POLITECNICO DI TORINO

Master degree course in Computer Engineering

Master Degree Thesis

Computational approaches for the identification of candidate chemotherapy-related lncRNAs in HGSOvCa

Supervisors

prof. Elisa Ficarra

prof. Sampsa Hautaniemi

prof. Rainer Lehtonen

Candidate

Maria Serena Ciaburri

matricola 231745

July 2018

Ai tre venti gentili che
soffiano nelle mie vele

Summary

High grade serous ovarian cancer (HGSOvCa) is a malignant tumor subtype that originates from the female reproductive system. The standard therapies prescribed to HGSOvCa patients include several chemotherapy cycles based on platinum-taxol drugs and a debulking surgery for removing cancer tissues.

A fundamental characteristic of this disease, that drastically decreases the 5-years survival rates, is the acquisition of chemotherapy resistance by the tumoral cells after the first-line treatment. Both the cancer aggressiveness and the development of the platinum resistance increase the necessity of a more effective and targeted therapy.

During the last 10 years, a branch of the cancer research has focused its attention on the genomic components called "long non-coding RNAs". These elements, originating from RNA molecules, do not encode for proteins and are composed by a number of nucleotides that ranges from 200 to 100000. Even if they do not have encoding properties, it was shown that those transcripts are actively involved in many cell functions and they are dysregulated during the genesis and the development of different tumors.

The main goal of this master thesis is to develop a pipeline for the automatic identification of long non-coding RNAs that can be possibly involved in the platinum-resistance process (generally called *drivers*). By knowing the drivers and the molecular processes that lead to chemotherapy resistance, it would be possible to identify efficient pharmacological targets and design a more effective therapy.

This thesis was conducted in collaboration with the System Biology Lab for Drug Resistance of the Helsinki University in Finland. The data employed in this analysis are clinical and genetic information regarding HGSOvCa patients enrolled in a chemotherapy treatment after the diagnosis of the disease. As genetic information the analysis uses the

expression levels computed from the total RNA-sequencing of the patients' samples. Expression levels measure the amount of the different genetic elements present in the sample and they consequently facilitate the identification of the processes in which those elements are involved.

In order to achieve the proposed goal, the analysis was focused on the identification of genes that show different behaviours (so different expression levels) in the chemo-resistant patients with respect to the chemo-sensitive ones. For this reason, samples were initially divided in two groups, according to the available clinical data.

The analysis was conducted by realizing a pipeline that integrates two different strategies: an unsupervised hierarchical clustering approach supported by statistical processing and a supervised procedure based on feature selection through machine learning methods.

With the first strategy, lncRNAs showing differences in the expression levels between the chemo-resistant and chemo-sensitive patients were extracted by considering the genes judged as statistically significant by the Mann-Whitney-Wilcoxon test. Those genes were successively employed in the unsupervised hierarchical clustering of the available samples, which produced two separate chemotherapy-related clusters of patients. Moreover, the application of the Kaplan-Meier analysis and the log-rank test revealed a significant difference between the survival rates of the two subgroups.

In the second approach, the identification of platinum resistance related lncRNAs was exploited by applying two feature selection methods. The first one, is a customized approach that involves several runs of the Random Forest (RF) algorithm and the employment of the leave one out cross-validation for assessing the model's accuracy. Differentially expressed long non-coding RNAs were extracted by considering the Mean Decrease Accuracy (or MDA) index computed in the RF learning phase. The second feature selection technique is, instead, an already available wrapper algorithm called *Boruta*, which is specifically created for feature selection. Also this method uses the MDA index as variable importance metric. In both cases, the application of unsupervised hierarchical clustering based on the retrieved lncRNAs produced two well-separated clusters: one containing chemo-resistant samples and the other containing chemo-sensitive ones.

The choice of combining different methods for the same analysis was raised by the necessity of having more confident results. The outcomes of the two employed methods,

in fact, were finally compared and only the long non-coding RNAs identified by all the techniques were taken into account.

From the analysis were retrieved 6 long non-coding RNAs that can potentially be related to the chemotherapy resistance process and that are currently under wet-lab validation. In this thesis are also highlighted the strength and the weaknesses of the adopted approaches.

Acknowledgements

Contents

List of Tables

List of Figures

Chapter 1

Introduction

High grade serous ovarian cancer (HGSOvCa) is an aggressive gynaecological malignancy that affects women worldwide. It is characterized by asymptomaticity, that leads to a delay in the diagnosis of the disease and consequently to a late beginning of the therapy. HGSOvCa is also characterized by the acquisition of chemotherapy resistance by the tumor cells after the treatment. This means that even if the first-line chemotherapy has a positive outcome, a relapse phase in which the disease shows up again is still possible. The recurrence of the disease happens in the 75% of the cases and it usually leads to death [1]. Both the asymptomaticity and the chemotherapy resistance have an highly impact on the increase of the mortality rate. In fact, the 5 year survival rate for this kind of malignancy is very low and it is extimated around 35%-40%.

The standard therapy for this kind of disease is not effective enough and there is the need to find a more accurate and targeted solution. It has to underlined that cancer in general is a malignancy that originates from a mutant cell and subsequently differentiate in sub-clones. Those sub-groups can develop in different ways and they can produce different outcomes to the same therapy. This heterogeneity raises the need of a better understanding of the underlying processes and a more focused therapy. Different aspects of this malignancy, from the genesis to the chemoresistance, are currently under study.

During the last ten years it emerged the hypothesis that among the different genomic elements that compose a cell, there are some components that do not encode for proteins. These elements generate from RNA molecules and they are called non-coding RNAs. It

was found that even if they do not encode for proteins, they play an active role in different cell functions and, consequently, they also take part to the development of cancer. By now, not all the functions of these non-coding regions are yet fully understood and classified.

On the basis of this hypothesis, this work of thesis is focused on the identification of long non-coding RNAs that can possibly be related to chemotherapy resistance in patients with HGSOvCa. For this purpose, it was realized a pipeline that integrates two different strategies: a feature selection technique and a statistical one. The combination of the two different approaches is adopted in order to obtain a bigger confidence in the results and eliminate possible false positives.

The data on which the analysis is based are the expression levels of the cell components. The expression levels are a measure of the cell activity. The challenge is to identify which processes in the cell activity, and therefore which elements, are responsible for the development of the chemoresistance. Once identified the drivers of this process, it is possible to understand in details the different biological phases that lead to the acquisition of chemotherapy resistance and find an adequate and targeted therapy.

Expression levels are highly informative but noisy data and they need to be carefully analyzed. One component of the noise has a biological nature and it is related to the huge amount of information contained in these data. The expression levels describe the whole set of processes ongoing in a cell, so it is necessary to isolate the one we are interested in. Through RNA-sequencing it is possible to obtain the amount of cell activity each RNA element is responsible for. Computational processing is then needed in order to identify which portion of this activity is actually involved in the chemoresistance process and, at the same time, which are the genomic elements responsible for it. The other component of the noise has a technological nature and it is due to the computational errors present in each step of the RNA-sequencing process.

Both the complexity of the data at disposal and their intrinsic technological errors lead to the necessity of using different approaches to obtain more confident results. Confidence is in this case gained by comparing the results obtained with both techniques. The analysis is based on primary samples taken from the patient when the disease is diagnosed. Those samples are divided in two categories on the basis of clinical data, in order to distinguish chemoresistant patients from chemosensitive ones. The common goal for the two

approaches is to identify long non-coding RNAs showing different expression levels between the two groups of patients, in order to highlight a set of genes that may be involved in the chemotherapy resistance. The long non-coding RNAs found with both approaches are currently under validation in wet lab.

The following chapters describe the analytical work in details. In the second chapter there is an overview on the biological background, with the description of the high grade serous ovarian cancer characteristics, the long non-coding RNAs functions and the RNA-sequencing technique. In the third chapter, there is the introduction of concepts about machine learning and statistical theory that will be used during the analysis. The fourth chapter is dedicated to the illustration of the input data used and the results obtained and the description in details of the approaches and the choices adopted in this study. The conclusive chapter comments the results of each analytical step from a critical point of view and presents the limitation encountered during the study.

Chapter 2

Biological background

2.1 High grade serous ovarian cancer

Cancer is a set of multifactorial diseases characterized by an uncontrolled growth of cells that can affect organs and tissues. The reasons why cancer develops can be ascribed to genetic mutations, that can be both hereditary or due to environmental causes. Because of alterations in the genomic sequence, cells can lose their habitual properties and functions and they can gain new different ones. A list of the capabilities acquired by cancer cells during the malignant progression was first compiled in 2000 by Hanan and Weinberg under the name of "Hallmarks of cancer" [2] and then reviewed in 2011 [3]. The list is shown in the figure below and it includes processes that allow deregulated growth and sustainment of tumor cells and that promote tissue invasion and metastatization.

Ovarian cancer is a malignant tumor subtype that involves the female reproductive system. According to the site from which the tumors presumably originate [4], it can be classified in:

- *surface epithelial-stroma tumor*, if it arises from the cells that constitute the surface epithelium;
- *sex cord-stroma tumor*, if it arises from the cells that constitute the inner tissue of the ovary;
- *germ cells tumor*, if it arises from the cells of the germ line.

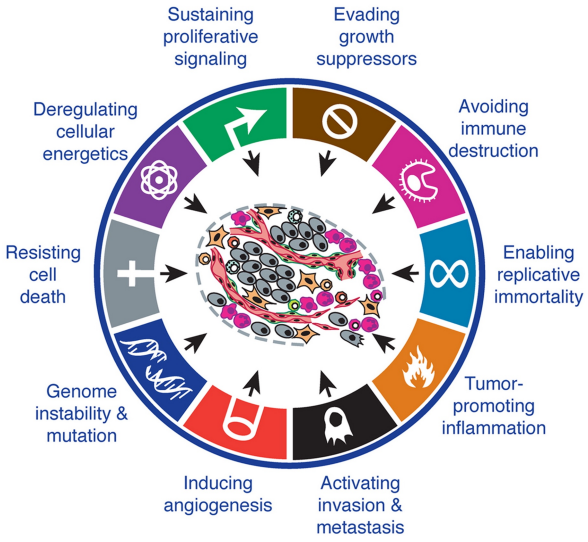
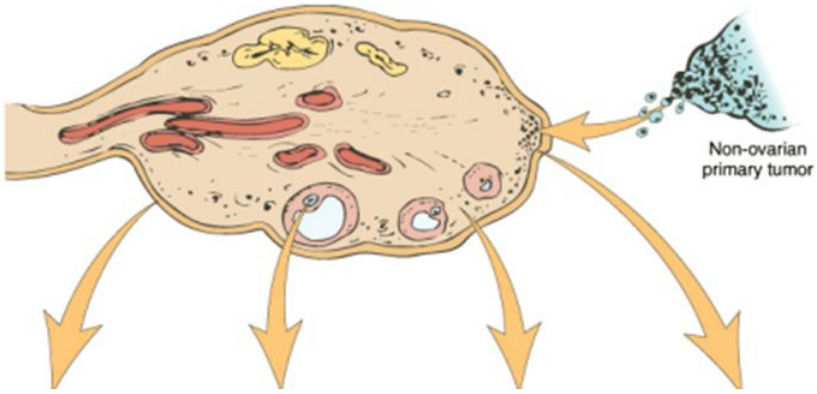


Figure 2.1. Hallmarks of cancer (2011)[3]

The epithelial tumors (or epithelial ovarian cancers, EOC) are almost the 85%-90% of all ovarian cancers [5] and one of their subclassification is the *serous ovarian cancer*, that can be again distinguished in *low grade serous ovarian cancer* (LGSOvCa) and *high grade serous ovarian cancer* (HGSOvCa) according to the aggressiveness of the tumor cells. This type of disease originates from the fallopian tubes and it is known for being asymptomatic. This means that in the majority of the cases, the discovery of the cancer coincides with the latest stages of the disease. Because the five-year survival rates get worse with the progression of the tumor, the HGSOvCa is also characterized by high mortality. In fact, the 5 year survival rate for this kind of disease is extimated around 35%-40% [6].

Once discovered, the HGSOvCa can be treated with a primary debulking surgery (PDS) to remove the cancerous tissues, followed by adjuvant chemotherapy (ACT)[7]. In standard chemotherapy are employed platinum-based drugs (e.g. carboplatin and cisplatin) in combination with paclitaxel (also known as "taxol"). If it is not possible to have a complete resection of the malignant tissues during PDS because the cancer is too vast, the patient can undergo to a first neoadjuvant chemotherapy (NACT) followed by an interval debulking surgery (IDS)[8]. This kind of surgery is performed to remove all the residual tumoral tissues, in order to raise the chances of a good therapy response.

Anyway, even if patients initially present a good response to the treatment, the majority



ORIGIN	SURFACE EPITHELIAL CELLS (Surface epithelial-stromal cell tumors)	GERM CELL	SEX CORD-STROMA	METASTASIS TO OVARIES
Overall frequency	65% - 70%	15% - 20%	5% - 10%	5%
Proportion of malignant ovarian tumors	90%	3% - 5%	2% - 3%	5%
Overall frequency	20+ years	0-25+ years	All ages	Variable
Types	- Serous tumor - Mucinous tumor - Endometrioid tumor - Cell clear tumor - Brenner tumor - Cystadenofibroma	- Teratome - Dysgerminoma - Endometrioid sinus tumor - Choriocarcinoma	- Fibroma - Granulosa - theca cell tumor - Sertoli - Leydig cell tumor	

Figure 2.2. Ovarian cancer subtypes

of them have a relapse phase after few months (tipically 6 - 12 months) in which the disease come back[9]. This is due to the fact that ovarian cancer cells acquire drug resistance after the chemotherapy. The acquisition of platinum resistance prevent the cells to be sensitive to the treatment and leaves to the cancer the possibility to grow. Until now, the reasons and the processes for the acquisition of drug resistance are still not completely clear.

2.2 Long non-coding RNAs

During the last 10 years, it was found that only 1-2% of the whole genome encodes for proteins while at least the 75% of the remaining part encodes for regulatory RNA[10]. RNA molecules that lack in protein coding capabilities are collectively referred to as *non-coding RNA*. These non-coding RNAs are functionally divided into *housekeeping non-coding RNA*

and *regulatory RNA*. The non-coding RNAs are also classified according to their molecular size into:

- *short non-coding RNAs*, ranging in length from 20 to 200 nucleotides;
- *long non-coding RNAs*, ranging in length from 200 to 100 000 nucleotides.

The long non-coding RNA topic is still quite new, so the functional role of many non-coding RNAs is still unknown or not well defined. By now, it is known that several lncRNAs are involved in different biological processes and regulate growth, differentiation and establishment of cell identity. The main functions in which long non-coding RNAs are involved in are [11]:

- the regulation of gene expression at transcriptional level, involving chromatin re-modelling and histone modification achieved through the interaction with protein complexes;
- the regulation of gene expression at post-transcriptional level, in which long non-coding RNAs may function as endogenous sponges and down-regulate a series of microRNAs;
- the assembly of protein complexes, in which they can act as scaffold to bring together different proteins in the same location.

Because these processes are commonly deregulated in several kinds of diseases, like cancer, lncRNAs play an important role in this field. As protein coding genes, in fact, long non-coding RNAs can act as oncogenes (genes that promotes the development of the disease) or tumor suppressor genes (genes that goes against the development of the disease) and influence several hallmarks of cancer. For example, HOTAIR, one of the most famous lncRNAs, participate to the promotion of angiogenesis[12],that ensure nutrient suppliers for tumor cells, and to the promotion of tissue invasion, that leads to metastasization[13].

2.3 Experimental procedures

2.3.1 RNA-sequencing

RNA sequencing[14] is an approach used for the study of the transcriptome (the whole set of RNA molecules we can find in one or more cells) that take advantage of the use of next generation sequencing technologies (NGS). NGS technologies are the second generation of sequencing techniques, characterized by high scalability and speed. This new generation of machines is able to sequence huge amount of data in parallel. The RNA-sequencing process is used in order to:

- understand the transcriptional structure of the genes, to be able to identify gene mutations or fusions;
- quantify the expression levels of a gene over time or among different groups of donors;
- classify all the genomic elements of a transcript.

Some steps or methods in the RNA-seq process can vary according to the experimental goals, but the general flow can be described as follows. The first phase of the sequencing process is the creation of a library made of *cDNA fragments*. CDNA stands for "complementary DNA", that is the DNA obtained from the synthesis of RNA molecules. A library is a pool of DNA fragments with adaptors. During the composition of the cDNA library, large RNA molecules are isolated from the original sample, reverse transcribed to cDNA and then fragmented in smaller pieces long 200-500 bp. Adaptors (that act as hooks for the sequencing platform) are ligated to both ends of these fragments and the final library is constructed.

In the second phase, the library is provided as input for the NGS machine, that sequences the fragments and generates as output a file containing the corresponding set of reads. A *read* is the sequence of nucleotides that constitute a fragment, inferred from the sequencing steps.

The last phase is dedicated to data analysis. When cDNA is fragmented, the order of all the pieces is lost and also the reads are not ordered. In this step, to reconstruct the whole sequence, reads can be aligned by mapping them to a reference genome. Once

aligned, they can be assembled in transcripts. Transcripts can be reconstructed through *alignment*, by inferring the transcriptome sequence from the abundance of reads aligned to a reference genome, or they can be reconstructed *de novo*, by using a reference genome or annotations as a guide.

At this point, the expression levels of the genes can be computed by counting the number of mapped read that align with the obtained transcriptomes. Because the number of reads is influenced, among other variables, by the library size and the gene length, the raw counts must be corrected with ad-hoc metrics. The final expression levels are usually computed as FPKM (fragment per kilobase of transcripts per million mapped reads) or RPKM (reads per kilobase of transcripts per million mapped reads).

2.3.2 RNA-sequencing decomposition

In the traditional sequencing methods are used millions of cells coming from the same sample. If the purpose of the sequencing is analysing samples coming from a patient affected by cancer and obtaining the expression levels of the genes in those samples, it is necessary to highlight that only a portion of the whole sample is actually constituted by cancerous cells.

When a single sample is taken, in fact, it is composed in various proportions by tumor cells, fibroblast, immune cells and other kind of cell types. This means that when the expression levels are computed, they are not referred only to the cancerous component in the sample, but to the whole ensemble of cells. The analysis made on these data needs of course to take this heterogeneity into account, because the results would be influenced by it. A way to retrieve only the expression levels of a certain cell type is to use the *single cell sequencing*, in which is sequenced only one cell at a time. This kind of analysis is more challenging than the one previously described.

A possible alternative to single cell analysis is the *decomposed RNA-seq analysis*. This kind of process tries to identify the sample composition and the cell-type specific expression patterns that compose each sample. The portion of each cell type in the original RNA-seq sample is not known a priori, but can be inferred through the use of an iterative expectation maximization algorithm [15] from single cell data. Once the composition is known, it is

possible to decompose the expression levels of the original sample and obtain cell-type specific values.

The basic idea that underlies this procedure is shown in the figure 2.3. On the left side of the figure there are the expression levels of the original samples and on the right side of the figure there are the decomposed ones. The decomposed expression levels obtained with this technique are 4 and are related to the epithelial ovarian cancer cells, the fibroblasts, the immune cells and the remaining cell types.

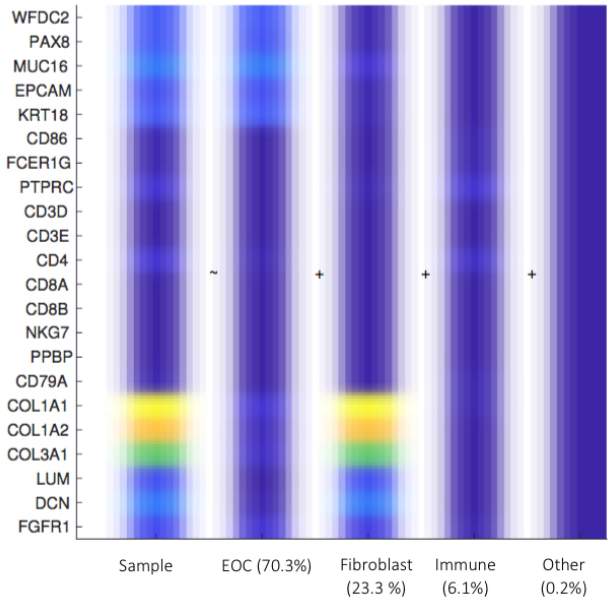


Figure 2.3. Decomposed RNA-seq sample

Chapter 3

Background on computational science methodologies

3.1 Machine learning

Machine learning is a branch of artificial intelligence whose goal is building a system that is able to learn how to solve a task without actually being programmed for it. The system, by adopting a bottom-up approach, is in fact able to extract the rules necessary to solve a problem from experimental data and experience. There are 3 different types of learning:

- the *supervised learning*, in which the learning process is supported by some previous knowledge about the data. The system, in fact, is trained with data for which the final outcome is known. These data are called "labelled";
- the *reinforcement learning*, based on the development of a system that continuously self-improves by interacting with the environment. Each time the system takes a decision, it receives a reward/punishment value in return as a measure of the goodness of the choice taken. The learning method is based on the trial-and-error approach and the system gains knowledge on the basis of the feedbacks it receives after each decision;
- the *unsupervised learning*, in which the algorithm simply explores the data at disposal

without any guidance. It is employed when the relationships existing between the data are not known and there is no previous knowledge on the dataset.

3.1.1 Supervised learning

The supervised learning approach is used when it is necessary to predict the outcome of unlabelled data on the basis of the knowledge acquired with the labelled ones. For this reason, it is employed both in classification and in regression problems. The term "classification" is used when the outcome of the prediction is a categorical class label, while the term "regression" is used when the prediction outcome is a continuous value.

A machine learning model that uses a supervised approach is composed by several phases: preprocessing, learning, evaluation and prediction. The objects that constitute the dataset are represented by a certain number of attributes or measurements called *features* and each of them has a *label* that represents the final expected outcome.

During the preprocessing phase, the complete dataset is partitioned into a training and a test set. The training set is the ensemble of data instances used for training the machine learning algorithm and building the model, while the test set is the ensemble of instances used for testing the obtained model and assess its performances. Usually, the training set is composed by the 70% of the initial data set and the test set is composed by the remaining 30%. Preprocessing is also necessary for elaborating the data before dividing them in the two set. This elaboration can include: the rescaling of the numerical values of the features in a different range, the encoding of the categorical features as numerical quantities and the reduction of the number of features to be taken into account by the system. The last problem is known as "feature reduction" problem and it is faced when there is an high number of features for each observation. This can limit the both the storage space and the computational performances of the machine learning algorithm. The two main approaches exploited for feature reduction are the dimensionality reduction and the feature selection. In the first approach, the original features are mathematically recombined in a new set of features having lower dimensionality. Well-known examples of dimensionality reduction algorithms are the Principal Component Analysis (or PCA) and the Linear Discriminant Analysis (or LDA). In the second approach, instead, a subset of the most informative

features is extracted from the original features set. According to the literature [30], feature selection methods are divided in:

- *wrapper methods*, which involve the usage of a ML algorithm. Subsets of features are iteratively used to train models and features are selected according to the models' performances;
- *filter methods*, which involve statistical steps that compute the variable correlation with the correspondent outcome. They do not involve any learning algorithm;
- *embedded methods*, which are a combination of filter and wrapper techniques.

The learning phase is the core phase of a machine learning system. In this step, the selected machine learning algorithm is trained with the training set and the final model is built. Among the numerous machine learning algorithms that can be used in this phase, it is worth to mention the Random Forest and the Support Vector Machine. There is not an optimal algorithm for all the problems, so each time a machine learning system is built it is necessary to understand which algorithm has the best behaviour according to the data at disposal and according to the task to solve.

During the evaluation phase, there is the assessment of the performances of the model built in the previous step. Quantifying the performances of the obtained model is necessary to understand the confidence of the following results and evaluate which is the best machine learning algorithm for a specific problem. In this phase, the instances composing the test set are used as input for the model and the predictions obtained as output are compared to the data labels. Ideally, it would be desirable to have a large number of data to divide in training and test sets but in the reality there is not always this possibility. In this case, it is possible to iteratively sample the initial dataset in order to create at each iteration different combinations of the two sets. This implies that several models are generated and tested. The most straightforward way to do it is by using the crossvalidation. This procedure partitions the labelled dataset in a certain number of subsets (also called "folds") and then starts an iterative procedure in which, at each iteration, one of this folds is used as testing set and all the others are used as training set. Well-known cross-validation techniques are:

- *k-fold cross-validation*, in which the number of created folds is equal to k;

- *leave one out cross-validation*, in which the number of folds is equal to the number of instances in the dataset.

The evaluation of performances implies the choice of a performance metric. Among all the existing metrics that can be used, the most famous ones are the confusion matrix, the ROC curve and the accuracy.

The prediction phase is the last phase of the machine learning system and it is the phase in which the model is actually used to predict the outcome of unlabelled data.

3.1.2 Unsupervised learning

The unsupervised learning approach is employed when the relationships existing between the data are not known and there is no previous knowledge on the dataset. The algorithm simply explores the data at disposal without any guidance. This technique is majorly employed for clustering and dimensionality reduction purposes.

With "clustering" it is intended the partitioning of data instances into a certain number of classes in order to minimize the similarities between the elements of different classes and maximize the similarity among the elements of the same class. Data objects are usually seen as point in the clusterization space and the similarity measure is usually seen as the distance existing between these points. The types of clustering methods can be classified, according on how they divide the clustering space, into:

- *hierarchical methods*, in which the obtained clusters are hierarchically organized in a tree;
- *partitioning methods*, in which objects are divided into non-overlapping clusters. Each object belongs to one and only one cluster.

3.2 Statistical concepts

In a statistical hypothesis test are generally compared two groups of samples in order to accept or dismiss a certain statement, called *null hypothesis* or H_0. The null hypothesis is compared to the *alternative hypotesis* or H_1. A result is statistically significant if it has low probability to happen under the null hypothesis. The significance of the result can be computed through the use of the *p-value*. The p-value is a numerical value between 0 and 1. It is defined as the probability to obtain the observed result or a more extreme one, under the condition that the null hypothesis is true. This means that if it is obtained a p-value which is higher than a significance threshold, the null hypothesis cannot be rejected; while if the p-value is lower than the significance threshold, the null hypothesis is rejected because there is strong evidence against it. As a rule of thumb, the significance level is usually set to 0,05.

Different statistical hypothesis tests exist and they can be divided in two major categories, according to the assumptions they made. It is possible to have a *parametric* test, in which a normal data distribution is assumed, or a *non-parametric* test, in which no assumption on the shape of the data distribution is made. This means that the last category of tests can deal with non-normal or highly skewed data. The most famous parametric test used is the Student t-test, which verifies that the mean value of a distribution differs from a certain value. Among the non-parametric tests, instead, the most used ones are the Kruskall-Wallis and the Mann-Withney-Wilcoxon tests.

When a statistical result is computed, there is always the possibility to obtain small p-values just by chance. In those cases the results are affected by errors. In particular, the term "type I error" (or *false positive*) is used when there is the erroneously rejection of the null hypothesis, while the term "type II error" (or *false negative*) is used when there is not the rejection of the null hypothesis, even if this last one is false.

When multiple statistical tests are performed, the chances to obtain false positives increase. This means that a portion of the observations believed statistically significant is actually wrong. In order to correct the obtained results, it is possible to follow two main approaches: controlling of the *Family-Wise Error Rate* or controlling the *False Discovery Rate*. In the first case, the Family-Wise Error Rate (or FWER) is the probability of

making at least one type I error in a multiple statistical test. Controlling the FWER means applying the Bonferroni correction, in which a new statistical test is performed on each hypothesis of the family by setting a lower significance threshold. In the second case, the False Discovery Rate (or FDR) is the proportion of false positives among all the rejected null hypothesis. For controlling it, the Benjamini-Hochberg correction can be applied.

Chapter 4

Materials and Methods

4.1 Workflow

In the diagram in figure 4.1 it is possible to see the workflow of the analysis described in the following.

The analysis is divided in two major branches according to the input data used. The first one (in red) is based on the employment of the expression levels obtained from RNA sequencing and it consists of a basic explorative step made to understand the information contained in the data. In the second step (the one in purple in the figure), decomposed RNA-seq data are used to find candidate genes that can be chemotherapy-related in HGSOvCa. The analytic approaches adopted in this phase are two: an unsupervised technique based on statistical data processing and hierarchical clustering (depicted in light purple in the figure) and a supervised technique based on feature selection through machine learning methods (in dark purple). All the scripts realized to perform this analysis are written in Python and R.

Before the description of the analysis, there is an overview of the sample collection and the available input data.

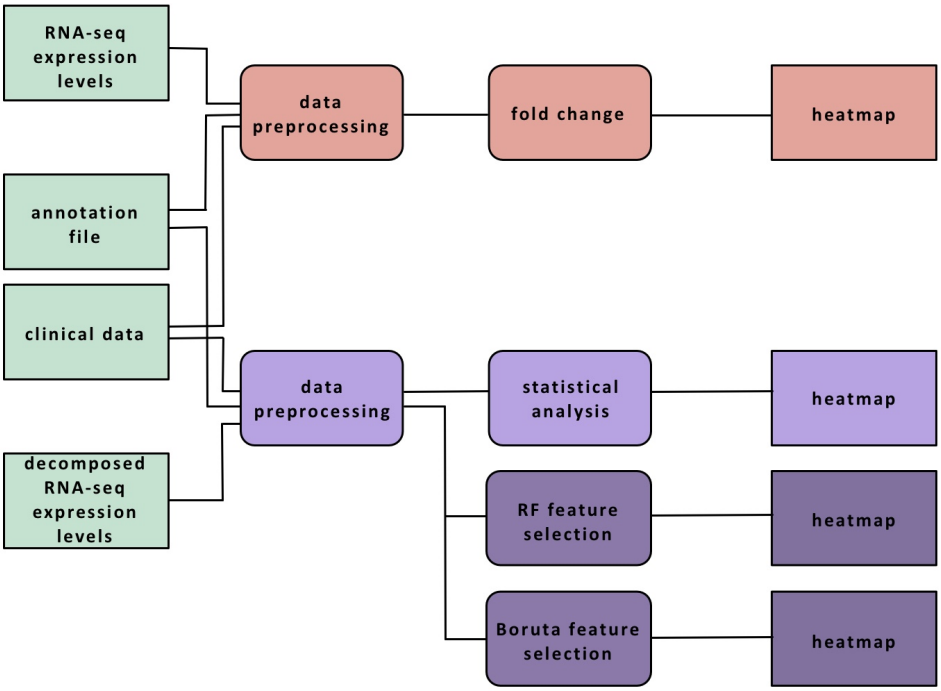


Figure 4.1. Workflow of the analysis

4.2 Input data description and pre-processing

4.2.1 Sample collection

The data used for this analysis come from high grade serous ovarian cancer patients. After the disease is diagnosed, the patient is enrolled in a NACT-IDS or a PDS treatment. According to the type of treatment, the patient can undergo to a first set of neoadjuvant chemotherapy cycles or to a primary debulking surgery. In any case, before the beginning of the therapeutic process, tumor samples are collected and those samples are called "primary". In case of NACT-IDS, after the chemotherapy, the patient faces an interval debulking surgery to remove the tumoral tissues. After this surgical procedure, other samples are collected. These are called "interval" samples. Interval samples are also sometimes gathered after the first adjuvant chemotherapy cycles in case of a PDS treatment. Once those samples are collected, the patient undergoes to another adjuvant chemotherapy treatment, after which the follow-up period starts. In this period of time, the patient performs

a series of periodic controls to monitor the treatment outcome and the disease progression. If the tumor shows up again, new samples are taken from the patient and these ones are called “relapse” samples.

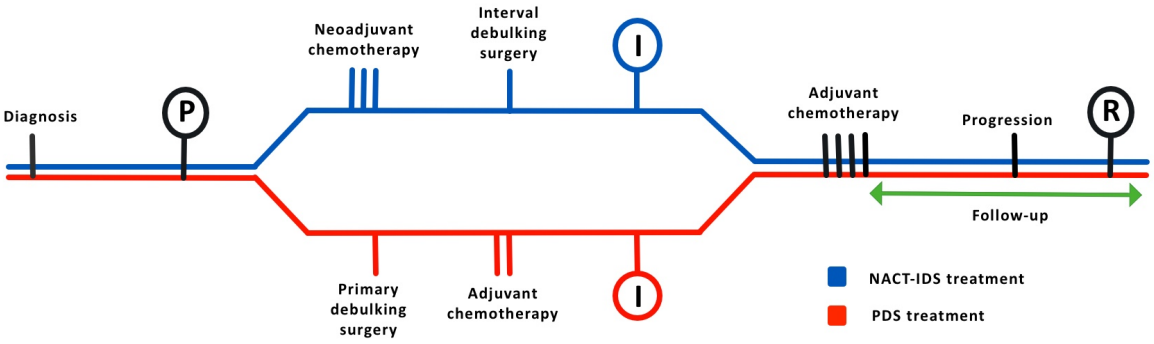


Figure 4.2. Samples collection time graph

Several samples can be collected at the same time from a single patient and, in order to distinguish them, each sample and each patient has a unique identifier. Patients are identified by an ID composed by:

- one or two capital letters (M, H or OC);
- an integer number of three digits.

Samples, instead, are identified by:

- the patient ID;
- an underscore character;
- a letter that identifies the time point in which the sample was taken (“p” for primary, “i” for interval and “r” for relapse);
- an acronym for the tissue type from which the samples was taken (“Mes” for Intestine/Mesenterium; “Napamet” for a metastasized tissue; “Ome” for Omentum; “OvaR”/”OvaL” for Ovary right or left; “Adn” for Adnex; “Per” for Peritoneum; “Asc” for Ascites; “LN” for Lymphonode; “Ute” for Uterus; “TubR”/”TubL” for Fallopian Tube right or left; “PerFd” for Peritoneum/Fossa Douglas);

- an integer number identifying the number of samples taken.

The three kind of collected samples are really different between them. This is due to the fact that cancer is on its own a disease that continuously evolves during its development. The cells that compose the tumor when it is established, in fact, are really different from the ones that could be found during oncogenesis [42]. Even after the establishment, cancer cells continuously undergo to genetic mutations and epigenetic alterations that contribute to this heterogeneity. In addition to that, primary, interval and relapse samples are collected in different time points during a treatment that, of course, causes changes in the cells. Because interval samples are taken after a first chemotherapy cycle (NACT or ACT), in fact, they contain a lower percentage of cancer cells with respect to primary samples. Also, it is not always possible to take interval samples from the same tissues of the primary ones, because some tissues could have been removed with the surgery. This means that, in this cases, it is impossible to compare the same tissue before and after the chemotherapy. Relapse samples are again different from the primary and the interval ones because the cells that compose them are chemoresistant cells, completely different from the original tumoral cells populations.

The analysis described in the following chapters relies on 130 samples from 41 patients. After the samples are collected, they are sequenced through RNA-Sequencing techniques and the expression levels of the genes are computed. During the whole therapeutic process, clinical data are also collected.

4.2.2 Expression level data

The expression level data are contained in two CSV tab-separated files. Each entry of these files represents a genomic element, identified in the first two columns by its *gene ID* and its *gene name*. The expression levels are computed after RNA-sequencing, quantified by using *eXpress*[45] and transformed from raw counts to log2(TPM). TPM stands for "Transcripts Per Million" and it is a quantity obtained by normalizing the Read Per Kilobase (RPK) for the sequencing depth. RPK is computed by dividing the read counts for the gene length expressed in basepair. The sequencing depth, instead, is given by the sum of all the RPK values in a sample, divided by one million.

One file contains the original expression levels obtained by sequencing the samples and the other file contains their decomposed version. In this last file, for each sample and for each lncRNA are stored: the expression level in the original sample, the one related to the epithelial ovarian cancer profile, the one related to the fibroblast profile, the one related to the immune profile and the one related to the remaining cell types.

4.2.3 Data annotation

Data annotation are additional information on the genomic region for which expression levels are computed. This kind of data are stored in a separate file produced by the GENCODE consortium [43] of the National Human Genome Research Institute (NHGRI) within the ENCODE project [44]. One of the file format used to store the annotation data is called "BED" (Browser Extensible Data). A generic BED file can have from three to twelve columns. In fact, this standard imposes the use of three required fields ("chromosome name", "genomic start location" and "genomic end location") and nine additional optional fields. The number of fields per line is consistent throughout any single set of data in an annotation track and fields are tab-delimited.

In the annotation file at disposal for this analysis there are 10 fields and the last one contains some additional information. The additional information field is generally made by an arbitrary number of key-value pairs. The pairs have a *<key "value"; >* format. Also this field is composed by several mandatory information and several optional ones. Each one of the 2 579 817 entries stored in the file contains the information for a RNA region. Examples of the possible information that can be retrieve from this file are: the name of the chromosome from which the sequence is extracted, the strand, the starting and the ending positions of the region within the chromosome. These additional data, together with the expression levels, can be used to better identify a genomic region and the biological functions in which it is involved.

For this analysis, the only additional information used is the *gene biotype*. The gene biotype describes the genomic sub-type of a specific RNA region. This fields is used in order to extract from the expression level files only the entries corresponding to long non-coding RNAs. Among all the possible biotypes, the ones of interest for this work of thesis

are the ones listed below:

- *non_coding*, generic elements that do not code for protein;
- *3prime_overlapping_ncRNA*, lncRNAs located within the 3' UTR of protein-coding genes;
- *antisense/antisense_RNA*, lncRNAs overlapping the genomic span of a protein-coding locus on the opposite strand;
- *lincRNA*, lncRNAs that can be found in evolutionarily conserved, intergenic regions;
- *sense_intronic*, llncRNAs within a coding gene that does not overlap any exons;
- *sense_overlapping*, lncRNAs within the intron of any protein-coding on the coding strand;
- *macro_lncRNA*, unspliced lncRNAs that are several kilobase in lenght;
- *bidirectional_promoter_lncRNA*, divergently transcribed lncRNAs which originate from the promoter region of a protein-coding gene;
- *misc_RNA*, miscellaneous RNAs.

4.2.4 Clinical data

The original clinical data file is a CSV file storing the clinical information of each patient participating to the study. Clinical data are collected throughout the duration of the study and they have different nature. Some of them are personal data (like age), others regard the past clinical history of the patient and others are related to the therapy the patient is following. In the clinical data file at disposal, there is one entry for each sample and the clinical information of interest in this analysis are: the patient and the sample identifiers, the *time to progression* and the *follow up time*. The time to progression is a float number that identifies the number of months for which the patient is free from the disease after the end of the therapy. If this information is not available, the corresponding field in the file contains the value NA. If, instead, the information is available, it shows that the patient had a relapse phase. The follow up time, instead, is a float number that specifies the

number of months for which the patient has been followed up. Those two values are used in order to classify patients (and therefore samples) in two groups: the ones that develop resistance to the treatment and the ones that do not. In this analysis, if the recurrence of the disease happens within a year after the end of the treatment, the patient is considered *chemoresistant*, because she developed chemotherapy resistance in few months. Instead, if the disease does not show up again or there is a relapse after more than one year, the patient is considered *chemosensitive*. Due to the aggressiveness of this malignancy and its high mortality rates, that lead to death in really short time, in fact, a relapse after more than twelve months is considered as a positive result. By expressing the resistance/sensitivity to chemotherapy in terms of months in which the patients is exempt from the development of resistance to platinum-based drugs after the treatment, it is possible to use the term *platinum-free interval*. In this way, patients identified by a *short* platinum-free interval are chemoresistant patients, while the ones characterized by a *long* platinum-free interval are chemosensitive ones.

4.2.5 Data preprocessing

The first step of the analysis is to retrieve from the initial input files only the information needed for the following processing. According to the biotypes in the annotation data, from the expression level files are extracted only the rows corresponding to long non-coding RNAs. In this way are obtained two files: one containing the original expression levels of the lncRNAs and another containing the tumoral decomposed version of the original expression levels for each lncRNA. Both files have 27 141 rows (the number of extracted lncRNAs) and 132 columns (the number of samples plus a column for the gene ID and another for the gene name).

The 130 samples at disposal come from 41 patients, 3 different time points (primary, interval and relapse) and 13 different tissues. 14 of these samples have cell line origin, while the remaining 116 have tissue origin. In the whole analysis, only tissue origin samples are taken into consideration. According to time point, the samples can be divided in 81 primary samples, 41 interval samples and 8 relapse samples. The investigation focuses only on the primary set.

4.3 RNA-seq data exploration

In this section it is described the first step of the analysis, in which input data are explored using hierarchical clustering and visualisation methods. In this phase, the output of the unsupervised clustering is analysed to understand the content and the characteristics of the input dataset.

Hierarchical clustering is a kind of unsupervised clustering method that divides the input observations in a set of groups hierarchically organized. According to the kind of strategy used to create the hierarchy of clusters, we can distinguish between a *top down* (also called "*agglomerative*") approach or a *bottom up* (also known as "*divisive*") approach.

In the first one, the algorithm starts by considering N different clusters, each one containing one of the N available observations. From this, it aggregates at each step the two clusters that have the smallest intergroup dissimilarity until all the observations are in a unique cluster. In the second method, the algorithm starts by considering one cluster containing all the N observations. At each step, it splits the cluster in two sets of observations having the largest between-group dissimilarity. The measure of dissimilarity between groups of observations is decided by setting a *metric* and a *linkage criterion*. A metric is a measure of distance between two observations and in the following we have examples of some of the most commonly used ones:

- the euclidean distance, computed as $d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$;
- the square euclidean distance, computed as $d^2(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n}(q_i - p_i)^2$;
- the manhattan distance, computed as $d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{q}, \mathbf{p}\|_1 = \sum_{i=1}^{n} \|q_i - p_i\|$;
- the canberra distance, computed as $d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$.

The linkage criterion, instead, is the measure of the distance between groups of observation as a function of the chosen metric. The main alternatives are:

- *maximum or complete linkage clustering*, in which the distance between two clusters is computed as the maximum distance between two observations belonging to the two different clusters. In formula: $L(r, s) = max(D(x_r i, x_s j))$;

- *minimum or single-linkage clustering*, in which the distance between two cluster is computed as the minimum distance between two observations belonging to the two different clusters. In formula: $L(r,s) = min(D(x_r i, x_s j))$;

- *mean or average linkage clustering*, in which the distance between two cluster is computed as the average of all the distances between all the observations belonging to one cluster and all the observations belonging to the other one. In formula: $L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_r} D(x_{ri}, x_{sj})$.

In all the options, the distance is the metric chosen to compute the dissimilarity between two observations. The choice of the metric and the linkage criterion influences the shape and the organization of the final hierarchy. There is not a winning combination of the two criteria that works in a perfect way for all the data. Each set of data and each research question are unique and require to be properly analyzed.

The results of the hierarchical cluster algorithm can be presented through a binary tree called *dendrogram*. The highest level of the tree, the root, contains one cluster with all the observations. The lowest levels, which correspond to the leaves, contain as many cluster as the number of observations and each cluster contains just one element. In a dendrogram, the height of each node (the distance between two nodes in a branch - so the distance between a child node and its father) is proportional to the amount of dissimilarity between two siblings nodes (the two child nodes of the father node). The number of levels in the hierarchical tree is N-1, where N is the number of initial observations. Hierarchical clustering do not require to specify a priori the number of clusters to be obtained, but by looking at the dendrogram it is possible to retrieve a certain number of clusters by cutting the binary tree in an horizontal way. This procedure is equivalent to setting a dissimilarity threshold and stopping the hierarchical clustering algorithm once that threshold is reached.

In bioinformatics, hierarchical clustering is often used in combination with a heat map, in order to understand if from the unsupervised clustering method emerge some groups of observations with some particular characteristics. A heat map (or heatmap) is a two-dimensional graphic visualization of data contained into a matrix, whose values are represented through colors. Those maps are widely employed for having an intuitive visualization of gene expression data. When the hierarchical clustering is applied to a heatmap, the data

inside the map on which the algorithm is applied are re-organized, clustered together and a dendrogram is shown.

In this analysis, to identify chemotherapy-related genes, patients and samples are divided in two groups on the basis of the clinical information. The distinction is made according to the *platinum free interval* (PFI), the amount of time after the treatment in which patients are free from the disease. Samples coming from patients who had a relapse in the 12 months following the therapy are classified as "short PFI", while the ones coming from patients who had a relapse phase after 12 months or who did not have progression of the disease after a year are classified as "long PFI". The division is made by considering the "time to progression" and the "follow up time" fields of the clinical data file.

In order to reduce the dimensionality of the problem and facilitate the visualization, in this phase it is considered only a subset of the lncRNA biotypes. This subset was decided with the supervision of a geneticist and it includes the most well-known lncRNAs biotypes in literature[46]:

- non-coding RNAs;
- long intergenic non-coding RNAs;
- macro long non-coding RNAs;
- bidirectional promoter long non-coding RNAs;
- antisense RNAs;
- miscellaneous RNAs;

Always in the view of reducing the dimensionality of the data, also the lncRNAs having in all samples expression levels lower than 1 are removed from the dataset. These genes, in fact, are judged as not informative for the research question because they do not show high differences in the expression levels across all the samples, but they just contribute to the computational load.

From the pool of remaining samples are extracted the ones belonging to primary time point. Then, for each lncRNA are computed:

- the average of the expression levels in the "long PFI" samples;

- the average of the expression levels in the "short PFI" samples;
- the absolute difference of these two values.

The last quantity computed is also called *fold change* and below there is an histogram of the fold change values obtained for the primary samples.

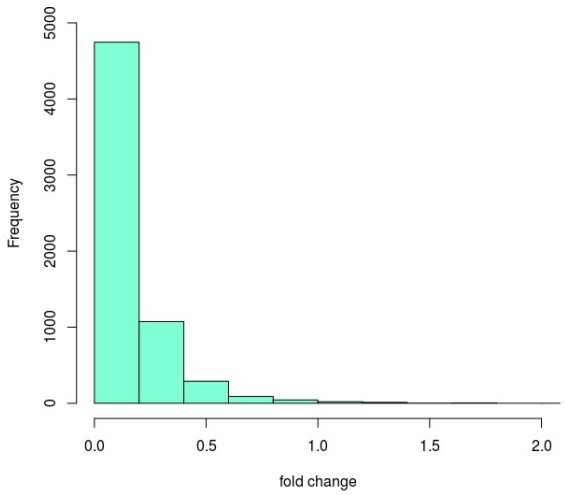


Figure 4.3. Histogram of the fold change values in primary samples

The graph shows that the majority of the lncRNAs have small differences in the averages of the expression levels between the two groups. In order to visualize only the genes having notably differences in the expression levels between the two sets, are selected the lncRNAs having fold change value bigger than 1.2. The expression levels of the resulting 20 genes in the primary samples, together with the correspondent clinical data, are visualized using a heatmap that is shown in the figure 4.4.

For the hierarchical clustering function it is possible to specify the metric and the linkage. In this case, it is decided to use the *euclidean distance* as metric and the *complete clustering* as linkage method. In this way, the distance between groups of observations is computed as the maximum euclidean distance between two observations belonging to two different clusters.

By looking at the plot in figure 4.4, it is possible to see that in the upper part of the heatmap, in correspondence of each sample, are displayed some of the clinical data

information: the progression interval, the survival, the primary therapy outcome, the follow up time and the progression. The legend on the right explains for each variable all the possible values. In this heatmap, as well as in all the following plots, the patient ID is replaced for privacy reasons with an alternative name, computed as the concatenation of the string “patient_” and a progressive number.

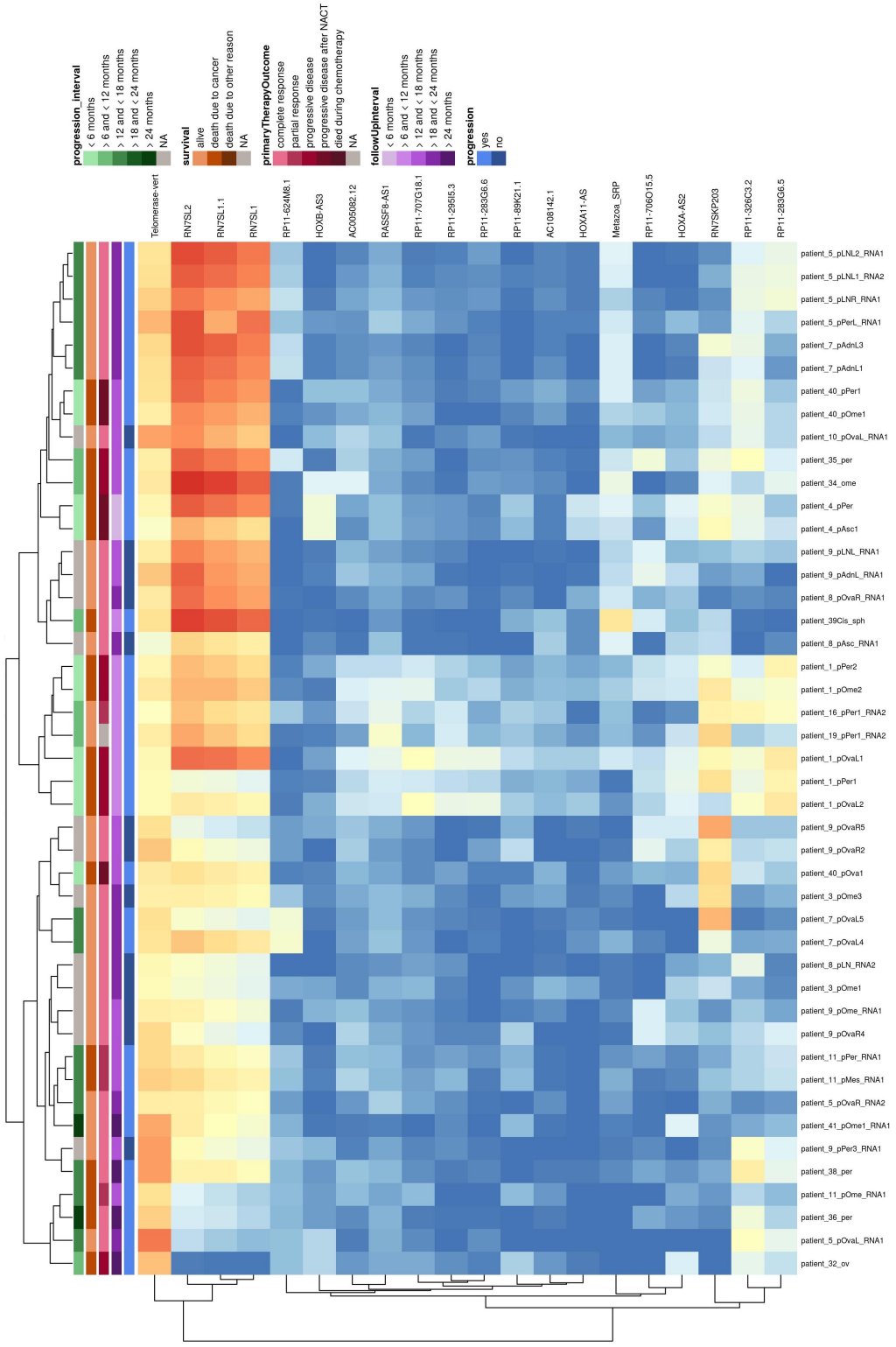


Figure 4.4. Heatmap of lncRNAs having fold change > 1.2 - primary samples

4.3.1 Results

From the upper dendrogram in the primary samples heatmap, it is possible to see that the hierarchical clustering mainly divide the samples in two groups. By cutting the dendrogram at the first level, in fact, it is possible to see that the obtained sample sets almost follow the division between "long PFI" and "short PFI" that was previously made by hand. With some exceptions, in fact, the left side of the heatmap groups together the samples having a longer platinum free interval, while the right side of the map groups together the ones having shorter PFI. By analysing the lower levels of the dendrogram, it is possible to understand that samples coming from the same patient mainly cluster together. If similar samples group together and we have different numbers of samples per patients, it means that patients with an higher number of samples will drive the clusterization, so the results can be actually affected by some patients more than others.

Finally, looking at the dendrogram on the left, the one referred to the long non-coding RNAs, it is possible to see that there are two clear groups of genes. The one at the top of the heatmap is the most interesting one, because it contains long non-coding RNAs that clearly shows high expression values in "short PFI" primary samples and lower expression values in "long PFI" primary samples.

4.4 Decomposed RNA-seq analysis

4.4.1 Data selection

From the previous step it can be seen that the obtained clusters of patients are not perfect and that clusterization is driven by patients having an high number of similar samples. For these reasons, it is decided to perform the next investigation step with a different approach. The idea is to consider just one sample per patient, by averaging the expression levels of the samples coming from the same patient and the same time point. For doing that, it is necessary to further investigate the similarities and the differences in the expression levels of the samples to average. To prevent data alteration, in fact, only samples containing similar expression levels can be averaged. In this way, it is possible to investigate the overall behaviour of similar samples. To analyse the data, the Principal Component Analysis is performed.

The Principal Component Analysis (or PCA)[16][17] is a mathematical procedure that aims to find a new coordinate system of uncorrelated variables starting from a dataset described by correlated parameters. The new set of variables has the same dimensionality of the original one, with the advantage that it can be reduced in a way in which the remaining components describe most of the variation in the data. In this sense, the PCA can be used in a machine learning system as an unsupervised dimensionality reduction technique. The assumptions of the principal component analysis are that the variables are normally distributed and the data are represented by real and continuous values.

From the mathematical point of view, the new variables are a linear combination of the original ones. In order to obtain orthogonal variables related to the variance of the data, the PCA is based on the computation of the eigenvectors and eigenvalues of the covariance matrix associated with the dataset. Because covariance is used, the original dataset is transformed in a mean-zero matrix by subtracting the variable's means to each variable value. Then, the associated covariance matrix, eigenvalues and eigenvectors are computed. The eigenvalues quantify the variance expressed by the eigenvectors; while the eigenvectors, once normalized, represent the coordinates of the new reference system. The eigenvectors are ranked in descendent order according to variance they express and

organized in a matrix of column vectors. At this point, the new variable are computed as:

$$\mathbf{Y} = \mathbf{XW} \tag{4.1}$$

where X is the (n x m) matrix containing the mean-adjusted dataset; W is the (m x m) matrix of eigenvectors (PCA coefficients); Y is the (n x m) matrix containing the new set of coordinates.

The new set of variables can be restricted by considering only the first p eigenvectors that express most of the variance in the data.

The principal component analysis is useful not only for reducing the dimensionality of data, that -among other things- helps the visualization of data, but also because it facilitates the detection of subsets and outliers. In addition, it contributes in the definition of the scaling relations existing among data and the correlation between the original variables.

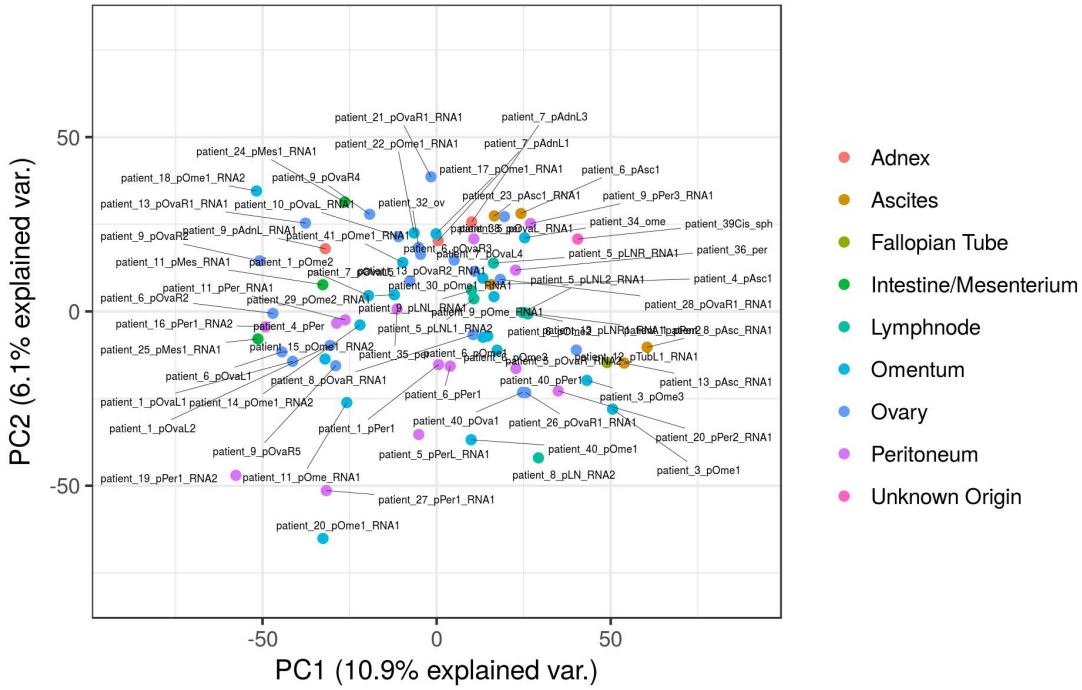
In this analysis, the Principal Component Analysis is used because there is the need to evaluate the similarities between the samples and decide if it is possible to merge them together or not. Each sample is described by several feature values that in this case are the expression levels of the long non-coding RNAs in the sample. Because the interest is focused on the overall behaviour of each sample, the PCA can be used to ease the analysis by reducing the number of features that describe the sample's characteristics.

Unlike the previous analytical step, in this one (and in all the following ones) are taken into account all the different lncRNA biotypes without distinctions. In this part of the analysis is used the decomposed version of the RNA-seq data, in order to take into account only the expression levels related to the tumoral component of each sample. In the previous analysis, in fact, the expression level values did not refer only to the tumor cells but to whole ensemble of cells types in the sample. Again, in order to reduce the dimensionality of the data, lncRNAs having expression levels lower than 1 in all the samples are excluded.

Samples are divided according to time point and the principal component analysis is applied on the primary set. Only the first two principal components are taken into account. The results of this step can be see in the figure 4.5.

From the plot, it can be seen that also with the decomposed expression level data, the samples coming from the same patient tend to cluster together. The clusterization, of course, is not perfect and this can be due to the sample's tissue type. Samples coming

Figure 4.5. PCA results for primary samples



from ascites and adnex, fox example, do not group with the others and because of those differences they have to be removed from the sample set. Biologically speaking, there are also differences between the tissues involved in the early stages of the disease (the tissues from which the cancer originates, like ovary and fallopian tubes) and the ones involved in the late stages (the tissues where the tumor spread). It was already said, in fact, that this kind of disease is in constant evolution and that it continuously produces changes. Changes in the biological processes lead to changes in the expression levels. The final decision is to take into account only the samples coming from the tissues involved in the late stages of the disease: peritoneum, peritoneum/fossa douglas, intestine/mesenterium and omentum. The samples coming from those tissues are extracted from the initial sample set and for each lncRNA are averaged the expression levels in the samples belonging to the same patient and the same time point. The results are stored in a new table containing the lncRNAs on the rows and the new collapsed samples on the columns.

4.4.2 DEG extraction through statistical analysis and unsupervised clustering

As shown in the workflow at the beginning of this chapter, different approaches are adopted to retrieve differentially expressed genes related to the chemotheraphy response. In this section it is described the one based on statistical processing and unsupervised hierarchical clustering. The underlying idea is to use a statistical test in order to reduce the dimensionality of the problem and identify a set of long non-coding RNAs that show significantly differences in the expression levels between the chemoresistant and the chemosensitive patients. In order to confirm that these genes are actually related to the platinum-free interval, it is shown that they can be used as clusterization features. In fact, if the hierarchical clustering algorithm, on the basis of the lncRNAs selected in the statistical step, is able to cluster the samples as they were initially grouped according to the PFI, it shows that those genes are really correlated with the chemotherapy resistance.

A statistical test is performed in order to compute the p-value related to the difference in the expression levels of each long non-coding RNA and to extract only the significative ones. To apply it, data are again divided into two populations according to the PFI value. Among the 26 new collapsed primary samples, only 14 have the needed clinical data (time to progression and follow up time). The expression levels of these samples are retrieved from the new expression level table and the samples are divided as before in “short PFI” and “long PFI”. The obtained groups are perfectly balanced, because each of them contains 7 samples. For each long non-coding RNA it is then computed the average of the expression levels in both groups and the fold change. The distribution of the data is plotted in order to understand the distribution of the values. The plots are represented in the figure 4.6 and the results show highly skewed data in both the populations.

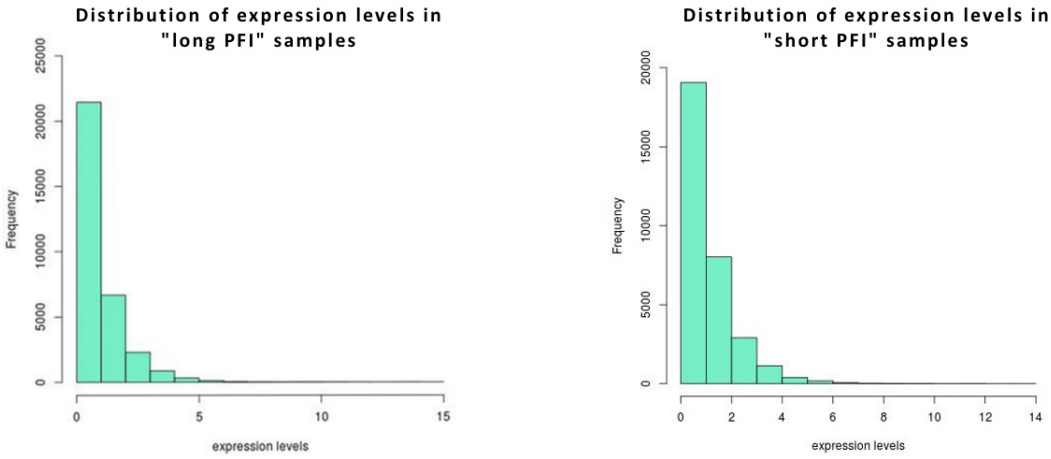


Figure 4.6. Distributions of the expression levels in primary samples

The shape of the distribution influences the choice of the statistical test to use. In this case, because of the skewness of the data, it is not possible to assume a normal distribution and consequently it is not possible to apply any parametric statistical test (like the commonly used Student-Welch t-test). The non-parametric Mann-Whitney-Wilcoxon test is then employed for the computation of the p-values for each long non-coding RNA.

The Mann-Whitney-Wilcoxon test (also known as "Mann-Whitney U-test" or" Wilcoxon rank sum test")[18] is a statistical hypothesis test. Because this test does not have limitations regarding the distribution underlying the data, it is classified as a *non-parametric* test. The Mann-Whitney-Wilcoxon test has three assumptions:

- the two sets of data have the same distribution;
- the two sets of data have the same variation (property called *homoscedaticity*);
- all the observations in both sets are independent from each other.

The null hypothesis supports the equality of the distribution of the scores for the two groups. If it is also true that the distribution of the two sets are equally shaped, as in this analysis, the test can determine if the medians of the two distributions are statistically different or not. The null hypothesis, in this case, supports the equality of the two distributions while the alternative one supports the inequality of the medians.

By considering the usual statistical significance threshold at 0.05, it is possible to retrieve from the results of the Mann-Whitney U-test 246 long non-coding RNAs that show differential expression between chemoresistant and chemosensitive patients. Anyhow, since the test involves an high number of genes, a multiple testing correction needs to be performed after it. The main alternatives in this case are the application of the Bonferroni correction or the application of the Benjamini-Hochberg method. The Bonferroni correction is a conservative technique and it is necessary when a single false discovery in the results can be dangerous. Nevertheless, when the number of multiple comparisons is too high, as in this case, it can be too stringent and it can lead to a very high rate of type II errors. For this reason, it is chosen to apply the Benjamini-Hochberg correction to compute the adjusted p-values.

The Benjamini-Hochberg approach was designed to control the false discovery rate, that is the proportion false positives among all the rejected null hypothesis[19]. This method ranks the p-values obtained from a statistical hypothesis test and then compute for each of them a quantity that can be expressed by the following formula:

$$\frac{i}{m}Q \tag{4.2}$$

where i s the ranking value, m is the total number of performed tests and Q is the chosen false discovery rate. All the observations having a p-value $< \frac{i}{m}Q$ are considered significant. The choice of the false discovery rate value (Q) depends on the data and the situation under study. It is possible to compute another quantity called "adjusted p-value" by considering the original p-value (or "raw p-value") multiplied by the $\frac{m}{i}$ quantity. In this case, the observations having adjusted p-values smaller than the chosen false discovery rate are considered significant. The only assumption of this method is the independence of all the individual tests.

The results of the Benjamini-Hochberg correction show that the smallest corrected p-value is 0.68, really far from the significance threshold. This means that they cannot be interpreted as statistically significant. For these reason, it was decided to rely on the raw p-values and the fold change values for the extraction of the differentially expressed genes.

The figure 4.7 shows a volcano plot obtained by representing the $log2$ of the fold change on the x-axis and the $-log10$ of the raw p-value on the y-axis. In blue are highlighted the

lncRNAs having fold change greater than 1 and raw p-value lower than 0.05. Those genes are, according to this kind of analysis, the ones showing higher differences in the expression levels between a group of chemoresistant patients and a group of chemosensitive ones. The p-value is taken into account in order to retrieve the statistically significant differentially expressed lncRNAs, while the fold change value is considered in order to chose the genes having a considerable difference between the average of the expression levels in the two groups.

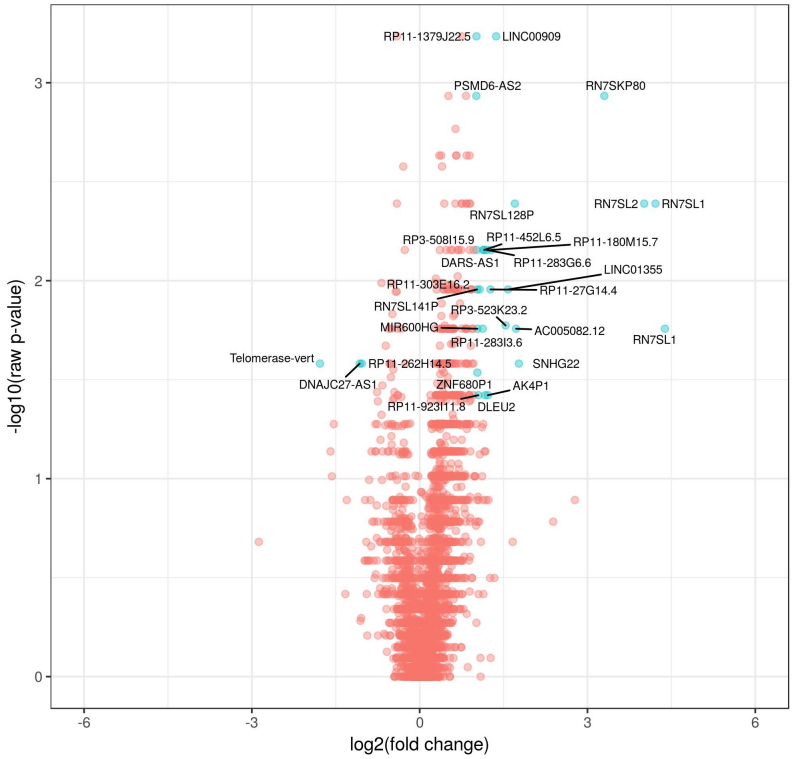


Figure 4.7. Volcano plot

The expression levels of those lncRNAs are then used as input data for the heatmap in figure 4.8. This step is performed in order to check if the unsupervised hierarchical clustering, based on these expression level values, classifies the samples in the same two groups.

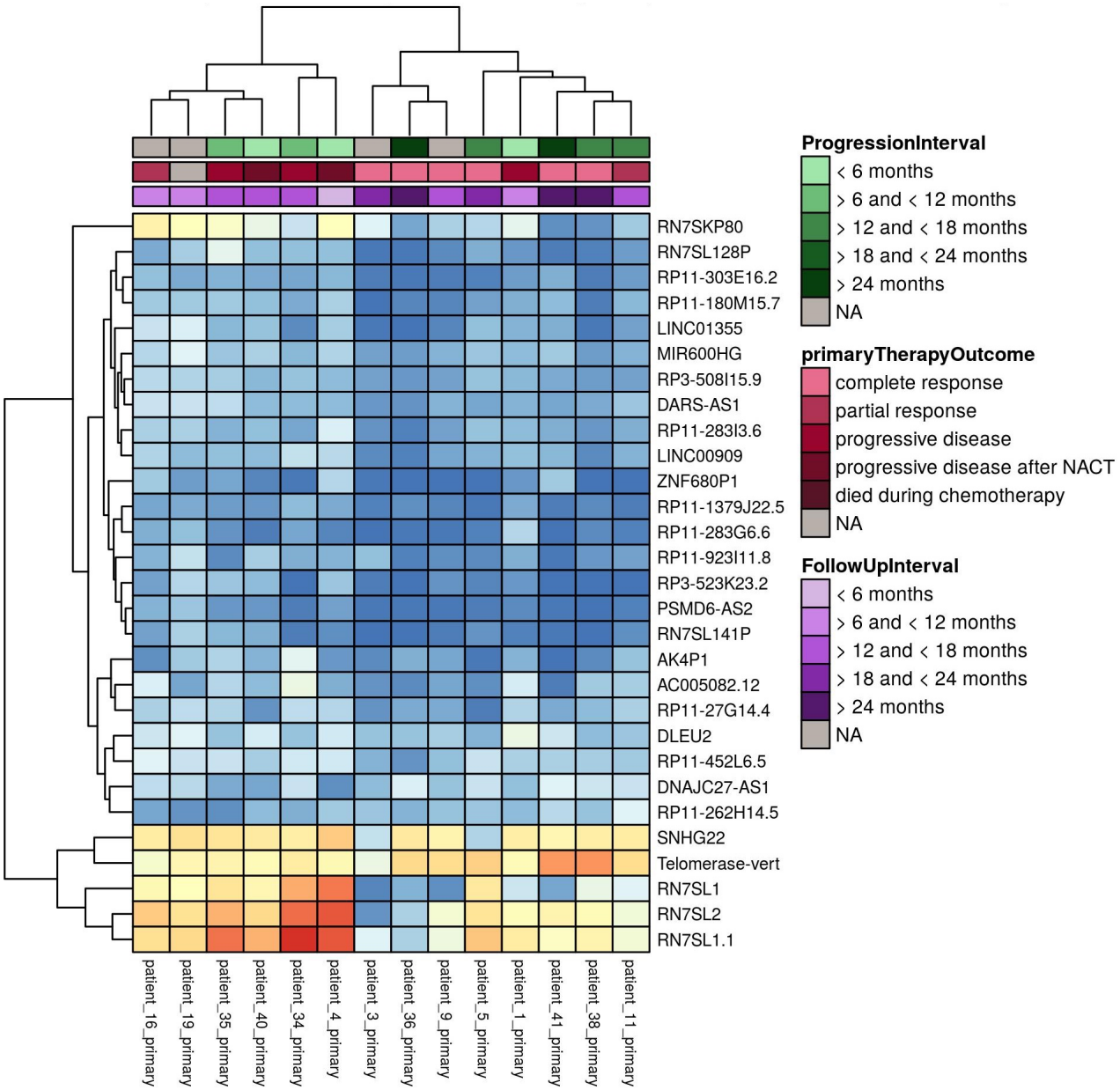


Figure 4.8. RNAs from decomposed RNA-seq samples having raw p-values <0.05 and fold change > 1

From the horizontal dendrogram in the heatmap it is possible to see that the patients are almost correctly divided into short PFI and long PFI, with one exception.

As already said in the previous chapters, the survival of HGSOvCa patients is influenced by the acquisition of platinum resistance. In order to evaluate the relationship between the survival rates and the chemosensitive/chemoresistant patients, a Kaplan-Meier plot is computed.

The Kaplan-Meier estimator (KM, or "product-limit estimator")[21] is a non-parametric statistic used in survival analysis. It is a method adopted when it is necessary to assess the influence of an event upon the survival. The data used in this statistics are lifetime data. Usually this kind of data are censoring data, so data for which the value of the measurement or observation is only partially known. It is known that the value is above (*right censored*) or below (*left censored*) a certain threshold, but the exact quantity cannot be accurately determined. In case of lifetime data, examples of right censored data are the ones in which[22]:

- the subjects withdraw the study before its conclusion;
- the patients are lost to follow-up;
- the study ends before the subjects had the event under investigation;
- the required information are not available for some reasons.

It is known that those subjects were alive before a certain time point, but it is not know for how long they lived thereafter. An advantage of the Kaplan-Meier estimator is that it takes into account all these kinds of incomplete observations. The Kaplan-Meier curve consists of a sequence of steps with different heights and lengths. The step's length depends on the dimension of the temporal interval, while the step' s height depends on the change in the cumulative survival rate. The number of the intervals in the diagram is linked to the number of events that happen during the time sequence under study. Each time there is an occurrence of the event under investigation, there is a new step in the plot. The survival rate corresponding to an interval is the percentage of alive patients in that period of time and it considered constant for all the duration of the interval. This value is computed as [23]:

$$S(t_i) = \frac{n_i - d_i}{n_i} \tag{4.3}$$

where n_i is the number of subjects living at risk (the ones that survived) before the time t_i and d_i is the number of events happened during the interval t_i.

When the event under investigation occurs, it is marked on the diagram by a vertical tick. Because the probability that one patient is alive in a certain interval is related to the probability that the same patient in the previous intervals, it is possible to compute the *cumulative probability* as function of the interval survival. In particular the cumulative survival rate is given by the product of the survival probability in the interval of interest and all the survival probabilities in the preceding intervals. This means that in the first interval the survival probability and the cumulative probability coincide. In formula:

$$S(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i} \tag{4.4}$$

where, again, n_i is the number of subjects living at risk (the ones that survived) before the time t_i and d_i is the number of events happened during the interval t_i.

The assumption of this method are that the time at which events happens are specified, the censoring is not related with the prognosis and the survival probabilities are the same for the all subjects recruited for the study, independently from the time in which they are recruited.

In order to make a comparison between the survival of the two sets it is usually used the *log rank test* [24]. The log rank test (or Mantel-Cox test) is a non-parametric hypothesis test that compares the survival distribution of two groups of samples under study. The null hypothesis states there is no difference between the two curves. The log rank statistic is computed as the sum of the differences between the observed and the expected number of events in a group, computed each time an event occurs, under the null hypothesis. In formula[25]:

$$\chi^2(log_rank) = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \tag{4.5}$$

where E_1 is the total number of expected events in the first group; E_2 s the total number of expected events in the second group; O_1 is the total number of observed events in the first group; O_2 s the total number of observed events in the second group;

The Pearson's chi-square (χ^2) test is used to compute the p-value and assess the statistical significance. This means that the two survival curves are significantly different if the log rank test has a p-value lower than 0.05 (the commonly used significance threshold).

In medical research and bioinformatics, the Kaplan-Meier curves are usually employed in order to estimate the effectiveness of a treatment by looking at the recovery rates of patients after that treatment, compare the survival rates of two groups of patients with different characteristics, understand the correlation between the event under study and the survival probabilities.

In this analysis, samples are divided according to the hierarchical clustering obtained from the dendrogram. The curve shows the relation between the two obtained groups of patients and the survival probabilities. The "long PFI" group contains 8 samples while the "short PFI" group contains 6 samples.

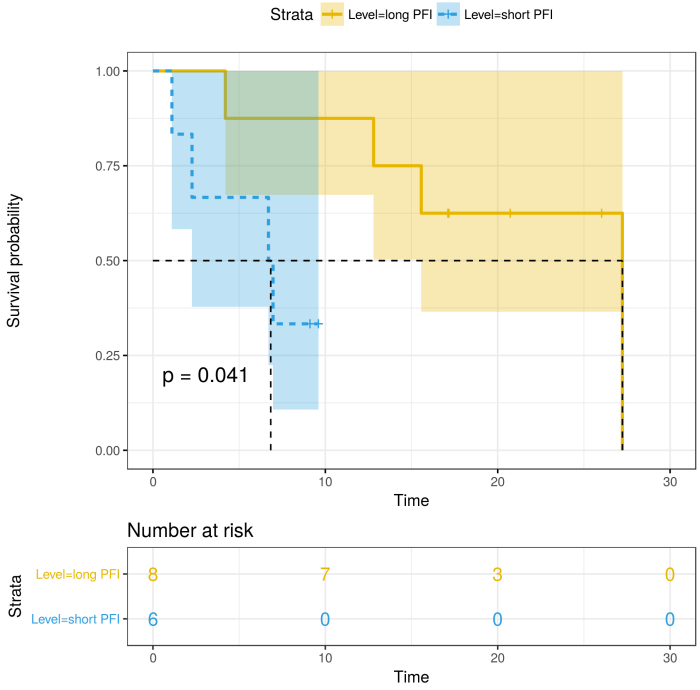


Figure 4.9. Kaplan-Meier plot for all the primary samples

The plot shows a clear difference between the two groups of patients. The ones classified as "long PFI" have a greater probability to survive than the ones classified as "short PFI". The log-rank test performed to compare the survival distributions of the two populations reveals a p-value of 0.041. Because for this kind of statistical test the null hypothesis states that the two groups have identical survival functions, such a lower p-value highlight that null assumption can be rejected.

Because of the high adjusted p-values resulting from the Benjamini-Hochberg correction after the Mann-Whitney-Wilcoxon test, the obtained set of lncRNAs is further analyzed in order to verify if there is a link with the chemotherapy resistance. For this reason, for each one of the differentially expressed lncRNAs is then computed the *correlation coefficient* between the platinum free interval value in months and the expression levels. The correlation is calculated by exploiting the Kendall method and the correspondent p-value for each gene is saved. A correlation coefficient is a value that expresses the degree of relation between two statistical variables. If the values of the two variables for N observations are ranked, it is possible to compute the correlation coefficient between the variables by looking at the order their values have in the rank. The Kendall coefficient[20] (also called "Kendall tau") is a type of rank correlation coefficient. According to it, once ranked the values, if the value of the first variable in the $i-th$ observation has the same rank of the value of the second variable in the $i-th$ observation, the variables are called "concordant". If not, they are called "discordant". The tau coefficient is then computed as:

$$\tau = \frac{(nc) - (nd)}{n(n-1)\frac{1}{2}} \tag{4.6}$$

where nc is the number of concordant pairs, nd is the number of discordant pairs and n is the total number of observations. As it is possible to see from the formula, tau can assume values that go from -1 to 1. If the coefficient value is equal to 1 the variables are *perfectly positively correlated*, while if the coefficient value is equal to -1 the variables are *perfectly negatively correlated*. A non parametric statistical test based on the computation of the τ is called a "tau test".

In the table 4.1 are listed, for the 29 differentially expressed lncRNAs, the obtained raw p-values for the Mann-Whitney-Wilcoxon test, the Kendall tau values and p-values for the Kendall correlation test.

Gene ID	Gene name	MWW p-values	Kendall tau	Kendall p-values
ENSG00000264247.1	LINC00909	0.0005827506	-0.6703296703	0.0004511262
ENSG00000250999.1	RP11-1379J22.5	0.0005827506	-0.4945054945	0.0138377949
ENSG00000202058.1	RN7SKP80	0.0011655012	-0.4945054945	0.0138377949
ENSG00000239653.1	PSMD6-AS2	0.0011655012	-0.5164835165	0.0097530433
ENSG00000276168.1	RN7SL1	0.0040792541	-0.6263736264	0.0012344453
ENSG00000274012.1	RN7SL2	0.0040792541	-0.5824175824	0.0030248117
ENSG00000240869.3	RN7SL128P	0.0040792541	-0.5384615385	0.0067423387
ENSG00000275560.1	RP11-180M15.7	0.006993007	-0.4505494505	0.0263997632
ENSG00000278095.1	RP11-283G6.6	0.006993007	-0.4945054945	0.0138377949
ENSG00000260267.1	RP11-452L6.5	0.006993007	-0.4945054945	0.0138377949
ENSG00000231890.7	DARS-AS1	0.006993007	-0.4945054945	0.0138377949
ENSG00000228274.3	RP3-508I15.9	0.006993007	-0.3406593407	0.1010208746
ENSG00000261326.2	LINC01355	0.0110722611	-0.4285714286	0.0355656671
ENSG00000282221.1	RP11-27G14.4	0.0110722611	-0.4285714286	0.0355656671
ENSG00000261061.1	RP11-303E16.2	0.0110722611	-0.4505494505	0.0263997632
ENSG00000243398.3	RN7SL141P	0.0110722611	-0.4505494505	0.0263997632

Continued on next page

Table 4.1 – *Continued from previous page*

Gene ID	**Gene name**	**MWW p-values**	**Kendall tau**	**Kendall p-values**
ENSG00000261116.1	RP3-523K23.2	0.0168366431	-0.3979585917	0.0522036353
ENSG00000283029.1	RN7SL1	0.0174825175	-0.5604395604	0.0045659834
ENSG00000226816.2	AC005082.12	0.0174825175	-0.5384615385	0.0067423387
ENSG00000261799.1	RP11-283I3.6	0.0174825175	-0.5164835165	0.0097530433
ENSG00000236901.5	MIR600HG	0.0174825175	-0.3626373626	0.0794568992
ENSG00000277925.1	Telomerase-vert	0.0262237762	0.3626373626	0.0794568992
ENSG00000267322.2	SNHG22	0.0262237762	-0.4065934066	0.0471759991
ENSG00000224165.5	DNAJC27-AS1	0.0262237762	0.5384615385	0.0067423387
ENSG00000229422.1	RP11-262H14.5	0.0262237762	0.1868131868	0.3879883441
ENSG00000228014.1	ZNF680P1	0.0291249031	-0.2905436016	0.1523577489
ENSG00000263535.1	AK4P1	0.0378787879	-0.3626373626	0.0794568992
ENSG00000231607.9	DLEU2	0.0378787879	-0.2967032967	0.157163016
ENSG00000278451.1	RP11-923I11.8	0.0378787879	-0.4725274725	0.0192786002

Table 4.1: Most significant lncRNAs

As visual examples of the results obtained, in the following are plotted the correlation plots of the first 3 differentially expressed lncRNAs having lower raw p-values and higher fold change values. The correlation plots show the correlation between expression levels (on the y axis) and the PFI (on the x axis). Ideally, if there is a correlation, the expression levels should decrease/increase according to the platinum-free interval. In the plot are also reported the Kendall tau, the Kendall p-value and the interval of confidence (in grey). Samples are ordered according to the PFI and distinguished by colour according to the group they belong to. The blue ones are the patients in the short PFI group, while the dark ones are the patients in the long PFI group. The green line is computed as the mean of all the expression levels in the samples and it is used to distinguish between high and low values of the gene expression.

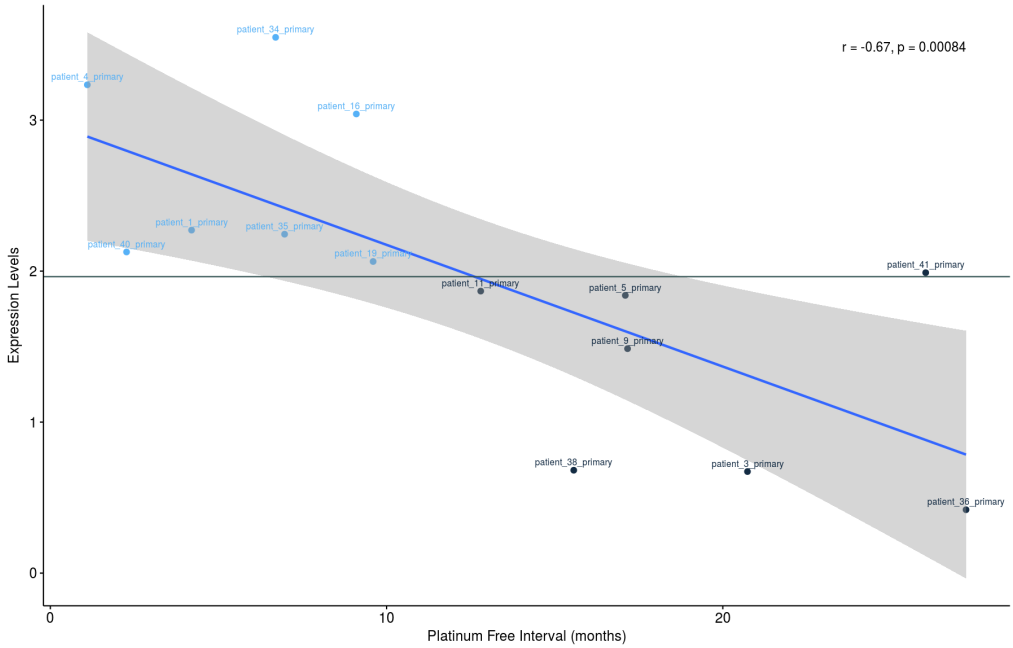


Figure 4.10. Correlation plot for the LINC00909 lncRNA

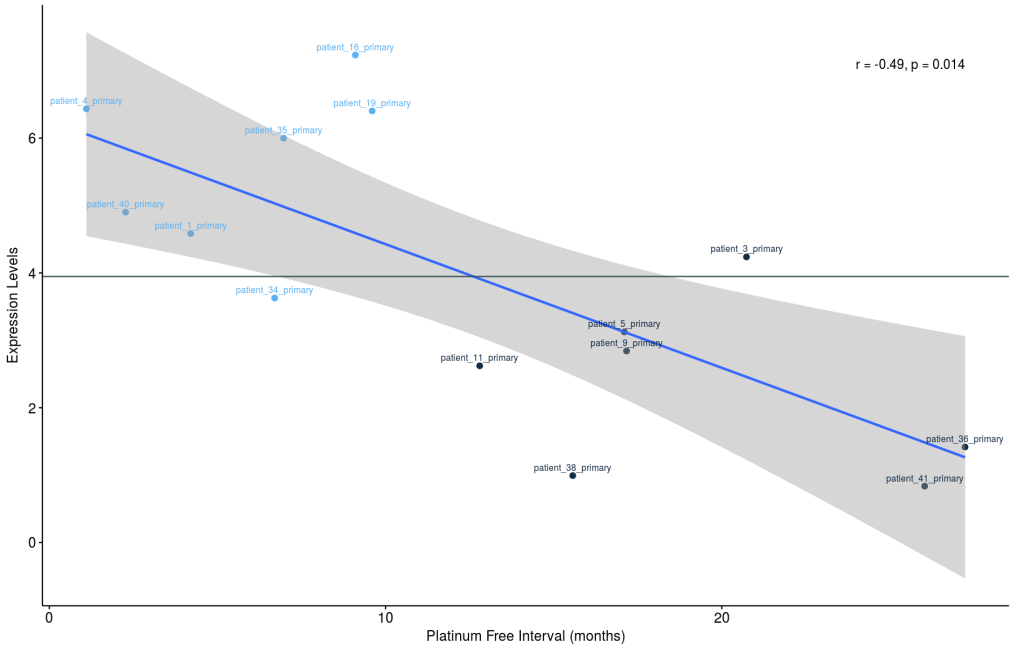


Figure 4.11. Correlation plot for the RN7SKP80 lncRNA

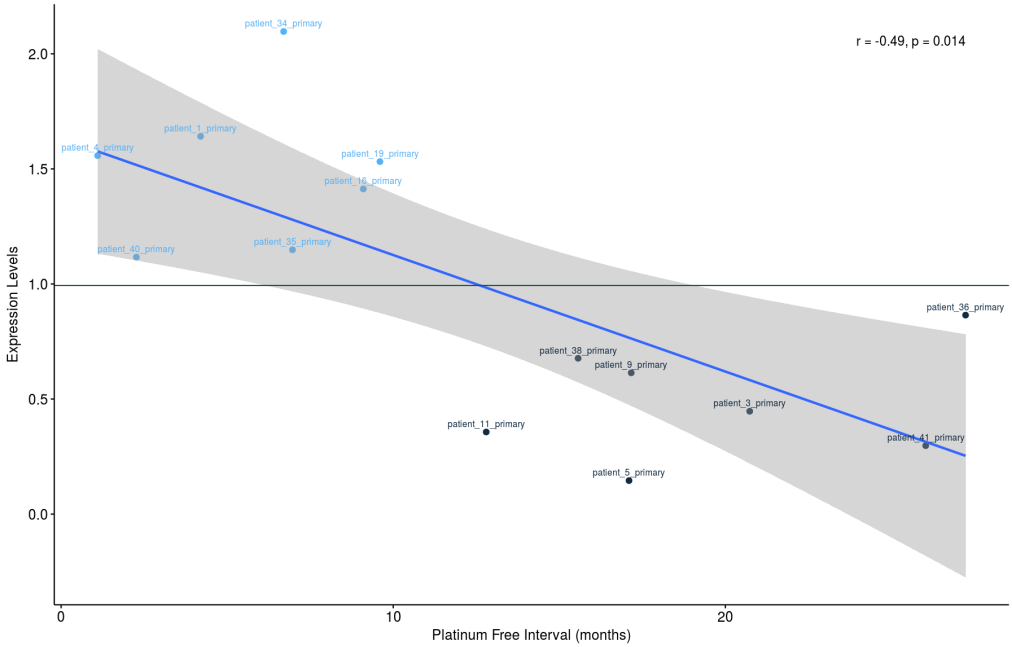


Figure 4.12. Correlation plot for the RP11-1379J22.5 lncRNA

4.4.3 DEG extraction through feature selection - Random Forest

In order to exploit an alternative way to extract differentially expressed genes that are chemotherapy related, two feature selection techniques based on the Random Forest algorithm are used. The first one, is a customized approach that involves several runs of the machine learning algorithm and the employment of the leave one out cross validation for assessing the model's accuracy. The second one is, instead, an already available algorithm specifically created for feature selection. This section describes the first approach, while the following section illustrates the second method.

Random forest[26] (RF) is a supervised machine learning algorithm for classification and regression. Its main goal is to predict the value of a target variable. The term "classification" is used when the predicted value is a label, a categorical attribute; while the term "regression" is used when the predicted value is a real number. Machine learning algorithms do not follow manually programmed rules or code instructions to catalogue a certain input, but they "build" a classification procedure starting from a training set of data for which the correct output is known. The input data are composed of a set of observations, defined by different values of a group of features. The output is a label or a real number. In order to build a machine learning model that is able to autonomously take decisions, it is necessary a *learning phase*. In this phase, the initial set of labelled data is divided in a *training* and a *test set*. By their names, it is intuitively understandable that the first set is used for building the model, while the second is used for testing it. After the training and the test phases, the ML algorithm is able to classify unlabelled data.

The Random Forest algorithm is called "ensemble method" because it exploits the use of a set of different *decision trees* to improve the predictive performances. A decision tree is a predictive model with a tree structure in which each node represents a decision point, each edge a possible decision outcome and each leaf an output variable used to take the final decision. Decision trees are known to be fast to construct and easy to interpret, but they may lack in accuracy and easily produce overfitting[27]. Because of that, the random forest algorithm trains several decision trees to mitigate those side effects. Of course, the adoption of this technique may lead to a significant reduction in speed. Random forest uses the *bagging* (or "bootstrap aggregating") strategy [28] to create different input dataset for

the trees. These new datasets (called "bootstraps") are created by sampling uniformly and with replacement the original input set. Each Random Forest model is then fitted with a different bootstrap and the final outcome is computed by averaging the single outputs in case of regression or by adopting a voting technique in case of classification.

The Random fForest algorithm can be described in pseudo-code as follow[27]:

---

for b = 1 to B: **do**

(a) Draw a bootstrap sample $\mathbf{Z}^{\star}$ of size N from the training data.

(b) Grow a random-forest tree T_b to the bootstrapped data, by re- cursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

i. Select m variables at random from the p variables.

ii. Pick the best variable/split-point among the m.

iii. Split the node into two daughter nodes.

end for

Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

Regression: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{rf}^B(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B$.

The existing implementations of this algorithm have three major parameters that can be tuned by the user, according to the kind of data at her disposal:

- *mtry*, the number of variable that have to be taken into consideration at each split. The value that is usually used in this case is the rounded square root of the total number of variables in the dataset;

- *ntree*, the number of trees in the forest;
- *nodesize*, the minimum size of the terminal nodes. The default value is 1.

In the machine learning methods that use bootstrap samples, it is possible to measure the prediction error of the fitted model by using the *out-of-bag error* (OOB error). Because in the bagging strategy some samples are left out, the OOB error for the random forest algorithm can be computed as the mean value of the prediction errors evaluated by using the left out samples on the trees that did not use those samples in the training phase [29].

The OOB estimation is also used to compute variable importance. The variable importance is a measure of how each variable contributes to the predictions made by the model. When a research question involves high dimensional data (data whose number of observations is much smaller than the number of features), it is possible that not all the features initially involved in the search are informative. In order to define which are the most informative features, it is possible to exploit a feature selection (FS) technique. The Random Forest algorithm is widely used for classification and regression purposes, but actually its structure gives the possibility to use it also as a feature selection method. In particular, the algorithm can be seen as a *wrapper* feature selection method, because it automatically computes variable importance while training the model. In the training phase of the Random Forest, when several decision trees are built, classification rules are constructed on the basis of the most informative features. A feature is judged as informative if, according to its value, the algorithm is able to correctly classify an instance. During training, for each feature is computed a measures of variable importance and that value is subsequently employed for the identification of the most significative features that are able to discriminate the input variables in the different classes. There are two types of scores for the computation of the variable importance in this algorithm:

- the mean decrease accuracy (MDA);
- the mean decrease impurity (MDI).

The MDA measure [26] relies on the idea that the permutation of a non-informative feature values does not decrease the model accuracy, while the permutation of important feature values affects it significantly. Model accuracy is computed by using the OOB error. The

values of each features are randomly permuted in the OOB samples and the difference between the correspondent OOB error and the one obtained without permutation is computed. The MDA index measured by averaging all the quantities computed all over the trees.

The MDI measure is based on the fact that each node of a tree in the forest evaluates a condition on a single feature and splits, according to the results, the ensemble of data in two different sets. The choice of the variable to be used for the splitting is based on the local extimation of the split purity. The purer are the subsets obtained after the split, the more informative is the used feature. The measure of the impurity reduction that each feature produces can be computed for each tree and than averaged over all the forest. When the Gini index is used as measure of importance, the metric is also called "Mean Decrease Gini" (MDG) or "Gini importance" [31].

Beyond the possibility of using Random Forest as a feature selection technique, this method is chosen, among all the machine learning algorithms, also because it is known for being robust[47] and having good performances [48] [49]. In this analysis, the features are the long non-coding RNAs at disposal, while the input instances are the available samples. The primary samples at disposal are again 14, always divided in two groups according to the platinum-free interval. In order to obtain more stable results with a such small number of samples, the feature selection algorithm is repeated several times. Because the Random Forest uses random variables, in fact, it does not return the same results with the same dataset all the times. Variability in the results is also given by the high number of features involved with respect to the number of samples. For this reason, similarly as before, the number of initial long non-coding RNAs is reduced by discarding the genes having expression levels lower than 1 in all the samples. The used implementation of the Random Forest algorithm leaves to the user the setting of the mtry and ntree paramters. In this case, it is chosen to use standard values for the parameters setting and to define mtry as the rounded square number of remaining lncRNAs and ntree as 501.

Because the procedure is repeated 10 times, 10 different seeds are computed and saved for reproducibility. For each one of the 10 seeds, the Random Forest algorithm is run for training a new model on all the 14 samples and the resulting features rankings are saved. The metric employed for the assessment of the variable importance is the *mean decrease*

accuracy. The resulting rankings of the features for each run and their corresponding MDA values are reported in the plots below. In each plot, are shown the 30 long non-coding RNAs having the highest values of mean decrease accuracy.

Once performed several runs of the algorithm, the final set of most informative long non-coding RNAs is retrieved by applying the major voting technique. The 30 most informative features of each run are merged together and for each of them it is computed a value indicating in how many runs that feature was judged as informative. This value is called "votes". Features are then ranked in descending order according to this value and from this rank are extracted the lncRNAs having more than 5 votes. As final result, it is obtained that 37 long non-coding RNAs are judged as informative in more than half of the cases. Those lncRNAs and their corresponding votes values are reported in the table 4.2. Those genes are then visually analysed through a heatmap shown in the figure (figure 4.15).

Gene ID	**Gene name**	**votes**
ENSG00000239653.1	PSMD6-AS2	10
ENSG00000250999.1	RP11-1379J22.5	10
ENSG00000265666.1	RARA-AS1	10
ENSG00000241187.1	CTC-209L16.1	9
ENSG00000259673.5	IQCH-AS1	9
ENSG00000265688.1	MAFG-AS1	9
ENSG00000277423.1	RP11-173P15.9	9
ENSG00000278811.4	LINC00624	9
ENSG00000219023.1	RP3-340B19.2	8

Continued on next page

Table 4.2 – *Continued from previous page*

Gene ID	**Gene name**	**votes**
ENSG00000239899.3	RN7SL674P	8
ENSG00000264247.1	LINC00909	8
ENSG00000267834.1	RP11-167N5.5	8
ENSG00000275142.1	RP5-999L4.2	8
ENSG00000282221.1	RP11-27G14.4	8
ENSG00000202058.1	RN7SKP80	7
ENSG00000228274.3	RP3-508I15.9	7
ENSG00000228613.1	AC144450.1	7
ENSG00000229473.2	RGS17P1	7
ENSG00000242170.3	RN7SL329P	7
ENSG00000249159.6	RP11-480D4.2	7
ENSG00000255067.1	RP11-47J17.1	7
ENSG00000260259.1	RP11-368I7.4	7
ENSG00000260597.1	AC012531.25	7
ENSG00000268080.2	RP11-388K12.3	7
ENSG00000268987.1	CTC-435M10.10	7

Continued on next page

Table 4.2 – *Continued from previous page*

Gene ID	**Gene name**	**votes**
ENSG00000278075.1	RP11-248M19.1	7
ENSG00000200488.1	RN7SKP203	6
ENSG00000228232.1	GAPDHP1	6
ENSG00000228280.1	RP11-367B6.2	6
ENSG00000231170.5	AC002451.3	6
ENSG00000234115.2	RP11-288G3.4	6
ENSG00000239726.3	RN7SL688P	6
ENSG00000245468.3	RP11-367J11.3	6
ENSG00000255468.6	RP11-867G23.8	6
ENSG00000262265.1	RP5-867C24.4	6
ENSG00000271984.1	RP3-337O18.9	6
ENSG00000277925.1	Telomerase-vert	6

Table 4.2: Most significant lncRNAs obtained with Random Forest feature selection

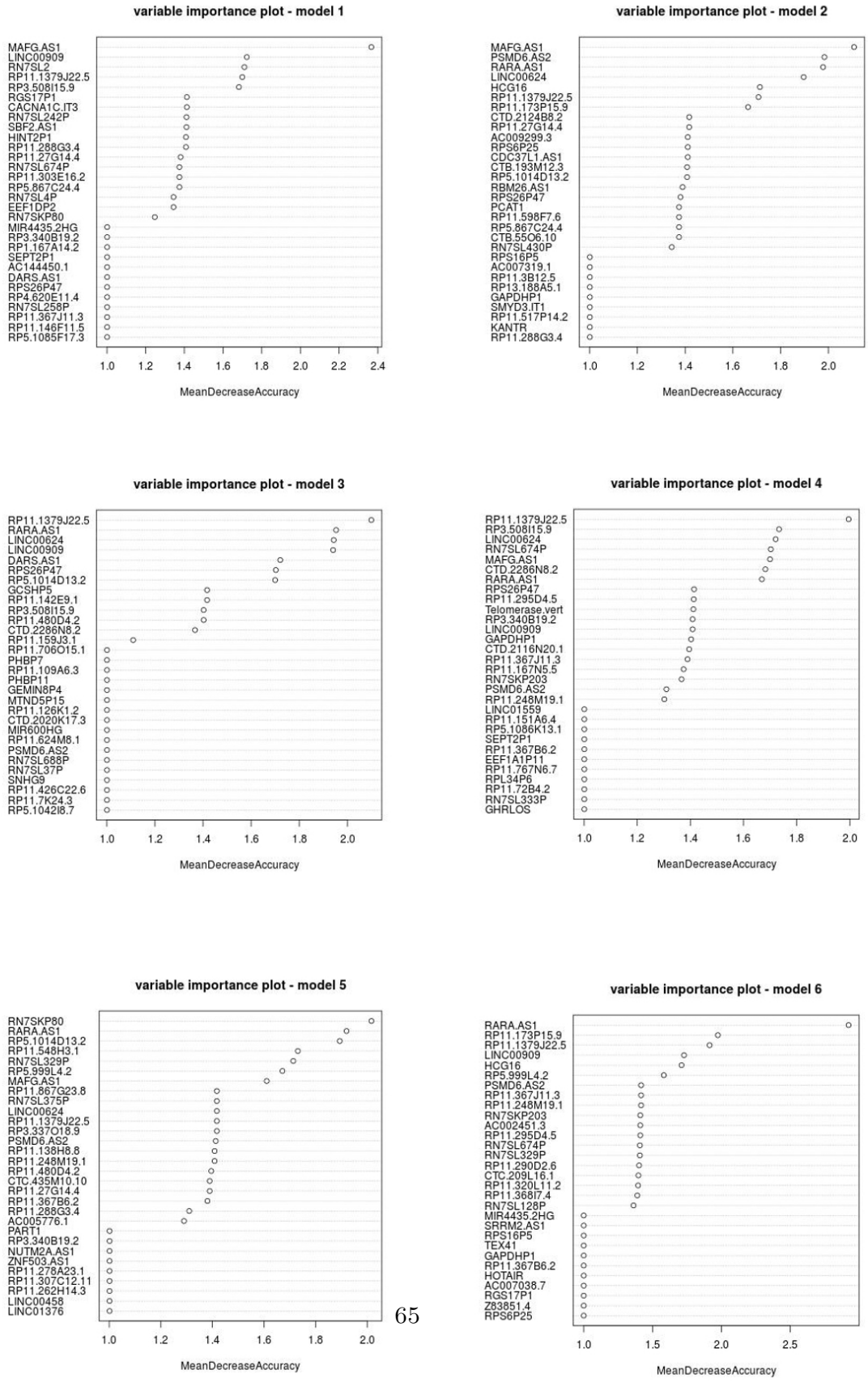


Figure 4.13. Variable importance plots of the first 6 models

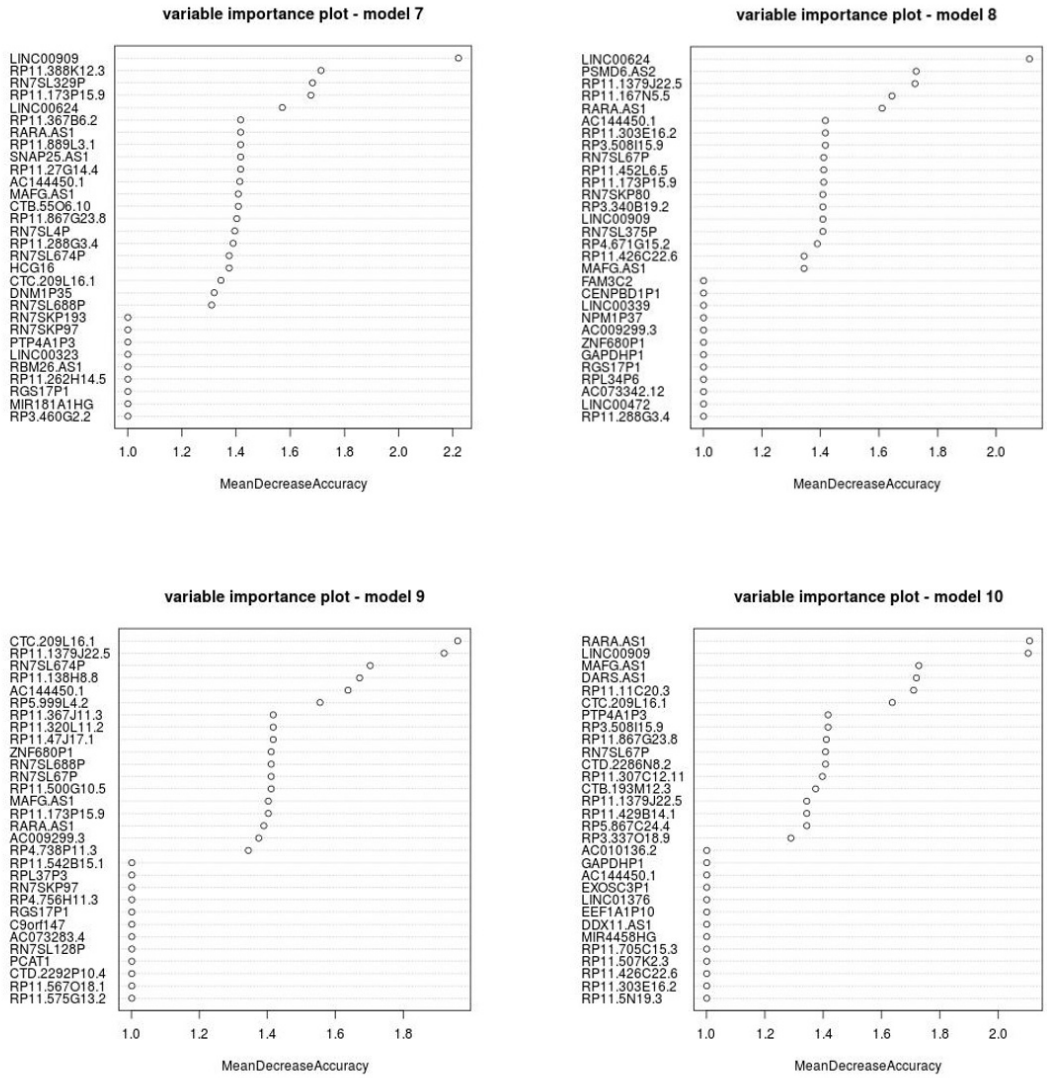


Figure 4.14. Variable importance plots of the least 4 models

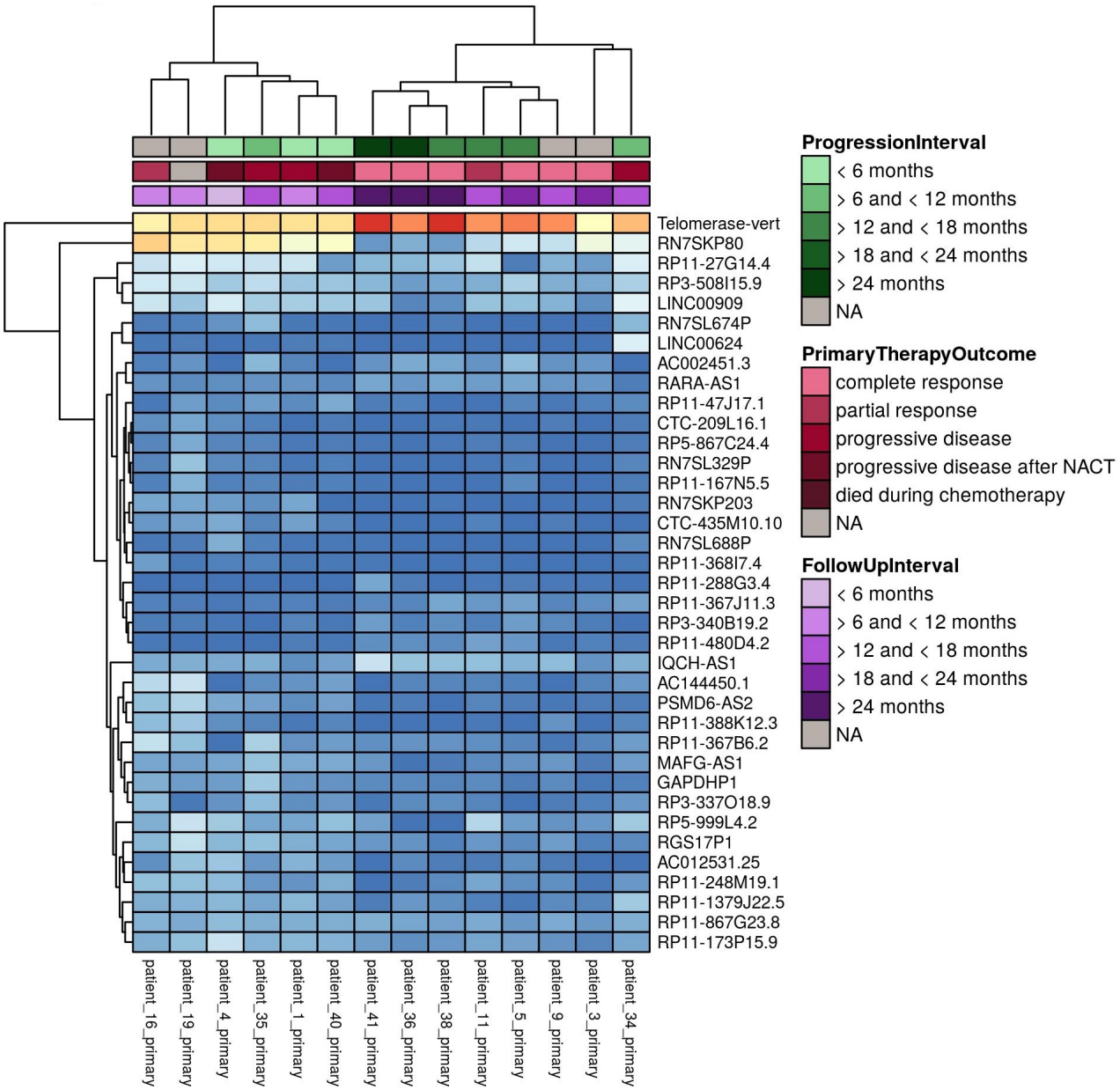


Figure 4.15. Significative lncRNAs obtained through Random Forest feature selection

The assessment of the model performances is achieved by using the *leave one out cross-validation*. The cross-alidation (CV) is a validation method that evaluates the degree of generalization of a statistical analysis when this is applied to a new independent set of data. In machine learning algorithms, CV methods are widely used for assessing the model performances when there are too few data and it is not possible to partition them in a test and a training sets. Testing a learning model on the same data used for training it, in fact, leads to over optimistic results [34]. the cross-validation adopts an iterative behaviour in which the initial dataset is first partitioned in a training and a test sets and then, the chosen learning algorithm is runned on them. This means that at each iteration a new model is built and validated. The outcome of each validation is stored and used for the computation of the final performance. The assumptions of cross-validation are the independency and the identical distribution of the data (i.i.d.). Without having i.i.d. data, it is not possible to create independent test sets so it is impossible to correctly evaluate the model performance. A particular kind of cross-validation is called "leave one out cross-validation" (or LOOCV), in which, at each step, the training set is created by excluding one observation from the original set. The correspondent test set, instead, is composed by the observation left out from the original set. In this way, at each step, mutually exclusive subsets of data are created. The number of iteration performed by this algorithm is equal to the number of original data. In this analysis the algorithm is repeated 14 times and each time the outcome of the prediction made on the test set is saved. In the end, all the saved predictions are used to compute the performance metrics. In this analysis, in order to evaluate the Random Forest performances, are computed the ROC curve, the area under the curve and the confusion matrix.

The *confusion matrix* (also called "contingency table") is a squared matrix that have on one dimension the predicted values and on the other the real values. Each dimension is divided in a set of classes, which correspond to the possible categorical outcomes. In the case study proposed in this work of thesis there will be only two classes, so the confusion matrix will be a 2x2 matrix, like the one in the figure 4.16:

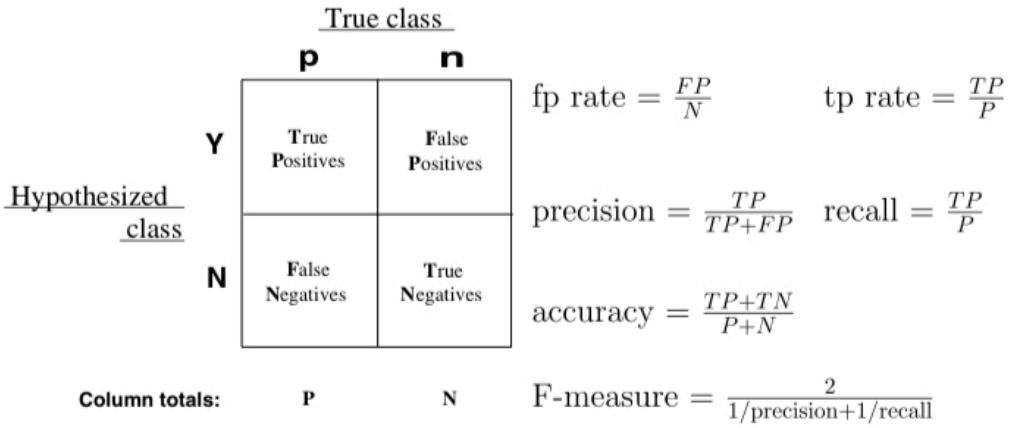


Figure 4.16. Confusion matrix [36]

Generally, the two classes are called "positive" and "negative". The terms contained in the matrix are the following ones:

- *true positives*, which express the number of cases in which the actual class value was positive and the case was correctly classified;
- *true negatives*, which express the number of cases in which the actual class value was negative and the case was correctly classified;
- *false positives*, which express the number of cases in which the actual class value was negative but it was classified as positive;
- *false negatives*, which express the number of cases in which the actual class value was positive but it was classified as negative;

From these terms, it is possible to compute some performance measurements. *Accuracy* is obtained by dividing the sum of the true positive values and the true negative ones by the sum of all the real positive values and all the real negative ones. *Specificity* is computed by dividing all the true negatives by the sum of all the true negatives and all the false positives. *Sensitivity,* also called *recall*, is computed as the ratio between the true positives and all the real positives.

The ROC curve (acronym for "Receiver Operating Characteristic curve") is a two-dimensional graphical representation having on the x axis the false positive rate (also known as "probability of false alarm", computed as 1-specificity) and on the y axis the true positive rate (also known as "probability of detection", the sensitivity). It is a cost-benefits graph for a binary classifier, in which costs are expressed in terms of false positives and benefits are expressed in terms of true positives. All the curves in the graph originate in the point (0,0) and terminate in the point (1,1). The diagonal line y=x is called the "line of no-discrimination", the ones that define a classification based on random guessing. If the curve is below that line, it means that the classification is worse than random; if the curve is above, it means that the classification is better than random.

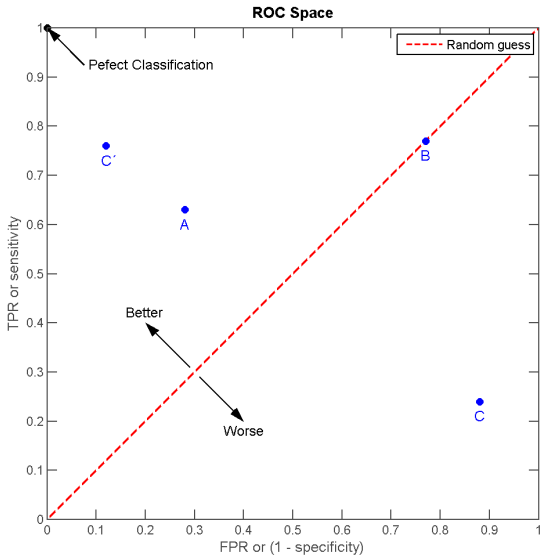


Figure 4.17. ROC space

The area under the ROC curve (AUC or better AUROC)[37] is a real number between 0 and 1. It expresses the portion of the graph area that is under the ROC curve, as it is guessable from the name. Its value is equal to the probability that "the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance" [36], assuming that the in the rank the positive instances are above the negative ones.

For the Random Forest algorithm employed in this thesis, the performance metrics

computed after the leave one out cross-validation step are the following ones:

Confusion matrix:

	X1	**X2**
X1	4	2
X2	3	5

Accuracy: 0.6429

Specificity: 0.7143

Sensitivity: 0.5714

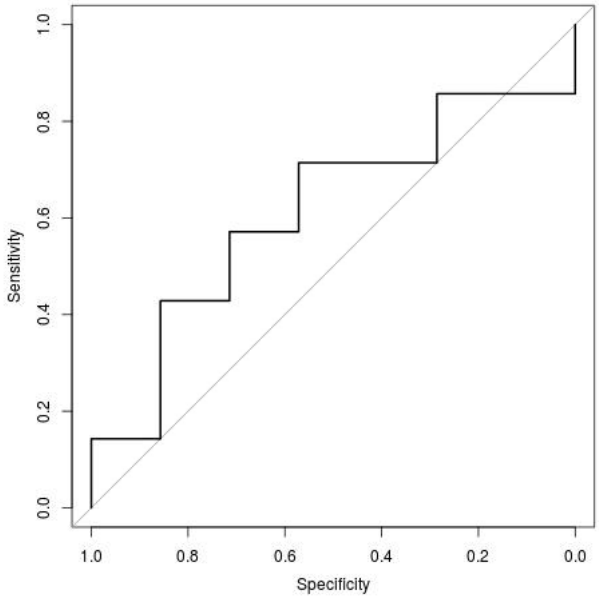


Figure 4.18. ROC curve

The area under the ROC curve is equal to 0,6122.

4.4.4 DEG extraction through feature selection - Boruta

The last approach used for the extraction of differentially expressed genes is based on the Boruta algorithm. This wrapper feature selection method is relatively new and it is chosen for having something comparable to the previous feature selection step. Also Boruta is, in fact, based on Random Forest and it also uses the Mean Decrease Accuracy index for ranking the most informative features. In addition to that, it is known from literature [41] that this feature selection method is a powerful and stable when applied on omics data. Feature selection can be used in order to find the smallest subset of features that ensure the best classification (minimal optimal problem) or to find the complete subset of features that are relevant for the classification (all-relevant problem)[38]. The Boruta was designed to overcome the second type of problems. The importance measure of an attribute on which this algorithm is based is the *Z score*. This quantity is computed as the ratio between the average accuracy loss and its standard deviation. For each tree in the forest, in fact, the loss in classification accuracy obtained by shuffling the feature values between objects of a tree is computed.

Boruta uses the significance of the Z score values to divide the features in "important" and "unimportant". In particular, the steps needed for achieving the attributes selection are the following ones[39]:

1. The initial dataset is extended with the addition of the copies of the attributes.Those copies are called "shadow attributes".

2. The values of the additional attributes are then shuffled to remove their correlation with the response and the Random Forest algorithm is runned several times on this new set.

3. For each RF run, the Z score values for both the original and the shadow attributes are computed. The maximum Z scores among all the shadow attributes is evaluated and this quantity is called MZSA. A hit is marked for all original features having Z score higher than the MZSA value.

4. A two-sided test is computed on all the original attributes. The null hypothesis is that there is the equality between the variable's importance and the MZSA value.

For each feature is computed the total number of hits in all the RF runs.

5. The attributes having a number of hits significantly higher than the expected value are marked as “important” (or “accepted”) in that run.

6. The attributes having a number of hits value significantly lower than the expected value are marked as “unimportant” (or “rejected”) and are removed from the extended dataset together with their correspondent shadow attribute.

The procedure is run a predefined number of times and the the stopping conditions are the following:

- the max number of iterations is reached;
- all the attributes are rejected before the max number of iterations is reached;
- all the attributes are accepted before the max number of iterations is reached.

The Boruta method was adopted with promising results in several bioinformatics studies [40][41] based on omics data and in particular for gene selection problems. It showed good computational time, easily interpretable parameters and good accuracy in variable selection.

As in the previous case, also this time are performed several runs of the algorithm in order to obtain more stable results. Being based on the Random Forest algorithm, in fact, also this methods internally sets some random variables and this lead to different results for different runs on the same dataset. As in the previous approach, the process is repeated 10 times on all the samples and the final results are computed by performing major voting on the different outcomes. Each run returns as output the gene identifiers of the lncRNAs that are judged as more informative. The only parameter that can be set by the user in the Boruta algorithm is the *maxRun* value (the maximun runs of the Random Forest algorithm), that in this analysis is set equal to 501.

The major voting is performed on the lncRNAs obtained from the 10 runs. The resulting long non-coding RNAs are merged together and, also in this case, for each feature (gene), it is computed the number of times it was judged as informative by the algorithm. This value is again called “votes” and the resulting long non-coding RNAs are ranked according

to it. The number of total features extracted from the results of the different runs are 56, but the number of genes that are judged as informative in more than 1 run is equal to 38. Those 38 genes and their "votes" values are reported in the table below.

Gene ID	**Gene name**	**votes**
ENSG00000250999.1	RP11-1379J22.5	10
ENSG00000264247.1	LINC00909	10
ENSG00000265666.1	RARA-AS1	10
ENSG00000265688.1	MAFG-AS1	10
ENSG00000278811.4	LINC00624	10
ENSG00000277423.1	RP11-173P15.9	8
ENSG00000202058.1	RN7SKP80	7
ENSG00000242170.3	RN7SL329P	6
ENSG00000239726.3	RN7SL688P	5
ENSG00000239899.3	RN7SL674P	5
ENSG00000234115.2	RP11-288G3.4	4
ENSG00000239653.1	PSMD6-AS2	4
ENSG00000241187.1	CTC-209L16.1	4
ENSG00000243854.3	RN7SL67P	4

Continued on next page

Table 4.3 – *Continued from previous page*

Gene ID	**Gene name**	**votes**
ENSG00000253438.2	6 PCAT1	4
ENSG00000255067.1	RP11-47J17.1	4
ENSG00000228274.3	RP3-508I15.9	3
ENSG00000228613.1	AC144450.1	3
ENSG00000244349.1	HCG16	3
ENSG00000261116.1	RP3-523K23.2	3
ENSG00000268987.1	CTC-435M10.10	3
ENSG00000279738.1	RP5-1014D13.2	3
ENSG00000219023.1	RP3-340B19.2	2
ENSG00000228280.1	RP11-367B6.2	2
ENSG00000229473.2	RGS17P1	2
ENSG00000234354.3	RPS26P47	2
ENSG00000244389.3	RN7SL242P	2
ENSG00000259673.5	IQCH-AS1	2
ENSG00000261662.1	RP5-1042I8.7	2
ENSG00000262380.1	CTB-193M12.3	2

Continued on next page

Table 4.3 – *Continued from previous page*

Gene ID	**Gene name**	**votes**
ENSG00000265745.2	RN7SL375P	2
ENSG00000267655.1	CTD-2286N8.2	2
ENSG00000267834.1	RP11-167N5.5	2
ENSG00000270558.1	CTD-2124B8.2	2
ENSG00000271984.1	RP3-337O18.9	2
ENSG00000275142.1	RP5-999L4.2	2
ENSG00000276529.1	AP001505.10	2
ENSG00000277925.1	Telomerase-vert	2

Table 4.3: Most significant lncRNAs obtained with the Boruta feature selection algorithm

The same set of genes is also used, together with the primary samples information and their correspondent clinical data, as input value for the heatmap shown in figure 4.19.

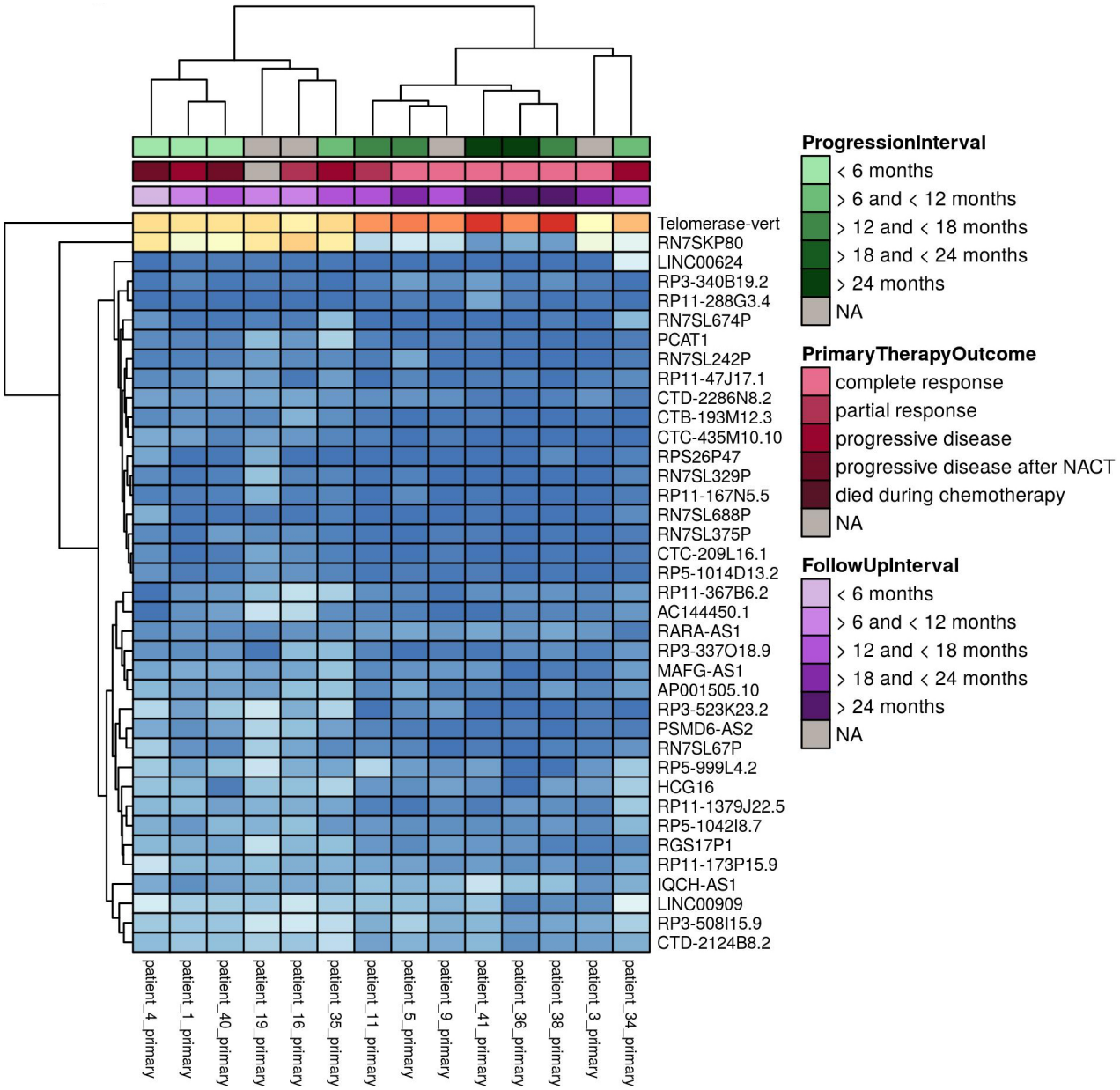


Figure 4.19. Significative lncRNAs obtained through the Boruta feature selection method

By examining only the results obtained with the two supervised machine learning methods, it is obtained that 25 long non-coding RNAs are identified as chemotherapy related by both the approaches. If are taken into account also the results obtained with the unsupervised clustering and the statistical analysis, instead, the number of common results drops

to 6. The long non-coding recognized as potentially related to chemotherapy resistance by all the approaches exploited in this analysis are listed below:

- ENSG00000239653.1, PSMD6-AS2
- ENSG00000250999.1, RP11-1379J22.5
- ENSG00000264247.1, LINC00909
- ENSG00000202058.1, RN7SKP80
- ENSG00000228274.3, RP3-508I15.9
- ENSG00000277925.1, Telomerase-vert

Chapter 5

Results and Comments

This work of thesis has the intention to present the possibility and the usefulness of using different analytical approaches for the identification of chemotherapy-related long non-coding RNAs in patients affected by high grade serous ovarian cancer. In particular, the approach adopted consists in the employment of an unsupervised technique supported by statistical processing and two different supervised feature selection methods based on the Random Forest algorithm. The results obtained with different methodologies are in the end integrated to increase the confidence level.

5.0.1 Evaluation of the RNA seq analysis

The explorative step of the analysis is conducted just by dividing samples in short platinum-free interval and long platinum-free interval, by computing the fold change as absolute difference between the average values of expression levels in the two groups for each long non-coding RNA and by using the genes having fold change values > 1.2 as input data for the hierarchical clustering. The results obtained show that samples belonging to the same patient and the same tissues tend to cluster together. This reveals two things: that a patients with a lot of samples can potentially drive more the clusterization with respect to the patients having one or few samples and that the sample's tissue type influences the expression levels.

The use of heatmaps combined with hierarchical clustering also reveals the difficulty

of analyzing in a visual way a huge amount of data. This raises the necessity to reduce the dimensionality of data. Having a great number of long non-coding RNAs in a single heatmap leads to interpretational problems and the presence of non informative data may not reveal the existence of patterns in the dataset. This means that before using visualization methods, it is necessary to preprocess the initial set of data in order to detect the data that are most informative for the research question.

5.0.2 Evaluation of the decomposed RNA seq analysis

In the core part on the analysis, it is decided to use decomposed data in order to decrease the variance and to focus only on the information related to the tumoral activity present in the samples. As already stated in the Materials and Methods chapter, it is also decided to average the expression levels of the samples coming from the same patients. In this way, even if the analysis is not performed on the original data, the overall behaviour remains unaltered.

The results obtained through the statistical processing and the unsupervised clustering technique reveal the necessity of further investigations. In the statistical analysis, performed with the Mann-Wilcoxon-Withney test and the Benjamini-Hochber correction, in fact, are obtained high p-values. The high difference between the adjusted p-values and the significance threshold lead to the impossibility to discard the null hypothesis. It is consequently not possible to state that the detected long non-coding RNAs are effectively chemotherapy related or not. The explanation for such high adjusted p-values may be linked to the small number of samples employed in the analysis. In addition to that, there is the problem that the Benjamini-Hochberg correction does not take into account the possible correlations existing among the variables. In this case, variables are long non-coding RNAs and these can be highly correlated because they can interact with each other and participate to the same processes.

Because it is not possible to retrieve statistically significant results with this approach, it is decided to further analyse the long non-coding RNAs with raw p-values smaller than 0.05 and fold change >1 and check if they are correlated with the platinum-free interval by using the Kendall tau correlation. By examining the Kendall p-values computed for those

genes and reported in table 4.1, it is possible to see that the majority of the long non-coding RNAs (21 ou of 29) present Kendall p-values smaller than 0.05. As shown in the correlation plots in section 4.4.2 of the previous chapter, there is effectively a correlation between the expression levels in the samples and the number of months in which the patient is free from the disease after the chemotherapy treatment.

The selected genes are also used as input values for the heatmap in figure 4.8. This plot shows a good separation between the samples belonging to the "long PFI" group and the "short PFI" group. The upper dendrogram of the heatmap, in fact, clearly divides the samples in two groups and, on the basis of the long non-coding RNAs selected, the hierarchical clustering is able to correctly group 13 samples out of 14. Of course, because of the small number of samples it is not possible to be sure that the same division would be kept with a larger dataset.

By considering the divisions in the two groups performed by the hierarchical clustering algorithm, it is also performed a survival analysis by reproducing a Kaplan-Meier plot. The graph in figure 4.9 shows a clear distinction between the survival curves of the patients having low values of PFI and patients having high values of PFI. The distance between the two curves highlight the difference existing among the two groups: patients classified as "long PFI" have greater probability to survive for a longer time with respect to patients classified as "short PFI". The p-value resulting from the log-rank test has a value of 0.041 and it clearly rejects the null hypothesis for which there is no difference between the two curves.

Talking about the analysis performed in the supervised approach with the Random Forest algorithm, from the heatmap in figure 4.15 it is possible to see that the long non-coding RNAs extracted with the feature selection process are actually able to correctly divide the samples in chemosensitive and chemoresistant. The exiguous number of samples, anyway, is a strong limit for the Random Forest approach. Machine learning algorithms need a relatively big number of data during the training phase in order to produce a good classifier. In the ROC curve in figure 4.18, it is shown that the general performances of the algorithm are better than the random guessing (AUC = 0.5), but the results are still far from the best case. Having such a small dataset on which to perform the training, the algorithm is not able to learn all the main distinctive characteristics of the data and,

consequently, it is not able to well classify them.

With the last approach, the one based on the Boruta algorithm, are obtained results similar to the ones obtained with the Random Forest approach. By looking at the heatmap in figure 4.19, in fact, it is possible to verify that the patients are divided in the same way in both the cases. In addition to that, as already mentioned in the previous chapter, the results of the two methods partially overlaps and they identify 25 common long non-coding RNAs.

5.0.3 Limitations of the study and comments

The major limitation of this work of thesis is constituted by the number of samples at disposal. With the available data it is possible to show that different techniques for the identification of candidate chemotherapy-related long non-coding RNAs are possible, but further analysis with a larger set of data are still needed. Given the complexity of information contained in the dataset, even with a larger number of samples, the combination of different techniques may be a useful approach for reaching a bigger confidence in the results. A biological assessment in wet-lab is then needed to definitively validate the identified chemotherapy-related long non-coding RNAs.

Moreover, it is possible to understand from the analysis that it would be desirable to have more samples equally distributed with respect to the platinum-free interval variable.

Bibliography

[1] C. Hoppenot, M. A. Eckert, S. M. Tienda, E. Lengye *Who are the long-term survivors of high grade serous ovarian cancer?*, Gynecologic oncoogy, 2017.

[2] D. Hanahan and R. A. Weinberg, *The hallmarks of cancer*, Cell, 2000.

[3] D. Hanahan and R. A. Weinberg, *Hallmarks of cancer: the next generation*, Cell, 2011.

[4] V.W.Chen et al., *Pathology and classification of ovarian cancer tumors*, Cancer, 2003.

[5] E. S. Hosseini et al.,*Dysregulated expression of long noncoding RNAs in gynecologic cancers*, Therapeutic advances in medical oncology, 2014.

[6] M. J. J. Berns and D. D. Bowtell,*The changing view of high-grade serous ovarian cancer*, Cancer Research, 2012.

[7] S. Pignata et al., *Carboplatin Plus Paclitaxel Versus Carboplatin Plus Pegylated Liposomal Doxorubicin As First-Line Treatment for Patients With Ovarian Cancer: The MITO-2 Randomized Phase III Trial*, Journal of clinical oncology, 2011.

[8] S. Sato and H. Itamochi,*Neoadjuvant chomotheraphy in advanced ovarian cancer: latest results and place in theraphy*, Cancer Research, 2015.

[9] M. Meryet-Figuière et al.,*An overview of long non-coding RNAs in ovarian cancers*, Oncotarget, 2016.

[10] X. Yan et al., *Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers*, Cancer Cell, 2015.

[11] N. Bartonicek, J. L. V. Maag and M. Dinger, *Long noncoding RNAs in cancer: mechanisms of action and technological advancements*, Molecular Cancer, 2016.

[12] W. Fu et al. *Long noncoding RNA hotair mediated angiogenesis in nasopharyngeal carcinoma by direct and indirect signaling pathways*, Oncotarget, 2016.

[13] R. A. Gupta et al.,*Long noncoding RNA HOTAIR reprograms chromatin state to*

promote cancer metastasis, Nature, 2010.

[14] K. R. Kukurba and S. B. Montgomery,*RNA sequencing and analysis*, HHS Public Access, 2015.

[15] A. Dempster, N. Laird and D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, 1977.

[16] R. Broa and a. K. Smilde, *Principal component analysis*, Analytical Methods, 2014.

[17] M. Einasto et al., *SDSS DR7 superclusters. Principal component analysis*, Astronomy & Astrophysics, 2011.

[18] J. H. McDonald,*Handbook of Biological Statistics*, 3rd ed. Sparky House Publishing, 2014.

[19] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the royal society, 1995.

[20] M. G. Kendall, *Rank correlation methods*, 2nd edition, Hafner Publishing Company, 1955.

[21] E. L. Kaplan and P. Meier, *Nonparametric estimation from incomplete observations*, Journal of the American statistical association, 1958.

[22] J. T. Rich et al.,*A practical guide to understand Kaplan-Meier curves*, HHS Public Access, 2010.

[23] M. K. Goel, P. Khanna, and J. Kishore,*Understanding survival analysis: Kaplan-Meier estimate*

[24] V. Bewick, L. Cheek, and J. Ball,*Statistics review 12: Survival analysis*, Critical care, 2004.

[25] J. M. Bland and D. G. Altman*The logrank test*, BMJ, 2004.

[26] L. Breiman,*Random Forest*, University of California - Berkeley, 2001.

[27] T. Hastie, R. Tibshirani and j. Friedman, *The elements of statistical learning - data mining, inference and prediction*, 2nd edition, Springer Series in Statistics Springer, 2009.

[28] L. Breiman,*Bagging Predictors*, University of California - Berkeley, 1994.

[29] L. Breiman,*Out of bag estimation*, University of California - Berkeley, 1996.

[30] I. Guyon and A. Elisseeff, *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research, 2003.

[31] G. Louppe, L. Wehenkel, A. Sutera and P. Geurts,*Understanding variable importances in forests of randomized trees*, Advances in Neural Information Processing Systems 26 - NIPS, 2013.

[32] X. Chen*Random forests for genomic data analysis*, Genomics, 2012

[33] R. Díaz-Uriarte and S. Alvarez de Andrés, *Gene selection and classification of microarray data using random forest*, BMC Bioinformatics, 2006.

[34] S. Arlot and A. Celisse, *A survey of cross-validation procedures for model selection*, Statistic surveys, 2010.

[35] P. Refaeilzadeh, L. Tang and H. Liu*Cross Validation*, Encyclopedia of Database Systems, Springer, 2009.

[36] T. Fawcett,*An introduction to ROC analysis*, Pattern Recognition Letters, 2006.

[37] A. P. Bradley, *The use of the area under the ROC curve in the evaluation of machine learning algorithms*, Pattern Recognition, 1997.

[38] M. B. Kursa and W. R. Rudnicki,*Feature Selection with the Boruta Package*, Journal of statistical software, 2010.

[39] M. B. Kursa, A. Jankowski and W. R. Rudnicki, *Boruta – A System for Feature Selection*,Fundamenta Informaticae, 2010.

[40] M. B. Kursa, *Robustness of Random Forest-based gene selection methods*,BMC Bioinformatics, 2014.

[41] F. Degenhardt, S. Seifert and S. Szymczak, *Evaluation of variable selection methods for random forests and omics data sets*, Briefings in bioinformatics, 2017.

[42] H. Easwaran, H. Tsai anf S. B. Baylin,*Caner epigenetics: tumor heterogeneity, plasticity of stem-like states and drug resistance*, Cell Press, 2014.

[43] J. Harrow, et al., *GENCODE: The reference human genome annotation for The ENCODE Project*, PubMed, 2012.

[44] The ENCODE Project Consortium, *An Integrated Encyclopedia of DNA Elements in the Human Genome*, Nature, 2012.

[45] A. Roberts and L. Pachter,*Streaming fragment assignment for real-time analysis of sequencing experiments*, Nature methods, 2013.

[46] L. Ma, V. B. Bajic, and Z. Zhang, *On the classification of long non-coding RNAs*, RNA Biology, 2013.

[47] Y. Saeys, T. Abeel and Y. Van de Peer ,*Robust Feature Selection Using Ensemble Feature Selection Techniques*Daelemans W., Goethals B., Morik K. (eds) Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, vol 5212. Springer, Berlin, Heidelberg, 2008.

[48] H. Abusamra,*A Comparative Study of Feature Selection and Classification Methods for Gene Expression Data of Glioma* , Proc. of 4th International Conference on Computational Systems-Biology and Bioinformatics, Procedia Computer Science, vol. 23, pp. 5-14,2013.

[49] K. Moorthy and M. S. Mohamad*Random forest for gene selection and microarray data classification*, Knowledge Technology, Springer Berlin Heidelberg, 174-183, 2012.