinous adenocarcinoma (MAC), and signet ring cell carcinoma (SRCC). While Adenocarcinoma is the most prevalent colorectal cancer, the other two subtypes are rare and differ in characteristics [2].

This project aims to uncover new findings about Colorectal cancer (CRC) and seeks to advance the understanding of this tumor by establishing a set of variables capable of discriminating between cancer and normal samples. The workflow for this project can be summarized into three main steps:

- **Exploratory Analysis**: In the first part of the workflow, an exploratory data analysis was performed using unsupervised methods.
- **Supervised Learning**: The second part involved utilizing supervised learning methods as classifiers to identify the most significant variables.
- **Functional Enrichment Analysis**: Finally, it was used the important variables found in the previous step as inputs for various tools that perform Over-Representation Analysis and Network-based Analysis to identify enriched terms.

Chapter 2

Materials and Methods

2.1 Dataset Description

The dataset used for this project was taken from the scientific paper *Vlachavas EI et al.* (2019) [3], which is freely available on GEO (*Gene Expression Omnibus*) with accession number GSE110225, present 60 samples divided equally between tumor and control cases from 30 patients with confirmed primary colorectal adenocarcinomas.
The chosen dataset presents the characteristic to be composed of two 'sub-datasets': GSE110223 and GSE110224. The first one is generated from the Affymetrix Human Genome U133A Array and contains 26 samples which are evenly distributed between tumor and control, the latter is generated from the Affymetrix Human Genome U133 Plus 2.0 Array and consists of 34 samples (17 control, 17 tumor). Of particular note, for each patient, a tumor and a control sample were obtained from the colon through surgical intervention.

2.2 Unsupervised Learning Methods

Principal Component Analysis (PCA), **K-means** and **Hierarchical Clustering** have been used to perform an initial exploratory analysis and verify whether the two groups can be correctly recognized and distinguished.

PCA is a statistical technique that transforms a data table with multiple variables into a smaller set of linearly uncorrelated components, known as principal components. These components capture the maximum variance present in the original variables [4]. In essence, PCA performs dimensionality reduction on complex data while retaining essential information and this analysis has been performed with the function `prcomp` from the library `stats`.

K-means clustering is an unsupervised learning technique that aims to partition a set of observations into k clusters. Each observation belongs to the cluster with the nearest mean (also known as the cluster center or centroid), which serves as a prototype for that cluster. The algorithm iteratively relocates data points to form these clusters. The number of initial centroids (k) must be chosen during initialization, often based on prior knowledge or empirical insights. K-means relies on a distance measure (usually Euclidean distance) to optimize the clustering. Its objective is to maximize the distance between clusters while minimizing the distance within each cluster [5]. For this task, the R package `stat` and the function `kmeans` were used.

Hierarchical Clustering is a data analysis technique that groups data points based on their similarity or distance. It creates a cluster tree, known as a dendrogram, illustrating the observations' hierarchical relationships. There are two possible approaches:

- *Agglomerative Clustering*: Each observation starts in its own cluster, and the algorithm iteratively merges the closest clusters until all data points belong to a single cluster.
- *Divisive Clustering*: All observations start in a single unique cluster, and the algorithm recursively splits the cluster into smaller subclusters based on similarity or distance.

The dendrogram can be sliced at different heights to determine the ideal number of clusters. This algorithm requires both a linkage criterion and a distance measure [6]. In this case *average*, *complete*, and *single linkage* methods along with Euclidean distance have been employed and the function used is `hclust` from the `stat` R package.

2.3 Supervised Learning Methods

Supervised learning encompasses a class of machine learning techniques that construct models based on labeled data. In this context, 'supervised' implies that the training data includes explicit labels that indicate the relationship between input features and corresponding outputs. These methods utilize both input observations and their associated labels to create accurate predictions for unseen data instances. The primary objective is to identify underlying patterns and relationships

within labeled datasets and generalize that knowledge to make predictions [7].
The data used as input for the models underwent a feature selection through a t-test on the rows (function `rowttests`) to identify the genes that looked significantly different in the two groups considering only p-values lower than 0.01.
This procedure resulted in a decrease in the number of variables in the dataset, from 22277 variables to 4944 variables (probe IDs).
Each method has been tested through 10-fold cross-validation to asses the best-performing one and the best parameters to be chosen (hyper-parameter tuning). The cross-validation was performed using the `train` function from package `caret`.

Random Forest is a widely-used machine learning algorithm that combines the output of multiple decision trees to reach a single result. These trees are created by resampling the input features using a "bootstrap" procedure. The predictions from these numerous trees are then merged calculating a mean value. Additionally, the final tree allows us to extract the most important feature (such as genes) that guide the prediction or clusterization [8].

Using the function `tuneRF` of `iRF` package the best parameters were chosen to train the model. The number of trees to be created by Random Forest was selected after several trials and the final value for parameter `ntree` is equal to 1000, then the chosen `mtry` parameter is 35 because it is the one linked with the lowest OBB error.
Linear Discriminant Analysis (LDA) is an approach used in supervised machine learning to solve multi-class classification problems by separating multiple classes with multiple features through data dimensionality reduction. It works by identifying a linear combination of features (or linear decision surface) used to separate or characterize two or more classes of objects or events and it does this by projecting data with two or more dimensions into one dimension so that it can be more easily classified [9]. There is the necessity to choose a prior probability to train the model, with the knowledge that the data are equally represented between Tumor and Control samples the parameter `prior` of the function `train` was set adequately based on that knowledge (`prior = c(0.5,0.5)`).
Lasso and **Ridge regression** are shrinkage methods documented in various research papers [10, 11] that improve the predictive accuracy of a model. They achieve this by constraining the coefficients of the model and so effectively reducing the model's complexity and preventing overfitting.

In **Lasso regression**, the constraint can drive some coefficients to exactly zero. This results in a sparse model, where irrelevant features are eliminated, leading to better interpretability and potential computational efficiency [11].

Ridge regression also shrinks coefficients towards zero but doesn't allow them to become exactly zero. This helps to address multicollinearity and improve model stability [12].

The key difference between **Lasso** and **Ridge** lies in the specific type of constraint they apply during this shrinkage process. Lasso can drive some coefficients all the way to zero, performing feature selection. Ridge regression, on the other hand, uses and shrinks all coefficients toward zero but keeps them in the model[13]. Through the utilization of the parameter `TuneGrid` of the function `train` it was possible to retrieve the best parameters to train the models. After the training step, the `lambda` chosen were 0.2405576 for **Lasso** and 0.07607098 for **Ridge**.
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 [14]. The best parameters to train the model were selected using a trial process via the `scudoModel` function, part of the `rSCUDO` package. The final values used for the model were `nTop = 100`, `nBottom = 100`, `N = 0.25`, `maxDist = 1` and `beta = 1`.

All the models were then compared based on their accuracy and from the best method, therefore the one with the highest accuracy, a list of the most significant probe IDs was extracted for further analysis through the utilization of the function `varImp` of the `caret` R library.

2.4 Functional Enrichment Analysis

Gene Set Analysis (GSA), also known as functional enrichment analysis, is a powerful approach used for analyzing high-throughput experimental data and interpreting gene expression data. The final goal of these analyses is to identify enriched or over-represented biological functions and pathways among a list of interest compared to a reference background. These methods are usually associated with a gene-level statistic that is used to highlight the significant expression patterns between different conditions [15]. Different tools have been used to perform this part of the analysis:

- *Over-Representation Analysis*: This analysis, as the name suggests, is used to determine if a pre-defined set is present more (over-represented) than expected in a subset of the input data. In this study, **gProfiler** (in R and online) 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.

Chapter 3

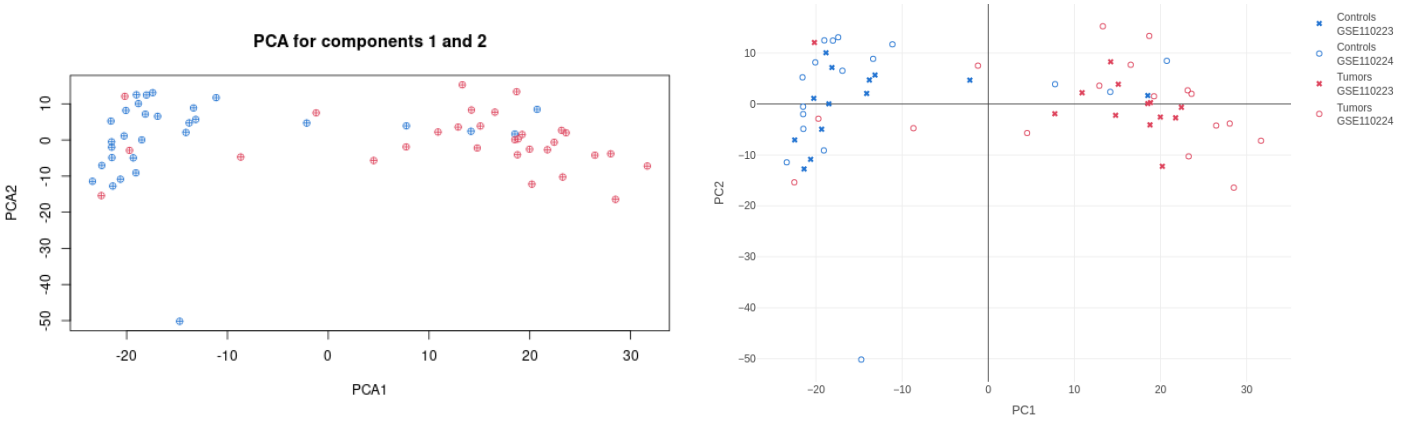
Results

3.1 Pre-processing and Data visualization

Before initiating the analysis workflow, it was necessary to pre-process the data. This involved merging the two datasets and retaining only the probe IDs present in both. To address batch effects commonly encountered in microarray experiments, *Combat* was applied to the data. Combat is a robust method that uses an empirical Bayes framework to correct for technical and non-technical biases [16]. Following batch correction, the data was normalized using *scale* and *median-to-zero* techniques. The effectiveness of these normalization methods can be visualized through the resulting boxplots (Appendix Figure 5.1).

3.2 Unsupervised Learning Methods

A **Principal Component Analysis** (PCA) was performed on the entire gene expression dataset and metadata information was incorporated to examine how the data get stratified. It's possible to see from the results in Figure 3.1(a), that the two groups are not perfectly separated, however, the first principal component seems to distinguish fairly well between tumor and control data except for some samples. It is also possible to see from Figure 3.1(b) that there is no principal component that separates the samples based on their data origin (GSE110223 and GSE110224). Other two PCAs were performed while keeping the two datasets separated (Appendix Figure 5.2), also in the two cases the Tumor and Control groups are discerned by the first principal components even though not perfectly.

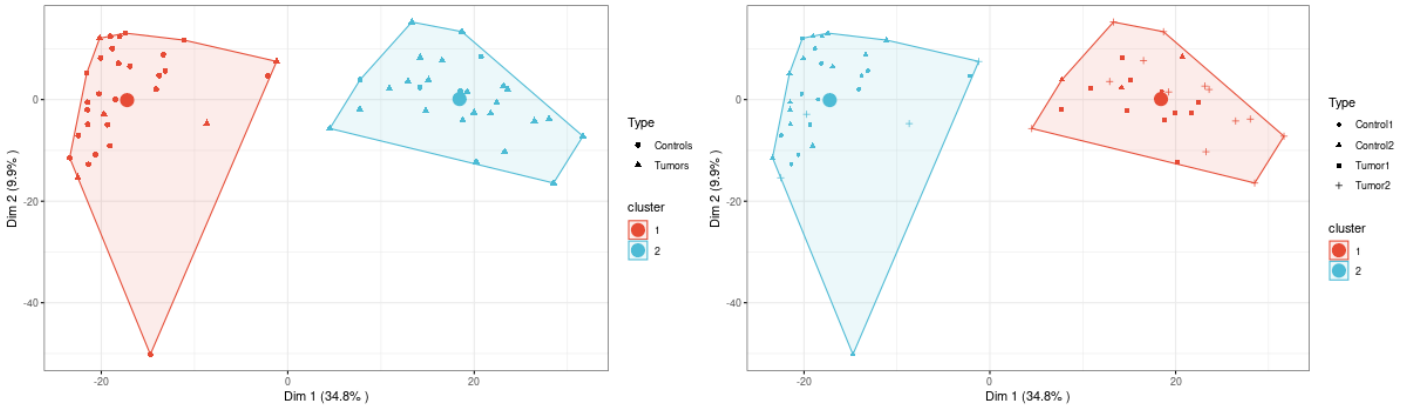


(a) *PCA 2D stratified Tumor vs Control*

(b) *PCA 2D stratified Tumor vs Control and also origin of the data*

Figure 3.1: **Principal Component Analysis**. (a) PCA plot in two dimensions, x-axes: PC1; y-axes: PC2; with the control group in blue and the tumor group in red. (b) PCA plots in two dimensions but in this case the data are also stratified based on the dataset of origin in addition to the stratification between Tumor and Control.

K-means has been performed to identify a putative pattern among observations, the model has been fitted with $k = 2$ since the dataset contains information samples of tumor and normal tissues. From Figure 3.2 it is possible to observe the presence of two clusters, in light blue the Controls and in red the Tumors, as observed in the PCA. Indeed, the samples are quite well separated except for a few ones that have been misclassified (with five Tumors misclassified as Control and four Controls misclassified as Tumors). This might be quite encouraging because K-means has identified some patterns in the data.

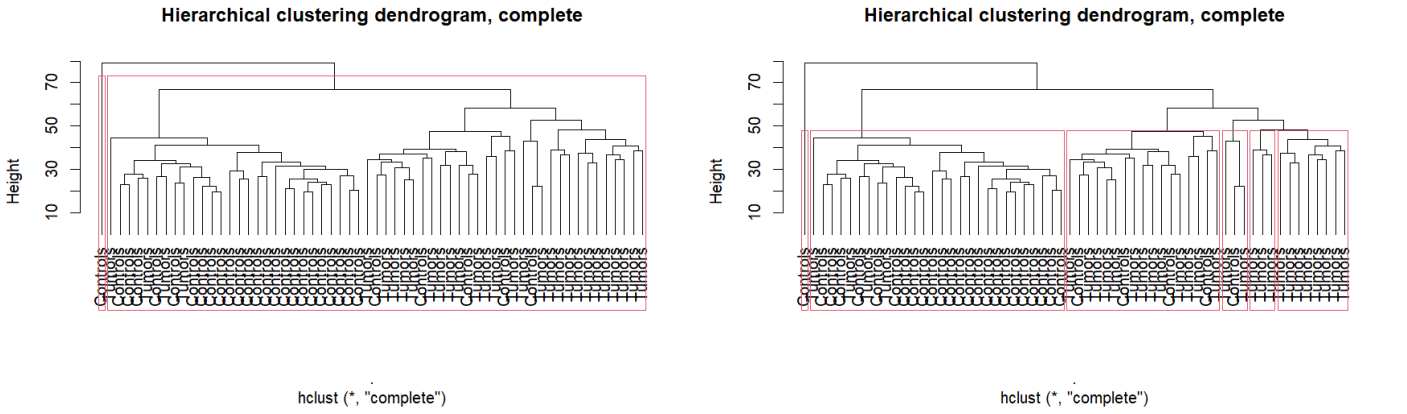


(a) *With information on Control and Tumor*

(b) *With information on Control, Tumor and also differentiate between the origin of the data*

Figure 3.2: **Cluster analysis**. K-mean clustering, where the axes are coordinates for the variables extracted from the first and the second PC.

Regarding **Hierarchical Clustering**, only results retrieved by setting the linkage measure to *complete* are shown and a major difference in the ability to cluster the data can be seen (Fig. 3.3). When k =*2* the resulting clustering cannot be considered informative since one cluster contains almost all the data points, while the other contains only one. Given these results, the number of clusters was modified, by setting $k = 6$, and as a result, a better clustering is observed with clusters presenting only a specific type of samples (Tumors or Controls). Some misclassifications are still present but these results can be considered more reliable, even though it's an unsupervised setting. It is also important to point out that Colorectal cancer is known for its high tumor heterogeneity that often leads to a complex interplay between clonal populations and metastasis generation [17] and because of this, it is possible to see from Figure 3.3(b) that Tumor samples are divided into four different clusters. To see if the **Hierarchical Clustering** generates the clusters based on the data's origin, information was added about which datasets the data comes from (Appendix Figure 5.3). It is possible to observe that there is no clutsering based on the origin datasets of the samples.



(a) *Setting k=2*

(b) *Setting k=6*

Figure 3.3: **Cluster analysis**. Hierarchical clustering, the red boxes represent the identified clusters. The x-axis represents the sample type (Tumors or Controls), and the y-axis the height.

3.3 Supervised Learning Methods

As explained in section 2.3 of *Chapter 2*, five models have been evaluated: **Random Forest**, **Linear Discriminant Analysis**, **Ridge Regression**, **Lasso Regression** and **SCUDO** (training and test networks are shown in Appendix Fig. 5.4). Overall, all methods performed relatively well with a good accuracy (Fig. 3.4). In particular, Random Forest presents the highest accuracy of all (0.7825).

By looking at Figure 3.5(a), which represents the distribution of the out-of-bag (OOB) score over the total number of trees used to train the **RF** model, it's possible to see that as the number of trees increases the OBB error rate, which can

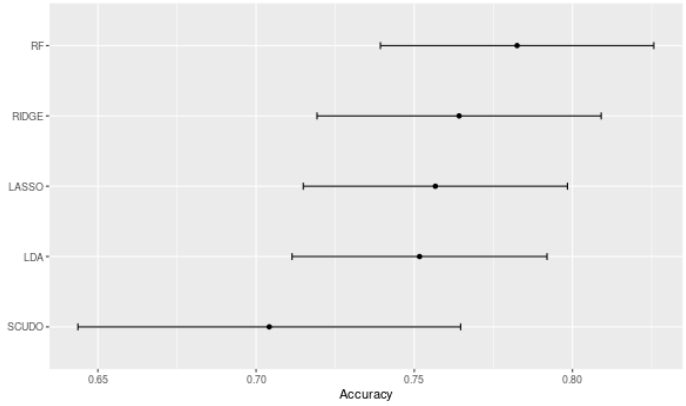
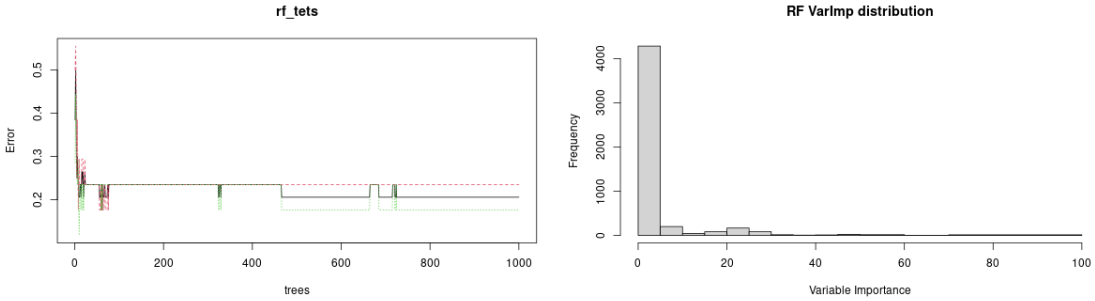
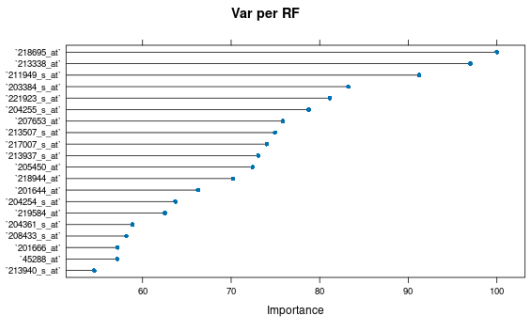


Figure 3.4: **Performance plot of the supervised models.**All models. x-axis: accuracy, y-axis: model name.

be seen as the number of wrongly classified observations, decreases and arrives at 0.2 and remained stationary after circa 750 trees.



(a) *Distribution OOB score over number trees to train* (b) *Variable importance distribution in Random Forest.*



(c) *Top 20 most important variables for the supervised models.*

Figure 3.5: **Random Forest Plots**.

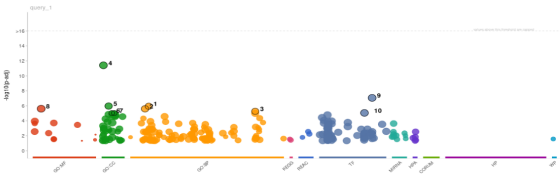
Then from the most accurate model, *Random Forest*, the most important variables were selected by establishing a threshold based on the distribution of the variable importance (Figure 3.5(b)). Indeed from the distribution it is visible that the majority of variables present an importance value that sits between 0 and 30, therefore the threshold set was equal to 30 and from Figure 3.5 it is possible to see the top 20 variables and their importance values. Through this passage, a list of 291 probe IDs was extracted as the most important features, unfortunately, some variables have been lost during the conversion of probe IDs to gene symbols resulting in 277 genes.

3.3.1 Functional Enrichment Analysis

As explained in the previous section, a list of 277 genes was extracted from the best-performing method for classification because the model considered these variables the most informative to conduct their classification task.

Over-Representation Analysis

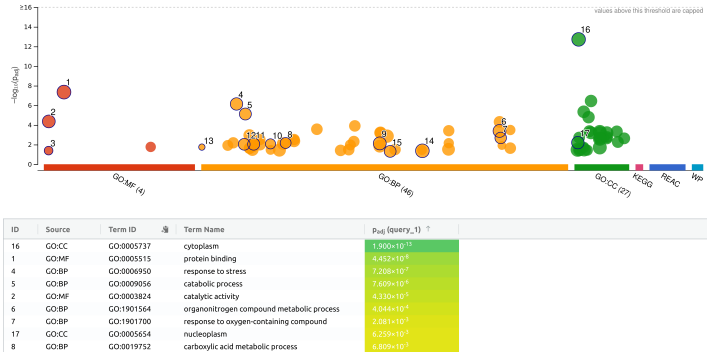
ORA (Over Representation Analysis) was performed using **DAVID** and **gProfiler**. The list of genes from *Random Forest* was given as input for both tools.
From Figures 3.6 and 5.5 it's possible to see that the outputs of the two tools are similar, indeed by considering the terms with the most significant p-values and adjusted p-value (ranked highest: $10^{-12} \sim 10^{-13}$ for *gProfiler* and 10^{-4} for *DAVID*) they concern mainly *extracellular exosome*, *protein binding* and *stress response*.



(a) *Output from R code plot.*

id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0009056	catabolic process	2556	1.1e-06
2	GO:BP	GO:0006950	response to stress	3855	2.5e-06
3	GO:BP	GO:1901575	organic substance catabolic process	2074	5.3e-06
4	GO:CC	GO:0005737	cytoplasm	12345	3.9e-12
5	GO:CC	GO:0031982	vesicle	4004	1.1e-06
6	GO:CC	GO:0042470	melanosome	111	9.1e-06
7	GO:CC	GO:0048770	pigment granule	111	9.1e-06
8	GO:MF	GO:0005515	protein binding	14838	2.4e-06
9	TF	TF:M00716_1	Factor: ZF5; motif: GSGCGCGR; match class: 1	14519	8.8e-08
10	TF	TF:M10072	Factor: sp4; motif: NNGNARGRGGCGGRGCNNRR	10844	8.8e-06

(b) *Output from R code table.*



ID	Source	Term ID	Term Name	P_{adj} (query_1)
16	GO:CC	GO:0005737	cytoplasm	1.900×10^{-13}
1	GO:MF	GO:0005515	protein binding	4.452×10^{-8}
4	GO:BP	GO:0006950	response to stress	7.208×10^{-7}
5	GO:BP	GO:0009056	catabolic process	7.609×10^{-6}
2	GO:MF	GO:0003824	catalytic activity	4.330×10^{-5}
6	GO:BP	GO:1901564	organonitrogen compound metabolic process	4.044×10^{-4}
7	GO:BP	GO:1901700	response to oxygen-containing compound	2.081×10^{-3}
17	GO:CC	GO:0005654	nucleoplasm	6.259×10^{-3}
8	GO:BP	GO:0019752	carboxylic acid metabolic process	6.809×10^{-3}

(c) *Output from online tool.*

Figure 3.6: **gProfiler outputs for Random Forest most important variables**.

Several studies have shown that several kinds of cell stress, including DNA damage stress, autophagy, and endoplasmic reticulum (ER) stress, can stimulate an increase in the number of extracellular vesicles released, including extracellular exosome, that shape immune responses [18]. Important examples are tumor-derived exosomes (Tex) that module the functions of Natural-Killer cells [19]. The literature supports the role of protein binding in colorectal cancer (CRC), indeed specific oncogenic-related RNA-binding proteins have been identified in CRC[20]. Also, cancer cells exhibit remarkable adaptability to various stress conditions, including hypoxia, nutrient deprivation, DNA damage, and immune responses. This *stress response* is a subject of intense research due to its critical role in cancer progression. By developing mechanisms to survive and adapt, cancer cells can evade growth inhibition and immune surveillance, leading to genomic instability, altered gene expression, and metabolic reprogramming [21].

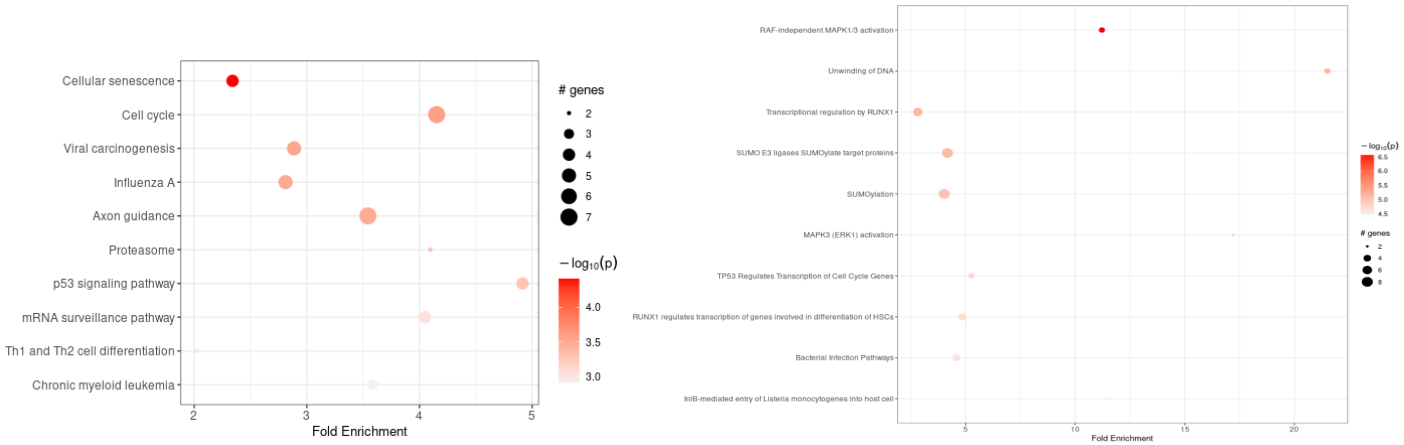
Network-based Analysis

Network-based analysis was conducted using **STRING**, **EnrichNet**, and **pathfindR** as the final step in the workflow. The input genes for these analyses were those identified by the Random Forest model.
STRING shows a network (Fig. 5.6(a)) that presents a PPI enrichment p-value of 2.5×10^{-4} and also has more interactions (521) than expected (445) indicating that the proteins are at least partially biologically connected, as a group. Also, it was tried to increase the *minimum required interaction score* to 0.7, which puts a threshold on the confidence score so only interactions above this score are included in the predicted network, this means that the higher the score the lower false positives are present [22]. The result is present in Figure 5.6(b) where the PPI enrichment p-value is 1.7×10^{-2} and, also in this case, the expected interactions (120) are lower than the ones present (144).

In both networks, it is possible to see terms that concern the *Regulation of cellular response to stress* in the Biological process, *Extracellular exosome* present in the Cellular component and *RNA binding* in the section of Molecular function, confirming the results of the Over-Representation Analysis (the STRING networks of these specific terms are present in the Appendix Figure 5.7).

Figure 3.7 shows the results obtained running **pathfindR** under two different conditions. On the left, by setting *KEGG* as the gene set, while on the right, the results were obtained by setting *Reactome* instead. From Figure 3.7(a) the most significant term is *Cellular senescence*, which is well known to play a fundamental role in cancer development [23]. From

recent research, it was possible to understand the role of such mechanism in the development of Colorectal cancer, highlighting new potential therapeutic targets for CRC treatment [24]. It is also important to notice that, the terms *Cell cycle* and *p53 signaling pathway* could correlate with what was previously found from the Over-Representation Analysis since it is well known that p53 has an impact on the cell cycle and its role is activated in response to stress and environmental changes [25].
From Figure 3.7(b) it is possible to observe that, among the most significant results, there are terms related to the process of *SUMOylation*, which is a process that regulates various aspects of protein function, such as transcription and sub-cellular localization. What is interesting to notice is that Colorectal cancer has shown higher *SUMOylation* levels compared to other cancers [26]. Indeed, an over-activation of this process positively correlates with an increase in cancer cell stemness, driving tumor progression and potentially causing relapses [27]. Another key term highlighted in the enrichment chart is *Unwinding of DNA*, a crucial step in *DNA replication* which can contribute to genome instability due to DNA damage stress, particularly when triggered by Environmental stressors [21]. Any condition that does not permit the normal replication procedure is referred to as *replication stress*, a feature of cancerous cells [28].
Again, it is possible to highlight terms related to the *Regulation of cellular stress response*. Indeed, it is possible to identify terms concerning *MAPK1/3 activation*, a signaling network activated in response to a variety of cellular stresses [29].



(a) *Gene set: KEGG* (b) *Gene set: Reactome*

Figure 3.7: **PathfindR** enrichment charts on important variables extracted from the Random Forest model.

EnrichNet was performed multiple times by utilizing three different annotation databases: *KEGG*, *Reactome*, and *GO (Biological Process)*.
In the case of *KEGG* unfortunately, there were no terms that resulted significantly due to low `XD-score` and high `q-value` (shown in Figure 5.9(e)). Even for *Ractome* and *GO* there aren't any terms that present a significant `q-value` (lower than 0.05), which represents the overlap-based Fisher exact test score, so with a lot of terms not overlapping with the input list of genes (in appendix Figures 5.9(b) and (d), indeed, we can see that a lot of terms in both scatterplots present a `q-value` of 1). However, by looking at the results presented in appendix Figure 5.9 there are terms for *Reactome* and *GO* that have high `XD-score` values: *Unwinding of DNA* and *hindlimb morphogenesis*.

Unwinding of DNA is a term that was already found in the section of **pathfindR**, so this highlights again terms that relate with *replication process* and *stress response*. Also, the genes from **EnrichNet** results that overlap this term are *MCM6*, *MCM7* and *CDC45*. All these three genes are known to be present in Colorectal cancer. Different studies have used these genes as possible cancer biomarkers, targets for therapies and their expression as potential prognostic genes [30–32].

The term *hindlimb morphogenesis* is not very informative in this project per se, however the three genes *PITX1*, *Twist1* and *PTCH1*, which take part in the regulation of hindlimb and pituitary development, are important for the characterization of CRC. Indeed *PITX1* besides being a key regulator in animal growth and development, is also known to interact with p53 and module crucial cellular processes among which cell cycle progression, apoptosis, and chemotherapy resistance [33]. In literature has been shown that *Twist1* is important in the carcinogenesis of many tumors including Colorectal cancer, indeed its positive expression is cell cycle progression in gastric cancer cells with a lower survival rate in patients with CRC and it's also shown that the knockdown of *Twist1* in CRC cell lines leads to a decrease of the proliferation of cancer cells and also increase the numbers of apoptotic cells [34]). *Patched1* (PTCH1) is a gene that is frequently altered in CRC and its mutations contribute to unregulated Hedgehog (Hh) signaling, which is critical for the formation of the brain, the distal limbs, and other organ systems but its mutation results also in uncontrolled cellular growth and tumor development [35].

Chapter 4

Discussion

The goal of this analysis is to determine a set of variables that can more accurately discriminate between Colorectal cancer patients and control subjects. The original dataset, formed by two sub-datasets, contained 60 samples of which 30 were tumoral and 30 controls.
The best-performing method, Random Forest, has been used to obtain a list of the most important variables, resulting in 277 genes.

The functional enrichment analysis results were highly concordant within the several tools used, with some small divergences from **Enrichnet** results. Overall it is possible to categorize the genes extracted from Random Forest as dealing with three main terms, specifically, *stress response*, *extracellular exosome* and *replication process*.
Indeed, it is possible to understand from the literature that all these terms are highly connected with Colorectal cancer and a connection between the three main terms can be made. Specifically, it is well-known that tumor progression requires bidirectional communication between tumor cells and the tumor microenvironment (TME).

In this sense, *extracellular vesicles* (EVs) often act as shuttles between tumor cells and TME, thanks to the release of regulatory molecules needed to alter the microenvironment and enable tumor progression. A recent publication highlights the role of a stressed microenvironment in the progression of tumors to a more malignant state. In response to stimuli generated from the stressed microenvironment, extracellular vesicles are secreted, containing specific immunoregulatory mediators that provoke changes in the immune status of the tumor microenvironment [18].
At the same time, the tumor microenvironment (TME) has been established as a critical factor in tumorigenesis and metastasis in colorectal cancer [36]. Therefore, it is possible to raise a hypothesis for which a stressed microenvironment could cause an increase in the production of exosomes able to enhance metastasis progression in CRC.
Regarding the connection of the term *replication process* with the previous terms, it is possible to find scientific articles explaining the effect of stress in the activation of the replication pathway guided by *P53* [37], where an aberrant replication is a fundamental hallmark of cancer. It is easy to highlight the role of stress as a trigger to a cascade of pathways connected to an enhanced generation of extracellular vesicles, causing the activation of an aberrant replication and proliferation that could culminate in the generation of metastasis.

From the original work of *Efstathios-Iason Vlachavas et al.* [3] an initial gene signature of 94 differential expressed genes was used to extract 12 features. Interestingly, the selected genes are highly connected with terms dealing with biological processes such as *cellular response to stimulus*, *cell cycle regulation*, *chemotaxis* and *DNA biosynthesis*. Terms that can be easily compared to what found in this analysis, as with the term *response to stimuli* a wide range of factors are covered, including stress stimuli. The same can be said with the term *chemotaxis*, as it has been shown that cancer cell chemotaxis relies mainly on the secretion of exosome-type extracellular vesicles [38], and also for *DNA biosynthesis* which is correlated with the *replication process*.

Overall the findings of this analysis enhance the importance of *stress response* in Tumor cells and the connection of various cancerous characteristics due to the *replication process* and to the *tumor microenvironment*, especially in the case of Colorectal cancer, and by looking at the latest papers support the route that the therapeutic research is focusing in the latest years to contrast CRC.

Bibliography

1. MS, H. Colorectal Cancer: A Review of Carcinogenesis, Global Epidemiology, Current Challenges, Risk Factors, Preventive and Treatment Strategies. *Cancers (Basel)* (2022).
2. X, W. Prognoses of different pathological subtypes of colorectal cancer at different stages: A population-based retrospective cohort study. *BMC Gastroenterol* (2019).
3. EI, V. Radiogenomic Analysis of F-18-Fluorodeoxyglucose Positron Emission Tomography and Gene Expression Data Elucidates the Epidemiological Complexity of Colorectal Cancer Landscape. *Comput Struct Biotechnol J* (2019).
4. M, G. Principal component analysis. *Nat Rev Methods Primers* (2022).
5. M., A. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences* (2023).
6. Ran, X. Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review* (2022).
7. Sotiris, K. Supervised Machine Learning: A Review of Classification Techniques. *Informatica (Slovenia)* (2007).
8. Ali, J. Random Forest and Decision Trees. *International Journal of Computer Science Issues (IJCSI)* (2012).
9. Tharwat, A. Linear discriminant analysis: A detailed tutorial. *Ai Communications* (2017).
10. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** (1996).
11. Hastie, T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer Series in Statistics* (2009).
12. Gareth, J. An introduction to statistical learning with applications in R. *Springer* (2022).
13. Melkumova, L. Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering* (2017).
14. Lauria, M. SCUDO: a tool for signature-based clustering of expression profiles. *Nucleic Acids Research* (2015).
15. Subramanian, A. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* (2005).
16. Johnson, W. E. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* (2017).
17. Dedrick, K. H. C. Tumour heterogeneity and evolutionary dynamics in colorectal cancer. *Oncogenesis* (2021).
18. Q., W. Extracellular vesicles and immunogenic stress in cancer. *Cell Death Dis* (2021).
19. E., V. Cancer Exosomes as Conveyors of Stress-Induced Molecules: New Players in the Modulation of NK Cell Response. *Int J Mol Sci* (2019).
20. JM., G.-C. Post-transcriptional Regulation of Colorectal Cancer: A Focus on RNA-Binding Proteins. *Front Mol Biosci* (2019).
21. M, C. Therapeutic targeting of cellular stress responses in cancer. *Thorac Cancer* (2018).
22. Bork, P. *STRING* `http://version10.string-db.org/help/getting_started/`.
23. S., C. A. Senescence and cancer — role and therapeutic opportunities. *Nature reviews* (2022).
24. C., Z. METTL3 promotes cellular senescence of colorectal cancer via modulation of CDKN2B transcription and mRNA stability. *Oncogene* (2024).
25. L., F. The p53 endoplasmic reticulum stress-response pathway evolved in humans but not in mice via PERK-regulated p53 mRNA structures. *Cell Death and Differentiation* (2023).
26. Q., Z. Protein sumoylation in normal and cancer stem cells. *Front. Mol. Biosci.* (2022).
27. L., W. Cancer Stem Cells—Origins and Biomarkers: Perspectives for Targeted Personalized Therapies. *Front. Immunol.* (2020).

28. H., G. Replication stress and cancer. *Nat Rev Cancer* (2015).

29. J., D. N. The role of MAPK signalling pathways in the response to endoplasmic reticulum stress. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* (2014).

30. Y, H. Potential Prognostic and Diagnostic Values of CDC6, CDC45, ORC6 and SNHG7 in Colorectal Cancer. *Onco Targets Ther* (2019).

31. Zeng, T. The DNA replication regulator MCM6: An emerging cancer biomarker and target. *Clinica Chimica Acta* (2021).

32. X, L. PRMT5 promotes colorectal cancer growth by interaction with MCM7. *Journal of Cellular and Molecular Medicine* (2021).

33. J, Z. PITX1 plays essential functions in cancer. *Front Oncol* (2023).

34. Zhu, D.-J. Twist1 is a potential prognostic marker for colorectal cancer and associated with chemoresistance. *American Journal of Cancer Research* (2015).

35. David Rimoin Reed Pyeritz, B. K. *Emery and Rimoin's Principles and Practice of Medical Genetics* chap. Hedgehog Signaling Pathway (Academic Press, 2013).

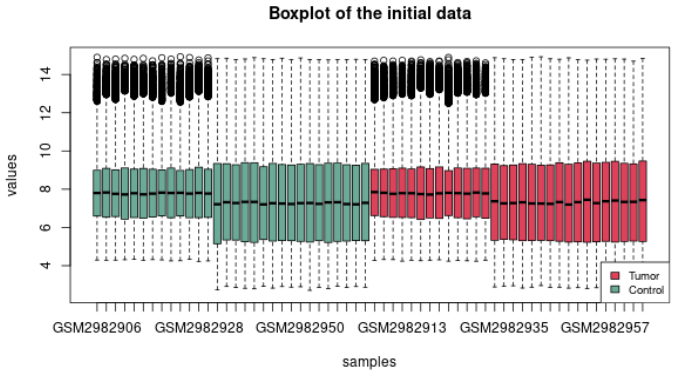
36. Z., W. Exosomes in metastasis of colorectal cancers: Friends or foes? *World J Gastrointest Oncol* (2023).

37. Lindström, M. p53 at the crossroad of DNA replication and ribosome biogenesis stress pathways. *Cell Death Differ* (2022).

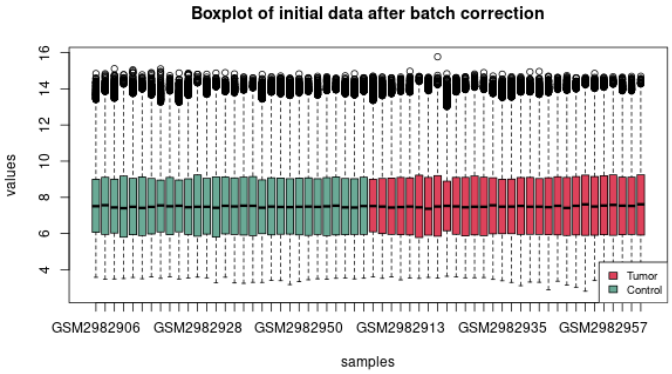
38. Sung BH, W. A. Exosome secretion promotes chemotaxis of cancer cells. *Cell Adh Migr.* (2017).

Chapter 5

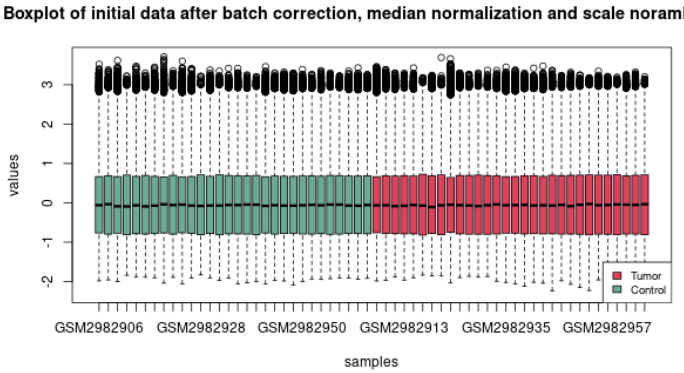
Appendix



(a) *Boxplots data before any type of normalization*



(b) *Boxplots data after batch effect correction*



(c) *Boxplot data after pre-processing*

Figure 5.1: Comparison between boxplots to see how the data changes during the pre-processing part.

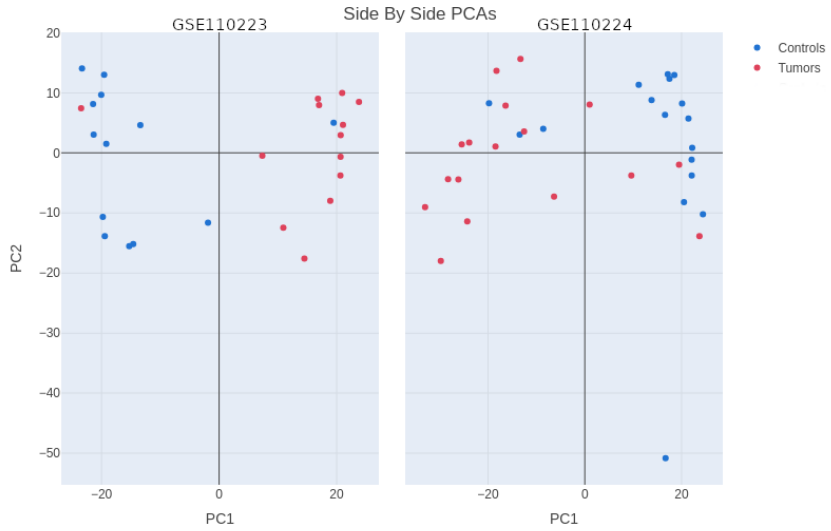


Figure 5.2: Comparison of the PCAs of the two distinct datasets.

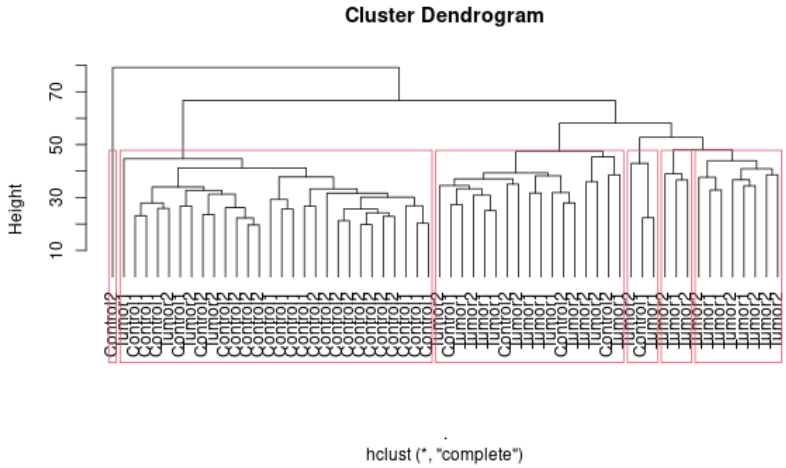


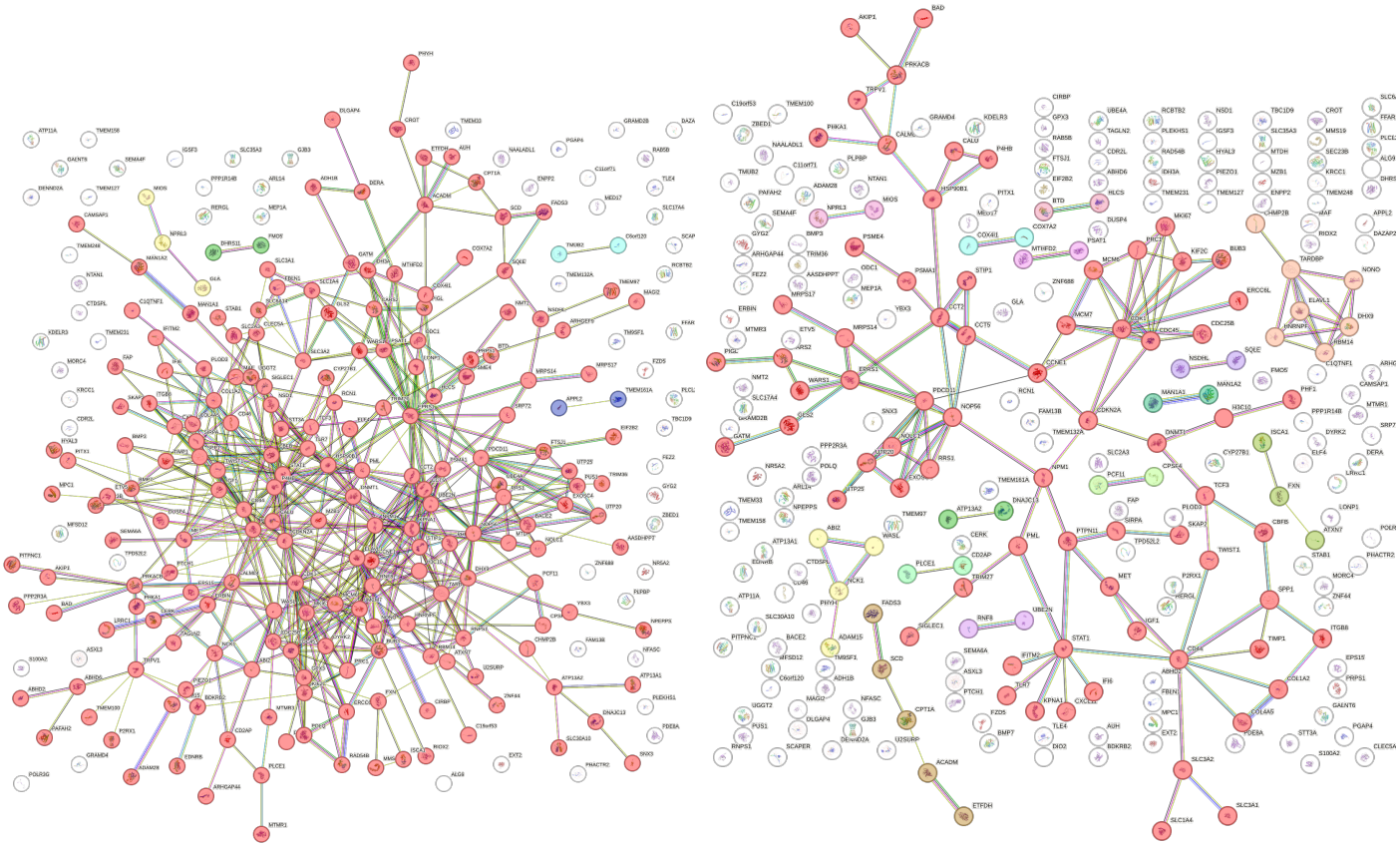
Figure 5.3: **Hierarchical clustering**. Setting $k = 6$, with information about data origin.

(a) *Training network obtained from SCUDO* (b) *Test network obtained from SCUDO*

Figure 5.4: **SCUDO**. Training and test networks that were obtained for the SCUDO model.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_MF_DIRECT	protein binding	RT		205	77.4	1.0E-6	5.9E-4
	GOTERM_CC_DIRECT	membrane	RT		94	35.5	9.7E-6	3.4E-3
	GOTERM_CC_DIRECT	endoplasmic reticulum	RT		32	12.1	6.1E-5	1.1E-2
	GOTERM_CC_DIRECT	extracellular exosome	RT		48	18.1	2.2E-4	2.1E-2
	GOTERM_CC_DIRECT	nucleoplasm	RT		74	27.9	2.5E-4	2.1E-2
	GOTERM_CC_DIRECT	melanosome	RT		8	3.0	3.0E-4	2.1E-2
	GOTERM_CC_DIRECT	mitochondrial matrix	RT		15	5.7	7.7E-4	4.5E-2
	GOTERM_CC_DIRECT	endoplasmic reticulum lumen	RT		12	4.5	1.7E-3	8.3E-2
	GOTERM_CC_DIRECT	cytosol	RT		92	34.7	1.9E-3	8.3E-2
	GOTERM_CC_DIRECT	endoplasmic reticulum membrane	RT		27	10.2	2.7E-3	1.1E-1
	GOTERM_CC_DIRECT	neuromuscular junction	RT		6	2.3	5.6E-3	1.9E-1
	GOTERM_CC_DIRECT	cytoplasm	RT		90	34.0	7.3E-3	1.9E-1
	GOTERM_CC_DIRECT	postsynaptic specialization	RT		3	1.1	7.7E-3	1.9E-1
	GOTERM_CC_DIRECT	endoplasmic reticulum chaperone complex	RT		3	1.1	7.7E-3	1.9E-1
	GOTERM_CC_DIRECT	CMG complex	RT		3	1.1	7.7E-3	1.9E-1
	GOTERM_CC_DIRECT	early phagosome	RT		3	1.1	9.2E-3	2.1E-1
	GOTERM_CC_DIRECT	protein-containing complex	RT		17	6.4	1.1E-2	2.4E-1
	GOTERM_CC_DIRECT	early endosome membrane	RT		8	3.0	1.4E-2	3.0E-1
	GOTERM_CC_DIRECT	mitochondrial membrane	RT		7	2.6	1.6E-2	3.0E-1
	GOTERM_CC_DIRECT	mitochondrion	RT		29	10.9	1.6E-2	3.0E-1
	GOTERM_BP_DIRECT	cartilage development	RT		7	2.6	1.7E-4	3.0E-1
	GOTERM_CC_DIRECT	glutamatergic synapse	RT		12	4.5	1.8E-2	3.2E-1
	GOTERM_CC_DIRECT	RNA polymerase II transcription regulator complex	RT		6	2.3	2.1E-2	3.4E-1
	GOTERM_CC_DIRECT	acrosomal vesicle	RT		6	2.3	2.4E-2	3.6E-1
	GOTERM_CC_DIRECT	vesicle	RT		7	2.6	2.4E-2	3.6E-1
	GOTERM_CC_DIRECT	lysosomal membrane	RT		11	4.2	2.8E-2	4.1E-1
	GOTERM_CC_DIRECT	lamellipodium membrane	RT		3	1.1	3.2E-2	4.5E-1
	GOTERM_CC_DIRECT	collagen trimer	RT		5	1.9	3.5E-2	4.7E-1
	GOTERM_CC_DIRECT	secretory granule membrane	RT		5	1.9	3.7E-2	4.8E-1
	GOTERM_CC_DIRECT	centrosome	RT		15	5.7	3.9E-2	4.8E-1
	GOTERM_CC_DIRECT	fibrillar center	RT		6	2.3	4.0E-2	4.8E-1

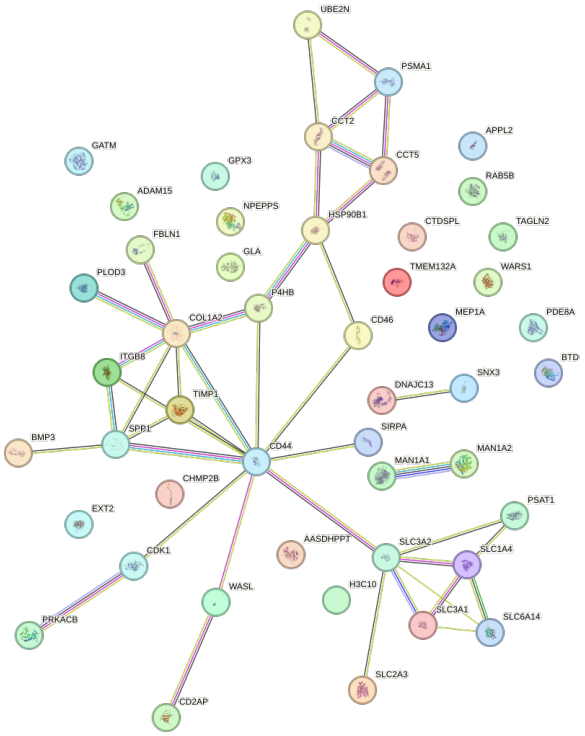
Figure 5.5: **David** chart for **Random Forest** most important variables.



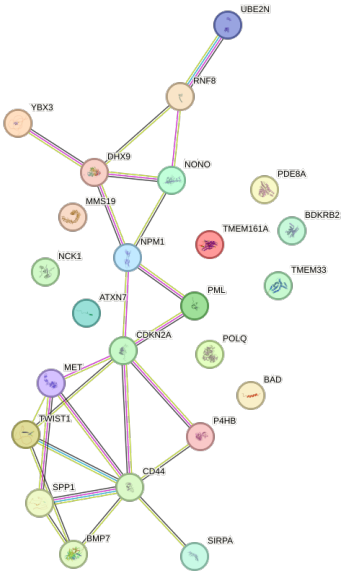
(a) ***STRING*** *network: minimum required interaction score 0.4*

(b) ***STRING*** *network: minimum required interaction score 0.7*

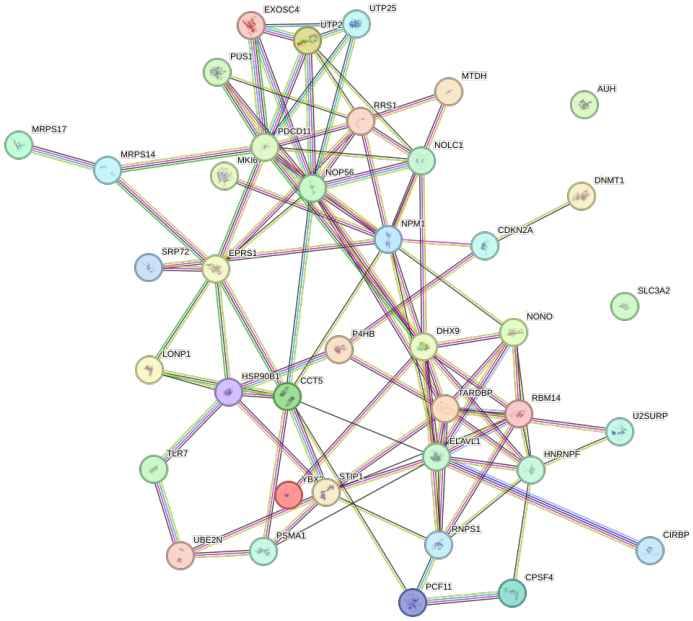
Figure 5.6: In figure (a) the cluster most populated is indicated in red with 195 genes compared to the total number of genes which is 263. There are four other clusters with a couple of genes each.
In Figure (b) also the most populated cluster is in red with 71 genes and other 14 clusters that present a gene count between 2 and 7 each.



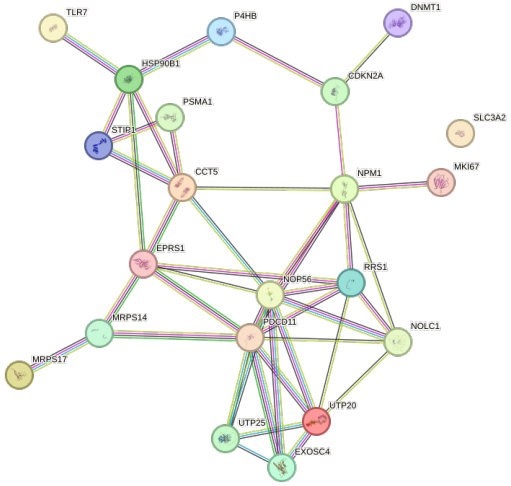
(a) ***STRING*** *sub-network of the term* ***Extracellular exosome*** *present in both networks with minimum interaction score at 0.4 and 0.7*



(b) ***STRING*** *sub-network of the term* ***Regulation of cellular response to stress*** *present in both networks with minimum interaction score at 0.4 and 0.7*

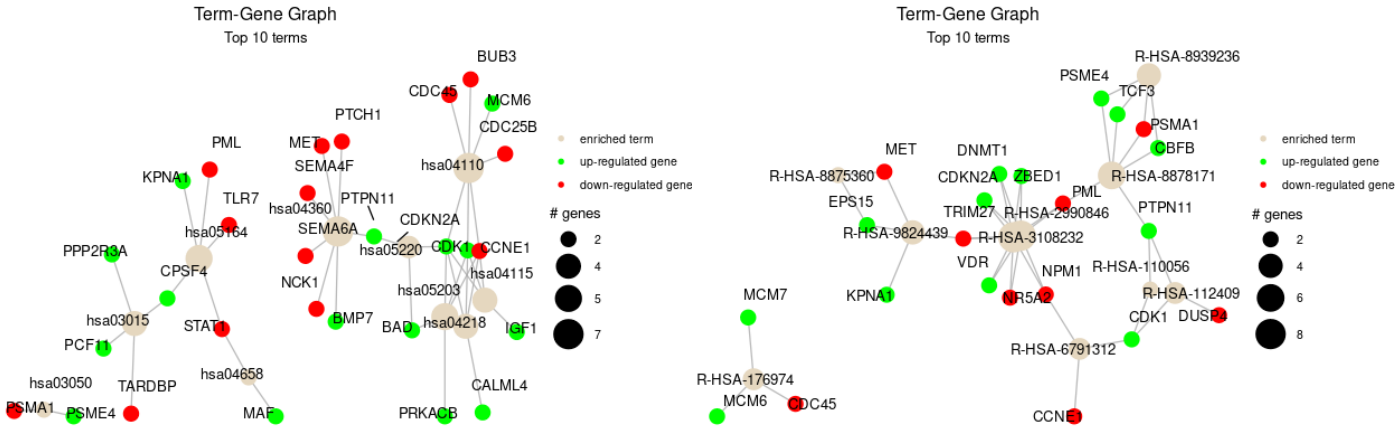


(c) ***STRING*** *sub-network of the term* ***RNA Binding*** *present in the network with minimum interaction score 0.4*



(d) ***STRING*** *sub-network of the term* ***RNA Binding*** *present in the network with minimum interaction score 0.7*

Figure 5.7: Networks specific of the concerning terms



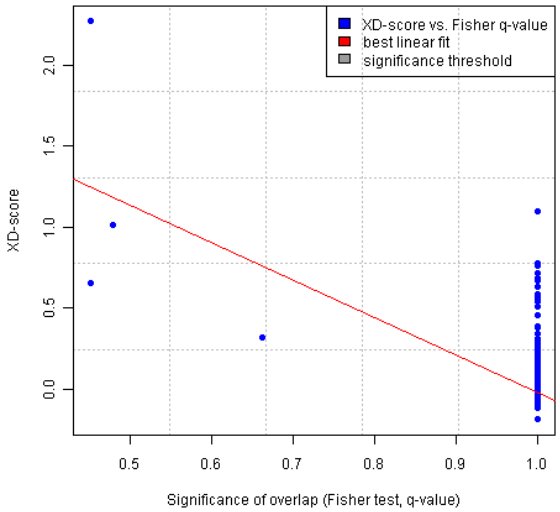
(a) *Gene set:* ***KEGG***

(b) *Gene set:* ***Reactome***

Figure 5.8: **PathfindR** networks of interactions.

Annotation (pathway/process)	Significance of network distance distribution (XD-Score)	Significance of overlap (Fisher-test, q-value)	Dataset size (uploaded gene set)	Dataset size (pathway gene set)	Dataset size (overlap)	Tissue-specific XD-scores
UNWINDING OF DNA compute graph visualization see mapped genes	2.2681	0.45	229	11	3 (show)	show tissue specificity
GLYCOGEN BREAKDOWN GLYCOGENOLYSIS compute graph visualization see mapped genes	1.0993	1.00	229	14	2 (show)	show tissue specificity
AMINO ACID TRANSPORT ACROSS THE PLASMA MEMBRANE compute graph visualization see mapped genes	1.0135	0.48	229	30	4 (show)	show tissue specificity
ASSOCIATION OF TRIC CCT WITH TARGET PROTEINS DURING BIOSYNTHESIS compute graph visualization see mapped genes	0.7778	1.00	229	28	3 (show)	show tissue specificity

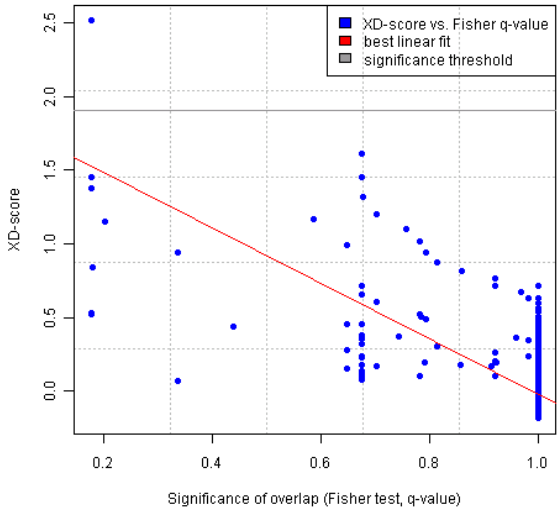
(a) *Table of terms with database annotation* ***Reactome***



(b) *Scatterplot of terms with database annotation* ***Reactome***

Annotation (pathway/process)	Significance of network distance distribution (XD-Score)	Significance of overlap (Fisher-test, q-value)	Dataset size (uploaded gene set)	Dataset size (pathway gene set)	Dataset size (overlap)	Tissue-specific XD-scores
hindlimb morphogenesis compute graph visualization see mapped genes	2.516*	0.18	229	10	3 (show)	show tissue specificity
protein heterotrimerization compute graph visualization see mapped genes	1.616	0.67	229	10	2 (show)	show tissue specificity
negative regulation of multicellular organism growth compute graph visualization see mapped genes	1.616	0.67	229	10	2 (show)	show tissue specificity
iron-sulfur cluster assembly compute graph visualization see mapped genes	1.616	0.67	229	10	2 (show)	show tissue specificity

(c) *Table of terms with database annotation* ***GO(Biological Process)***



(d) *Scatterplot of terms with database annotation* ***GO(Biological Process)***

Annotation (pathway/process)	Significance of network distance distribution (XD-Score)	Significance of overlap (Fisher-test, q-value)	Dataset size (uploaded gene set)	Dataset size (pathway gene set)	Dataset size (overlap)	Tissue-specific XD-scores
Steroid biosynthesis compute graph visualization see mapped genes	0.87054	1	229	17	2 (show)	show tissue specificity
Pentose phosphate pathway compute graph visualization see mapped genes	0.50403	1	229	26	2 (show)	show tissue specificity
Fatty acid metabolism compute graph visualization see mapped genes	0.47025	1	229	41	3 (show)	show tissue specificity
Glycosphingolipid biosynthesis - globo series compute graph visualization see mapped genes	0.45458	1	229	14	1 (show)	show tissue specificity

(e) *Table of terms with database annotation* ***KEGG***

Figure 5.9: **EnrichNet**.Results from utilizing different database annotations. The scatterplots present as x-axes: q-value that represent the Significance Overlap, y-axis: XD-score.