Analysis of an Esophageal cancer dataset

Exam: Network-based Data Analysis

Gloria Lugoboni

13 June 2024

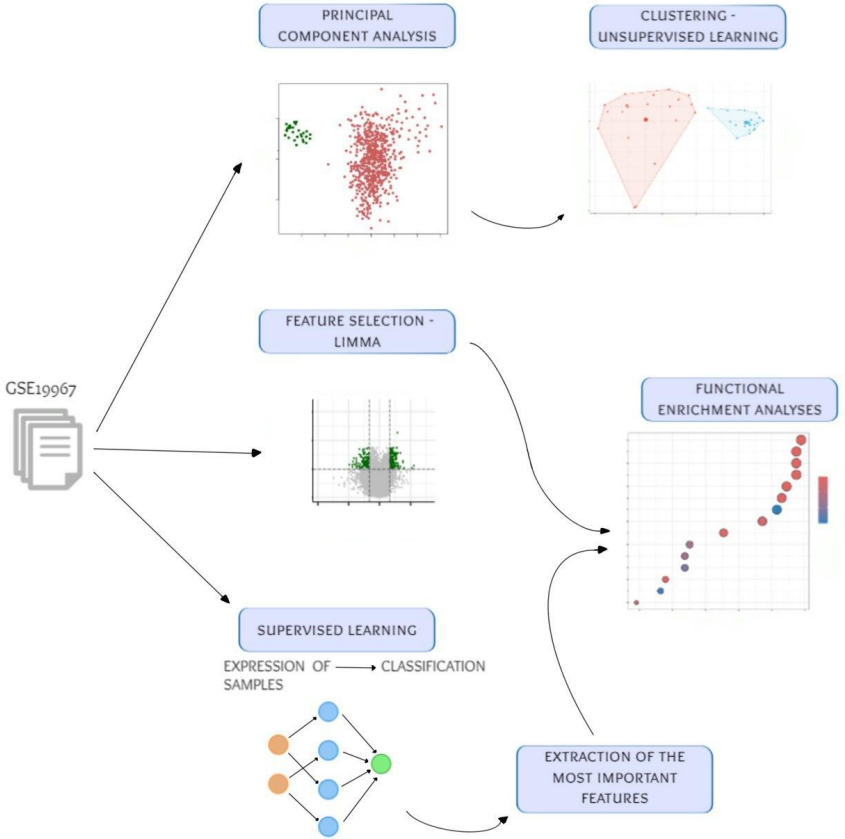
Selected dataset and motivation

Series GSE199967	Query DataSets for GSE199967
Status	Public on Apr 03, 2022
Title	Tumor tissues vs normal tissues from 21 cases ESCA
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Transcriptional profiling of tumor tissues and normal tissues from ESCA.
Overall design	Tumor tissues vs normal tissues. Biological replicates: 21.
Platforms (1)	GPL6480 Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version)

- Esophageal cancer (ESCA) is among the 10 highest mortality cancers worldwide
- 5-year overall survival (OS) of circa 20%
- High probability of developing metastasis and resistance to chemo-radiotherapy

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).

Methods



More in detail:

- Clustering:
  - K-means clustering
  - Hierarchical Clustering
- Supervised Learning methods
  - Random Forest
  - LDA
  - LASSO and Ridge
  - SCUDO
- Functional enrichment analyses
  - Over-Representation Analysis
    - gProfiler
    - DAVID
  - Network-based Analysis
    - pathfindR
    - enrichNet
    - STRING

Results - PCA

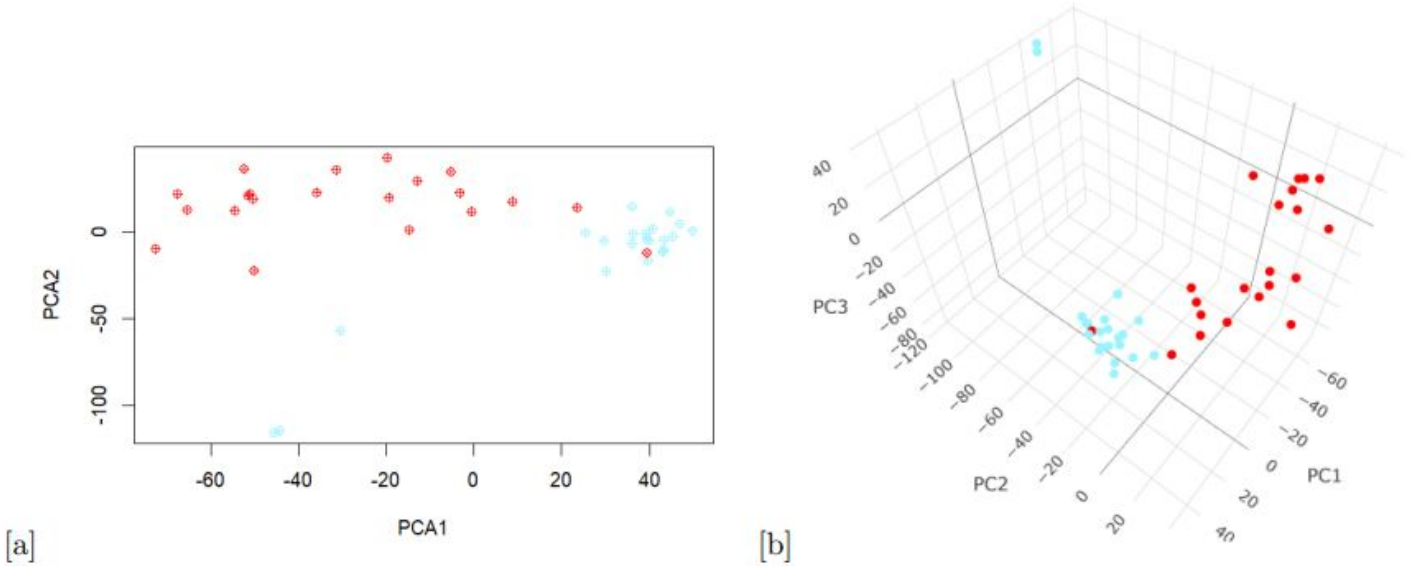


Figure 1: **Principal Component Analysis**. (a) PCA plot in two dimensions, x-axes: PC1, y-axes: PC2, (b) PCA plot in three dimensions, x-axes: PC1, y-axes: PC2, z-axis: PC3. Controls are shown in light blue while tumors are shown in red.

Results - Unsupervised clustering

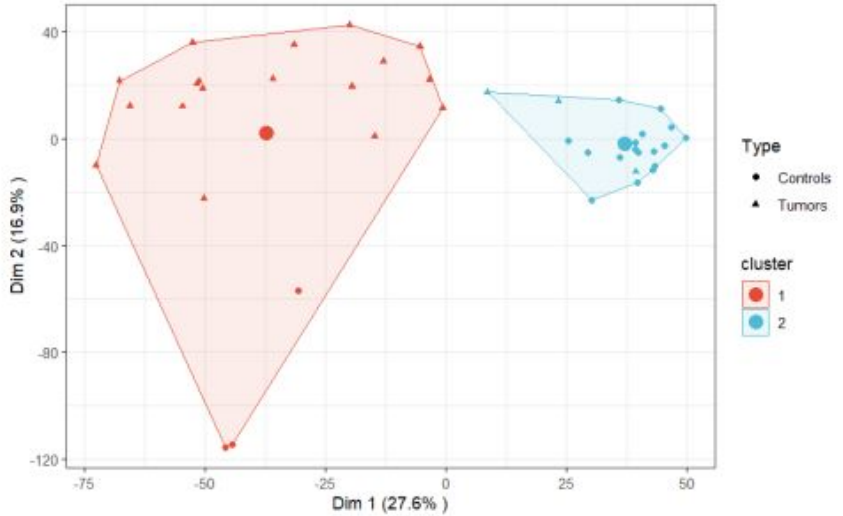


Figure 2: **Cluster analysis**. K-mean clustering, axes: Coordinates for the variables extracted from the first and the second PC

Results - Unsupervised clustering

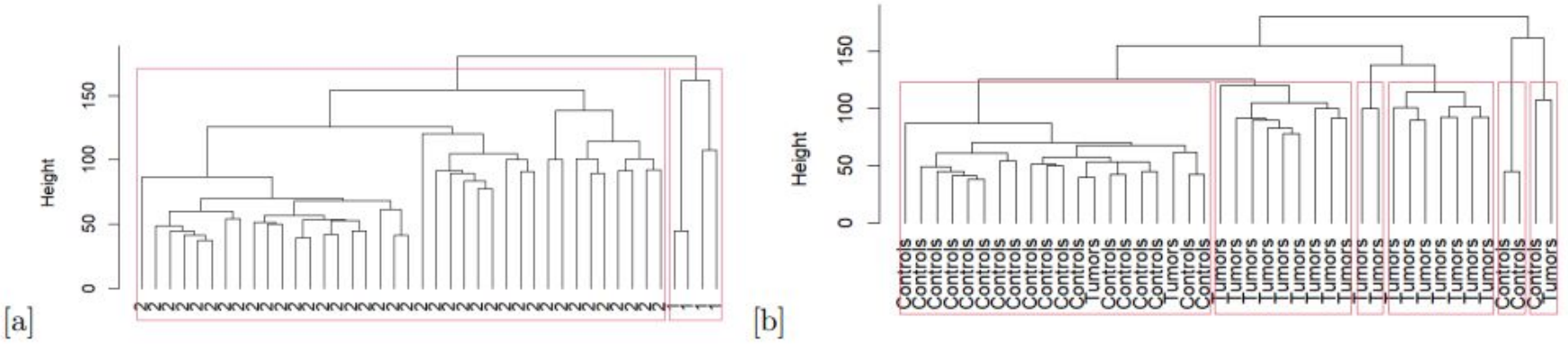
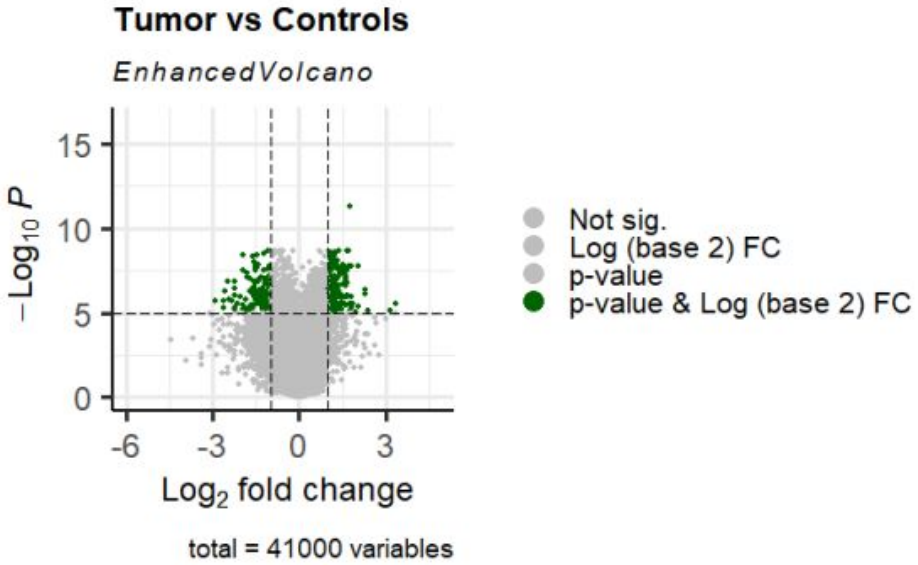


Figure 3: **Cluster analysis**. (a) Hierarchical clustering, x-axes: Identified clusters, y-axes: Height. k=2, (b) Hierarchical clustering, x-axes: Sample type (Tumor or Control), y-axes: Height. k=6.

Results - Feature selection with LIMMA



265 genes DE
79 up-reg
186 down-reg

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).

Results - Supervised learning methods

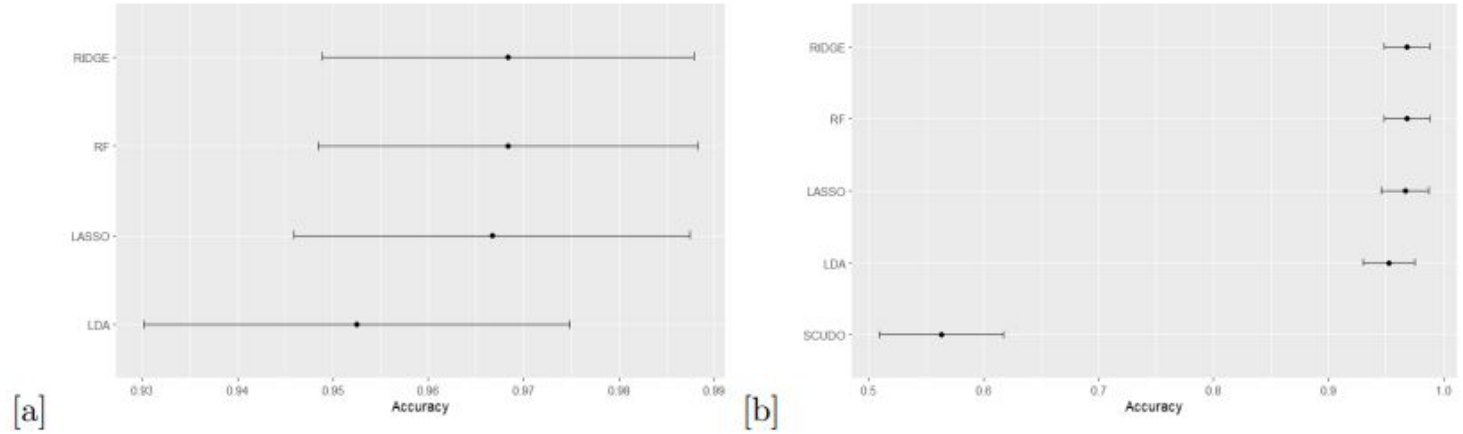


Figure 5: **Performance plot of the supervised models.**(a) All models except SCUDO. x-axis: accuracy, y-axis: model name. (b) All models. x-axis: accuracy, y-axis: model name.

Among all, random forest showed the highest accuracy, equal to 0.9683333, followed directly by Ridge with an accuracy of 0.9658333.

Results - Supervised learning methods

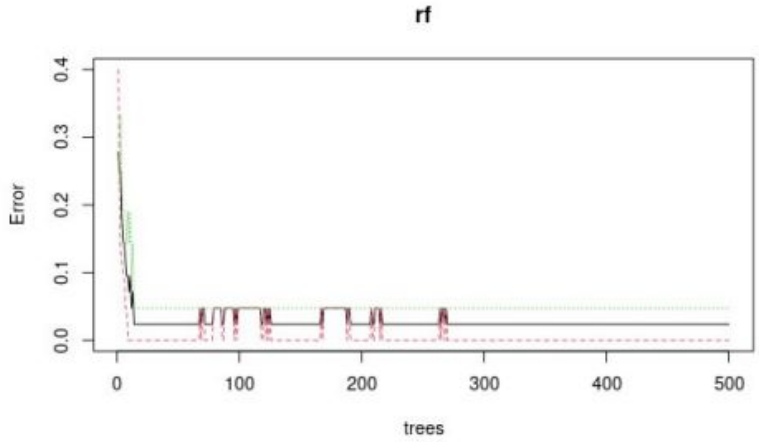


Figure 6: **Distribution of the out-of-bag (OOB) score over the total number of trees used to train the RF model.**

OOB score: number of wrongly classificated observations. The lower, the more accurate the model is.

Results - Supervised learning methods

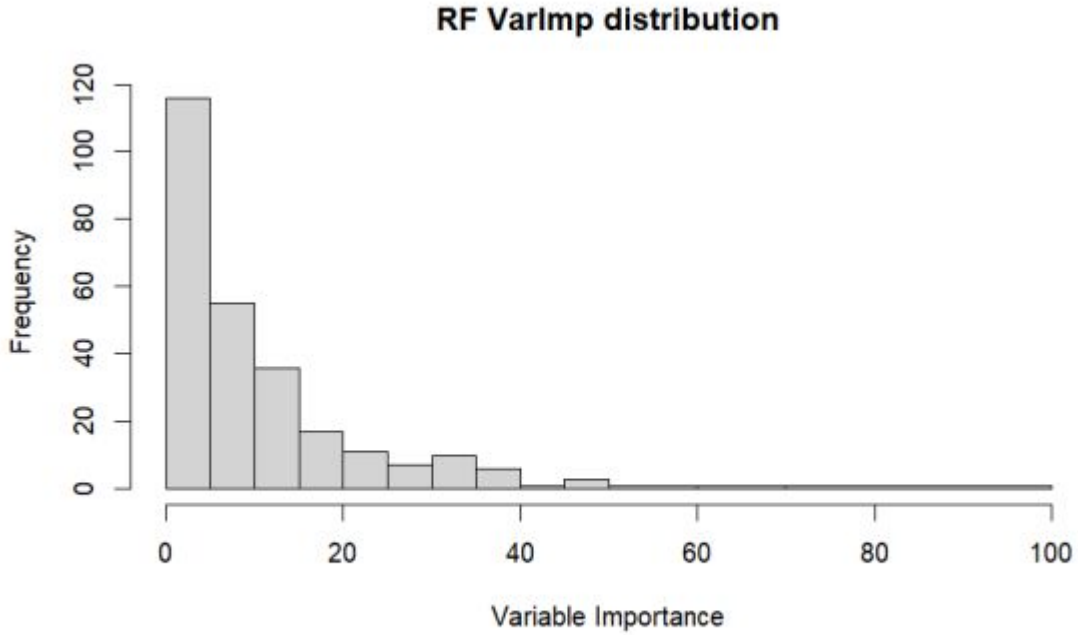
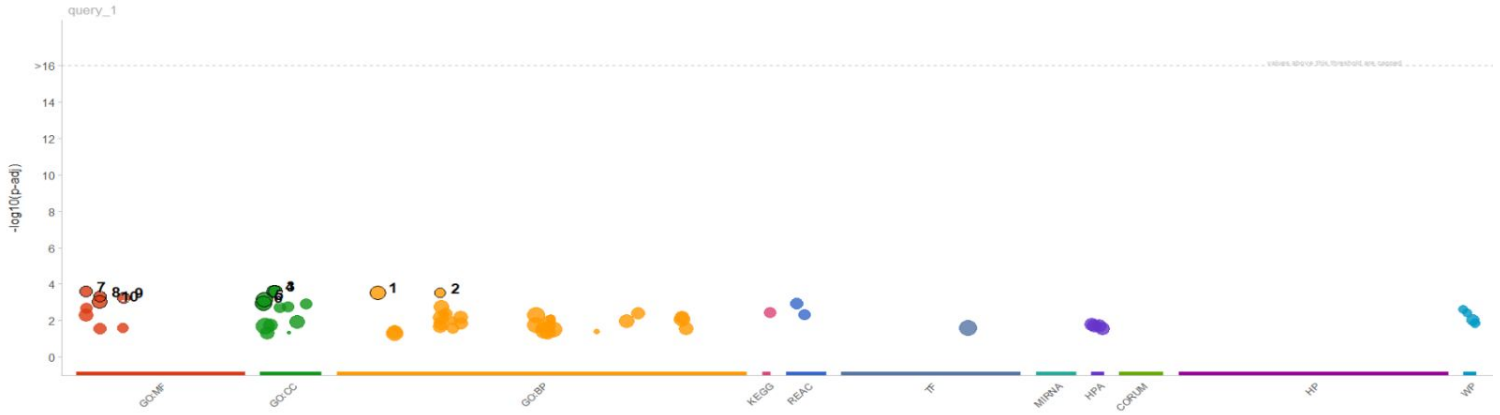


Figure 7: **Variable importance distribution in Random Forest.**

Results - Functional Enrichment Analyses

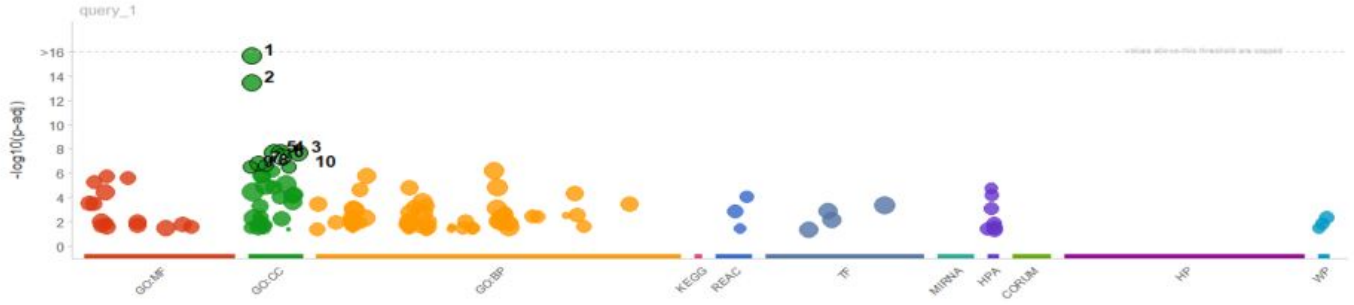


id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0007155	cell adhesion	1512	3.1e-04
2	GO:BP	GO:0022617	extracellular matrix disassembly	61	3.2e-04
3	GO:CC	GO:0031012	extracellular matrix	555	2.5e-04
4	GO:CC	GO:0030312	external encapsulating structure	556	2.5e-04
5	GO:CC	GO:0005615	extracellular 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

Extracellular matrix organization, peptidase activity and proteases activity

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

Results - Functional Enrichment Analyses



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)

Extracellular matrix and peptidase activity

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

Results - Functional Enrichment Analyses

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**

Extracellular matrix organization, peptidase activity and proteases activity

Results - Functional Enrichment Analyses

☐	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

Extracellular matrix and peptidase activity. Palmoplantar keratoderma

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

Results - Functional Enrichment Analyses

Statistic and functional enrichment retrieved from STRING networks.

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

Results - Functional Enrichment Analyses

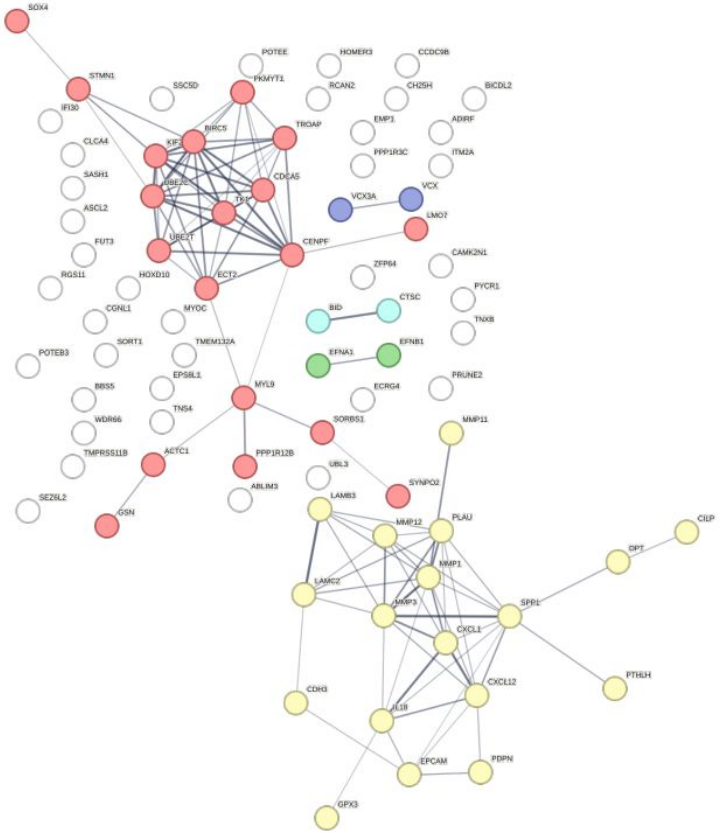


Figure 9: **Clusters obtained from RF STRING-network**. In red, cluster 1, with a total number of gene equal to 19 (genes regulating replication). In yellow, is cluster 2, with a total number of genes equal to 18 (gene regulating extracellular matrix degradation). Overall number of nodes: 94

Results - Functional Enrichment Analyses

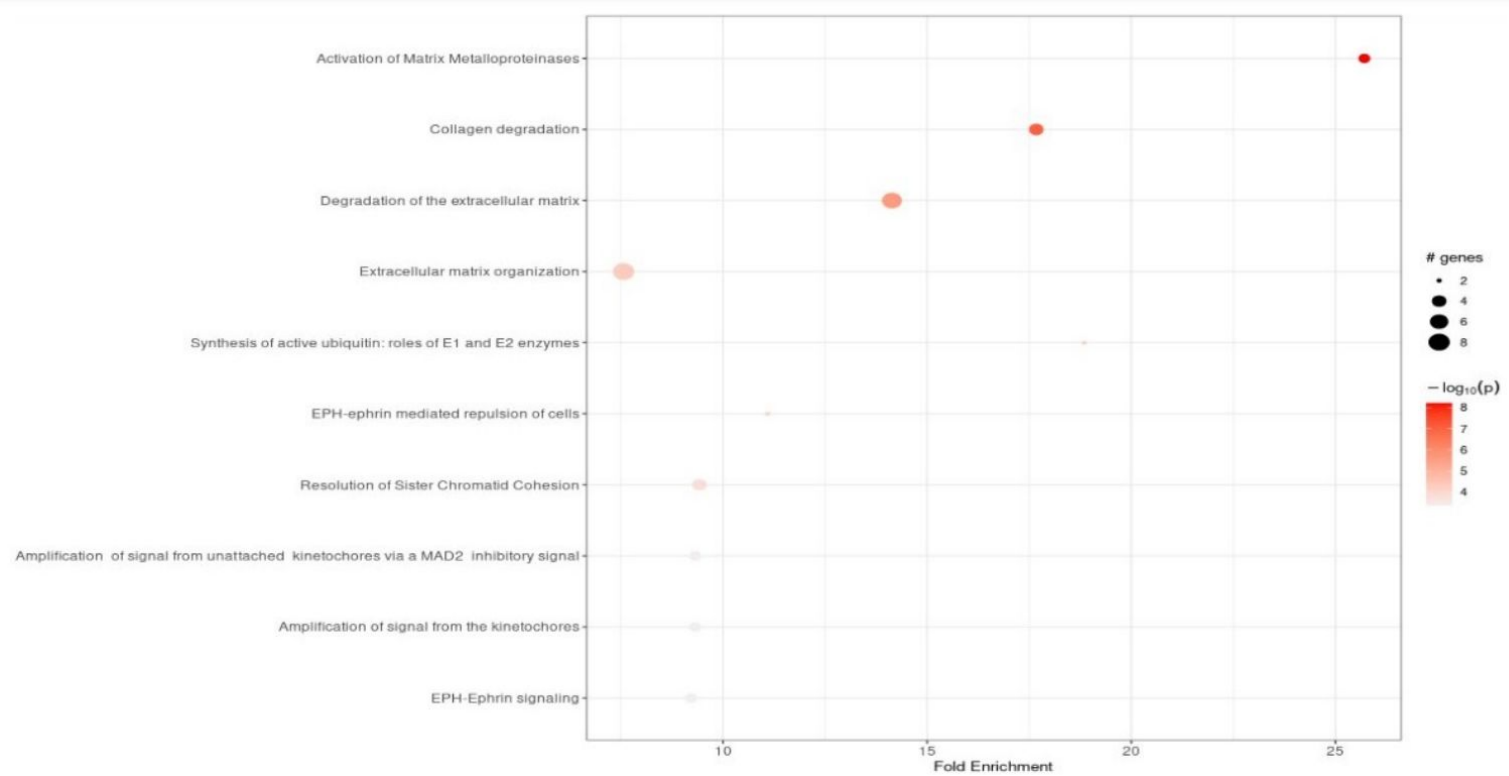


Figure 10: **PathFindR enrichment chart on important variables extracted from the Random Forest model.** Result obtained selecting Reactome as geneset.

Discussion

- This analysis aims to find the set of variables that better distinguish esophageal cancer patients respect control ones.
- It was possible to observe that for certain models, the most important variables represent genes that are differentially expressed.
- The functional enrichment analysis highlighted terms and pathways related to matrix degradation and DNA replication.
- It has been shown in several studies that an enhanced activity of matrix metalloproteinases (MMPs) plays an important role in esophageal carcinogenesis.
- Aberrant cell proliferation is a well-known hallmark of cancer.
- 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.

Thank you for the attention!