Universidad Politécnica de Cartagena

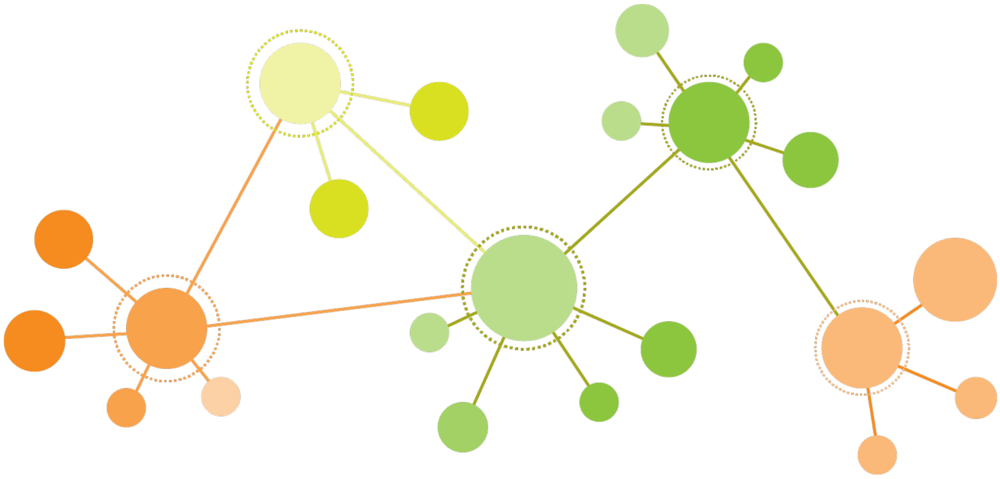
UNIVERSIDAD DE MURCIA

University of Murcia

Biology Faculty

Facultad de Biología

Statistical approach based on survival analysis to evaluate biomarkers in the diagnosis and prognosis of lung adenocarcinoma



Author:
Sergio Vela Moreno

Tutors:
Juana María Vivo Molina
Manuel Franco Nicolás

Year 2021-2022

Contents

List of Figures

Abstract

In developed countries, cancer is one of the main causes of death, being lung cancer the one with a higher death rate, and lung adenocarcinoma the leading cause of cancer-related deaths worldwide.

In recent years, research on different types of cancer has advanced in order to find suitable biomarkers for early diagnosis and cancer prognosis, as well as new targets for treatment therapies against them.

Among the different methods used for such purpose, different types of analysis are carried out, such as differential gene expression analysis of lncRNAs, miRNAs, and mRNAs, genome mutation analysis or differential methylation analysis. Key biomarkers identified with these analysis have been used to statistically evaluate cancer survival, typically by an *Overall survival (OS)*, which is worldwide considered a standard endpoint as it is a direct measure of clinical benefit. However, an extended follow-up period is required for its application, which might dilute its estimation. To overcome such requirement, alternative predictive endpoints which are also less time-consuming are performed as surrogates for *OS*, such as the *Disease free survival* (*DFS*) analysis.

In this study, differential expression analysis of mRNA, as well as differential methylation analysis of CpG sites between lung adenocarcinoma tumor and non tumor samples have been performed in order to find differentially expressed genes (*DEGs*) and differentially methylated probes (*DMPs*) among both types of samples. Transcriptome and methylome analysis results have been integrated in order to find suitable biomarkers for early lung cancer diagnosis. As a result, 47 *DMPs* have been identified as possible diagnostic indicators of lung adenocarcinoma. Afterwards, an *OS* and a *DFS* analysis have been performed in order to integrate identified *DEGs* and *DMPs* with clinical data to study possible relationships between these predictors and *OS* and *DFS* time and status respectively.

Finally, the models obtained from both survival analysis have been compared, determining that the *DFS* improved slightly the accuracy in prognosis outcomes for lung adenocarcinoma patients, explaining therefore better such relationships.

Resumen

En los países desarrollados, el cáncer es una de las principales causas de muerte, siendo el cáncer de pulmón el que presenta una mayor tasa de mortalidad, y el adenocarcinoma de pulmón la primera causa de muerte relacionada con el cáncer a nivel mundial.

En los últimos años se ha avanzado en la investigación sobre diferentes tipos de cáncer con el fin de encontrar biomarcadores adecuados para el diagnóstico precoz y el pronóstico del cáncer, así como nuevas dianas para terapias de tratamiento frente a los mismos.

Entre los diferentes métodos utilizados para tal fin, se llevan a cabo distintos tipos de análisis, como el análisis diferencial de expresión génica de lncRNAs, miRNAs y mRNAs, el análisis de mutaciones del genoma o el análisis diferencial de metilación. Los biomarcadores clave identificados con estos análisis se han utilizado para evaluar estadísticamente la supervivencia del cáncer, generalmente mediante un análisis *Overall Survival (OS)*, que se considera en todo el mundo como un criterio de valoración estándar, ya que es una medida directa del beneficio clínico. Sin embargo, se requiere un período de seguimiento prolongado para su aplicación, lo que podría diluir su estimación. Para superar este requisito, se realizan criterios de valoración predictivos alternativos que consumen menos tiempo como sustitutos de *OS*, como el análisis *Disease-Free Survival* (*DFS*).

En este estudio, se han realizado análisis de expresión diferencial de mRNA, así como análisis de metilación diferencial de sitios CpG entre muestras tumorales y no tumorales de adenocarcinoma de pulmón para encontrar genes expresados diferencialmente. (*DEGs*) y posiciones metiladas diferencialmente (*DMPs*) entre ambos tipos de muestras. Posteriormente, se han integrado los resultados de los análisis de transcriptomas y metilomas para encontrar biomarcadores adecuados para el diagnóstico precoz de este cáncer. Como resultado, 47 *DMPs* han sido identificados como posibles indicadores de diagnóstico de adenocarcinoma de pulmón.

Tras ello, un análisis *Overall survival (OS)* y *Disease free survival* (*DFS*) se ha llevado a cabo para integrar los *DEGs* y *DMPs* identificados con datos clínicos para estudiar posibles relaciones entre predictores y tiempo y estado *OS* y *DFS* respectivamente.

Finalmente, se han comparado los modelos obtenidos de ambos análisis de supervivencia, concluyendo que *DFS* mejoró ligeramente la precisión en los resultados del pronóstico para pacientes afectos por adenocarcinoma de pulmón, explicando por tanto mejor dichas relaciones.

1 Introduction

Cancer is a group of several diseases related to an uncontrolled development of the cells of multicellular organisms over time. According to recent studies, among the different types of existing cancer, breast, lung and colorectal cancer are the most frequent worldwide.

In developed countries, cancer is one of the main causes of death, being lung cancer the one with a higher death rate, despite been recently surpassed as the most frequent cancer by breast cancer. Among the different types of lung cancer, lung adenocarcinoma, which originates in the glandular tissue recovering this organ, is the leading cause of cancer-related deaths worldwide, being responsible for around 1.600.000 deaths every year[30].

As there are many different kinds of cancer, despite that the basic processes for its proliferation are quite similar, each one has its own features and characteristics. Different studies have shown that cancer development is not only related to genetic mutations or transcriptomic alterations[39], such as *mRNA*, *miRNA*[40] and *lncRNA* expression[22], but also with epigenetic alterations[2, 36], which are DNA modifications that do not change DNA sequence but can affect gene activity, being DNA methylation one of the most important epigenetic alterations, which is associated with the regulation of gene transcription. Moreover, the relationship between cancer development and its microenvironment has also been proven, as cells and molecules entering tumor microenvironment might play an important role in cancer malignancy[8, 11].

The different nature of processes involved in cancer proliferation and progression make its diagnosis and particularly its prognosis a complex task. Among the different *Survival Analysis* which are performed to study cancer prognosis, *Overall Survival*(*OS*) is considered as the standard endpoint [10]. *OS* is defined as the time from diagnosis of a disease that patients are still alive, which makes it a direct measure of clinical benefit. However, it requires an extended follow-up period, which is a disadvantage as it is not always possible. To overcome such requirement, alternative predictive endpoints are performed as surrogates for *OS*, such as the *Disease free survival* (*DFS*) and the *Progression free survival* (*PFS*) analysis [16]. *DFS* is the amount of time that a patient lives without any symptoms of a certain cancer after ending treatment for that specific cancer, whereas, *PFS* is the amount of time that a patient lives without cancer progression after ending treatment, despite the appearence of any possible cancer signs. These endpoints could be considered surrogates of *OS* to estimate cancer prognosis as they require a shorter follow-up period of patients.

The aim of this study is to explore DFS as a suitable surrogate for OS in the prognosis of lung adenocarcinoma (LUAD), through the determination of independent biomarkers which affect the evaluation of the OS and DFS of patients suffering from this type of cancer, as well as to determine which biomarkers could be used for early diagnosis.

For this purpose, transcriptome and methylome data of LUAD [4] samples have been used in order to identify differentially expressed genes (*DEGs*) as well as differentially methylated probes (*DMPs*) between tumour and non tumour samples.

Afterwards, samples containing both gene expression and methylation data are used to integrate observed values of their DEGs and DMPs, in order to find possible biomarkers for LUAD diagnosis.

Integrated data is then used to fit a *logistic regression model* able to predict whether a patient suffers from lung adenocarcinoma or not. Predictor covariates used for the generation of this model may be considered as suitable biomarkers for the early diagnosis of this cancer.

Next, both *Survival Analysis* are carried out: *OS* and *DFS* analysis are performed in order to integrate identified *DEGs* and *DMPs* with clinical data to study possible relationships between these predictors and *OS* and *DFS* time and status respectively. *DFS* analysis is presented as an alternative to *OS* analysis, often used to evaluate targeted therapies. A *Cox proportional hazards model* is fit for both cases in order to obtain estimated overall survival and disease free status probabilities of patients suffering from LUAD[15, 27].

Finally, both *Cox proportional hazards models* are compared in order to find out which one can determine prognosis for this type of cancer more accurately.

The code used for performing the different analyses described in this paper is available for reproductivility purposes and can be found in a Rmarkdown file at the Github repository *https://github.com/SergioVela17*. All the analyses have been developed using *RStudio*[24].

2 Materials and methods

2.1 Data source and data processing

Data used for this study has been obtained from the TCGA-LUAD project, which can be found in The Cancer Genome Atlas (TCGA) *(https://portal.gdc.cancer. gov/ projects/ TCGA-LUAD).* DNA methylation data and mRNA expression data of LUAD patients has been directly downloaded from the TCGA repository using the R package *TCGABiolinks* version 2.24.3 [3] in Bioconductor.

The mRNA expression data has been obtained following the *STAR-Counts* workflow[5]. This data belongs 537 patients and includes *Transcriptome Profilings* of 596 different samples, 537 of which are collected from primary tumor and 59 from normal solid tissue. Each *Transcriptome Profiling* includes expression values for 60660 human genes. When mRNA expression data is downloaded using *TCGABiolinks*, clinical information is added to the different samples automatically, among which sample type is included.

DNA methylation data has been obtained using the *Illumina Human Methylation 450* platform [12] and consists of methylation arrays that include methylation beta values from 505 samples, 473 of which are LUAD samples and the remaining 32 are control samples. Each array includes methylation beta values of 485577 CpG sites identified with their respective cg target ID. However, in this case, clinical data is not added automatically as when downloading mRNA expression. Therefore, methylation data for both types of samples has been queried and downloaded separately and binded afterwards, in order to easily identify samples belonging to each condition.

The reference genome used for RNA-Seq reads alignment has been *GRCh38.d1.vd1*, which can be obtained from The Cancer Genome Atlas as well *(https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files).* Additional clinical data associated with this project (TCGA-LUAD) has been downloaded from the cBioportal repository *(http://www.cbioportal.org).* The reason for using these data is that the original clinical data obtained from *TCGA* is incomplete and does not present information about disease-free time and status of patients. Therefore, data downloaded from this repository is used for both *OS* and *DFS* analysis.

In order to obtain *Transcriptome Profiling* files, which have been directly downloaded from TCGA, the raw reads obtained from paired-end RNA-Sequencing of lung adenocarcinoma cancer samples have previously undergone a quality control analysis. Afterwards, they have been aligned to the human reference genome using the splice-aware alignment tool *STAR (Spliced Transcripts Alignment to a Reference)* [3] and the reads associated to each gene have been counted following a *STAR-Counts workflow*, according to the TCGA pipeline.

As a result, *Transcriptome Profiling* files for each sample are obtained, which are *Gene Expression Quantification* files containing the number of reads aligned to every gene, as well as well as *RNA-Sequencing* metrics such as *TPM (Transcripts Per Kilobase Million)* and *FPKM (Fragments Per Kilobase Million).*

The *Transcriptome Profiling* files belong to both solid normal tissue and primary tumor samples and were queried and downloaded using the package *TCGABiolinks.*

With gene expression values contained in these files, a *DGE* analysis [33] was performed, comparing mRNA expression in tumor and non tumor samples. Among the different data forms used by TCGA to quantify gene expression, *FPKM* has been chosen, which is the result of first normalizing the reads aligned to each gene by differences in sequencing depth and afterwards normalizing by gene length. Despite the variety of packages for *DGE* analysis [25, 28], it has been performed using the package *NOISeq* version 2.40.0 [32, 33] in *Bioconductor* and the cutoff to consider a gene as *DEG* has been 0.9. The parameter *"norm"* has been set as "n" because genes expression has been previously normalized and the parameter *"replicates"* has been set as "no", as there are no actual biological replicates. The rest of parameters have been set as default. An overview of the main steps followed in RNA-Seq analysis can be found in Figure 1.

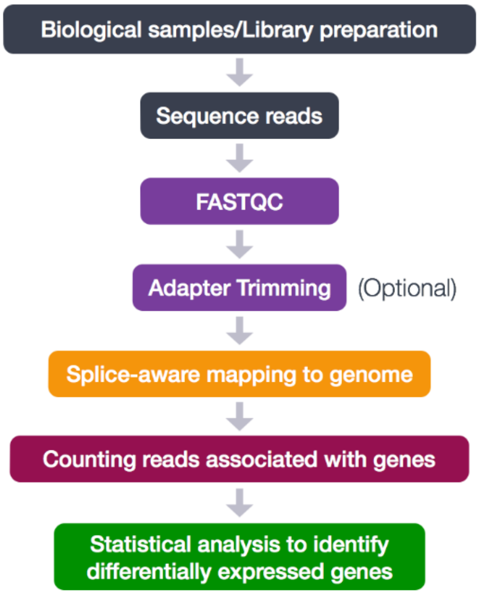

Biological samples/Library preparation
Sequence reads
FASTQC
Adapter Trimming
(Optional)
Splice-aware mapping to genome
Counting reads associated with genes
Statistical analysis to identify differentially expressed genes

Figure 1: Main steps followed in RNA-Seq analysis

DNA Methylation data was queried and downloaded using the package *TCGABiolinks* as well. Methylation data is then filtered in order to eliminate missing values using the function '*na.omit*' from the package *data.table* version 1.14.2 [6]. After data filtering, methylation values were normalized and differential methylation analysis between tumor and non tumor samples was performed using *ChAMP* package version 2.26.0 [35], in order to obtain differential *DMPs*. The parameters for this analysis have been the data frame containing normalized methylation values, the vector which indicates the phenotype of each sample and *arraytype="450K"*. This function implements the package *limma* to automatically calculate the adjusted p-value for differential methylation using a linear model. Among all the CpGs identified as DMPs, only those belonging to enhancer regions were selected for downstream analysis.

Samples containing both mRNA expression and DNA methylation data were selected in order to integrate both forms of data in a single data frame.

2.2 Correlation and Functional Enrichment Analysis

Once data had been integrated, a *Functional Enrichment Analysis* was performed using the function '*gost*' from the package *gprofiler2* version 0.2.1 [19], in order to identify biological processes related to *DEGs*. The parameters introduced for this function were the list of identified *DEGs* and *organism="hsapiens"* as organism. The rest of parameters were set as default, in order to obtain all results and use all the available data sources for research.

Afterwards, a Pearson correlation analysis was performed with package *caret* version 6.0-92 [17] to find strong relationships (cutoff=0.9) between *DEGs* expression and *DMPs* methylation values.

2.3 Cross-validation logistic regression model via elastic-net penalty

Non correlated *DEGs* and *DMPs* have been used to generate a *logistic regression model* for tumor state prediction using function '*cv.glmnet*' from the package *glmnet* version 4.1-4 [26].

In order to generate this model, data samples have been split into train and test data using the function *'createDataPartition'* from the package *caret*, using 75% of the samples for training and the remaining 25% for testing. The arguments used to generate this model have been the matrix containing train data values (x), a vector containing the response variable (y), the parameter *"model family"* which has been set as *"binomial"*, as this model aims to classify samples into two groups, and *"alpha=0.5"*) which is the *"ElasticNet mixing parameter"* and is related to the penalty. The rest of parameters were set as default, which means that the *number of folds* for cross-validation is *nfolds=10*, running the function 11 times.

2.4 Survival analysis

Finally, *DEGs* expression and *DMPs* methylation values were integrated with clinical data on survival and disease-free time and status to perform an overall survival and disease free survival analysis, using the package *survival* version 3.3-1 [34], in order to study relationships between predictors and *OS* and *DFS* status and time of LUAD patients.

Firstly, *Kaplan-Meier* curves were obtained using the function *'survfit'* in the package *survival.* The parameters used were survival time and status as independent variables, as well as, clinical data such as *age* and whether patients had received or not *targeted molecular therapy* as categorical variables to divide the samples into groups.

Afterwards, a *Cox proportional hazards model* [18, 26] for both analysis was fitted using the function 'glmnet' in the package *glmnet*, which handles ties in survival with the *Breslow* approximation. The parameters used for both analysis were a matrix containing the data of all the possible predictor variables considered, a matrix containing response variables (this is a 2-column matrix: first column corresponds to survival time, whereas, second column corresponds to status) and the parameter *"family"*, which has been set as *"cox"* as these models aim to study the relationship between predictor variables and survival time, setting the rest of parameters as default. Covariates included in models obtained this way are considered significant for survival estimation, and were selected and included in new models using the function '*coxph*' in package *survival*, that provides statistical characteristics of the model such as the Harrell C index, which is similar to the area under the curve (AUC) measure of concordance for survival data and is an estimate of how well the model predicts individual survival.

Finally, statistics of models from both survival analysis were compared in order to determine which analysis estimates survival time more accurately.

The main workflow followed in this study is displayed in Figure 2.

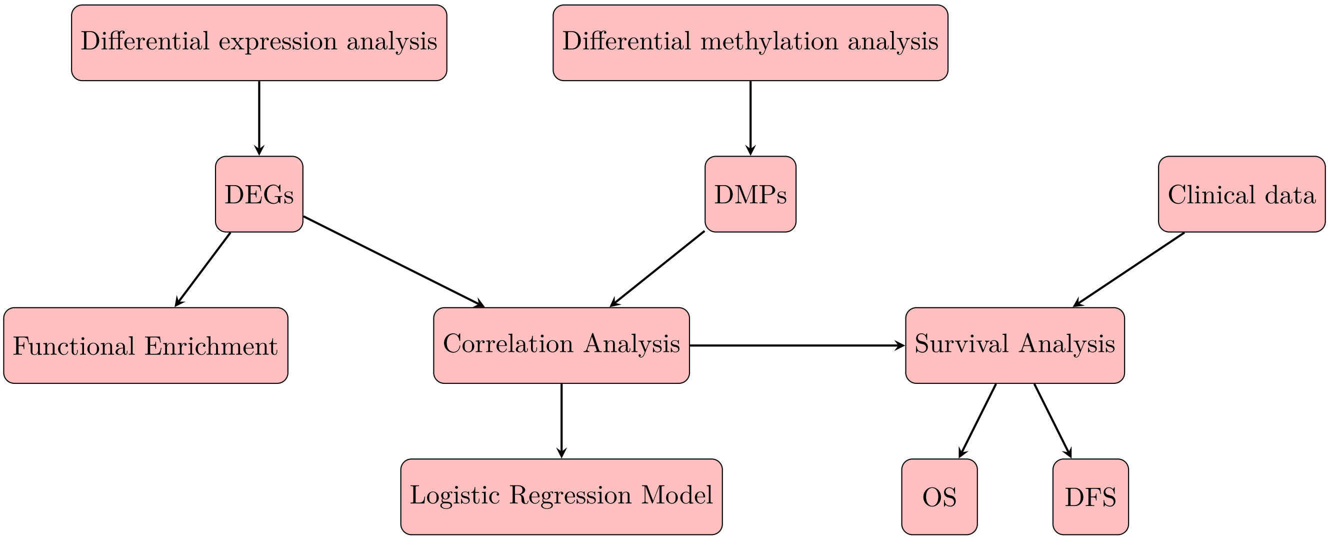

Differential expression analysis
Differential methylation analysis
DEGs
DMPs
Clinical data
Functional Enrichment
Correlation Analysis
Survival Analysis
Logistic Regression Model
OS
DFS

Figure 2: General workflow followed during this study

3 Results

3.1 Identification of DEGs and DMPs

After performing data filtering to eliminate samples containing missing values in mRNA expression data frame, the *DEG* analysis reported seven genes which have been identified as differentially expressed between cancerous and healthy samples. These genes were *FF2, RNY3, RNU5B-1, RNA5-8SP2, LINC00676, REG4* and *DEFA5*, all of which appear to be overexpressed in tumor samples.

Methylation beta values contained in a data frame are filtered in order to eliminate missing values and afterwards, differential methylation analysis is performed between tumor and non tumor samples, identifying 150903 DMPs between both groups of samples among the 325821 unique CpG sites finally used for the analysis.

An overview of *DMPs* identified with *ChaMP* is shown in Figure 3, which can be obtained using the function '*CpG.GUI*' and displays CpGs grouped by chromosome, genomic and messenger RNA region they are located in.

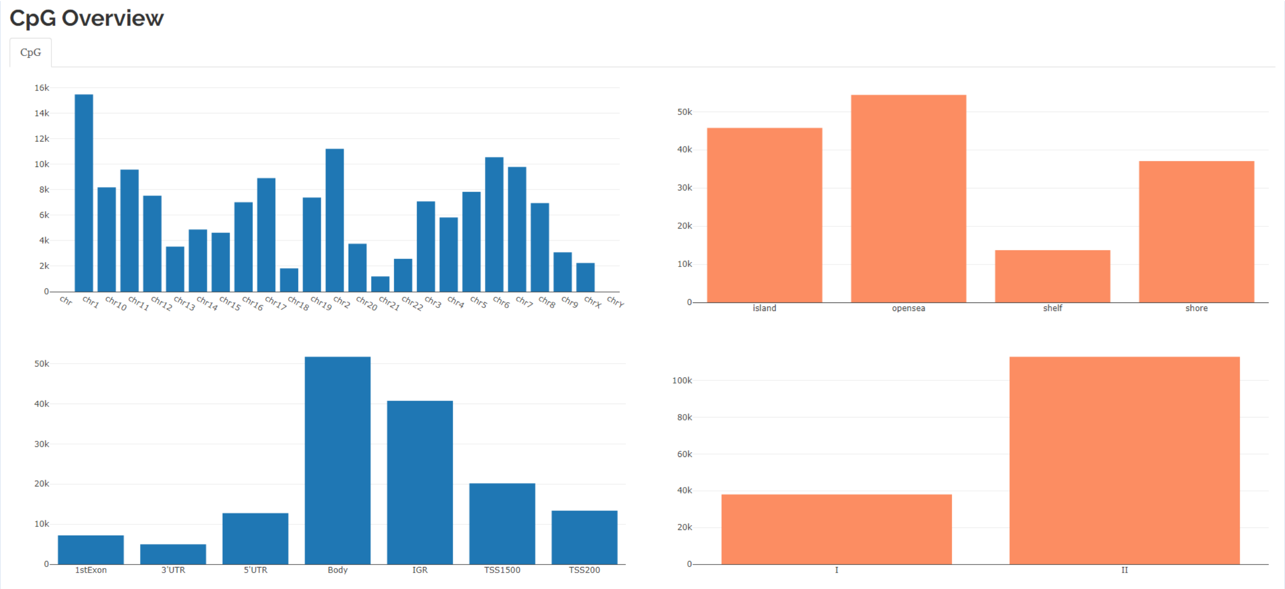
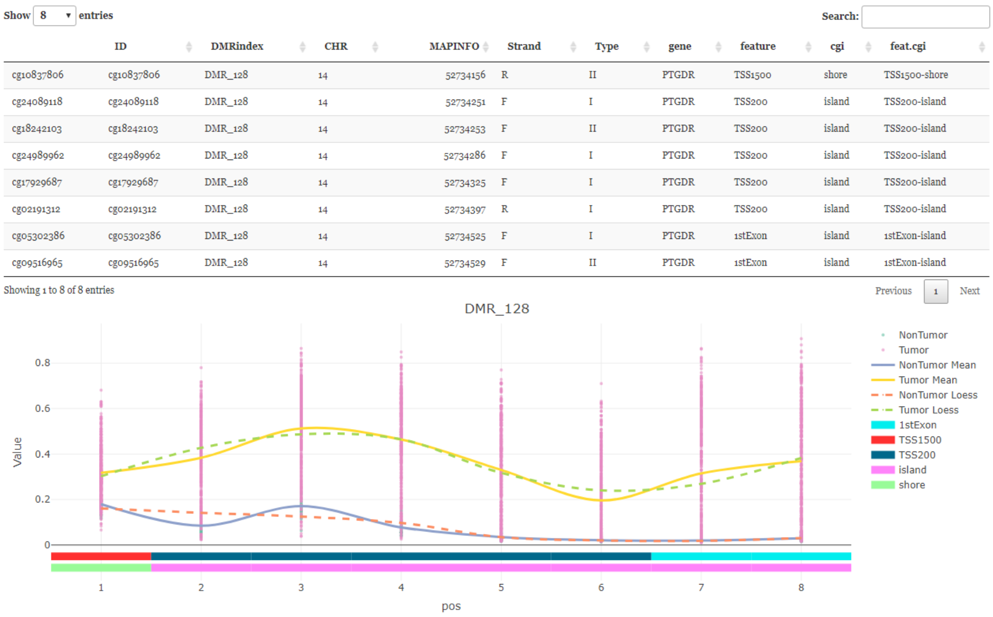


Figure 3: Overview of *DMPs* distribution using function CpG.GUI

Most *DMPs* identified between both sample types are located in chromosomes 1, 2, 6 and 7, and they mainly belong to gene body and intergenic regions of mRNA sequences.

However, only the *DMPs* located in enhancer regions are used for further analysis because these regions methylation is directly related to gene expression from a biological perspective as its high methylation associates with gene silencing[7, 31]. Therefore, only 37936 of the 150903 identified DMPs are considered downstream.

ChaMP also allows to identify differentially methylated regions (*DMRs*) conformed by *DMPs* using function '*champ.DMR*'. The total count of DMPs is distributed into 880 different DMRs, which can be observed using function '*DMR.GUI*' that also allows to visualize methylation values of DMPs conforming each region, as in the example shown in Figure 4.



Show 8 entries

Search:

	ID	DMRindex	CHR	MAPINFO	Strand	Type	gene	feature	cgi	feat.cgi
cg10837806	cg10837806	DMR_128	14	52734156	R	II	PTGDR	TSS1500	shore	TSS1500-shore
cg24089118	cg24089118	DMR_128	14	52734251	F	I	PTGDR	TSS200	island	TSS200-island
cg18242103	cg18242103	DMR_128	14	52734253	F	II	PTGDR	TSS200	island	TSS200-island
cg24989962	cg24989962	DMR_128	14	52734286	F	I	PTGDR	TSS200	island	TSS200-island
cg17929687	cg17929687	DMR_128	14	52734325	F	I	PTGDR	TSS200	island	TSS200-island
cg02191312	cg02191312	DMR_128	14	52734397	R	I	PTGDR	TSS200	island	TSS200-island
cg05302386	cg05302386	DMR_128	14	52734525	F	I	PTGDR	1stExon	island	1stExon-island
cg09516965	cg09516965	DMR_128	14	52734529	F	II	PTGDR	1stExon	island	1stExon-island

Showing 1 to 8 of 8 entries

Figure 4: Example of a *DMR* characterization

3.2 Correlation and Functional Enrichment Analysis

The list of identified *DEGs* is used to perform a *Functional Enrichment Analysis* and, as a result, *DEGs* identified appear to be only related to the *Molecular Function* with ID *GO:0031723* from the database *Gene Ontology*, as shown in Figure 5, which corresponds to *"CXCR4 chemokine receptor binding"*. The CXCR4 receptor has been related to different types of diseases, such as many types of cancer, being responsible for directly controlling cell proliferation of non-hematopoietic cells and promoting tumor growth [1].

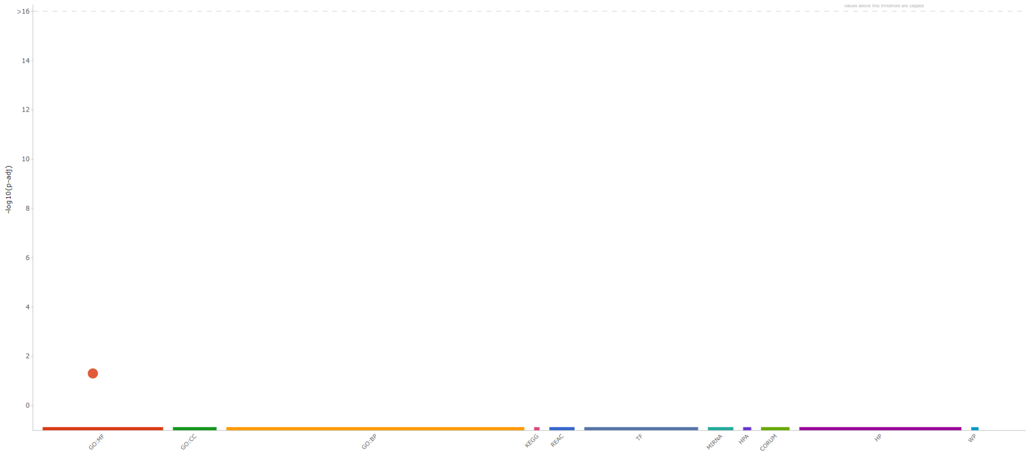


Figure 5: Functional Enrichment Analysis result

Once both differential analysis are performed, *DEGs* expression and methylation values of *DMPs* from samples with both kinds of data is obtained. After data integration, *Pearson* correlation analysis is carried out to identify possible relationships between genes and methylation probes included in enhancer regions.

Highly correlated covariates (cutoff=0.9) are removed from the data frame, as well as covariates with near zero variance, remaining a total of 35190 covariates (5 of which are *DEGs* and 35185 are *DMPs*) and a single class variable from 483 samples (21 coming from non tumor tissue and 462 from tumor tissue).

3.3 Identification of the potential diagnostic DEGs and DMPs candidates by cross-validation logistic regression model via elastic-net penalty

Samples contained in the resulting data frame are divided into training and test data in order to generate a *logistic regression model with elastic net penalty* for tumor state prediction. Prediction results obtained with this model as well as test data actual state are compared using a confusion matrix, which is represented in supplementary Table S1, verifying that logistic regression model generated successfully classified all 120 test cases (5 non tumor and 115 tumor cases). Therefore, the model obtained has proven to be effective for lung cancer diagnosis when applied to the data contained in this project.

Among the 35190 different covariates used to generate this model, only 47 of them are actually identified as possible biomarkers that contribute to state prediction, all of which are *DMPs*. The list of *DMPs* identified as suitable biomarkers as well as their respective coefficients is included in supplementary Table S2.

Probes identified this way can be displayed using package *ChaMp* to verify their distribution, as shown in Figure 6. Most of these *DMPs* are located in chromosomes 2, 1, 10 and 4, belong to gene body and intergenic type regions of mRNA and appear to be hypermethylated in tumor cases when compared with non tumor cases. Some genetic changes in previously mentioned chromosomes have been associated with other types of cancer already, such as chromosome 2 which associates with a form of blood cancer known as acute myeloid leukemia and chromosome 10, whose genetic changes have often been associated with brain gliomas and its genes correct expression is suggested to be critical in controlling cells division and growth rate.

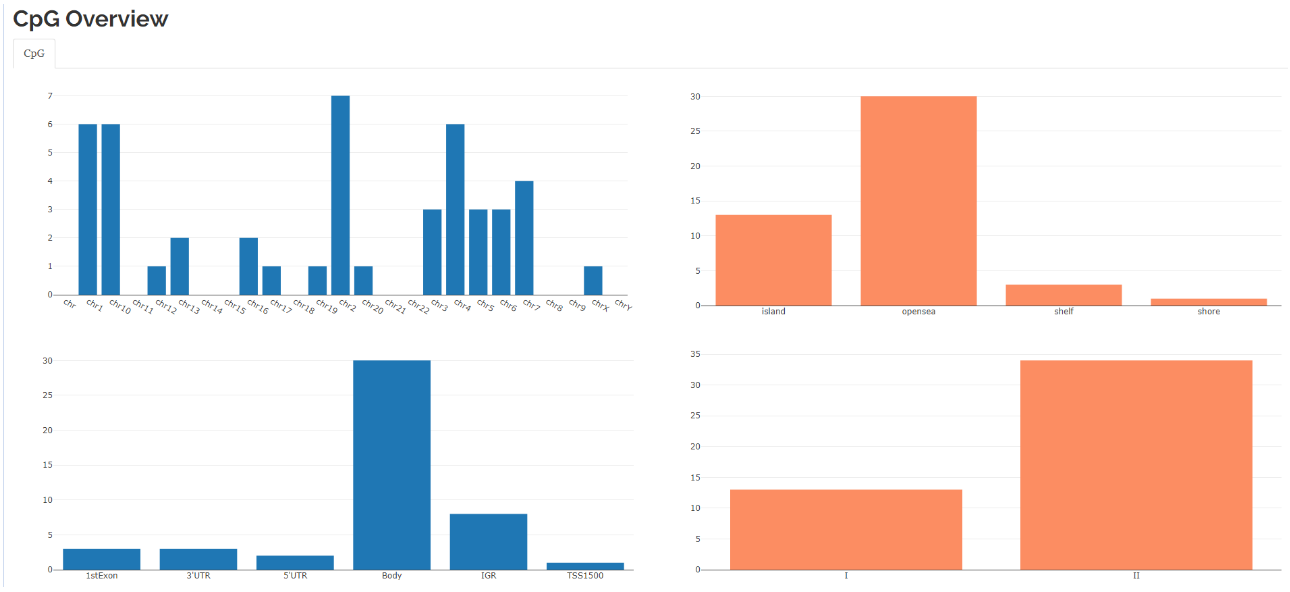


Figure 6: Overview of *DMPs* identified as possible diagnosis biomarkers distribution

3.4 Overall survival analysis

Clinical data of patients obtained from *cBioportal* contains information from 586 patients affected of LUAD. After data filtering and cleaning of patients lacking *OS* and *DFS* time and status, data from 436 cancer affected patients (108 of which have deceased) is kept . The distribution of *OS* status against time (measured in months) has been plotted in Figure 7, where deceased patients are represented in blue while living patients are represented in red. At the moment of the study, although many patients deceased at an earlier survival time, most living patients are observed to have higher *OS* times, being the mean *OS* time of 31 months.

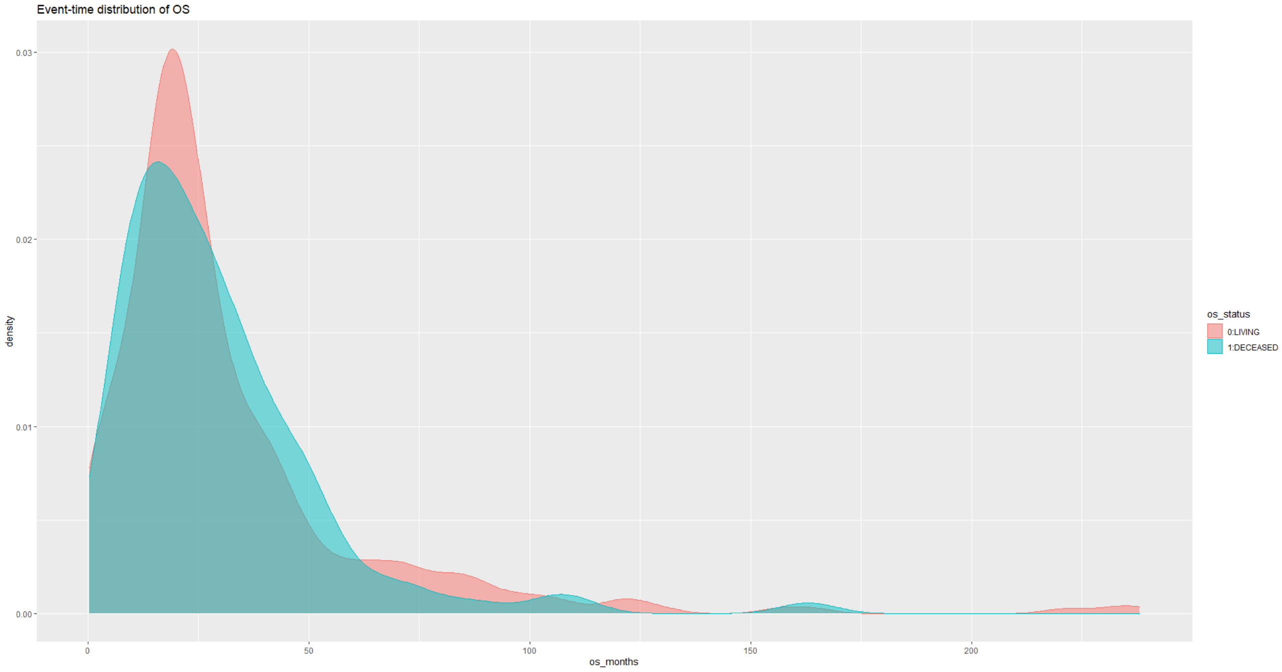


Figure 7: *OS* of lung adenocarcinoma patients grouped by *OS months*

3.4.1 Kaplan-Meier method for OS

The Kaplan-Meier method allows to make a first estimate that relates the probability of survival against time. In Figure 8, it can be observed how survival probability (represented as ”0:LIVING”) dicreases through time, until it meets the death probability curve. This happens at almost 72 months of survival and, after this time, death probability is higher than survival probability for this type of cancer.

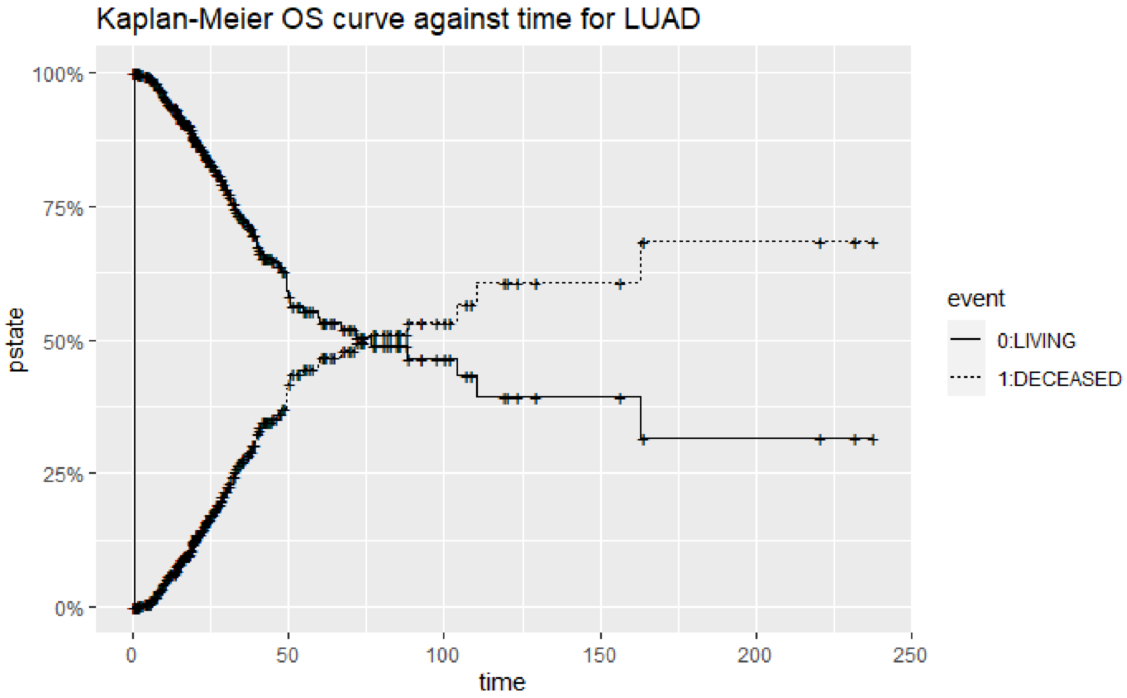


Figure 8: *Kaplan-Meier OS* curve against time in *LUAD* patients

If an independent variable is selected, this type of curve also allows to relate the proportion of recurrent cases against time, grouped based on the independent variable, such as the age of the patients studied. In Figure 9, patients are grouped based on whether they are older or younger than 60 years, an age close to which some types of cancer are usually diagnosed. It can be seen that in most cases of decease, the patients are over 60 years old but they have a higher *OS* mean time than the younger group of patients, as reported in supplementary Table S3.

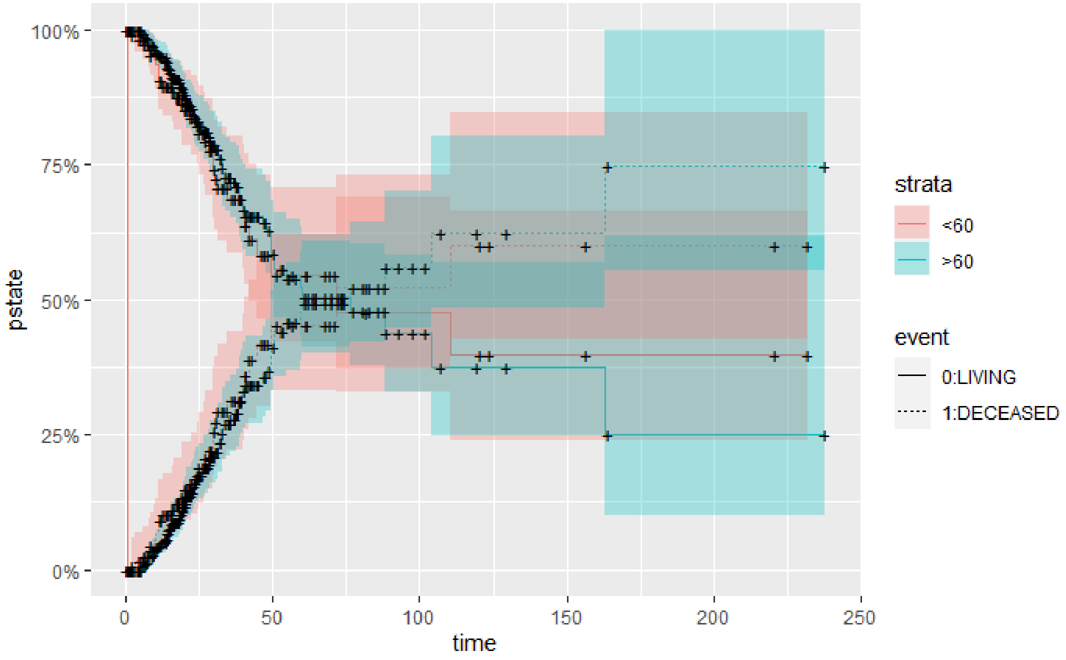


Figure 9: *Kaplan-Meier OS* curve against time in *LUAD* patients grouped by age

Similarly, patients can be grouped based on whether or not they have received targeted molecular therapy. As shown in Figure 10, mostly only those patients whose survival time is relatively short have undergone this treatment. According to the results shown in supplementary Table S4, only 30 patients have undergone targeted molecular treatment, 6 of which have deceased, and they have a higher *OS* mean time.

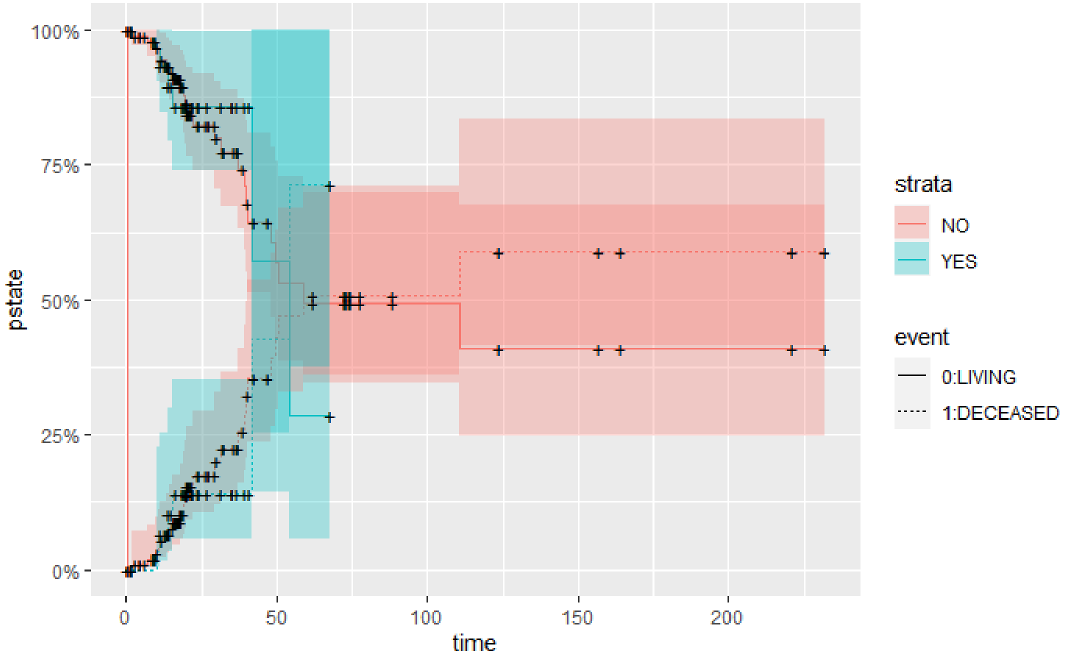


Figure 10: *Kaplan-Meier OS* curve against time in *LUAD* patients grouped by targeted molecular therapy

3.4.2 Cox proportional hazards model

Next, a *Cox proportional hazards model* is generated, which is frequently used to relate predictor variables with survival time and status.

After cleaning and filtering, the clinical data has been integrated with DEGs and the DMPs information, obtaining a data frame that contains 132 rows, each one corresponding to a case of lung adenocarcinoma, and 35,194 columns, 4 of which correspond to clinical data (time and status for overall survival, age and treatment with target molecular therapy), 5 to expression of DEGs, and the rest to methylation values in the selected probes.

The resulting Cox proportional hazards model has a concordance equal to 0.512 and uses 97 covariates of the data frame provided to determine *OS* time and status of patients. All of the covariates used to generate the model are DMPs.

3.5 Disease free survival analysis

Same clinical data can be used to plot the distribution of *DFS* status against time, as shown in Figure 11, where disease-free patients are represented in red and patients whose disease has recurred or progressed are represented in blue. Out of the 436 patients considered, cancer has recurred or progressed in 189 of them. At the moment of the study, most patients affected by LUAD experimented cancer recurrence at low *DFS* times. However, some patients remained disease free more time, with a mean *DFS* time of 26 months.

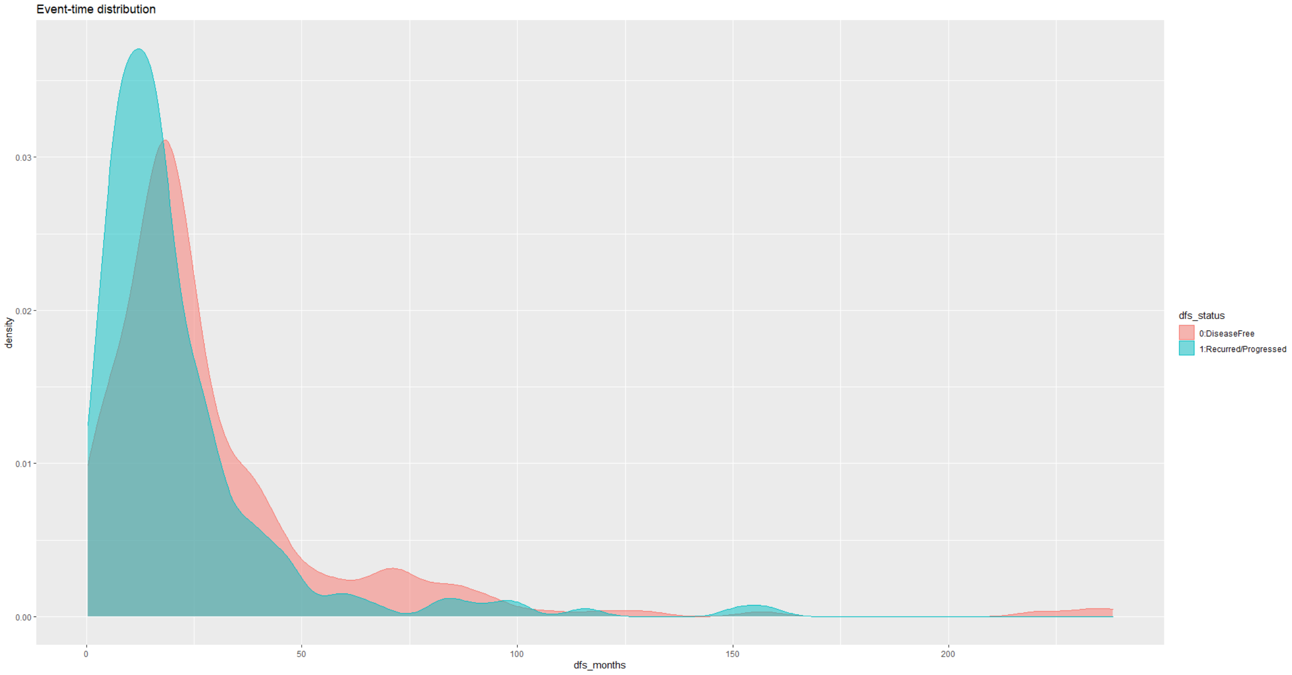


Figure 11: *DFS* of lung adenocarcinoma patients grouped by *DFS months*

3.5.1 Kaplan-Meier method for DFS

The Kaplan-Meier method is used to make a first estimate that relates the probability of recurrence against time. In Figure 12, it can be observed how disease free probability (represented as "0:DiseaseFree") decreases through time, being recurrence or progression of this type of cancer more likely after about 35 months of *DFS*.

When patients are grouped based on whether they are older or younger than 60 years in Figure 13, it seems that in most cases of recurrent disease, the patients are over 60 years old,

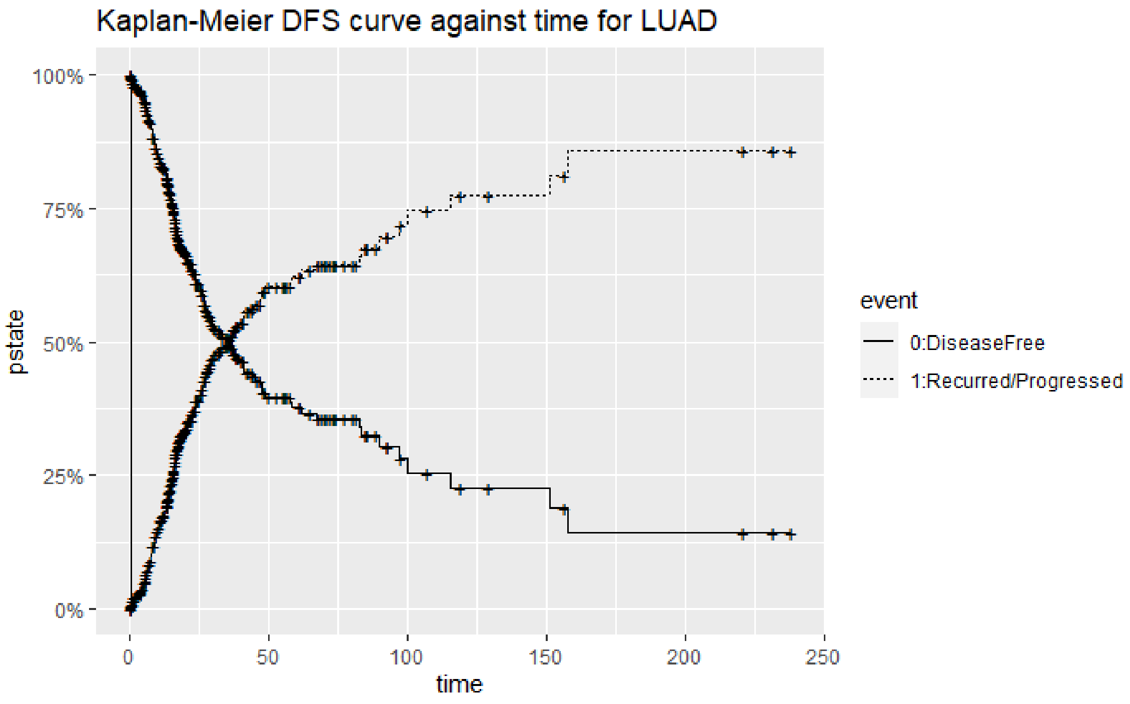


Figure 12: *Kaplan-Meier DFS* curve against time in *LUAD* patients

although recurrence in both groups occurs and *DFS* mean time is higher in older patients, as shown in supplementary Table S5.

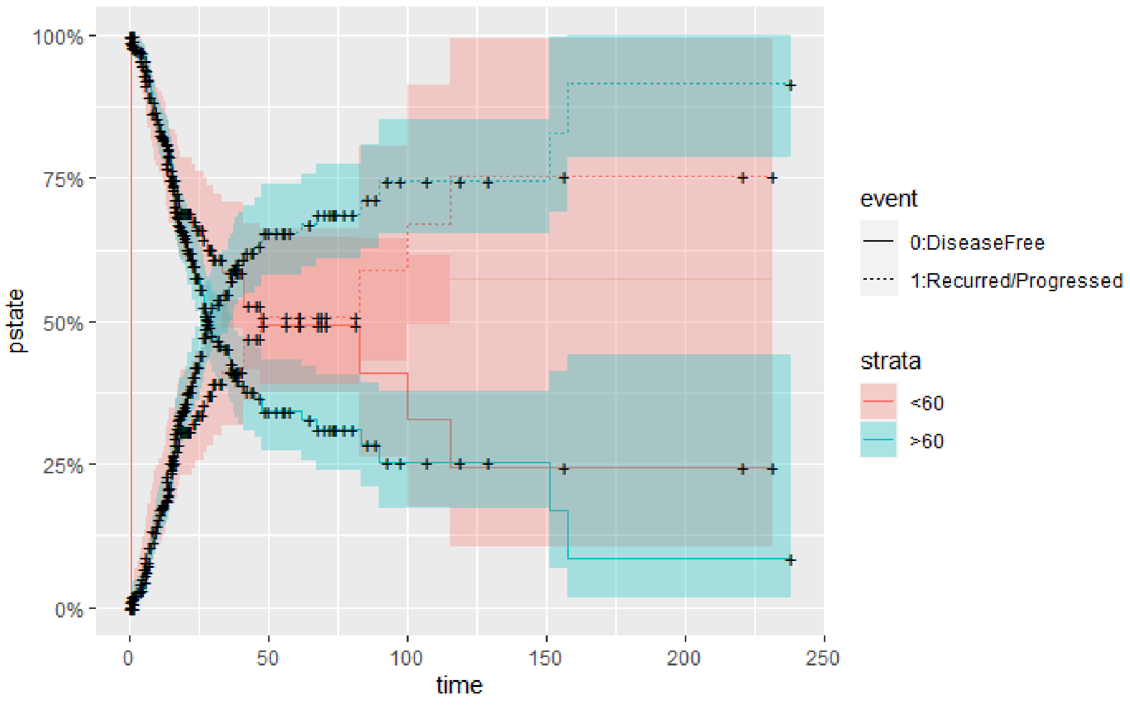


Figure 13: *Kaplan-Meier DFS* curve against time in *LUAD* patients grouped by age

When grouping patients based on targeted molecular therapy in Figure 14, mostly only those patients whose disease-free time is relatively short have undergone this treatment. Recurrence in treated patients appears to be slightly higher but they have higher *DFS* mean time according to the results in supplementary Table S6.

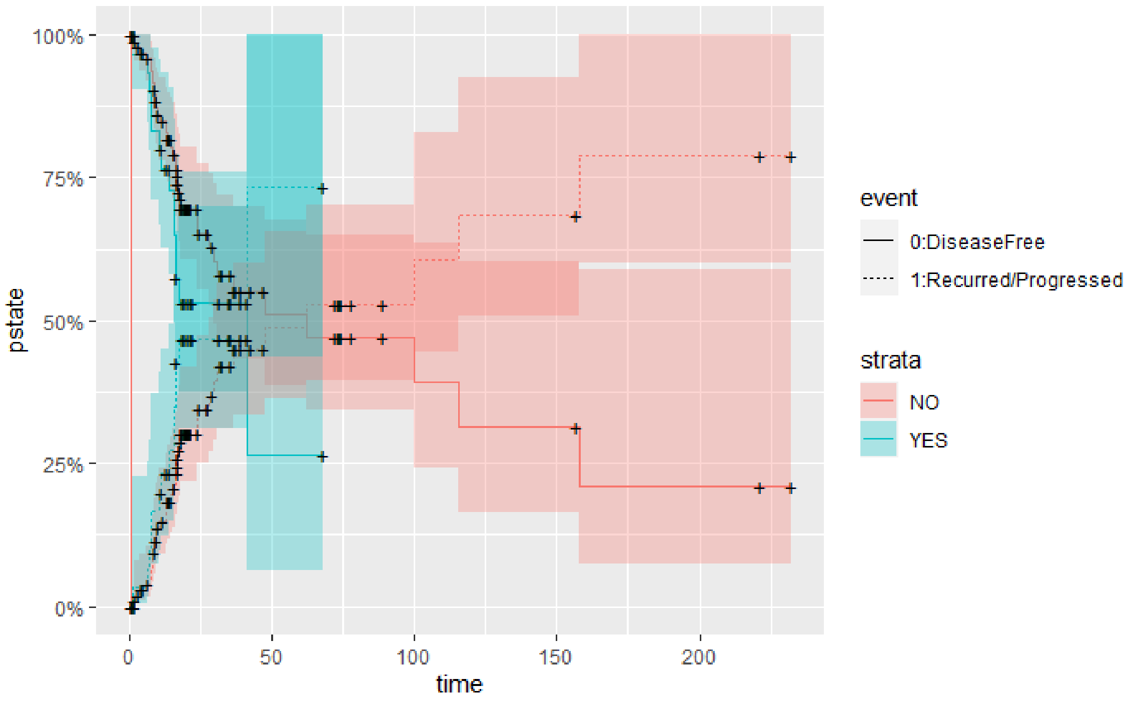


Figure 14: *Kaplan-Meier DFS* curve against time in *LUAD* patients grouped by targeted molecular therapy

3.5.2 Cox proportional hazards model

Next, a *Cox proportional hazards model* is generated for the DFS analysis. In this case, after cleaning and filtering of clinical data, it has been integrated with data from *DEGs* and *DMPs* into another data frame consisting of 35,194 columns. However, columns containing clinical information in this dataframe correspond to time and status for disease-free survival, age and treatment with target molecular therapy.

The resulting Cox proportional hazards model has a concordance equal to 0.574 and uses 94 covariates of the data frame provided to determine *DFS* time and status of LUAD patients. All of the covariates used to generate the model are DMPs.

As the model generated this way has a higher concordance than the OS Cox model, covariates included in this model are considered better prognosis biomarkers, including them and their relevance in the model in supplementary Table S7.

DMPs identified in this model are visualized with *ChaMP* in Figure 15. Most of these *DMPs* are located in chromosomes 1 and 2 and belong to gene body and intergenic type regions of mRNA. As previously mentioned, genetic changes in chromosome 2 has been associated with different types of cancer state as for chromosome 1, it has been proven to be associated with different types of cancers such as neuroblastoma and an increased risk of developing leukemia, as it is believed to contain a gene that prevents cells from rapidly growing and hypermethylation of enhancer DMPs in this chromosome might interfere with this tumor suppressor gene expression.

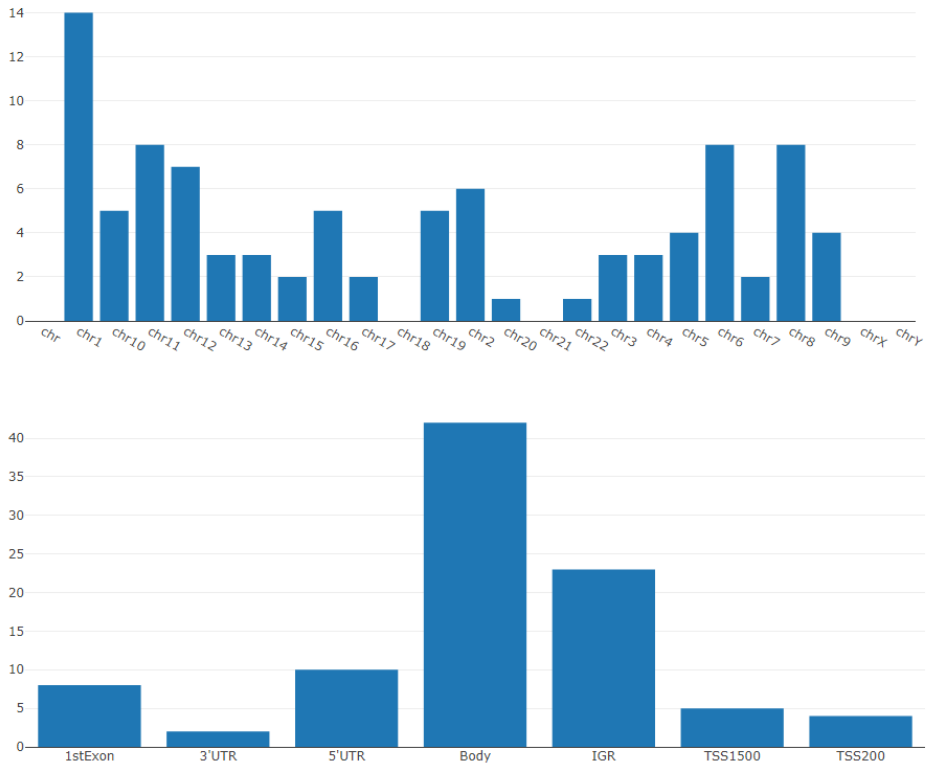


Figure 15: Overview of *DMPs* identified as possible treatment prognosis biomarkers distribution

4 Discussion

Although 7 *DEGs* have been found between tumor and non tumor samples with *DGE* analysis, none of them appears to be a suitable biomarker for lung adenocarcinoma diagnosis. However, they have been related to binding with *CXCR4 chemokine receptor*, which is known for controlling cell proliferation and promoting tumor growth[1]. In fact, overexpression of *CXCR4* has been related with many different types of cancer like breast cancer, where CXCR4 expression is upregulated by the effect of a *lncRNA*[29]. Therefore, further research on biological processes related to this receptor might be crucial to understand LUAD and other cancers.

On the other hand, some *DMPs* have proven to successfully diagnose cancer for this particular dataset, as represented in Table S1. The 47 *DMPs* identified as possible diagnosis biomarkers are located in enhancer regions and most of them are hypermethylated in tumor samples when compared to non tumor samples. Aberrant DNA methylation is one of the biggest epigenetic alterations, as it has been proven to have a significant effect on gene expression, which might lead to several disease conditions, such as different types of cancers like colorectal[38] and oesophageal cancer[20]. For instance, hypermethylation of enhancer regions in different types of genes, such as tumor supressor genes, has been related to their silencing and consequently to tumor development[37]. Identified *DMPs* are mainly distributed in chromosomes 2, 1, 10 and 4, where some genetic alterations have been previously associated with different types of cancer. For instance, correct expression of genes in chromosome 10 is suggested to be critical in controlling cells division and growth rate changes, and genetic alterations within this chromosome have often been associated with brain gliomas[13]. Moreover, the rearrangement of the RET gene, located in this chromosome, has been associated with thyroid cancer[23]. Changes in chromosome 2 have also been associated with different types of cancer, like a form of blood cancer known as acute myeloid leukemia[21]. As covariates are standardized before fitting the *logistic regression model* when the package *glmnet* is used, covariates with higher absolute value coefficients can be considered as the most relevant diagnosis biomarkers[9]. Therefore, *DMPs* identified as better diagnosis biomarkers are *cg15073624* and *cg11653410*, which can be tracked using the *UCSC human genome browser*[14]. These *DMPs* are found in enhancer regions of chromosomes 19 and 1 respectively.

When comparing *OS* and *DFS* analysis, *DFS* appears to be slightly more accurate to explain LUAD prognosis, as *Cox proportional hazards model* generated for this analysis has a higher Harrell C index, which is similar to the area under the curve (AUC) measure of concordance for survival[26], and therefore can be considered as an estimate of how well the model predicts individual survival. According to the *DFS* analysis, identified *DEGs*, as well as clinical data included, such as age, are not suitable biomarkers to estimate *DFS* time and status in LUAD patients, as they have not been included in the resulting *Cox proportional hazards model.* However, these covariates might be useful to understand cancer mechanisms.

On the contrary, among the different *DMPs* considered, 94 of them appear to be possible biomarkers for time and status determination. Most of these *DMPs* are located in previously mentioned chromosomes, such as chromosomes 1 and 2. Therefore, further studies of *DMPs* located in these chromosomes could be relevant for understanding LUAD prognosis as well as diagnosis. Based on the Pr(> |z|) value of covariates included in the supplementary Table S7, variables with the lowest p-value are considered as the most significant ones to explain LUAD prognosis, such as *cg16474118* in chromosome 2 and *cg17149019* in chromosome 3.

5 Conclusions

Cancer is a complex multifactorial disease associated to different variations, among which changes in genome, transcriptome and methylome can be found.

In order to early diagnose LUAD, the study of the chromosomes in which *DMPs* identified as potential biomarkers are located might be of special interest. Further research on these *DMPs* should be done to report gene expression affected by their methylation.

DFS analysis appears to be slightly more accurate for LUAD prognosis than *OS*, according to the concordance index of the generated Cox proportional hazards models. Therefore, *DFS* appears to be a suitable surrogate of *OS* and independent covariates identified as relevant in *DFS* prognosis model could be relevant cancer biomarkers. However, further data collection and analysis must be performed in order to determine if identified predictors are actually relevant biomarkers for LUAD diagnosis and prognosis.

Integration of other data such as *long non-coding RNA (lncRNA)* and *micro RNA (miRNA)* expression quantification might be useful for the study of this cancer. Coexpression analysis with tumor microenvironment infiltrating cells such as inmune cells could be as well relevant when studying this type of cancer. *DMPs* located in non enhancer regions should as well be included for further analysis.

Future research lines also include the implementation of other survival analysis such as *PFS*, which should be performed to determine if they constitute a better endpoint for LUAD survival estimation.

6 References

-

Bianchi, M. E., & Mezzapelle, R. (2020). The Chemokine Receptor CXCR4 in Cell Proliferation and Tissue Regeneration. *Front. Immunol.*, *11*, 2109. https://doi.org/10.3389/FIMMU.2020.02109/XML/NLM

Bjaanæs, M. M., Fleischer, T., Halvorsen, A. R., Daunay, A., Busato, F., Solberg, S., Jørgensen, L., Kure, E., Edvardsen, H., Børresen-Dale, A. L., Brustugun, O. T., Tost, J., Kristensen, V., & Helland, Å. (2016). Genome-wide DNA methylation analyses in lung adenocarcinomas: Association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. *Mol. Oncol.*, *10*(2), 330–343. https://doi.org/10.1016/j.molonc.2015.10.021

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., Ceccarelli, M., Bontempi, G., & Noushmehr, H. (2016). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, *44*(8), e71. https://doi.org/10.1093/nar/gkv1507

Collisson, E. A., Campbell, J. D., Brooks, A. N., Berger, A. H., Lee, W., Chmielecki, J., Beer, D. G., Cope, L., Creighton, C. J., Danilova, L., Ding, L., Getz, G., Hammerman, P. S., Hayes, D. N., Hernandez, B., Herman, J. G., Heymach, J. V., Jurisica, I., Kucherlapati, R., . . . Cheney, R. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nat. 2014 5117511*, *511*(7511), 543–550. https://doi.org/10.1038/nature13385

Dobin, A. (2019). Manual_STAR, 1–50. https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

Dowle, M., & Srinivasan, A. (2021). *Data.table: Extension of 'data.frame'* [R package version 1.14.2]. https://CRAN.R-project.org/package=data.table

Feinberg, A. P., & Tycko, B. (2004). The history of cancer epigenetics. *Nat. Rev. Cancer 2004 42*, *4*(2), 143–153. https://doi.org/10.1038/nrc1279

Hanahan, D., & Coussens, L. M. (2012). Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. https://doi.org/10.1016/j.ccr.2012.02.022

Hastie, T., Martin, R. T., Hastie, W., Tibshirani, •., & Wainwright, •. (n.d.). Statistical Learning with Sparsity The Lasso and Generalizations Statistical Learning with Sparsity.

Hess, L. M., Brnabic, A., Mason, O., Lee, P., & Barker, S. (2019). JournalofCancer Relationship between Progression-free Survival and Overall Survival in Randomized Clinical Trials of Targeted and Biologic Agents in Oncology. *10.* https://doi.org/10.7150/jca.32205

Hong, W., Liang, L., Gu, Y., Qi, Z., Qiu, H., Yang, X., Zeng, W., Ma, L., & Xie, J. (2020). Immune-Related lncRNA to Construct Novel Signature and Predict the Immune Landscape of Human Hepatocellular Carcinoma. *Mol. Ther. - Nucleic Acids*, *22*(December), 937–947. https://doi.org/10.1016/j.omtn.2020.10.002

Illumina. (2012). Infinium Human Methylation 450 data sheet. *Illumina*, 3.

Kannan, S., Murugan, A. K., Balasubramanian, S., Munirajan, A. K., & Alzahrani, A. S. (2022). Gliomas: Genetic alterations, mechanisms of metastasis, recurrence, drug resistance, and recent trends in molecular therapeutic options. *Biochem. Pharmacol.*, *201*, 115090. https://doi.org/10.1016/J.BCP.2022.115090

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler, & David. (2002). The Human Genome Browser at UCSC. *Genome Res.*, *12*(6), 996–1006. https://doi.org/10.1101/GR.229102

Li, M., Zhang, L., Feng, M., & Huang, X. (2022). m6A-Related lncRNA Signature Is Involved in Immunosuppression and Predicts the Patient Prognosis of the Age-Associated Ovarian Cancer. (F. Wang, Ed.). *J. Immunol. Res.*, *2022*, 3258400. https://doi.org/10.1155/2022/3258400

Li, N., & Services, C. P. (2019). Progression-Free Survival (PFS) Analysis in Solid Tumor Clinical Studies.

Max, A., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Ziem, A., Scrucca, L., Hunt, T., & Kuhn, M. M. (2022). *Package ' caret ' R topics documented :*

McGough, S. F., Incerti, D., Lyalina, S., Copping, R., Narasimhan, B., & Tibshirani, R. (2021). Penalized regression for left-truncated and right-censored survival data. *Stat. Med.*, *40*(25), 5487. https://doi.org/10.1002/SIM.9136

Peterson, H., Kolberg, L., Raudvere, U., Kuzmin, I., & Vilo, J. (2020). gprofiler2 – an R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler. *F1000Research*, *9*, 1–27. https://doi.org/10.12688/f1000research.24956.2

Poosari, A., Nutravong, T., Namwat, W., Wasenang, W., Sa-ngiamwibool, P., & Ungareewittaya, P. (2022). The relationship between P16INK4A and TP53 promoter methylation and the risk and prognosis in patients with oesophageal cancer in Thailand. *Sci. Reports 2022 121*, *12*(1), 1–10. https://doi.org/10.1038/s41598-022-14658-0

Quessada, J., Cuccuini, W., Saultier, P., Loosveld, M., Harrison, C. J., & Lafage-Pochitaloff, M. (2021). Cytogenetics of pediatric acute myeloid leukemia: A review of the current knowledge. *Genes*, *12*(6). https://doi.org/10.3390/genes12060924

Ren, S., Peng, Z., Mao, J. H., Yu, Y., Yin, C., Gao, X., Cui, Z., Zhang, J., Yi, K., Xu, W., Chen, C., Wang, F., Guo, X., Lu, J., Yang, J., Wei, M., Tian, Z., Guan, Y., Tang, L., … Sun, Y. (2012). RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res.*, *22*(5), 806. https://doi.org/10.1038/CR.2012.30

Romei, C., & Elisei, R. (2021). A narrative review of genetic alterations in primary thyroid epithelial cancer. *International Journal of Molecular Sciences*, *22*(4). https://doi.org/10.3390/ijms22041726

RStudio Team. (2022). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Seyednasrollah, F., Laiho, A., & Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. https://doi.org/10.1093/bib/bbt086

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.*, *39*(5), 1–13. https://doi.org/10.18637/jss.v039.i05

Song, B., Tian, L., Zhang, F., Lin, Z., Gong, B., Liu, T., & Teng, W. (2022). A novel signature to predict thyroid cancer prognosis and immune landscape using immune-related LncRNA pairs. *BMC Med. Genomics*, *15*(1), 183. https://doi.org/10.1186/s12920-022-01332-7

Stupnikov, A., McInerney, C. E., Savage, K. I., McIntosh, S. A., Emmert-Streib, F., Kennedy, R., Salto-Tellez, M., Prise, K. M., & McArt, D. G. (2021). Robustness of differential gene expression analysis of RNA-seq. *Comput. Struct. Biotechnol. J.*, *19*, 3470–3481. https://doi.org/10.1016/j.csbj.2021.05.040

Sun, T., Wu, Z., Wang, X., Wang, Y., Hu, X., Qin, W., Lu, S., Xu, D., Wu, Y., Chen, Q., Ding, X., Guo, H., Li, Y., Wang, Y., Fu, B., Yao, W., Wei, M., & Wu, H. (2020). LNC942 promoting METTL14-mediated m6A methylation in breast cancer cell proliferation and progression. *Oncogene 2020 3931*, *39*(31), 5358–5372. https://doi.org/10.1038/s41388-020-1338-9

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.*, *71*(3), 209–249. https://doi.org/10.3322/CAAC.21660

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet. 2008 96*, *9*(6), 465–476. https://doi.org/10.1038/nrg2341

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, *43*(21), e140–e140. https://doi.org/10.1093/nar/gkv711

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Res.*, *21*(12), 2213–2223. https://doi.org/10.1101/gr.124321.111

Therneau T. (2021). A Package for Survival Analysis in R; Version 3.2-11. https://cran.r-project.org/package=survival

Tian, Y., Morris, T. J., Webster, A. P., Yang, Z., Beck, S., Feber, A., & Teschendorff, A. E. (2017). ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics*, *33*(24), 3982–3984. https://doi.org/10.1093/bioinformatics/btx513

Titus, A. J., Way, G. P., Johnson, K. C., & Christensen, B. C. (2017). Deconvolution of DNA methylation identifies differentially methylated gene regions on 1p36 across breast cancer subtypes. *Sci. Rep.*, *7*(1). https://doi.org/10.1038/S41598-017-10199-Z

Wajed, S. A., Laird, P. W., & DeMeester, T. R. (2001). DNA Methylation: An Alternative Pathway to Cancer. *Ann. Surg.*, *234*(1), 10. https://doi.org/10.1097/00000658-200107000-00003

Xing, X. L., Yao, Z. Y., Xing, C., Huang, Z., Peng, J., & Liu, Y. W. (2021). Gene expression and DNA methylation analyses suggest that two immune related genes are prognostic factors of colorectal cancer. *BMC Med. Genomics*, *14*(1), 1–8. https://doi.org/10.1186/s12920-021-00966-3

Yao, Y., Zhang, T., Qi, L., Liu, R., Liu, G., Wang, J., Song, Q., & Sun, C. (2020). Comprehensive analysis of prognostic biomarkers in lung adenocarcinoma based on aberrant lncRNA-miRNA-mRNA networks and Cox regression models. *Biosci. Rep.*, *40*(1), 1–13. https://doi.org/10.1042/BSR20191554

Yao, Z. Y., Xing, C. Q., Zhang, T., Liu, Y. W., & Xing, X. L. (2021). MicroRNA related prognosis biomarkers from high throughput sequencing data of Kidney renal papillary cell carcinoma. *Eur. Rev. Med. Pharmacol. Sci.*, *25*(5), 2235–2244. https://doi.org/10.26355/eurrev_202103_25255

7 Supplementary material

List of Abbreviations

- *(AUC) Area under the curve*
- *(DEG) Differentially expressed gene*
- *(DFS) Disease free survival*
- *(DGE) Differential gene expression*
- *(DMP) Differentially methylated probe*
- *(DMR) Differentially methylated region*
- *(FPKM) Fragments per kilobase million*
- *(lncRNA) Long non-coding RNA*
- *(LUAD) Lung adenocarcinoma*
- *(miRNA) Micro RNA*
- *(mRNA) Messenger RNA*
- *(OS) Overall survival*
- *(PFS) Progression free survival*
- *(STAR) Spliced Transcripts Alignment to a Reference*
- *(TCGA) The Cancer Genome Atlas*
- *(TPM) Transcripts Per Kilobase Million*

Supplementary tables

Confusion Matrix for diagnosis		
StatePred	NonTumor	Tumor
NonTumor	5	0
Tumor	0	115

Table S1: Confusion Matrix of logistic regression model for LUAD diagnosis

DMPs identified as diagnosis biomarkers	
CpG ID	Model Coefficient
cg06801994	5.61565074
cg11114465	3.40582677
cg08943107	2.21385330
cg15986164	3.46400976
cg03915012	0.61774993
cg11207652	4.99318092
cg11855693	5.26973393
cg00310375	-1.50359358
cg15488143	1.21026221
cg15948504	3.07261341
cg12589486	3.01130232
cg18613031	1.25748402
cg01943585	0.36457853
cg03292206	0.99791131
cg09010802	0.26247728
cg04864807	1.63852091
cg00615919	0.69863243
cg09848693	0.89596988
cg22827250	0.31796098
cg14410016	0.84055492
cg18910243	2.85169987
cg10592926	0.89364256
cg14156751	0.93279806
cg00754604	0.24500921
cg00224911	-4.91619173
cg07300534	0.17074527
cg09765089	0.23087391
cg06821075	5.20145880
cg03190219	0.15414409
cg02671826	0.29669343
cg03622664	0.32819214
cg23627350	5.76536493
cg08173709	0.81313575
cg13775533	2.85729975
cg12625139	0.74978068
cg21842916	1.57501634
cg08384322	0.01185672
cg00485849	-1.92416107
cg00530870	3.40624697
cg22698604	-0.30344842
cg15073624	-11.62449822
cg16844053	0.42912638
cg00035945	-6.70401118
cg05348535	-1.45258555
cg11653410	-7.28021445
cg23898701	0.01237718
cg19587168	-0.55864715

Table S2: DMPs identified as diagnosis biomarkers

OS grouped by age			
	Number of patients	Deceased	Mean OS time (months)
< 60	122	32	117.18
> 60	304	73	129.55

Table S3: Distribution of LUAD patients according to overall survival and age

OS grouped by treatment			
	Number of patients	Deceased	Mean OS time (months)
Non treated	104	24	110.56
Treated	30	6	136.39

Table S4: Distribution of LUAD patients according to overall survival and treatment

DFS grouped by age			
	Number of patients	Recurrence	Mean DFS time (months)
< 60	122	44	144.81
> 60	304	139	170.91

Table S5: Distribution of LUAD patients according to disease free survival and age

DFS grouped by treatment			
	Number of patients	Recurrence	Mean DFS time (months)
Non treated	104	37	137.70
Treated	30	14	153.84

Table S6: Distribution of LUAD patients according to disease free survival and treatment

DMPs identified as DFS prognosis biomarkers	
CpG ID	Pr(> |z|)
cg04377850	< 2e-16
cg22916646	0.345398
cg16474118	< 2e-16
cg04181528	< 2e-16
cg14519917	0.111815
cg06606207	0.052782
cg11667020	< 2e-16
cg27158867	9.01e-13
cg04737087	0.145695
cg03649796	< 2e-16
cg23844904	< 2e-16
cg00472277	9.12e-06
cg09088580	0.656746
cg07868273	0.000636
cg15020568	0.035269
cg14484885	0.029470
cg07105272	0.014318
cg11174998	< 2e-16
cg12174804	3.76e-09
cg15844611	0.925243
cg16754857	< 2e-16
cg17149019	< 2e-16
cg05065690	3.04e-09
cg10919522	0.045425
cg10604851	1.81e-07
cg00937392	< 2e-16
cg08955358	< 2e-16
cg25024442	1.12e-15
cg20053110	5.34e-09
cg15110403	5.78e-08
cg04793813	< 2e-16
cg22237937	0.005714
cg12062198	1.53e-05
cg22812684	< 2e-16
cg05395321	0.226504
cg08778805	< 2e-16
cg04617905	< 2e-16
cg06732395	0.075979
cg13458384	< 2e-16
cg00402068	< 2e-16
cg25958158	< 2e-16
cg14270890	6.84e-09
cg16516691	1.17e-10
cg18846883	< 2e-16
cg02809409	0.000332
cg19084629	< 2e-16
cg23060872	< 2e-16
cg02375178	< 2e-16
cg15989436	1.05e-11

cg12298212	< 2e-16
cg15411984	1.03e-05
cg25246084	2.70e-10
cg05412222	< 2e-16
cg02980621	0.023184
cg01613870	0.329289
cg24892571	< 2e-16
cg19476368	0.028925
cg14076971	< 2e-16
cg07115035	< 2e-16
cg16311740	0.898153
cg11467381	< 2e-16
cg03271150	< 2e-16
cg22335076	0.230125
cg12911763	< 2e-16
cg01939980	0.085581
cg18582180	9.51e-05
cg01065938	< 2e-16
cg11955768	< 2e-16
cg23924137	4.61e-06
cg27216937	< 2e-16
cg08826066	< 2e-16
cg24360651	< 2e-16
cg25683012	0.247590
cg15193782	1.00e-09
cg06935979	< 2e-16
cg15671725	0.075112
cg06056929	0.062398
cg13455439	0.153922
cg17512922	1.07e-05
cg08707123	0.000133
cg02827328	1.15e-07
cg24085436	6.35e-08
cg25024143	0.450292
cg01844866	4.03e-06
cg04573845	< 2e-16
cg02349096	< 2e-16
cg19814116	< 2e-16
cg14722140	< 2e-16
cg22307471	< 2e-16
cg16460157	< 2e-16
cg19151030	< 2e-16
cg09023869	< 2e-16
cg09630479	< 2e-16
cg18319799	0.004689

Table S7: DMPs identified as DFS prognosis biomarkers