Analysis of Colorectal cancer from Microarray data

Network Based Data Analysis project

Andrea Tonina

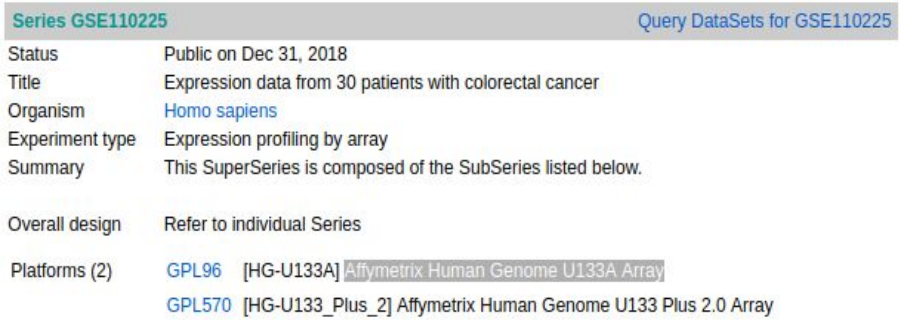
3 September 2024

Introduction

- ➔ Colorectal cancer (CRC) is the third most commonly diagnosed cancer and second leading cause of cancer-related deaths
- ➔ In 2020, approximately 9.4% of cancer-related deaths were attributed to CRC
- ➔ CRC present a significant heterogeneity, adenocarcinoma is the most prevalent colorectal cancer
- ➔ Three primary histological subtypes: adenocarcinoma (AC), mucinous adenocarcinoma (MAC) and signet ring cell carcinoma (SRCC)

X, W. Prognoses of different pathological subtypes of colorectal cancer at different stages: A population-based retrospective cohort study. BMC Gastroenterol (2019).

Dataset

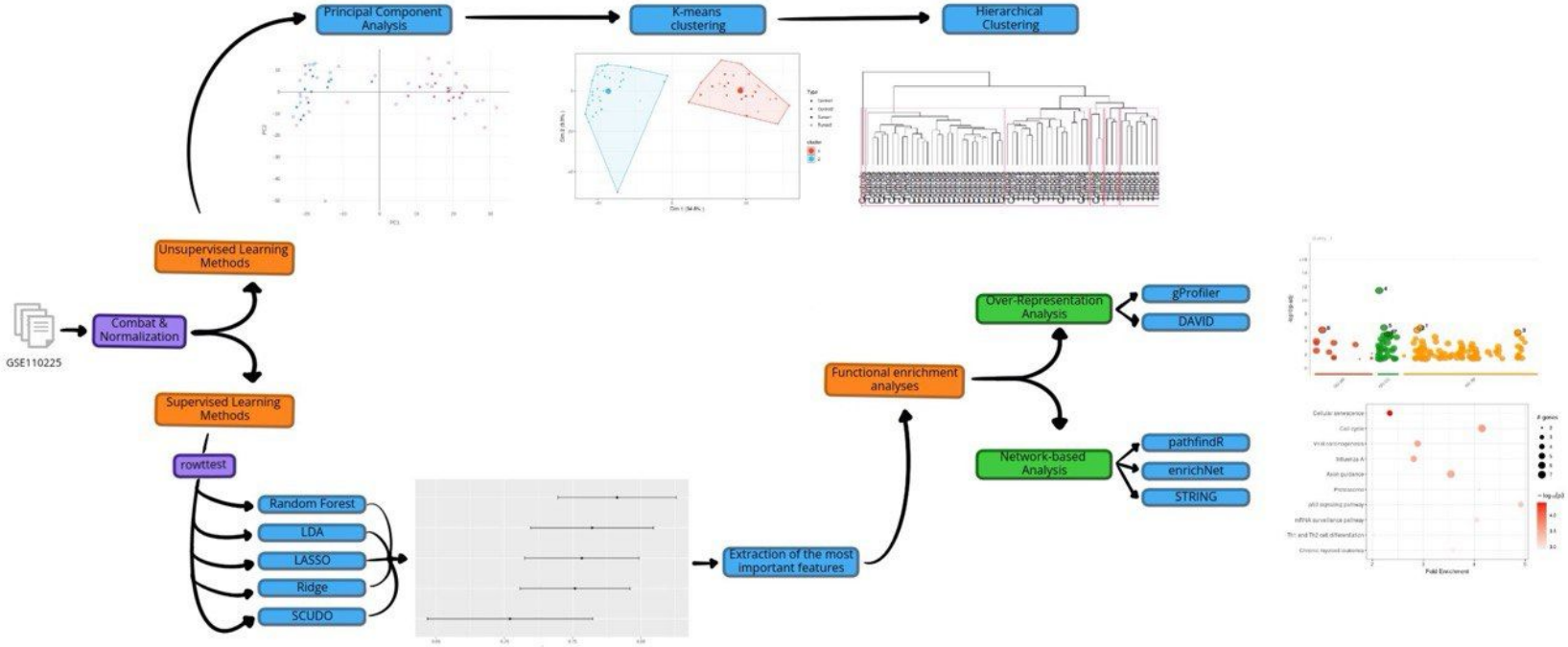


Series GSE110225	Query DataSets for GSE110225
Status	Public on Dec 31, 2018
Title	Expression data from 30 patients with colorectal cancer
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	This SuperSeries is composed of the SubSeries listed below.
Overall design	Refer to individual Series
Platforms (2)	GPL96 [HG-U133A] Affymetrix Human Genome U133A Array
	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

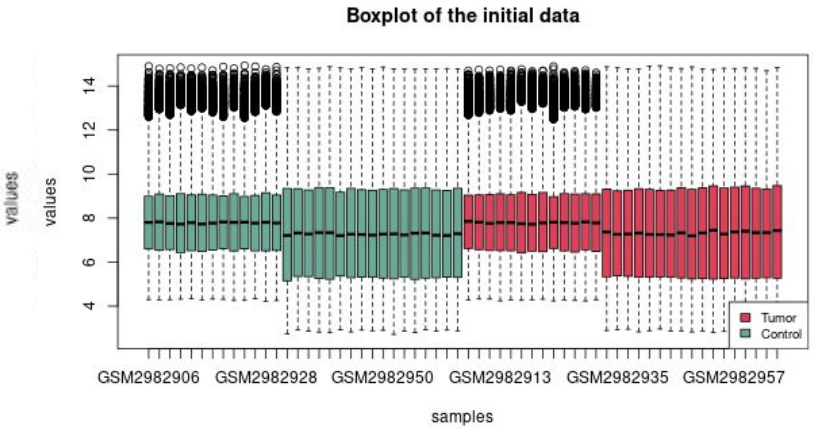
- ➔ GSE110225 composed of two sub-datasets: GSE110223 and GSE110224
- ➔ 30 patients with histologically confirmed, primary, untreated colorectal adenocarcinomas
- ➔ For each patient tumor and control sample were obtained from the colon through surgical intervention

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).

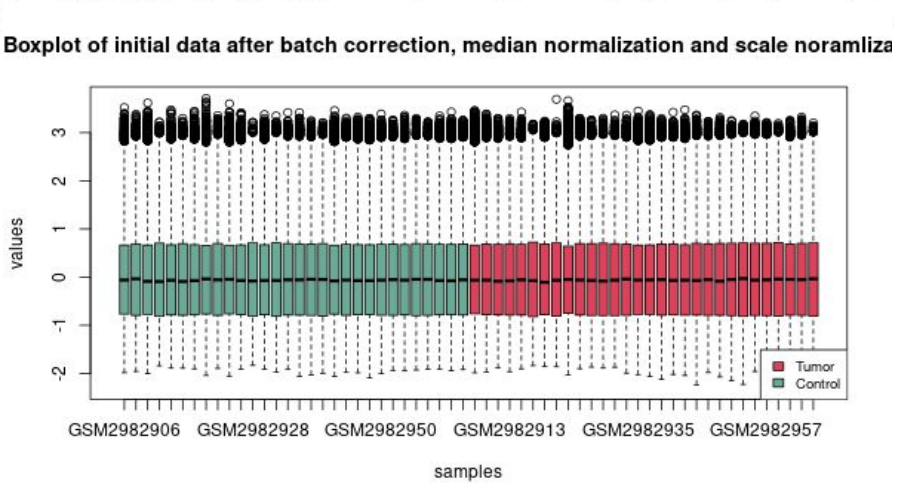
Methods


Principal Component Analysis
K-means clustering
Hierarchical Clustering
Unsupervised Learning Methods
GSE110225
Combat & Normalization
Supervised Learning Methods
rowttest
Random Forest
LDA
LASSO
Ridge
SCUDO
Extraction of the most important features
Functional enrichment analyses
Over-Representation Analysis
gProfiler
DAVID
Network-based Analysis
pathfindR
enrichNet
STRiNG

Results - Combat & Normalization

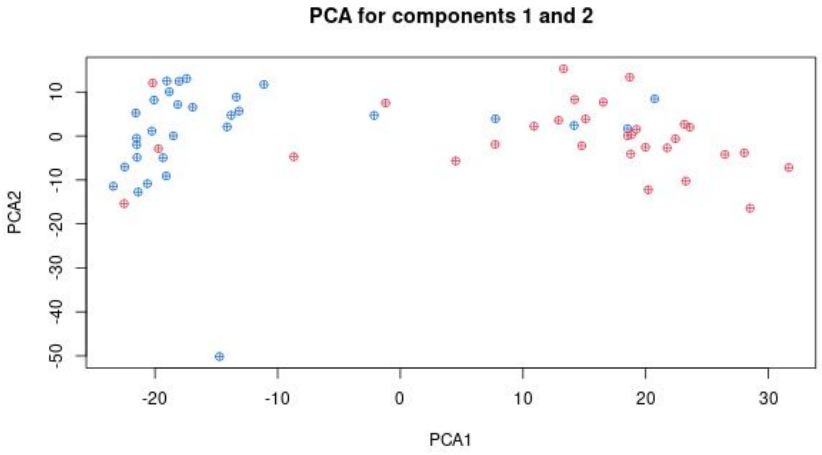


(a) *Boxplots data before any type of normalization*

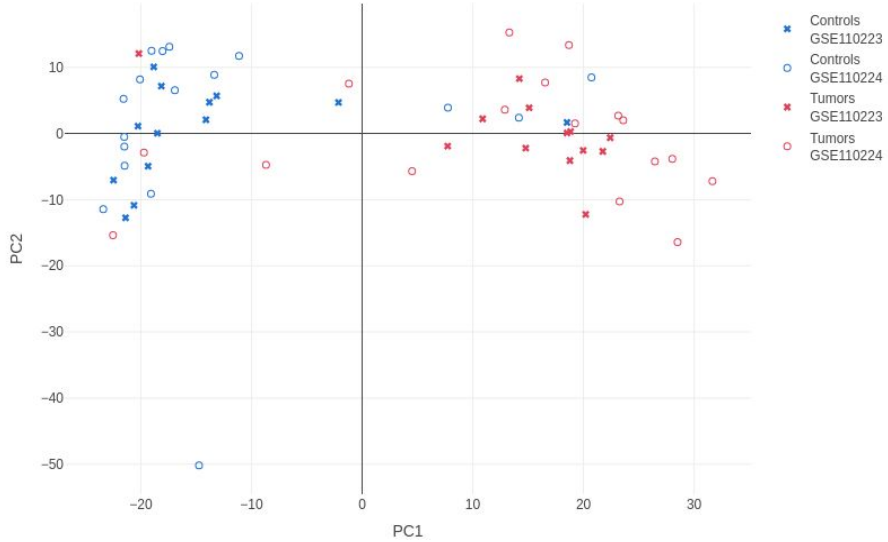


(c) *Boxplot data after pre-processing*

Results - PCA



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



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

Results - Unsupervised clustering

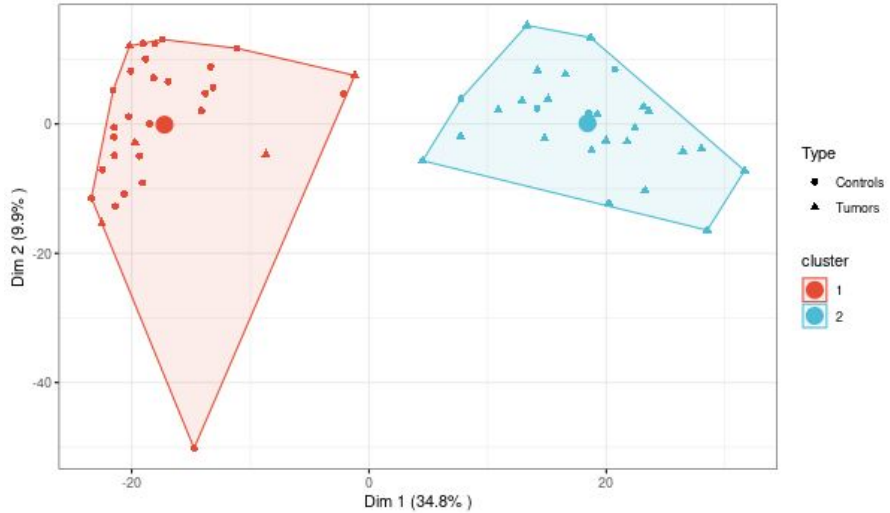


Figure 3.2: **Cluster analysis**. K-mean clustering, where the axes are coordinates for the variables extracted from the first and the second PC.

Results - Unsupervised clustering

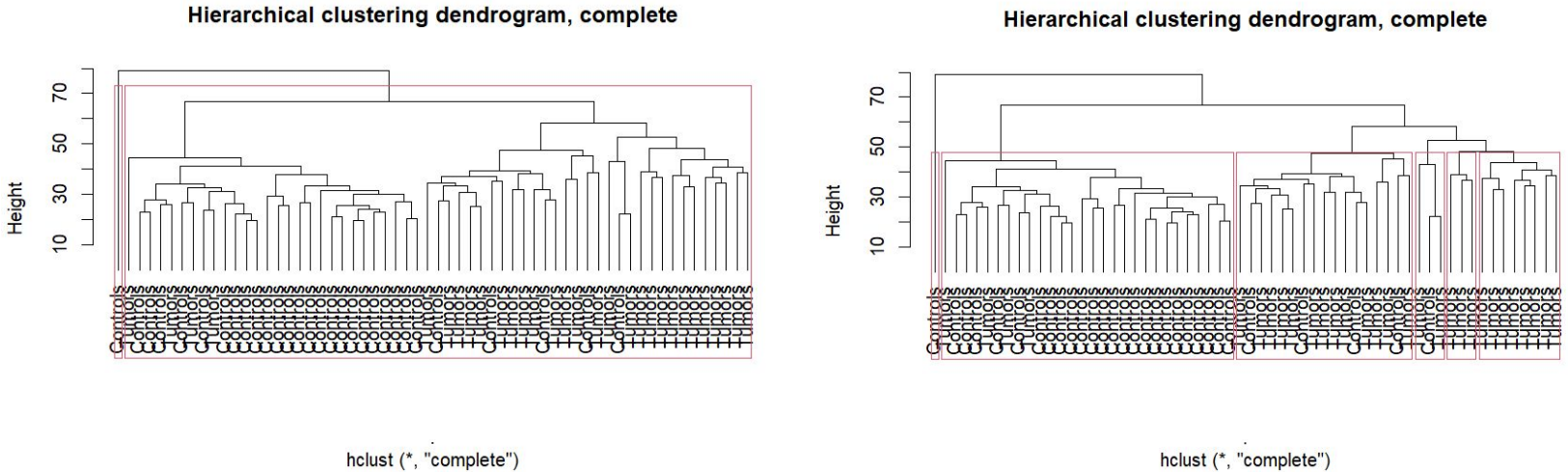


Figure 3.3: **Cluster analysis**. Hierarchical clustering, the red boxes represent the identified clusters. The x-axis represents the sample type (Tumors or Controls), and the y-axis the height.

Results - Supervised learning methods

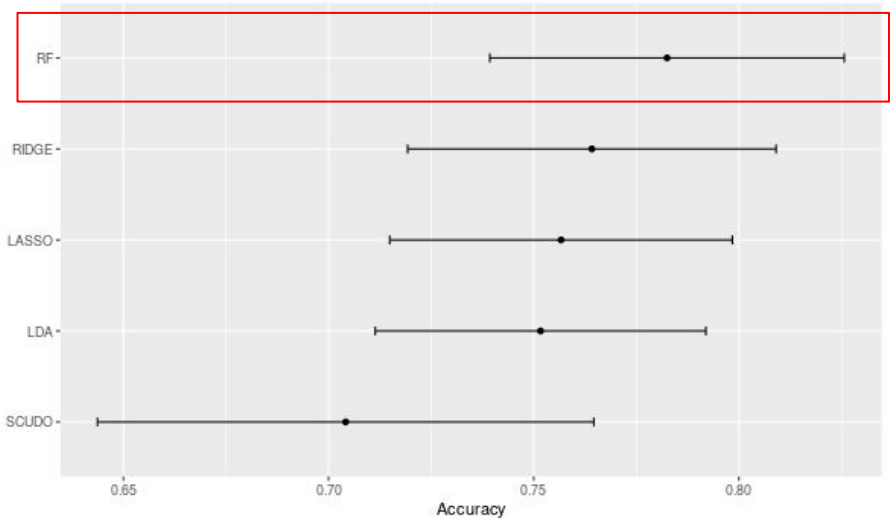
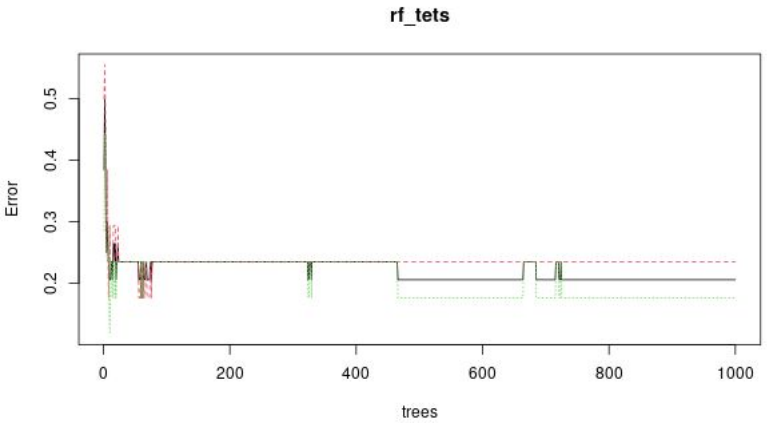
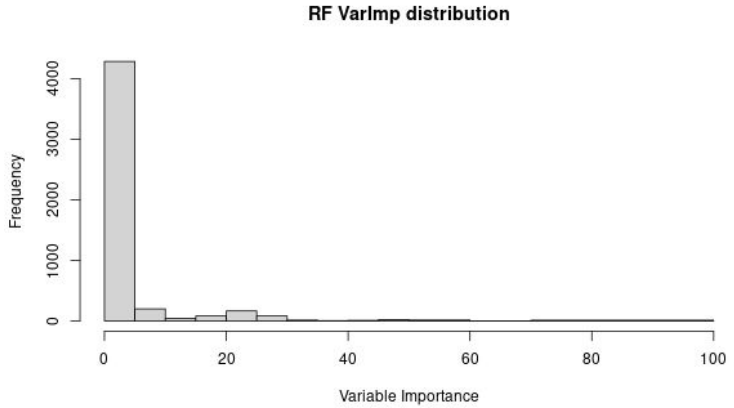


Figure 3.4: **Performance plot of the supervised models.** All models. x-axis: accuracy, y-axis: model name.

- ➔ rowttest from 22277 variables to 4944 variables
- ➔ Random Forest presents the highest accuracy, equal to 0.7825

Results - Supervised learning methods

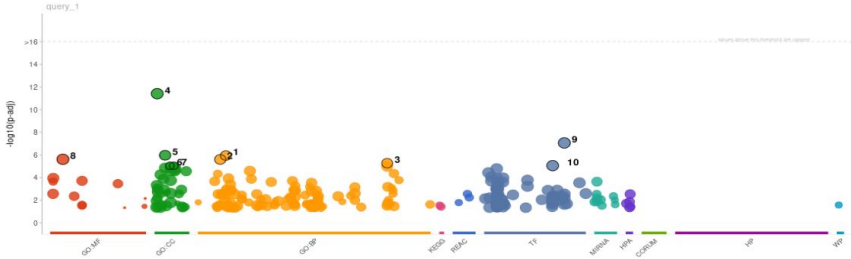




- ➔ OOB score over the total number of tree used to train the RF model
- ➔ OOB score: number of wrongly classificated observations. The lower, the more accurate the model is.

- ➔ Most of the variables have an importance that sits between 0 and 30

Results - ORA

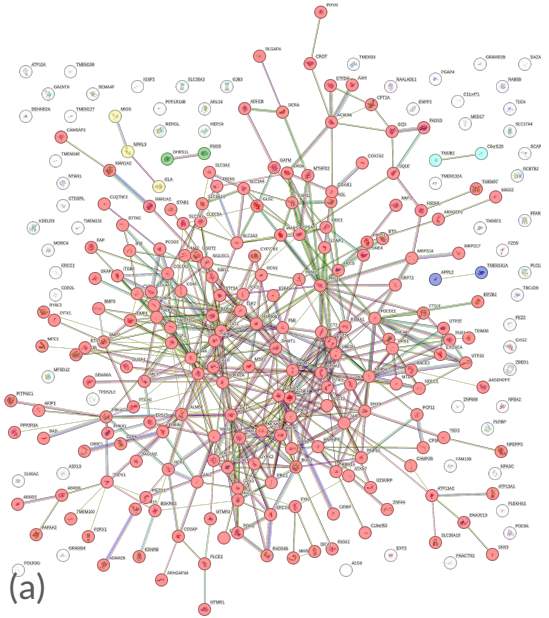


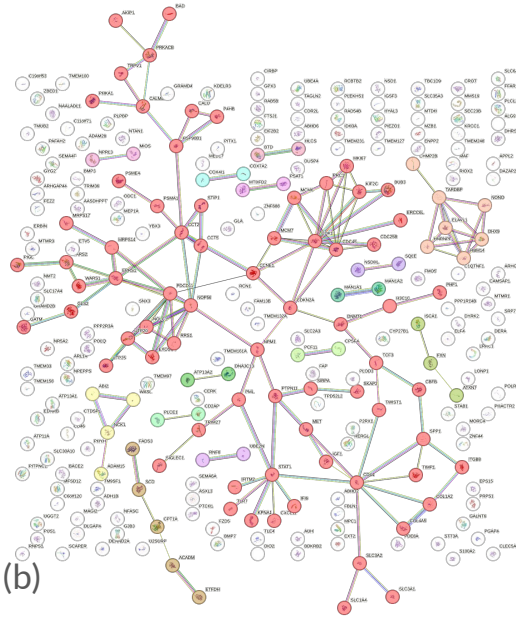
id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0009056	catabolic process	2556	1.1e-06
2	GO:BP	GO:0006950	response to stress	3855	2.5e-06
3	GO:BP	GO:1901575	organic substance catabolic process	2074	5.3e-06
4	GO:CC	GO:0005737	cytoplasm	12345	3.9e-12
5	GO:CC	GO:0031982	vesicle	4004	1.1e-06
6	GO:CC	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

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

➔ extracellular exosome, protein binding and stress response

Results - NBA (STRING)

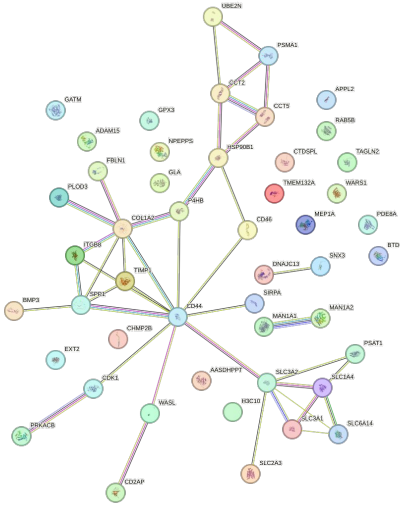




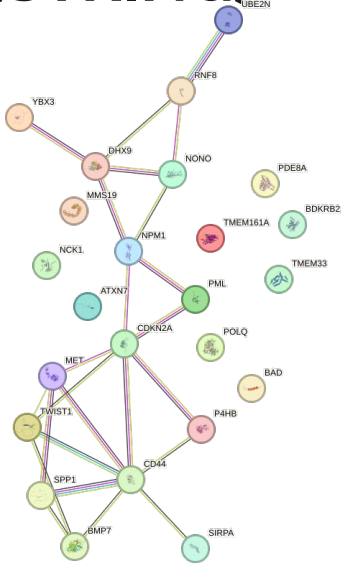
- (a) STRING network: minimum required interaction score* 0.4
- (b) STRING network: minimum required interaction score* 0.7
- Regulation of cellular response to stress, extracellular exosome and RNA binding
- confirm the results of the Over-Representation Analysis

* minimum required interaction score puts a threshold on the confidence score, such that only interaction above this score are included in the predicted network

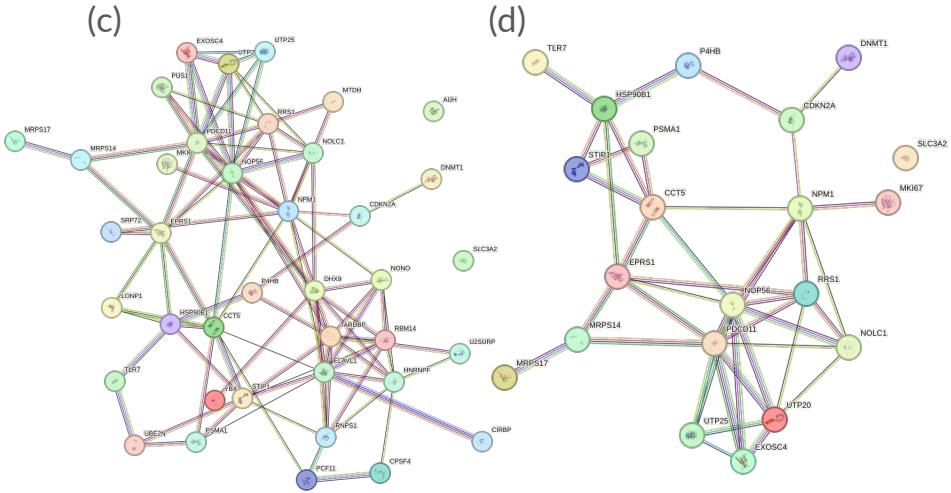
Results - NBA (STRING)



STRING sub-network of the term Extracellular exosome

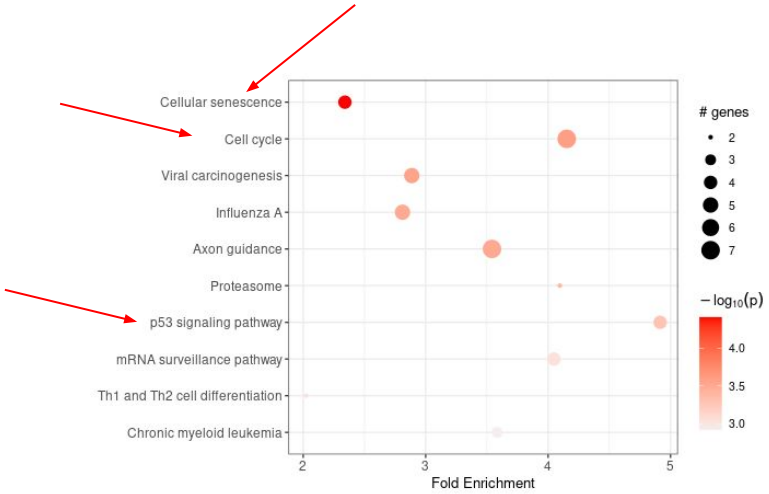


STRING sub-network of the term Regulation of cellular response

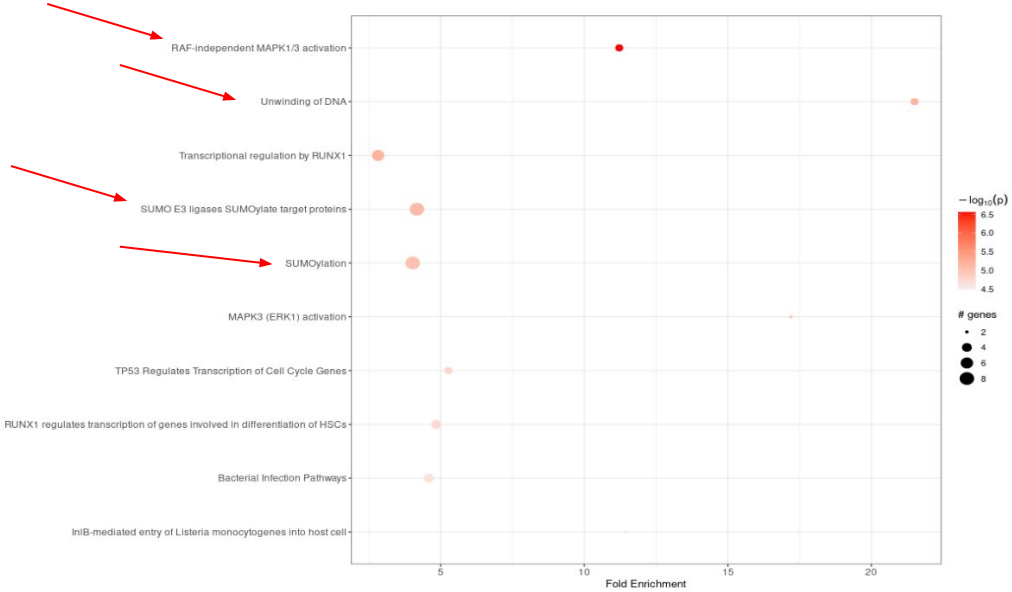


(c) STRING sub-network of the term RNA Binding present in the network with minimum interaction score 0.4. (d) STRING sub-network of the term RNA Binding present in the network with minimum interaction score 0.7.

Results - NBA (PathfindR)

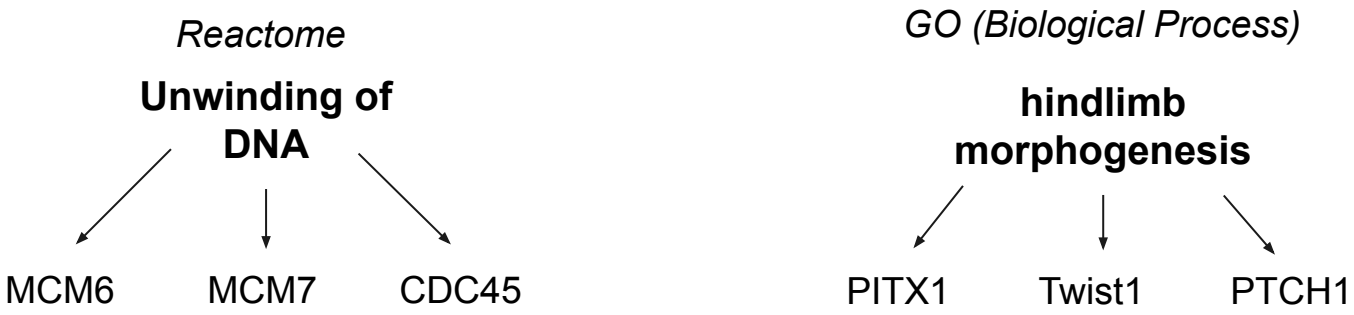


Gene set: KEGG



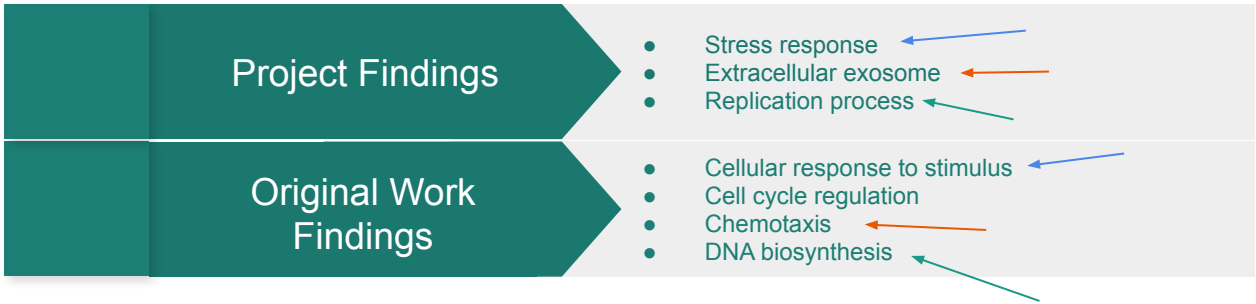
Gene set: Reactome

Results - NBA (EnrichNet)



➔ KEGG unfortunately, there were no terms that resulted significantly

Comparison with original Work



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).

Discussion

- ➔ These analysis aim to find the set of variables that better distinguish Colorectal cancer samples respect control ones.
- ➔ The functional enrichment analysis highlighted terms and pathways related to stress response, extracellular exosome and replication processes.
- ➔ The terms found in the original study match pretty well with the ones found in this project
- ➔ Aberrant cell replication is a well-known hallmark of cancer.
- ➔ The literature highlight the importance of extracellular exosome for the progression of the disease
- ➔ With this findings it is possible to highlight the role of stress as a trigger to a cascade of pathways connected with generation of extracellular vesicles, aberrant replication and proliferation

Thank you for the attention!