Universitat Oberta de Catalunya

uoc.edu

Finding a predictive gene signature in pancreatic cancer using gene expression

Sabela de la Torre Pernas

Master in Data Science

Machine learning in Cancer

Carles Barceló

Ferran Prados Carrasco

2021/06/06

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Finding a predictive gene signature in pancreatic cancer using gene expression*
Nombre del autor:	*Sabela de la Torre Pernas*
Nombre del consultor/a:	*Carles Barceló*
Nombre del PRA:	*Ferran Prados Carrasco*
Fecha de entrega (mm/aaaa):	06/2021
Titulación:	*Master in Data Science*
Área del Trabajo Final:	*Machine Learning in Cancer*
Idioma del trabajo:	***English***
Palabras clave	*Pancreatic cancer, gene expression, predictive model*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

El cáncer de páncreas es uno de los cánceres más agresivos que afectan al ser humano con una tasa de supervivencia a los 5 años inferior al 10%. En el momento del diagnóstico, para muchos pacientes es ya demasiado tarde y no pueden ser sometidos a cirugía o no puede beneficiarse de tratamientos con quimioterapia.

El objetivo del presente trabajo es estudiar los perfiles de expresión génica des de diferentes perspectivas, para generar varias firmas génicas con poder predictivo. Estas firmas podrían ayudar a los médicos tanto en la prognosis de la enfermedaad como para determinar tratamientos según el paciente.

Los procesos biológicos de los genes en estas firmas génicas también han sido analizados, encontrando que la mayoría de ellos están relacionados con processos celulares y metabólicos.

Además, el análisis de supervivencia con Kaplan-Meier Plotter ha mostrado que muchos de estos genes están significativamente correlacionados con la supervivencia en el cáncer de páncreas.

Finalmente, el poder predictivo de las diferentes firmas génicas ha sido estudiado mediante un algoritmo de Machine Learning. Concretamente, varios modelos de Random Forest han sido entrenados y evaluados con diferentes configuracions. La mejor precisión (62%) se obtuve con la firma génica fruto de la intersección de las firmas obtenidas a partir de los dos grupos estudiados: Tratamiento vs resultado y Expressión génica vs resultado.

Abstract (in English, 250 words or less):

PDAC is one of the most aggressive human cancers with a 5-year overall survival rate lower than 10%. At the time of diagnose, it is already too late for many patients who can’t benefit from surgical procedure or chemotherapy.

The objective of the present work is to study the gene expression profiles from different perspectives, and generate several signatures with predictive power. This could allow physicians to improve prognosis and find better treatments according to each patient.

The biological processes where the genes in these signatures participate were studied, showing that most of the genes are related to cellular and metabolic processes.

The survival analysis with Kaplan-Meier Plotter showed that many of these genes had a significant correlation with survival in pancreatic cancer.

Finally, the predictive power of the different signatures was assessed using a Machine Learning algorithm. Specifically, several Random Forest models were trained and evaluated with different configurations. The best accuracy (62%) was obtained with the common signature, which included the intersection of the genes in the signatures of each of the groups studied, Treatment vs outcome and Gene expression vs outcome.

Table of contents

1 Introduction....1
1.1 Keywords....1
1.2 Abstract....1
1.3 Description and justification....2
1.4 Personal motivation....2
1.5 Goals....3
1.6 Planning....3
2 State of the art....6
2.1 Gene expression studies....6
2.2 Other data sources studies....7
2.3 Application studies....8
2.4 Other cancers....8
2.5 Summary of techniques....8
3 Methodology....10
3.1 Data acquisition....10
3.2 Data analysis....12
4 Results....14
4.1 DEG analysis....14
4.2 Kaplan-Meier survival analysis....17
4.3 Functional analysis....20
4.4 Evaluation of signatures as predictive models for prognosis using machine learning....27
5 Conclusions....50
6 Future work....51
7 Glossary....52
8 References....58
9 Annexes....61
9.1 DEG analysis....61

9.2 Gene signatures.....73
9.3 ML model training and evaluation.....76
9.4 Code repository.....78

List of figures

Figure 1: Timeline of this project.....3

Figure 2: Summary of tasks.....4

Figure 3: Stage-specific survival for histologically confirmed PC derived from SEER data.....15

Figure 4: Biological processes of genes in “Treatment vs outcome” signature.....20

Figure 5: Details of “cellular process” category.....21

Figure 6: Details of "metabolic process" category.....21

Figure 7: Biological processes of genes in “Gene expression vs outcome” signature. .22

Figure 8: Details of "cellular process" category.....23

Figure 9: Details of "cellular metabolic process" category.....23

Figure 10: Biological processes of genes in common signature.....24

Figure 11: Details of “cellular process” category.....25

Figure 12: Details of “cellular metabolic process” category.....25

Figure 13: Details of “metabolic process” category.....26

Figure 14: Heat map of the correlation matrix using the signature of “Treatment vs outcome” and all the samples available.....29

Figure 15: Heat map of the correlation matrix using the signature of “Treatment vs outcome” and samples corresponding to deceased patients.....30

Figure 16: Heat map of the correlation matrix using the signature of “Gene expression vs outcome” and all the samples available.....31

Figure 17: Heat map of the correlation matrix using the signature of “Gene expression vs outcome” and samples corresponding to deceased patients.....32

Figure 18: Heat map of the correlation matrix using the common signature and all the samples available.....33

Figure 19: Heat map of the correlation matrix using the common signature and samples corresponding to deceased patients.....34

Figure 20: Histogram of PFS_MONTHS.....35

Figure 21: Correlation matrix with the subset of “Treatment vs outcome” signature….38

Figure 22: Correlation matrix with the subset of “Gene expression vs outcome” signature.....40

Figure 23: Correlation matrix with a subset of the common signature.....42
Figure 24: Variable importance of 60/40 model.....45
Figure 25: Variable importance of 70/30 model.....46
Figure 26: Variable importance of 90/10 model.....47
Figure 27: Variable importance of 70/30 model.....48
Figure 28: Box plot for GSE112282.....61
Figure 29: Venn diagram for GSE112282.....62
Figure 30: Box plot for GSE45757.....63
Figure 31: Venn diagram for GSE45757.....63
Figure 32: Box plot for GSE14426.....64
Figure 33: Venn diagram for GSE14426.....64
Figure 34: Box plot for GSE21501.....65
Figure 35: Venn diagram for GSE21501.....66
Figure 36: Box plot for GSE28735.....67
Figure 37: Venn diagram for GSE28735.....68
Figure 38: Box plot for GSE62165.....69
Figure 39: Venn diagram for GSE62165.....70
Figure 40: Box plot for GSE71729.....71
Figure 41: Venn diagram for GSE71729.....71
Figure 42: Box plot for GSE56560.....72
Figure 43: Venn diagram for GSE56560.....72

List of tables

Table 1: Summary of the selected series *in* Treatment vs Outcome.....10

Table 2: Summary of the selected series in Gene expression vs Outcome.....11

Table 3: Summary of DEG analysis for Treatment vs Outcome.....14

Table 4: Summary of DEG analysis for Gene expression vs Outcome.....15

Table 5: MST vs Stage.....15

Table 6: Genes of “Treatment vs outcome” signature significantly correlated with poor survival.....17

Table 7: Genes of “Gene expression vs outcome” signature significantly correlated with poor survival. In orange, upregulated genes. In black, downregulated genes.....18

Table 8: Genes of the common signature significantly correlated with poor survival. In orange, upregulated genes. In black, downregulated genes.....19

Table 9: Summary of the selected study in cBioPortal.....27

Table 10: Analysis of missing values.....28

Table 11: Proportion of alive and dead patients.....30

Table 12: Missing values in “Treatment vs outcome” signature.....36

Table 13: Missing values in “Gene expression vs outcome” signature.....36

Table 14: Missing values in common signature.....37

Table 15: Sizes of train and test sets.....44

Table 16: Models’ metrics using “Treatment vs outcome” signature.....44

Table 17: Confusion matrix of 60/40 model.....44

Table 18: Models metrics using “Gene expression vs outcome” signature.....46

Table 19: Confusion matrix of 70/30 model.....46

Table 20: Confusion matrix of 90/10 model.....47

Table 21: Models metrics using the common signature.....48

Table 22: Confusion matrix of 70/30 model.....48

Table 23: Common DEGs to series in Treatment vs outcome.....73

Table 24: Common DEGs to series in Gene expression vs outcome.....74

Table 25: Common DEG to both groups.....75

1 Introduction

1.1 Keywords

Pancreatic cancer, personalized medicine, gene expression, data mining, predictive model.

1.2 Abstract

Pancreatic ductal adenocarcinoma, is one of the most aggressive human cancers with a 5-year overall survival rate lower than 10%[1]. Traditional treatments, like chemotherapy, surgery and radiation, have not proved significant to improve survival[2]. Only some chemotherapy agents (FOLFIRINOX, gemcitabine) and one targeted therapy (erlotinib) have shown some degree of efficacy, but a few rate of patients responded to these drugs. On the other hand, developing new drugs will also be needed, as there has not been much improvement in the treatment of the PDAC for the last 10, or even 20 years[3].

This cancer is also very difficult to diagnose, most of the tumours can't be resected, are locally advanced or have metastasised by the time the disease is detected. There is a variety of mutations involve in PDAC, and each of them is present in a small fraction of patients, which makes very difficult to apply targeted therapies. Also, a small fraction (0.5-1%) of the pancreatic cells are cancer stem cells (CSC), which have an increased capacity for self-renewal and have specific properties, e.g. chemo-resistance, that allow them to escape treatments. This cancer also has, metastatic potential and an overdeveloped tumour micro-environment (TME) which hinders drug activity because they can't penetrate the stroma.

For all these goals, developing new drugs, targeting treatments and early prognosis, it is necessary to molecularly characterize these tumours. So far, the AJCC-TNM classification is the only tool used by physicians to guide the treatment and evaluate the survival rate of a patient. But this method constantly fails when evaluating this type of cancer[4],[5].

In this project we want to study the relation between treatments, their outcome and gene expression to generate a predictive gene signature which aims at helping physicians choosing a specific treatment for each patient, that is, applying targeted therapy.

1.3 Description and justification

As PDAC has such a low survival rate, using targeted therapy might work better than current standard treatments and, hopefully, increase the survival chances of the patients.

The purpose of this predictive gene signature is to predict an outcome based on the gene expression and treatment. In order to generate this signature, we will try to find the relevant processes and the best therapeutic window by analysing different datasets which compare treatments and outcome, and gene expression and outcome.

If this project is successful and we develop a good predictive gene signature, it might help physicians to decide what treatments they should use in each patient.

On the other hand, the outcome of this project might not be of any help for current patients, as expecting to generate a useful tool is probably too ambitious considering the short amount of time that we have, but might be a base line for further research.

1.4 Personal motivation

I studied computer science, although I was also very interested in sciences in general and biology in particular. My degree's final year project was about building a Desktop application which could compare a pair of images from the surface of Jupiter and calculate some metrics which were of interest to astronomers studying this planet.

After finishing my degree, I worked in a couple of private companies, but the projects were not meaningful enough to me, and I tried to find something different and more interesting. Then, I started to work at CRG (Centre de Regulació Genòmica[1]) as a software programmer. We are involve in many projects related to genomics and we develop tools which can help the research community.

I chose this project because I think it is an opportunity to contribute to research, it is challenging and, finally, it is about a topic which I am not familiar with, but I would like to learn.

1 https://www.crg.eu/

1.5 Goals

The main goals of this project are:

- Generate three different signatures using gene expression data in studies with different approaches, specifically 1) Treatment vs outcome, and 2) Gene expression vs outcome.
- Explore the biological processes of the genes in these signatures.
- Examine the prognostic significance of the genes in these signatures.
- Study the predictive power of these signatures for pancreatic cancer prognosis using Machine Learning.

1.6 Planning

The following Figure 1 depicts the different phases of this project, which are described below:

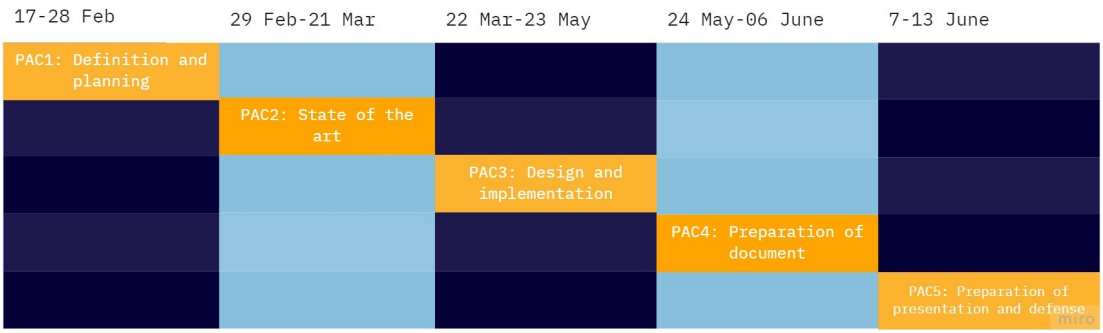

Figure 1: Timeline of this project

1. Definition and planning (11 days)

In this phase, the scope of this project is defined, as well as, the personal motivation.

2. State of the art (20 days)

First step was to analyse the state of the art of this topic. That was, searching and reading papers, studies, etc., which seemed to be related with what this project wants to achieve.

3. Design and implementation (62 days)

Next step consisted in defining the tasks which need to be done in order to accomplish the goals defined in phase 1.

The following figure Figure 2 depicts these tasks and their ordering:

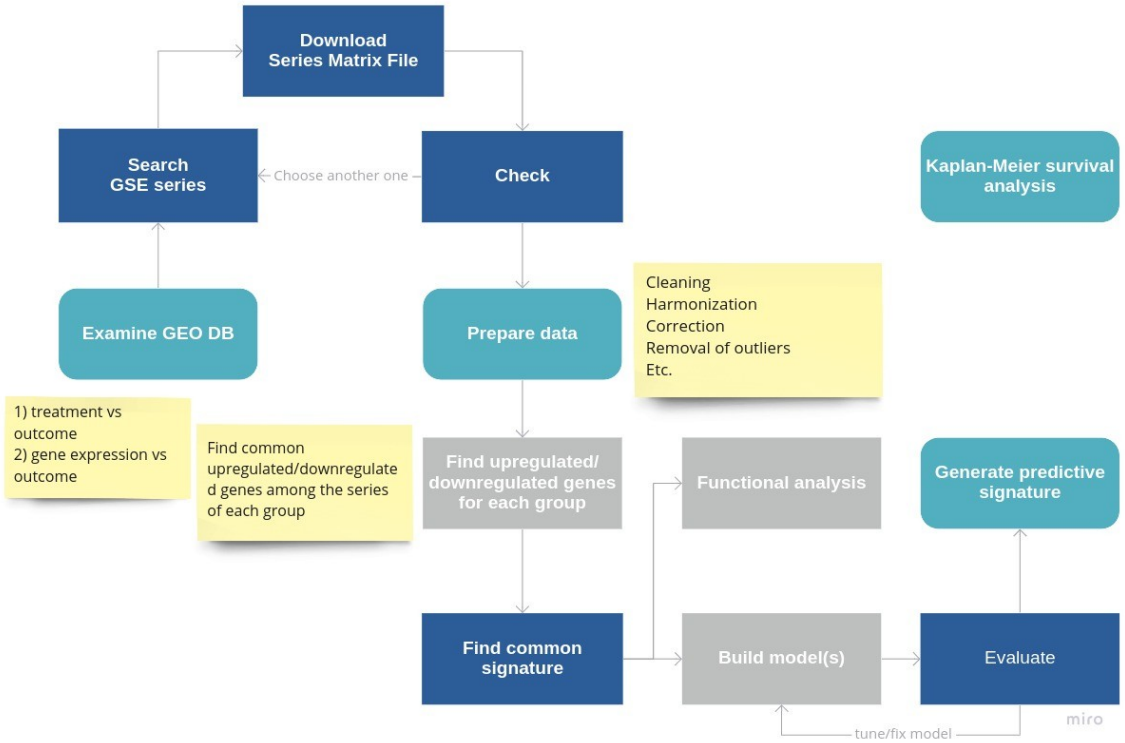

Figure 2: Summary of tasks

Each of these tasks is now briefly described:

i. Examine GEO DB: get familiar with this repository. Understand what data is stored.

ii. Search GSE Series: using the GEO search tool, try to find series matching the two groups of interest: 1) Treatment vs outcome, and 2) Gene expression vs outcome.

iii. Download Series Matrix File: download the expression data, as well as, the metadata.

iv. Check: some series didn't have the expression data available. Other series didn't have clinical data among the metadata.

v. Prepare data: some arrangements were necessary to be done in the data in order to do the Differentially expressed genes analysis.

vi. Find upregulated/downregulated genes for each group: data was analysed in order to find these DEGs.

vii. Find common signature: a gene signature was generated for each group, as well as, a final signature.

viii. Build model: different models were trained using the different signatures.

ix. Evaluate: the different models were evaluated using unseen data.

x. Generate final signature: considering the model, the most significant genes were selected for this final signature.

xi. Functional analysis: this allowed learning about the biological processes related to the genes in the signatures.

xii. Kaplan-Meier survival analysis: this is an interesting tool to compute the effect of different variables to the survival.

4. Preparation of the document (13 days)

In this step, final results are obtained and this document is finished.

5. Preparation of the presentation and defence (6 days)

Final phase consists of preparing and recording a presentation, and defend the project in front of the committee.

2 State of the art

During the last two decades there has been a huge progress in computing. Machine Learning algorithms are now available for anybody with many languages providing implementations of the most common ones. Hence, it is not strange to find so many studies where these techniques have been used in some way or another.

In this chapter, we take a brief look at some studies related to finding predictive models for pancreatic cancer,

2.1 Gene expression studies

There have been many efforts to generate a predictive model based on gene expression.

Artificial Neural Networks were used to create a model to diagnose PDAC using the 5 best genes among all DEG which were found overexpressed in microarray expression data. The model generated could classify samples with a sensitivity of 87.6, and specificity of 83.1[6].

Another study built a predictive nomogram integrating clinicopathological information and the risk score based on survival-related genes through a univariate Cox regression analysis. This study identified four new biomarkers, and the nomogram showed robust performance, making it an effective and reliable guide for prognosis assessment and treatment decision-making in the clinic[7].

There is a paper where a novel hybrid framework based on data mining techniques was proposed. They also defined two methods of gene selection, and included the age of patients as an additional factor. The results were very promising, and showed their biological validity. Interestingly, they also found that the age is not relevant in the expression changes of the selected genes when the pathology has already developed[8].

There is also a very interesting study where they found a gene signature to identify early-stage PDAC. They took a different approach based on within-sample relative expression orderings (REOs), and used a selection technique called minimum redundancy maximum relevance (mRMR) to pick out the optimal REOs. They also compared the performance of different classification algorithms, being SVM the best one. Finally, they defined a 9 gene pairs' signature, which was validated with data from different platforms, namely, microarray and RNA-Seq. This study was also interesting because raised a concert regarding the existing diagnostic signatures. They claimed that the batch effect could influence the choice of these signatures because they are basically obtained by using signature genes' absolute expression value. This is why they chose using REO, which is highly robust to experimental batch effects and

platform differences, and has already been successfully used to identify the early diagnosis signature of malignant carcinoma[9].

2.2 Other data sources studies

There are also some studies which built predictive models using microRNAs (miRNAs) [10]–[12], long non-coding RNAs (lncRNAs)[3], and proteomics[13] data.

For example, in one of these studies, they selected 27 miRNA signatures through a differential expression analysis, and used these as the input data for two independent feature selection algorithms (PSO + ANN, and NCA). The final model consisted of five miRNAs and showed great diagnostic results, including the validations set (0.93 accuracy, and 0.93 sensitivity). However, they also claimed that experimental testing is needed[11].

Other studies used a multi-omics approach to the problem, integrating data from different sources[10],[14]. One of these found five candidate genes what were mutated in the early stages, and had high cellular prevalence (CP). They then integrated data from different sources: RNA sequencing, microRNA sequencing, and DNA methylation. They reduced the dimensions with an autoencoder, and used K-means to cluster the patients in two subgroups. They found that patients in these subgroups had significant differences in survival and recurrence. They finally developed a prediction model for prognosis using two biological features (weather a sample had mutations in candidate genes, and the subgroup assigned by K-means clustering), and 9 clinical features (e.g. sex, grade, AJCC cancer stage, age, treatment, etc.). Among all the trained models, logistic regression showed the best performance for both DFS and OS[10].

In the other multi-omics study mentioned, they used next-generation sequencing, transcriptome meta-analysis, and immunohistochemistry, combined with statistical learning, to validate multiplex biomarkers candidates. They applied several statistical methods (Kaplan-Meier survival analysis, multivariate Cox regression analysis, Pearson's correlation analysis and Spearman's rank correlation analysis), and, finally, built a Random Forest classification model based on 11 potential genes. It showed excellent performance, and found that four genes were the most important variables of the model. Also, an interesting fact they rose is that samples usually stem from patients with the advanced disease, but biomarkers are expected to be for early, curable stages, which could be a problem when generating these signatures[14].

We also found an interesting study where they used data from blood tests (comprising cancer antigens, hemoglobin, leukocytes, hematocrit, and platelets), to train a Twin Support Vector Machine (TWSVM) model, giving an accuracy of 98%, and sensitivity of 100%. There was no validation with another dataset(s), but results seemed interesting enough to further investigation, considering that blood tests are very easy to obtain and non-invasive[15].

2.3 Application studies

A paper was found where they studied and listed the many applications of Artificial Intelligence in pancreatic cancer: (1) in molecular/imaging/pathological diagnosis; (2) in radiotherapy, to both use an appropriate dose and an accurate location; (3) in chemotherapy, to guide it based on the cancer subtypes; (4) in surgical treatment through the usage of robot-assisted pancreatic cancer surgery (RDP); (5) in prognosis[16].

2.4 Other cancers

In the context of breast cancer, there is a very interesting study where they developed several multi-gene expression signatures to help predicting different factors of interest for the patient: risk factors (e.g. hereditary cancer predisposition), recurrence or metastasis (e.g. probability for distant recurrence for 10 years, distant metastasis within 5 years), and response to therapies (e.g. guide chemotherapy decisions, detection of deficiencies in some genes which produce a wrong activation of some drugs). These signatures have improved the diagnose of breast cancer, prognosis of tumours, identification of therapeutic targets, and prediction of response to adjuvant systemic therapies[17].

There are also similar studies regarding lung cancer. In one of these, they used several machine learning algorithms to investigate the gene expression profiles of two types of lung cancer. They used MCFS to find the informative features, and fed this list into the IFS method to extract the optimal features. Then, they constructed an optimal SVM classifier, which showed very high performance in distinguishing the two types of lung cancer[18].

2.5 Summary of techniques

All the studies mentioned above, used different methods for feature selection, algorithms to create the models, statistical measures, etc. In this section, we will list all these tools, as they can be a very helpful source of ideas when developing our own project.

2.5.1 Machine learning algorithms

- K-means
- K-Nearest Neighbour (KNN)
- Artificial Neural Network (ANN)
- Support Vector Machine (SVM)
- Twin Super Vector Machine (TWSVM)

- Random Forest (RF)
- Logistic Regression (with or without regularization, e.g. L2)

2.5.2 Dimensionality reduction

- Principal Component Analysis (PCA)
- Neighbourhood Component Analysis (NCA)
- Monte Carlo feature selection (MCFS)
- Incremental Feature Selection (IFS)
- Particle Swarm Optimization (PSO)
- Minimum Redundancy Maximum Relevance (mRMR)
- Autoencoder

2.5.3 Statistical measures

- Pearson's correlation coefficient
- Spearman's rank correlation
- Mann-Whitney test
- Kruskal-Wallis test

2.5.3.1 Survival analysis

- Kaplan-Meier
- Cox regression

3 Methodology

3.1 Data acquisition

All data expression used in this project to find the gene signatures has been retrieved from the Gene Expression Omnibus (GEO) repository, and has been downloaded, processed and analysed using R programming language.

On the other hand, data (expression and clinical) used to train the Machine Learning algorithm has been retrieved from the cBioPortal for Cancer Genomics repository. The programming language used for this step has also been R.

3.1.1 Gene expression data for Treatment vs Outcome

- GSE112282[2]: BET inhibitor GSK525762 and MEK inhibitor trametinib treatments. Array: Affymetrix Human Genome U133 Plus 2.0.
- GSE45757[3]: MEK inhibitor CI-1040 treatment. Array: Affymetrix Human Genome U133A 2.0.
- GSE14426[4]: retinoic acid treatment. Array: Illumina human-6 v2.0 expression beadchip.

Series	Total samples	Samples used	Array
GSE112282	48	48	Affymetrix Human Genome U133 Plus 2.0
GSE45757	141	132	Affymetrix Human Genome U133A 2.0
GSE14426	30	12	Illumina human-6 v2.0 expression beadchip

Table 1: Summary of the selected series in Treatment vs Outcome

2 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112282
3 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45757
4 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14426

3.1.2 Gene expression data for Gene expression vs Outcome

- GSE21501[5]: expression and clinical data. Array: Agilent-014850 Whole Human Genome Microarray 4x44K G4112F.
- GSE28735[6]: expression and clinical data. Array: Affymetrix Human Gene 1.0 ST.
- GSE62165[7]: expression and clinical data. Array: Affymetrix Human Genome U219.
- GSE71729[8]: expression and clinical data. Array: Agilent-014850 Whole Human Genome Microarray 4x44K G4112F.
- GSE56560[9]: expression and clinical data. Array: Affymetrix Human Exon 1.0 ST.

Series	Total Samples	Samples used	Array
GSE21501	132	102	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F
GSE28735	90	42	Affymetrix Human Gene 1.0 ST
GSE62165	131	118	Affymetrix Human Genome U219
GSE71729	357	125	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F
GSE56560	35	28	Affymetrix Human Exon 1.0 ST

Table 2: Summary of the selected series in Gene expression vs Outcome

5 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21501
6 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28735
7 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62165
8 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71729
9 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56560

3.2 Data analysis

3.2.1 Differentially expressed genes analysis

In order to find a common signature with potential predictive power, gene expression data was analysed to find the differentially expressed genes (DEG) for each of the following types of study: 1) treatment vs outcome, and 2) gene expression vs outcome.

The steps followed to perform this analysis with *limma* package in R, were:

1. Verification that data is normalised and log2.

It could be easily verified by plotting a box plot of the data with *boxplot()* function from *graphics* R package.

2. Selection of platform characteristics.

In most of the platforms (a.k.a. arrays), the ID column corresponded to the probe ID, but we were interested in genes. The information available in the platforms was used to convert from the probe ID to GenBankID.

In some cases, it was also necessary to split one row into many because that probe matched with more than one gene. For other series, it was also necessary to collapse several rows pointing to the same gene into a single one. For the latter, the average was calculated on the expression data using *avereps()* function from *limma* package.

3. Selection of sample and characteristics.

For example, in some series only samples corresponding to tumour samples were selected, and the *survival* value was used to perform the DEG analysis.

4. Design and execution of the DEG analysis.

The DEG analysis was performed using *model.matrix()* from *stats* package, *makeContrasts()*, *lmFit()*, *eBayes()*, and *decideTests()* functions from *limma* package in R. Specific details about this and previous steps for each of the series can be found in annex DEG analysis.

5. Intersection of DEG in common between series of each group.

This generated two lists of DEG, one for each of the groups.

6. Intersection of DEG in common between the two groups.

This intersection was done between the two lists produced in the previous step. The signatures produced in both steps 6 and 7 can be found in annex Gene signatures.

3.2.2 Correlation analysis

Several correlation analysis were performed using the different signatures obtained after the DEG analyses for each of the groups. These analyses also included prognostic parameters like OS or PFS. In all cases, *cor()* function of *stats* package, as well as, *ggplot()* function of *tidyverse* package, were used: the former to calculate the correlation matrix, and the latter to nicely print it as a heat map. Finally, Pearson coefficient was chosen as the test statistic.

3.2.3 Kaplan-Meier survival analysis

The effect of the genes of the different signatures on survival was assessed using the Kaplan-Meier Plotter[10] online tool[24]. The data sources of this tool include GEO, EGA and TCGA databases. It includes gene expression and clinical data in order to analyse the prognostic value of a particular gene by comparing the two patient cohorts (e.g. high vs low expression) in a Kaplan-Meier survival plot, as well as, the hazard ratio with 95% confidence intervals and *logrank P* value.

3.2.4 Functional analysis with Panther

In order to understand the biological processes of the genes involved in the different signatures, the Panther Classification System[11] online tool[25] was used. This tool classifies the provided genes by the function of the protein, considering also the interaction with other proteins to accomplish some goal at the level of the cell or organism.

3.2.5 Machine Learning

The algorithm chosen was Random Forest, which is a supervised algorithm for classification and regression tasks.

In this project, classification was used in order to predict the class of interest. Different proportions of training and test data were used, namely 60%/40%, 70%/30%, 80%/20% and 90%/10%, respectively. Also, cross-validation method with 10 folds was used, in conjunction with 3 and 10 repeats. Different values of *mtry* and *ntree* were tried, and default values were used for *maxnode* and *nodesize*. Finally, for the evaluation of the different models, accuracy, specificity and sensitivity were calculated.

10 https://kmplot.com/
11 http://pantherdb.org/

4 Results

4.1 DEG analysis

The DEG analysis was performed on 3 series for “Treatment vs outcome” and 5 series for “Gene expression vs outcome”.

The goal was to detect the genes differentially expressed considering different criteria.

4.1.1 Treatment vs Outcome

This table summarises the analyses performed on the three series:

Series	Sample characteristics	Contrasts	p.value	Total DEG
GSE112282	Cell line, replicate and treatment	BET - VEHICLE, BETMEK – VEHICLE, MEK - VEHICLE	0.05	1044
GSE45757	Treated with, cell line	Untreated - Treated	0.05	5450
GSE14426	Source name (only 24hr and 168hr)	Vehicle168h – ATRA168h, Vehicle24h - ATRA24h	0.3	19

Table 3: Summary of DEG analysis for Treatment vs Outcome

All the details about each of these analyses, as well as, the list of genes obtained can be found in annex Treatment vs outcome.

4.1.2 Gene expression vs Outcome

This table summarises the analyses performed on the five series:

Series	Sample characteristics	Contrasts	p.value	Total DEG
GSE21501	Risk	LowRisk - HighRisk	0.05	1205
GSE28735	Survival, tissue (only tumour)	Early - Advanced	0.25	46
GSE62165	Stage, tissue (only tumour)	Early - Advanced	0.6/0.4	199/49
GSE71729	Survival, tissue (only tumour)	Early - Advanced	0.6	83
GSE56560	Grade	G3 - G2	0.99	153

Table 4: Summary of DEG analysis for Gene expression vs Outcome

The following Table 5 has been used to convert the survival value (matched to Median Survival Time, MST) to a stage, based on the figure Figure 3 below[23]:

MST (months)	Stage
21.44	I
11.84	II
8	III
3	IV

Table 5: MST vs Stage

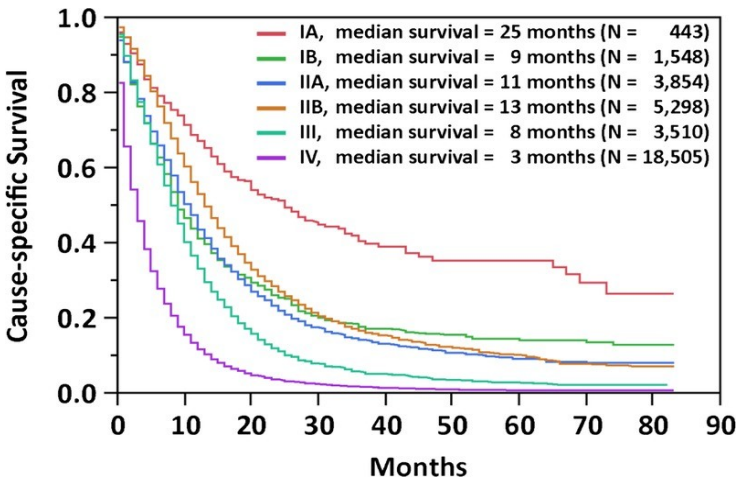

Figure 3: Stage-specific survival for histologically confirmed PC derived from SEER data

All the details about each of these analyses, as well as, the list of genes obtained can be found in annex Gene expression vs outcome.

4.1.3 Common signatures

Once the DEG analyses were run, a signature for each group was calculated. Several approaches were tested in order to find the biggest overlap of genes:

- selecting top X genes (e.g. top 500 genes with lowest *p.value*)
- modifying the *p.value* threshold applied in the *decideTests()* function of *limma* package in R.

The best results were obtained with the latter, and two signatures with 9 (Treatment vs outcome) and 17 (Gene expression vs outcome) genes were generated.

Finally, a common signature to both groups was also calculated, producing a signature of 17 genes.

These three signatures, and more details about the procedure followed, can be found in annex Gene signatures.

4.2 Kaplan-Meier survival analysis

A survival analysis for each signature (see annex Gene signatures) was performed with the Kaplan-Meier Plotter online tool using pancreatic cancer data in order to asses their correlation with prognostic parameters.

Note that only significant genes are displayed in the following tables.

4.2.1 Treatment vs outcome signature

The survival analysis based on the expression data of the 9 genes in the signature (see Treatment vs outcome signature) was done returning that the low expression of 5 of them was significantly (*p.value* < 0.05) correlated with poor survival. Data is presented in Table 6:

Gene	Hazard-Ratio (HR)	logrank P	Median survival in low expression cohort (months)	Median survival in high expression cohort (months)
CREB1	0.85 (0.73 – 0.99)	0.039	35	42.05
IDS	0.87 (0.79 – 0.96)	0.0061	191.21	216.66
BCKDHB	0.75 (0.68 – 0.83)	3.8e-8	39	62.36
HNRNPA2B1	0.57 (0.49 – 0.66)	3.7e-13	122.64	171.43
MYO9A	0.84 (0.72 – 0.98)	0.0221	33	43

Table 6: Genes of “Treatment vs outcome” signature significantly correlated with poor survival

In this case, low expression of CREB1, IDS, BCKDHB, BCKDHB and MYO9A was found to be related to a poorer median survival.

4.2.2 Gene expression vs outcome signature

The survival analysis based on the expression data of the 17 genes in the signature (see Gene expression vs outcome signature) was done returning that the high or low expression of 9 of them was significantly (*p.value* < 0.05) correlated with poor survival. Result is presented in Table 7:

Gene	Hazard-Ratio (HR)	logrank P	Median survival in low expression cohort (months)	Median survival in high expression cohort (months)
HMGA2	0.84 (0.72 – 0.98)	0.0244	33.64	43
C1orf21	0.59 (0.53 – 0.68)	<1e-16	184.04	216.66
CBX6	0.79 (0.71 – 0.87)	2.8e-6	42	60
E2F7	1.36 (1.17 – 1.59)	5.0e-5	49.2	30.42
KCNB1	0.77 (0.66 – 0.9)	0.0007	35	44
MAMSTR	0.61 (0.52 – 0.7)	7.3e-11	28	58
MT1H	1.18 (1.06 – 1.3)	0.0016	216.66	191.21
KIAA0238	0.73 (0.66 – 0.8)	5.7e-10	40.8	61.64
TSPAN3	0.69 (0.6 – 0.81)	2.1e-6	30.42	53.56

Table 7: Genes of "Gene expression vs outcome" signature significantly correlated with poor survival. In orange, upregulated genes. In black, downregulated genes.

In this case, high expression of E2F7 and MT1H was found to be related to a poorer median survival, whereas low expression of HMGA2, C1orf21, CBX6, KCNB1, MAMSTR, KIAA0238 and TSPAN3 was related to a poorer median survival.

4.2.3 Common signature

The survival analysis based on the expression data of the 17 genes in the signature (see Common signature) was done returning that the high or low expression of 7 of them was significantly (*p.value* < 0.05) correlated with poor survival. Result is presented in Table 8:

Gene	Hazard-Ratio (HR)	logrank P	Median survival in low expression cohort (months)	Median survival in high expression cohort (months)
HMFT1638	1.18 (1.07 – 1.31)	0.001	216.66	228.85
CDT1	1.57 (1.35 – 1.83)	4.0e-9	57.6	28
C12ORF2	0.85 (0.77 – 0.94)	0.0012	228.85	216.66
F12	1.14 (1.03 – 1.26)	0.011	53.56	45
MAPK13	1.21 (1.09 – 1.34)	0.0002	59	41.42
RAB4B	0.77 (0.69 – 0.85)	2.8e-7	228.85	216.66
BCEI	0.71 (0.65 – 0.79)	6.0e-11	191.21	216.66

Table 8: Genes of the common signature significantly correlated with poor survival. In orange, upregulated genes. In black, downregulated genes.

In this case, high expression of HMFT1638, CDT1, F12 and MAPK13 was found to be related to a poorer median survival, whereas low expression of C12ORF2, RAB4B and BCEI was related to a poorer median survival.

4.3 Functional analysis

Each of the previous signatures generated was analysed with Panther Classification System online tool in order to learn about the biological processes in which these genes participate.

4.3.1 Treatment vs outcome signature

The biological processes of the 9 genes in this signature (see annex Treatment vs outcome signature) were examined and summarised in Figure 4:

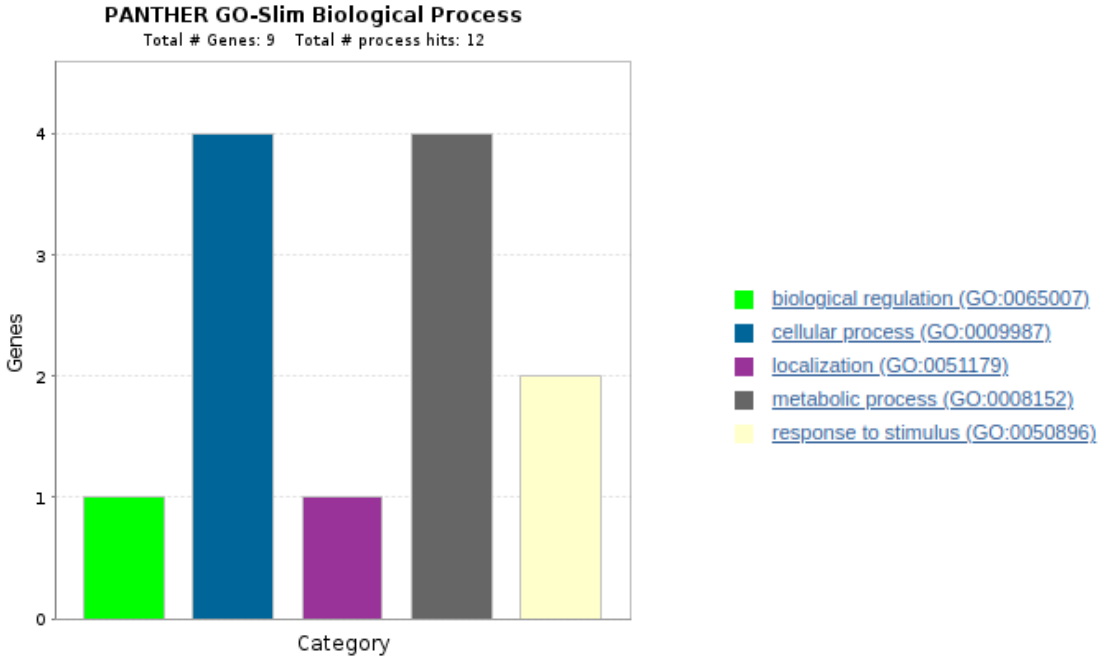

Figure 4: Biological processes of genes in "Treatment vs outcome" signature

The top three processes were: "cellular process" with 4 out of 9 genes involved, "metabolic process" with 4 genes, and "response to stimulus" with 2 genes (note that some genes are involved in several biological processes).

Details of the genes in category “cellular process” can be found in Figure 5, where 4 genes are involved in “cellular metabolic process”.

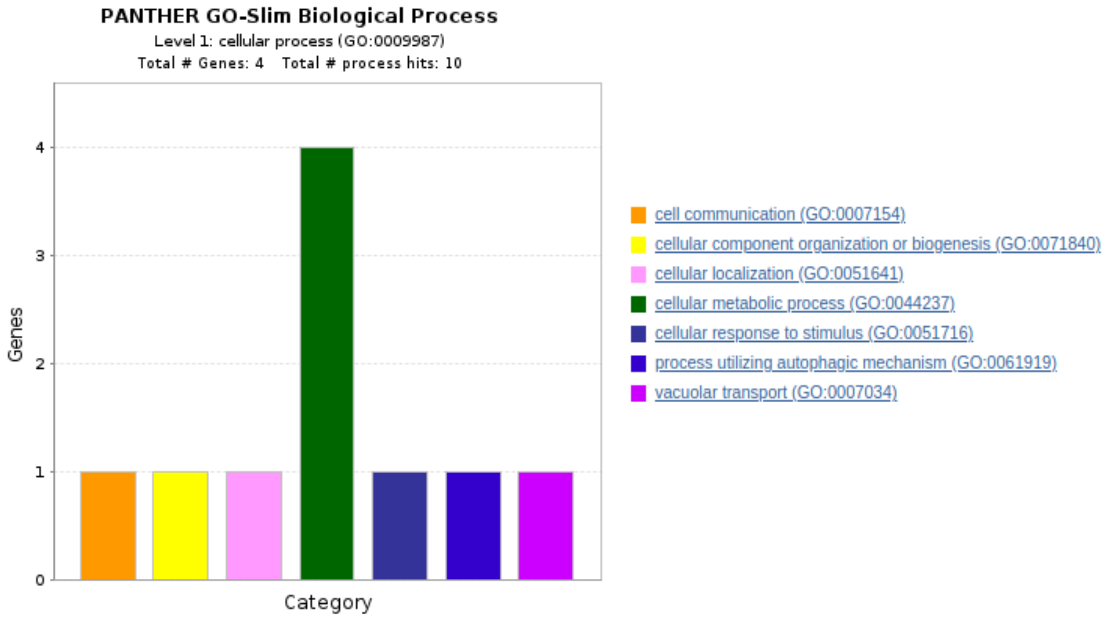

Figure 5: Details of “cellular process” category

In this other Figure 6, the details of category “metabolic process” are displayed. 4 genes participate in “cellular metabolic process”, and 3 genes are involved in “nitrogen compound metabolic process”, “organic substance metabolic process” and “primary metabolic process”.

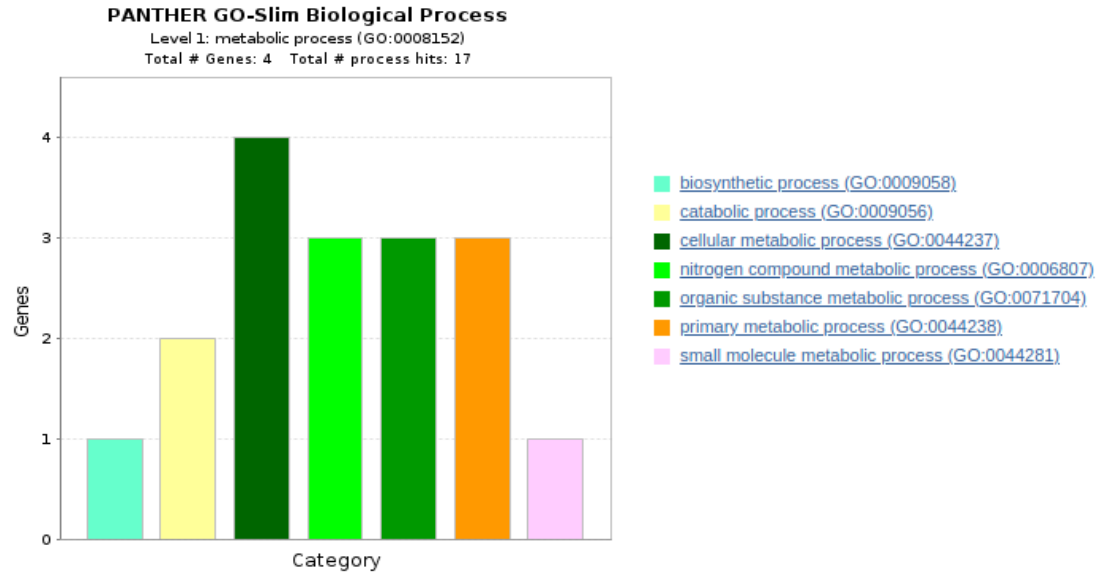

Figure 6: Details of "metabolic process" category

4.3.2 Gene expression vs outcome signature

The biological processes of the 17 genes in this signature (see annex Gene expression vs outcome signature) were examined and summarised in Figure 7:

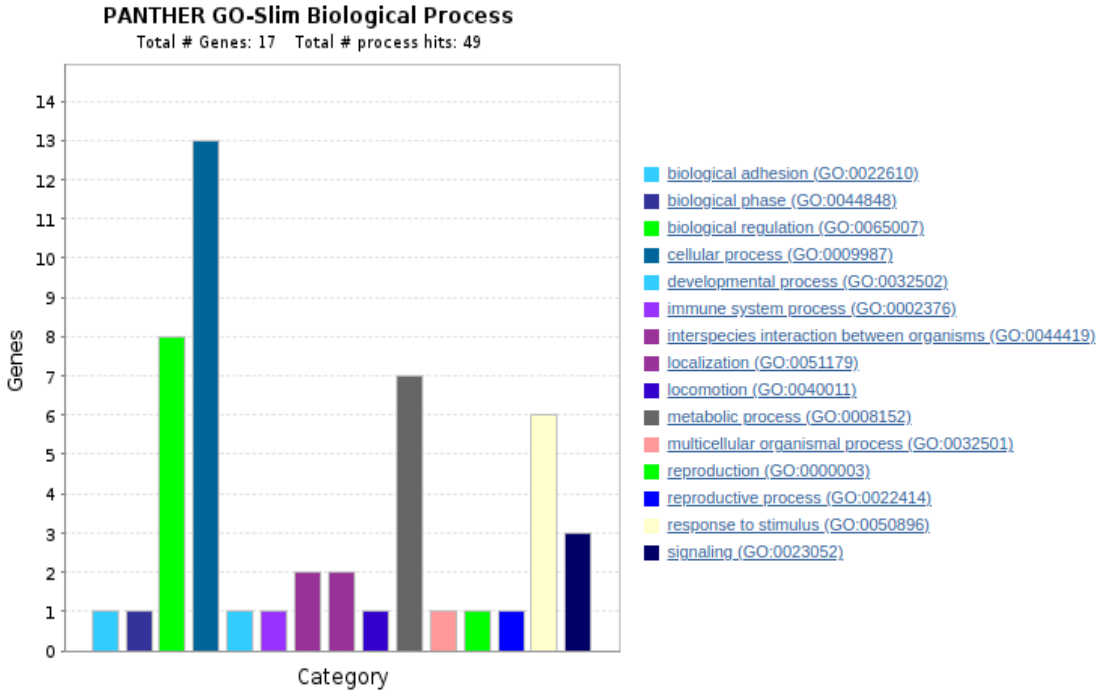

Figure 7: Biological processes of genes in "Gene expression vs outcome" signature

The top three processes were: "cellular process" with 13 out of 17 genes involved, "biological regulation" with genes, and "metabolic process" with 7 genes (note that some genes are involved in several biological processes).

In the next figure Figure 8 the details of the processes in “cellular process” category were examined. 7 genes fell in the category of “cellular metabolic process”, while 3 genes were related to “cell communication” and “cellular response to stimulus”, respectively.

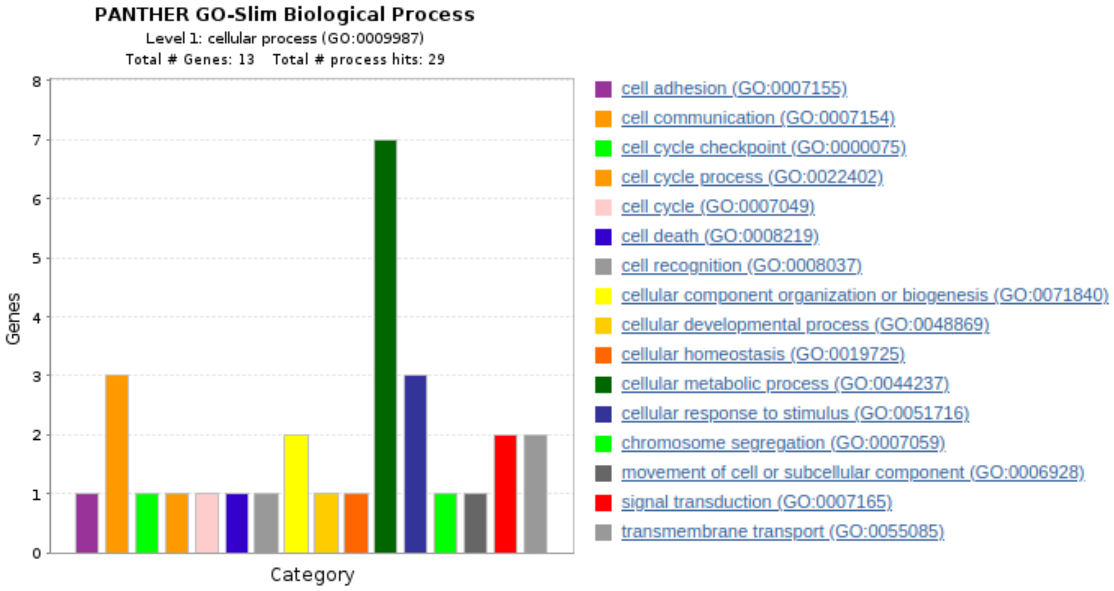

Figure 8: Details of "cellular process" category

If further details are requested for category “cellular metabolic process”, 6 genes are classified into “cellular macromolecule metabolic process”, as displayed in Figure 9.

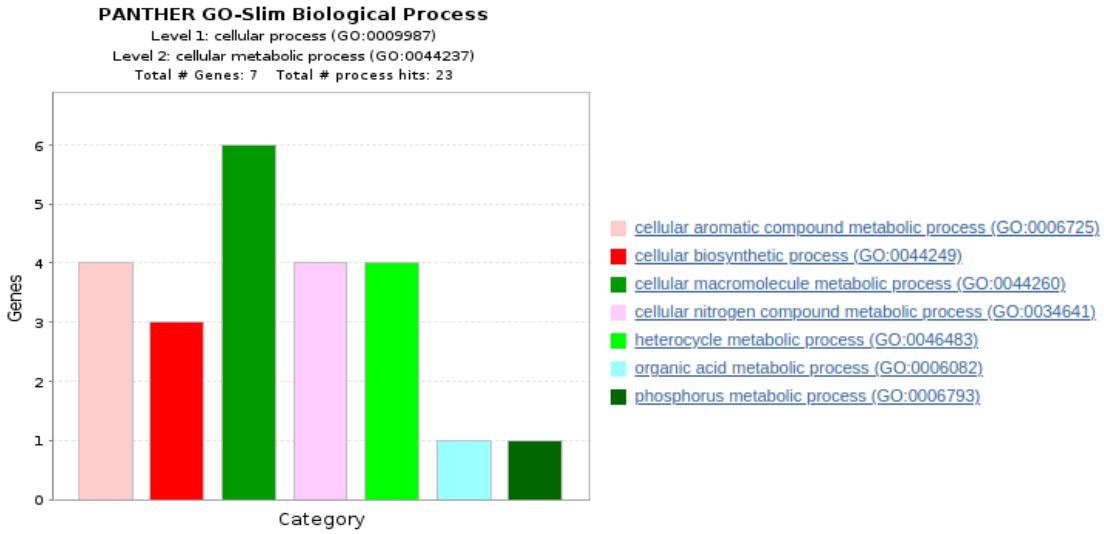

Figure 9: Details of "cellular metabolic process" category

4.3.3 Common signature

The biological processes of the 17 genes in this signature (see annex Common signature) were examined and summarised in Figure 10:

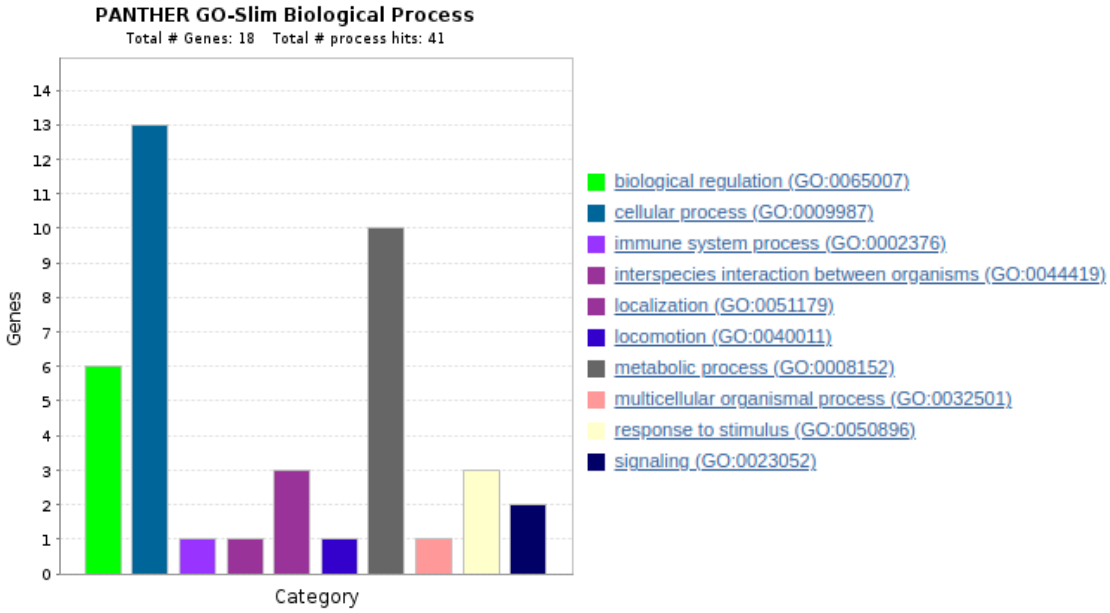

Figure 10: Biological processes of genes in common signature

The top three processes were: “cellular process” with 13 out of 17 genes involved, “metabolic process” with 10 genes, and “biological regulation” with 6 genes (note that some genes are involved in several biological processes).

Requesting further details about the largest category, “cellular process”, genes were classified into the following categories: “cellular metabolic processes” (8 genes out of 13), “cellular component organization or biogenesis” (3 genes), and the rest of categories comprised 1 or 2 genes, like “signal transduction” (2 genes), “vesicle-mediated transport” (2 genes), “cellular localization” (1 gene), etc. All these categories are presented in Figure 11:

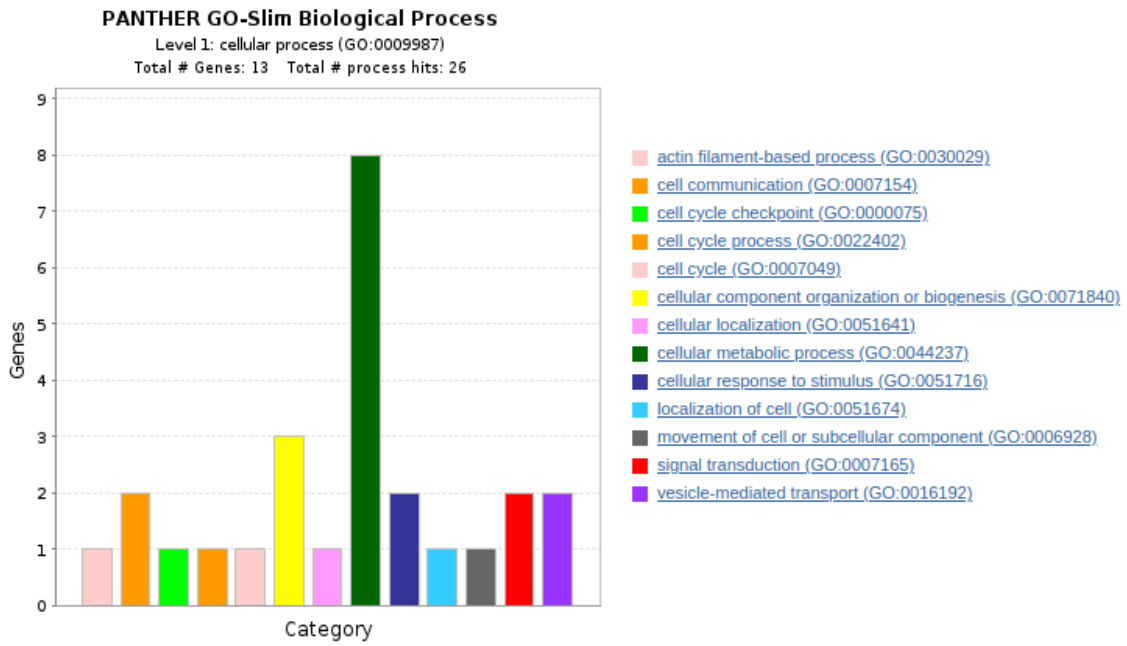

Figure 11: Details of "cellular process" category

The details about the genes in "cellular metabolic process" can be checked in Figure 12, where 5 out of 8 genes were involved in "cellular nitrogen compound metabolic process", and 4 out 8 genes participated in "cellular aromatic compound metabolic process", "cellular biosynthetic process", and "heterocycle metabolic process".

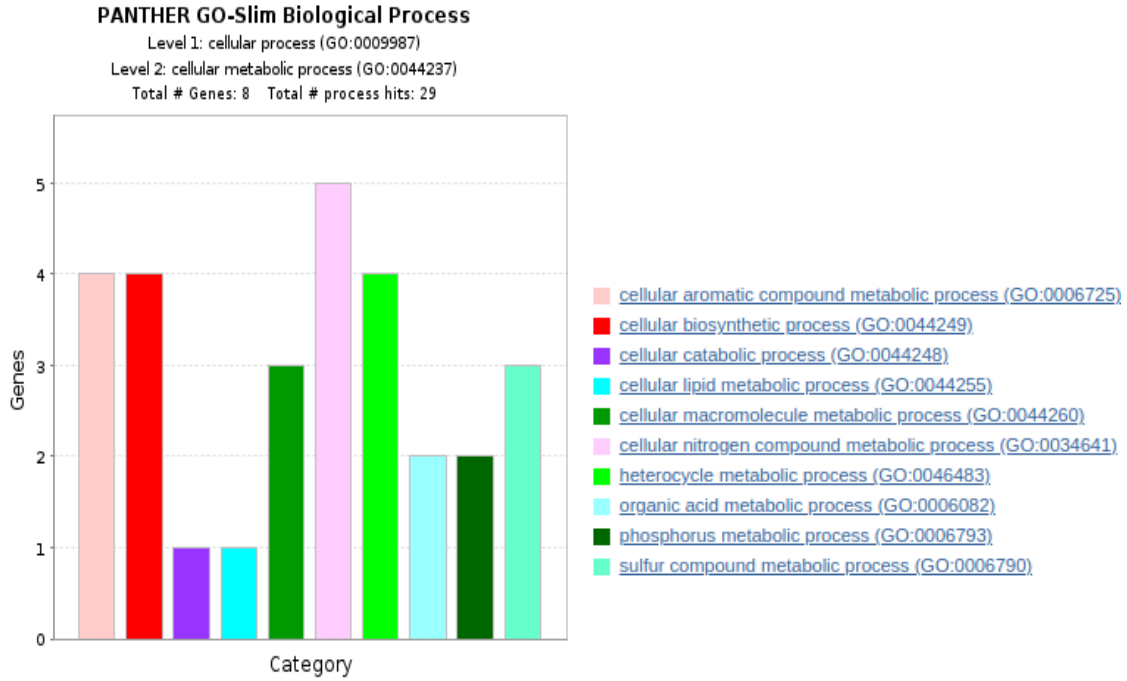

Figure 12: Details of "cellular metabolic process" category

In the next Figure 13, the details regarding “metabolic process” are displayed. The 10 genes are classified into 7 categories, with some of them falling into more than one category. For example, “organic substance metabolic process” concentrates 10 genes, but “primary metabolic process”, “nitrogen compound metabolic process” and “cellular metabolic process” have 8 genes each.

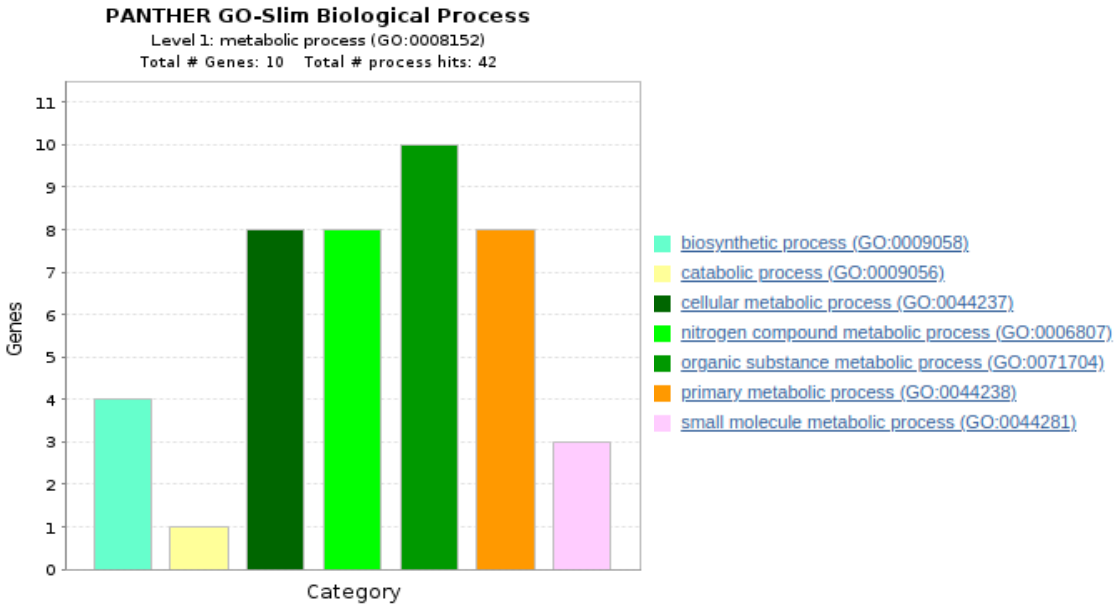

Figure 13: Details of “metabolic process” category

4.4 Evaluation of signatures as predictive models for prognosis using machine learning

After getting the three signatures, they needed to be tested with new data in order to assess their predictive power. Therefore, the new data had to include gene expression data for a high number of genes, provide some prognostic parameters, and a significant number of samples.

In cBioPortal for Cancer Genomics a study was found which fulfilled this requirements. The following Table 9 summarises the information available:

Pancreatic Adenocarcinoma (TCGA, PanCancer Atlas)	
Cancer type	Pancreatic Adenocarcinoma
Samples	177 (out of 184)
Genes	20531
Prognostic parameters	OS_STATUS (living/deceased) OS_MONTHS DSS_STATUS (alive or dead tumor free/dead with tumor) DSS_MONTHS DFS_STATUS (disease free/recurred or progressed) DFS_MONTHS PFS_STATUS (censored/progression) PFS_MONTHS
Demographic parameters	age, sex, race, ethnicity
Clinical parameters	stage[12]

Table 9: Summary of the selected study in cBioPortal

12 Neoplasm Disease Stage American Joint Committee on Cancer Code

4.4.1 Model selection

The objective is to assess the predictive power of the signatures, therefore supervised algorithms seemed a valid choice. As shown in the previous section, the are numerical and categorical variables, allowing to choose between a regression or a classification model, depending on the data type of the variable to predict.

In the following sections, the available variables are analysed in different ways in order to choose the best algorithm for this study.

4.4.1.1 Analysis of missing values

The following Table 10 shows the result of analysing the number of missing values for the prognostic parameters available:

OS_STATUS	OS_MONTHS	DSS_STATUS	DSS_MONTHS
0	0	7	0

DFS_STATUS	DFS_MONTHS	PFS_STATUS	PFS_MONTHS
115	115	0	0

Table 10: Analysis of missing values

Considering data is missing for most of the samples (115 out of 177), variables DSS_STATUS and DFS_MONTHS were dropped from further analysis.

On the other hand, DFS_STATUS was kept as only around 12% of the samples (7 out of 177) were missing this value.

4.4.1.2 Correlation analysis

This section presents all the correlation analyses performed using the three different signatures (see annex Gene signatures for further information).

The objective of these analyses is to find what prognostic variables are better candidates to be used in the machine learning phase.

Treatment vs outcome signature

A correlation analysis using the signature obtained for this group (see annex Treatment vs outcome signature) was performed and represented in a graphical way which is shown in Figure 14:

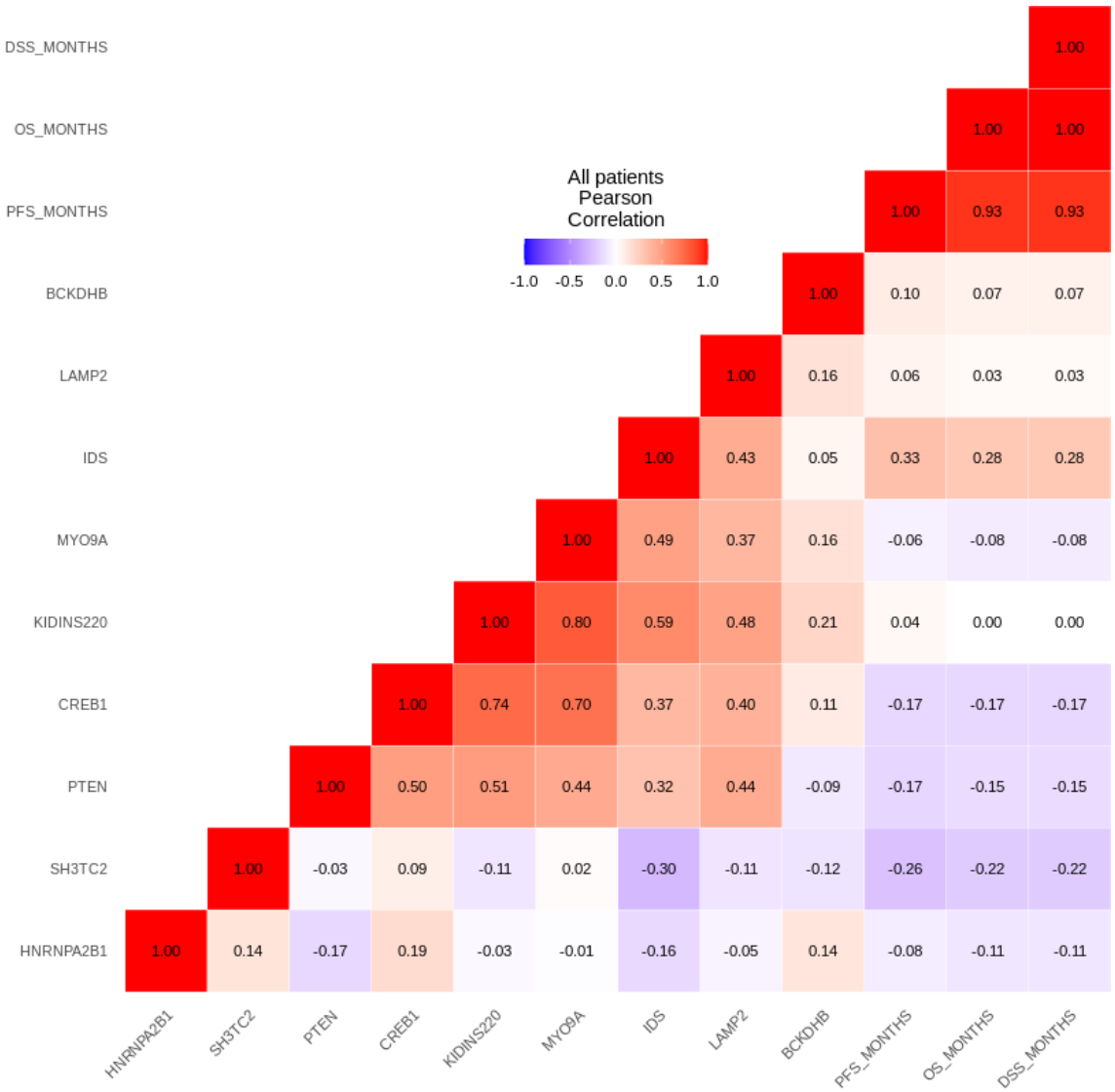

Figure 14: Heat map of the correlation matrix using the signature of "Treatment vs outcome" and all the samples available

There is a higher (negative in most cases) correlation with PFS_MONTHS than OS_MONTHS or DSS_MONTHS. This correlation is specially high among the negative correlations for SH3TC2. Among the positive, highest correlation is presented for IDS.

This plot has been drawn using the whole dataset, which includes dead and alive patients. The proportion of each status is shown in Table 11:

Living	Deceased
85	99

Table 11: Proportion of alive and dead patients

A second correlation analysis was performed using the fraction of deceased patients, to assess if there was any difference in the correlation values using this subset.

In the following figure Figure 15, the new correlation matrix was plotted:

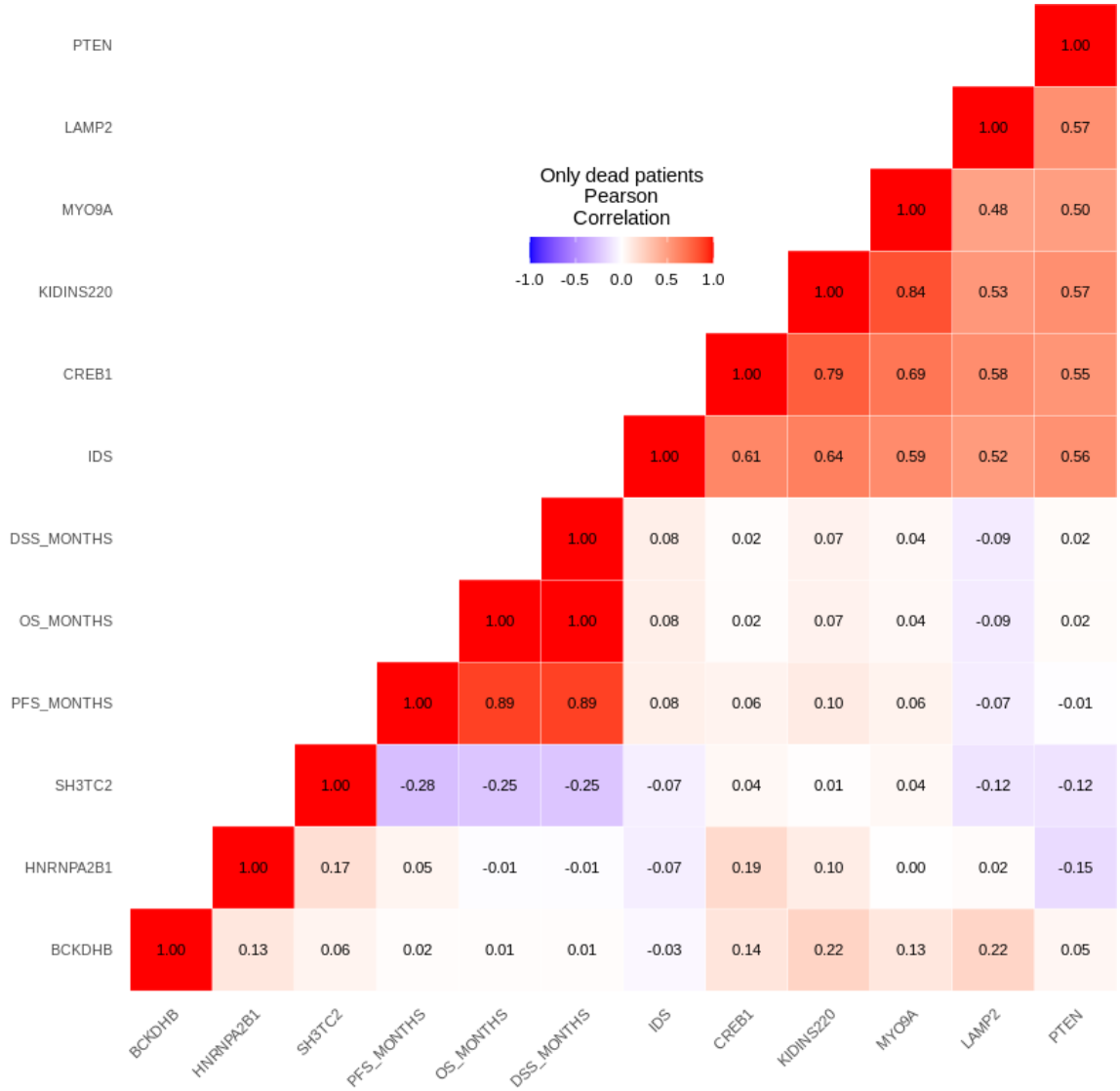

Figure 15: Heat map of the correlation matrix using the signature of "Treatment vs outcome" and samples corresponding to deceased patients

This new plot shows that correlation between PFS_MONTHS and SH3TC2 is slightly stronger (-0.28 vs -0.26), but correlations with IDS are largely weaker (0.08 vs 0.33).

Gene expression vs outcome signature

A correlation analysis using the signature obtained for this group (see annex Gene expression vs outcome signature) was performed and represented in a graphical way which is shown in Figure 16:

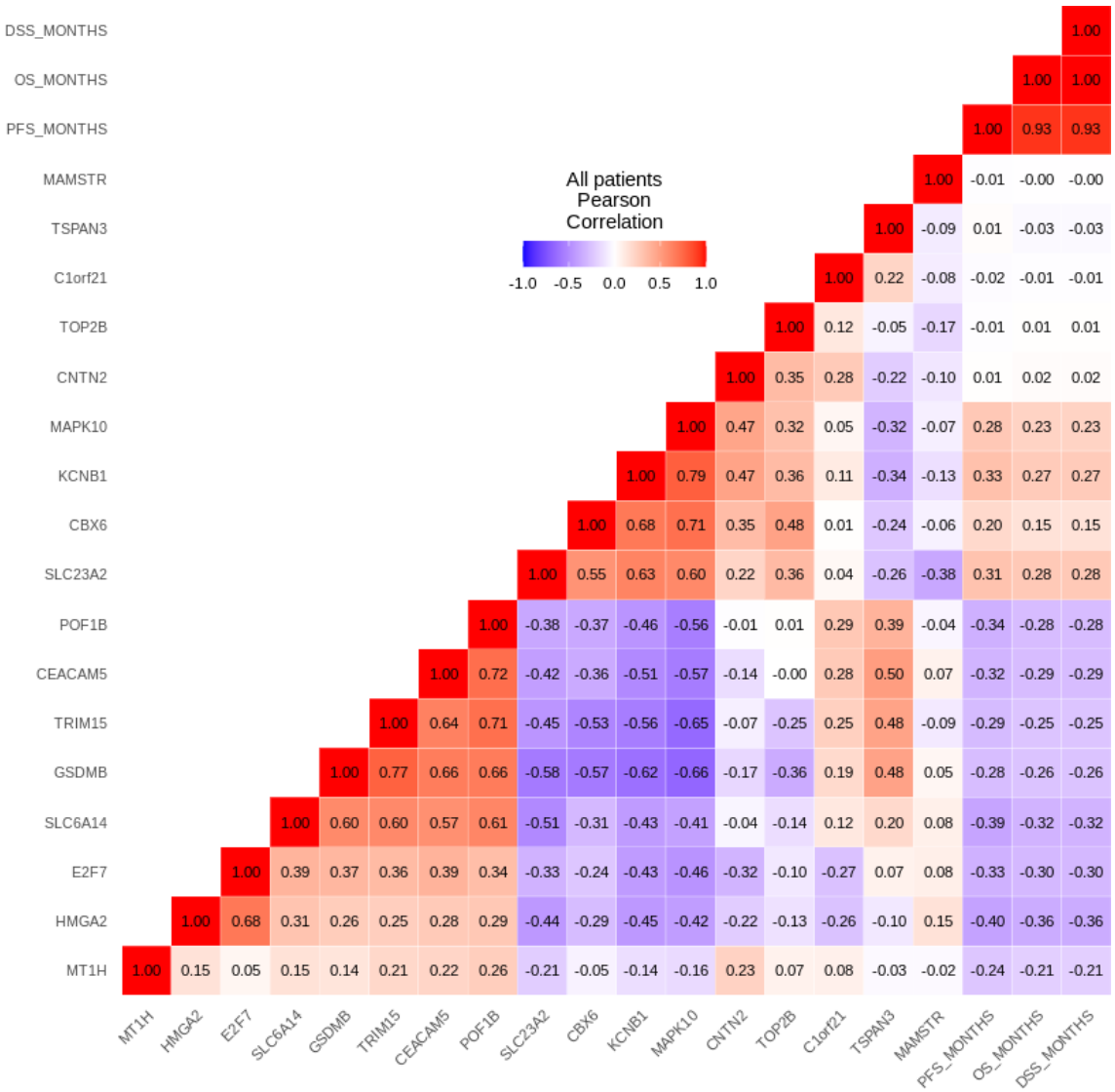

Figure 16: Heat map of the correlation matrix using the signature of "Gene expression vs outcome" and all the samples available

There is a higher (negative in most cases) correlation with PFS_MONTHS than OS_MONTHS or DSS_MONTHS. This correlation is specially high among the negative correlations for HMGA2 (-0.4), SLC6A14 (-0.39) and POF1B (-0.34). Among the positive, highest correlations are presented for KCNB1 (0.33) and SLC23A2 (0.31).

In the following figure Figure 17, a new correlation matrix was plotted using only the fraction of deceased patients, as explained in the previous section:

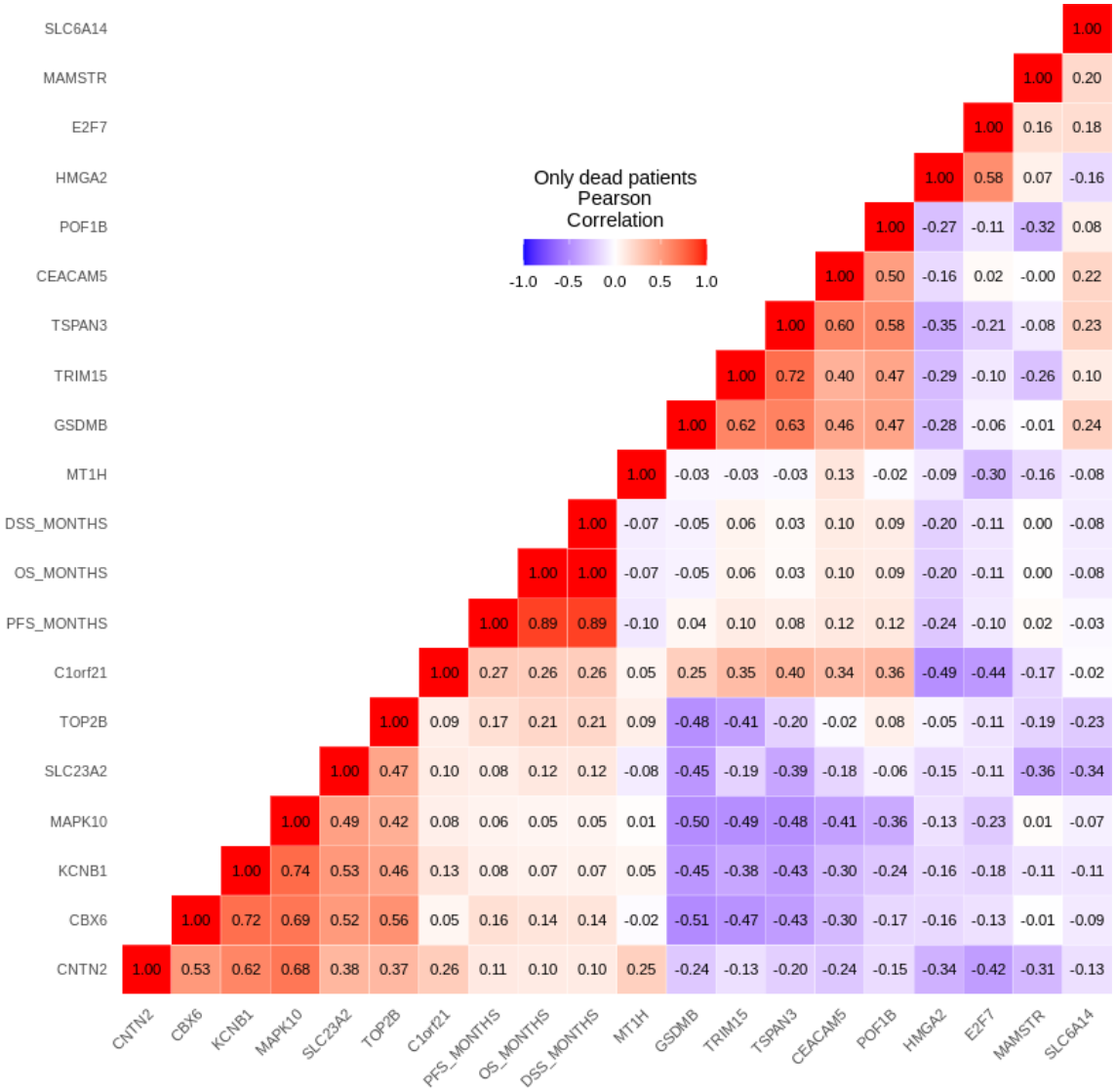

Figure 17: Heat map of the correlation matrix using the signature of "Gene expression vs outcome" and samples corresponding to deceased patients

In this new plot, correlations with the prognostic variables were weaker than the observed values when using the whole dataset (e.g. -0.24 vs -0.4 for HMGA2 and PFS_MONTHS).

Common signature

A correlation analysis using the final signature common to all groups (see annex Common signature) was performed and represented in a graphical way which is shown in Figure 18:

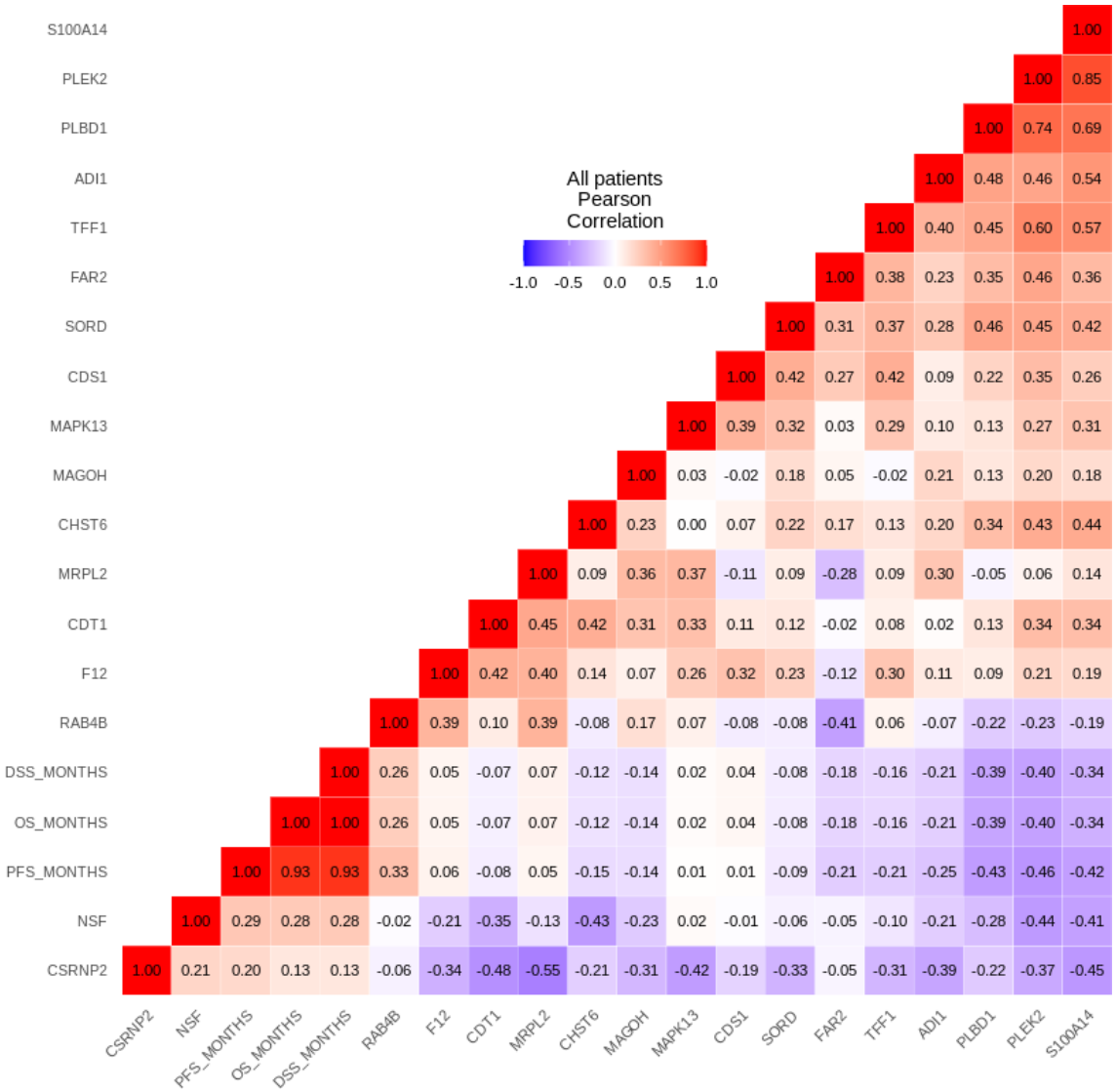

Figure 18: Heat map of the correlation matrix using the common signature and all the samples available

There is a higher (negative in most cases) correlation with PFS_MONTHS than OS_MONTHS or DSS_MONTHS. This correlation is specially high among the negative for PLEK2 (-0.4), PLBD1 (-0.39) and S100A14 (-0.34). Among the positive, highest correlations are presented for NSF (0.29) and RAB4B (0.26).

In the following Figure 19, a new correlation matrix was plotted using only the fraction of deceased patients, as explained in the previous sections:

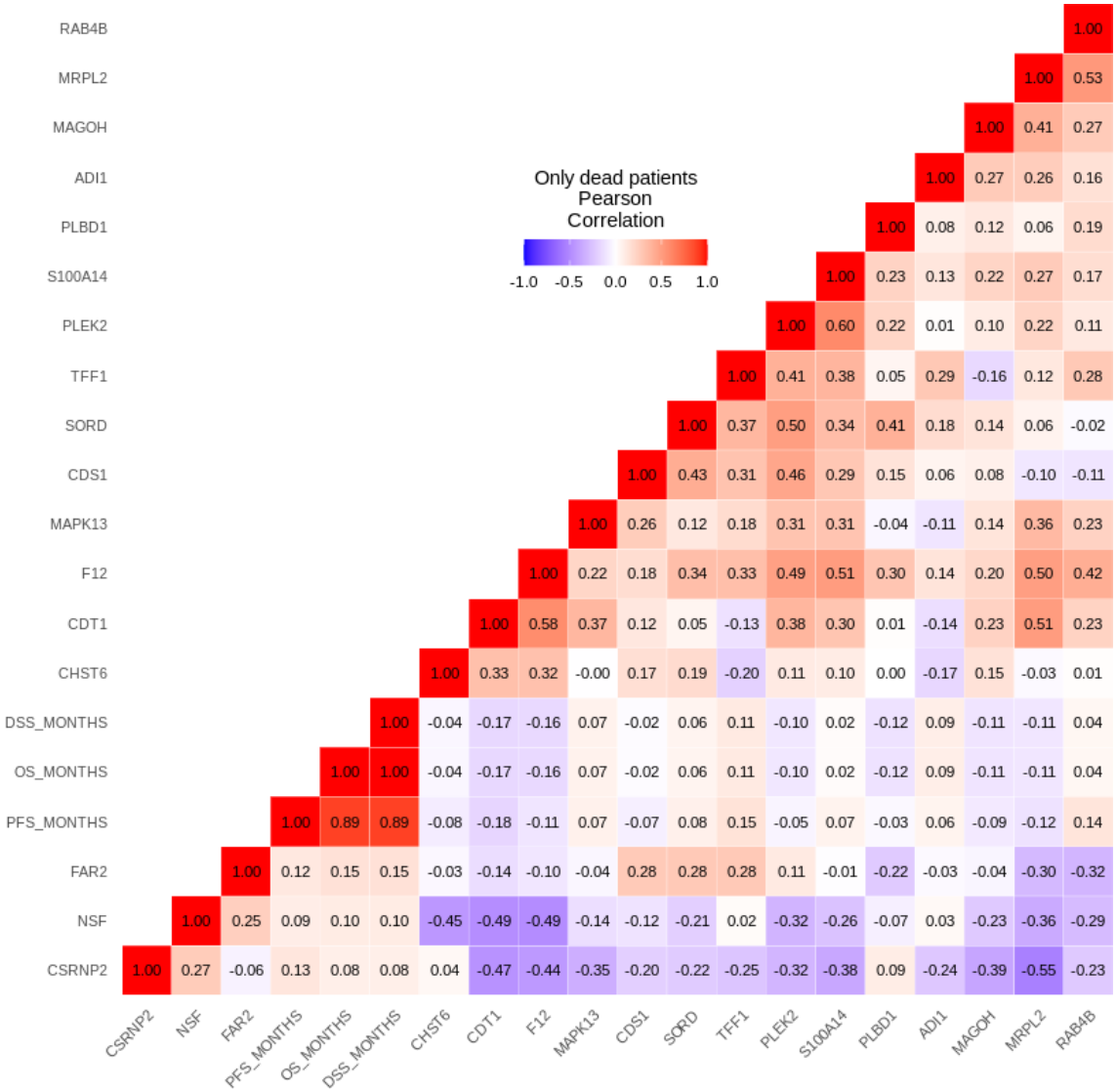

Figure 19: Heat map of the correlation matrix using the common signature and samples corresponding to deceased patients

In this new plot, correlations with the prognostic variables were largely weaker than the observed values when using the whole dataset (e.g. -0.05 vs -0.4 for PLEK2 and PFS_MONTHS).

Results

Considering the results of all these analyses, all samples were used in the following sections. Also, a classification algorithm seemed a good candidate, and Random Forest was selected.

In order to split the samples in two classes to be predicted by the model, the distribution of values of PFS_MONTHS was further analysed, as shown in Figure 20:

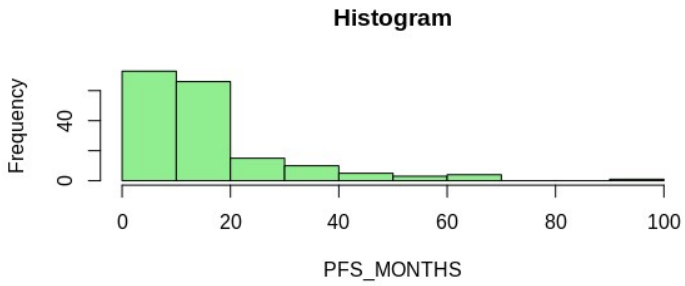

Figure 20: Histogram of PFS_MONTHS

According to this and the median 12.03274, data was split into the following two groups:

- X0 or Good progression. Those patients with PFS_MONTHS >= 12.
- X1 or Bad progression. Those patients with PFS_MONTHS < 12.

4.4.2 Feature selection

In the following sections, the three signatures generated were analysed separately, in order to find what genes could be included in the model, depending on factors like the absence of values or the high/low correlation between some genes.

4.4.2.1 Analysis of missing values

Treatment vs outcome signature

Considering this signature (see annex Treatment vs outcome signature), it is analysed if there are missing values for any of these genes. The result is shown in Table 12:

BCKDHB	CREB1	HNRNPA2B1	IDS	KIDINS220
0	0	0	0	0

LAMP2	MYO9A	PTEN	SH3TC2
0	0	0	0

Table 12: Missing values in "Treatment vs outcome" signature

Gene expression vs outcome signature

Considering this signature (see annex Gene expression vs outcome signature), it is analysed if there are missing values for any of these genes. The result is shown in Table 13:

HMGA2	C1orf21	CBX6	CEACAM5	CNTN2	E2F7
0	0	0	0	0	0

GSDMB	KCNB1	MAMSTR	MAPK10	MT1H	POF1B
0	0	0	0	0	0

SLC23A2	SLC6A14	TOP2B	TRIM15	TSPAN3
0	0	0	0	0

Table 13: Missing values in "Gene expression vs outcome" signature

Common signature

Considering this signature (see annex Common signature), it is analysed if there are missing values for any of these genes. The result is shown in Table 14:

ADI1	CDS1	CDT1	CHST6	CSRNP2	F12
0	0	0	0	0	0

FAR2	MAGOH	MAPK13	MRPL2	NSF	PLBD1
0	0	0	0	0	0

PLEK2	RAB4B	S100A14	SORD	TFF1
0	0	0	0	0

Table 14: Missing values in common signature

Results

None of the signatures presented missing values for any of the genes, so no other gene was excluded.

4.4.2.2 Correlation analysis

Treatment vs outcome signature

As previous step Analysis of missing values did not discard any gene, the correlation matrix is the same shown in Figure 14.

Correlation with PFS_MONTHS

First, the correlation values of the genes with the prognostic variable PFS_MONTHS was analysed. Values comprised a range between 0.04 and 0.33 among the positive correlations, and between -0.06 and -0.26 among the negative.

Weakest correlations with PFS_MONTHS were discarded using a cut-off of 10%.

Correlations between genes

Second, the correlation values between genes was examined in order to find too strong correlations among them. The correlation matrix after discarding those genes with a too low correlation with PFS_MONTHS is showed in Figure 21:

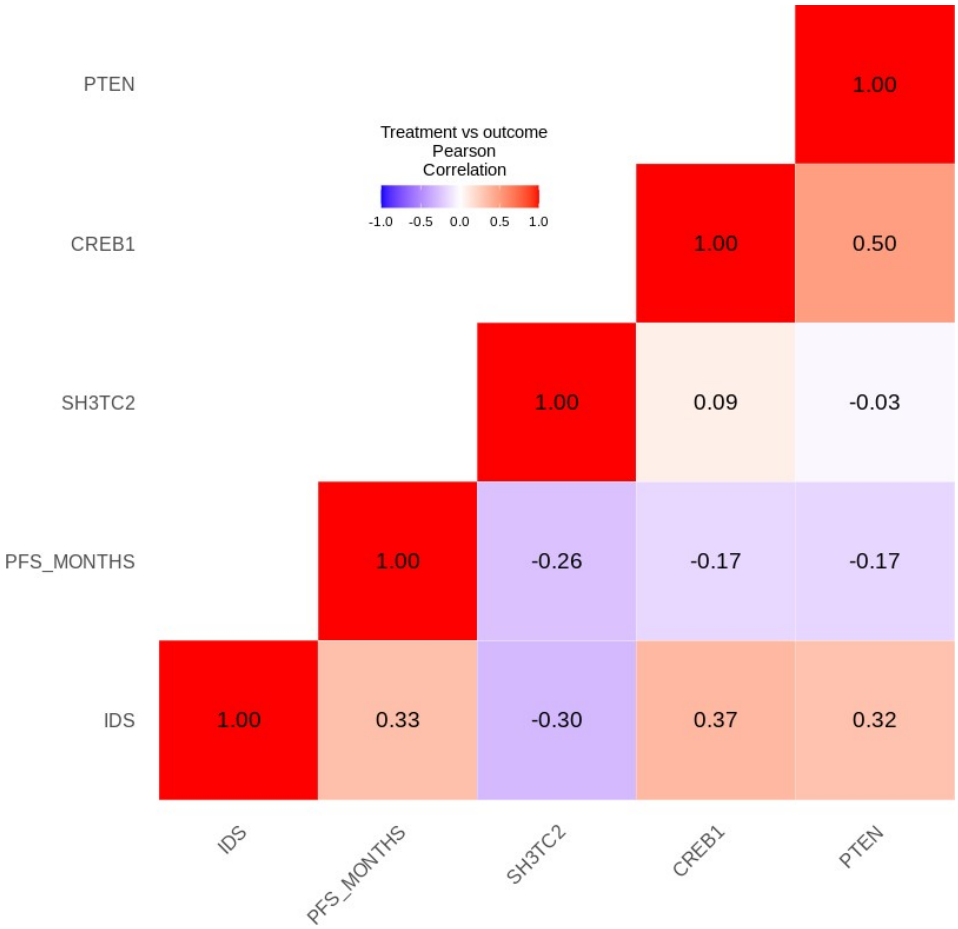

Figure 21: Correlation matrix with the subset of "Treatment vs outcome" signature

Observing the values in the figure, the highest positive correlation (0.5) could be found between CREB1 and PTEN, and the highest negative (-0.3) between IDS and SH3TC2.

As there is no correlation higher than 90%, no other gene was discarded. Therefore, the final signature comprised the following 4 genes:

CREB1	PTEN
IDS	SH3TC2

Gene expression vs outcome signature

As previous step Analysis of missing values did not discard any gene, the correlation matrix is the same shown in Figure 16.

Correlation with PFS_MONTHS

First, the correlation values of the genes with the prognostic variable PFS_MONTHS was analysed. Values comprised a range between 0.01 and 0.33 among the positive correlations, and between -0.01 and -0.40 among the negative.

Weakest correlations with PFS_MONTHS were discarded using a cut-off of 10%.

Correlations between genes

Second, the correlation values between genes was examined in order to find too strong correlations among them. The correlation matrix after discarding those genes with a too low correlation with PFS_MONTHS is showed in Figure 22:

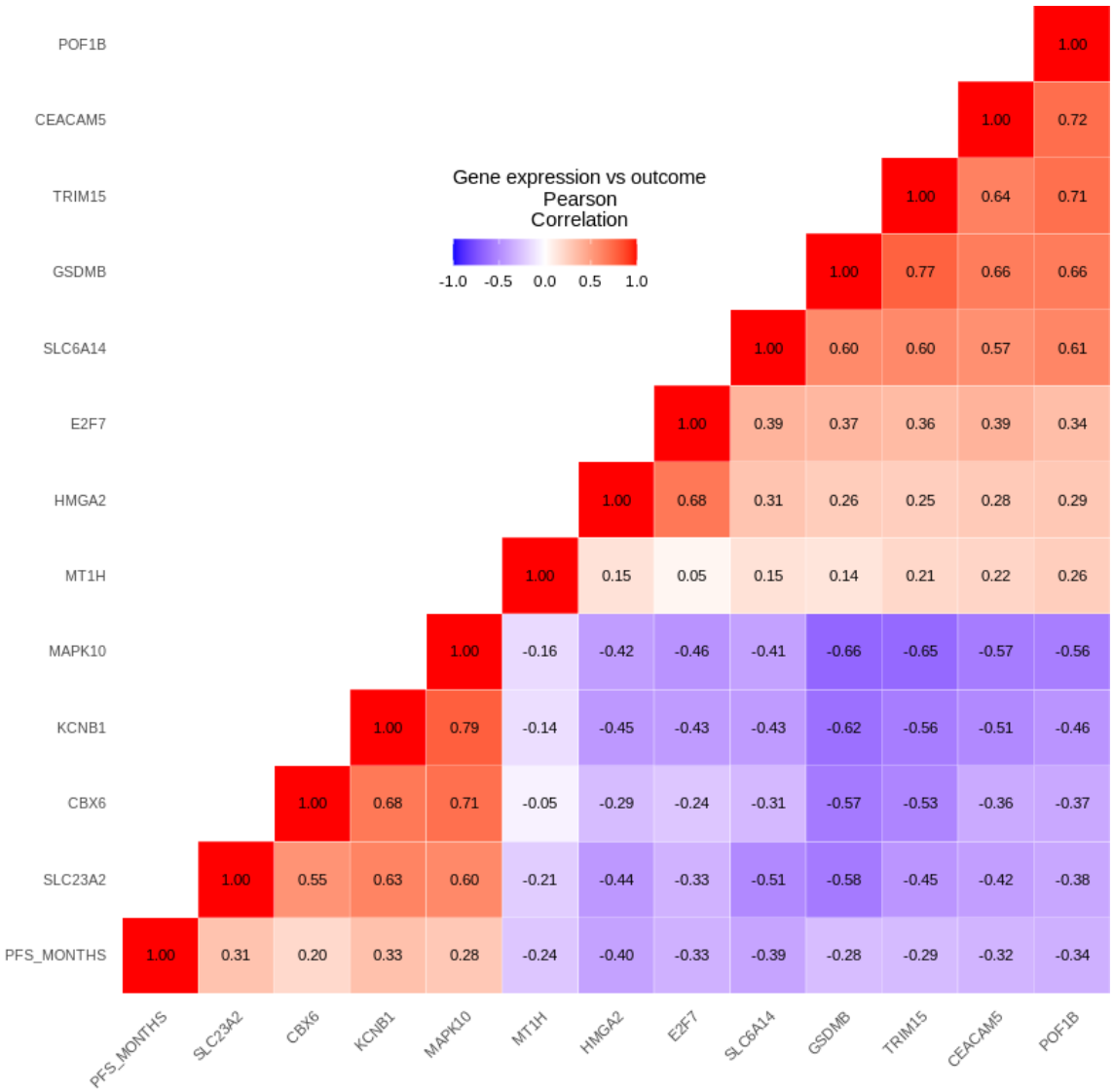

Figure 22: Correlation matrix with the subset of "Gene expression vs outcome" signature

Observing the values in the figure, the highest positive correlation (0.79) could be found between KCNB1 and MAPK10, and the highest negative (-0.66) between MAPK10 and GSDMB.

As there is no correlation higher than 90%, no other gene was discarded. Therefore, the final signature comprised the following 12 genes:

HMGA2	E2F7	MAPK10	SLC23A2
CBX6	GSDMB	MT1H	SLC6A14
CEACAM5	KCNB1	POF1B	TRIM15

Common signature

As previous step Analysis of missing values did not discard any gene, the correlation matrix is the same shown in Figure 18.

Correlation with PFS_MONTHS

First, the correlation values of the genes with the prognostic variable PFS_MONTHS was analysed. Values comprised a range between 0.01 and 0.33 among the positive correlations, and between -0.08 and -0.46 among the negative.

Weakest correlations with PFS_MONTHS were discarded using a cut-off of 10%.

Correlations between genes

Second, the correlation values between genes was examined in order to find too strong correlations among them. The correlation matrix after discarding those genes with a too low correlation with PFS_MONTHS is showed in Figure 23:

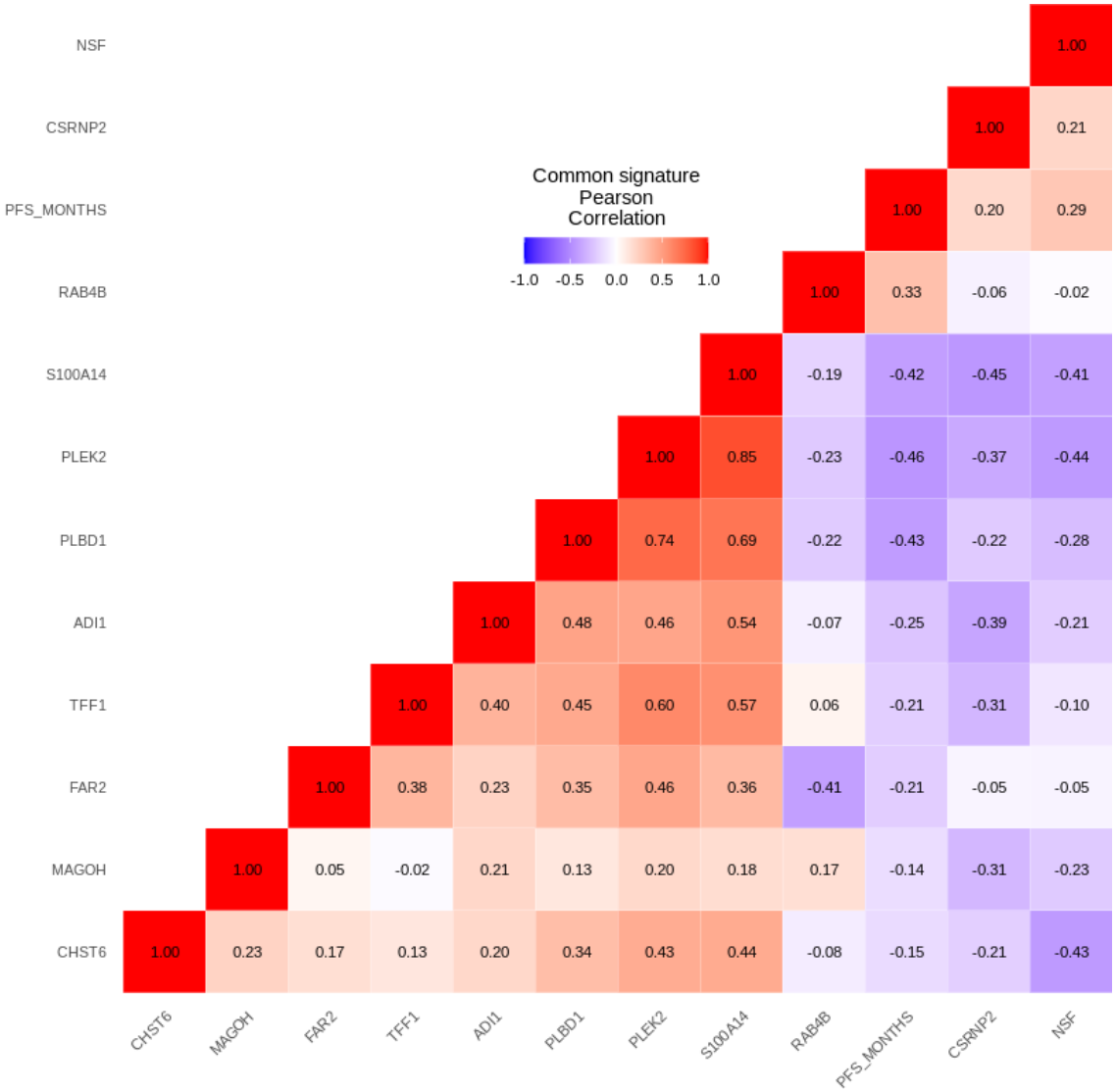

Figure 23: Correlation matrix with a subset of the common signature

Observing the values in the figure, the highest positive correlation (0.85) could be found between PLEK2 and S100A14, and the highest negative (-0.45) between S100A14 and CSRNP2.

As there is no correlation higher than 90%, no other gene was discarded. Therefore, the final signature comprised the following 11 genes:

ADI1	FAR2	PLBD1	S100A14
CHST6	MAGOH	PLEK2	TFF1
CSRNP2	NSF	RAB4B	

4.4.3 Evaluation of the Random forest model

In this section, the results from training and evaluating different models with the different signatures produced in the previous section Feature selection, are presented.

Specific details about the procedure followed can be found in annex ML model training and evaluation. Although different combinations of parameters were tested (e.g. *mtry* and *ntree*), only the best models are showed here.

Finally, sizes of train and test sets used in the next subsections are displayed in Table 15:

	Train	Test
60/40	106	71
70/30	124	53
80/20	142	35
90/10	159	18

Table 15: Sizes of train and test sets

Treatment vs outcome signature

Results of training and evaluating different models using this signature is showed in Table 16:

Training/Test size	mtry	ntree	Accuracy	Sensitivity	Specificity
60/40	4	2000	0.45	0.5	0.4
70/30	3	800	0.42	0.52	0.3
80/20	4	400	0.4	0.39	0.41
90/10	2	250	0.44	0.55	0.33

Table 16: Models' metrics using "Treatment vs outcome" signature

Confusion matrix of the best model (60/40) is presented in Table 17 below:

Prediction / Reference	X0	X1
X0	18	21
X1	18	14

Table 17: Confusion matrix of 60/40 model

Also, the importance of each predictor in this model is plotted in Figure 24:

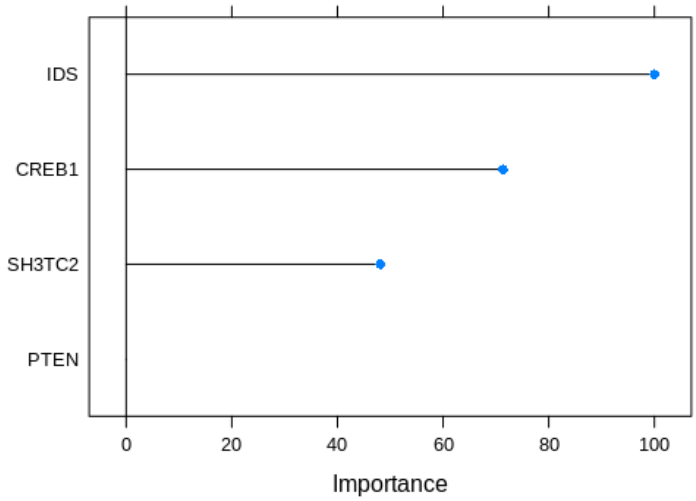

Figure 24: Variable importance of 60/40 model

The variable with major importance is IDS, which is consistent with the correlation analysis (see Treatment vs outcome signature) were IDS was the gene presenting the highest correlation value with prognostic variable.

Gene expression vs outcome signature

Results of training and evaluating different models using this signature is showed in Table 18:

Training/Test size	mtry	ntree	Accuracy	Sensitivity	Specificity
60/40	10	600	0.4	0.36	0.46
70/30	4	550	0.58	0.7	0.46
80/20	2	500	0.49	0.61	0.35
90/10	5	2000	0.97	0.94	1

Table 18: Models metrics using "Gene expression vs outcome" signature

Confusion matrix of the best model (70/30) is presented in Table 19 below:

Prediction / Reference	X0	X1
X0	19	14
X1	8	12

Table 19: Confusion matrix of 70/30 model

Also, the importance of each predictor in this model is plotted in Figure 25:

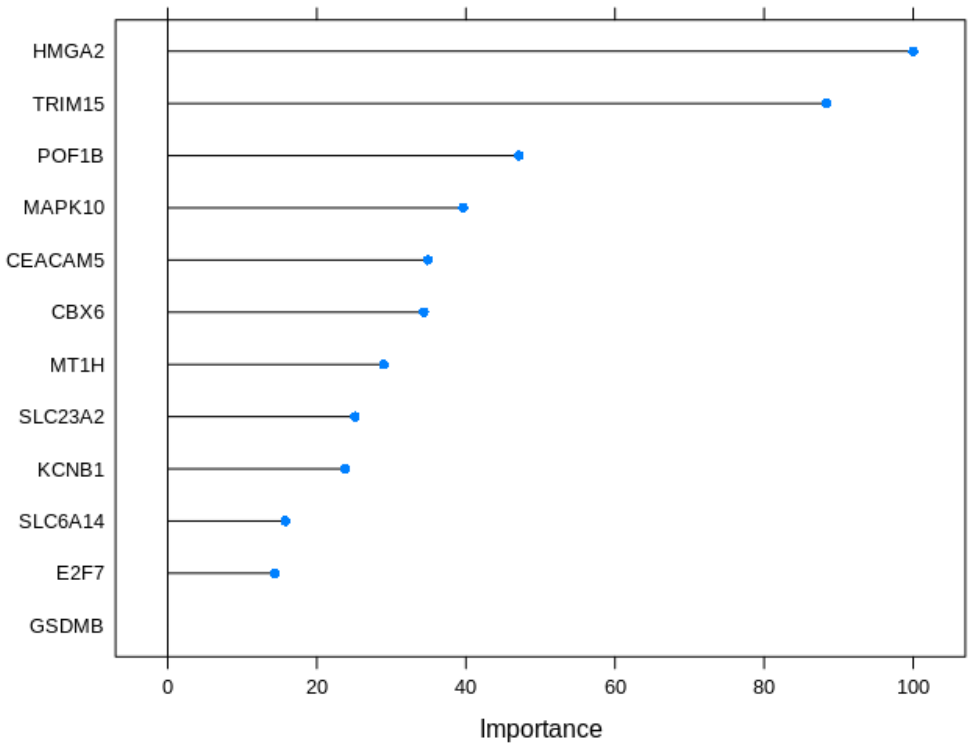

Figure 25: Variable importance of 70/30 model

The variables with major importance are HMGA2 and TRIM15, followed by POF1B, which is consistent with the correlation analysis (see Gene expression vs outcome signature) were HMGA2 was the gene presenting the highest correlation value with prognostic variable (-0.4), and POF1B was the third one (-0.34).

Model 90/10 showed an awesome accuracy, but it cannot be considered as the best without further validation due to the small size of the test set. This model should be tested against new data with a larger size in order to validate its performance.

However, its confusion matrix is presented in Table 20 below:

Prediction / Reference	X0	X1
X0	17	0
X1	1	17

Table 20: Confusion matrix of 90/10 model

Also, the importance of each predictor in this model is plotted in Figure 26:

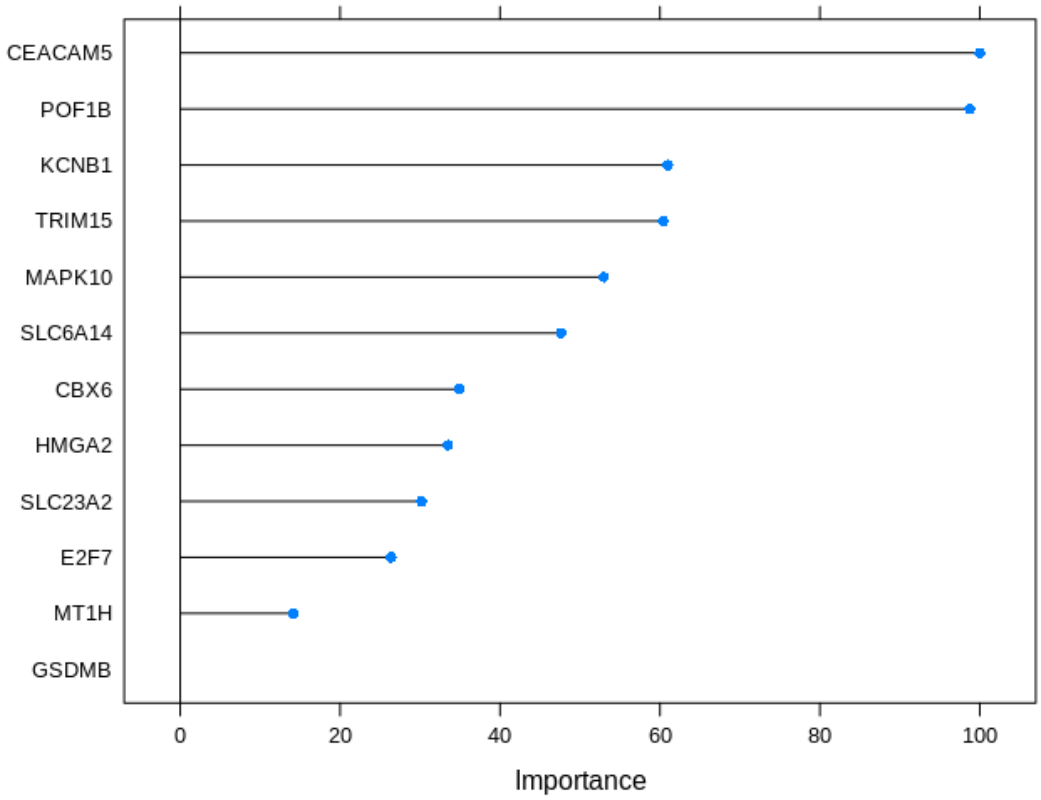

Figure 26: Variable importance of 90/10 model

The variables with major importance are CEACAM5 and POF1B, followed by KCNB1 and TRIM15.

Common signature

Results of training and evaluating different models using this signature is showed in Table 21:

Training/Test size	mtry	ntree	Accuracy	Sensitivity	Specificity
60/40	11	1000	0.60	0.56	0.66
70/30	10	800	0.62	0.59	0.65
80/20	10	450	0.57	0.61	0.53
90/10	7	2000	0.5	0.33	0.67

Table 21: Models metrics using the common signature

Confusion matrix of the best model (70/30) is presented in Table 22 below:

Prediction / Reference	X0	X1
X0	16	9
X1	11	17

Table 22: Confusion matrix of 70/30 model

Also, the importance of each predictor in this model is plotted in Figure 27:

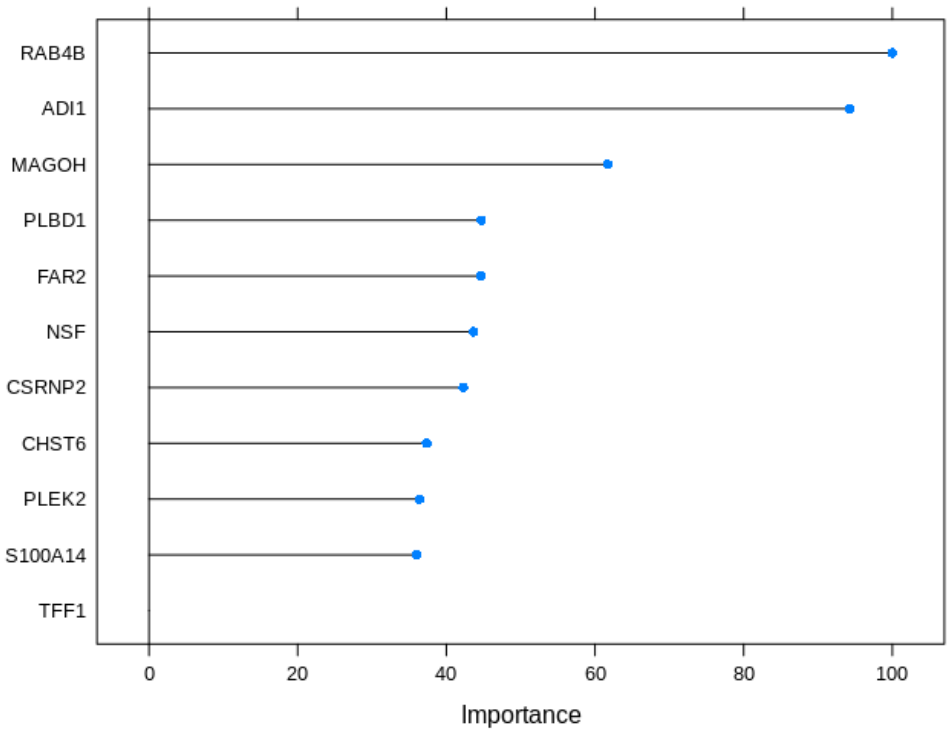

Figure 27: Variable importance of 70/30 model

The variables with major importance are RAB4B and ADI1, followed by MAGOH, which is consistent with the correlation analysis (see Common signature) were RAB4B was the gene presenting the highest positive correlation value with prognostic variable (0.33).

5 Conclusions

The impossibility of finding large datasets which included gene expression data, as well as, prognostic variables, has been a big obstacle in order to train the model because, the more data the better for any machine learning algorithm. As showed in section Evaluation of the Random forest model, the training and test sets were not large enough to train better models.

Also, during the DEG analyses of the gene expression data, it was really difficult to find differentially expressed genes among the data in the series. Many series were analysed and later discarded, making this part of the project the hardest and longest in both effort and time.

Nevertheless, this project has drawn the following conclusions:

- Up to 4 genes out of 9 included in the signature obtained after analysing the gene expression data grouped under the type “Treatment vs outcome”, showed a correlation higher than 10% with the prognostic variable progression-free survival (PFS) using Pearson coefficient. Gene showing highest correlation value was IDS (0.33).

- Up to 12 genes out of 17 included in the signature obtained after analysing the gene expression data grouped under the type “Gene expression vs outcome”, showed a correlation higher than 10% with the prognostic variable progression-free survival (PFS) using Pearson coefficient. Genes showing highest correlation values were HMGA2 (-0.4), SLC6A14 (-0.39) and KCNB1 (0.33).

- Up to 11 genes out of 17 included in the signature generated after intersecting the signature obtained from the gene expression data in “Treatment vs outcome” with the signature obtained from data in “Gene expression vs outcome”, showed a correlation higher than 10% with the prognostic variable progression-free survival (PFS) using Pearson coefficient. Genes showing highest correlation values were PLEK2 (-0.46), PLBD1 (-0.43) and S100A14 (-0.42), and RAB4B (0.33).

- The three signatures analysed, comprising 4, 12 and 11 genes, obtained an accuracy of 45%, 58% and 62%, respectively. Although none showed a significant performance to be used as a classifier, the latter could be further investigated (e.g. tested with other independent datasets).

6 Future work

There are different tasks that could be done in order to extend the work done in this project:

- Find more expression data to be included in “Treatment vs outcome” group, in order to obtain a larger signature and, at the same time, maximise the genes in common.

- Add more variables to train the model. For example, age, sex and stage could be interesting factors that could improve the prediction power of the signature(s) for prognosis in pancreatic cancer.

- Train and test other Machine Learning algorithms, like KNN.

7 Glossary

ANN, Artificial Neural Network: An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain[13].

Autoencoder: is a type of artificial neural network used to learn efficient data codings in an unsupervised manner. The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore signal “noise”[14].

cBioPortal for Cancer Genomics: is an open-access, open-source resource for interactive exploration of multidimensional cancer genomics data sets. The goal of cBioPortal is to significantly lower the barriers between complex genomic data and cancer researchers by providing rapid, intuitive, and high-quality access to molecular profiles and clinical attributes from large-scale cancer genomics projects, and therefore to empower researchers to translate these rich data sets into biologic insights and clinical applications[15].

CG, Core genes: Genes identified to be differentially expressed (DEG) among the datasets under study.

Cox regression: The log-rank test and KM curves don't work easily with quantitative predictors such as gene expression, white blood count, or age. For quantitative predictor variables, an alternative method is Cox proportional hazards regression analysis[16].

DEG, Differentially Expressed Genes: A gene is declared differentially expressed if a difference or change observed in read counts or expression levels/index between two experimental conditions is statistically significant. Transcription is the expression analysis of population of genes or analysis of differences in expression of gene populations under different environments, conditions, treatments, and stages. Several statistical methods are there for gene expression analysis. Statistical distributions are used to approximate the pattern of differential gene expression. Such genes are selected based on a combination of expression change threshold and score cutoff, which are usually generated by statistical modeling[19].

DFS, disease-free survival: The length of time after primary treatment for a cancer ends that the patient survives without any signs or symptoms of that cancer. In a clinical trial, measuring the disease-free survival is one way to see how well a new treatment works[17].

13 https://en.wikipedia.org/wiki/Artificial_neural_network
14 https://en.wikipedia.org/wiki/Autoencoder
15 https://www.cbioportal.org/
16 https://en.wikipedia.org/wiki/Survival_analysis#Cox_proportional_hazards_(PH)_regression_analysis
17 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/dfs

Gene expression: Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product that enables it to produce protein as the end product. These products are often proteins, but in non-protein-coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA[18].

GEO, Gene Expression Omnibus: GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array-and sequence-based data are accepted[19].

GES, gene expression signature, gene signature: A gene signature or gene expression signature is a single or combined group of genes in a cell with a uniquely characteristic pattern of gene expression that occurs as a result of an altered or unaltered biological process or pathogenic medical condition. This is not to be confused with the concept of gene expression profiling. Activating pathways in a regular physiological process or a physiological response to a stimulus results in a cascade of signal transduction and interactions that elicit altered levels of gene expression, which is classified as the gene signature of that physiological process or response[20].

GSEA, Gene Set Enrichment Analysis: It is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes)[21].

IFS, Incremental Feature Selection: Feature selection is a problem of finding relevant features. When the number of features of a dataset is large and its number of patterns is huge, an effective method of feature selection can help in dimensionality reduction. An incremental probabilistic algorithm is designed and implemented as an alternative to the exhaustive and heuristic approaches[20].

K-means: The k-means algorithm is an unsupervised clustering algorithm. It takes a bunch of unlabeled points and tries to group them into “k” number of clusters. It is unsupervised because the points have no external classification. The “k” in k-means denotes the number of clusters you want to have in the end[9].

Kaplan-Meier: It s a non-parametric statistic used to estimate the survival function from lifetime data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment[22].

KNN, K-Nearest Neighbour: It is a supervised classification algorithm. It takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point which are its nearest neighbors, and has those neighbors vote[23].

18 https://en.wikipedia.org/wiki/Gene_expression
19 https://www.ncbi.nlm.nih.gov/geo/
20 https://en.wikipedia.org/wiki/Gene_signature
21 https://www.gsea-msigdb.org/gsea/index.jsp
22 https://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator
23 https://becominghuman.ai/comprehending-k-means-and-knn-algorithms-c791be90883d

Kruskal-Wallis test: is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing two or more independent samples of equal or different sample sizes. It extends the Mann–Whitney U test, which is used for comparing only two groups. The parametric equivalent of the Kruskal–Wallis test is the one-way analysis of variance (ANOVA)[24].

Logistic Regression: It is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one[25].

Mann-Whitney test: is a non-parametric test of the null hypothesis that, for randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X[26].

MCFS, Monte Carlo feature selection: It is an algorithm for feature selection and attribute ranking[21].

Metastasis: The spread of cancer cells from the place where they first formed to another part of the body. In metastasis, cancer cells break away from the original (primary) tumor, travel through the blood or lymph system, and form a new tumor in other organs or tissues of the body. The new, metastatic tumor is the same type of cancer as the primary tumor. For example, if breast cancer spreads to the lung, the cancer cells in the lung are breast cancer cells, not lung cancer cells[27,28].

mRMR, Minimum Redundancy Maximum Relevance: Feature selection, one of the basic problems in pattern recognition and machine learning, identifies subsets of data that are relevant to the parameters used and is normally called Maximum Relevance. These subsets often contain material which is relevant but redundant and mRMR attempts to address this problem by removing those redundant subsets. mRMR has a variety of applications in many areas such as cancer diagnosis and speech recognition[29].

NCA, Neighbourhood Component Analysis: It is a supervised learning method for classifying multivariate data into distinct classes according to a given distance metric over the data[30].

NGS: Massive parallel sequencing or massively parallel sequencing is any of several high-throughput approaches to DNA sequencing using the concept of massively

24 https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance
25 https://en.wikipedia.org/wiki/Logistic_regression
26 https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
27 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/metastasis
28 https://www.cancer.gov/types/metastatic-cancer
29 https://en.wikipedia.org/wiki/Minimum_redundancy_feature_selection
30 https://en.wikipedia.org/wiki/Neighbourhood_components_analysis

parallel processing; it is also called next-generation sequencing (NGS) or second-generation sequencing[31].

OS, overall survival: The length of time from either the date of diagnosis or the start of treatment for a disease, such as cancer, that patients diagnosed with the disease are still alive. In a clinical trial, measuring the overall survival is one way to see how well a new treatment works[32].

PCA, Principal Component Analysis: It is used in exploratory data analysis and for making predictive models. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible[33].

PDAC, PAAD, pancreatic ductal adenocarcinoma: This is the most common form of pancreatic cancer. In fact, it makes up more than 80 percent of diagnosed cases of pancreatic cancer. A type of exocrine pancreatic cancer, PDAC, grows from cells lining small tubes, called ducts, in the pancreas. These tubes carry the digestive juices, which contain enzymes, into the main pancreatic duct and then on into the first part of the small intestine, called the duodenum[34].

Pearson's correlation coefficient: It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation[35].

PFS, progression-free survival: The length of time during and after the treatment of a disease, such as cancer, that a patient lives with the disease but it does not get worse. In a clinical trial, measuring the PFS is one way to see how well a new treatment works[36].

Prevalence: In medicine, a measure of the total number of people in a specific group who have (or had) a certain disease, condition, or risk factor (such as smoking or obesity) at a specific point in time or during a given period of time. For example, the prevalence of breast cancer may show how many women in the U.S. were diagnosed with breast cancer within the past 10 years, including those who are receiving treatment and those who are considered cured, and are still alive on a certain date[37].

PSO, Particle Swarm Optimization: It is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. It solves a problem by having a population of candidate solutions,

31 https://en.wikipedia.org/wiki/Massive_parallel_sequencing
32 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/os
33 https://en.wikipedia.org/wiki/Principal_component_analysis
34 https://umiamihealth.org/sylvester-comprehensive-cancer-center/treatments-and-services/pancreatic-cancer/pancreatic-ductal-adenocarcinoma
35 https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
36 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pfs
37 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/prevalence

here dubbed particles, and moving these particles around in the search-space according to simple mathematical formula over the particle's position and velocity. Each particle's movement is influenced by its local best known position, but is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move the swarm toward the best solutions[38].

Recurrence: Cancer that has recurred (come back), usually after a period of time during which the cancer could not be detected. The cancer may come back to the same place as the original (primary) tumour or to another place in the body. Also called recurrent cancer[39,40].

Relapse: The return of a disease or the signs and symptoms of a disease after a period of improvement[41].

REO, Relative Expression Ordering: the within-sample relative expression orderings (REOs) of gene pairs, which is also called Relative Expression Analysis (RXA), are robust against experimental batch effects and invariant to monotone data transformation. Besides, the within-sample REOs of gene pairs are robust against variations of the tumor epithelial cell proportions in tissues sampled from different sites of a tumor, partial RNA degradation in the sample preparation process and during the storage stage and amplification bias for minimum specimens even with about 15–25 cancer cells, which are also important factors leading to the failure of validation and clinical application of the quantitative transcriptional signatures. The robustness property of the within-sample REOs enables researchers to integrate multiple datasets produced by the same or similar platforms for selecting disease signatures and training classifiers, which makes it more likely to find robust signatures[22].

RF, Random Forest: The are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees[42].

RFS, relapse-free survival: see Disease-free survival (DFS).

Spearman's rank correlation: The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not)[43].

Stroma, stromal cell: In cancer, during normal wound healing processes, the local stromal cells change into reactive stroma after altering their phenotype. However,

38 https://en.wikipedia.org/wiki/Particle_swarm_optimization
39 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/recurrence
40 https://www.cancer.org/treatment/survivorship-during-and-after-treatment/understanding-recurrence/what-is-cancer-recurrence.html
41 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/relapse
42 https://en.wikipedia.org/wiki/Random_forest
43 https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

under certain conditions, tumor cells can convert these reactive stromal cells further and transition them into tumor-associated stromal cells (TASCs). In comparison to non-reactive stromal cells, TACs secrete increased levels of proteins and matrix metalloproteinases (MMPs). These proteins include fibroblast activating protein and alpha-smooth muscle actin. Furthermore, TACs secrete many pro-tumorigenic factors such as vascular endothelial growth factor (VEGF), stromal-derived factor-1 alpha, IL-6, IL-8, tenascin-C, and others. These factors are known to recruit additional tumor and pro-tumorigenic cells.[44]

SVM, Support Vector Machine: These are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis[45].

TME: Tumour micro-environment.

TNM: A system to describe the amount and spread of cancer in a patient's body, using TNM. T describes the size of the tumor and any spread of cancer into nearby tissue; N describes spread of cancer to nearby lymph nodes; and M describes metastasis (spread of cancer to other parts of the body). This system was created and is updated by the American Joint Committee on Cancer (AJCC) and the International Union Against Cancer (UICC). The TNM staging system is used to describe most types of cancer. Also called AJCC staging system[46].

TWSVM, Twin Support Vector Machine: aims to find two symmetry planes such that each plane has a distance close to one data class and as far as possible from another data class. On several benchmark data sets, TWSVM is not only fast, but shows good generalization. The kernel functions commonly used for SVM methods are the linear kernel, polynomial kernel, and radial basis function (RBF) kernel[15].

44 https://en.wikipedia.org/wiki/Stromal_cell#In_Cancer
45 https://en.wikipedia.org/wiki/Support-vector_machine
46 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/tnm-staging-system

8 References

[1] L. Rahib, B. D. Smith, R. Aizenberg, A. B. Rosenzweig, J. M. Fleshman, and L. M. Matrisian, ‘Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the united states’, *Cancer Res.*, vol. 74, no. 11, pp. 2913–2921, 2014, doi: 10.1158/0008-5472.CAN-14-0155.

[2] P. Sarantis, E. Koustas, A. Papadimitropoulou, A. G. Papavassiliou, and M. V. Karamouzis, ‘Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy’, *World Journal of Gastrointestinal Oncology*, vol. 12, no. 2. Baishideng Publishing Group Co, pp. 173–181, Feb. 15, 2020, doi: 10.4251/wjgo.v12.i2.173.

[3] J. Song *et al.*, ‘Five key lncRNAs considered as prognostic targets for predicting pancreatic ductal adenocarcinoma’, *J. Cell. Biochem.*, vol. 119, no. 6, pp. 4559–4569, 2018, doi: 10.1002/jcb.26598.

[4] D. J. Birnbaum, F. Bertucci, P. Finetti, D. Birnbaum, and E. Mamessier, ‘Molecular classification as prognostic factor and guide for treatment decision of pancreatic cancer’, *Biochimica et Biophysica Acta - Reviews on Cancer*, vol. 1869, no. 2. Elsevier B.V., pp. 248–255, Apr. 01, 2018, doi: 10.1016/j.bbcan.2018.02.001.

[5] H. Ying *et al.*, ‘Genetics and biology of pancreatic ductal adenocarcinoma’, *Genes and Development*, vol. 30, no. 4. pp. 355–385, 2016, doi: 10.1101/gad.275776.115.

[6] P. P. Almeida, C. P. Cardoso, and L. M. de Freitas, ‘PDAC-ANN: An artificial neural network to predict Pancreatic Ductal Adenocarcinoma based on gene expression’, *bioRxiv*, pp. 1–11, 2019, doi: 10.1101/698209.

[7] X. Yan *et al.*, ‘Importance of gene expression signatures in pancreatic cancer prognosis and the establishment of a prediction model’, *Cancer Manag. Res.*, vol. 11, pp. 273–283, 2019, doi: 10.2147/CMAR.S185205.

[8] J. Ramos, J. A. Castellanos-Garzón, J. F. de Paz, and J. M. Corchado, ‘A data mining framework based on boundary-points for gene selection from DNA-microarrays: Pancreatic Ductal Adenocarcinoma as a case study’, *Eng. Appl. Artif. Intell.*, vol. 70, no. June 2016, pp. 92–108, 2018, doi: 10.1016/j.engappai.2018.01.007.

[9] Z. M. Zhang, J. S. Wang, H. Zulfiqar, H. Lv, F. Y. Dao, and H. Lin, ‘Early Diagnosis of Pancreatic Ductal Adenocarcinoma by Combining Relative Expression Orderings With Machine-Learning Method’, *Front. Cell Dev. Biol.*, vol. 8, no. October, pp. 1–9, 2020, doi: 10.3389/fcell.2020.582864.

[10] B. Baek and H. Lee, ‘Prediction of survival and recurrence in patients with pancreatic cancer by integrating multi-omics data’, *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-76025-1.

[11] B. Alizadeh Savareh *et al.*, ‘A machine learning approach identified a diagnostic model for pancreatic cancer through using circulating microRNA signatures’, *Pancreatology*, vol. 20, no. 6, pp. 1195–1204, 2020, doi: 10.1016/j.pan.2020.07.399.

[12] X. H. Shi *et al.*, ‘A Five-microRNA Signature for Survival Prognosis in Pancreatic Adenocarcinoma based on TCGA Data’, *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018, doi: 10.1038/s41598-018-22493-5.

[13] M. Sinkala, N. Mulder, and D. Martin, ‘Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics’, *Sci. Rep.*, vol. 10, no. 1, pp. 1–14, 2020, doi: 10.1038/s41598-020-58290-2.

[14] N. P. Long *et al.*, ‘An integrative data mining and omics-based translational model for the identification and validation of oncogenic biomarkers of pancreatic cancer’, *Cancers (Basel).*, vol. 11, no. 2, 2019, doi: 10.3390/cancers11020155.

[15] W. Sadewo, Z. Rustam, H. Hamidah, and A. R. Chusmarsyah, ‘Pancreatic cancer early detection using twin support vector machine based on kernel’, *Symmetry (Basel).*, vol. 12, no. 4, Apr. 2020, doi: 10.3390/SYM12040667.

[16] H.-M. Lin, X.-F. Xue, X.-G. Wang, S.-C. Dang, and M. Gu, ‘Application of artificial intelligence for the diagnosis, treatment, and prognosis of pancreatic cancer’, *Artif. Intell. Gastroenterol.*, vol. 1, no. 1, pp. 19–29, 2020, doi: 10.35712/wjg.v1.i1.19.

[17] N. R. Latha *et al.*, ‘Gene expression signatures: A tool for analysis of breast cancer prognosis and therapy’, *Crit. Rev. Oncol. Hematol.*, vol. 151, no. April, p. 102964, 2020, doi: 10.1016/j.critrevonc.2020.102964.

[18] F. Yuan, L. Lu, and Q. Zou, ‘Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms’, *Biochim. Biophys. Acta - Mol. Basis Dis.*, vol. 1866, no. 8, p. 165822, 2020, doi: 10.1016/j.bbadis.2020.165822.

[19] A. Anjum, S. Jaggi, E. Varghese, S. Lall, A. Bhowmik, and A. Rai, ‘Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach’, *J. Comput. Biol.*, vol. 23, no. 4, pp. 239–247, 2016, doi: 10.1089/cmb.2015.0205.

[20] H. Liu and R. Setiono, ‘Incremental Feature Selection’, *Appl. Intell.*, vol. 9, 1998, doi: 10.1023/A.

[21] A. Rada-iglesias, S. Enroth, and C. Wadelius, ‘Monte Carlo feature selection for supervised classification’, *Bioinformatics*, vol. 24, no. 2008, pp. 110–117, 2008, doi: 10.1093/bioinformatics/btm486.

[22] Q. Guan *et al.*, ‘Quantitative or qualitative transcriptional diagnostic signatures? A case study for colorectal cancer’, *BMC Genomics*, vol. 19, no. 1, pp. 1–11, 2018, doi: 10.1186/s12864-018-4446-y.

[23] Chari S, Kelly K, Hollingsworth M et al, ‘Early Detection of Sporadic Pancreatic Cancer’, *Pancreas*, vol. 44, no. 5, 2015, doi: 10.1097/MPA.0000000000000368.

[24] Nagy A, Munkacsy G, Gyorffy B, ‘Pancancer survival analysis of cancer hallmark genes’, *Sci. Rep.*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-84787-5

[25] Huaiyu Mi, Anushya Muruganujan, J Xiaosong Huang, Dustin Ebert, Caitlin Mills, Xinyu Guo and Paul D Thomas, ‘Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0)’, *Nat. Protoc.,* vol. 14, no. 3, pp. 703-721, 2019, doi: 10.1038/s41596-019-0128-8

9 Annexes

9.1 DEG analysis

9.1.1 Treatment vs outcome

9.1.1.1 GSE112282

Total of 1044 unique DEG.

Procedure

Verification that data is normalized:

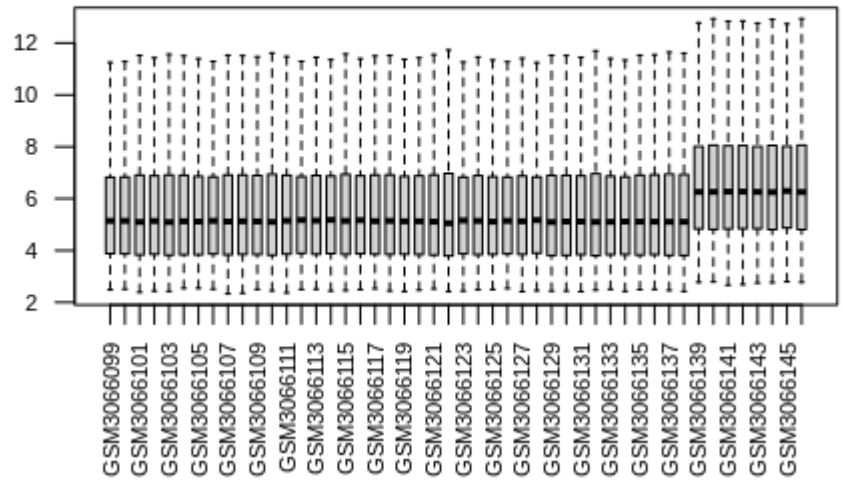

Figure 28: Box plot for GSE112282

Samples' metadata selection (*cell line*, *replicate* and *treatment*):

```
sampleInfo <- select(sampleInfo,
                     "cell line:ch1",
                     "replicate info:ch1",
                     "treatment:ch1")
sampleInfo <- rename(sampleInfo,
                     line="cell line:ch1",
                     replicate="replicate info:ch1",
                     treatment="treatment:ch1")
```

Design (*treatment, line, replicate*) and contrasts for DEG analysis:

```
design_colnames <- c("BET","BETMEK","MEK","VEHICLE","COLO201",
                     "HPAFII","NCIH510","RKO","Replicate2")
design <- model.matrix(~0+sampleInfo$treatment
                       +sampleInfo$line
                       +sampleInfo$replicate)
colnames(design) <- design_colnames
contrasts <- makeContrasts(BET - VEHICLE,
                           BETMEK - VEHICLE,
                           MEK - VEHICLE,
                           levels=design)
```

Result of DEG analysis with default *p.value=0.05*:

```
> table(results)
results
    -1      0      1
 18640 126966  18419
```

Venn diagram:

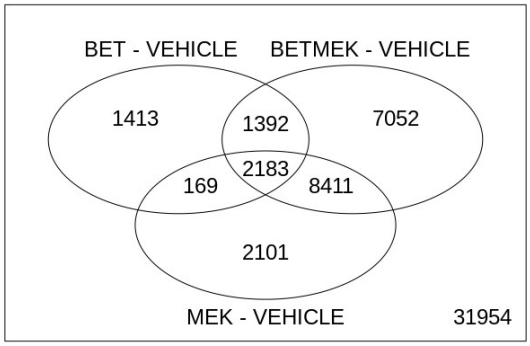

Figure 29: Venn diagram for GSE112282

There are 2147 unique GenBankIDs out of 2183, and they map to 1056 ENTREZIDs, and to 1044 HGNC Ids.

9.1.1.2 GSE45757

Total of 5450 unique DEG.

Procedure

Verification that data is normalized after applying log2:

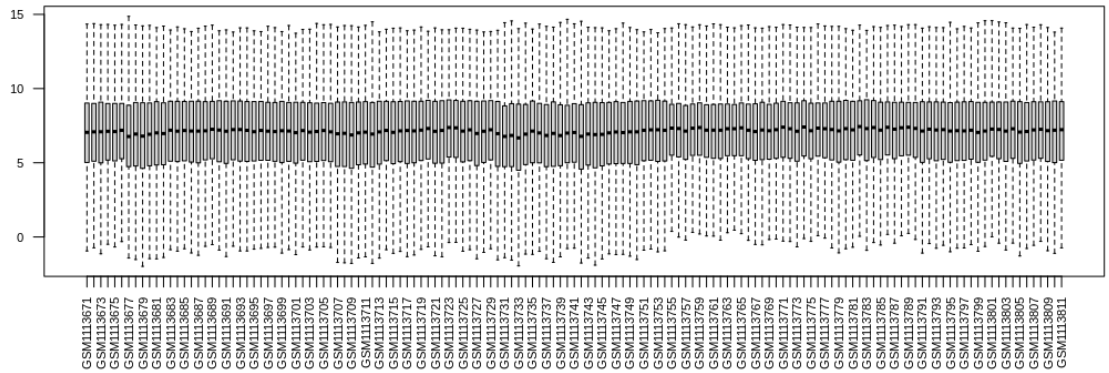

Figure 30: Box plot for GSE45757

Samples' metadata selection:

```
sampleInfo <- select(sampleInfo,
                              "treated with:ch1", "cell line:ch1")
sampleInfo <- rename(sampleInfo,
                              treated="treated with:ch1",
                              line="cell line:ch1")
```

Design and contrasts for DEG analysis:

```
design <- model.matrix(~0+sampleInfo$treated
                       +sampleInfo$line)
colnames(design) <- design_colnames
contrasts <- makeContrasts(Untreated - Treated,
                           levels=design)
```

Result of DEG analysis with default *p.value=0.05*:

```
results
   -1     0     1
 5510 13058  3709
```

Venn diagram:

Figure 31: Venn diagram for GSE45757

There are 8732 unique GenBankIDs out of 9219, and they map to 5498 ENTREZIDs, and to 5450 HGNC Ids.

9.1.1.3 GSE14426

Total of 19 unique DEG.

Procedure

Verification that data is normalized after applying log2:

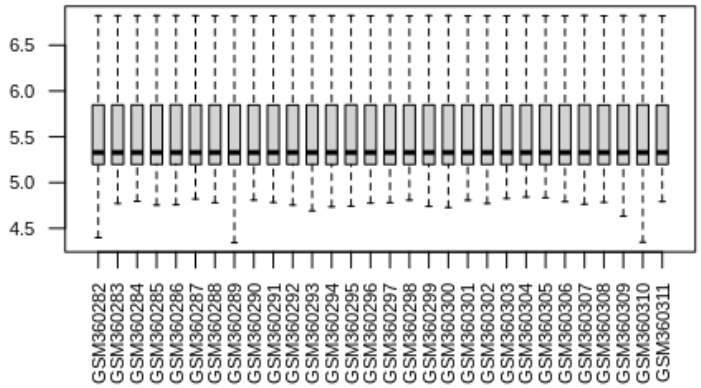

Figure 32: Box plot for GSE14426

Samples' metadata selection (only *24hr* and *168hr*):

```
sampleInfoSubset <- sampleInfo[str_detect(sampleInfo$source_name_ch1,
                                          "24hr|168hr"), ]
sampleInfo <- select(sampleInfoSubset, "source_name_ch1")
sampleInfo <- rename(sampleInfo, source="source_name_ch1")
```

Design and contrasts for DEG analysis:

```
design <- model.matrix(~0+sampleInfo$source)
design_colnames <- c("ATRA168h","ATRA24h","Vehicle168h","Vehicle24h")
colnames(design) <- design_colnames
contrasts <- makeContrasts(Vehicle168h - ATRA168h,
                           Vehicle24h - ATRA24h,
                           levels=design)
```

Result of DEG analysis with a modified *p.value=0.3*:

```
> table(results)
results
   -1     0     1
 2770 91910  2722
```

Venn diagram:

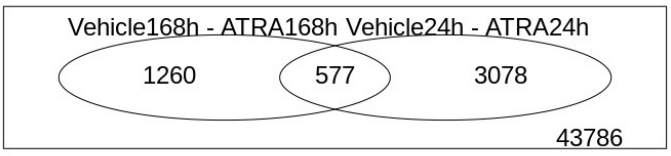

Figure 33: Venn diagram for GSE14426

There are 565 unique GenBankIDs out of 577, and they map to 20 ENTREZIDs, and to 19 HGNC IDs.

Lower values (e.g. *p.value=0.05*) were also tried, returning 137 GenBankIDs, but they mapped to only 2 HGNC IDs.

9.1.2 Gene expression vs outcome

9.1.2.1 GSE21501

Total of 1205 unique DEG.

Procedure

Verification that data is normalized after applying *log2*:

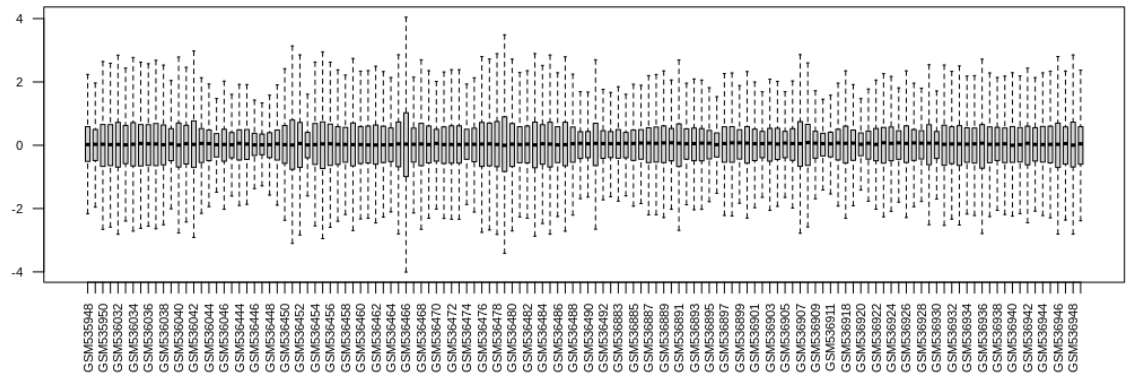

Figure 34: Box plot for GSE21501

Samples' metadata selection (*risk*):

```
sampleInfo <- dplyr::select(sampleInfo,
                      "characteristics_ch2.5",
                      "characteristics_ch2.6")
sampleInfo <- dplyr::rename(sampleInfo,
                      risk="characteristics_ch2.5",
                      risk2="characteristics_ch2.6")

# Information is misplaced in these samples
sampleInfo["GSM536946","risk"] <- sampleInfo["GSM536946","risk2"]
sampleInfo["GSM536892","risk"] <- sampleInfo["GSM536892","risk2"]
sampleInfo <- dplyr::select(sampleInfo, risk)

# Remove samples with empty value
sampleInfo[sampleInfo == ""] <- NA
sampleInfo <- na.omit(sampleInfo, "risk") # 102 samples
```

Design (*risk*) and contrasts for DEG analysis:

```
design <- model.matrix(~0+sampleInfo$risk)
design_colnames <- c("HighRisk","LowRisk")
colnames(design) <- design_colnames
contrasts <- makeContrasts(LowRisk - HighRisk,
                           levels=design)
```

Result of DEG analysis with default *p.value=0.05*:

```
> table(results)
results
   -1     0     1
 1093 43263   864
```

Venn diagram:

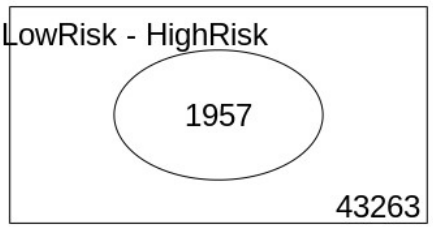

Figure 35: Venn diagram for GSE21501

There are 1424 unique IDs out of the 1957 GenBankIDs, and they map to 1235 ENTREZIDs, and to 1205 HGNC IDs.

9.1.2.2 GSE28735

Total of 2025 unique DEG.

Procedure

Verification that data is normalized after applying *log2*:

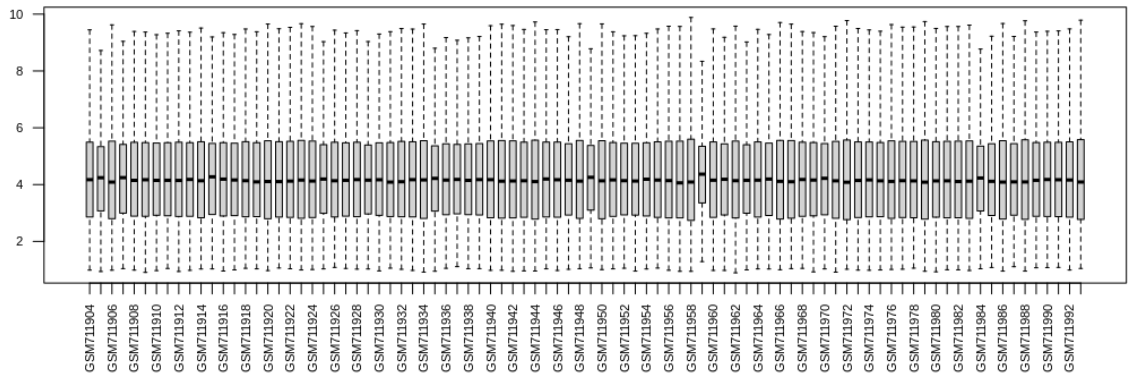

Figure 36: Box plot for GSE28735

Samples' metadata selection (*survival* and *tissue*):

```
  sampleInfo <- dplyr::select(sampleInfo,
                              "survival_month:ch1",
                              "tissue:ch1")
  sampleInfo <- dplyr::rename(sampleInfo,
                              OS="survival_month:ch1",
                              tissue="tissue:ch1")
  sampleInfo$OS <- as.numeric(sampleInfo$OS)
  sampleInfo <- na.omit(sampleInfo, "OS")
  # Select only Tumor samples
  sampleInfo <- sampleInfo[sampleInfo$tissue == "T", ] # 42 samples

  samples_to_keep <- row.names(sampleInfo)

 sampleInfo$stage[sampleInfo$OS <= stages_mst["III","MST+40%"]]
            <- 'Advanced'
 sampleInfo$stage[sampleInfo$OS > stages_mst["III","MST+40%"]]
            <- 'Early'
```

Design (*stage*) and contrasts (1) for DEG analysis:

```
  design <- model.matrix(~0+sampleInfo$stage)
  colnames(design) <- c("Advanced","Early")
  contrasts <- makeContrasts(Early - Advanced,
                             levels=design)
```

Result of DEG analysis with a modified *p.value=0.4*:

```
> table(results)
results
    -1      0      1
  7459 128241   5966
```

Venn diagram:

Figure 37: Venn diagram for GSE28735

There are 13425 unique GenBankIDs, and they map to 2122 ENTREZIDs, and to 2025 HGNC Ids.

9.1.2.3 GSE62165

Total of 596 unique DEG.

Procedure

Verification that data is normalized:

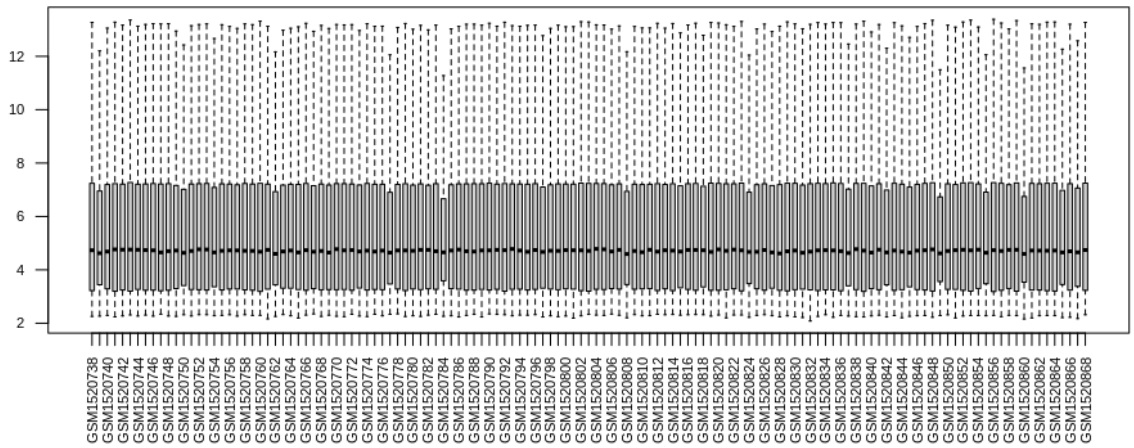

Figure 38: Box plot for GSE62165

Samples' metadata selection (*stage*):

```
sampleInfo <- dplyr::select(sampleInfo,
                            "Stage:ch1", "tissue:ch1")
sampleInfo <- dplyr::rename(sampleInfo,
                            stage="Stage:ch1",
                            tissue= "tissue:ch1")
sampleInfo <- sampleInfo[sampleInfo$tissue == "pancreatic tumor", ]

sampleInfo$stage[sampleInfo$stage == "1a"] <- "Early"
sampleInfo$stage[sampleInfo$stage == "1b"] <- "Early"
sampleInfo$stage[sampleInfo$stage == "2a"] <- "Early"
sampleInfo$stage[sampleInfo$stage == "2b"] <- "Early"
sampleInfo$stage[sampleInfo$stage == "3"] <- "Advanced"
sampleInfo$stage[sampleInfo$stage == "4"] <- "Advanced"
```

Design (*stage*) and contrasts (1) for DEG analysis:

```
design <- model.matrix(~0+sampleInfo$stage)
design_colnames <- c("Advanced","Early")
colnames(design) <- design_colnames
contrasts <- makeContrasts(Early - Advanced,
                           levels=design)
```

Result of DEG analysis with a modified *p.value=0.8*:

```
> table(results)
results
   -1     0     1
  239 33101   136
```

Venn diagram:

Figure 39: Venn diagram for GSE62165

There are 1051 unique GenBankIDs, which map to 604 ENTREZID, and to 596 HGNC Ids.

9.1.2.4 GSE71729

Total of 749 unique DEG.

Procedure

Verification that data is normalized:

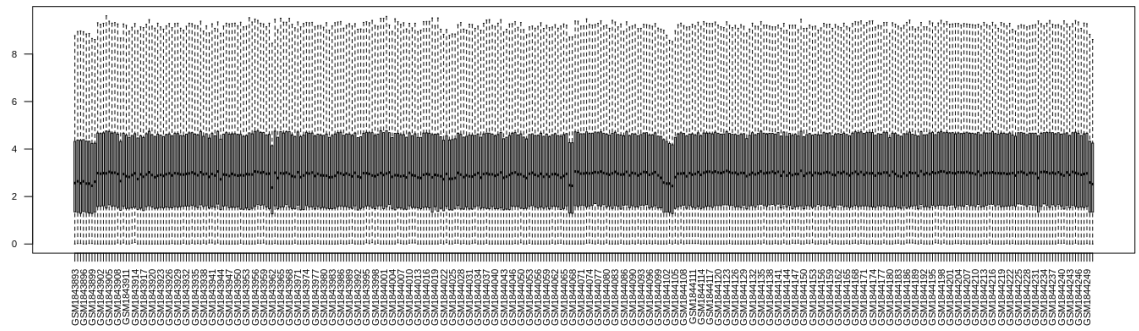

Figure 40: Box plot for GSE71729

Samples' metadata selection (*survival*):

```
sampleInfo <- dplyr::select(sampleInfo,
                            "survival_months:ch2")
sampleInfo <- dplyr::rename(sampleInfo,
                      OS="survival_months:ch2")
sampleInfo$OS <- as.numeric(sampleInfo$OS)
sampleInfo <- na.omit(sampleInfo, "OS") # 125 samples

samples_to_keep <- row.names(sampleInfo)

sampleInfo$stage[sampleInfo$OS <= stages_mst["III","MST+40%"]]
          <- 'Advanced'
sampleInfo$stage[sampleInfo$OS > stages_mst["III","MST+40%"]]
          <- 'Early'
```

Design (*stage*) and contrasts for DEG analysis:

```
design <- model.matrix(~0+sampleInfo$stage)
colnames(design) <- c("Advanced", "Early")
contrasts <- makeContrasts(Early - Advanced,
                           levels=design)
```

Result of DEG analysis with a modified *p.value=0.7*:

```
> table(results)
  -1     0     1
 336 19000   413
```

Venn diagram:

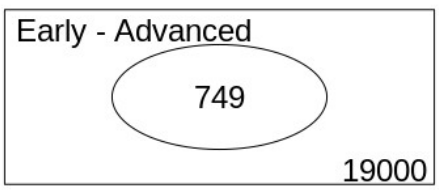

Figure 41: Venn diagram for GSE71729

There are 749 unique HGNC IDs.

9.1.2.5 GSE56560

Total of 153 unique DEG.

Procedure

Verification that data is normalized:

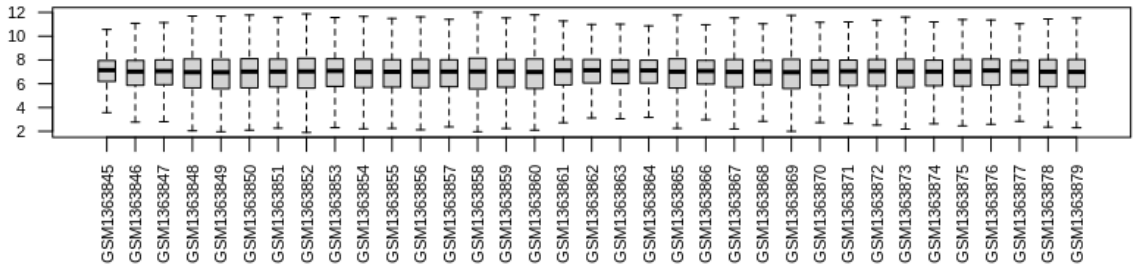

Figure 42: Box plot for GSE56560

Samples' metadata selection (*grading*):

```
  sampleInfo <- dplyr::select(sampleInfo,
                              "grading:ch1")
  sampleInfo <- dplyr::rename(sampleInfo,
                       grading="grading:ch1")
  sampleInfo$grading[sampleInfo$grading == "N/A"] <- NA
  sampleInfo <- na.omit(sampleInfo, "grading") # 28 samples
  samples_to_keep <- row.names(sampleInfo)
```

Design (*grading*) and contrasts for DEG analysis:

```
  design <- model.matrix(~0+sampleInfo$grading)
  design_colnames <- c("G2", "G3")
  colnames(design) <- design_colnames
  contrasts <- makeContrasts(G3 - G2,
                             levels=design)
```

Result of DEG analysis with a modified *p.value=0.99*:

```
> table(results)
results
   -1     0     1
  299 61676   200
```

Venn diagram:

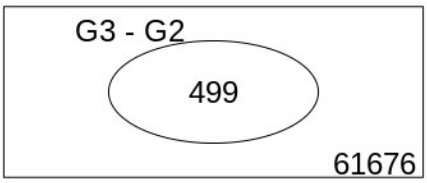

Figure 43: Venn diagram for GSE56560

There are 499 unique GenBankIDs, and they map to 159 ENTREZIDs, and to 153 HGNC IDs.

9.2 Gene signatures

The objective was to maximise the number of genes in common and obtain a signature smaller than 25 genes. To accomplish this, several iterations were performed trying different approaches:

- narrowing the number of DEG obtained in some series by selecting the top X (e.g. top 500 genes with lowest *p.value*)

- modifying the *p.value* in *decideTests()* function of *limma* package in R, in order to get a higher number of DEG

The latter worked better because was able to maximise the overlap of genes in all the signatures that had to be generated.

The next sections showed the results obtained, as well as, the full list of genes for each signature.

9.2.1 Treatment vs outcome signature

The intersection of the DEG obtained for each of the series, produced the values showed in Table 23:

1	2	3
5998	742	9

Table 23: Common DEGs to series in Treatment vs outcome

There are 9 DEG in common to the three series in this group. The full list of genes follows:

BCKDHB	IDS	MYO9A
CREB1	KIDINS220	PTEN
HNRNPA2B1	LAMP2	SH3TC2

9.2.2 Gene expression vs outcome signature

The intersection of the DEG obtained for each of the series, produced the values showed in Table 24:

1	2	3	4
4016	330	16	1

Table 24: Common DEGs to series in Gene expression vs outcome

There is 1 DEG in common to four series (out of 5) and 16 DEG in common to three series in this group. The full list of genes follows:

HMGA2	GSDMB	SLC23A2
C1orf21	KCNB1	SLC6A14
CBX6	MAMSTR	TOP2B
CEACAM5	MAPK10	TRIM15
CNTN2	MT1H	TSPAN3
E2F7	POF1B	

9.2.3 Common signature

Finally, a common signature was generated by doing the intersection between the two groups. In order to maximise the number of genes in common, any gene in at least two series (see Table 23 and Table 24) was considered. The result is shown in Table 25:

1	2
1064	17

Table 25: Common DEG to both groups

There are 17 genes in common between the two groups. The full list of genes follows:

ADI1	FAR2	PLEK2
CDS1	MAGOH	RAB4B
CDT1	MAPK13	S100A14
CHST6	MRPL2	SORD
CSRNP2	NSF	TFF1
F12	PLBD1	

9.3 ML model training and evaluation

Different tools of the *caret* package in R were used for this phase. The following steps were followed to train and evaluate each model using the different signatures obtained in Feature selection:

1. Split data into training and test sets using *sample.split()* function of *caTools* package:

```
sample_final <- sample.split(my_data_signature_final$class,
                             SplitRatio = .6)
training_common <- subset(my_data_signature_final,
                          sample_final == TRUE)
test_common <- subset(my_data_signature_final,
                      sample_final == FALSE)
```

The proportion of classes in each set was verified to confirm they were similar:

```
table(training_common$class)
table(test_common$class)
```

2. Define the control using *trainControl()* function of *caret* package:

```
trControl <- trainControl(method = "repeatedcv",
                          number = 10,
                          repeats = 10,
                          search = "grid")
```

3. Search best *mtry* using *tuneGrid* parameter in *train()* function of *caret* package:

```
tuneGrid <- expand.grid(.mtry = c(2,3,4,5,6,7,8,9,10,11))
rf_default <- train(class~.,
                    data = train,
                    method = "rf",
                    metric = "Accuracy",
                    trControl = trControl,
                    tuneGrid = tuneGrid,
                    importance = TRUE)
```

4. Search best *ntree* by manually running the *train()* function with different values for this parameter:

```
tuneGrid <- expand.grid(.mtry = best_mtry)
store_maxtrees <- list()
for (ntree in c(250, 300, 350, 400, 450, 500, 550,
                600, 800, 1000, 2000, 3000)) {
  # Run the model
  rf_maxtrees <- train(class~.,
                       data = train,
                       method = "rf",
                       metric = "Accuracy",
                       trControl = trControl,
                       tuneGrid = tuneGrid,
                       ntree = ntree,
                       importance = TRUE)
  key <- toString(ntree)
  store_maxtrees[[key]] <- rf_maxtrees
}
results_tree <- resamples(store_maxtrees)
summary(results_tree)
```

5. Train model with best settings for *mtry* and *ntree*:

```
fit_rf <- train(class~.,
```

```
                train,
                method = "rf",
                metric = "Accuracy",
                tuneGrid = tuneGrid,
                trControl = trControl,
                importance = TRUE,
                ntree = 1000)
```

6. Evaluate using *predict()* function and the test set:

```
prediction <- predict(fit_rf, test)
```

7. Get model metrics using *confusionMatrix()* function:

```
confusion_matrix <- confusionMatrix(prediction, test$class)
confusion_matrix$overall
confusion_matrix$byClass
confusion_matrix$table
```

8. Plot the importance of each variable using *plot()* and *varImp()* functions:

```
plot(varImp(fit_rf))
```

9.4 Code repository

All the code used to process the data (e.g. perform the DEG analysis), generate the signatures, perform the correlation analyses, and train/evaluate the models can be found in this public repository: https://github.com/uruloki85/tfm