VNiVERSiDAD
D SALAMANCA

INSTITUTO UNIVERSITARIO DE
BIOLOGIA MOLECULAR Y CELULAR DEL CANCER (IBMCC)

CENTRO DE INVESTIGACION DEL CANCER (CiC-IBMCC)

DOCTORADO EN BIOCIENCIAS:
BIOLOGIA Y CLINICA DEL CANCER Y MEDICINA TRASLACIONAL

Tesis Doctoral

Development and application of bioinformatic methods to analyze genome-wide expression and survival data from cancer patients

Santiago Bueno Fortes

Dr. Javier De Las Rivas
Director

Dr. Manuel Martín-Merino Acera
Codirector

Salamanca, July 2019

Dr. Javier De Las Rivas Sanz, con D.N.I. 15949000H, Investigador Científico del Consejo Superior de Investigaciones Científicas (CSIC), director del grupo de Bioinformática y Genómica Funcional en el Centro de Investigación del Cáncer (CiC-IBMCC), y profesor del Programa de Doctorado y del Máster de Biología y Clínica del Cáncer de dicho Centro y la Universidad de Salamanca (USAL).

Y el **Dr. Manuel Martín-Merino Acera**, con D.N.I. 07967144J, Catedrático de la Facultad de Informática, Universidad Pontificia de Salamanca (UPSA).

CERTIFICAN

que han dirigido la Tesis Doctoral titulada: **"Development and application of bioinformatic methods to analyze genome-wide expression and survival data from cancer patients / Desarrollo y aplicación de métodos bioinformáticos al análisis de datos de expresión genómica y datos de supervivencia de pacientes con cáncer"**; realizada por **D. Santiago Bueno Fortes**, dentro del programa de doctorado Biociencias: Biología y Clínica del Cáncer y Medicina Traslacional del Centro de Investigación del Cáncer (CiC-IBMCC, CSIC/USAL).

Y AUTORIZAN

la presentación de la misma, considerando que reúne las condiciones de originalidad y contenidos requeridos para optar al grado de Doctor por la Universidad de Salamanca.

En Salamanca, a 10 de Julio de 2019

Dr. Javier De Las Rivas
Director

Dr. Manuel Martín-Merino Acera
Codirector

Para la realización de esta Tesis Doctoral, el doctorando Santiago Bueno Fortes obtuvo en concurso público una Ayuda destinada a financiar la Contratación Predoctoral de Personal Investigador, cofinanciadas por el Fondo Social Europeo (FSE) y convocadas por la Junta de Castilla y León (Orden EDU/310/2015, de 10 de abril de 2015).

Contents

List of Figures

List of Tables

Abbreviations

AUC	Area Under Curve
BRCA	Breast Cancer
CMS	Consensus Molecular Subtypes (CRC subtypes)
CPM	Counts Per Million
CRC	ColoRectal Cancer
DECO	Decomposing heterogeneous Cohorts by Omic data profiling method.
EMT	Epithelial–mesenchymal transition
ER+	ER positive breast cancer subtype
FDR	False discovery rate
FPR	False positive rate
FPKM	Fragments Per Kilobase per Million mapped reads
GEO	Gene Expression Omnibus database
GEP	Genome-wide expression profiling
HR++	Hormone Receptor positive ER+ and PR+
HER2+	HER2 positive breast cancer subtype
IHC	Immunohistochemistry
IQR	Interquartile range
KM	Kaplan-Meier
KO	Knock Out
mRNA	Messenger RNA
NUSE	Normalised Unscaled Standard Errors
OR	Odds ratio
OS	Overall survival
PCA	Principal Component Analysis
PR+	PR positive breast cancer subtype
RFS	Relapse Free Survival
RMA	Robust Multiarray Average normalisation method
RNA	Ribonucleic acid
RNAseq	RNA sequencing
RPKM	Read Per Kilobase per Million mapped reads
ROC	Receiving operative curve
SAM	Significance Analysis of Microarrays
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
TMM	Weighted trimmed mean of M-values
TNBC	Triple Negative Breast Cancer
TPM	Transcripts Per Million
TPR	True positive rate

Chapter 1

Introduction, hypothesis and objectives

1.1 General introduction

The development of robust omic technologies (genomics, transcriptomics, proteomics, etc) to generate and understand genome-wide alterations is already having an impact on health care, with a particular relevance on cancer and oncology. Within the current context of Personalised Medicine, Precision Medicine and Genomic Medicine (Roden and Tyndale, 2013), modern cancer research has to be done considering an adecuate use of the large-scale data derived from these new omic technologies. Some of these technologies, such as transcriptomic expression profiling, have been already applied to thousands of human samples (see public database GEO (NCBI, 2019)), and provide information about the expression status of all the known genes in the analised individuals. In order to be useful and applicable to medical research, such omic data should be integrated with the corresponding clinical data using adequate computational and bioinformatic tools and methods. This is a main framework where the current Doctoral Thesis work is proposed.

Incidence of cancer in Europe

With respect to the study in the area of cancer, the work in this PhD is done with 2 major types of cancer: Breast Cancer (BRCA) and ColoRectal Cancer (CRC). These cancer types are nowdays the most frequent in Europe, representing together the largest proportion of all cancers (https://ecis.jrc.ec.europa.eu/). In particular in Europe in 2018 the most common cancer types, according to the body location, were cancers of the female breast (523,000 cases), followed by colorectal (500,000), lung (470,000) and prostate cancer (450,000). The global numbers for Europe in 2018 estimated 3.91 million new cases of cancer and 1.93 million deaths from cancer (Ferlay et al., 2018). Since the European population is close to 513 millions, having each year about 4 million new cases and about 2 million deaths, cancer represents the second most important cause of death and morbidity in Europe.

Considering just the specific numbers for breast cancer and colorectal cancer, the estimated incidence of these cancers in Europe in 2018 by country is presented

in Figure 1.1, that shows the map of Europe coloured according to the level of such incidences in each country. It is quite remarkable that breast cancer has a higher impact in Holland and Belgium, and in some north countries like Sweden, Finland, Denmark, United Kingdom and Ireland. By contrast, colorectal cancer has a quite higher impact in some specific countries that are: Norway and Hungary.

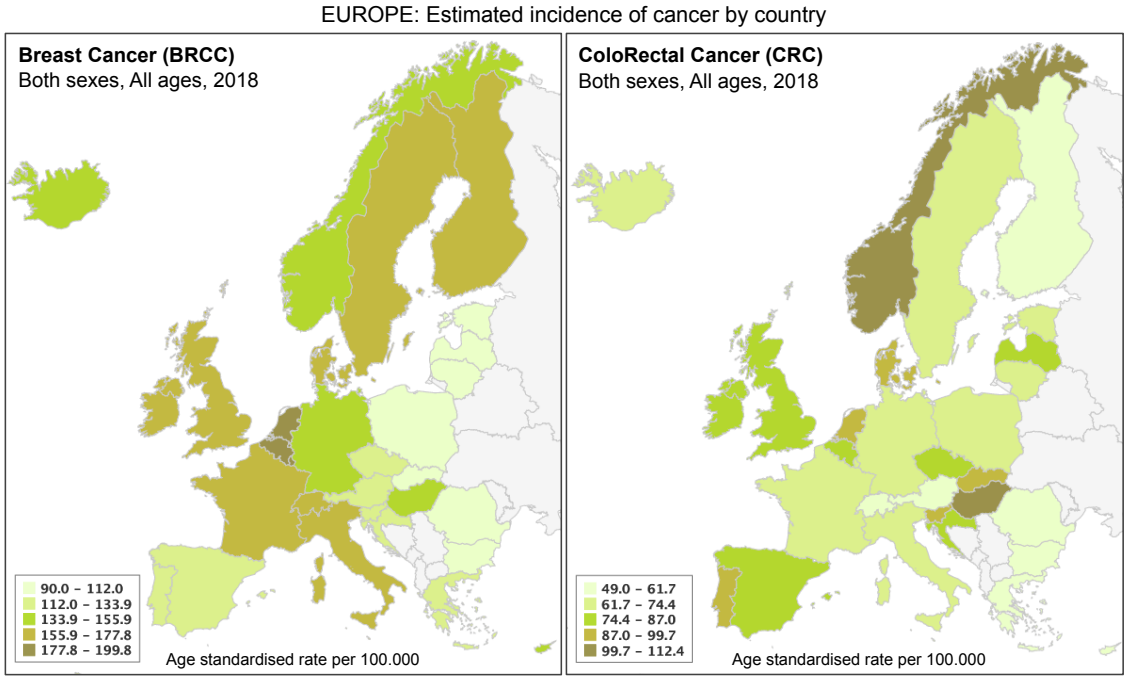


Figure 1.1: Estimated incidence of breast cancer and colorectal cancer in Europe by country (source: European Cancer Information System, ecis.jrc.ec.europa.eu).

Cancer a genomic disease driven by mutations

As indicated above, current cancer research is very concern about the value and power of omic technologies applied to the advance of medical and clinical oncology. The application of genome-wide omic technologies to the study of cancer over last 20 years has generated a new understanding of this complex disease that can not longer be called a "genetic disease", since it is more properly a "genomic disease". In fact, over the past two decades, comprehensive sequencing efforts have revealed the genomic landscapes of common forms of human cancer (Gerlinger et al., 2012). These studies have revealed about 140 human genes that, when altered by intragenic mutations, can promote or “drive” tumourigenesis and cell malignancy. A typical tumour contains two to ten of these “driver gene” mutations; the remaining mutations are passengers that confer no selective growth advantage. In common solid tumours (such as those derived from the colon, breast, lung, brain, or pancreas), an average of 25 to 75 genes display subtle somatic mutations that would be expected to alter their protein products (Gerlinger et al., 2012).

Figure 1.2 presents schematically the complexity of cancer that can affect to many different cell types, tissues types and organs in the human body from children

to adults. This complexity is due not only to all the different types of cancers originated in different locations of the body, but also due to the large number of somatic mutations that have been found thanks to the genome-wide scale analyses of all the human genes in many thousands of samples of tumours (Gerlinger et al., 2012). The results of these large-scale analyses on tumour mutations have been included in a reference data portal called COSMIC (the Catalogue Of Somatic Mutations In Cancer), that is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer, and can be accessed freely at: cancer.sanger.ac.uk/cosmic.

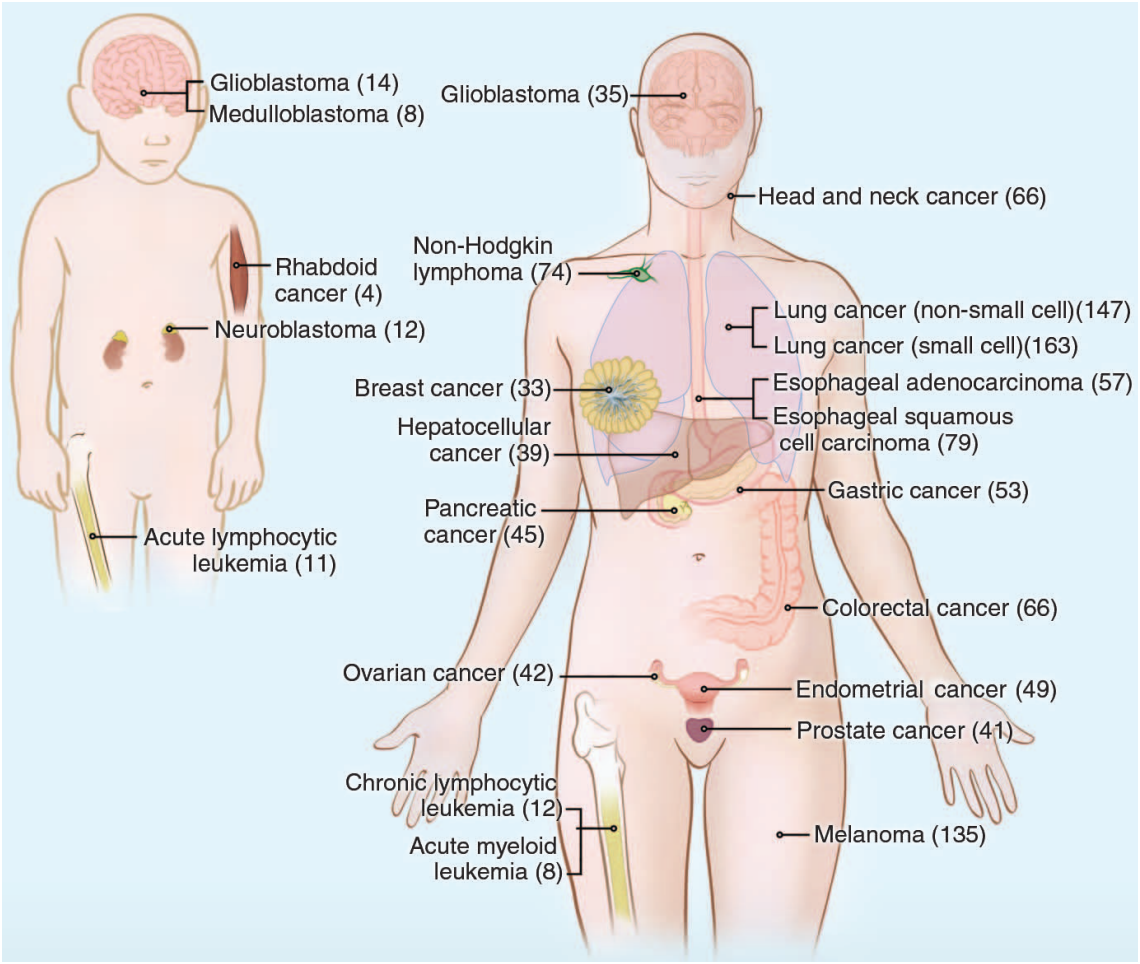


Figure 1.2: Average number of somatic mutations in a representative collection of human cancers, detected by genome-wide DNA sequencing studies. The genomes of a diverse group of adult (right) and pediatric (left) cancers have been analysed. The numbers in parentheses indicate the median number of non-synonymous mutations per tumour. (source: Vogelstein et al. 2013 *Science*, www.sciencemag.org).

Cancer heterogeneity: a challenge for the genomic era

Cancer is a heterogeneous disease with unique genetic and phenotypic features that differ between individual patients and even among individual tumour regions (Dagogo-Jack and Shaw, 2018). The observation of individual heterogeneity in can-

cer has been many times described, but a breakthrought was achieved when intratumour heterogeneity and branched evolutionary tumour growth was proven using the omic technology of whole-exome multiregion spatial DNA sequencing (Gerlinger et al., 2012). This "intratumour heterogeneity" added a new level of complexity to the study of cancer, which already had a complex nature due to the many possible tissue origins of the tumours (Figure 1.2). Moreover, cancer is a very dynamic disease where tumour cells proliferate and evolve over time even within the same location; so "temporal heterogeneity" should be added to "spatial heterogeneity" (Dagogo-Jack and Shaw, 2018).

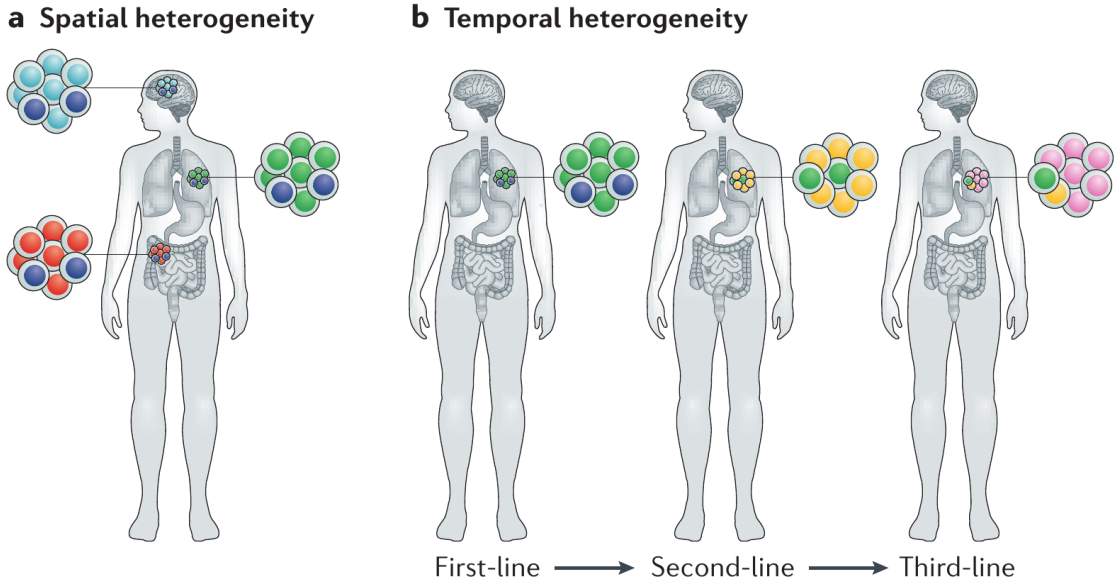


Figure 1.3: Spatial heterogeneity **(a)**: uneven distribution of cancer subclones across different regions of the primary tumour or the metastatic sites. Temporal heterogeneity **(b)**: variations in the molecular makeup of a single lesion over time, either as a result of natural progression of the tumour or as a result of exposure to selective pressures created by clinical interventions. (source: Dagogo-Jack et al. 2018 *Nat Rev Clin Oncol*, www.nature.com/nrclinonc).

Combination of omic data plus survival data, key way for beating cancer

All the levels of tumour complexity and heterogeneity described above make it very difficult to identify stable biomarkers for the different types and subtypes of cancer. It is clear that some genes (such as the top-10 genes detected as the most mutated in COSMIC, which are: BRAF, JAK2, KRAS, TP53, EGRF, FLT3, PIK3CA, TERT, IDH1 and KIT) are well known cancer genes, drivers of many specific tumours (Tate et al., 2019). However, COSMIC database and the Cancer Genome Atlas (TCGA) project revealed that about 40% of the studied tumours do not have any mutation in any of the 576 humas genes included in the current Cancer Gene Census (Sondka et al., 2018).

Our work in this Doctoral Thesis started addressing this research problem by using the power of new omic technologies (mainly transcriptomic profiling of tumour

samples from cancer patients), combined with useful medical data about those patients (i.e. accessible medical data that can provide information about the actual clinical status and the evolution of patients). The most useful clinical information directly related to patient prognosis and disease outcome is the "survival" information (either the disease-free survival, DFS, the relapse-free survival, RFS, or the overall survival, OS). In this way, the combination of large-scale gene expression data from tumour samples with survival data from the same patients can be a very powerful strategy to improve the indentification and discovery of new biomarkers for specific cancer subtypes. This is the main approach explored and investigated in the present work. This approach could only be adequately addressed using powerful computational and bioinformatic tools and methods.

Oncogenic driver pathways: a change of paradigm

Biomedical cancer research and clinical oncology have been working so far on the assumption that the biological nature of the different types and subtypes of cancer are primarily defined by the tissue and organ origin of the tumours. This approach have provided the main cancer classification: breast cancer, colorectal cancer, lung cancer, liver cancer, etc.

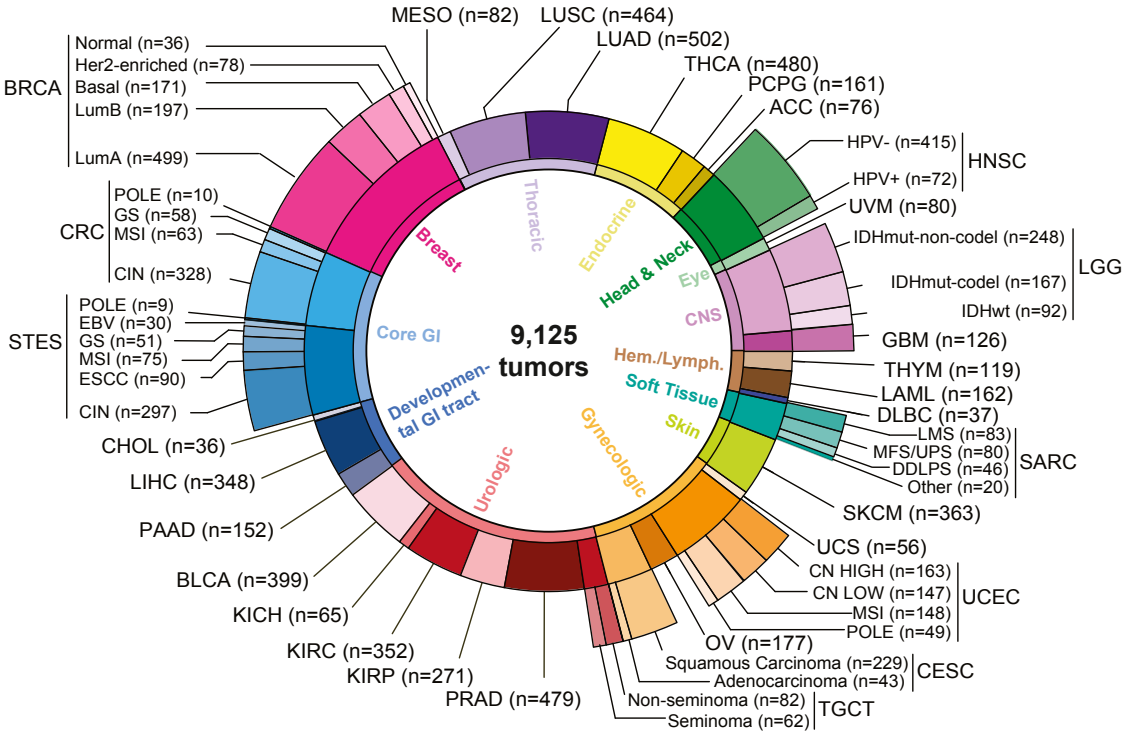


Figure 1.4: Distribution of cancer types in The Cancer Genome Atlas cohort, including molecular subtypes analysed. TCGA pancancer atlas contains 9,125 tumour samples. (source: Sánchez-Vega et al. 2018 *Cell*, www.cell.com).

However, the results of the worldwide effort done by the application of genome-wide omic technologies to the study of many types and subtypes of cancer (for example in The Cancer Genome Atlas, TCGA, project) (Figure 1.4) are providing a new

deeper molecular understanding of the cancer biology, suggesting that the nature of tumours is best explained when they are associated with specific molecular gene signatures and to specific biological pathways, instead of the classical association with the "cell-of-origin" (Hoadley et al., 2018).

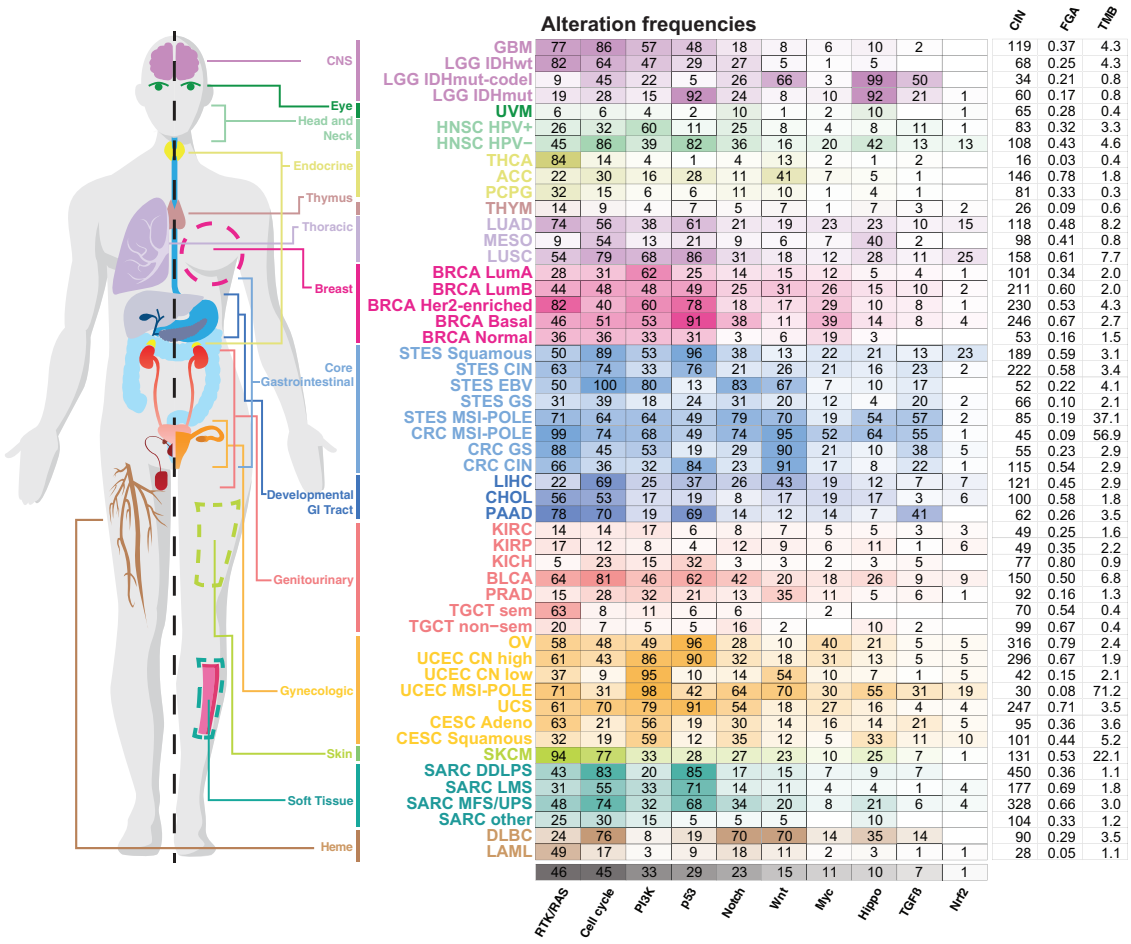


Figure 1.5: Fraction of altered samples of the TCGA cohort per pathway and tumour subtype. Pathways are ordered by decreasing median frequency of alterations. Increasing colour intensities reflect higher percentages. Average mutation count, as well as number of unbalanced segments and fraction genome altered (two measures of the degree of copy-number alterations) per cancer subtype are also provided. (source: Sánchez-Vega et al. 2018 *Cell*, www.cell.com).

Using somatic mutations, copy-number alterations, mRNA expression changes and DNA methylation modifications detected in 9,125 tumours (profiled by TCGA), Nik Schultz, Chris Sander and collaborators (Sanchez-Vega et al., 2018) analysed the mechanisms and patterns of somatic alterations, identifying ten canonical pathways as the ones that hold the major part those alterations: **(1)** cell cycle, **(2)** Hippo signaling, **(3)** MYC signaling, **(4)** NOTCH signaling, **(5)** oxidative stress response/NRF2, **(6)** PI-3-Kinase signaling, **(7)** receptor-tyrosine kinase (RTK)/RAS/MAP-Kinase signaling, **(8)** TGF-β signaling, **(9)** TP53 and **(10)** beta-catenin/WNT signaling. These ten pathways contain the biological processes that can explain most

of the malignant alterations observed in the cancer cells (Figure 1.5).

This paradigm sift in cancer has strong support by the demonstration that genetic alterations in signaling pathways that control cell-cycle progression, apoptosis and cell growth are common hallmarks of cancer. The extent, mechanisms, and co-occurrence of these alterations in these pathways differ between individual tumours and tumour types, but are the common molecular features that allow nowadays the identification of similar cancer types independent of the tissue or organ of origin. In this way, comprehensive, integrated molecular analysis of tumour samples can identify molecular relationships across a large diverse set of human cancers, suggesting future directions for exploring clinical actionability in cancer treatment independent of the "cell-of-origin" of the tumours (Sanchez-Vega et al., 2018).

Finding new biomarkers: robust data analysis in the era of big data

The amount of omic data generated over recent years in the study of cancer, using patient-derived samples, is huge. This is not only due to the large-scale international projects, like TCGA and the International Cancer Genome Consortium (ICGC); but also due to several efforts initiated by different leading countries around the world during the last decade (Hudson et al., 2010). Figure 1.6 presents the numbers about the amount of data included in TCGA project.

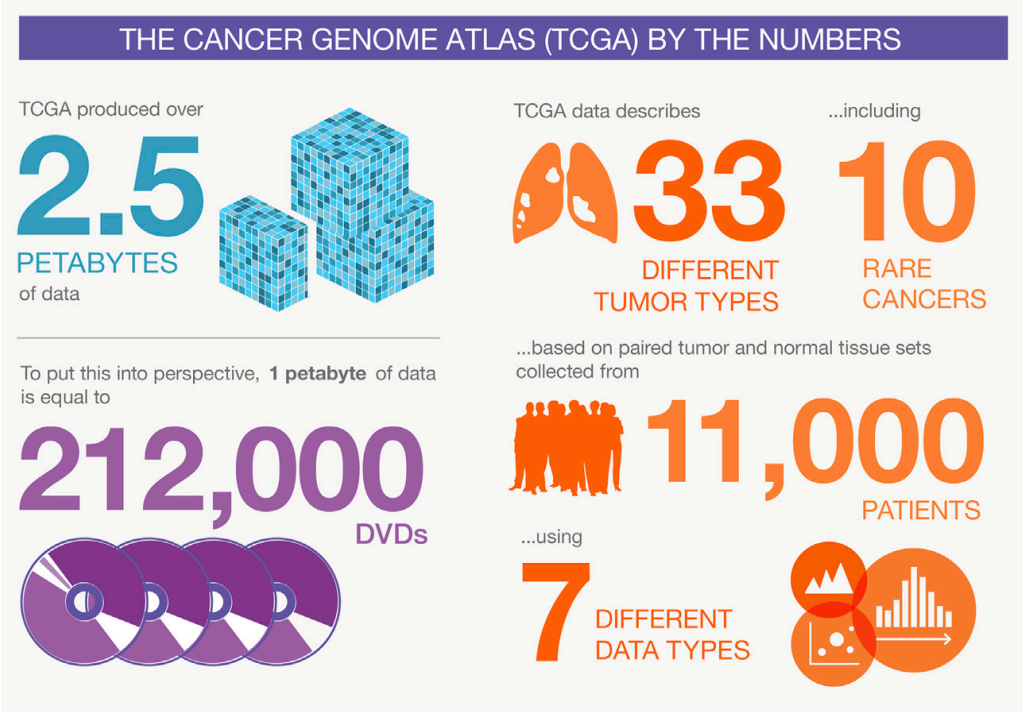


Figure 1.6: Information about the amount of data included in TCGA. (source: Hutter and Zenklusen 2018 *Cell*, www.cell.com).

TCGA began in 2006 as a pilot project focused on 3 cancer types (lung, ovarian, and glioblastoma), but due to the success of the initial efforts, it was reauthorised for a full production phase in 2009. In the following decade, TCGA collected more than

11,000 cases across 33 tumour types and generated a vast, comprehensive dataset describing the molecular changes that occur in cancer (Figure 1.6).

In this scenario, a key challenge for modern cancer research and clinical oncology is to use and apply smart robust bioinformatic methods to analyze the "big data" produced, in order to find new biomarkers for many tumours that are not yet well characterised and, in this way, create a lasting value beyond mere data.

The development and application of computaional and bioinformatic methods applied to genomic and clinical data derived from samples of cancer patients is a key strategy to find new "**actionable molecular biomarkers**". This approach is the main one proposed and followed in this work, that has a clear "**patient-centric**" scope and it is applied to the two most frequent malignant pathologies: breast cancer and colorectal cancer (Figure 1.1). The type of clinical data that can be better correlated with the prognosis of the disease and the outcome of the patients are the survival data. When we say "**survival**", we refer not only to the time to the death of the patients, but to the time that measures different critical clinical events, such as: the response to treatment, the relapse in the disease, the appearence of resistance to the drugs or the appearance of metastasis. The combination of the survival data with omic data is not trivial, because often there are not enough samples to achieve a solid inference of the biomolecular features (i.e. genes, gene alterations, proteins, etc) that mark a better or worse survival. Also because the scale of omic data (which provides information on thousands of putative markers) requires the use of multivariate statistical analysis and robust cross-validation strategies.

1.2 Hypothesis

The work presented in this PhD is centered in the field of **Bioinformatics and Computational Biology** applied to **Cancer Research**, with a particular focus on the analysis and integration of omic data and clinical data to improve the discovery and identification of new molecular biomarkers related to the prognosis of the cancer patients.

In particular, our main hypothesis to start and develop our research was the following: "**We consider that a thorough and detailed analysis of the survival data of cancer patients combined with the transcriptomic data derived from the tumour biopses of such patients should be a very powerful and meaningful way to discover new genetic biomarkers directly related to the nature and prognosis of each patient's specific cancer**".

To prove and develop this hypothesis, we worked in this PhD with samples from two major types of cancer: **Breast Cancer** (**BRCA**, i.e. invasive breast carcinoma) and **ColoRectal Cancer** (**CRC**). These cancer types are nowdays the most frequent in Europe (https://ecis.jrc.ec.europa.eu/), representing together the largest proportion of all cancers.

A first critical challange to carry out this work was to collect and integrate into uniform data sets a **large number of cancer samples** (i.e., more than a thousand) that included **genome-wide expression and survival data**. We say this because, most of the survival analyses that we found in the literature were restricted to smaller data sets (that is, to hundreds of samples, or even less). Many researchers in the field do not realize that the statistical power and significance of all the survival algorithms and methods depend, critically, on the number of samples studied together.

A final point with respect to the hypotesis, is that we did not propose just the application of standard computational methods for survival analysis, but we wanted to develop and apply new bioinformatic algorithms to do so. In particular, in order to improve the way survival data can be integrated with genome-wide tested expression data to discover new **gene survival markers**.

1.3 Objectives

Once we have described the main hypothesis of our Doctoral Thesis, we must present the objectives that describe in a more tangible way the particular work carried out and the specific challenges we faced during the four years of our doctorate. The objectives are divided into two main groups: **(i)** the first two objectives (1st and 2nd) correspond to work done with **Breast Cancer** data; and **(ii)** the second group of objectives (3rd and 4th) correspond to work done with **ColoRectal Cancer** data. These **four objectives** are presented in this disertation as **four separated chapters** following this one.

The OBJECTIVES:

Objective 1.- Generation of a large homogeneous data set of Breast Cancer (BRCA) samples that include genome-wide expression data and patient survival data; and discovery of BRCA survival markers associated with the three currently standard clinical markers (ER, PR and HER2) through the development and application of robust algorithms for survival analysis based on transcriptomic profiling.

Objective 2.- Unravel and discovery of positive gene markers and regulators of Triple Negative Breast Cancer (TNBC) using transcriptomic and gene regulatory profiling combined with survival analysis. Study done by comparison and contrast of TNBC with the most frequent subtype of breast cancer, that is luminal BRCA.

Objective 3.- Generation of a large homogeneous data set of ColoRectal Cancer (CRC) samples that include genome-wide expression data and patient survival data; and discovery of new CRC survival marker genes derived from a robust integration and meta-analysis of multiple transcriptomic data sets.

Objective 4.- Integrative analyses of the transcriptomic profiles of multiple Colorectal Cancer (CRC) samples in order to identify and characterize the four Consensus Molecular Subtypes (CMS1, 2, 3, 4); integration of this transcriptomic data with the survival data of patients; and relative characterisation of an EMT gene signature associated to P21 knockout (i.e. CDKN1A KO gene) obatined in a human cell-line.

Chapter 2

Discovery of Breast Cancer (BRCA) survival markers associated to standard clinical markers and robust algorithms for survival analysis based on transcriptomic profiling

2.1 Motivation

Breast cancer treatment is determined by a standard categorisation of tumours in four groups. The classification is carried out considering mainly, three clinical markers: ER (ESR1), PR (PGR), and, HER2 (Saini et al., 2011) (ERBB2 or NEU) obtained by inmunohistochemistry (IHC). The markers define the subclases; Luminal A, Luminal B, HER2 enriched, and triple negative (TNBC). Some complementary markers such as AURKA or MKI67 are recently being considered to improve the risk prediction.

However, errors in the estimation of standard clinical markers are more extensive than expected (Li et al., 2010). This error may lead to the wrong treatment of the patient. Besides, the groups obtained by only three markers are frequently too heterogeneous (Venet et al., 2011) (Bartlett et al., 2016) (Mertins et al., 2016). The identification of genes related to the clinical markers may help to improve the stratification and treatment of patients providing new terapeutic targets.

Several comercial platforms that consider a multivariate gene signature have been developed. However, the overlaping among the gene signatures and the risk groups is low (Venet et al., 2011) (Bartlett et al., 2016) (Mertins et al., 2016). The platform works as a black box and the decisions can not be interpreted in terms of standard clinical markers. This prevents the application in clinical practice. Investigations focusing on the influence of feature selection method on performance and stability of the signature are lacking (Haury et al., 2011).

Moreover, a large number of prognostic gene signatures have been proposed in the literature. The consensus among them is quite small and frequently they are sample dependent (Mertins et al., 2016). That is, the algorithm retrieves a different subset of genes regarding the dataset considered. Some authors such as (Ein-Dor et al., 2005) have studied several gene signatures, and they have concluded that the stability, reproducibility and robustness remains a challenging problem. To overcome this, a validation in an independent RNAseq series is another objective.

In this chapter a robust multivariate gene signature is obtained that is interpretable in terms of the clinical markers. This gene signature provides alternative targets to develop new treatments. Besides, it will allow to estimate the status of clinical markers with smaller error and to improve the stratification of patients according to their risk.

To achive this goal, a dataset with large number of samples is built by integration of different studies. Robust algorithms are developed in order to identify stable markers for the whole population.

Several new robust strategies have been developed in order to improve four main steps: normalisation, differential expression, feature selection, and univariate or multivariate prediction models.

2.2 Material and Methods

In this section, we introduce several robust algorithms to identify survival gene markers related to the standard clinical markers in breast cancer. Several methods to improve the risk prediction and the patient stratification are presented. To ensure the generability and robustness of the biological findings, a standard method was modified to integrate an extensive collection of samples coming from different studies.

2.2.1 Datasets description

Breast cancer survival datasets

The first two datasets are based on microarray technology. We have considered microarrays because it is a widely studied technology and there are a large number of datasets available that can be integrated. This chapter is devoted to the discovery of prognostic genes associated to standard clinical markers. Therefore, all the series considered, should contain the following meta-data: (i) Survival time. (ii) Status (if the data is censored or not at the end of patient's follow up). (iii)IHC measurement if possible, for the primary markers of BRCA: ER, PR, and HER2.

These series were obtained mainly from the Gene Expression Omnibus (GEO) (NCBI, 2019), using the GEO search tools such as the getGEO function from **GEOquery** package (Davis and Meltzer, 2007) function for R.

All the studies integrated in each dataset should comply with the following criteria: (i) GeneChip: Affymetrix Human Genome U133a and hgu133plus2, Plus 2.0 Array annotation data. (ii) Both platforms (plus2 and 133a) may be used to carry out independent studies.

The series selected and filtered from publicly available databases are defined in the following table, Tab: 2.1.

GEO ID	Orig N	Final N	Surv Type	PMID	Year	Journal	Description
GSE6532	87	87	RFS, DMFS	17401012	2007	J Clin Oncol	Molecular subtypes in estrogen receptor positive breast carcinomas.
GSE12276	204	204	MFS	19421193	2009	Nature	Genes that mediate breast cancer metastasis to the brain
GSE19615	115	115	RFS, MFS	20098429	2010	Nat Med	Chemotherapy resistance and recurrence of BRCA.
GSE17907	55	39	MFS	20932292	2010	BMC Cancer	Genome profiling of ERBB2-amplified breast cancers
GSE20685	327	327	OS, MFS	21501481	2011	BMC Cancer	BRCA molecular subtypes and clinical outcomes: treatment optimisation.
GSE21653	266	252	DFS	20490655	2011	BRCA Res Treat	A gene expression signature identifies two prognostic subgroups of basal BRCA
TOTAL	1054	1024					

Table 2.1: Compilation of BRCA microarray series, sources, number of samples and, description.

We have a total of 1024 samples from Plus2, a subset of 380 with IHC values for ER, PR, and HER2. For the other 644 samples, the IHC values are missing or incomplete, but all have survival time and status value.

The RNAseq dataset here is in whole or part based upon data generated by the TCGA Research Network (TCGA, 2019).

This dataset and phenodata are provided to the user by **RCurl** (Lang and the CRAN team, 2019), **curatedTCGAData** (Ramos, 2019), and **TCGAutils** (Ramos et al., 2019) R packages. The packages allow us to access the raw counts, the FPKM or RPKM matrix and all the clinical information available. RPKM (Reads Per Kilobase Million) and FPKM (Fragments Per Kilobase Million) are very similar, but there are differences. RPKM was made for single-end RNA-seq, every read corresponded to a single sequenced fragment. FPKM was made for paired-end RNA-seq, two reads can correspond to a single fragment, or one if one of the reads fail at mapping. So the main difference is that FPKM takes into account that two reads may map the same fragment. We chose the one recommended by TCGA (in each series).

Only the samples that contain the following meta-data are considered: (i) Survival time. (ii) Status (if the data is censored or not at the end of patient's follow up). (iii) IHC measurement in as many patients as possible for the main markers of BRCA: ER, PR, and HER2.

Once this compilation of 879 samples was done, the normalisation process started. This is explained in the next section.

Ensembl (Flicek et al., 2014) have developed tools and data resources to facilitate genomic analysis in chordate species with an emphasis on human. The Ensembl identifiers are choosen to be used in our studies. Among the high amount of ways to identify a gen (which is a big problem per se) Ensembl identifiers were selected, which are the less ambiguous.

2.2.2 Normalisation methods

The integration of heterogeneous series obtained under different protocols, conditions, and laboratories provides a gene expression signal not directly comparable for different studies. In this section, a normalisation method is applied in order to obtain a signal that is comparable for samples from different studies.

Normalisation methods for microarray gene expression data

First, the "cel" datasets downloaded from GEO are loaded in R, using the tools provided in their web and GEO's Bioconductor packages in this case, the **ReadAffy** function from affy package (Gautier et al., 2004).

The first issue addressed is the bath problem which arises when samples come from different series. Then, gene signals for different samples are considered similar just because they belong to the same study/batch.

Several algorithms have been proposed in the literature to remove the batch effect. In particular, **RMA + COMBAT** (from **inSilicoMerging** R package) (Kupfer et al., 2012) has been widely applied with encouraging results. However, for our problem, a personalised batch effect correction is needed. The fact that the sample sizes differ between series, the conditions of the experiments, and the year of the study could lead to a stronger batch effect.

In (McCall et al., 2010a) the **fRMA** algorithm developed by Irizarry and his group from Johns Hopkins University is introduced. This algorithm outperforms other approaches but should be adapted for our problem. We focused on the array quality metrics, and in the creation of our own vector of weights to be used by fRMA. To normalise our data, the vector will give different weights to modify the expression values for each dataset in order to make different datasets comparable. Irizarry's team provide some example vectors, generated through compilation of a high amount of data. These vectors can be used to normalise the datasets if they meet certain requeriments. However, our dataset does not comply with the previous requeriment.

Since the function of the normalisation vector is to provide a metric for standarisation of the batches, a factor will be computed for each batch. This factor, will be taken into account when performing the normalisation process, multiplying the expression value in each sample by the factor. The gene signal obtained is then comparable for the different studies integrated.

We designed diferent approaches, changing the number of subsets for each series, and the number of samples included in the subset provided to generate the fRMA vector. The aim is to make the series comparables and to avoid overfitted expression distributions which may destroy the signal in our data. We made several tests and then selected a randomised sample size of 5 for each mini-subset in order to create the vector. The number of mini-subsets done for each series varied depending on the number of samples as explained in the Tab: 2.2. The resulting vector will allow not only to normalise data but also to add individual samples later.

Series	Samples	Number of batches
AffyBatchEsetGSE6532	87	1
AffyBatchEsetGSE6532	204	2
AffyBatchEsetGSE19615	115	1
AffyBatchEsetGSE19615	39	1
AffyBatchEsetGSE20685	327	3
AffyBatchEsetGSE21653	252	3

Table 2.2: The fRMA batches size and samples, 5 for each new batch.

Another algorithm that has been sucessfully applied to avoid the batch effect is ComBat. This algorithm considers an Empirical Bayes method to adjust for potential batch effects. Empirical Bayesian ComBat is an approach that assumes that the batch effect may affect many genes in similar ways, and thus the algorithm adjusts for these batch biases which are standard across genes.

ComBat differs from previous methods in its ability to adjust data whose batch sizes are small. There are two estimation methods, a parametric one which computes prior probability distributions and a non-parametric one which makes no prior assumptions. The second one is more dependent on computing time.

ComBat should be used in data that has been already preprocessed and normalised gene by gene such that they have similar mean and variance. It also includes covariates in analysis when possible. The proportion of treatment/control samples also may derivate in the removal of the biological signal.

Quality control metrics

Normalised Unscaled Standard Errors (NUSE) can be used for assessing the quality of the gene signal. The gNUSE function estimates the standard error for each gene in each array using RLE (Relative Log Expression). RLE method does a log scale estimates of expression $\hat{\theta}_{gi}$ for each gene g on each array i, then computes median across arrays for each gene m_g and defines the relative expression as $M_{gi} = \hat{\theta}_{gi} - m_g$.

To account for the fact that variability differs between genes, it standarises these estimated values so that the median of standard errors is 1 for each gene. In the example in **Fig: 2.1** the sample to filter is displayed:

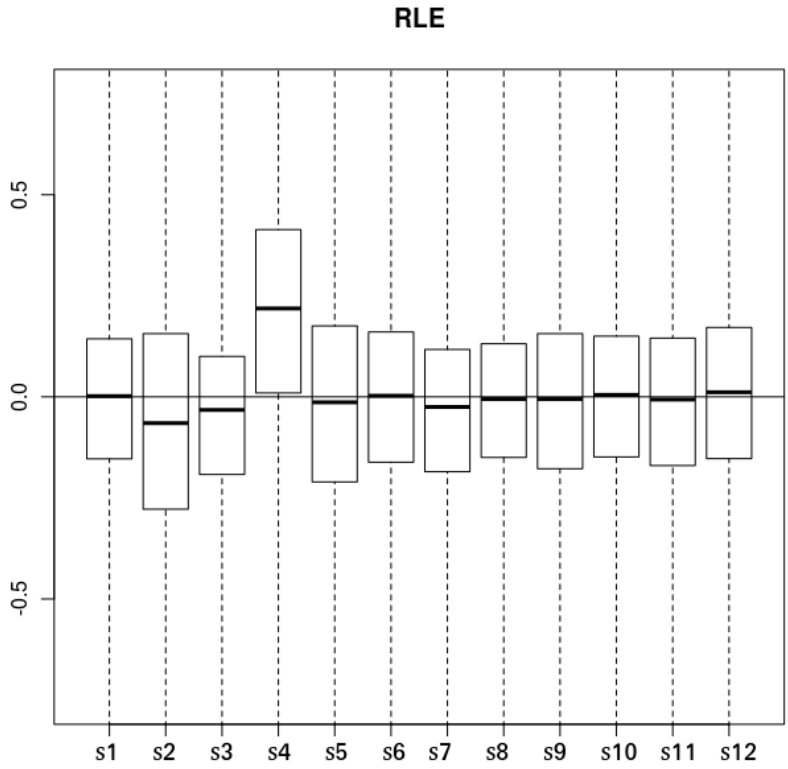


Figure 2.1: RLE plot from NUSE example. (source: bioinformatics.knowledgeblog.org)

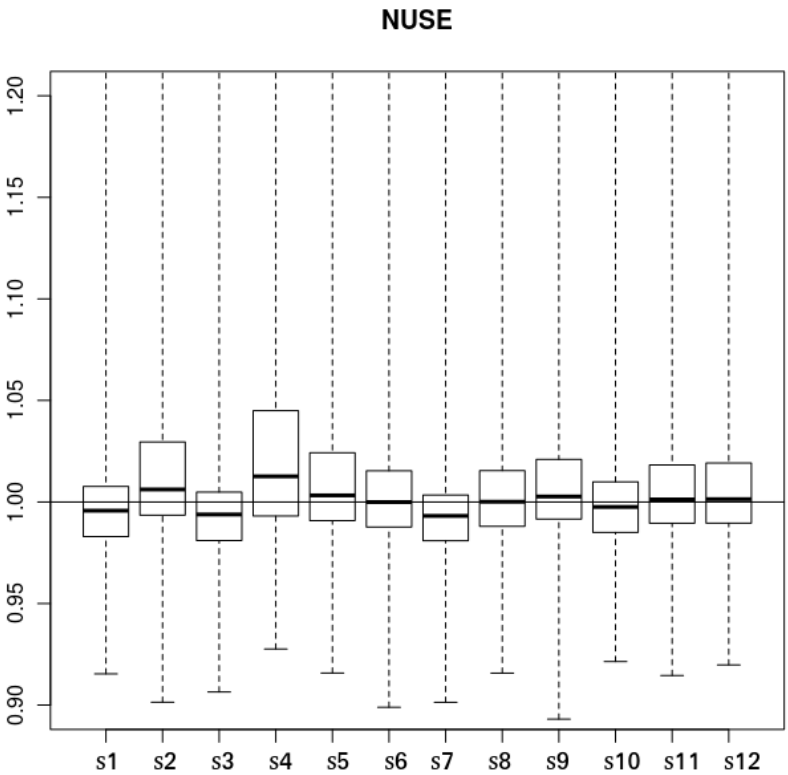


Figure 2.2: NUSE plot from NUSE example. (source: bioinformatics.knowledgeblog.org)

The function represents each sample with a boxplot. The boxplots that differ from the usual distribution represent samples that may be discarded. In this example, only one is removed because the average variation of genes in that sample is higher than in the rest. Conversely, if the interquantile range for a boxplot is significantly larger than for the rest, the corresponding array is removed.

The R function **gNUSE** (McCall et al., 2011), was used to check the quality of the arrays.

RNAseq expression data normalisation

In RNAseq data, another approach is needed; the use of FPKMs or RPKMs has been widely discussed (Conesa et al., 2016). In our case, the dataset in colon and breast cancer was created following the guidelines provided by limma users guide (Ritchie et al., 2015).

The protocol used was downloaded from TCGA (TCGA, 2019) using the tools provided by the **curatedTCGAData** and **TCGAutils** R packages as previously explained.

The phenoData matrix is obtained from a mix of sources in order to keep the maximum amount of information.

Provided the count matrix previously downloaded, the **edgeR** (McCarthy et al., 2012) package was used to remove rows that consistently have zero or very low counts. Using the **filterByExpr** function, we determine which genes have sufficiently large counts to be retained in a statistical analysis thus filtering the zero values in the rows (genes) and the ones that are below a computed minimum. The function keeps rows that have worthwhile counts in a minimum number of samples (two samples in this case because the smallest group size is two).

The filtering is performed disregarding the group, each sample belong to such that no bias is introduced.

Scale normalisation has been widely applied in RNA-seq read counts, and the **TMM normalisation method** (Robinson and Oshlack, 2010) has been found to perform well in comparative studies.

TMM method filters the genes by computing the trimmed mean using M-values(a weighted trimmed mean of the log expression ratios). This strategy makes the overall expression levels of genes comparable under the assumption that most of them are not differentially expressed.

TMM is the recommended scaling method for most RNA-Seq data when a lot of not differentially expressed pairs of samples is expected (which is very usual in RNAseq datasets). In RNAseq datasets, there are usually some genes that have both, a low expression count and a low standard deviation. These genes usually have not enough counts to be said that they are expressed.

2.2.3 Algorithms for status prediction and discovery of IHC markers

As explained earlier, errors in the determination of standard markers by inmunohistochemistry may have a substantial impact on patient health (Venet et al., 2011). In particular, the value of these markers allow us to classify breast cancer patients into four groups.

To avoid this problem, a robust predictor was depeloped to reduce the errors in the determination of IHC markers. Although several predictors have been applied to the estimation of IHC markers using the gene expresion profiles (Bartlett et al., 2016), they are not able to reduce the errors significantly in a consistent way (Mertins et al., 2016).

Our approach is based on the idea that clinical markers are determined by a set of pathways and coregulated genes. Therefore, a small subset of features that are strongly associated to clinical markers is selected, removing noisy genes. This subset of genes may provide alternative targets to standard clinical markers. Next, an ensemble of linear classifiers is built using a bagging strategy (Mbogning and Broet, 2016).

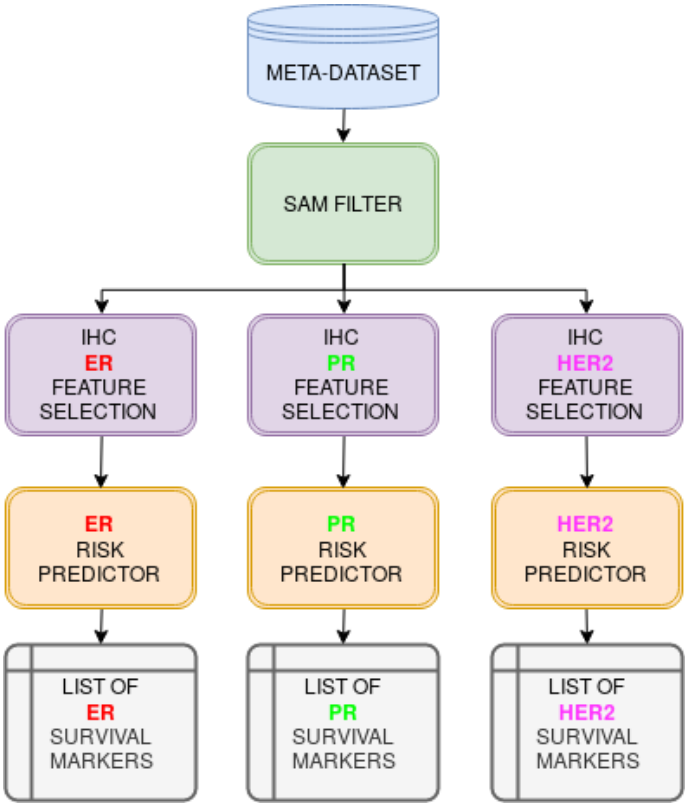


Figure 2.3: Discovery of IHC marker genes, feature prediction algorithm.

The figure (**Fig: 2.3**) describes our approach. First, a robust SAM algorithm is applied to remove noisy features. Next, for each marker, ER, PR, and HER2, a robust ensemble of classifiers is trained to obtain a related list of genes and to

improve the prediction of standard clinical markers.

Noise reduction by robust differential expression

Significance Analysis of Microarrays (SAM) (Tusher et al., 2001) from **siggenes** (Schwender, 2012) package is used in order to pre-filter genes; the whole procedure is bootstrapped (Efron and Tibshirani, 1993) with replacement.

SAM is a robust algorithm that determines if the difference between the gene expression means for two groups of samples is statistically significant. The algorithm is non-parametric and computes the null hypothesis by permutation of the sample labels. It has two relevant parameters; The first one, Δ, is a threshold that determines when the alternative hypothesis of being the difference statistically significant is true. This threshold is set up by trial and error, larger values of Δ reduce the FDR (False Discovery Rate). The second one is the FDR that determines the rate of false positives for those genes considered differentially expressed by the test ($\hat{FDR} = \frac{\hat{FP}}{\hat{TP}}$).

To improve the stability and robustness of this feature selection step, the SAM algorithm is bootstrapped (Efron and Tibshirani, 1993). Bootstrap resamples iteratively the original set of patients with replacement. Thus, for each iteration a different random sample of patients is considered. Then, the SAM algorithm is run and the list of genes is recorded. The resulting ensemble of lists is combined using as metric the number of resamples in which each gene is considered significant. The final list of genes will be stable and independent of the particular sample considered.

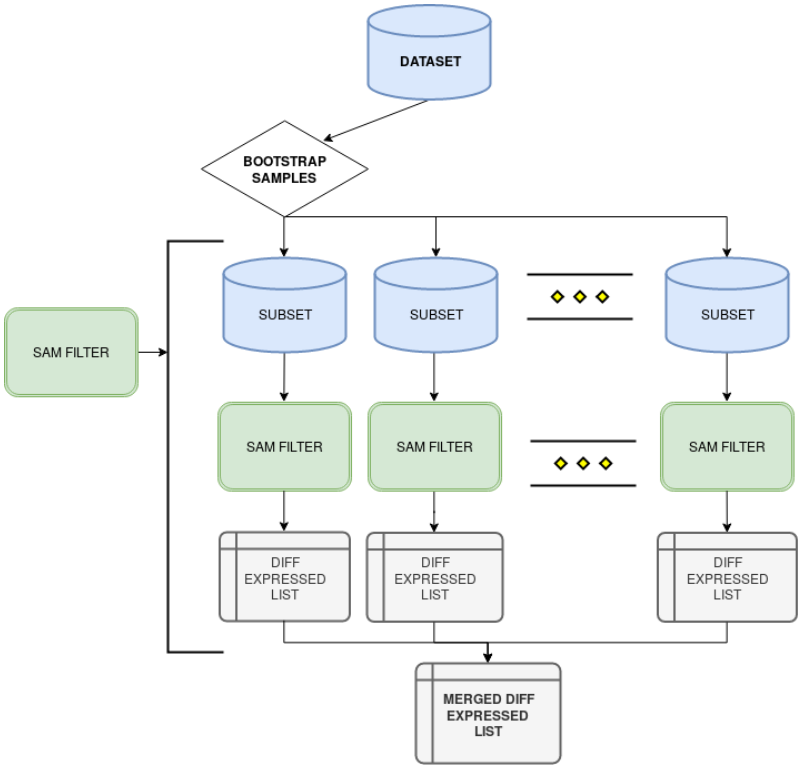


Figure 2.4: Robust differential expression SAM filter. Workflow of the algorithm.

This is the workflow of the previously described function, which will be included in the R package (R code is described in Appendix: 6.2).

For each list, the optimal Δ threshold was calculated by repeating the process using different FDR values and selecting the one that minimises the p-value. The final list included only the genes that are present in at least 10% of the iterations. This step improves the stability and reproducibility of the candidate genes and will remove noise.

IHC status prediction and discovery of class marker genes

Errors in the determination of IHC clinical or phenotypical status are high (Venet et al., 2011) (Bartlett et al., 2016) (Mertins et al., 2016). The gene expression of the corresponding markers can be considered to predict the IHC status. However, the concordance between IHC status and gene expression is low because, in IHC techniques, a doctor considers heuristic knowledge not available in microarray data. Some authors have proposed algorithms to determine the IHC status by a subset of genes instead of using a single one. In particular, Prosigna is frequently considered in Breast Cancer (NANOSTRING, 2019) (Jensen et al., 2018).

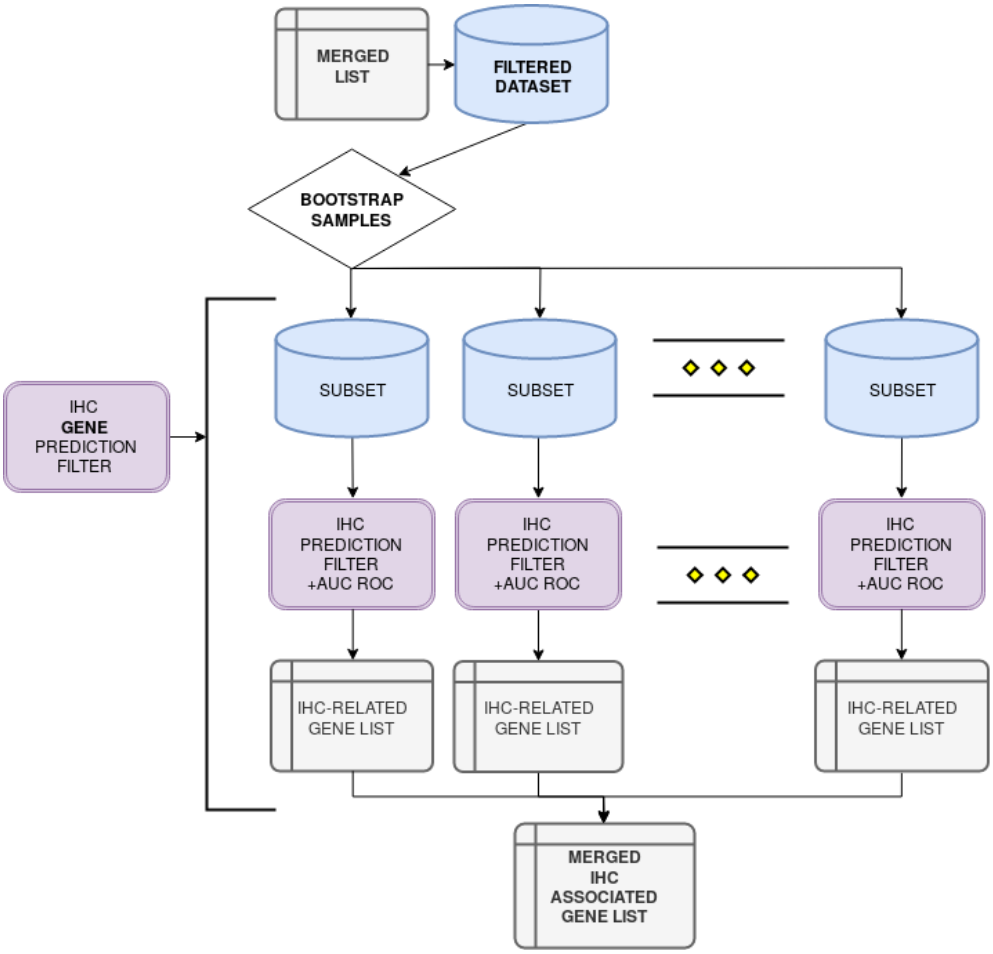


Figure 2.5: IHC prediction filter. Workflow of the algorithm.

However, the identification of subsets of genes related to IHC status suffers from a severe inestability and reproducibility problem (Mertins et al., 2016). In this section, a robust and stable feature selection algorithm is described. The resulting list of genes is considered to implement a predictor of IHC status, reducing the errors of standard approaches.

The procedure is done for each one of the IHC markers, and our approach using cross-validation will provide the method with the desired robustness. Using two-thirds of the dataset as training and the rest as validation, several predictors are computed(100 iterations in this case) and then the lists of marker genes are merged taking only the most robust ones. The performance of the whole process is evaluated computing the ROC curves for each instance of the predictor and the AUC (area under the ROC curve).

The R package glmnet (Lasso and Elastic-Net Regularised Generalised Linear Models) (Friedman et al., 2010) is used to predict the clinical status. Functions for the methods presented here are shown in the appendix. Method workflow is resumed in the following figure, and R code is described in Appendix: 6.2.

Objective measures to evaluate the predictions

Once the predictor is trained, the error prediction is evaluated. We have two types of statistical errors. Type I errors, or false positives and type II errors or a false negative. Let define the following meassures:

- Sensitivity: It gives the probability that a sample considered positive is detected by the predictor. $S = \frac{TP}{FP+TP}$
- Specificity: It gives the probability that a sample considered negative is categorised as negative by the predictor. $SP = \frac{TN}{FN+TN}$

	True condition		
Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$
Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$
Predicted condition negative	**False negative**, Type II error	**True negative**	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	

Figure 2.6: Contingency table, as example, showing the different errors. Source: "https://en.wikipedia.org/wiki/Sensitivity_and_specificity"

The false positive errors give the probability that the predictor detects a sample considered negative. Conversely, the false negative errors give the probability that a sample considered positive is categorised as negative by the predictor. Reducing false positive errors frequently increases false negative ones. Both false positive and false negative have a strong impact on patient health, and both should be considered. False positive errors are measured by $1 - SP$ and false negative ones by the $1 - S$. In **Fig: 2.6**, the table shows the different errors and metrics.

A ROC curve (Receiver Operating Characteristic) is a function that evaluates the sensibility of detection against $1 - SP$ that is false positive errors. Those predictors that achieve high sensitivity keeping a high specificity are preferred because they minimize both types of errors. In this research, the area under the ROC curve (AUC) as an objective measure to evaluate the predictors is considered as shown in **Fig: 2.7**. Those predictors with larger area achieve the right balance between both types of statistical errors and are preferred.

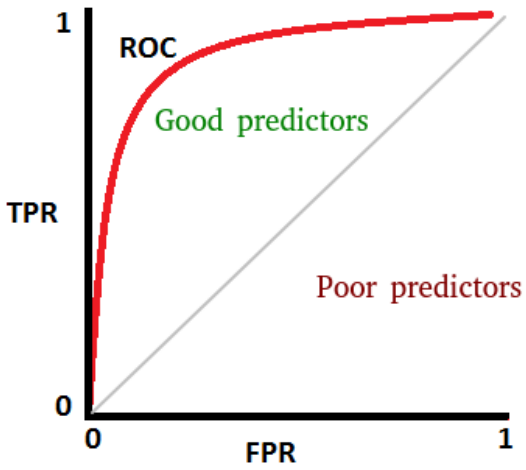


Figure 2.7: A ROC curve example showing good and poor prediction curve areas.

Elastic net algorithm for clinical status prediction

The determination of the status by inmunohistochemistry suffers from significant errors (Bartlett et al., 2016). In this section a robust linear predictor is considered to estimate the clinical markers using the gene expresion profiles and to identify alternative markers to test in the laboratory.

Linear predictors allow us to estimate the IHC status for each sample considering a small subset of genes. Besides, the same predictor can be applied to select a small number of genes strongly associated with the clinical markers. However, IHC status is frequently recorded only for a small subset of samples. Therefore, even linear predictors suffer from the small sample size problem and are prone to overfitting. To avoid this, the elastic net algorithm, introduced in (Friedman et al., 2010) is applied.

The status is a categorical variable. For simplicity, let G be the response variable that takes on values in $\mathcal{G} = \{1, 2\}$. The model can be extended easily for a more substantial number of classes. A binary variable can be aproximated by a linear model via a regularised logistic regression. Logistic regression assumes that:

$$\log \frac{Pr(G = 1|x)}{Pr(G = 2|x)} = \beta_0 + \beta^T x \tag{2.1}$$

Where $Pr(G = i|x)$ is the a posteriori probability of class i and β is the vector of linear coefficients associated with the predictors.
It can be easily shown that the a posteriori probability given the value of the gene expression is given by:

$$Pr(G = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \tag{2.2}$$

$$Pr(G = 2|x) = \frac{1}{1 + e^{+(\beta_0 + \beta^T x)}} \tag{2.3}$$

The model is adjusted by maximizing the regularised binomial likehood. If the predictor variables are standarised, this function is written as:

$$\max_{(\beta_0, \beta) \in R^{n+1}} \frac{1}{N} \sum_{i=1}^{N} \{I(g_i = 1) \log p_i + I(g_i = 2) \log(1 - p_i) + \lambda P_\alpha(\beta)\} \tag{2.4}$$

where $y_i = I(g_i = 1)$ is the class label, $p_i = Pr(G = 1|x_i)$ is the a posteriori probability for sample x_i.

$P_\alpha(\beta)$ is a regularization term that plays a critical role to avoid overfitting. For the elastic net algorithm this term is defined as:

$$P_\alpha(\beta) = \sum_{j=1}^{p} \frac{1}{2}(1 - \alpha)\beta_j^2 + \alpha|\beta_j| \tag{2.5}$$

The α parameter determines the type of regularisation considered and has a strong impact on the solution. Thus, for $\alpha = 0$, the model is equivalent to ridge regression. The solution is dense and a large number of β_i will become small but not equal to zero. For $\alpha = 1$ the model reduces to Lasso. The solution is sparse and most of the β_i will become zero. The non-zero coefficients correspond to the predictor variables strongly associated to the class status. This kind of model is able to select a small subset of genes associated to the IHC classes.

Finally, λ is a regularisation parameter that should be determined by nested cross-validation to avoid overfitting.

To improve the robustness of the status prediction and the gene list obtained, a bagging strategy is implemented. Therefore, diversity between classifiers is induced by a bootstrap resampling technique. Then, an ensemble of predictors is built that allow us to reduce the errors in the estimation of the status and to improve the stability of the gene lists.

2.2.4 Survival analysis methods

The identification of genes related to survival is probably a critical point to detect new terapeutic targets. In this section, the discovery of stable and reproducible survival markers that can be interpreted in terms of standard clinical markers is considered. We have considered two approaches: The first one is univariate, and try to retrieve individual genes strongly associated to survival. The second one is multivariate and takes into account additive interactions between genes and their correlation structure.

Bootstrapped univariate Cox regression to discover survival markers

The Cox regression approach is preferred against non-parametric techniques because it is less sensitive to the small sample size problem and it does not assume a particular distribution for data. Besides, the univariate approach is less prone to overfitting.

Let $h(t|\mathbf{x})$ be the conditional hazard function given the value of the variables $\mathbf{x}$. The hazard function can be interpreted as the probability of failure at time t given that the patient is alive before t and once the value of the variables is observed.
The Cox regression is a semi-parametric regression algorithm that assumes that the logarithmic transformation of the hazard ratio for two states of the gene expression can be aproximated by a linear model:

$$\log\left(\frac{h(t|\mathbf{x}_2)}{h(t|\mathbf{x}_1)}\right) = \beta^T(\mathbf{x}_2 - \mathbf{x}_1) \tag{2.6}$$

Where β is a coefficient that determines the change in the logarithmic hazard ratio by each unity of increment in the expression level.
This kind of model is quite robust to overfitting because it is linear and univariate. Univariate models are more robust to the small sample size problem. Any parametric form for the hazard ($h(t|\mathbf{x})$) is not assumed.
For each β_i, the Wald statistic is computed in order to determine the probability of being $\beta_i = 0$. Those genes with a strong association with the hazard ratio will have smaller p-value. If $\beta_i > 0$ then, the overexpressed genes increment the risk of failure while when $\beta_i < 0$ the overexpressed genes reduce the risk of failure.
The β parameters are determined by optimisation of the partial log-likehood. The partial log-likehood takes into account censored patients that could not be followed until the end of the study. Once the parameters are optimised, several statistical test evaluate the adjustment of the model to the data. Likehood ratio and log-rank test are included.

Although Cox regression has been widely applied to gene selection and risk survival prediction in the literature, the stability and reproducibility of the list of genes obtained should be improved. To this aim, a bootstraped version of the original Cox algorithm is considered. In particular, the set of patients is resampled with replacement 100 times. For each sample, a ranked list of genes is obtained. Finally, an ensemble strategy is applied and the set of list are merged into a single one employing several metrics.

Univariate cox regression has been implemented using the R function **coxph** from survival package (Therneau, 2014). The appendix 6.2 shows the code developed for this problem.

A non parametric strategy to improve risk prediction and patient stratification

In this section, a univariate and non-parametric strategy to rank genes is considered according to their ability to predict survival.

Let T be the non-negative random variable that represents lifetime in a population. The survival function is defined $S(t) = p(T > t)$ as the probability that an individual survives to time t. $S(t)$ can be approximated by the non-parametric Kaplan-Meier estimator.

Let T_i be the life time for individual i and C_i the time for which the individual i get out of the study. The following pairs of variables (Y_i, δ_i) are observed for each:

$$Y_i = \min\{T_i, C_i\} \qquad \delta_i = \begin{cases} 1 & T_i < C_i \\ 0 & C_i < T_i \end{cases} \tag{2.7}$$

Where δ_i determines if the event is censored or not. Consider the following notation:

$y_{(i)}$ is the time for the ordered censored or uncensored observations.
$n_i = \mathcal{R}_{y_{(i)}}$ denotes the subset of patients at risk before $y_{(i)}$.
d_i = number of failures at time $y_{(i)}$.
$p_i = P(T > y_{(i)} | T > y_{(i-1)})$.
$q_i = 1 - p_i$.

Then, the survival function can be estimated as:

$$S(t) = P(T > t) = \prod_{y_{(i)} \leq t} p_i \tag{2.8}$$

Considering that $\hat{q}_i = \frac{d_i}{n_i}$ and $\hat{p}_i = 1 - \frac{d_i}{n_i} = \frac{n_i - d_i}{n_i}$, the Kaplan Meier estimaton can be written as:

$$\hat{S}(t) = \prod_{y_{(i)} \leq t} \left(\frac{n_i - d_i}{n_i} \right) \tag{2.9}$$

A Kaplan-Meier curve shows the estimated survival function by plotting the estimated survival probabilities against time. The estimated survival probability is constant between the events. Therefore, the curve is a step-function in which each vertical drop indicates the occurrence of one or more events. The right censored data are represented with a vertical mark in the curve.

We have developed a robust algorithm which have been designed to compute Kaplan-Meier curves with a different approach. This algorithm will be a part of a complete R Bioconductor (Huber et al., 2015) package which is actually in late phase development.

The ability of each gene to predict survival is evaluated as follows. First, the patients are splitted into two groups according to the gene expression level. Next, a log-rank test is computed to evaluate the difference between the kaplan Meier curves for each group.

This statistical test is non-parametric and makes no explicit assumptions about the form of the survival curves. The algorithm looks for the optimal splitting that maximises the separability between the Kaplan-Meier curves. Genes that split the patients into groups of different prognosis with statistically significant p-value are considered as markers of survival.

In order to improve the robustness, a variant of the logrank and the Kaplan-Meier estimator is developed. It has been reported that Kaplan-Meier and logrank lack reproducibility and give different results depending the dataset considered.

This observation is true even if the datasets share a significant amount of samples (Raman et al., 2019). In order to deal with this problem the following method is developed.

Patient stratification has usually been done in the literature comparing the expression level of a given gene with the median for the entire group of individuals (Klein and Moeschberger, 1997). This method fails for non-standard problems.

Next,the method that we have developed based on the log-rank statistics is considered:

(i) Patients are ranked according to the gene expression level for a given gene. Let g_{ij} denotes the gene expression for gene i and patient j.

(ii) Define the threshold $\theta_{ij} = g_{ij}$ for gene i and patient j. Define the group variable as:

$$G(g_{ij'}) = \begin{cases} 1 & g_{ij'} > \theta_{ij} \\ 0 & g_{ij'} < \theta_{ij} \end{cases} \tag{2.10}$$

This variable splits the patients in two risk groups according to the expression level for gene i.

(iii) For each j compute the p-value using the log-rank statistic.

(iv) Build the log-rank curve. The lowest p-value defines the optimal splitting.

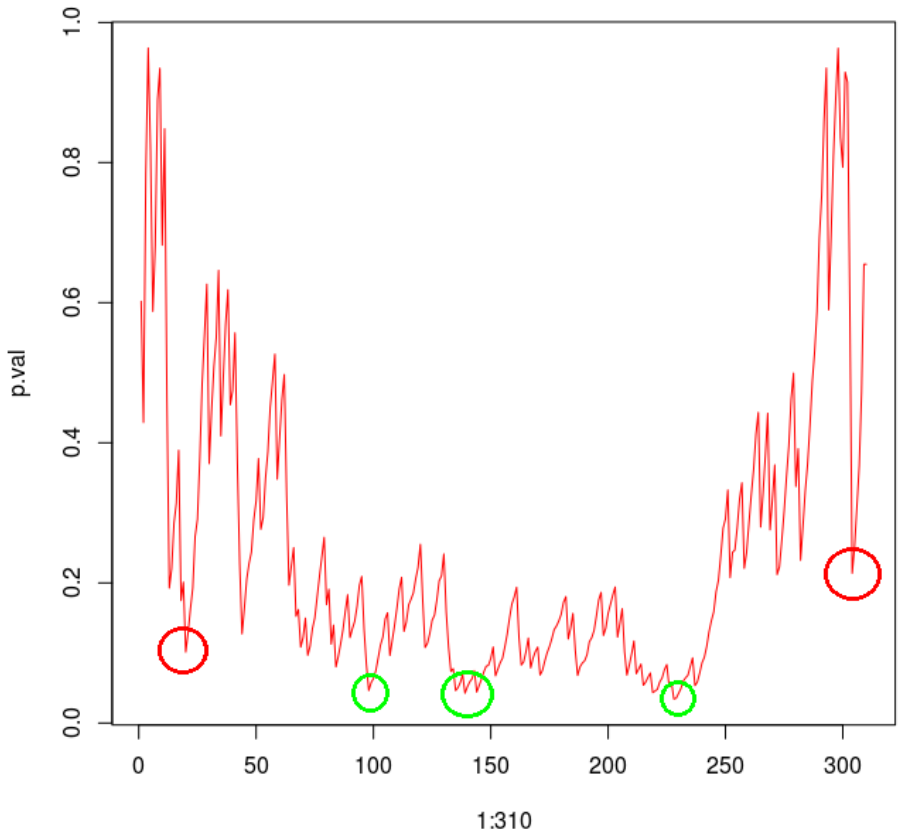


Figure 2.8: P-value distribution ordered by expression level. Relative minima in red, candidates to absolute minimum in green.

Figure 2.8 shows that the method introduced is susceptible to local minima, giving rise to meaningless solutions.

Logrank efficiency and optimisation

In this section, a method to improve the computational efficiency of the logrank algorithm and to avoid local minima is developed. R code is described in Appendix: 6.2.

The new approach will compute the threshold θ_{ij} in step (ii) considering only patients between the 25 and 75 quantiles. This step will avoid highly umbalanced groups that correspond to local minima of the log-rank curve. Besides, it will improve the computational eficiency.

The other modification is done in order to re-calculate the group membership for each central sample. The problem arises when a large proportion of samples around the optimal threshold have similar value for the gene expression.

In order to better identify those patients, a bootstrapped logrank is performed accounting for a more robust algorithm at the time of categorising a sample in a good or bad prognosis group. The algorithm resamples iteratively the original set of patients and classifies each individual.

The frequency with which each patient has been assigned to each prognostic group is computed. This will allow us to estimate a membership probability to each group. This feature will help the medical doctors to choose the right treatment for each patient.

In **Fig: 2.9**, the "group membership probability" is represented with colours in a metric that is more representative.

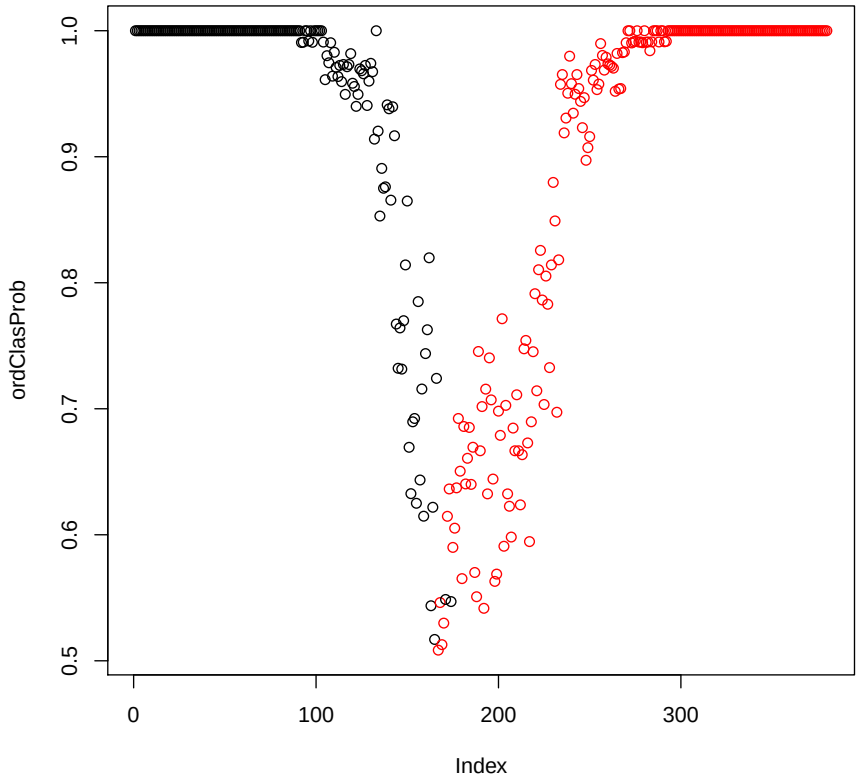


Figure 2.9: Class membership, black and red colours define the new groups as high and low expression.

Patients in this plot are ordered by the expression level. Samples from left to right (coloured in black) are the ones assigned with a class membership probability higher than 0.5 (CMp $>$ 0.5) to a low expression group. Patients coloured in red are assigned to the high expression group.

The class membership probability is computed from the number of times that a sample has been assigned to one or other group in all the bootstrap runs. As it can be seen a single run of the logrank optimised method is not enough to provide a proper classification.

2.2.5 Multivariate approach: risk prediction and gene selection

Survival in cancer is determined by multiple genes that cooperate in carrying out biological functions. In this section, a multivariate approach is considered with two objectives: the first one is to improve the risk prediction and patient stratification. The second one is to identify gene markers related to survival, considering additive interactions and the corregulation structure of genes.

Cox-multivariate proportional hazards model

The risk prediction power is a milestone in cancer markers (Chibon, 2013). In order to evaluate the prognostic value of the best markers, a robust version of the multivariate Cox proportional hazards regression model with L_1 norm penalty is developed. This penalty shrink to zero the coefficients which are useless to predict the hazard of death. This algorithm will be a part of a complete R Bioconductor (Huber et al., 2015) package which is in late phase development.

Let $X_1, X_2, \ldots, X_p$ be the expression levels of the p genes. For each sample i, let (t_i, δ_i) be the survival time and the censoring indicator for patient i respectively. The hazard of death at time t given the observed values of the gene expression $(\lambda(t|x))$ can be modelled using a Cox regression:

$$\lambda(t|x) = \lambda_0(t) \exp\left(\sum_j \beta_j X_j\right) = \lambda_0(t) \exp\left(\beta^T X\right) \tag{2.11}$$

Where $\lambda_0(t)$ is an unspecified baseline hazard function, $(\beta_1, \beta_2, \ldots, \beta_p)$ are the regression coefficients and $(X_1, X_2, \ldots, X_p)$, are the gene expression levels. $f(X) = \beta^T X$ is the linear risk score for the corresponding patient.

The β_j coefficients are estimated by maximizing the partial log-likehood with L_1 norm penalty:

$$l(\beta) = \sum_{j=1}^{p} \sum_{k=1}^{n} \left(x_{kj}\beta_j - \log \sum_{m \in \mathcal{R}_k} \exp(x_{mj}\beta_j \right) - \lambda \sum_j |\beta_j| \tag{2.12}$$

Where $\mathcal{R}_k$ is the set of patients at risk for time t_k and λ a regularisation parameter that is estimated by ten-fold cross-validation. This parameter allows us to shrink most of the β_j coefficients to zero.

Note that the norm penalty might not make sense if the predictors are in different units. Therefore, all the predictors are standardised by the Fisher Information matrix which allow us to interpret the coefficients as the predictive power of gene j.

The algorithm is explained in **Fig: 2.10**. First, feature selection is implemented considering the IHC associated list of genes. Double nested cross-validation is applied. The inner loop estimates the optimal λ regularisation parameter for each predictor by ten-fold cross-validation. In the outer loop, ten-fold cross-validation

predict the risk score considering non-overlapping training and test sets. This strategy helps to avoid overfitting improving the stability and reproducibility of the experiments. The risk score is scaled to 0-100. Next section presents an algorithm to stratify the group of patients considering the risk curve. R code is described in Appendix: 6.2.

Figure 2.10: Risk prediction groups using cross validation. Workflow of the algorithm.

Improving risk stratification for medical decission making

The usual method for risk group classification is based in heuristic thresholds. This strategy does not perform well when the groups of risk are unbalanced. The number of patients usually asigned to the mid-risk group is almost 50%. Therefore, an algorithm that estimates the optimal thresholds and reduce the "twilight zone" of intermediate risk is developed.

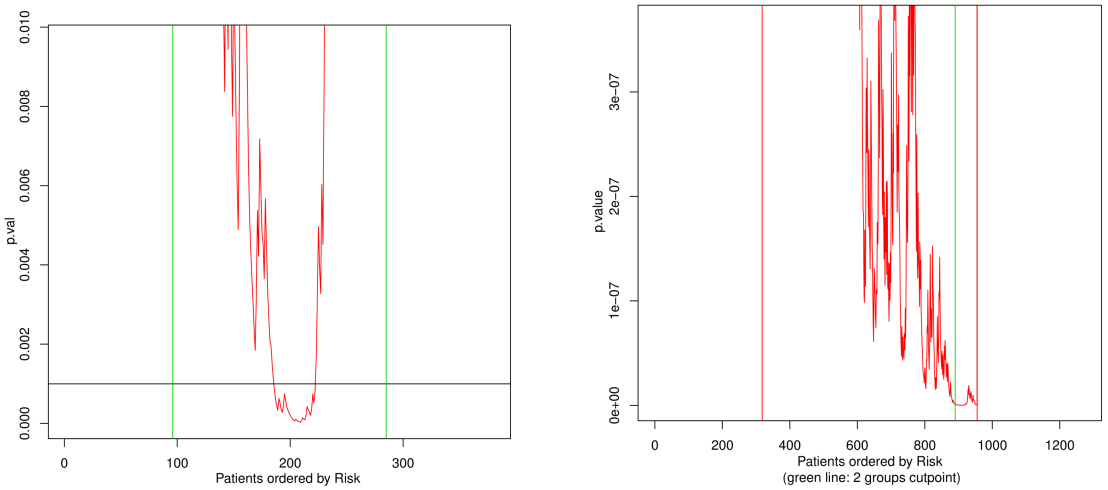
The new contribution to the model is the way Risk Score is treated to make different groups. Risk score output from cross-validated uniCox (Tibshirani, 2009) R function is ordered, transformed to a 0-100 interval and stratified into regions of risk: low, mid, and high. The algorithm classifies the mid-risk samples as follows.
(i) Patients are ranked according to the risk score.
Let R_j be the predicted risk for patient j.

(ii) Define the threshold $\theta_j = R_j$. Let the group variable be:

$$G(R_{j'}) = \begin{cases} 1 & R_{j'} > \theta_r \\ 0 & R_{j'} < \theta_r \end{cases} \tag{2.13}$$

This variable splits the patients in two prognostic groups acording to the predicted risk.
(iii)For each j compute the p-value using the log-rank statistic.
(iv)Build the log-rank curve. The lowest p-value defines the optimal splitting.

We provide several plots for further understanding of the capability of the selected genes to identify risk groups.

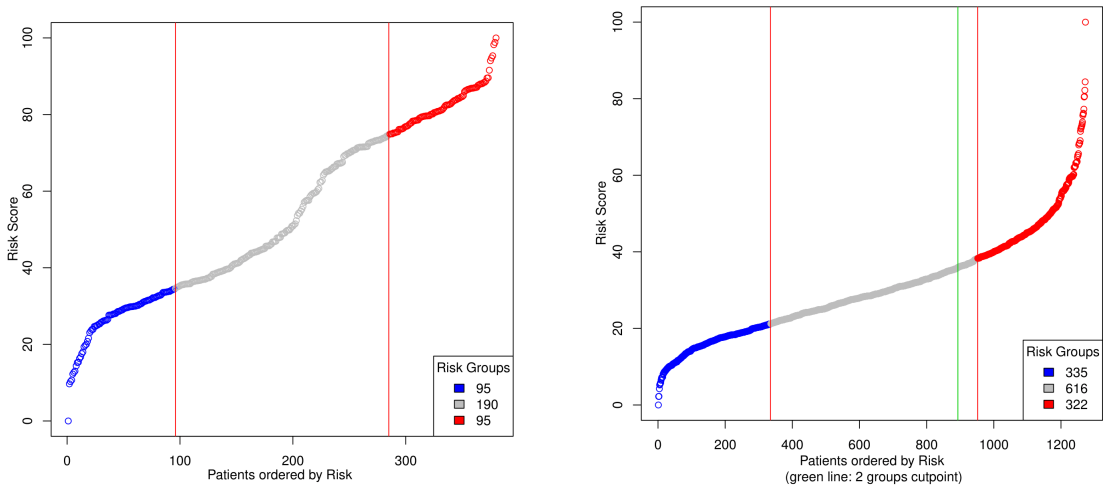


(a) Ideal case **(b)** Local minima case

Figure 2.11: Risk prediction output, ordered p-values by risk.

In a simple problem, **Fig: 2.11(a)** the logrank curve is a quadratic function with a minimum. This minimum is the selected cutpoint to stratify risk in training and validation sets.

The **Fig: 2.11(b)** shows a spiked and wide curve. It means that defining two risk groups is not easy and a lot of relative minima in the risk groups selection nay be present.



(a) Ideal case **(b)** Less defined risk curve

Figure 2.12: Ordered risk score prediction from multivariate predictions.

Figure 2.12(a) shows the stratification of the risk score for the ideal log-rank curve 2.11(a). Similarly, figure 2.12(b) shows the stratification for curve 2.11(b).

The **Fig: 2.12** shows the ordered risk score and the regions, as previously described. The mid region is the ambiguous one, with samples that may be classified to both groups with almost the same probability.

While the **Fig: 2.12(a)** (risk curve with less sigmoid shape) shows an ideal case in which the risk score is more clearly defined and a strong change in the slope for the mid group samples is present. The **Fig: 2.12(b)** shows a less differentiated or defined risk curve, more difficult to stratify.

2.3 Results

In this section, the biological contributions in breast cancer are presented as obtained from the algorithms and data described in section 2.2. Discoveries in the form of new marker genes, the compilation of series, tables and, plots are explained further in this chapter.

2.3.1 Quality control of normalised gene expression data

We have obtained a curated meta-dataset which integrates the biggest amount of microarrays with survival data publicly available as described in Section 2.2.1, and more specifically in the Tab: 2.1. This meta-dataset has been normalised and standardised in a way that guarantees a batch effect free compilation.

The quality of the signal obtained by the normalisation algorithm is analysed. **Fig: 2.13** shows the distribution of expression values for each one of the batches:

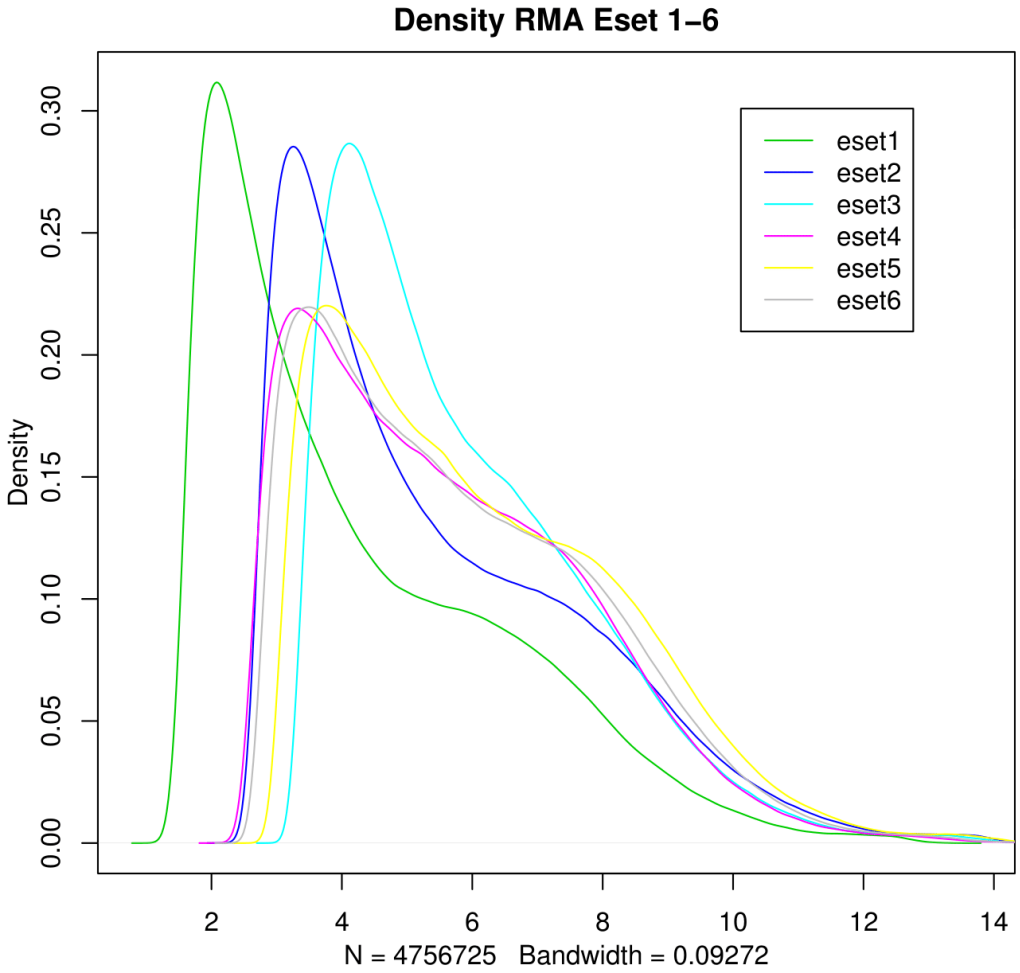


Figure 2.13: Esets representing each series batch (not normalised).

The Fig. 2.13, represents the expression with RMA normalisation proposed by Irizarri. The differences between the series are too high and a study performed with this normalization method would give meaningless biological results.

The batch effect in this case will be high, as a bias is introduced in the compilation by each one of the esets.

By contrast, the following figure shows the method fRMA+COMBAT considered in this thesis.

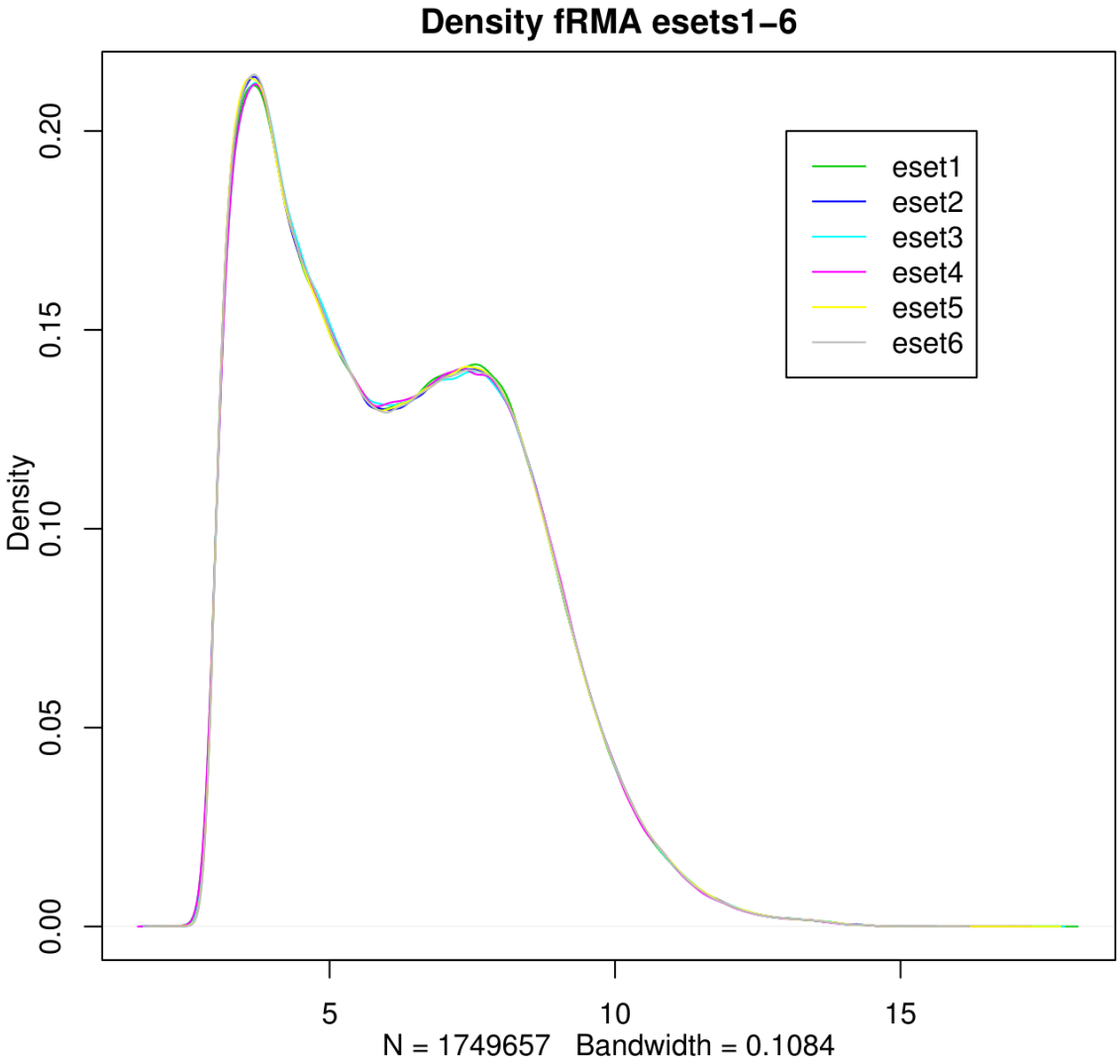


Figure 2.14: Esets representing each series batch (normalised).

As it can be readily appreciated, the expression distribution for each one of the batches or series is similar. The subsequent analysis within this data is guaranteed to be batch effect free.

The second quality control test is based on a hierarchical clustering algorithm. When the different series are not properly normalized, the samples cluster together just because they belong to the same bath or series.

Fig. 2.15, shows a hierarchical clustering algorithm of all microarrays, using Ward method and Manhattan distance.

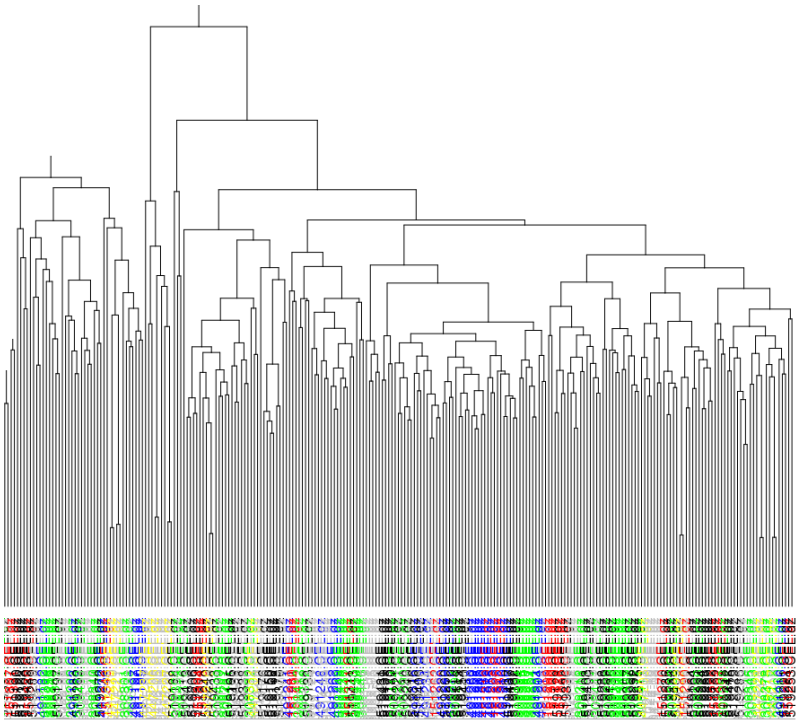


Figure 2.15: Dendogram showing clustering of samples.

Samples from different batches do not cluster together, which suggest the normalisation algorithm has removed the batch problem.

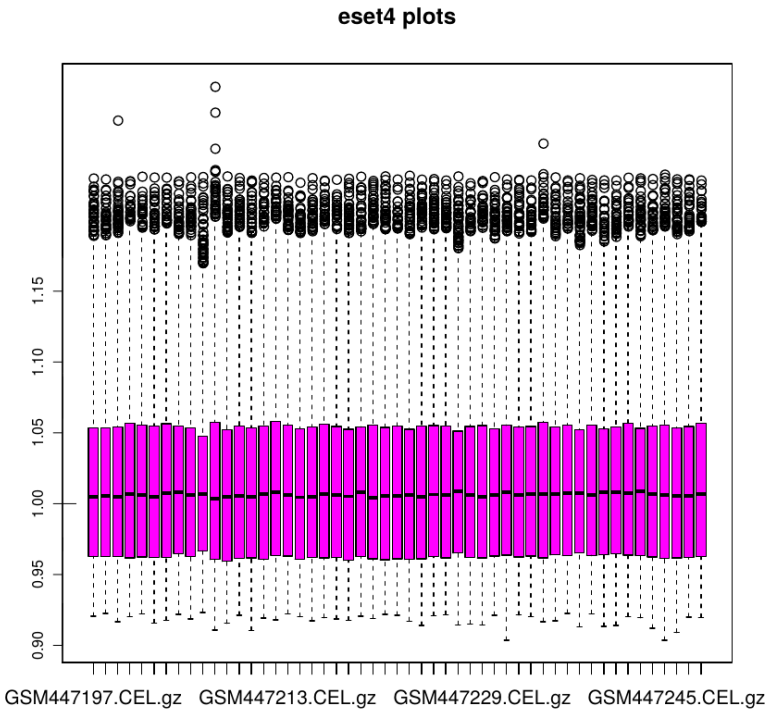


Figure 2.16: NUSE representation of our samples.

Finally, in figure 2.16, gNUSE algorithm is applied.No statistical differences among the medians and interquantile ranges of different boxplots can be appreciated.

Once a normalisation algorithm is applied, a well designed and normalised matrix is obtained with complete clinical information and a sample size of 1024.

2.3.2 A survival gene signature related to standard clinical markers

Several survival gene signatures of breast cancer have been proposed in the literature. Even more, standard platform such as oncotype or prosigna have their own heuristic signature which is applied to stratify patients according to their risk. However, as several authors have mentioned, the overlapping of the different gene signatures is very small and frequently, the resulting genes can not be related to the standard clinical markers. This is a serious drawback to transfer the biological findings to the clinical practice.

The main contribution of this section is to propose a robust and stable list of survival markers that can be interpreted in terms of the standard clinical ones. Moreover, some of the genes discovered may suggest alternative targets to clinical markers easier to detect in the laboratory.

The methods developed in this chapter provide three tables of risk and class markers for breast cancer. Each IHC marker that defines an intrinsic subtype have a table with a ranked list of survival markers.

geneName	KMpvalTrain	betaCoxph	p.coxph	betaUniCox	se.beta.uniCox
TBC1D9	0.00000057	-0.273	0.000067	-0.0461	0.0119
SUSD3	0.00000880	-0.298	0.000080	-0.0490	0.0146
SLC39A6	0.00006319	-0.225	0.001527	-0.0374	0.0127
GFRA1	0.00000189	-0.177	0.001640	-0.0260	0.0079
SOX11	0.00000027	0.154	0.003205	0.0250	0.0097
GATA3	0.00249397	-0.154	0.012805	-0.0254	0.0100
SLC15A2	0.00831951	0.316	0.017415	0.0779	0.0416
C6orf97	0.00029611	-0.244	0.017643	-0.0494	0.0213
NANOS1	0.00153739	0.153	0.019000	0.0264	0.0138
ZNF552	0.00003888	-0.182	0.019474	-0.0345	0.0156
ESR1	0.00057954	-0.239	0.030505	-0.0489	0.0218
NAT1	0.00459318	-0.098	0.032377	-0.0133	0.0061
NME3	0.00342702	-0.303	0.037352	-0.0799	0.0426
DNALI1	0.00320825	-0.112	0.068916	-0.0189	0.0109
AGR3	0.00628798	-0.056	0.070142	-0.0071	0.0042
CA12	0.00012777	-0.104	0.079446	-0.0167	0.0099

Table 2.3: ER survival markers table.

The following metrics and statistical tests have been computed:

The **KMpvalTrain** is the p-value for the non-parametric bootstraped Kaplan Meier method. The **betaCoxph** is the beta value assigned by **coxph** function for univariate Cox regression. A positive value indicates that the overexpression increment the risk of failure. A negative value suggests that overexpression reduces the risk of failure. The **p coxph** is the Wald statistic p-value that indicates how significative is the relationship between the individual gene expression and risk. The **betaUniCox** is the coefficient for multivariate cox regression; in this column the higher is the absolute value the higher is the relevance of the gene in the proposed model. The **se.beta.uniCox** measures how much may vary the beta from previous column. Sometimes a beta with high SD may be worse than a higher absolute beta.

geneName	KMpvalTrain	betaCoxph	p.coxph	betaUniCox	se.beta.uniCox
PNMT	0.0155	0.140	0.049	0.0358	0.023
CWC25	0.1768	0.234	0.088	0.0706	0.059
C17orf37	0.0056	0.118	0.162	0.0258	0.023
SIRT3	0.0093	-0.371	0.207	-0.1940	0.133
MED1	0.3471	0.116	0.268	0.0301	0.036
ERBB2	0.0279	0.069	0.357	0.0149	0.018
KMO	0.2884	0.063	0.366	0.0152	0.017
PSMD3	0.0149	0.081	0.424	0.0167	0.033
GRB7	0.0244	0.049	0.511	0.0094	0.016
CRKRS	0.3292	0.069	0.556	0.0214	0.043
PGAP3	0.1891	0.046	0.599	0.0093	0.024
KLC4	0.1549	-0.053	0.867	-0.0399	0.203
STARD3	0.0476	0.017	0.869	0.0016	0.036
CNKSR1	0.8930	0.014	0.936	0.0037	0.074

Table 2.4: PR survival markers table.

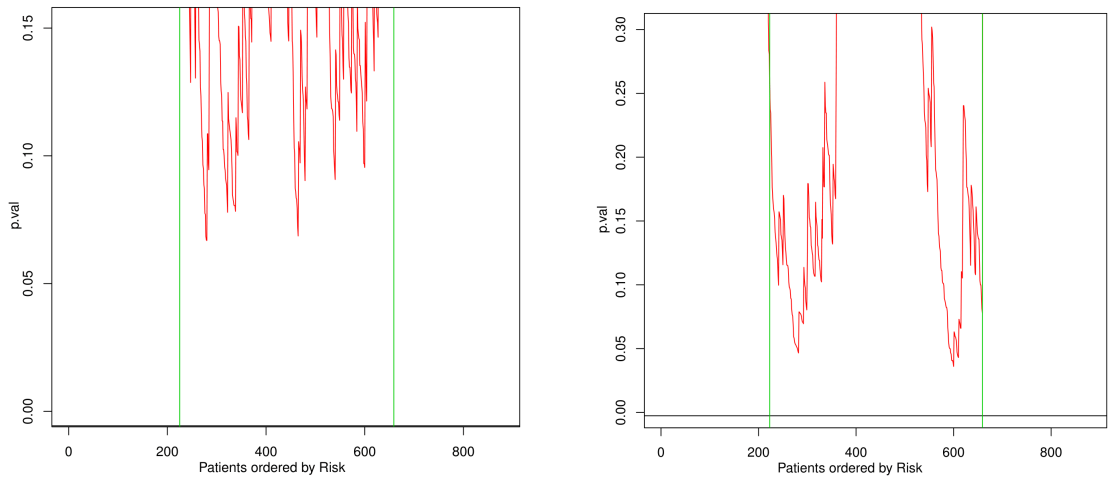
geneName	KMpvalTrain	betaCoxph	p.coxph	betaUniCox	se.beta.uniCox
SUSD3	0.0000088	-0.298	0.00008	-0.0445	0.0133
GFRA1	0.0001724	-0.177	0.00164	-0.0233	0.0071
PGR	0.0001936	-0.161	0.01368	-0.0217	0.0100
C6orf97	0.0002847	-0.244	0.01764	-0.0456	0.0196
ESR1	0.0002989	-0.239	0.03050	-0.0452	0.0202
NAT1	0.0013054	-0.098	0.03238	-0.0119	0.0055
DNALI1	0.0023350	-0.112	0.06892	-0.0171	0.0099
AGR3	0.0011249	-0.056	0.07014	-0.0063	0.0037
CA12	0.0001278	-0.104	0.07945	-0.0151	0.0089
TFF1	0.0037668	-0.053	0.09463	-0.0064	0.0037

Table 2.5: HER2 survival markers table.

Tabs: 2.3, 2.4 and 2.5 show the list of genes related to each one of the IHC markers (ER, PR, and HER2). We remark that the genes shown in previous tables are quite stable and independent of the particular sample considered. Besides, all the genes have been ranked high according to several metrics.
Next section comments the relevance of the survival markers discovered.

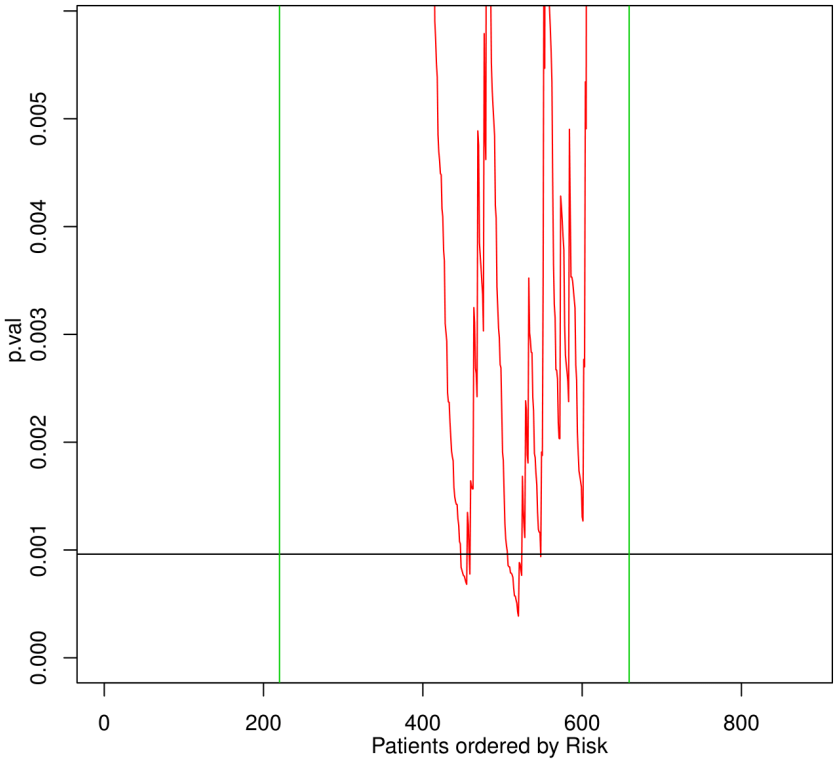
2.3.3 A new survival signature that outperforms Oncotype and Prosigna

Concordance between our marker gene list and oncotype and Prosigna is low (see table 2.6). Several questions will be answered in this section: Is the new signature proposed better to predict the risk and to stratify the patients? Our signature is able to reduce the uncertainty in the categorization of patients according to their risk? The markers discovered can be validated using other technologies such as RNA-seq and independent datasets?



(a) Oncotype

(b) Prosigna

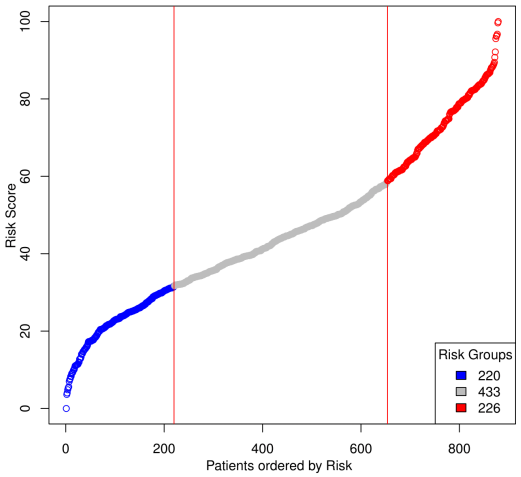


(c) Our approach. 34 markers

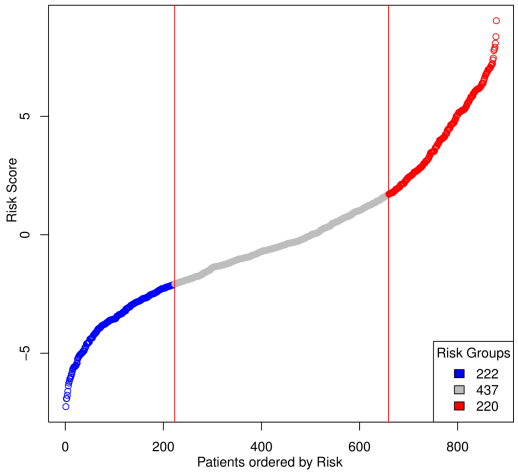
Figure 2.17: Robust p-value calculation for mid risk group. Comparison of markers.

Our list of genes is compared with standard platforms for risk prediction such as Oncotype and Prosigna. The list of genes considered by each platform are defined in Tab: 2.6. To compare the risk prediction ability of each gene signature the same multivariate predictor introduced in this chapter is applied.

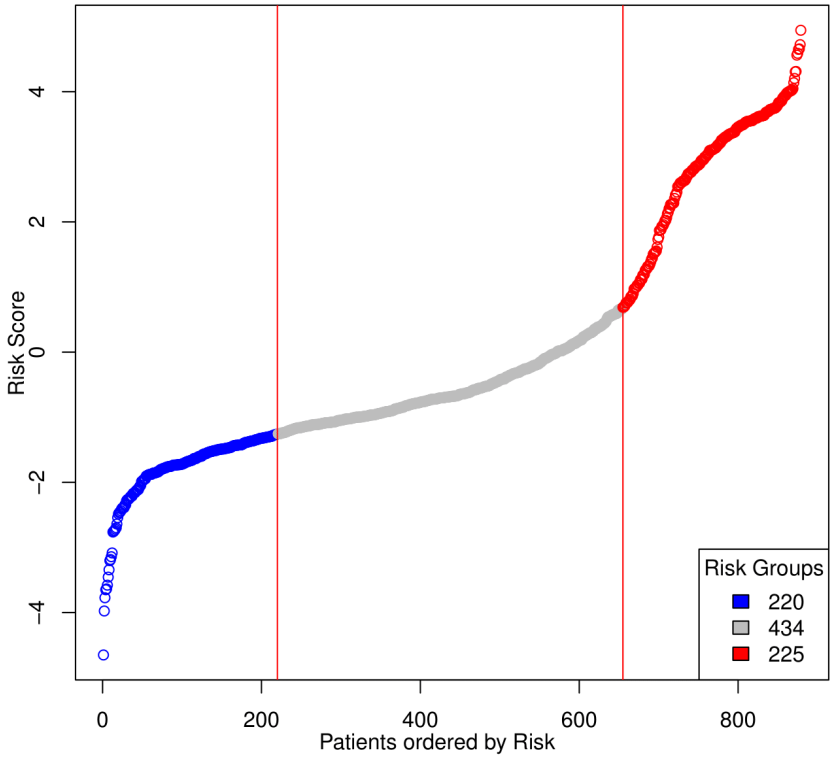
The **Fig: 2.17** compares the p-values distribution for central mid-risk score group, which is portrayed in **Fig: 2.18**. The results can be evaluated using the following guidelines:



(a) Oncotype



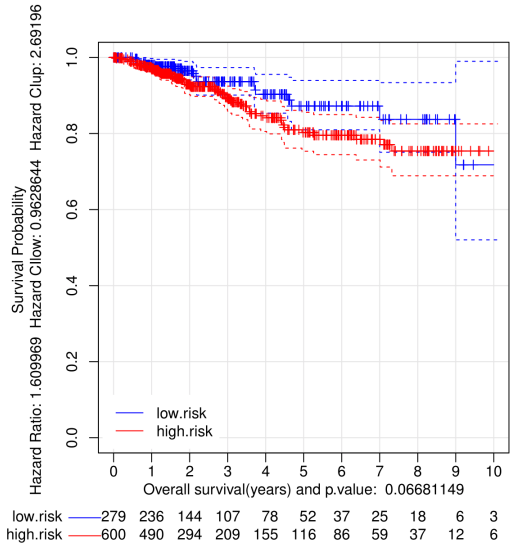
(b) Prosigna



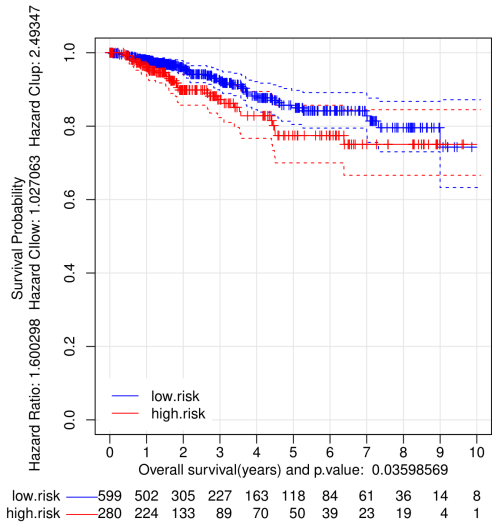
(c) robust 34 markers

Figure 2.18: Ordered risk score curves. Comparison of markers.

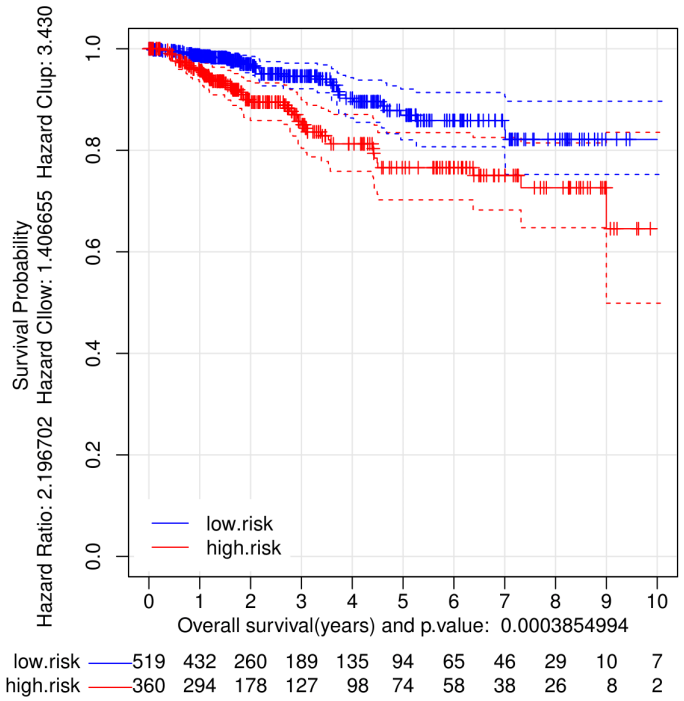
If the p-val distribution follows a quadratic like distribution, and the relative minima and absolute minimum are in a narrow window (case **Fig: 2.17(c)**), a better risk prediction is obtained. If the window is wider, and there are a lot of relative minima across it, then we will have a poorer prediction (case **Fig: 2.17(c)**). Notice that the interval using our genes is better than the others, showing a shorter interval width. This result responds to the second question. Our gene signature is able to reduce the uncertainty in the categorization of patients in risk groups. Usually the shape of the ordered risk score curve follows a sigmoid-like distribution. If this distribution is closer to a straight line, then the risk stratification will be poorer (case **Fig: 2.17(c)**).



(a) Oncotype (p=0.0668)



(b) Prosigna (p=0.0359)



(c) Robust 34 markers (p=0.00038)

Figure 2.19: Kaplan-Meier curves, high risk and low risk. Comparison of markers.

The previous qualitative results can be corroborated with the Fig. 2.19. In this figure, the cases (a) and (b) show a poorer curve separability between the Kaplan-Meier curves. Instead, the (c) case has a better separability and a much better p-value than the commercial chips Prosigna and Oncotype.

The hazard ratio for the survival curves obtained with the gene signature proposed is significantly better than for Prosigna and Oncotype. This result gives the answer to question one. Our gene signature is able to improve the risk prediction and the stratificacion of patients according to their risk.

The genes used in each one of the predictors are described in the following table. The Prosigna and OncotypeDX gene list were obtained from a previous study (Bartlett et al., 2016). They are described in the following Table: 2.6.

Prosigna(49g)		**Oncotype(16g)**	**New Markers(34g)**	
ACTR3B	KRT17	BAG1	AGR3	SIRT3
ANLN	KRT5	BCL2	AURKA	SLC15A2
BAG1	MAPT	BIRC5	MIEN1	SLC39A6
BCL2	MDM2	CCNB1	CCDC170	SOX11
BIRC5	MELK	CD68	CA12	STARD3
BLVRA	MIA	CTSV	CNKSR1	SUSD3
CCNB1	MKI67	ERBB2	CDK12	TBC1D9
CCNE1	MLPH	ESR1	CWC25	TFF1
CDC20	MMP11	GRB7	DNALI1	ZNF552
CDC6	MYBL2	GSTM1	ERBB2	
CDH3	MYC	MKI67	ESR1	
CENPF	NAT1	MMP11	GATA3	
CEP55	ORC6L	MYBL2	GFRA1	
CXXC5	PGR	PGR	GRB7	
EGFR	PHGDH	SCUBE2	KLC4	
ERBB2	PTTG1	AURKA	KMO	
ESR1	RRM2		MED1	
EXO1	SFRP1		MKI67	
FGFR4	SLC39A6		NANOS1	
FOXA1	TMEM45B		NAT1	
FOXC1	TYMS		NME3	
GPR160	UBE2C		PGAP3	
GRB7	UBE2T		PGR	
KIF2C	PSMC4		PNMT	
KRT14			PSMD3	

Table 2.6: Breast cancer discovered marker genes.

2.3.4 Relation between the survival signature proposed and the IHC markers

This section tries to answer a relevant question from a clinical point of view: Is the survival gene signature related to the standard clinical markers ? If this is true, the list of genes can be consider to estimate the value of the clinical status reducing the errors of current techniques. Besides, some genes may constitute alternative targets to IHC markers.

ER

		CLINICAL ER	
		T	F
KM BOOTSTRAP ER	T	150	31
	F	17	182

Table 2.7: Confusion Matrix ER clinical vs bootstrap.

Confusion Matrix and Statistics

Accuracy : 0.8737
95% CI : (0.836, 0.9054)

Sensitivity : 0.8545
Specificity : 0.8982

		CLINICAL ER	
		T	F
RISK PRED ER	T	196	10
	F	17	157

Table 2.8: Confusion Matrix ER clinical vs risk prediction.

Confusion Matrix and Statistics

Accuracy : 0.9289
95% CI : (0.8983, 0.9527)

Sensitivity : 0.9401
Specificity : 0.9202

ER Tab: 2.7 and 2.8 show, respectively, the confusion matrix for the risk groups obtained using the ER expression vs the IHC marker and the multivariate predictor considering our gene signature vs the IHC marker.
Notice that the multivariate gene signature proposed is able to predict the ER status much better than the expression of the corresponding gene. The accuracy is improved up to a 3% and both, sensitivity and specificity are significantly bigger.

PR

		CLINICAL PR	
		T	F
KM BOOTSTRAP PR	T	153	30
	F	37	160

Table 2.9: Confusion Matrix PR clinical vs bootstrap.

Confusion Matrix and Statistics

Accuracy : 0.8237
95% CI : (0.7816, 0.8607)

Sensitivity : 0.8421
Specificity : 0.8053

		CLINICAL PR	
		T	F
RISK PRED PR	T	175	28
	F	15	162

Table 2.10: Confusion Matrix PR clinical vs risk prediction.

Confusion Matrix and Statistics

Accuracy : 0.8868
95% CI : (0.8506, 0.9169)

Sensitivity : 0.8526
Specificity : 0.9211

PR Tab: 2.9 and 2.10 shows the same results as the previous tables. The multivariate method (section: 2.2.5) performs better.

HER2

		CLINICAL HER2	
		T	F
KM BOOTSTRAP HER2	T	221	25
	F	67	67

Table 2.11: Confusion Matrix HER2 clinical vs bootstrap.

Confusion Matrix and Statistics

Accuracy : 0.7579
95% CI : (0.7116, 0.8001)

Sensitivity : 0.7283
Specificity : 0.7674

		CLINICAL HER2	
		T	F
RISK PRED HER2	T	244	18
	F	44	74

Table 2.12: Confusion Matrix HER2 clinical vs risk prediction.

Confusion Matrix and Statistics

Accuracy : 0.8368
95% CI : (0.7958, 0.8726)

Sensitivity : 0.8043
Specificity : 0.8472

HER2 Tab: 2.11 and 2.12 shows a different distribution. In this case, the groups are deeply unbalanced. The KM with membership probability method (section: 2.2.4) is capable to improve the classification, but the multivariate method (section: 2.2.5) performs much better.

2.3.5 Survival genes discovered are related to relevant cancer biological functions

Several genes discovered are related to cancer or hormone receptors. In the following section,some of the marker genes we discovered and their effect and relation with cancer is explained.

ER and PR markers

CA12

Carbonic anhydrases (CAs) are a large family of zinc metalloenzymes that catalyse the reversible hydration of carbon dioxide. They participate in a variety of biological processes, including respiration, calcification, acid-base balance, bone resorption, and the formation of aqueous humor, cerebrospinal fluid, saliva, and gastric acid. This gene product is a type I membrane protein that is highly expressed in normal tissues, such as kidney, colon, and pancreas.

As a membrane protein, it is a good candidate for any kind of drug targeting and have already been reported (Kobayashi et al., 2012), but because it is overexpressed in a bunch of other tissues high toxicity is assumed if the cell functions through this path are altered.

CA12 also has relation with multidrug chemoresistance phenotype in cancer, which is closely related to survival (Kopecka et al., 2015). The relation with survival (Chien et al., 2012) (Yoo et al., 2010) and with breast cancer (Chen et al., 2018) has been reported too.

This gene havs been found to be closely related to **ER** expression in our studies, and it has been already reported (Barnett et al., 2008).

SUSD3

SUSD3 is a novel promoter of estrogen-dependent cell proliferation and regulator of cell-cell and cell-substrate interactions and migration in breast cancer. It may serve as a novel predictor of response to endocrine therapy and potential therapeutic target because it is located on cell surface (Moy et al., 2015) (Zhao et al., 2015).

As reported by literature it was found closely related to ER in our analysis.

SLC15A2(PEPT2)

SLC15A2 or PEPT2 is a proton-coupled peptide transporter that is responsible for the absorption of small peptides, as well as beta-lactam antibiotics and other peptide-like drugs, from the tubular filtrate.

As our analysis sugest, it has the strongest relation with survival from all our ER markers, which has been already reported (Lee et al., 2015). Also, a relationship with a different kind of cancer has been discovered (prostate (Tai et al., 2013)).

SLC39A6(ZIP6)

SLC39A6 or ZIP6 belongs to a subfamily of proteins that show structural characteristics of zinc transporters. Zinc is an essential cofactor for hundreds of enzymes. It is involved in protein, nucleic acid, carbohydrate, and lipid metabolism, as well as in the control of gene transcription, growth, development, and differentiation.

In our analysis it is more related to survival than to ER, the relation of this gene with survival has also been defined (Cheng et al., 2017), even it is trongly related to survival in breast cancer (Matsui et al., 2017). There are reports of its relation with cancer too (Lopez and Kelleher, 2010).

TBC1D9(MDR1, GRAMD9)

TBC1D9 is other of our best ER and risk related markers. This gene is not deeply studied and therefore is an excellent candidate to validate its role as ER related gene and survival marker. A study reports its relation with clinical outcome in gastric cancer and thus making this gene more interesting (Li et al., 2011).

NME3(NDPKC)

NME3 or NDPKC is a kinase which is higly expressed across all cancer types. In our study it shows the best capability of risk prediction in ER markers and strong relation with the ER marker.

This gene is related with colorectal cancer (Qu et al., 2013), DNA repair (Tsao et al., 2016) and neuroblastoma (Negroni et al., 2000). However, the relationship with Breast Cancer has not been described yet.

C6orf97(CCDC170)

As explained in NCBI database, the function of this gene and its encoded protein is not known. Several genome-wide association studies have implicated the region around this gene to be involved in breast cancer and bone mineral density, but no link to this specific gene has been found. A posible relation with the Golgi-Microtubule Network disruption in BRCA has been reported (Jiang et al., 2017), the same happens with survival (Hong et al., 2014) even with specifically ER positive BRCA (Veeraraghavan et al., 2014). In the other hand, the association between this gene and ER marker has also been described (Luo et al., 2014).

AGR3

This gene encodes a member of the disulfide isomerase (PDI) family of endoplasmic reticulum (ER) proteins that catalyze protein folding and thiol-disulfide interchange reactions. It is reported to be overexpressed in cancer (breast, ovarian, and prostrate).

In our analysis, this gene is one of the main ER-related marker, and that fact

also resembles in literature (Fletcher et al., 2003), being cited as survival-related (King et al., 2011) too.

HER2 markers

PNMT

PNMT (Phenylethanolamine N-methyltransferase) is related to HER2 marker. This gene is thought to play a key step in regulating epinephrine production. The product of this gene catalyses the last step of the catecholamine biosynthesis pathway, which methylates norepinephrine to form epinephrine (adrenaline).

PNMT is related with pheochromocytoma/paraganglioma (Lee et al., 2016), it is also related as mentioned in a study of ERBB2 amplicon (Benusiglio et al., 2006)

CWC25(CCDC49)

This gene encodes a factor that is part of the multi-protein C complex involved in pre-mRNA splicing. It is usually overexpressed in cancer and particularly in bone marrow related cancer. In our study, it shows the strongest risk prediction power, so it is one of the main survival related genes from our markers.

MIEN1(C17orf37, C35)

Overexpressed in all cancer types, MIEN1 is deeply related to cell viability, invasion, and migration of breast carcinoma cells (Che et al., 2017) (Zhao et al., 2017) (Dong et al., 2015). The relation with ERBB2 is also probed (Katz et al., 2010).

SIRT3

The encoded protein is found exclusively in mitochondria, where it can eliminate reactive oxygen species, inhibit apoptosis, and prevent the formation of cancer cells. SIRT3 has far-reaching effects on nuclear gene expression, cancer, cardiovascular disease, neuroprotection, aging, and metabolic control.

SIRT3 is related to breast cancer (Pinterić et al., 2018), pancreatic cancer, survival (Huang et al., 2017), and cancer in general (Yu et al., 2016)

GRB7

The product of this gene belongs to a small family of adapter proteins that are known to interact with a number of receptor tyrosine kinases and signaling molecules. This gene encodes a growth factor receptor-binding protein that interacts with epidermal growth factor receptor (EGFR) and ephrin receptors.

The developed algorithms (as already seen in other genes) identify relations between IHC markers and our markers, as described for GRB7 in this cases (Bivin et al., 2017) (Lim et al., 2014) showing the corelation with HER2, and with survival (Ramsey et al., 2011)

CRKRS(CDK12)

CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation (Krajewska et al., 2019). It is a potential novel biomarker for DNA damage response targeted therapies (Naidoo et al., 2018) and cell invasion (Tien et al., 2017) in Breast Cancer.

KLC4

Members of the kinesin-8 motor class have the remarkable ability to both walk towards microtubule plus-ends and depolymerise these ends on arrival, thereby regulating microtubule length (Peters et al., 2010). In cancer, genes that regulate microtubule functioning or are related with that machinery (like AURKA) have a huge impact in the probabilities of the generation of DNA failure when a cell is divided, which flows into more complex tumours, chemoresistance, metastasis...

2.4 Discussion

Most survival breast cancer studies rely on a small set of patients. Therefore, the results are not generalizable to the whole population and can not be transferred to the clinical practice.

In this chapter, we have built three gene expression datasets that integrate a large number of samples with survival and clinical data. They are based on affymetrics microarray and illumina RNAseq technologies. The biological results obtained with these datasets are robust and generalizable to the whole population.

One of the key points in breast cancer research is the discovery of survival cancer genes that allow to predict the risk and to stratifiy the patients. New molecular targets are needed to improve the treatments. Although a large variety of survival gene signatures have been proposed, the overlaping between them is small and they are inestable and sample dependent. Moreover, comercial platforms work as a black box and the decision making and the gene signatures considered can not be interpreted in terms of standard clinical markers.

In this chapter, a list of gene markers is proposed with the following properties: First, the experimental results have shown that the list of survival markers is robust, stable and generalizable to other groups of patients. The list of genes has small overlaping with oncotype or prosigna signatures, but the risk prediction and patient stratification is significantly better. The uncertainty for the classification of patients in two risk groups is reduced. Besides, the list of genes can be interpreted in terms of standard clinical markers managed by the clinitian and can be considered to reduce the errors in the estimation of IHC status. Finally, several survival markers are related to biological cancer functions relevant in cancer disease. This suggests that some markers should be tested in laboratory as alternative genes to standard markers.

Finally, the experimental results have shown that the bioinformatics methods

developed to predict the survival function and risk of patients are robust and stable. The method proposed in this chapter improve the performance of two widely used platforms such as Oncotype and Prosigna.

Chapter 3

Unravel positive markers and regulators of Triple Negative Breast Cancer (TNBC) using transcriptomic and regulatory profiling combined with survival analysis

3.1 Motivation

Breast Cancer (BRCA) is classified as Triple Negative (TNBC); when it does not show significant expression of the estrogen receptor (ER) or the progesterone receptor (PR) and does not overexpress human epidermal growth factor receptor 2 (HER2). The presence of these molecular markers is done by immunohistochemistry (IHC) and Fluorescence In Situ Hybridisation and has been shown to have significant inter-lab variability.

This is problematic as HER2+ or hormone receptor positive (HR+) samples with false-negatives via these analyses are at risk of being classified as TNBC, and corresponding patients would be given an incorrect prognosis and denied viable treatment.

Therefore, samples designated as TNBC could be subjected to a verification step to ensure that tumours have not been incorrectly classified as TNBC. As the "exclusionary"; definition of TNBC is the source of potential misclassification, it seems prudent that further classification of TNBC be based on an "inclusive"; criteria (i.e. "positive"; biomarkers).

To identify potential biomarkers, the novel bioinformatics tool, DECO (Decomposing heterogenous Cohorts using Omic data profiling) was used to identify 24 genes whose differential upregulation best characterizes TNBC and not hormone receptor positive (HR+) or HER2+ BRCAs. Identified biomarkers can be used for classification and prediction purposes.

Furthermore, Viper (Virtual Inference of Protein-activity by Enriched Regulon Analysis) was used to determine which transcription factors (TFs) are differentially more active in TNBC than in HR++ BRCA. Viper identified BCL11A and FOXC1 as potential drivers of TNBC. These TFs could serve as potential therapeutic targets for TNBC.

Also, as an addition, the relationship between the discovered markers and survival and risk will be analysed. This risk and survival result will provide relevance and importance to the markers.

3.2 Material and Methods

3.2.1 General workflow of the study

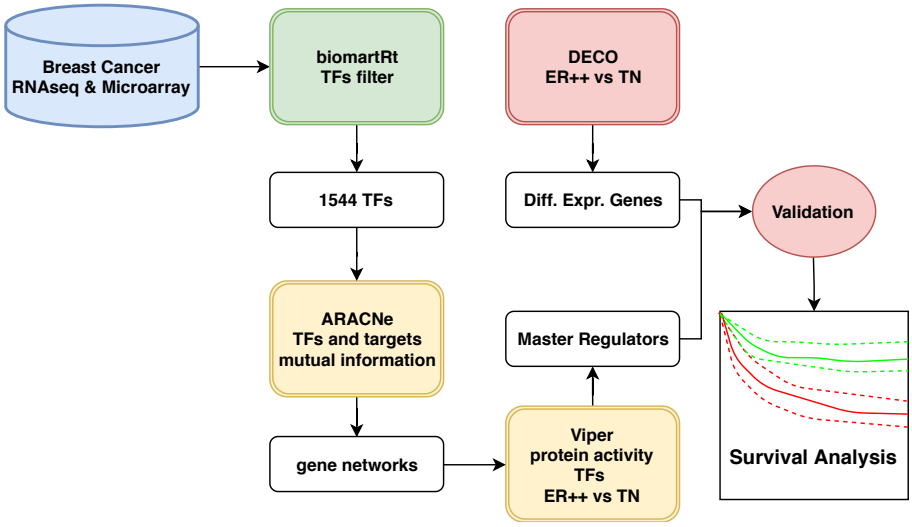


Figure 3.1: ER++ vs TNBC methodology and workflow.

In **Fig: 3.1**, the process is portrayed. First, the data from the compilation described in Chapter 2 microarray and RNAseq series. The first step is the filter to identify previously described TFs, it was performed using **biomartRt** (Durinck et al., 2005) (Durinck et al., 2009).

The selected TFs were provided to **ARACNe** (Margolin et al., 2006) (He et al., 2017) R package and the gene TFs and targets network was created. Then the **Viper** (Alvarez et al., 2016) R package was used in order to obtain the TF master regulators signature.

Furthermore, an independent DECO analysis was performed in order to select genes equivalent to TFs.

The markers relationship with survival was checked using the tools developed in Chapter 2.

3.2.2 Data used in the study

The datasets used for this study are described in the previous chapter in Section: 2.2.1. A 1024 microarray series, and a 879 RNAseq series which will refer henceforth as microarray series and RNAseq series respectively. RNAseq data allows the validation of the experimental results obtained with microarrays. Each technology requires the development of diferent preprocesing and normalisation techniques.

The dataset of methylation in BRCA available in TCGA (TCGA, 2019) was also downloaded following the same protocols and tools used for dowloading the RNAseq dataset. The methylation data was obtained for the same samples described in our phenodata.

3.2.3 Gene Interaction Analysis, TF Mapping and Differential Protein Activity Profiles

ARACNe

The **ARACNe** algorithm is used to perform a global gene correlation based on mutual information analysis using expression matrices (Margolin et al., 2006) (He et al., 2017). ARACNe-AP (Adaptive Partitioning) was used for the analysis (He et al., 2017). It required a list of TFs and RNA expression data as input.

Therefore, the **biomaRt** R package is used to obtain a list of 1,544 known TFs to supply the algorithm (Durinck et al., 2005) (Durinck et al., 2009). We subsequently generated a robust reverse engineered breast cancer gene network.

This interactome identified interactions between gene regulators (i.e., TFs) and gene targets and provided scores based on mutual information analysis that represented the strength of the interactions (He et al., 2017).

Viper

Viper (Alvarez et al., 2016) algorithm allows for a network-based inference of protein activity and identification of TFs that show a significant change in activity levels between two different phenotypes.ARACNe-AP produced gene networks was used and corresponding RNA expression data as input for Viper analysis.

We compared TNBC (tumours that had a clear ER-PR-HER- status determined by IHC) to the most common BRCA subclass, ER+PR+HER2- (HR++), in order to identify regulators of the triple negative phenotype. The analysis was done independently for both datasets.

For these comparisons, there were 113 TNBC samples and 148 HR++ samples used in the analysis of microarray series. In the analysis of RNAseq series, 150 samples comprised the TNBC class and 470 comprised the HR++ class.

3.2.4 Tumor Sample Categorisation and Differential Feature Analysis

DECO

DECO (Campos-Laborie et al., 2019) (developed by our group) identifies and categorises biomarkers that are most significantly associated to specific phenotypic conditions based on a Recurrent Differential Analysis (RDA) integrated with a Non-Symmetrical Correspondence Analysis (NSCA).

We utilised this novel bioinformatic tool on the data set with the best gene coverage (RNAseq series). Multiclass analysis was chosen and run on a total of 361 samples from the RNAseq series. The number of samples used in this analysis was reduced concerning the original dataset due to the subclasses chosen for this analysis.

However, the proportion of each tumour subtype is very similar in clinical practice. Only samples that were ER-PR-HER- (TN), ER-PR-HER+ (HER2+), or ER+PR+HER- (HR++) were considered. The number of tumour samples designated for analysis included in each of these classes was: for HR++ 271 samples; for HER2+ 22 samples; and for TNBC 68 samples.

3.3 Results

3.3.1 Finding TNBC Regulators by contrast with the Major Subtype ER+PR+

Our analyses with ARACNe and VIPER found 10 TFs to be differentially more active in the TNBC condition than in the HR++ condition utilising microarray series (FDR < 0.05). In RNAseq series, 26 TFs were identified as differentially more active in the TNBC condition (FDR < 0.05).

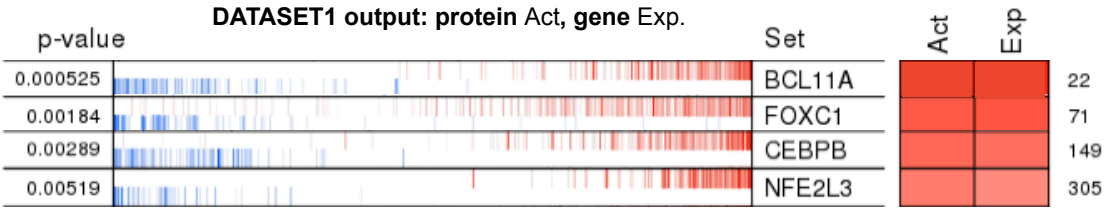


Figure 3.2: VIPER, output top TFs in the TNBC samples.

Microarray series produced a bipartite network composed of 1,243 regulators and 19,954 targets with 272,967 interactions; while the network created with RNAseq series was comprised of 1,355 regulators and 18,880 targets with 239,515 interactions.

As shown in **Fig: 3.2**, each row characterises one TF that VIPER inferred to be significantly more active in the TNBC condition in microarray series (FDR<.05). The regulon of each TF (set of gene targets) is represented by the barcode on the x-axis with each vertical line representing one target gene.

Genes located on the far-left end of the x-axis represent genes that were the most downregulated in the TNBC, while genes on the far right represent the genes that were most upregulated. Genes depicted in red reflect genes that are induced by the TF and genes depicted in blue represent those that the TF represses.

Differential RNA expression is represented in the far-right box and is represented by the depth of the red colouration. Inferred protein activity lies on the box directly to the left and is also represented by depth of red colouration.

DATASET1: RNA-microarrays (1024 samples)

Number	GeneSymbol	RegulonSize	NES(score)	p.value	FDR(adj.p.value)
1	KLF5	129	3.66	0.000249	0.00356
2	ELF5	72	3.54	0.000402	0.00503
3	BCL11A	452	3.47	0.000525	0.00564
4	PAX6	58	3.45	0.000564	0.00564
5	FOXC1	444	3.11	0.001840	0.01310
6	L3MBTL4	58	3.03	0.002440	0.01430
7	CEBPB	323	2.98	0.002890	0.01430
8	ZFP69B	275	2.80	0.005100	0.02260
9	NFE2L3	217	2.79	0.005190	0.02260
10	ARNTL2	405	2.67	0.007540	0.03140
11	NFIL3	386	2.45	0.0143	0.05140
12	E2F3	546	2.34	0.0191	0.05960
13	YBX3	229	2.31	0.0208	0.06300
14	CBFB	354	2.28	0.0225	0.06330
15	SOX11	352	2.23	0.0254	0.06690
16	NFIB	157	2.13	0.0335	0.08370
17	TP53	287	2.06	0.0394	0.09160
18	KLF11	219	1.98	0.0478	0.106
19	CEBPG	497	1.94	0.0519	0.109
20	PPARA	69	1.94	0.0523	0.109
21	TEAD4	208	1.92	0.0551	0.112
22	EN1	298	1.89	0.0589	0.118
23	ZBED4	481	1.86	0.0623	0.120
24	TCF7L1	234	1.85	0.0642	0.121
25	ZIC1	92	1.79	0.0729	0.127
26	SOX10	344	1.76	0.0788	0.127
27	SOX6	121	1.68	0.0936	0.140
28	GRHL3	101	1.66	0.0960	0.140
29	ZBTB24	367	1.65	0.0981	0.140
30	SOX9	184	1.64	0.1010	0.142
31	GATA6	63	1.59	0.1120	0.151
32	GRHL1	418	1.58	0.1150	0.153
33	TCF3	344	1.49	0.1350	0.170
34	HIVEP2	262	1.49	0.1360	0.170
35	ZNF391	62	1.46	0.1430	0.174
36	ZNF711	310	1.44	0.1490	0.179
37	POU4F1	127	1.37	0.1710	0.193

Figure 3.3: VIPER, differentially most active TFs in the TNBC samples as compared to the HR++ samples in microarray.

DATASET2: RNA-seq (879 samples)

Nº	Gene Symbol	Regulon Size	NES (score)	p.value	FDR (adj.p.value)
1	TLX1	71	5.20	0.0000002	0.000012
2	PRDM13	30	4.78	0.0000018	0.000051
3	ZIC1	52	4.75	0.0000020	0.000051
4	LIN28B	25	4.62	0.0000038	0.000060
5	SIX3	38	4.60	0.0000042	0.000060
6	ZIC4	36	4.42	0.0000097	0.000099
7	PDX1	39	4.32	0.0000153	0.000139
8	DMRT1	46	4.23	0.0000237	0.000198
9	NKX2-5	42	4.02	0.0000590	0.000421
10	RCOR2	166	3.89	0.0000987	0.000642
11	OLIG2	45	3.85	0.0001200	0.000693
12	LHX5	32	3.84	0.0001250	0.000693
13	NR2E1	62	3.45	0.0005520	0.002420
14	POU4F1	37	3.45	0.0005570	0.002420
15	ALX1	26	3.37	0.0007570	0.003030
16	TP53	311	3.29	0.0010200	0.003920
17	ATOH7	118	3.26	0.0011200	0.004130
18	YBX1	689	3.23	0.0012400	0.004290
19	SOX30	96	3.15	0.0016400	0.005290
20	RAX	65	2.93	0.0034300	0.010100
21	BCL11A	826	2.91	0.0036400	0.010200
22	GATA6	30	2.91	0.0036700	0.010200
23	HOXD13	28	2.75	0.0059900	0.015400
24	ZIC5	44	2.70	0.0068500	0.016700
25	FOXC1	1004	2.63	0.0086100	0.020000
26	POU5F1	105	2.38	0.0175000	0.036500
27	PPARD	262	1.98	0.0476	0.09160
28	OTX1	147	1.95	0.0517	0.09580
29	CTCFL	93	1.88	0.0597	0.105
30	CEBPB	269	1.70	0.0884	0.143
31	FOSL1	213	1.67	0.0948	0.144
32	CBFB	571	1.54	0.1230	0.174
33	TEAD4	207	1.51	0.1310	0.175
34	NFE2L3	292	1.48	0.1380	0.182
35	HMGA1	482	1.46	0.1440	0.187
36	E2F3	647	1.42	0.1550	0.195
37	NFIL3	190	1.41	0.1570	0.195
38	E2F4	357	1.41	0.1580	0.195

Figure 3.4: VIPER, differentially most active TFs in the TNBC samples as compared to the HR++ samples in RNAseq.

As shown in Figs: 3.3 and 3.4 The genes included in the table are the most significant found in microarray series 3.3 and in RNAseq series 3.4. Genes in common to two datasets are marked in yellow. Regulon size depicts how many genes are in the TF regulon (how many genes it regulates). The Normalised Enrichment Scores (NES) is a measure of differential activity. A positive NES denotes increased protein activity seen in the TNBC condition as compared to that of the HR++ condition, while negative denotes reduced.

B cell leukemia 11A (BCL11A) (FDR=0.0056 in microarray series, FDR=0.0102 in RNAseq series) and Forkhead Box C1 (FOXC1) (FDR=0.0131 in microarray series, FDR=0.02 in RNAseq series) were found to be significantly more active in the TNBC condition in both data sets.

Both data sets also agreed that the activity level of 12 TFs was significantly reduced in the TNBC condition as compared to that of the HR++ which is represented in the following table in **Tab: 3.1**. Among these TFs were ER, PR, MYB Proto-oncogene (MYB), and the Androgen Receptor (AR).

Regulon	Size	NES	p-value	FPDR
KLF5	129	3.66	0.000249	0.00356
ELF5	72	3.54	0.000402	0.00503
BCL11A	452	3.47	0.000525	0.00564
PAX6	58	3.45	0.000564	0.00564
FOXC1	444	3.11	0.00184	0.0131
L3MBTL4	58	3.03	0.00244	0.0143
CEBPB	323	2.98	0.00289	0.0143
ZFP69B	275	2.8	0.0051	0.0226
NFE2L3	217	2.79	0.00519	0.0226
ARNTL2	405	2.67	0.00754	0.0314
ZNF484	113	–2.53	0.0114	0.044
NR4A2	66	–2.66	0.00785	0.0314
AR	363	–2.97	0.003	0.0143
ZNF844	90	–2.98	0.00285	0.0143
BHLHE40	195	–2.99	0.00277	0.0143
TADA2B	297	–3.01	0.00263	0.0143
TOX3	40	–3.02	0.00249	0.0143
MYB	280	–3.25	0.00115	0.00887
ZNF442	63	–3.27	0.00109	0.00887
PGR	370	–3.42	0.000629	0.00572
XBP1	717	–3.69	0.00022	0.00356
FOXA1	585	–3.84	0.000121	0.00242
ESRI	457	–4.18	0.0000297	0.000742
GATA3	915	–4.31	0.0000165	0.000549
ZNF552	278	–4,54	0.00000551	0.000275
AFF3	460	–4.78	0.00000178	0.000178

Table 3.1: TFs with differential activity levels, TNBC vs ER+PR+ tissue in microarray series.

Differential activity levels are inferred by Normalised Enrichment Scores (NES). A positive NES denotes increased protein activity seen in the TNBC condition as compared to that of the HR++ condition. A negative NES denotes increased TF activity in the HR++ relative to the TNBC condition.

The absolute value of the NES depicts activity level, with a larger NES depicting a very active TF. TFs that were found among both datasets are highlighted in gray, tables in Tab: 3.1 and 3.2. There are 12 TFs that are shared among both dat sets that show increased activity in the HR++ condition compared to the TNBC.

Regulon	Size	NES	p-value	FPDR
TLX1	71	5.2	1.99E-07	0.0000121
PRDM13	30	4.78	0.00000176	0.000051
ZIC1	52	4.75	0.00000204	0.000051
LIN28B	25	4.62	0.00000378	0.0000604
SIX3	38	4.6	0.00000423	0.0000604
ZIC4	36	4.42	0.00000969	0.0000989
PDX1	39	4.32	0.0000153	0.000139
DMRT1	46	4.23	0.0000237	0.000198
NKX2–5	42	4.02	0.000059	0.000421
RCOR2	166	3.89	0.0000987	0.000642
OLIG2	45	3.85	0.00012	0.000693
LHX5	32	3.84	0.000125	0.000693
NR2E1	62	3.45	0.000552	0.00242
POU4F1	37	3.45	0.000557	0.00242
ALX1	26	3.37	0.000757	0.00303
TP53	311	3.29	0.00102	0.00392
ATOH7	118	3.26	0.00112	0.00413
YBX1	689	3.23	0.00124	0.00429
SOX30	96	3.15	0.00164	0.00529
RAX	65	2.93	0.00343	0.0101
BCL11A	826	2.91	0.00364	0.0102
GATA6	30	2.91	0.00367	0.0102
HOXD13	28	2.75	0.00599	0.0154
ZIC5	44	27	0.00685	0.0167
FOXC1	1004	2.63	0.00861	0.02
POU5F1	105	2.38	0.0175	0.0365
ZNF563	69	-2.26	0.0241	0.0472
RORC	249	-2.27	0.0232	0.0464
AR	527	-2.35	0.0189	0.0386
SALL2	282	-2.39	0.0168	0.0358
ZNEF396	397	-2.41	0.016	0.0348
MYB	260	-2.54	0.011	0.0244
FOXP1	278	-2.59	0.00955	0.0217
AFF3	661	-2.66	0.00778	0.0185
GCM1	168	-2.72	0.00645	0.0161
S0X13	106	-2.75	0.00595	0.0154
PAX2	76	-2.83	0.00462	0.0125
TADA2B	274	-2.99	0.0028	0.00849
LMX1B	304	-3.1	0.00192	0.00599
HOXB2	55	-3.18	0.00148	0.00492
PGR	521	-3.23	0.00122	0.00429
FOXN1	78	-3.44	0.000582	0.00243
TRERF1	90	-3.46	0.000536	0.00242
EMX1	69	-3.69	0.000226	0.00113
GATA3	1121	-3.74	0.000182	0.000958
ESR1	865	-3.88	0.000103	0.000642
XBP1	482	-4.08	0.0000443	0.000341
ZNF552	219	-4,42	0.00000989	0.0000989
FOXA1	1333	-4.52	0.00000614	0.0000768
ZNF442	111	-4.65	0.00000332	0.0000604
BHLHE40	84	-5.16	2.43E-07	0.0000121

Table 3.2: TFs with differential activity levels, TNBC vs ER+PR+ tissue in RNAseq series.

VIPER further interrogated all differentially active TFs to determine which TFs synergistically regulate the TNBC gene expression signature. Thus, TFs who shared a significant proportion of targets that were more enriched in the TNBC gene expression signature relative to their exclusive targets, were identified as "synergy pairs"; (Aytes et al., 2014). Using microarray series, 73 synergy pairs were identified among the 26 differentially active TFs (p-value$<$0.05). In RNAseq series, 160 synergy pairs were identified among the 51 differentially active TFs (p-value$<$0.05) (Figure2). Of note, 21 synergy pairs were seen in both data sets, including the following: BCL11A with FOXC1, BCL11A with AR, BCL11A with ER, and BCL11A with ER and GATA3.

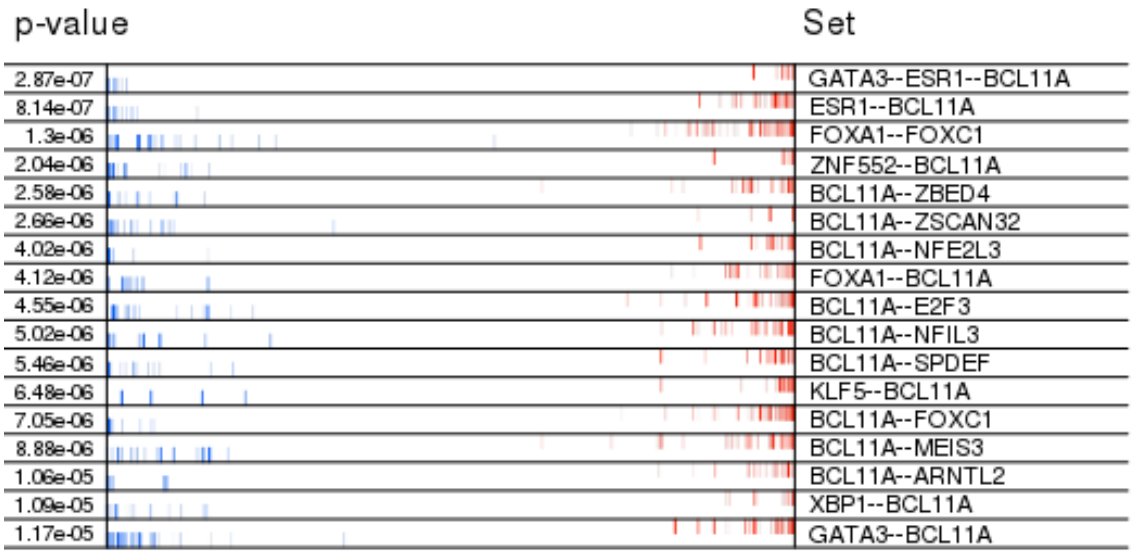


Figure 3.5: Viper output depicting the synergistic relationships among FOXC1 and BCL11A with other significant TFs in microarray series.

The 26 significant TFs from microarray series were evaluated for possible synergistic relationships amongst one another. A total of 73 pairs were found to exhibit significant synergy. As FOXC1 and BCL11A were found to be differentially more active in both data sets, the synergistic relationships between these TFs and other significant TFs found in microarray series is highlighted.

Each row in **Fig: 3.5** characterises one synergistic relationship with the p-value of the relationship found on the far left. The shared gene targets that are enriched in the TNBC gene expression signature are represented by the barcode on the x-axis with each vertical line representing one shared target gene. Genes located on the far-left end of the x-axis represent genes that were the most downregulated in the TNBC, while genes on the far right represent the genes that were most upregulated. Genes depicted in red reflect genes that are induced by the synergistic relationship and genes depicted in blue represent those that the synergistic relationship represses.

Differential RNA expression of the shared TFs is represented in the far-right box and is represented by the depth of the red colouration. The boxes appear gray

as these TFs work together to show increased protein activity relative to mRNA expression. Inferred protein activity lies on the box directly to the left and is also represented by depth of red colouration. All boxes appear in deep read as they show high levels of activity when working together.

3.3.2 Expression Profiling to find a Signature of Upregulated Genes in TNBC

DECO utilised a differential expression analysis to identify genes whose expression differed significantly between the HR++, HER2+, and TNBC subclasses. It then quantified the association between these genes and their membership in each one of the three subclasses. Finally, it measured the difference in expression between each sample and the mean expression for that gene.

The results of these analyses culminated in the development of an h-statistic for every differentially expressed gene in respects to each of the subclasses. This h-statistic represents how well a gene's expression level characterises a given subtype. The sign in front of the h-statistic signifies if the relative up-regulation (+) or down-regulation (-) of a gene marks a subclass. The larger the absolute value of the h-statistic, the greater a gene characterises a certain subtype.

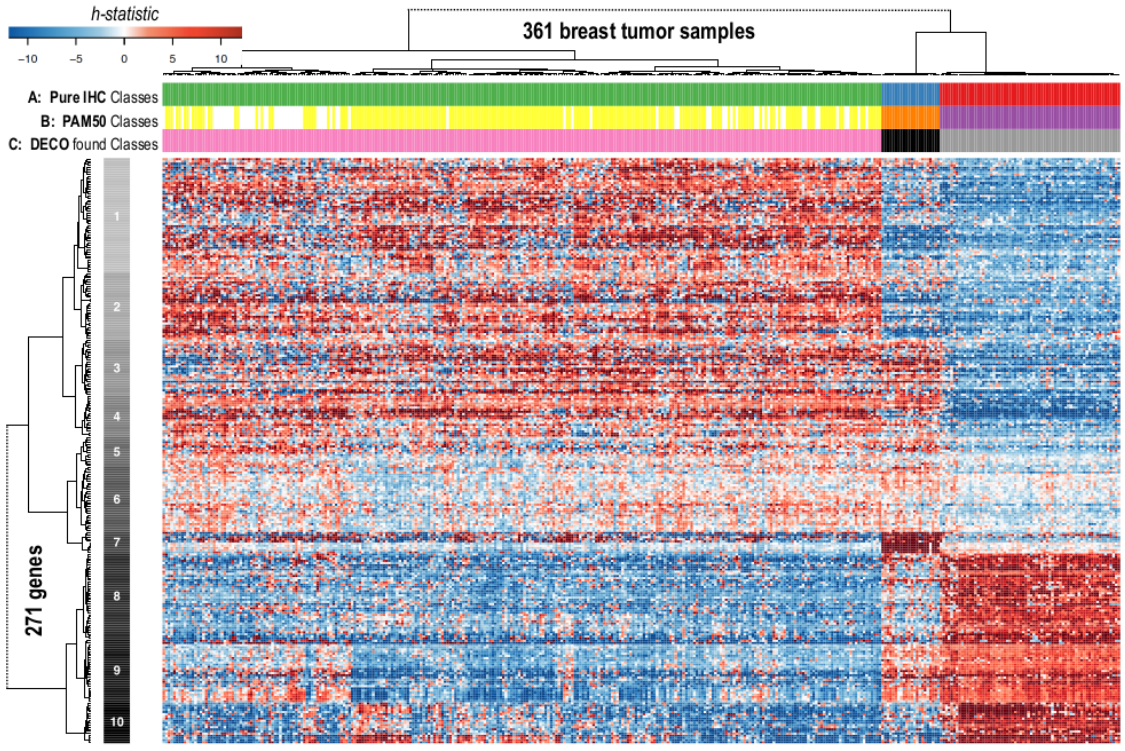


Figure 3.6: Heatmap clustering of 361 breast tumour samples derived from IHC subclass(HR++, TNBC, HER2+).

Since the transcriptomic omic data does not follow a normal distribution, the strength of DECO is that it outperforms other methods of comparative analysis when the data have minority changes (a subgroup of samples from one group is differenced from the rest and mark a class that otherwise would not be relevant). DECO is also capable of differentiate two groups that have changes in some of their samples, which other methods would not notice due to the medians being equal.

DECO analysis found 271 genes to be significant in the characterisation of HR++, HER2+, and TNBC. Hierarchical Bi-clustering of significant genes revealed 10 groups that exhibited a similar pattern of how h-statistic signal varied over all samples **Fig: 3.6**. The TNBC class showed increased signal in groups 8 and 10, while the ER+PR+ and HER2+ classes showed decreased signal.

We utilised the following input parameters for DECO analysis: RDA r = 5, combinations=9999, adjusted p.value < 0.01; NSCA variability explained = 90.197%, feature threshold=20 differential events in at least 5% samples.

The heatmap represents the h-statistic parameter, provided by DECO algorithm. The clustering identifies 10 groups of genes. The triple negative tumours (type 000) are characterised by increased signal by the groups of genes in 8, 9 and 10. Genes in group 8 and 10 show increased signal in TNBC and decreased signal in HR++ and HER2+ and could serve as viable "positive" biomarkers for TNBC.

Genes which showed a positive h-statistic for the TNBC subclass and a negative h-statistic for the HER2+, and HR++ subclasses were identified as potential "positive" biomarkers for TNBC as shown in Table: 3.3. Thirty-eight genes were identified as prospective biomarkers with 37 coming from groups 8 and 10, while 1 came from group 9.

Gene	HR++ BRCA	TNBC	HER2+ BRCA	Group
ROPN1	-6.35	15.39	-7.22	10
HORMAD1	-6.65	14.48	-1.65	8
ZIC1	-8.20	14.46	-0.28	8
GABRP	-5.81	14.35	-4.02	10
MSLN	-6.68	13.27	-3.66	8
MIA	-4.40	12.73	-10.23	10
ROPN1B	-5.99	12.68	-2.90	10
SOX10	-4.09	12.49	-10.56	10
TTYH1	-5.47	12.04	-1.44	10
UGT8	-5.21	12.03	-3.96	10
FOXC1	-3.82	11.54	-5.41	8
LEMD1	-5.49	11.53	-2.37	8
FBN3	-5.66	11.45	-1.93	8
PRSS33	-5.95	11.38	-1.32	8
MAPK4	-5.48	11.32	-2.58	10
SCRG1	-5.38	11.25	-3.51	10
ABCA13	-5.37	11.12	-2.48	10
CHODL	-5.48	10.99	-1.29	8
SFRP1	-4.06	10.99	-4.30	10
SLC6A15	-5.06	10.88	-4.52	10
DLX6	-4.95	10.03	-1.94	8
OPRK1	-5.39	9.89	-0.32	8
SHC4	-4.53	9.75	-3.02	10
RASGEF1C	-4.95	9.52	-1.47	8
KCNK5	-4.56	9.30	-1.90	8
GDF5	-3.79	8.91	-5.27	8
KCNQ4	-4.52	8.68	-0.73	8
RGMA	-3.82	8.66	-1.95	8
CHST4	-4.20	8.66	-2.38	10
FAM171A1	-3.82	8.11	-0.33	8
PM20D2	-3.91	8.07	-2.55	10
L3MBTL4	-3.57	7.80	-2.62	10
LDHB	-3.91	7.55	-0.04	9
TFCP2L1	-2.65	7.48	-14.59	10
SNX32	-3.43	7.00	-2.26	8
PRTFDC1	-2.97	6.44	-2.13	8
TAF4B	-3.29	6.31	-0.68	8
ANP32E	-3.19	5.95	-1.33	8

Table 3.3: "Positive" biomarkers found to characterise TNBC via DECO analysis.

These 38 genes show a positive h-statistic in TNBC and a negative h-statistic in HER2+ BRCA and HR++ BRCA. Dendrogram group affiliation is denoted on the far right. We then looked for genes which had a larger h-statistic in the TNBC subclass than ER showed for the HR++ subclass, which was h=9.39 (**Fig: 3.7**). The h-statistic of ER was chosen as the threshold of interest as ER has been both successfully targeted by therapeutics and evaluated by IHC and RNA-seq. 24 genes showed an h-statistic for TNBC greater than this landmark: ROPN1, HORMAD1, ZIC1, GABRP, MSLN, MIA, ROPN1B, SOX10, TTYH1, UGT8, FOXC1, LEMD1, FBN3, PRSS33, MAPK4, SCRG1, ABCA13, CHODL, SFRP1, SLC6A15, DLX6, OPRK1, SHC4, and RASGEF1C.

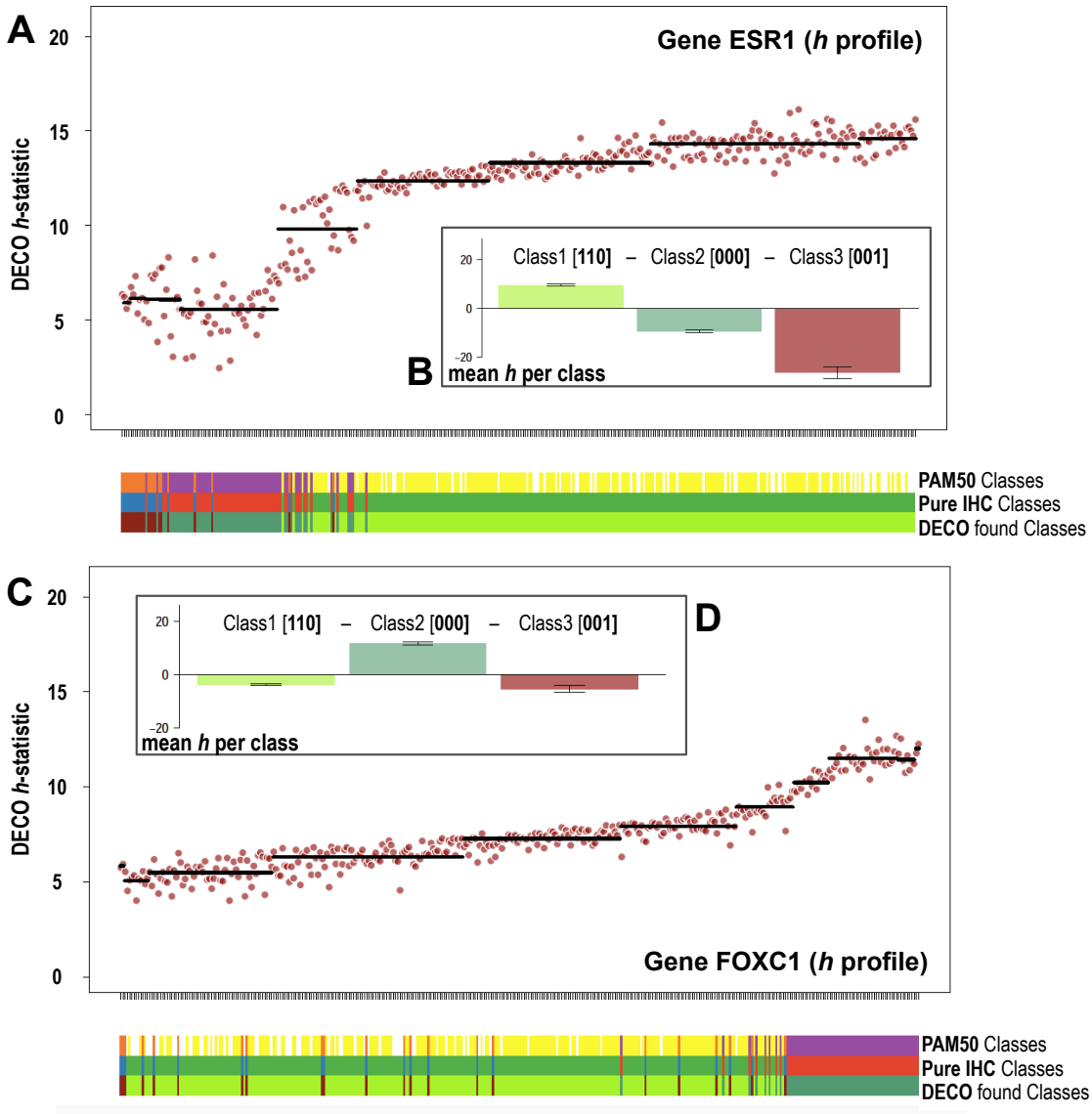


Figure 3.7: DECO output highlighting the distribution of the h-statistic for FOXC1 in comparison to ESR1 (ER).

Fig: 3.7 A) shows the plot of the h-statistic per sample for ESR1. The scale below designates samples in respects to "Pure IHC" with HR++ (green), TNBC (blue), and HER2+ (red) seen. The "PAM50" scale depicts the molecular subtypes with basal in red, Luminal A and B in green, and Her2 in blue. DECO found subclasses are the basis for hierarchical gene clustering and the heatmap derived from the h-statistic seen in **Fig: 3.6**. Subclass 1 coincides with HR++, subclass 2 with TNBC, and subclass 3 with HER2+. The bar graph in **Fig: 3.7 B)** pictures the directionality of the h-statistic for ESR1 in each of the DECO found subclasses.

Fig: 3.7 C) pictures a plot of the h-statistic per sample for FOXC1. FOXC1 is one of the 38 genes that showed a positive h-statistic in TNBC and a negative h-statistic in HER2+ BRCA and HR++ BRCA. It is a member of dendrogram group 8. The **Fig: 3.7 D)** bar graph pictures the directionality of the h-statistic for FOXC1 in each of the DECO found subclasses.

3.3.3 Methylation in TNBC

Using the data available in TCGA (TCGA, 2019) with the series of BRCA methylation, some analysis were done which started by the comparison between our two classes, the HR++ and TNBC.

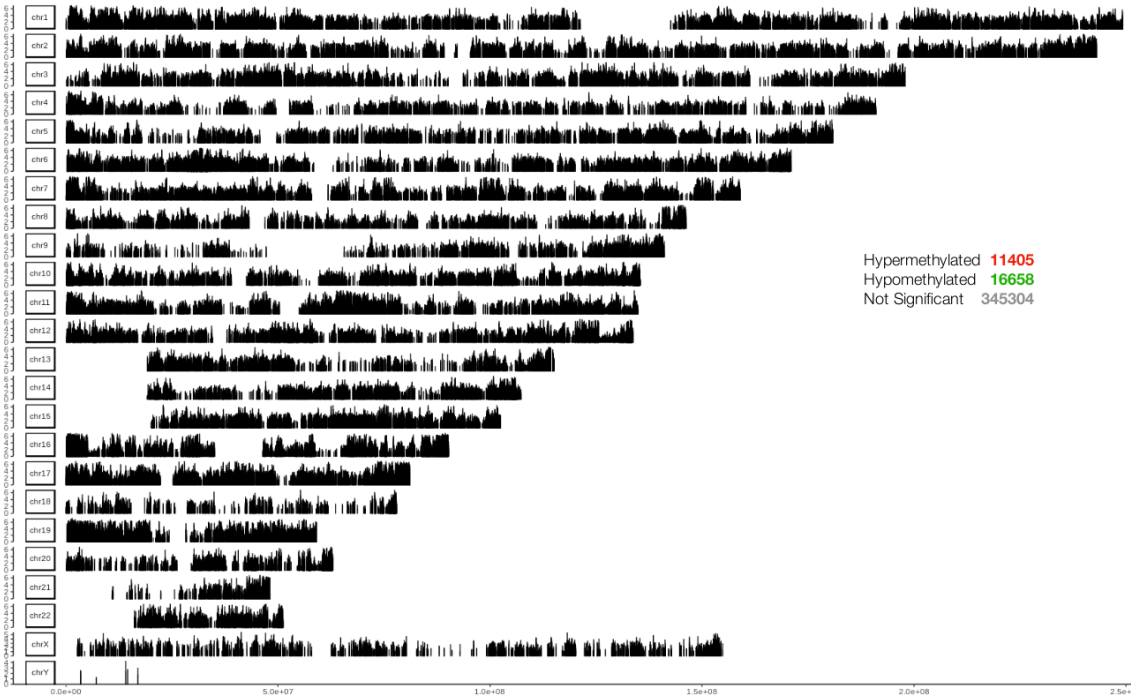


Figure 3.8: Distribution of methylated CpG islands in TNBC(000 patients) along hg19 chromosomes.

We detected hypermethylation in 11405 CpG islands (in red) and hypomethylation in 16658 (in green).

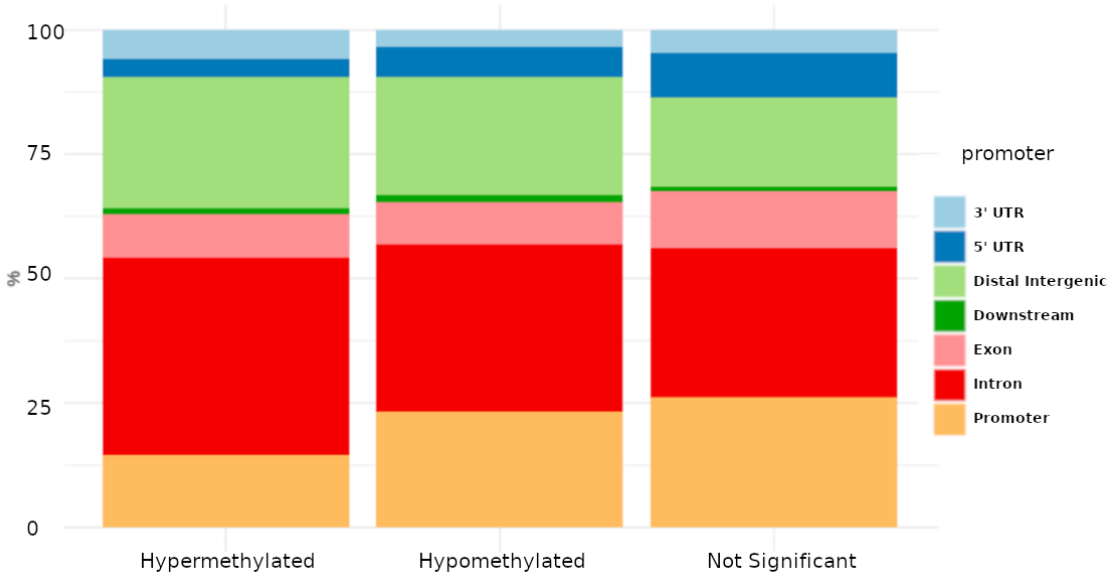


Figure 3.9: Hypermethylation and hypomethylation of promoters in TNBC data.

In general a higher count of hypomethylated CpG islands are found in out TNBC subset, the distribution of the islands is represented in **Fig: 3.9**, and shows that in hypermethylated regions there are more introns affected and less promoters affected than in hypomethylated regions.

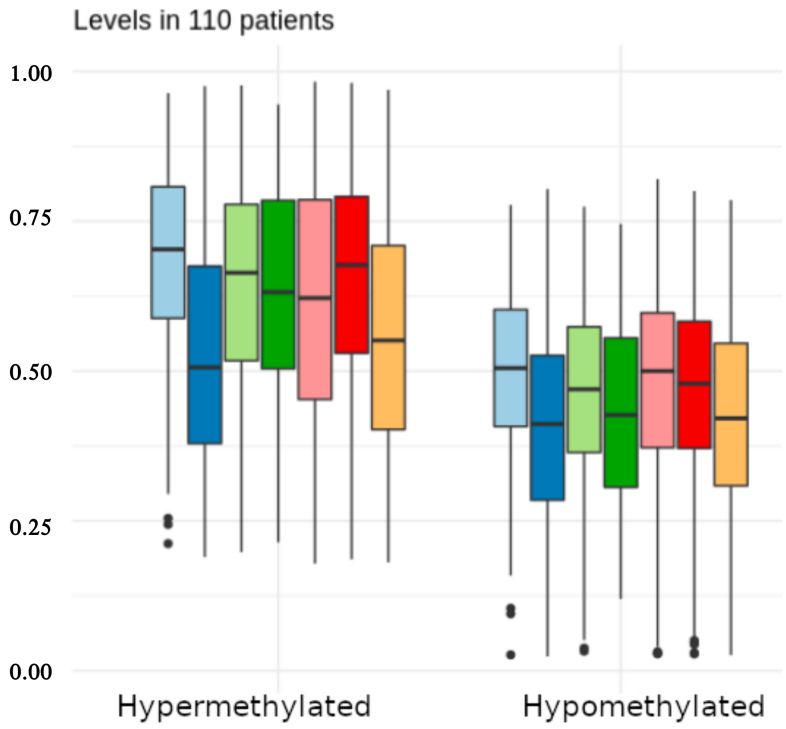


Figure 3.10: Hypermethylation and hypomethylation of promoters in TNBC(000) patients.

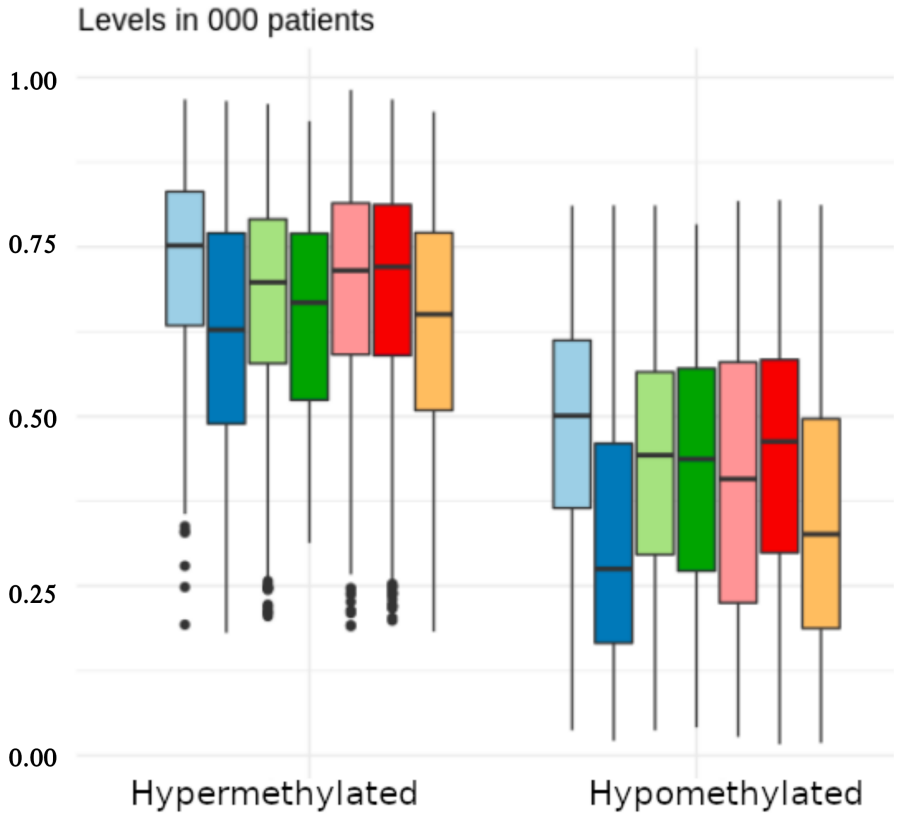


Figure 3.11: Hypermethylation and hypomethylation of promoters in HR++(110) patients.

As shown in **Fig: 3.10 and Fig: 3.11** global DNA methylation levels in TNBC is slightly lower in comparison with HR++ patients. Hypermethylation is higher in 000 patients and hypomethylation variation is less significant. The level of global methylation on CpG regions or promoter regions is very low in the TNBC patients which might be telling us that hypermethylation events are specific of 110 patients.

The global signal is lower than the signal for TNBC, and the signal for HR++ is the higher than both TNBC and global signal. The differences between both classes may suggest that in the case of triple negative BRCA patients there are more regions in the genome that are hypomethylated.

DNA hypomethylation in cancer has grown to be more and more important (Ehrlich, 2009), that epigenetic abnormality was often ignored but it is getting relevance again. DNA global methylation is reported to be lower in all tissues of cancer patients compared to control, and in some cases high hypomethylation in tumours (Kankava et al., 2019). Recent high-resolution genome-wide studies confirm that DNA hypomethylation is the almost constant companion to hypermethylation of the genome in cancer.

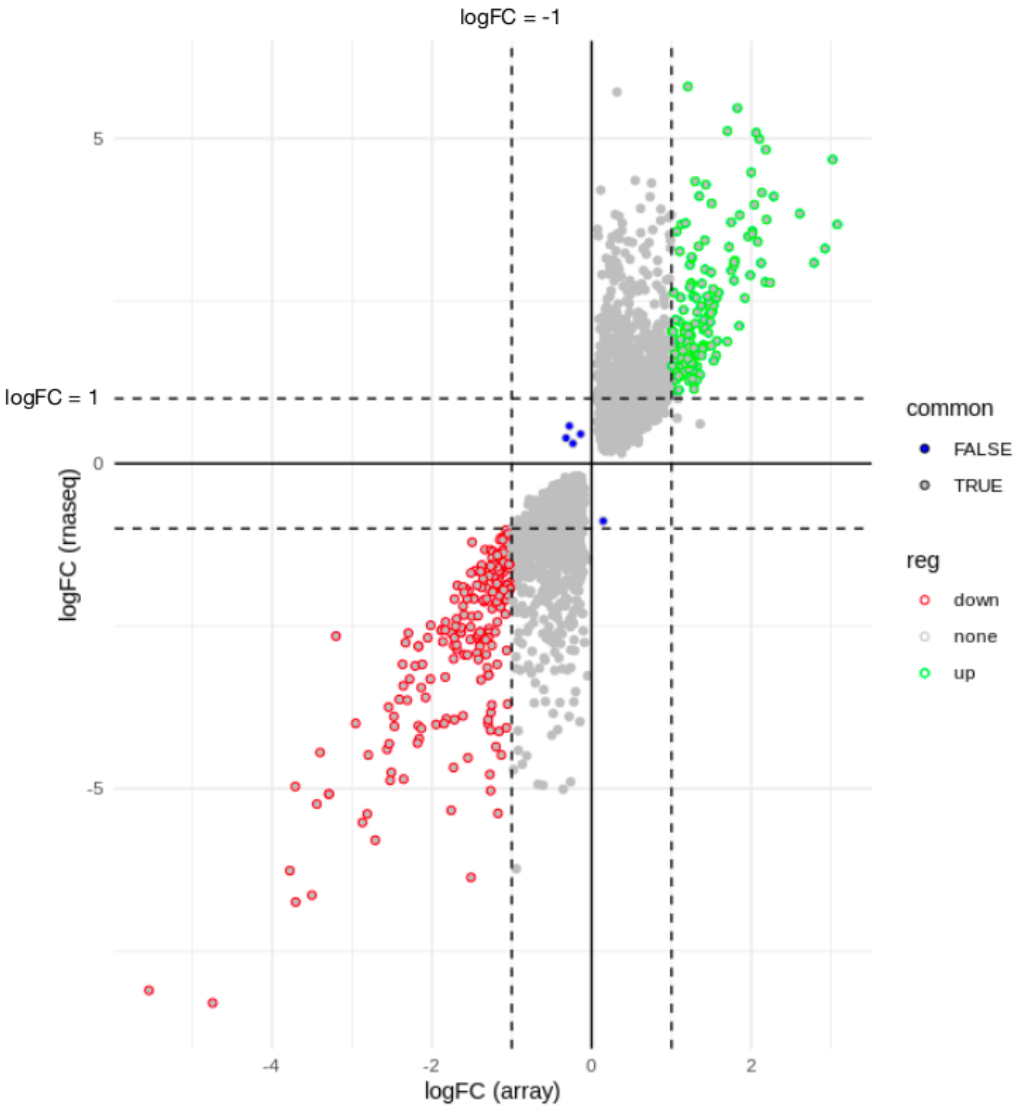


Figure 3.12: Differentially expressed genes with annotated promoters.

Concordance between the DECO discovered Master Regulators in microarray series and RNAseq series is almost total. In 3.12 the genes that dont overlap in the comparison are shown as blue dots. The data of the overlap is represented in the following table **Tab:** 3.4.

		RNAseq	
		T	F
Array	T	15766	2852
	F	1427	4120

Table 3.4: Confusion Matrix DECO microarray vs RNAseq.

Confusion Matrix and Statistics

Accuracy : 0.8229
95% CI : (0.8181, 0.8277)

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9170
Specificity : 0.5909
Pos Pred Value : 0.8468
Neg Pred Value : 0.7427

The accuracy, sensitivity, and positive and negative predicted value are strongly significative. In conclusion the capability of DECO to generate reproducible results is quite strong even for different platforms. Following this conclusion, the robustness of our proposed Master Regulators is guaranteed.

3.3.4 Risk and Survival analysis

This section analysis use the robust risk and survival tools developed as main part of this thesis(described in Chapter 2).

Multivariate risk prediction using TNBC markers

In this section, an analysis is done using the multivariate risk prediction algorithm defined in Chapter 2.

The genes discovered as TFs and TNBC marker genes are used in the multivariate approach to predict risk.

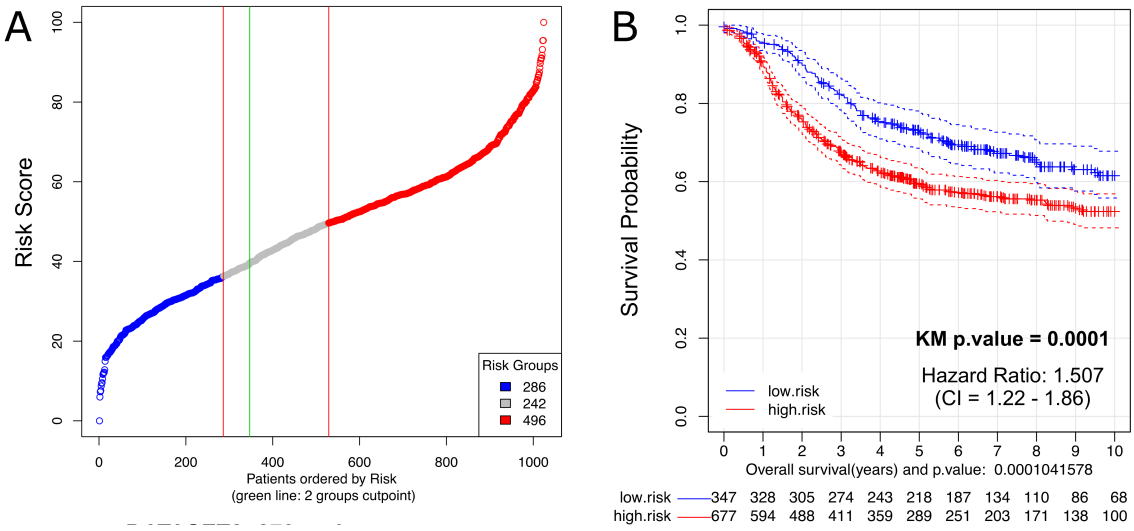


Figure 3.13: Risk predicion and Kaplan-Meier curves TNBC markers, microarrays.

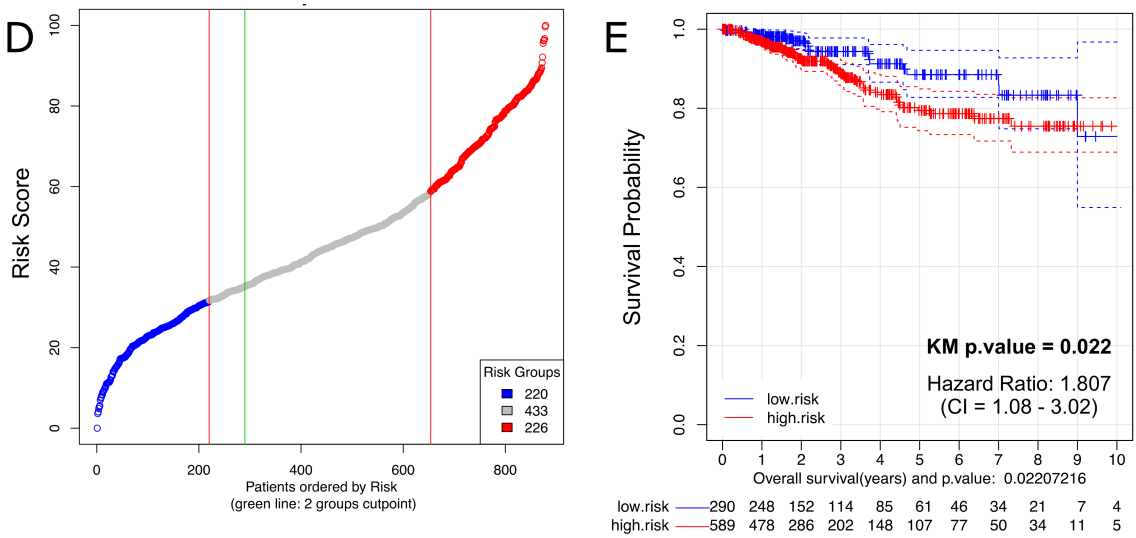


Figure 3.14: Risk predicion and Kaplan-Meier curves TNBC markers, RNAseq.

In **Figs: 3.13 and 3.14**, the analysis corresponding to microarray (**A** and **B**) and RNAseq (**C** and **D**) is pictured. The ordered risk curve in **A** divide by the green line the high and low risk patients from microarray series. These patients are represented in high and low risk groups in Kaplan Meier curves in **B**.

Besides, **C** displays the risk ordered patients from RNAseq series. As before, **D** displays the Kaplan-Meier curves for high and low risk groups.

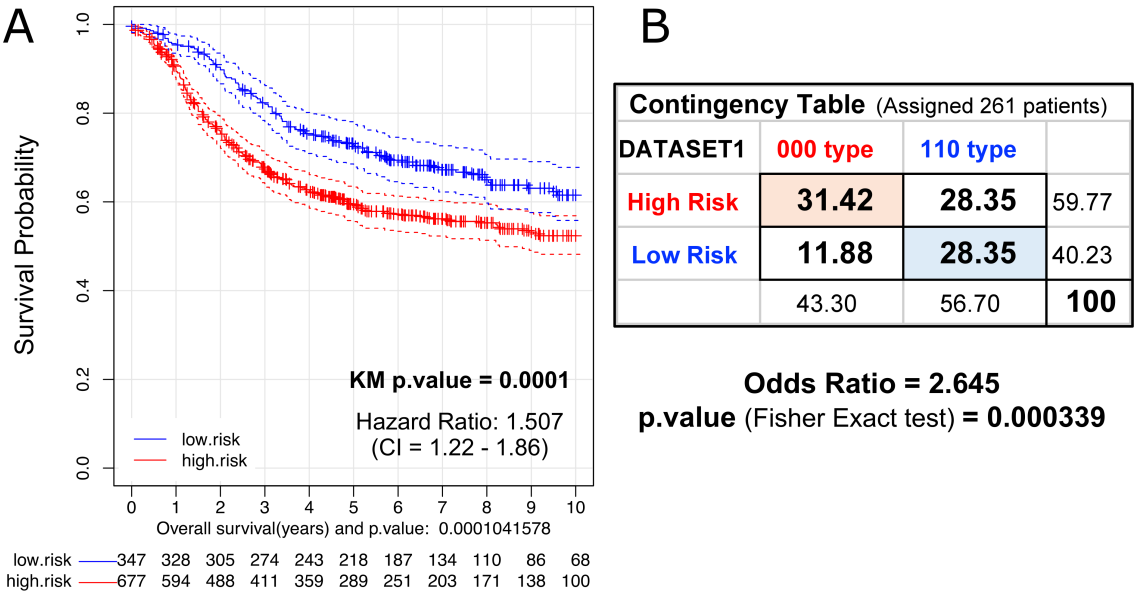
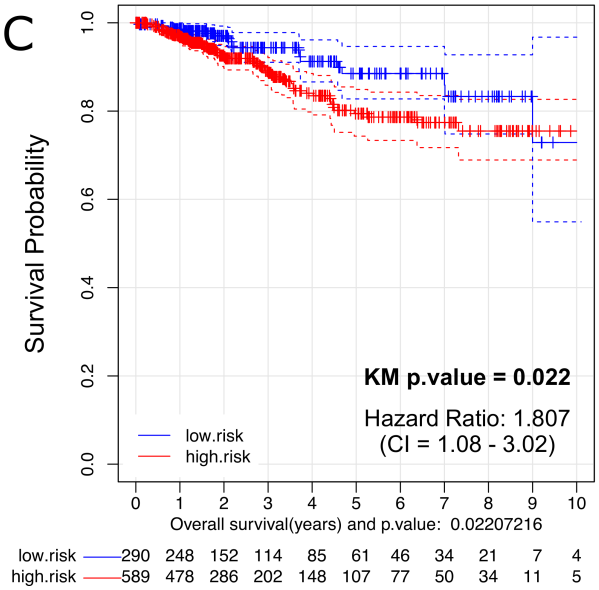


Figure 3.15: Contingency tables and Kaplan-Meier curves TNBC markers, microarrays.



D

Contingency Table (Assigned 620 patients)			
DATASET2	000 type	110 type	
High Risk	23.55	39.52	63.06
Low Risk	0.65	36.29	36.94
	24.19	75.81	100

Odds Ratio = 33.520
p.value (Fisher Exact test) **= 3.931e-23**

Figure 3.16: Contingency tables and Kaplan-Meier curves TNBC markers, RNAseq.

In **Figs: 3.15 and 3.16**, the **A** and **C** plots represent the same curves from previous figure to clarify the information in contingency tables. Contingency table **B** and **D** shows the coincidences between the group defined as TNBC(or 000) and ER++(or 110) and the risk groups computed and labeled as low or high risk. As said before, **B** corresponds with microarray data and **D** with RNAseq. The tables show the relationship between subtype TNBC and bad prognosis and risk, showing also the relationship between low risk and prognosis and ER++ subtype.

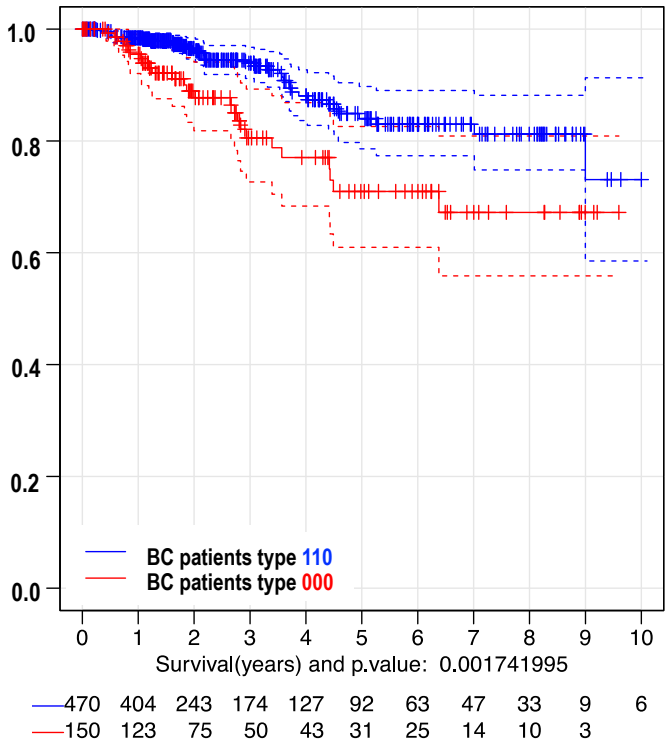


Figure 3.17: Kaplan-Meier curves in RNAseq dataset. 110 vs 000 subtypes.

In **Fig: 3.17**, the difference in survival from TNBC group and ER++ group is represented with a significant separability.

3.4 Discussion

BC tumour samples are currently evaluated by IHC and FISH to determine the presence of the three most critical and well-identified BRCA molecular markers: ER, PR, and HER2. When a sample does not show significant expression of any of these markers it is given the label of TNBC. Thus, HER2+ or HR++ samples that show false-negatives via these methods of analysis are at risk of being classified as TNBC and given an incorrect prognosis and denied viable treatment. It is therefore imperative that tissue samples be classified correctly.

We propose that samples designated as TNBC could be subjected to a verification step to ensure that tumours have not been incorrectly classified as TNBC. As the "exclusionary"; definition of TNBC is the source of potential misclassification, it seems prudent that further classification of TNBC be based on an "inclusive" criterion. The 24 genes identified to best classify TNBC could serve as potential biomarkers for this verification step. Each of these genes was shown to have differentially upregulated mRNA levels in TNBC as opposed to HER2+ or HR++ BRCA and each has been identified to better characterise TNBC than ER does for HR++ BRCA.

While RNA expression levels allow us to better categorise TNBC, RNA expression analysis does not measure protein activity and thus cannot directly tell us which proteins are playing a dominate role in the landscape of TNBC. In order to address this issue, Viper was used to evaluate differential protein activity in TNBC and HR++. Viper identified 12 TFs that were differentially more active in the HR++ samples as compared to the TNBC samples. Among the 12 TFs were ER and PR, which is to be expected based on the underlying comparison. These findings provide support to the efficacy of the methods utilised and provide justification for the exploration of other TFs on the list. As ER and PR have been successfully targeted with pharmaceuticals in HR++, it seems logical that TFs that were identified as differentially more active in the TNBC could serve as potential targets for pharmaceutical inhibition in TNBC. Ultimately, Viper identified 34 TFs with differentially increased activity in TNBC. BCL11A and FOXC1 were the focus because they were identified in both data sets.

BCL11A a BRCA cancer related transcriptional repressor

BCL11A encodes for a zinc finger TF that functions as a transcriptional repressor. BCL11A is also a proto-oncogene for hematologic cancers and is a proposed biomarker for non-small cell tumours of the lung (Jiang et al., 2013) (Weniger et al., 2006) (Nakamura et al., 2000). Its expression is essential for proper B cell and T cell development and has been found in low levels in the thymus, bone marrow,

and lymph nodes but in high levels in germinal center B cells and the fetal brain (Satterwhite et al., 2001) (Liu et al., 2003). BCL11A has also recently been found to be a regulator of normal mammary gland development and is required for the development of both mammary stem cells and luminal progenitor cells. However, high levels of BCL11A expression have been shown to promote tumourigenesis in TNBC and is negatively correlated with survival (Khaled et al., 2015).

BCL11A is one of the 271 genes that DECO identified as significant in the characterisation of HR++, HER2+, and TNBC BRCAs. It is part of dendrogram group 10 and is differentially upregulated in TNBC. BCL11A effects the TNBC phenotype not only by its increased activity but also through synergistic relationships. It takes part in at least 4 synergistic relationships to yield greater enrichment of its shared gene targets in the TNBC gene expression signature.

FOXC1, downregulated in both HER2+ and HR++ BRCA

FOXC1 was identified as the 7th greatest gene to exhibit differential expression among HR++, HER2+, and TNBC. It followed ER, which showed the greatest differential expression and ERBB2, which showed the 3rd greatest. FOXC1 is part of dendrogram group 8 and is differentially upregulated in TNBC and differentially downregulated in both HER2+ and HR++ BRCAs. The increased expression of FOXC1 is has been shown to be characteristic of TNBC, with an h-statistic of 11.54 (**Fig: 3.7 C & D**).

The FOX family of TFs are characterised by their Forkhead domain, a 100 amino-acid DNA-binding domain that has been conserved throughout evolution (Hannenhalli and Kaestner, 2009). Members of the FOX family play diverse roles in organogenesis, cell cycle regulation, and cell differentiation (Tuteja and Kaestner, 2007a) (Tuteja and Kaestner, 2007b). FOXC1 specifically has been associated with a number of cancers, including Hodgkin's Lymphoma, non-Hodgkin's Lymphoma, hepatocellular carcinoma, endometrial cancer, and breast cancer (Elian et al., 2018).

In TNBC, FOXC1 promotes metastatic change at least in part by its ability to activate chemokine receptor-4 (CXCR4), a well-known promoter of metastasis (Pan et al., 2018). In BLBC (a molecular subtype of BRCA which is made up of roughly 60-90% by TNBC), BRCA1 and GATA3 have been found to collectively repress FOXC1. Increased FOXC1 expression in BLBCs has also been associated with drug resistance, the epithelial-to-mesenchymal transition, and loss of E-Cadherin. Furthermore, FOXC1 has been found to have additional roles in the maintenance and proliferation of BLBC (Tkocz et al., 2012).

The transcription factors and their potential

We have shown that FOXC1 and BCL11A exert their influence on the TNBC phenotype through upregulated expression, increased TF activity, and synergistic relationships. These findings have led us to postulate that FOXC1 and BCL11A may be drivers of TNBC. The hypothesis is that these TFs could serve as viable targets

for the treatment of TNBC. Additional studies should be conducted using murine models and cell lines to further evaluate the role of putative drivers and their potential use in the targeted therapy of TNBC.

The relationship between discovered TNBC markers and risk prediction power and survival is a great addition to the study. The value that this relationship incorporates to the markers is triple. First, the possibility to compute and predict risk for new patients in clinic while the pertenency to a subgroup is evaluated. Second, the posibility of being used as factors that define the barely known triple negative BRCA subgroup. Finally, the proposed markers could be investigated to define subgroups inside the TNBC.

In conclusion, our work has allowed us to identify positive biomarkers in TNBC. These biomarkers could serve to confirm that samples designated as TNBC are truly TNBC and not HER2+ of HR++ BRCAs that were given false-negatives via FISH or IHC. Additionally, our work allowed for the identification of potential tumour drivers of TNBC. These potential targets should be investigated further to determine if their increased activity is in fact driving TNBC and if targeting these TFs would be a viable therapy in the treatment of TNBC.

Chapter 4

Survival marker genes of ColoRectal Cancer (CRC) derived from integration and meta-analysis of multiple transcriptomic datasets

4.1 Motivation

Colorectal cancer (CRC) is one of the most frequent tumours that causes great morbidity worldwide. It is the third most common cancer in men, the second most common cancer in women and the third leading cause of global cancer mortality (https://www.wcrf.org/).

CRC is a heterogeneous disease since from one patient to another it differs in clinical presentation, molecular characteristics, and prognosis (Linnekamp et al., 2015). The heterogeneity of CRC increases the complexity of this tumoural pathology, making subtyping and stratification a difficult task for therapeutic decisions.

In this way, personalised medicine for CRC is becoming increasingly needed, especially for targeted therapies where large variations between individual's treatment responses exist (Linnekamp et al., 2015) (Dienstmann et al., 2017). In this context, the need to find robust gene markers associated with specific subtypes of CRC led us to this study.

Furthermore, the specific purpose of our work was to find consistent biomolecular targets that, together to facilitate samples stratification, could be related to the prognosis of the disease using survival data.

The genomic and transcriptomic profiling of human cancer samples has been demonstrated over the last decade as an excellent way to obtain a better molecular characterisation of many tumour types and subtypes. While gene expression-based CRC classifications has been heavily approached (Dienstmann et al., 2017), little consensus in CRC standalone gene bio-marking has been achieved.

In fact, several studies have identified a broad variety of gene sets as gene expression profiles for classification and categorisation of this malignant disorder (Liu et al., 2017) (Guinney et al., 2015). Moreover, several transcriptomic-based tests oriented towards prognosis have also been investigated.

Some examples of these are: ColoLipidGene (Vargas et al., 2015), ColoGuidePro (Sveen et al., 2012) or ColoPrint (Kopetz et al., 2015); that include gene signatures associated with CRC survival in some specific biological contexts. Despite these efforts, at present there is not a clear compendium of gene markers for CRC survival and it is quite difficult to find consistency in the literature.

In the clinic, patients are classified into four CRC stages based in the anatomopathologycal characteristics of their tumours. It is common to use the TNM Staging System (where T stands for tumour, N for lymph node, and M for metastasis). The disease "staging" also allows grouping the patients in 4 progressive cancer stages, indicated by roman numerals: I, II, III, and IV (Society, 2017).

In this way, stages I and II correspond to cases which had not shown cancer cells beyond the tumour or blood. By contrast, stages III and IV correspond to individuals in where the cancer had diseminate to the lymph system or other organs in the body. This four stage categorisation represents significantly distinctive patients groups for final outcome or disease relapse, but the stages do not predict the risk of each individual patient because they are not directly associated to survival (Tauriello and Batlle, 2016).

Based on the described need and potential benefits to find survival marker genes correlated with high risk and poor prognosis in CRC; global gene expression profiles of colorectal tumours and its alteration throughout stages is investigated, to identify genes that could be levered as biomarkers of survival and prognosis for CRC in late stages (i.e., III and IV). To undertake this work we performed a deep analysis on a large cohort of human samples derived from a robust integration of several datasets that had transcriptomic and clinical survival data.

The integration provided a homogeneous and well-standardised meta-dataset that includes 1273 human colorectal samples. The identification of candidate markers was performed using an initial contrast between the gene expression of the subset of patients with CRC allocated by their clinical features to stages I and II versus the patients with tumours corresponding to stages III and IV.

Finally, after internal and external cross-validation, the genes selected as best survival markers were used to construct a risk predictor to allow stratification of the patients with respect to their relative risk.

4.2 Material and Methods

4.2.1 General workflow of the study

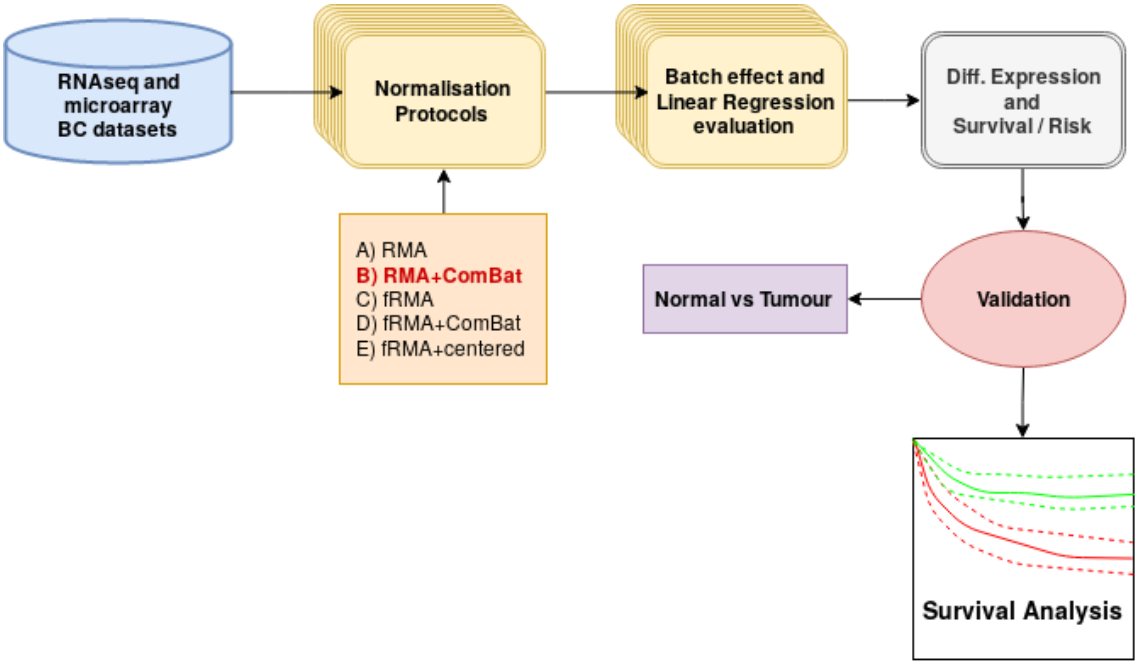


Figure 4.1: CRC methodology and workflow of the study.

In **Fig: 4.1**, the followed process is portrayed. First, the dataset is merged an normalised using five different protocols. The choice of the best performing method is further explained. **RMA + ComBat** method is selected, using this method the batch effect was eliminated.

Differential expression analysis was performed as described in Chapter 2 using **Limma** this time method. This provided a list of relevant genes which were further analysed using the survival tools described in the same chapter. The final list of markers is validated in two independent studies.

The first one, evaluates the multivariate relationship of the genes with risk, which is not analysed in previous step.

The second one, compares tumour samples and versus CRC samples. Colorectal samples from the series have higher expression than normal samples for our UP defined marker genes.

Moreover, the DOWN defined marker genes are downregulated in CRC samples.

4.2.2 Genome-wide expression data sets

In this study, seven data sets of CRC samples were analysed and integrated (Tab: 4.1).

GEO dataset	Description	Inital samples	PMID	Reference	Discarded samples	Final samples
GSE14333	primary colorectal	290	19996206	Jorissen RN et al. (2009)	64	226
GSE17536	colorectal	177	19914252	Smith JJ et al. (2010)	0	177
GSE31595	stage II and III colorectal	37	–	Thorsteinsson M et al. (2011)	0	37
GSE33113	stage II colorectal	90	22496204	Kemper K et al. (2012)	0	90
GSE38832	colorectal	122	25320007	Tripathi MK et al. (2014)	0	122
GSE39084	primary colorectal	70	25083765	Kirzin S et al. (2014)	1	69
GSE39582	colorectal	566	23700391	Marisa L et al. (2013)	14	552
Total number		**1352**				**1273**

Table 4.1: Series summary of colorectal cancer (CRC) samples integrated in the data set.

All data sets are available at GEO repository, corresponding to 7 series with the following accession numbers: GSE14333, GSE17536, GSE31595, GSE33113, GSE38832, GSE39084 and GSE39582. All these series included the raw expression signal and correspond to data obtained with the microarrays expression platform: Affymetrix GeneChip U133 Plus 2.0 for Homo sapiens.

The phenotypic information corresponding to all these series was analysed in order to select only the samples that included information regarding: the cancer stage and the Overall Survival (OS).

The samples that did not have any survival information were discarded from the study. In all cases only primary tumours samples were considered for our analysis; in this way individuals who had received preoperative chemotherapy and/or radiotherapy were also discarded.

For the external validation two independent datasets were used. A cohort of 276 colorectal carcinomas that had been studied using RNA-seq gene expression profiling, and that had survival data for 269 samples (Muzny et al., 2012). A second cohort of CRC samples from the platform SurvExpress (Aguirre-Gamboa et al., 2013).

This second dataset selected, called "Colon-Metabase-Uniformised", included 482 CRC samples with overall survival data and genome-wide expression determined with Affymetrix microarrays.

4.2.3 Expression data sets exploration and integrative normalisation

Previously, to make the best use of the information obtained from the microarrays, the importance to ascertain the quality of the data is considered. To assess the validity of generated microarray information a wide variety of quality assessment methods is performed, both in raw and pre-processed information. In this way, several explanatory data analysis were applied for the detection of problematic arrays.

The R function image was used to create chip images of the raw intensities to discover spatial artefacts in the samples. The distribution of probes intensities across all arrays were checked, using the boxplot method available for the Affybatch class. The Normalised Unscaled Standard Error (NUSE) (McCall et al., 2011) algorithm was applied to the samples. This quality assessment tool requires a previous PLM fitting procedure applied on the raw expression data. The function fitPLM provided in the affyPLM (Brettschneider et al., 2007) package was used to create the PLMset class object used as the input in the elaboration of the NUSE analysis. After applying the referred quality assessment methods, 79 of the initial samples collected were discarded and proceeded with the remaining 1273 (Tab: 4.1).

To create a table with all the phenotypic characteristics of the patients selected which involved all samples GSM accession numbers and related clinic variables in a consistent and homogenize way, getGEO and pData functions from GEOquery package were used. Regular expressions and common text manipulation R functions were used to solve the issue of formatting heterogenic data. Finally, a binary variable was created to label the patients and select them in a proper way during the hypothesis contrasts and statistical modeling.

4.2.4 Batch effect removal

Batch effect is one of the main problems when several datasets are combined to be studied together, because different batches usually add large unwanted variability to the data. To avoid this effect a combination of different pre-processing and normalisation algorithms was tested: Robust Multi-array Average **(RMA)** algorithm (Gautier et al., 2004); Combatting Batch effects **(ComBat)** algorithm from inSilicoMerging package; Frozen Robust Multi-array Average **(fRMA)** algorithm (McCall et al., 2010a). For the **fRMA** algorithm application, the frozen parameter vector was constructed using a training dataset in where we distributed randomly selected samples proportionally to each labelled group to obtain a balanced sample from the 7 batches of microarrays.

Another important issue addressed was the fact that the Affymetrix probe-sets included in the expression microarrays many times do not correspond to singular genes and some probes inserted in the defined probe-sets are ambiguous or inaccurate (Risueno et al., 2010). Affymetrix GeneChip is a popular and usefull platform for gene ex- pression profiling, but the use of its probes and probe-sets mapping has multiple inconveniences. In fact, the probe-sets for the Affymetrix Human

Genome U133 Plus 2.0 Array are based on UniGene database (Build 133, April 20, 2001) and considering how rapidly human genome has evolved many probes on the array are not correctly assigned. To avoid this problem, we used the updated probe alignment and gene mapping that is provided by the Chip Definition File (CDF): **hgu133plus2hsensgcdf** (downloaded from brainarray (Sandberg and Larsson, 2007)).

4.2.5 Batch effect removal evaluation

We performed unsupervised hierarchical clustering to observe unlikely clustering based on batches in those ex- pression value matrixes where batch effects remained after pre-processing. We used a 30-random sampling per batch, identifying each batch by a different colour.

The batch effect was also investigated using principal components analysis (PCA). PCA is a useful technique for exploratory data analysis, allowing you to better visualize the variation present in a dataset with many variables. It is particularly helpful in the case of "wide" datasets, where you have many variables for each sample.

A linear regression of average gene expression on array batch per pre-processing method was the final approach fulfilled to assure removal.

4.2.6 Differential expression analysis

For the identification of gene whose altered expression achieved statistical significance we used the R algorithm Linear Models for Microarrays (LIMMA package). We applied LIMMA (Ritchie et al., 2015) to the expression data matrix fixing an adjusted p-value threshold of $FDR \leqslant 0.01$ to select significant genes.

The comparison was done separating the samples according to their clinical and pathological stage (comparing CRC stages I and II versus III and IV). In this way we found a set of 2707 candidates genes, corresponding to 2524 protein-coding genes that were tested in the survival analysis (the rest were non-coding genes). In this work we focus only on the genes that encode proteins because we wanted to find CRC survival markers that later can be tested at protein level using, for example, immunohistochemistry (IHC) analysis.

4.2.7 Linear Regression analysis

In statistics, linear regression or linear adjustment is a mathematical model used to approximate the dependency relation between a dependent variable Y, the independent variables X_i and a random term ε. This model can be expressed as:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon \tag{4.1}$$

where:

Y_t: dependent variable.
$X_1, X_2, ..., X_p$: independent variable.
$\beta_0, \beta_1, \beta_2, ..., \beta_p$: regressions coefficients measuring the influence that independent variables have over the model.

Where β_0 is the intersection or "constant" term, the $\beta_i (i > 0)$ are the respective coefficents for each independent variable, and p is the number of independent parameters to be taken into account in the regression.

The independent variable X_i is each one of the batches or datasets, the model measures if the batches have influence over the regression. The independent term or β_0 is the coefficent that should be present if there are no batch effect. Otherwise, if any or all the coefficents β_i are not zero, then the batch effect will be present.

4.2.8 Survival analysis

Our intention in this research was to identify genes whose relative expression level affect survival and prognosis in CRC, once we had made a preselection in its behavior through stage evolution of 2524 protein-coding genes. The first step for the survival analysis was to define for each gene two separated distributions of high and low expression along the sample dataset investigated. This separation based in expression level determined the explanatory variable. The Kaplan-Meier and risk prediction techniques are the one described in the first chapter, section Sec: 2.2.4

For computing the time to event, the response variable in the models was the Overall Survival (OS) time. All the data sets that we integrated in our analyses had OS information. In some cases for some individuals, Disease Specific Survival (DSS) times or Relapse Free Survival (RFS) times were also provided with the original data, but we did not considered these time-events since we wanted to focus on OS to achieve a homogeneous analysis.

4.3 Results

4.3.1 A large dataset of CRC samples including global expression and survival data

We first built a large cohort of CRC samples collected from individuals that had clinical record with survival data times, as well as genome-wide expression profiles of their colorectal primary tumours at diagnosis (i.e. before any drug treatment). Our aim was to achieve a meta-dataset with at least 1 thousand samples and to demonstrate a good integration of the global transcriptomic profiles of different samples sets avoiding the typical batch-effects that can alterate any unified analysis.

Tab: 4.1 presents the datasets of CRC samples that were collected to produce the integrated dataset analysed in this work. All the CRC samples included in this meta-dataset were tested for global gene expression profiling using the platform of

high-density microarrays from Affymetrix: Human Genome U133 Plus 2.0. Using this platform, the probesets of the arrays were mapped to single genes (as indicated in Risueño et al.) (Risueno et al., 2010) and, in this way, each microarray measured the expression signal of 20,079 human genes (using the mapping provided by the Chip Description File, CDF v.21 from brainarray).

As a whole, Tab: 4.1 includes 7 series that were obtained from the Gene Expression Omnibus repository(GEO (NCBI, 2019)). These datasets included a total amount of 1352 CRC samples, but after collecting the clinical survival data and carrying out the integration and normalisation protocols we finished with 1273 samples, since we filtered 79 samples that did not have survival data or did not show comparable data distributions after normalisation.

4.3.2 Evaluation of normalisation procedures to integrate independent batches

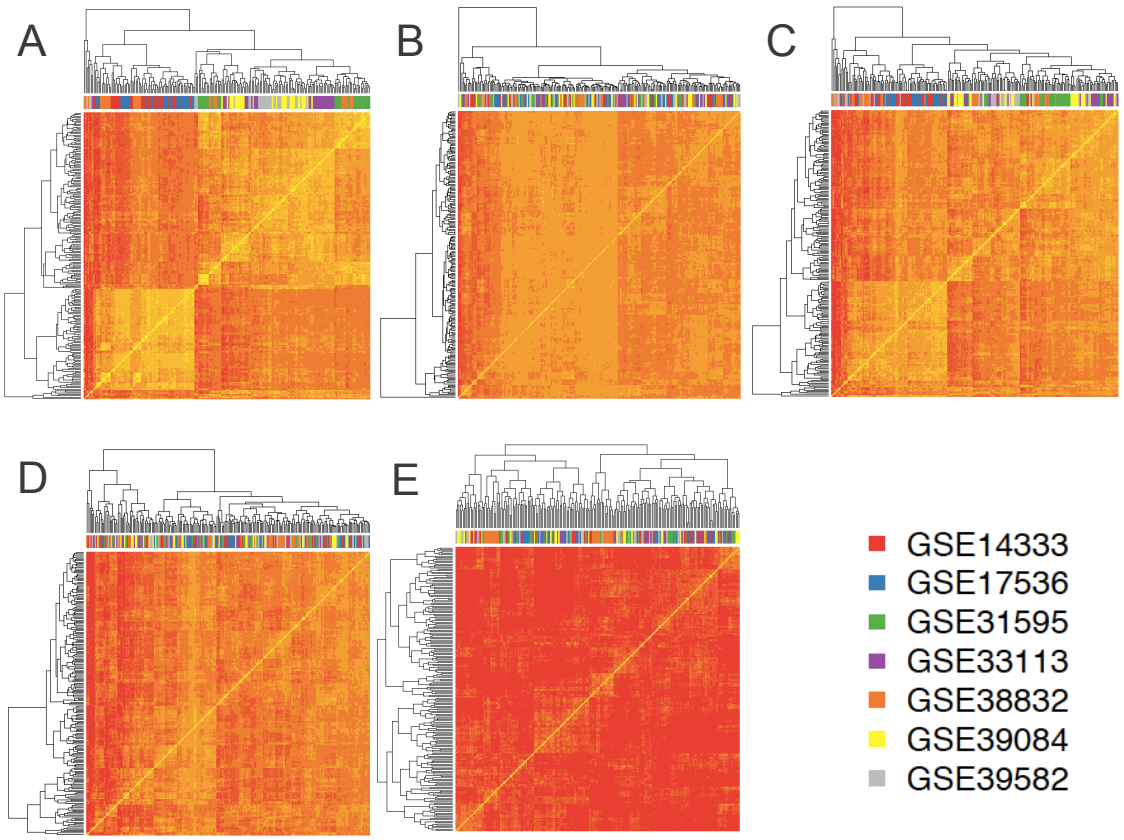


Figure 4.2: Symmetric heatmaps from different normalisation methods.

We performed the integration and combined normalisation of the CRC expression datasets using 5 different proce- dures. The procedures applied different normalisation algo- rithms to provide a homogeneous signal matrix, avoiding bias due to batch effect on the global expression profile of the CRC samples.

The procedures applied were: (i) Robust Multi-array Average (RMA) algorithm (Irizarry et al., 2003); (ii) RMA plus Combatting Batch effects (ComBat) algorithm (Stein et al., 2015); (iii) Frozen Robust Multi-array Average (fRMA) algorithm (McCall et al., 2010b); (iv) fRMA plus Combat; (v) fRMA plus scaling of the data using mean-centered expression values.

To evaluate and compare the results provided by each one of these 5 procedures we carried out several analyses. **Fig: 4.2** presents the heatmaps derived from an unsupervised clustering of the samples using in each case the expression data matrix derived from each one of the 5 procedures applied.

Due to the fact that each series has a different number of samples (one with more than 500 and several other with less than 100), we did a random selection of an even number of samples for each dataset to be included in the cluster analysis: 30 samples from each one. In this way, each heatmap is composed of 210 samples (30×7): 30 samples from each one of the 7 datasets (identified by the ID number, GSE, from GEO).

Symmetric heatmaps representing the similarity between the overall gene expression signal of the samples compared with each other. Each heatmap is composed of 210 samples (30×7, 30 samples random selected from each batch, i.e. from each one of the 7 GSE datasets). The samples of each batch are identified by a colour in the top bar below the top dendrograms (following the colours legend).

Each heatmap represents a different preprocessing and normalisation method performed to merge the datasets in one batch. The methods applied were: **A** RMA; **B** RMA plus ComBat, **C** fRMA, **D** fRMA plus ComBat, **E** fRMA plus scaling of the data using mean-centered expression values.

In **Fig: 4.2** the samples of each batch are identified by a colour that is indicated in the horizontal bar below the dendrograms. Each heatmap represents a different preprocessing and normalisation method performed to merge the datasets in one meta-dataset. The results shown in these clustering analyses indicate that in the case of methods that gave the heatmaps A, C and E, several samples of the same colour are grouped together showing that they have a common correlation profile within the global expression signature.

By contrast, in the case of methods that gave the heatmaps B and D, there is a clearer shuffling of all the colours, which reflects a homogenous mix of the overall expression signal coming from different datasets.

The clustering analysis presented in the symmetric heatmaps of **Fig: 4.2** was done using, for each sample, a vector including the expression signals along all genes and calculating with these vectors the pair-wise Pearson correlations between samples and the pair-wise distance matrix derived from such correlations. This approach can reveal major effects associated to the global expression signal of the samples, but it is not very sensitive to detect minor changes in a small number of genes.

For this reason we applied a second approach to compare the results provided by the 5 normalisation procedures in order to select the one that produces the best unification of the 7 CRC datasets, preserving a good signal to noise ratio in the expression distributions.

Algorithms of dimensionality reduction, such as PCA (Principal Component Analysis), allow exploring large datasets in an accurate way to identify factors that are relevant for the variance of studied variables (in our case the expression of the genes in the unified meta-dataset of 1273 samples).

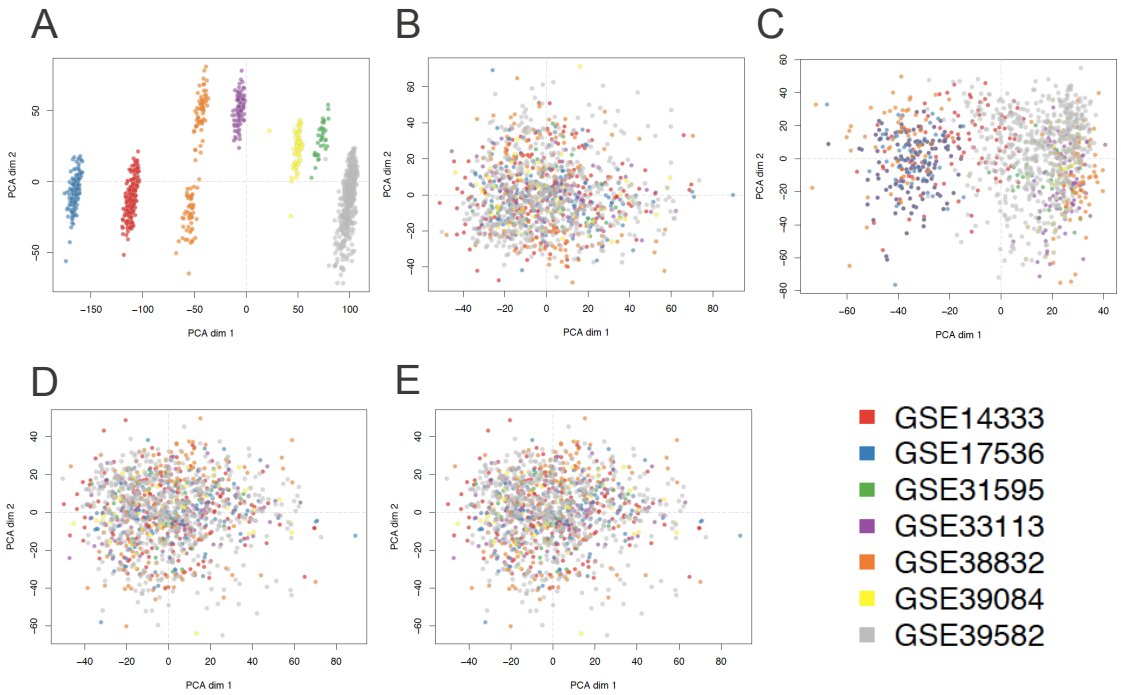


Figure 4.3: PCA representation of normalisation mehtods.

Plots presenting the distribution of the 1273 samples from 7 datasets (GSEs) obtained by Principal Component Analysis (PCA) of the global gene expression profile of each sample; that converts the signal of each sample using an orthogonal transformation in linearly uncorrelated variables called principal components or dimensions.

Each plot presents the values of the two main dimensions (dim 1 versus dim 2) and corresponds to the PCA results obtained using the expression data calculated with different preprocessing and normalisation methods. The methods applied were: **A** RMA; **B** RMA plus ComBat, **C** fRMA, **D** fRMA plus ComBat, **E** fRMA plus scaling of the data using mean-centered expression values. The samples of each batch are identified by colour dots following the colours legend.

Fig: 4.3 presents the plots derived from the PCA done over the 5 expression matrices (i.e. the signal of 20,079 genes in 1273 samples) obtained with 5 different normalisation approaches. These results show very clearly that the RMA method (**Fig: 4.3 A)** is not good to provide a proper normalisation of different batches,

since the samples keep a very strong signal associated to each batch.

The fRMA method (**Fig: 4.3 C)** neither is good, since some samples (specially the ones from the largest batch GSE39582) still keep a strong signal associated to their batch. By contrast, the analysis of the data provided by the other 3 procedures (RMA plus Combat, fRMA plus Combat and fRMA plus mean-centered scaling.

Fig: 4.3 B, D and E, respectively) showed an adequate mix of all the samples from different batches. Within these 3 procedures, the normalisation is very similar keeping a good signal to noise ratio along the genes and a small signal reduction.

We finally select option **Fig: 4.3 B**, RMA plus Combat, because the heatmap in **Fig: 4.2 B** showed the best mix between series and a better similarity between the samples (compared to options D or E).

As a final testing to identify the best integration and normalisation procedure of the 7 CRC expression datasets, we carried out a linear regression analyses (as described in equation: 4.1) on the global expression matrix considering as predictors 7 independent dummy variables or factors.

These variables correspond to the series from which each sample comes from. In this way, if these factors have a significant influence in the expression signal distributions, the linear regression analysis will show a significant p-value and correlation.

The results of this analysis are presented in Tab: 4.2 , that reveals again that only the data matrices produced by the methods B and D (RMA plus Combat and fRMA plus Combat, respectively) do not show a significant effect attributed to belonging to one of the series.

Finally, we choose B versus D as the final procedure applied because, despite being very similar, the application of RMA plus Combat provoked less dramatic changes with respect to the raw signal expression.

Factors considered	Est. coefficients	Std. error	p.value	Coefficients effect
(A) RMA				
Intercept	6.925	0.014	<2e-16	–
(GSE14333+) GSE17536	0.387	20.230	<2e-16	yes
GSE31595	– 1.212	0.019	<2e-16	yes
GSE33113	– 0.577	0.019	<2e-16	yes
GSE38832	– 0.355	0.019	<2e-16	yes
GSE39084	– 0.978	0.019	<2e-16	yes
GSE39582	– 1.375	0.019	<2e-16	yes
(B) RMA plus ComBat				
Intercept	6.219	0.013	<2e-16	–
(GSE14333+) GSE17536	0.000	0.001	0.999	no
GSE31595	0.002	0.019	0.903	no
GSE33113	0.001	0.019	0.959	no
GSE38832	– 0.001	0.019	0.973	no
GSE39084	0.002	0.019	0.927	no
GSE39582	0.001	0.019	0.977	no
(C) fRMA				
Intercept	6.535	0.015	<2e-16	–
(GSE14333+) GSE17536	– 0.011	– 0.553	0.580	no so much
GSE31595	0.089	0.021	0.000	yes
GSE33113	0.071	0.021	0.001	yes
GSE38832	0.054	0.021	0.008	yes
GSE39084	0.096	0.021	0.000	yes
GSE39582	0.089	0.021	0.000	yes
(D) fRMA plus ComBat				
Intercept	6.590	0.014	<2e-16	–
(GSE14333+) GSE17536	0.000	0.001	1.000	no
GSE31595	0.002	0.020	0.926	no
GSE33113	0.001	0.020	0.942	no
GSE38832	0.000	0.020	0.985	no
GSE39084	0.002	0.020	0.929	no
GSE39582	0.000	0.020	0.994	no
(E) fRMA plus centered				
Intercept	0.000	0.000	0.101	–
(GSE14333+) GSE17536	0.000	1.264	0.206	yes
GSE31595	0.000	0.000	0.773	no so much
GSE33113	0.000	0.000	0.108	yes
GSE38832	0.000	0.000	0.147	yes
GSE39084	0.000	0.000	0.940	no
GSE39582	0.000	0.000	0.163	yes

Table 4.2: Linear regression analysis depicting the coefficients(batches) relevance in the model.

Results of the linear regression analyses on the global expression matrix calculated for the 1273 samples from 7 datasets (GSEs) combined using 5 different preprocessing and normalisation methods.

4.3.3 Identification of genes associated to advanced CRC that mark survival differences

Once we produced a large and well-integrated metadataset of CRC samples, having global expression profiles and clinical survival data for all cases, we proceed to the identification of the subset of genes that suffer significant changes with colorectal tumour progression.

To do this, we explored the overall expression matrix to detect the genes that showed a significant expression change when comparing CRC tumours in early stages (stages I and II) versus CRC tumours in late or advanced stages (stages III and IV). This comparison was done applying LIMMA, differential expression algorithm, and retrieving all genes that gave a significant p-value (adjusted $p < 0.05$) in either direction (i.e., genes up-regulated with the progression of the disease, in late versus early CRC stages; or genes down-regulated with the progression of the disease).

Such differential expression analysis gave a subset of 2707 human genes: 2524 corresponding to protein-coding genes and the rest to non-coding genes (in this work we focused only in the protein-coding genes).

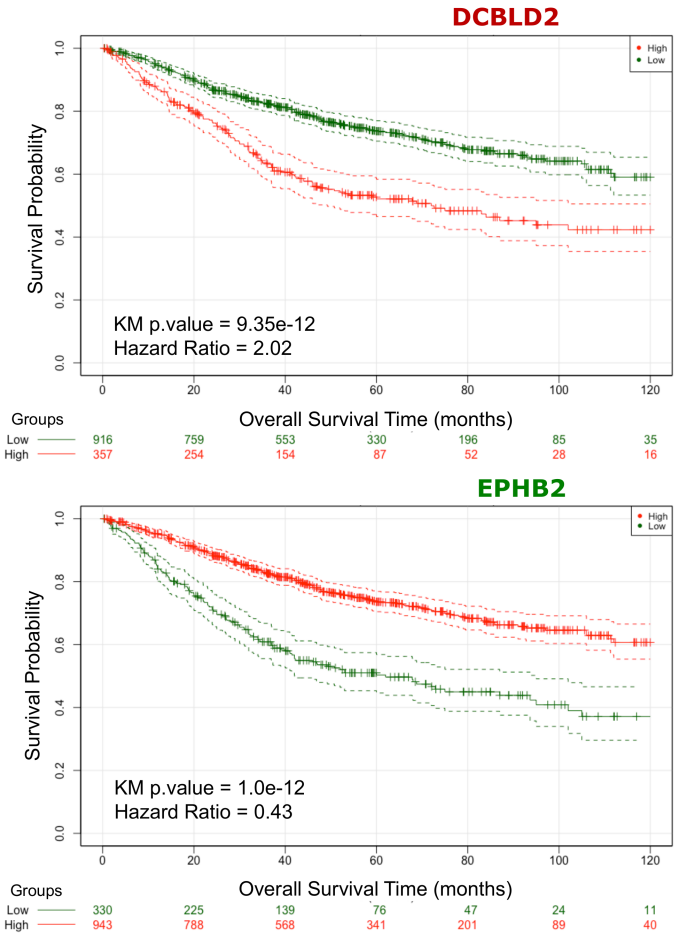


Figure 4.4: DCBLD2, EPHB2 survival plots.

Once we had the subset of genes that can be associated to advanced or progression of CRC, we perform a second analysis on these gene candidates to find out which ones can be correlated with the survival of the corresponding patient samples based on their expression signals.

To do this, we carried out Kaplan-Meier (KM) analysis of the survival times of the set of 1273 colorectal cancer samples for each one of the 2524 genes found in the previous exploration. In this analysis, the genes were ranked considering the non-parametric log-rank test that evaluates the separation between the two KM curves for two prognostic groups: one with good survival and another with poor survival.

To do this, our algorithm performs for each gene multiple splits of the sample cohort in two groups, and looks for the splitting that provides the best separation between groups (i.e. the best p-value). Then, a stringent cut-off value (adjusted p < 0.0003) was used to select the genes that are considered significant.

This allowed the identification of 429 significant genes in which the overexpression correlated with low survival, plus 336 significant genes where the repression correlated with low survival. These analyses were done in a univariate mode, considering each gene as an independent factor.

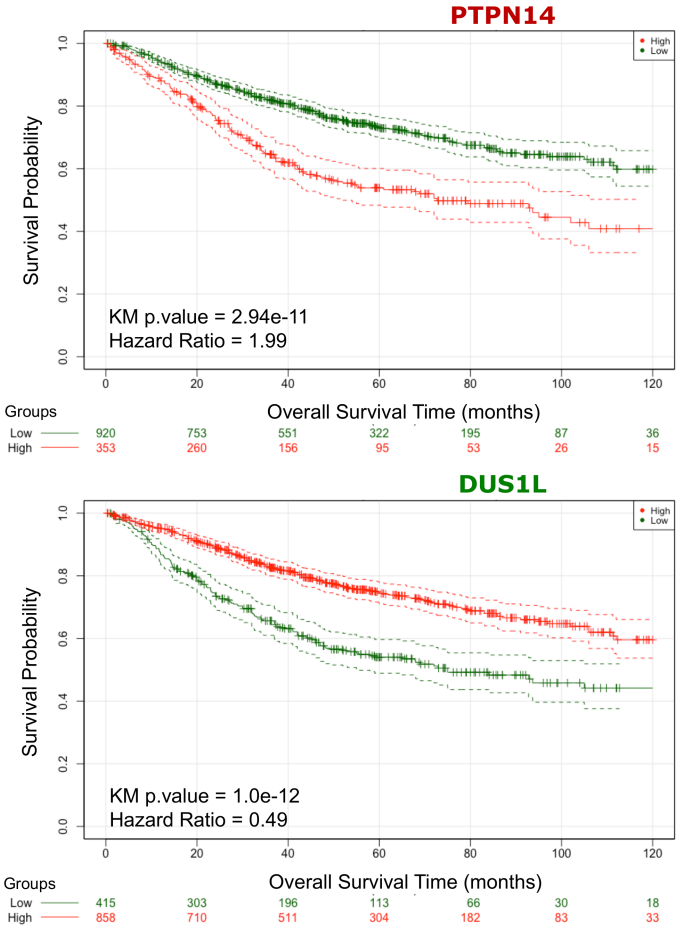


Figure 4.5: PTPN14, DUS1L survival plots.

Kaplan-Meier plots of the survival analysis of the set of 1273 samples from colorectal cancer (CRC) patients, **Figs: 4.4 and 4.5**. The patients are separated in two groups (high in red and low in green) according to the expression profiles of 4 genes: DCBLD2, PTPN14, EPHB2, and DUS1L. These genes provided the best split between patients of high and low risk based in their expression levels.

In the case of genes DCBLD2 and PTPN14 (labelled in red) the over-expression is correlated with poor survival; and in the case of genes EPHB2 and DUS1L (labelled in green) the over-expression is correlated with good survival. In all cases the adjusted p.values of the analyses are very significant (as indicated inside each plot), indicating that the two populations represented by the two curves have a very clear difference in their overall survival.

Figs: 4.4 and 4.5 show the Kaplan-Meier plots corresponding to the survival profiles of the two populations of individuals that were segregated according to the expression values of the gene tested. The 4 plots correspond to the top genes: DCBLD2 and PTPN14 with overexpression correlated to low survival; and EPHB2 and DUS1L with repression correlated to low survival.

The separation of the two populations in both cases is very significant, with KM p-values $<$ 1.0e-10 and Hazard Ratios (HR) around 2.0 for overexpression cases and around 0.45 for repression cases. These parameters were calculated using all the 1273 samples; however it was necessary to do an internal cross-validation of these results to assess how stable and reliable was the signal for each one of the selected genes.

We carried out a cross-validation of the top-200 genes selected in any of the two conditions (i.e. selected as survival markers when they were up-regulated for the cases of poor survival or when they were up-regulated for the cases of better survival). This internal cross-validation was done using for each gene a resampling strategy that randomly selected 80% of the sample 100 times (i.e. doing 100 iterations).

A short view of these data is shown in Tab: 4.3 that presents the 50 genes selected as best survival markers of CRC: the first part of the table corresponds to the top 25 genes, where up-regulation corresponds to shorter survival and higher risk (HR $>$ 1); the second part of the table corresponds to the top 25 genes, where up-regulation corresponds to longer survival and lower risk (HR $<$ 1).

The genes were ranked by their KM p-values and the HR values calculated for the whole dataset (i.e. for all the 1273 samples, all-dt). As indicated, the stability and robustness of the gene survival markers was assessed via a resampling strategy using the robust bootstrap strategy described in Chapter 1. For the final ranking of the genes included in these tables we also considered that they had to give a significant adjusted p-value in more than 80 out of 100 bootstrap iterations (i.e. N-sinf-in-100i $>$ 80).

Symbol	KM.p.value	N-in-100i	HR(in-100i)	Gene description
DCBLD2	0.0000000000	99	2.106	discoidin; CUB and LCCL domain containing 2
PTPN14	0.0000000000	99	2.082	protein tyrosine phosphatase; non-receptor type 14
LAMP5	0.0000000000	93	2.046	lysosomal associated membrane prot.member 5
TM4SF1	0.0000000001	93	2.031	transmembrane 4 L six family member 1
NPR3	0.0000000002	97	2.136	natriuretic peptide receptor 3
LEMD1	0.0000000003	85	1.937	LEM domain containing 1
LCA5	0.0000000003	97	2.021	LCA5; lebercilin
CSGALNACT2	0.0000000008	92	1.974	chondroitin sulfate N-acetylgalactosaminyltransferase 2
SLC2A3	0.0000000014	89	1.993	solute carrier family 2 member 3
GADD45B	0.0000000018	97	2.074	growth arrest and DNA damage inducible beta
SCEL	0.0000000018	87	1.928	sciellin
SIX4	0.0000000019	91	1.951	SIX homeobox 4
AKAP12	0.0000000028	95	2.092	A-kinase anchoring protein 12
COLEC12	0.0000000028	92	1.941	collectin subfamily member 12
PDLIM3	0.0000000047	91	1.985	PDZ and LIM domain 3
ITGB5	0.0000000049	88	1.911	integrin subunit beta 5
GULP1	0.0000000050	88	1.911	engulfment adaptor PTB domain containing 1
SCG2	0.0000000051	93	2.034	secretogranin II
AHNAK2	0.0000000066	87	1.896	AHNAK nucleoprotein 2
CYP1B1	0.0000000075	85	1.884	cytochrome P450 family 1 subfamily B member 1
PRKD1	0.0000000451	87	1.872	protein kinase D1
SPARCL1	0.0000000471	85	1.863	SPARC like 1
CDKN2B	0.0000000717	84	1.847	cyclin dependent kinase inhibitor 2B
MLLT11	0.0000001989	84	1.813	myeloid/lymphoid or mixed-lineage leukemia; t to 11
CD36	0.0000002751	85	1.891	CD36 molecule
EPHB2	0.0000000000	100	0.426	EPH receptor B2
DUS1L	0.0000000000	98	0.481	dihydrouridine synthase 1 like
NUAK2	0.0000000001	96	0.495	NUAK family kinase 2
FANCC	0.0000000002	95	0.498	Fanconi anemia complementation group C
CISD3	0.0000000002	87	0.511	CDGSH iron sulfur domain 3
TIMM13	0.0000000003	95	0.511	translocase of inner mitochondrial membrane 13
AGMAT	0.0000000005	95	0.515	agmatinase
MYB	0.0000000006	93	0.508	MYB proto-oncogene. Transcription factor
CHDH	0.0000000006	90	0.520	choline dehydrogenase
FHDC1	0.0000000008	96	0.505	FH2 domain containing 1
ZBED3	0.0000000009	88	0.522	zinc finger BED-type containing 3
NOL9	0.0000000015	92	0.527	nucleolar protein 9
GAR1	0.0000000017	99	0.479	GAR1 ribonucleoprotein
FAM83F	0.0000000019	93	0.518	family with sequence similarity 83 member F
TXN2	0.0000000036	88	0.527	thioredoxin 2
GALK1	0.0000000036	88	0.525	galactokinase 1
MLEC	0.0000000045	96	0.476	malectin
MAPKAPK3	0.0000000048	92	0.520	mitogen-activated protein kinase-activated 3
CASP1	0.0000000180	87	0.523	caspase 1
MCCC2	0.0000000183	93	0.516	methylcrotonoyl-CoA carboxylase 2
BEND3	0.0000000193	88	0.529	BEN domain containing 3
CISH	0.0000000216	87	0.508	cytokine inducible SH2 containing protein
LARS2	0.0000000239	91	0.528	leucyl-tRNA synthetase 2; mitochondrial
CDC25A	0.0000000481	90	0.539	cell division cycle 25A
L3MBTL4	0.0000000606	90	0.506	l(3)mbt-like 4 (Drosophila)

Table 4.3: Genes selected as top-50 best survival markers of colorectal cancer (CRC).

4.3.4 External validation of prognostic markers with a CRC cohort studied using RNA-seq

The analyses done so far provided a ranked collection of genes found as robust markers of survival in CRC. The consistency of the results obtained with the internal cross-validation gives strong support to the top genes found (presented in Tab: 4.3) , but we had to consider the value of using other external independent CRC cohorts to corroborate these findings.

As far as we could investigate we did not find other large CRC datasets (i.e., sets with more than one thousand samples) that included global gene expression data plus survival as part of the clinical characterisation of samples.

Despite this limitation, we look for independent datasets and found in The Cancer Genome Atlas (TCGA (TCGA, 2019)) a well-characterised cohort of 276 colorectal carcinomas that had been studied with several genome-scale technologies (including RNA-seq gene expression profiling) and that had survival data for 269 samples (Muzny et al., 2012).

We used these data to validate the top genes found as best survival markers in our previous analysis. The results indicated a good performance in more than two thirds of the genes tested. 7 genes of the top 10 for the case of up-regulation associated with poor survival (PTPN14, LAMP5, TM4SF1, LCA5, CSGALNACT2, SLC2A3 and GADD45B) and 6 genes of the top 10 previously found for the case of up-regulation associated with good survival (EPHB2, DUS1L, NUAK2, FANCC, MYB and CHDH).

4.3.5 External validation of prognostic markers using multivariate survival analysis

Up to now the search to find gene survival markers associated to the prognosis of CRC have been done using univariate analysis that look for the value and influence of each singular gene. The results presented provided multiple parameters to allow a proper statistical assessment and ranking of each gene survival markers proposed (Tab: 4.3).

To provide extra support to these results we did another external validation using a second independent cohort of CRC samples from the platform SurvExpress (Aguirre-Gamboa et al., 2013).

The CRC dataset selected was called “Colon-Metabase-Uniformised” and it included 482 samples with overall survival data and genome-wide expression determined with Affymetrix microarrays. We performed several multivariate survival analyses (OS, overall survival) on this dataset using combinations of the top genes proposed in Tab: 4.3.

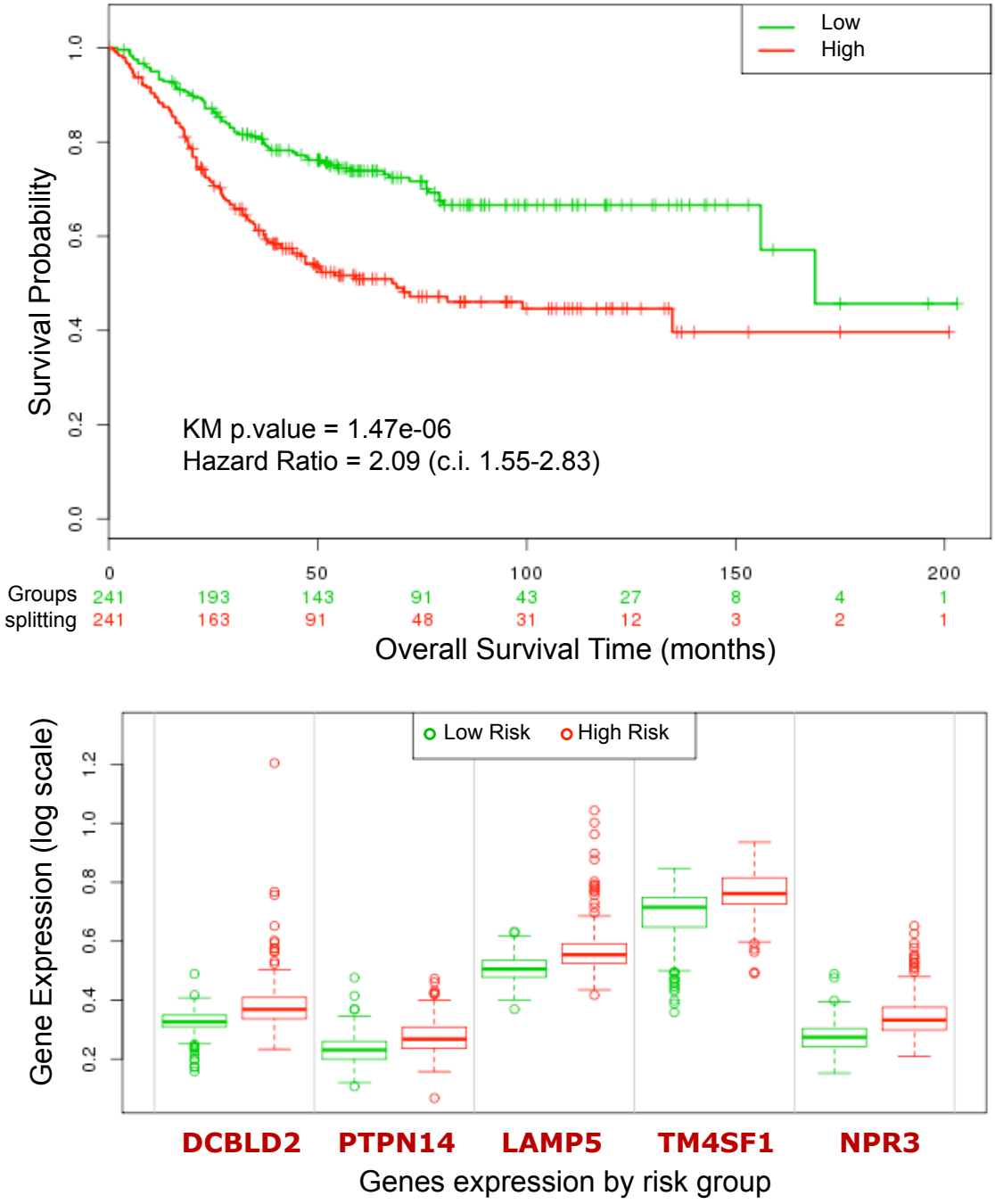


Figure 4.6: KM multivariate survival analysis. Top 5 genes.

As an example of these analyses we present the KM plot **Fig: 4.6** corresponding to the multivariate survival study done using the top 5 genes found up-regulated for poor survival (DCBLD2, PTPN14, LAMP5, TM4SF1 and NPR3).

It can be seen that the combination of these genes provides a very good separation of two CRC populations: one group of high-risk, associated to the overexpression (or up-regulation) of the genes; and another group of low-risk, associated to the lower expression (or down-regulation) of these genes.

This analysis was repeated with several other combinations of the top up-regulated genes associated with poor survival (present in Tab: 4.3), resulting in similar results. For example, combining DCBLD2, LAMP5, TM4SF1, NPR3 and GADD45B the separation of the high and low-risk groups improved a bit: KM p-value = 2.21e-07 and HR = 2.23 (95% confidence interval, CI: 1.65–3.02). Another combination that provided very good separation was using genes DCBLD2, LAMP5, TM4SF1, NPR3 and AKAP12: KM p-value = 2.51e-10 and HR = 2.74 (95% CI: 2.00–3.74).

4.3.6 Gene expression profiles of CRC tumour samples versus normal colorectal samples

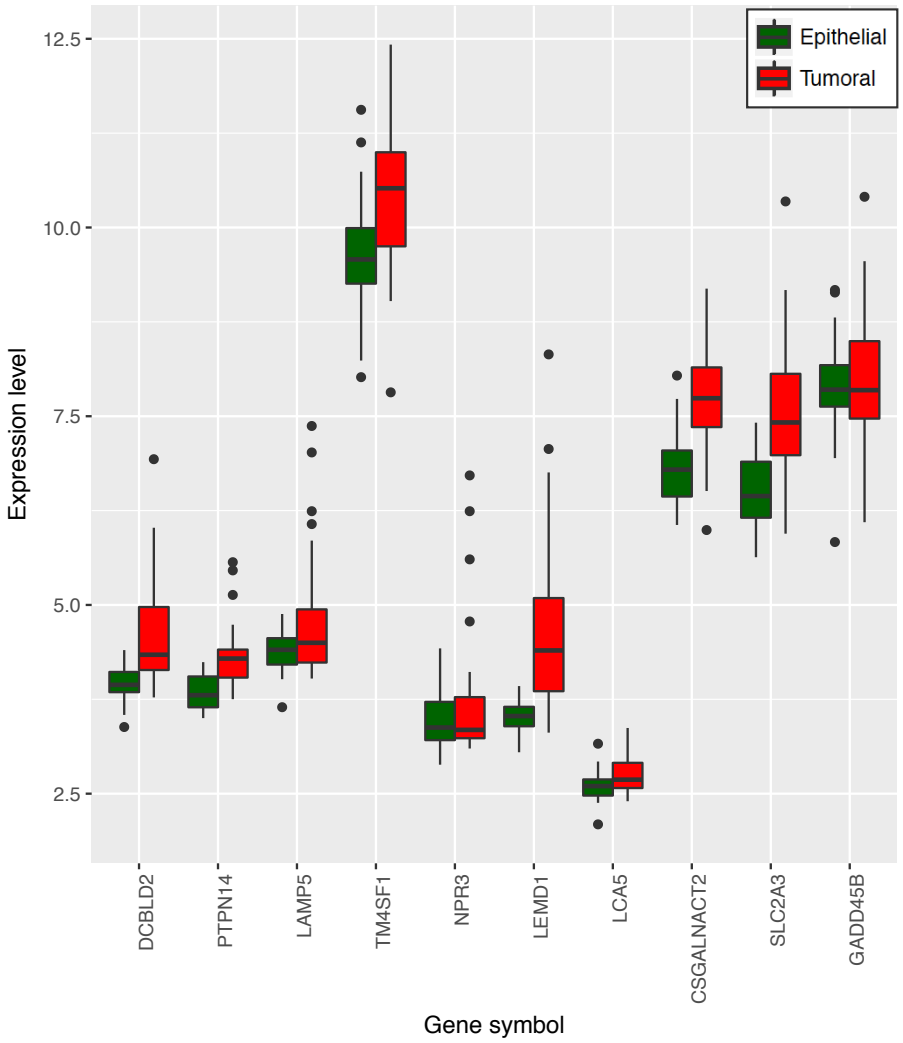


Figure 4.7: Gene markers up-regulated in CRC tumours vs normal.

All the integrated datasets, so far presented in this study corresponded to CRC samples, because we want to provide genes that are disease markers present in the transformed tumour cells of the intestinal epithelium, and genes that mark the progression and aggravation of this type of cancer.

In addition, we can only have survival information about patients since in healthy individuals survival time cannot be related to disease and there are not disease-associated events.

Despite this obvious consideration, it is interesting to explore what would be the level of expression of the genes, that we identified as survival markers, when they are analysed in normal colorectal tissue.

Exploring back on the experimental series used to create our meta-dataset of 1273 CRC samples, we found in series GSE33113 and GSE39582 a collection of 25 samples that corresponded to normal colorectal tissue. We took these samples and included then with our CRC dataset using the same normalisation protocol.

After this integration, we could explore the expression level of the top up-regulated genes (identified as markers of poor survival), comparing the expression distribution on a set of cancer samples versus a set of normal tissue samples.

In both cases the number of samples compared were 25, since this is the number of normal samples that we had. We did this comparison 20 times, random selecting each time a different subset of 25 cancer samples. The results were always very similar and the boxplots of the expression distributions for the top 10 genes are presented in **Fig: 4.7**.

These results indicate that the gene markers, identified in our survival studies, are most of the times also up-regulated in CRC tumours with respect to normal colorectal tissue.

4.3.7 Risk predictor score based in the multivariate analysis of candidate survival markers

Finally, to obtain a more accurate evaluation of the prognostic value of all the genes selected as best candidates, we performed another analysis of the candidate markers using a regularised multivariate Cox proportional-hazards regression with L1 norm penalty (Gui and Li, 2005), with the scope of building a multigenic "risk predictor".

This analysis was done on the cohort of 1273 samples of CRC patients, using for the multivariate analysis the top 100 genes that showed up-regulation correlated with poor prognosis (i.e. overexpressed in low survival cases).

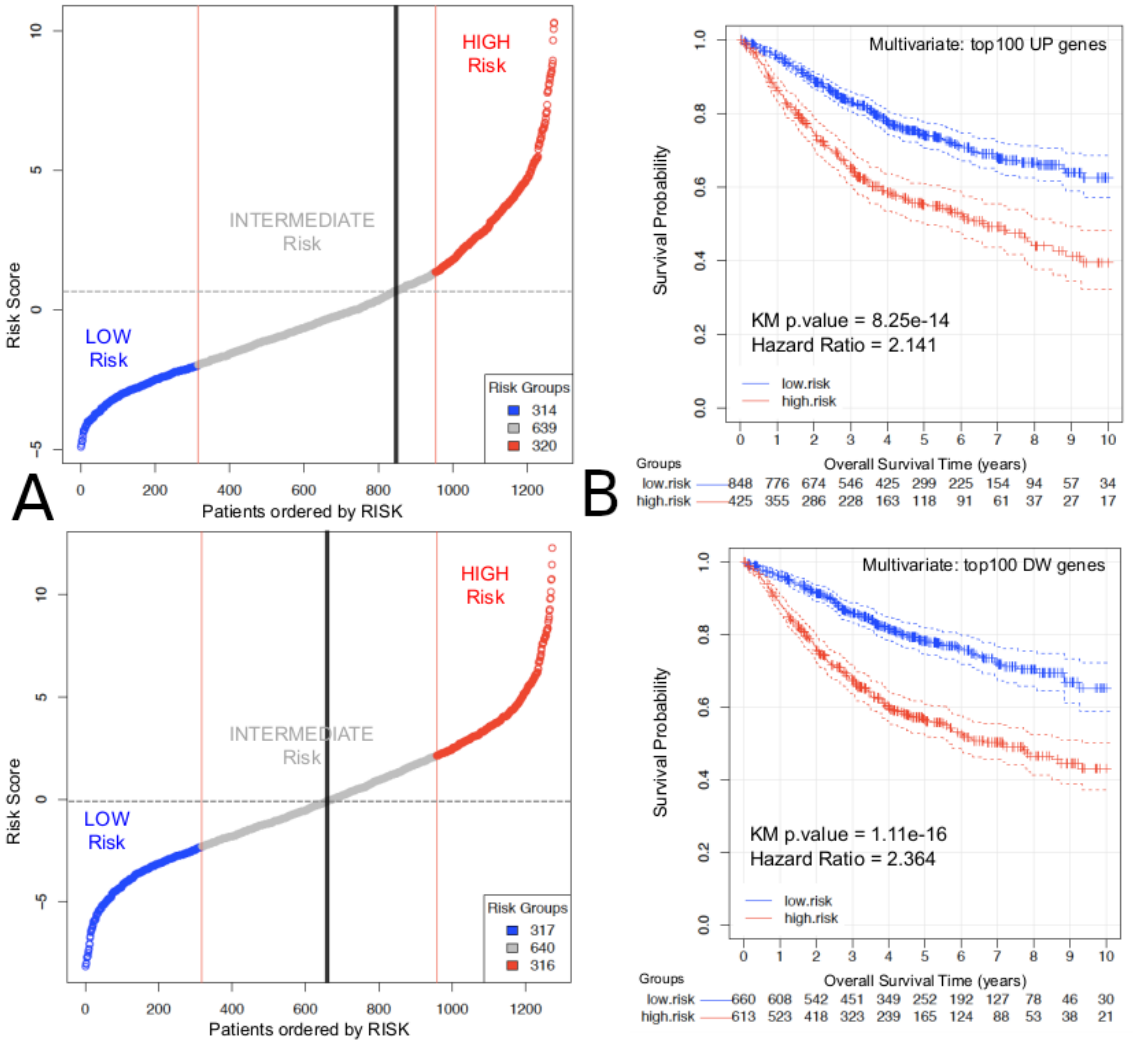


Figure 4.8: Risk and Survival analysis, top 100 UP and DOWN genes.

Risk prediction done for the cohort of 1273 patients of CRC based in the multivariate analysis using the top 100 genes that showed upregulation correlated with poor prognosis (i.e. overexpressed in low survival cases).

A Plot presenting the patients according to their risk score, from Low (blue) to High (red) risk. A recursive algorithm using 10-fold crossvalidation(algorithm described in Chapter 1) finds the value of risk score (marked with a vertical black line) that allows the best splitting of the cohort in two groups. **B** Kaplan-Meier plot showing the separation of these two groups: a high-risk group including 425 individuals (in red) and a low-risk group including 848 individuals (in blue).

The analysis has been done using a multivariate Cox proportional-hazards regression. As shown, the division is very significant (p-value = 8.25e-14) and allows an optimal separation of individuals according to their survival. The analysis of the beta factors assigned by the regression to each of the top 100 genes (i.e. to each variable within the multivariate vector) allows the identification of the genes that are the most influential factors in this risk analysis and therefore it helps in the

selection of the best "gene survival markers"

The results are presented in **Fig: 4.8** that shows a graph ordering the patients according to their risk score, from low-risk (blue) to high-risk (red), including also an intermediate region (grey) (**Fig: 4.8 A**). A recursive algorithm using 10-fold cross-validation was applied to find the value of risk score. The threshold (marked with a vertical black line) is obtained by maximizing the separability between the survival curves for the resulting groups. Therefore, it allows the best splitting of the cohort in two groups.

A Kaplan-Meier plot showing the separation of these two groups is also presented (**Fig: 4.8 B**); dividing the population into a high risk group including 425 individuals and a low risk group including 848 individuals. As shown, the division is significant (p-value = 8.25e-14) and allows an optimal separation of individuals according to their survival. The analysis of the beta factors assigned by the regression to each of the top 100 genes, i.e. to each variable within the multivariate vector, allows the identification of the genes that were the most influential factors in this risk analysis and therefore it facilitated the selection of the best "gene survival markers". As indicated in previous sections, the top 100 genes included in the construction of this multigenic risk predictor score were selected from the list of best markers found during the survival test with single genes.

4.4 Discussion

CRC is a complex disease composed of biologically and clinically diverse subtypes, which can originate in differ- ent ways provoking multiple clinical scenarios (Linnekamp et al., 2015) (Dienstmann et al., 2017). This complexity causes the molecular characterisation of CRC to remain deficient, with a lack of clear gene markers associated to specific CRC subtypes and to the prognosis of the disease (Sameer, 2013) (Fessler and Medema, 2016) (Bijlsma et al., 2017). In fact, current molecular phenotyping of colorectal tumours is usually linked to the traditional determination of somatic mutations in well- known oncogenes such as KRAS and BRAF (Kocarnik et al., 2015).

The recent advance of genomic and transcriptomic technologies applied to the study of clinical samples did open the way to obtain genome-wide expression profiles of multiple patient cohorts and correlate the expression of certain genes with different disease subtypes, disease stages and progression (Aibar et al., 2015) (Moreno and Sanz-Pamplona, 2015). This approach had been widely applied in cancer research in the last decade and is very powerful when the identification of marker genes is associated with survival time. The correlation between gene expression and survival is an excellent tool to investigate prognosis of the disease and to build risk predictors that will be applicable to individual patients.

The identification of molecular biomarkers with prognostic value in CRC has been a challenging task (Sanz-Pamplona et al., 2012) (George and Kopetz, 2011a)

(George and Kopetz, 2011b) (Das et al., 2017). Molecular prognosis of colorectal tumour samples by transcriptional profiling started about 15 years ago (see review (George and Kopetz, 2011a)), and in more recent years several specific gene signatures associated with CRC survival have been published (Vargas et al., 2015) (Sveen et al., 2012) (Kopetz et al., 2015) and (Salazar et al., 2011) (Nguyen et al., 2015) (Chen et al., 2017) (Tian et al., 2017) (Xu et al., 2017). Despite these efforts, at present there is not a clear compendium of gene markers for CRC survival and it is quite difficult to find consistency in the literature (George and Kopetz, 2011a). A clear limitation comes from the fact that, in most of previous studies, the number of tumour samples used to select the genes that enter into the construction of the prognostic predictors is small (i.e., the size of the patient cohorts rarely it is greater than a few hundred individuals).

For example, ColoPrint is a 18-gene signature for prognosis prediction of stage II and III CRC, that was identified using as training set tumour samples from 188 patients (Kopetz et al., 2015) (Salazar et al., 2011); a 113-gene expression signature for predicting prognosis in patients with CRC was built using 145 samples as dicovery set (Nguyen et al., 2015); a 7-gene signature to predict overall survival of CRC patients was based in an initial training set of 67 samples (Chen et al., 2017); a recurrence-associated CRC signature of 13 genes was developed using a screening set of 145 samples (Tian et al., 2017); a 15-gene signature for prediction of CRC recurrence and prognosis was elaborated using for the gene selection a set of 55 patients (Xu et al., 2017). In conclusion, we can say that as far as it is reflected in the current literature, the size of the initial training sets used to identify candidate gene markers for CRC survival is small and the overlap between the published gene signatures is very reduced and inconsistent. To address these critical problems, we constructed a large, well-standardised, integrated data set of 1273 tumour samples with survival information, which was used to identify genes that had a clear change in expression in the middle and late stages of CRC and were consistent markers of the disease-outcome and patient-risk.

With respect to the specific genes proposed as CRC survival markers, we want to underline that our study does not pretend to provide a fixed gene signature for prognosis and risk prediction, like the reported signatures of 7-genes, 15-genes or 113-genes (Nguyen et al., 2015) (Chen et al., 2017) (Xu et al., 2017) but instead we propose a robust set of genes ranked according to their predictive power of CRC survival. In this way, an ordered list of 200 genes including the best survival markers is presented: 100 genes for which up-regulation marks "poor survival" and 100 genes for which up-regulation marks "good survival". We think that this approach is more useful, since it allows an open selection of different number of genes for further purposes or investigations (for example, for additional tests with other CRC clinical cohorts). In fact, we used the 100 most significant genes, up-regulated with the progression of CRC, to build the risk predictor (presented in **Fig:4.7**); and we used the top 5 or top 10 genes of this list for the external validations with different independent datasets.

Another relevant comment is that, as reminded above, we constructed the risk

predictor using the genes that showed up-regulation correlated with poor prognosis. This was done because in the selection of biomarkers it is better to use the ones that provide a positive signal (i.e. "gain-of-function" factors) than the ones that provide a negative signal. Therefore, all the gene survival markers that we proposed were detectable as overexpressed in the CRC patients with high risk. The fact that they give a positive signal will also make easier their detection by standard biomolecular protocols (PCR, ELISA, immunohistochemistry, etc).

Finally, we are investigating the biological meaning of the genes found as best predictive and prognostic markers. We are focusing our efforts in the top 10 for which up-regulation marked poor survival: DCBLD2, PTPN14, LAMP5, TM4SF1, NPR3, LEMD1, LCA5, CSGALNACT2, SLC2A3, GADD45B. The analysis of the literature reveals some relevant observations. For example, the transmembrane protein DCBLD2 (ESDN), member of a family of neuropilin-like proteins, is a novel regulator of mitotic and metabolic effects of insulin, and it modulates signal transduction through regulation of the insulin receptor interaction with its adaptor proteins (Li et al., 2016). The importance of insulin regulation in the function of our digestive system is clear, and this adds extra value to the proposal of DCBLD2 as a CRC survival marker.

Other genes within the top rank have been recently involved in cancer progression, like the case of SLC2A3 (GLUT3) a glucose transporter that mediates glucose utilisation and glycogenolysis, which is induced during epithelial-mesenchymal transition and promotes tumour cell proliferation (Masin et al., 2014). Recent publications have also proposed the role of some other genes found as prognostic markers, like the case of LAMP5 that has been included in a multigenic assay to predict recurrence for gastric cancer patients after surgery (Lee et al., 2014). As a final example, GADD45B (growth arrest and DNA-damage-inducible 45 beta) is a gene that responds to environmental stresses, associated with cell growth control, apoptosis and DNA damage repair response. GADD45B overexpression has been recently correlated with shorter overall survival in colorectal carcinoma (Wang et al., 2012). Moreover, a recent integrative analysis of multiple colon cancer gene-expression-based subtype classifiers reported that one of the three highest scoring genes included in several classifiers was GADD45B (Sztupinszki and Győrffy, 2016).

Despite all these positive findings that correspond to the biological value and the support of the genes identified as most significant markers of CRC survival, there are some possible limitations of the results, beginning with the general observation about the frequent heterogeneity of the colorectal tumours (Linnekamp et al., 2015) (Sameer, 2013). In fact, it is clear from the anatomical pathology that CRC can affect quite different regions of the digestive tract: ascending colon, transverse colon, descending colon, sigmoid colon and rectum. The causal genes that drive tumours in these different regions may not be the same, and most CRC studies do not enter into a detailed separation of these regions (Bijlsma et al., 2017).

The variability due to the different staging of the tumours is another factor that can bring limitations to any CRC study; but in this case we clearly indicated

that our work searched for genes that were candidate prognostic markers for CRC in stages III and IV. A final reason for the limitations of the results may be an over-adjustment to the tested data sets. To avoid this kind of limitations, we built a large well-normalised data set with more than a thousand samples, performed a cross-validation analysis on that set, and also explored the validity of the gene markers in two other independent sets.

In conclusion, we consider that the results presented in this work provide strong support and a solid rationale for the prognostic value of a new set of genes in CRC and for their potential to predict colorectal tumour progression and evolution towards stages III and IV. The final proposed set of gene survival markers includes an open list of one hundred up-regulated genes, with a robust statistical estimation of the value of each one. In this way the set of genes is clearly ranked, being the top in the list the ones that provide best prognostic strength and the ones that can be introduced to build smaller predictors. In fact, our results showed that a selection of the top 5 genes applied to independent external cohorts provided very good separation of CRC samples in two distinct groups of high and low risk.

Chapter 5

Integrative transcriptomic profiling of Colorectal Cancer (CRC) Consensus Molecular Subtypes (CMS) with survival data and relative characterisation of a EMT gene signature associated to P21 knockout, CDKN1A (-/-)

5.1 Motivation

Metastasis formation is based on a multi-step process also known as the invasion-metastatic cascade. It starts with the dissemination of cancer cells from the primary tumour site, their survival in the circulatory system, extravasation, and eventually recolonisation at a distant organ site, thus eventually generating a secondary tumour.

All of the individual steps require specific features of tumour cells, which are largely connected to the epithelial-mesenchymal transition (EMT) and cancer stem cell phenotype. The progression of the healthy colon epithelium to invasive and metastatic carcinoma is strongly associated with the process of EMT, and the ability of tumour cells to survive under non-adherent conditions is well known.

However, the elucidation of the detailed mechanisms and regulators driving metastatic spread in patients remains a significant focus of translational CRC research.

The cyclin-dependent kinase inhibitor p21 represents a negative regulator of both cell cycle progression and gene expression, as reported in literature in (Abbas and Dutta, 2009).

An unregulated passage of the cell through the G1/S checkpoint by downregulation or loss of p21 might induce aberrant proliferation and thus trigger tumour transformation.

In CRC downregulation of p21, the expression has been reported to correlate with the development of metastases and poor patient survival (Abbas and Dutta, 2009). Thus, silencing of p21 seems to be of outstanding importance for the unrestricted proliferation of cancer cells.

Previous studies have reported that there is a direct relationship between cell lines HCT116 and isogenic HCT116 p21 KO with Colorectal Cancer (CRC) subtypes defined as Consensus Molecular Subtypes, CMS1 and CMS4, respectively. It has also been reported that some genes like VIM and p21 are related to Epithelial to Mesenchymal Transition (EMT), which is the main hallmark of CMS4. A difference in the gene expression between the subclasses is expected.

Thus, the central objective is to demonstrate whether there is a relationship between HCT116 and isogenic HCT116 p21 KO cellular lines and the subtypes CMS1 and CMS4, respectively. To corroborate our hypothesis, we use the microarray and RNAseq datasets described in Chapter 4 as validation for the results obtained from the analysis of differential expression.

We classify the samples in CMS subtypes using already developed classifiers. This allows us to verify if the biological observations in cell lines correspond with data obtained from human patients. Marker genes will be validated measuring if the genetic pattern found in HTC166 and HTC166 KO cell lines are similar to the one obtained in CMS1 and CMS4 subtypes, respectively.

At the same time, the relation of the discovered marker genes with risk and survival is evaluated. The already proposed difference in the survival outcome between CMS4 and the rest of the subtypes will be investigated as another validation point.

According to (Guinney et al., 2015) the CMS4 subtype reflects the gene signature of mesenchymal cells together with TGF-β signalling and matrix remodelling. Interestingly, CMS4 subtype was also majorly correlated with drug resistance and increased tumour budding (Trinh et al., 2018). In the consensus molecular subtype classification of colorectal adenoma, there was not CMS4 subtype since an invasion-associated stroma does not exist (Komor et al., 2018).

So far, the full gene signature of HCT116 p21 KO cells has not been determined. For a better understanding of the p21 KO model particularly as a preclinical model for the analysis of therapeutic response, we aimed to evaluate if the p21 KO is strong enough to switch the molecular subtype of the microsatellite unstable mutator HCT116 cell line.

5.2 Materials and Methods

5.2.1 General workflow of the study

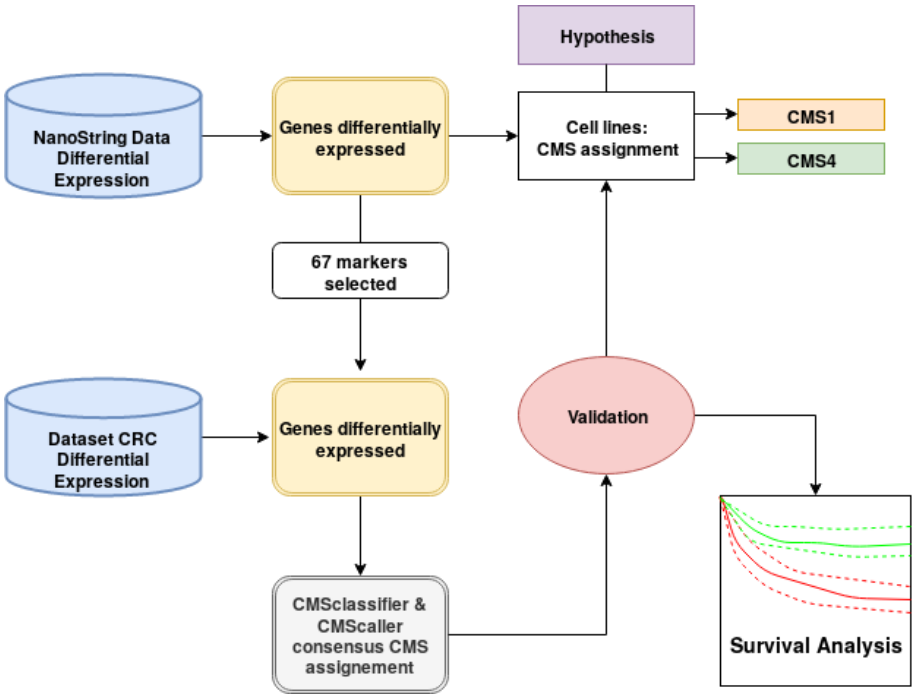


Figure 5.1: CRC methodology and workflow.

In **Fig: 5.1**, the process we followed is portrayed. First, we have the data from the comparison between cell lines and the microarray and RNAseq CRC series. The first step obtains the top genes that show differential expression between the WT cell line and the knock out.

This gave us 67 candidate markers which were related to CMS1 or CMS4 subtypes. If the expression of the gene is higher for the CMS1 like cell line, it is defined as a CMS1 positive marker. Otherwise, if the expression is higher for CMS4 like cell line (the p21 KO), then it is defined as a CMS4 positive marker.

The study involving microarray and RNAseq data is done in order to validate. For patient classification in CMS subtypes we used the consensus between two class predictors, CMSclssifier and CMSCaller which are described in the next sections.

5.2.2 Cell lines and biological study

HCT116 WT and isogenic HCT116 p21 KO cell lines were used for this study. HCT116 cells were obtained from the American Type Culture Collection (ATCC, Manassas, VA, USA), while the HCT116 p21 KO cells were obtained from Bert Vogelstein (The Johns Hopkins University School of Medicine, Baltimore, MD, USA). Mycoplasma-free status of the cell lines was confirmed and their genotypes were au-

thenticated using Multiplex Cell Authentication by Multiplexion (Heidelberg, Germany).

NanoString gene expression analysis

In this study, we will analyze the gene signature that characterizes HCT116 against HCT-116 p21 KO cell lines. The results complement the previous work which relate the HCT116 WT cell line with CMS1 colorectal cancer subtype and HCT-116 p21 KO cell line with CMS4 due to the metastasic capabilities of this modified cell line.

The genes discovered from the analysis of differential expression considering nanostring tools will provide a expression pattern that characterizes CMS1 against CMS4 cell lines.

Gene expression analyses were performed using the NanoString PanCancer Progression panel gene expression assay with 100 ng of total RNA as input according to the manufacturer's instructions. Raw data (counts per analysed gene) as obtained by the nCounter® FLEX Analysis System were then processed according to the following description.

Expression values were normalised to human B2M expression. Gene expression of HCT116 p21 KO cells is shown as the relative fold expression compared to HCT116 controls.

5.2.3 CMSclassifier

CMSclassifier is a multi-class classifier developed to predict cancer molecular subtypes (CMS) (Breiman, 2001). Two baseline models can be considered: Random Forest (RF) where data is row-centered and Single Sample Prediction (SSP) based on Pearson correlation similarity (Guinney et al., 2015).

The RF algorithm builds an ensemble of weak trees. For each tree, only a subset of random variables is allowed to split the branches of the tree. Besides, a bagging strategy is implemented by resampling the pattients with replacement. In this way, diversity among the classifiers is induced and the combinarion is expected to reduce the error prediction and the variance of the classifier (Chen and Ishwaran, 2012).

The generalisation error converges to a limit as the number of trees grows (Breiman, 2001). This type of classifiers have been widely applied to genomic problems, and performs well with high dimensional and correlated data.

CMSclassifier is able to work with unbalanced classes. This is a requirement for our problem because the proportion of CMS classes is unequal. This problem is overcome in Random Forest by resampling each class (Guinney et al., 2015).

5.2.4 CMSCaller

CMScaller is a CMS classifier optimised for pre-clinical models systems of colorectal cancer (CRC) (Eide et al., 2017). It was developed to make robust classification across gene expression platforms.

CMScaller exploits the evidence that the gene expression profiles are highly coregulated. Therefore, the dimensionality can be reduced removing redundant and noisy features. Erroneous or missing measurements can be estimated considering the observed value for co-regulated genes. To this aim, the nearest template prediction algorithm (NTP) is applied (Eide et al., 2017).

NTP algorithm is a flexible prediction method, less sensitive to difference in experimental or analytical conditions, applicable to single patient's gene expression, that provides measure of prediction confidence. It makes class prediction using only a list of signature genes and a test dataset, for each single patient's gene expression data. The method can be flexibly applied to cross-platform, cross-species, and multiclass predictions without any optimisation of analysis parameters (Hoshida, 2010).

Both algorithms provide a p-value for each sample measuring the confidence in the classification. If the p-value is above a certain threshold, the classification is rejected.

5.2.5 geNetClassifier

geNetClassifier is a network-oriented and data-driven bioinformatic tool implemented in an R package, named geNetClassifier. This library gives complementary information to the classification. Particularly, it allow us to search for association of genes and diseases based on the analysis of genome-wide expression data derived from next generation sequencing (NGS) technologies (microarrays or RNA-Seq). The algorithm proceeds as follows:

- Find genesets associated to a given pathological state.
- Identify minimal subsets of genes within these genesets that unequivocally differentiates and classifies the disease subtypes
- Provide a measurement of the discriminant power of these genes.
- Build a gene network that characterises each of the disease subtypes

The estimation of the gene networks related to specific disease subtypes that include parameters such as gene-to-gene association, gene disease specificity and gene discriminant power can be considered to draw gene-disease maps and to unravel the molecular features that characterize specific pathological states (Aibar et al., 2015).

5.2.6 Cox regression to detect epistatic interactions

This kind of algorithms are able to model the correlation structure among genes removing redundant features. However, it has been shown that the gene regulatory pathways of cancer involves complex non-linear interactions between genes. Although models based on linear combinations of genes may perform well for most complex interactions, the effect of non-linear interactions on survival should be studied in depth.

From a biological point of view, the study of epistatic interactions helps to develop efficient anticancer therapy. In particular, the current therapies have limited efficacy due to toxicity and to the development of drug resistance. Recently, some therapies have exploited the lethal interaction between certain pairs of genes. Synthetic lethality is a negative interaction between genes where the co-inactivation of two genes simultaneously results in celular death. This can be considered to kill selectively cancer cells.

The main challenge of this section is to discover relevant interactions composed of genes that do not have a significant association to survival by themshelves.

Non-linear interactions between genes may be modeled with a multivariate cox regression algorithm. To this aim, the product of individual variables is included as a new covariate. To avoid overfitting, the algorithm is run iteratively on a random subset of the input using a bootstrap strategy. Besides, the candidate genes that can be considered to detect interactions have been reduced using the additive models introduced in chapter 2. This preprocessing step will help to improve also the computational efficiency of the algorithm. Finally, the non-linear interactions are ordered considering the p-value for the coefficient in the Cox regression.

Let $h(t \mid \mathbf{x})$ be the hazard function at time t conditioned to the observed covariates $\mathbf{x}$. As we are looking for pair of variables, we have to adjust a model for each posible combination of two variables. The hazard function can be written as:

$$h(t \mid \mathbf{x}) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) \tag{5.1}$$

where each β_i determines the increment in the hazard ratio (log scale) if the expression of the gene x_i changes one unity. β_3 is the coefficient for the non-linear interaction term. If $\beta_3 > 0$ then, as the non-linear interaction term increases the risk grows. Conversely, if $\beta_3 < 0$ then the risk decreases with the interaction term.

The β_i parameters are determined by optimisation of the partial log-likehood as has been explained in chapter 2. The Cox regression will provide a risk score, that will allow us to stratify the patients according to their risk. A statistical test is computed for each β_i parameters that will determine if the interaction term is associated to the patient risk. Finally, a likehood ratio test will help to determine if the non-linear model considering interactions helps to explain better the risk of patients than the additive linear model.

The implementation of this method is based on the **coxph** function from **survival** R package (Therneau, 2014).

5.3 Results

5.3.1 Nanostring: differentially expressed genes

In this section, a list of genes is obtained that characterise colorectal cancer cell lines (HTC116 WT) against p21 KO (HTC116 p21 KO). The genes presented in Tab: 5.1 for the up-regulated genes and 5.2 for the down-regulated genes are the basis for our later analysis.

Symbol	DE logFC	DE statistic	DE adj.p.value	CA statistic	CA p.value	Assignment
VIM	7,8488	1642,29	0	-14,38531	3,66E-35	CMS4
SPARC	8,2699	1506,55	0	-16,85454	3,35E-45	CMS4
NRP1	3,3402	409,97	0	-4,56500	6,68E-06	CMS4
SNAI2	27,3859	333,53	0	-9,29921	1,88E-18	CMS4
THBS1	2,2342	281,48	0	-11,19640	2,88E-25	CMS4
TNC	32,2267	251,40	0	-6,66385	8,60E-11	CMS4
PLEKHO1	1,4848	140,07	0	-6,97527	1,36E-11	CMS4
FGF2	1,5965	124,21	0	-3,96364	9,13E-05	CMS4
FERMT2	1,3288	113,17	0	-19,35596	6,07E-58	CMS4
SPHK2	1,7297	103,01	0	-4,01767	7,11E-05	CMS4
PDGFA	1,4772	97,27	0	-4,10453	5,09E-05	CMS4
MMP13	29,2676	85,20	0	-3,75387	0,000199241	CMS4
NRP2	2,0484	75,50	0	-12,63576	5,39E-31	CMS4
CD24	1,0741	72,44	0	-3,53611	0,000463411	CMS4
FGF18	1,9665	69,92	7,08E-16	-8,42034	6,43E-16	CMS4
FST	33,0730	69,83	7,08E-16	-9,78526	1,95E-20	CMS4
CREBBP	0,9871	64,87	4,79E-15	-6,56140	1,76E-10	CMS4
PIK3CA	0,9651	59,67	6,75E-14	-7,47992	4,67E-13	CMS4
ZEB1	1,4050	52,43	2,51E-12	-17,30530	1,04E-50	CMS4
P3H1	0,8713	50,98	5,11E-12	-7,76292	7,27E-14	CMS4
FHL1	1,2050	50,10	7,81E-12	-17,49613	1,26E-51	CMS4
OLFML2B	1,6033	48,00	2,25E-11	-13,44126	8,70E-34	CMS4
LAMC1	0,8418	47,21	3,33E-11	-15,85795	1,57E-42	CMS4
PLCG1	0,8340	47,14	3,44E-11	-12,75009	1,76E-31	CMS4
CYP1B1	5,4332	42,84	2,97E-10	-10,25702	4,30E-22	CMS4

Table 5.1: UP regulated genes associated to CMS4 subtype.

Table: 5.1 shows the list of genes differentially expressed and up regulated obtained through Nanostring tool.

DE logFC is the fold change in logarithmic scale, **DE statistic** is the value of the statistic for the diferential expression **(DE)** analysis of nanostring, and DE adj.p.value is the p-value, adjusted for multiple tests.

CA statistic and **CA p.value** correspond to the statistic computed for the classification of each sample and p-value of the statistical test.

Symbol	DE logFC	DE statistic	DE adj.p.value	CA statistic	CA p.value	Assignment
ID2	-4,6551	972,56	0	3,40259	0,000754208	CMS1
TP53	-4,0625	804,76	0	5,09264	5,66E-07	CMS1
IL18	-4,7962	734,54	0	8,07327	9,92E-15	CMS1
RAMP1	-32,1818	667,64	0	3,52531	0,000485142	CMS1
CDH1	-3,3914	617,59	0	-1,01830	0,309246382	NA
CDKN1A	-3,6018	552,81	0	3,53396	0,000474390	CMS1
ITGA3	-3,0876	352,34	0	5,35542	1,65E-07	CMS1
TJP3	-3,2495	350,49	0	7,29310	2,30E-12	CMS1
S100A14	-3,2206	294,47	0	11,67354	1,10E-26	CMS1
GRHL2	-3,8916	291,45	0	4,57220	6,43E-06	CMS1
AP1M2	-2,0948	275,11	0	7,59930	2,04E-13	CMS1
TMC6	-3,1659	232,26	0	5,20052	3,31E-07	CMS1
HKDC1	-7,4257	211,12	0	3,69186	0,000253606	CMS1
TOM1L1	-1,5895	160,92	0	8,63937	1,54E-16	CMS1
CD44	-1,6368	157,63	0	6,24127	1,31E-09	CMS1
GDF15	-4,4624	148,17	0	2,67067	0,007888663	CMS1
MET	-1,5415	143,88	0	3,41445	0,000712745	CMS1
RAB25	-2,0369	143,43	0	5,24247	2,57E-07	CMS1
PRSS22	-3,0175	137,87	0	2,63069	0,008901344	CMS1
EIF4EBP1	-1,4701	134,55	0	8,53346	3,18E-16	CMS1
UTS2	-5,5983	131,46	0	2,88800	0,004193874	CMS1
EPS8L1	-1,9951	117,66	0	5,65831	3,82E-08	CMS1
PDGFC	-1,2773	111,44	0	-17,91141	4,43E-49	CMS4
SLC12A6	-1,3482	111,10	0	2,75854	0,006134769	CMS1
ADAP1	-1,2536	96,11	0	4,60927	5,58E-06	CMS1
PTK2B	-1,3964	93,79	0	5,55089	5,57E-08	CMS1
ANXA2P2	-1,3230	87,76	0	8,67459	1,39E-16	CMS1
TSPAN1	-2,0643	87,18	0	8,39079	9,52E-16	CMS1
ICAM1	-8,8530	86,79	0	5,15507	4,62E-07	CMS1
IRF6	-1,1005	84,27	0	3,09659	0,002095301	CMS1
CXCL8	-2,6455	82,36	0	7,15129	4,11E-12	CMS1
ITGA2	-1,0669	79,92	0	3,68571	0,000261304	CMS1
CDS1	-1,0436	75,07	0	6,67325	1,02E-10	CMS1
ITGB4	-1,3160	69,67	7,08E-16	6,66385	1,11E-10	CMS1
DSC2	-0,9811	67,80	1,40E-15	7,64402	1,99E-13	CMS1
ADAM15	-0,9488	56,97	2,62E-13	2,75734	0,006115982	CMS1
SH2D3A	-0,9293	52,79	2,12E-12	6,75447	5,91E-11	CMS1
EGLN3	-2,6701	50,90	5,29E-12	9,87423	2,70E-20	CMS1
SLPI	-3,2923	50,06	7,90E-12	4,31395	2,11E-05	CMS1
FRAS1	-1,0718	43,20	2,50E-10	4,52480	8,60E-06	CMS1
TYMP	-0,9639	42,90	2,90E-10	7,81649	1,10E-13	CMS1
PRF1	-2,1367	42,56	3,38E-10	10,41441	5,82E-22	CMS1

Table 5.2: Nanostring genes defined as DOWN regulated.

Table: 5.2 corresponds to the list of genes down regulated obtained through Nanostring tool.

The association between the CMS subtypes and the previous list genes is studied.

The list of genes will be validated considering the gene expression patterns in the colorectal cancer series described in Chapter 3. The amount of data that these series have will provide an excellent workspace for analysing the diferences between subtypes and the capability of these genes to describe differences.

5.3.2 Validation of gene markers

In order to study the difference between both cell lines, a differential expression analysis was done for each class. The hypothesis is that CMS4 subtype is strongly related to p21 KO cell line.

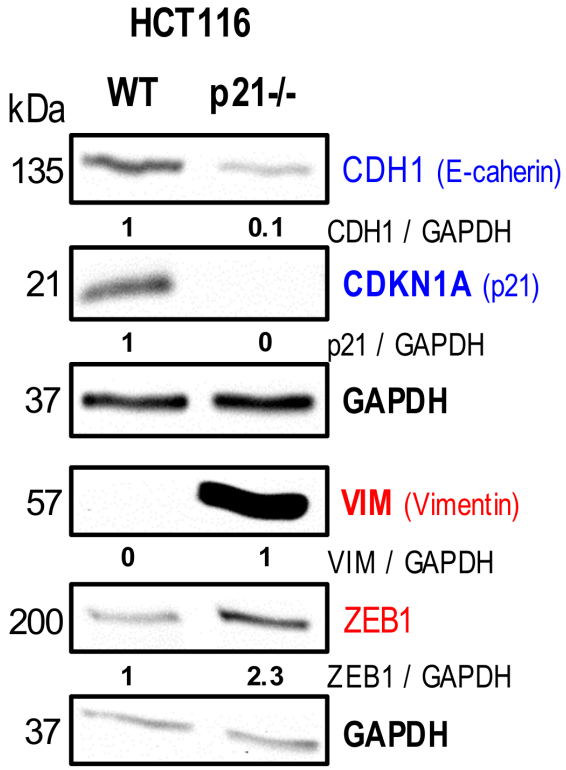


Figure 5.2: Comparison of colon cancer HTC116 cells wild type (WT) versus p21ko (KO), standard markers.

Colon cancer HCT116 WT cells versus p21 KO cells

In Fig: 5.2. fepresentative western blot of HCT116 WT and HCT116 p21 KO cells for genetic background and 3 EMT markers (ECadherin, Vimentin, ZEB1); n $\geq$ 3. Fold expression is represented relative to GAPDH loading control; n $\geq$ 3. The EMT markers are usually used to define cells associated to CMS4.

In this case, ECadherin is more present in WT cells, the function of this gene is related with membrane adhesion. The the lower expression level of this gene in p21 KO contribute to cancer progression by increasing proliferation, invasion, and/or metastasis. p21 is not present in KO as expected.

Vimentin overexpression is expected. It is an organizer of a number of other

critical proteins involved in cell attachment, migration, and signaling.

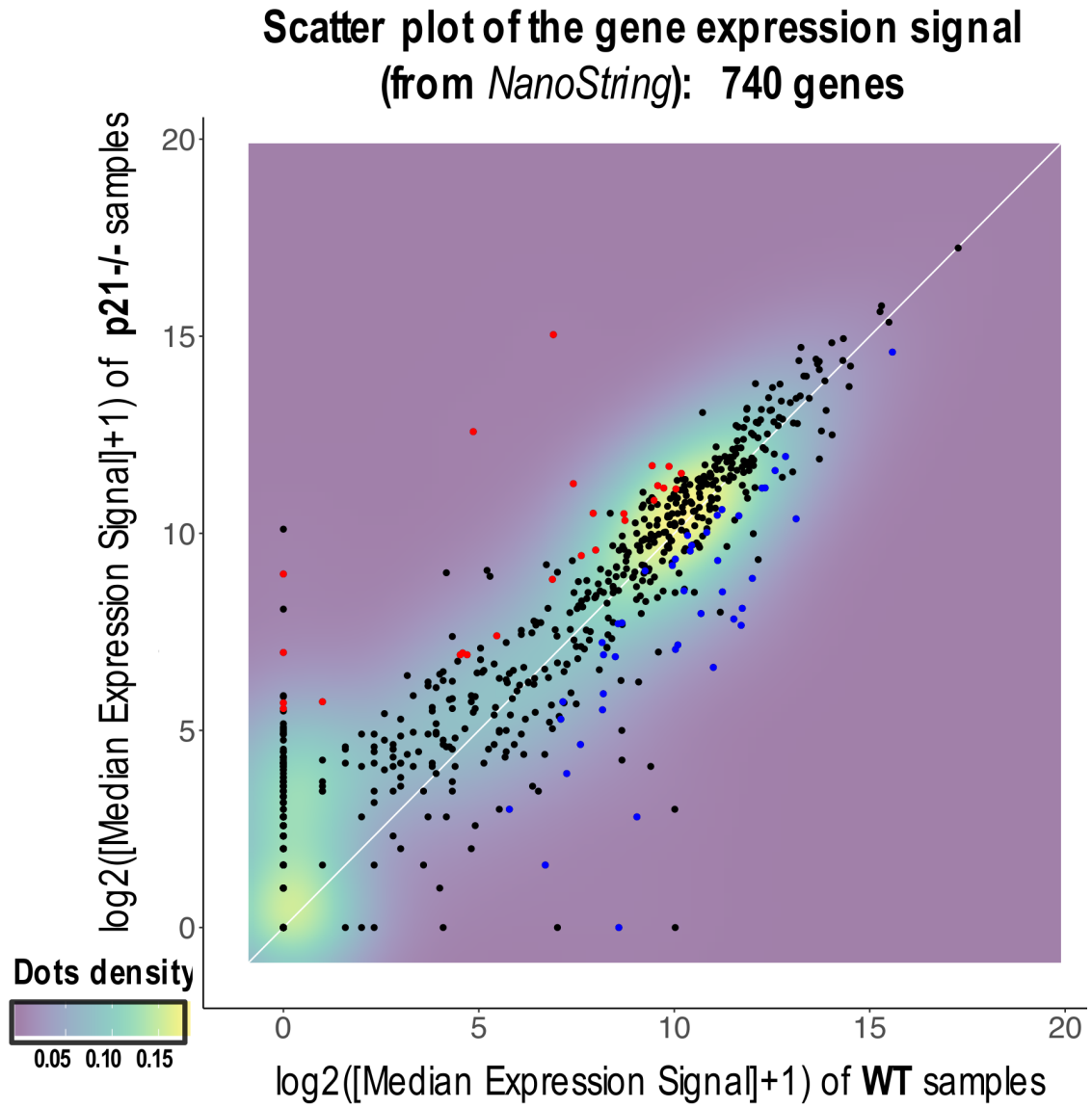


Figure 5.3: Comparison of colon cancer HTC116 cells wild type (WT) versus p21ko (KO), scatter plot of the global expression signal.

Fig: 5.3. Scatter blot of the global expression signal. The graph shows a linear relation between the WT and p21 KO. The red and blue dots are the up regulated and down regulated gene markers described in previous tables. The blue dots are genes overexpressed in wild type cell line. The red dots are genes overexpressed in p 21 KO cell line.

The dots in values x=0 or y=0 are typical in distributions of this kind of platforms, meaning a lack of lectures for that genes in one or another cell line.

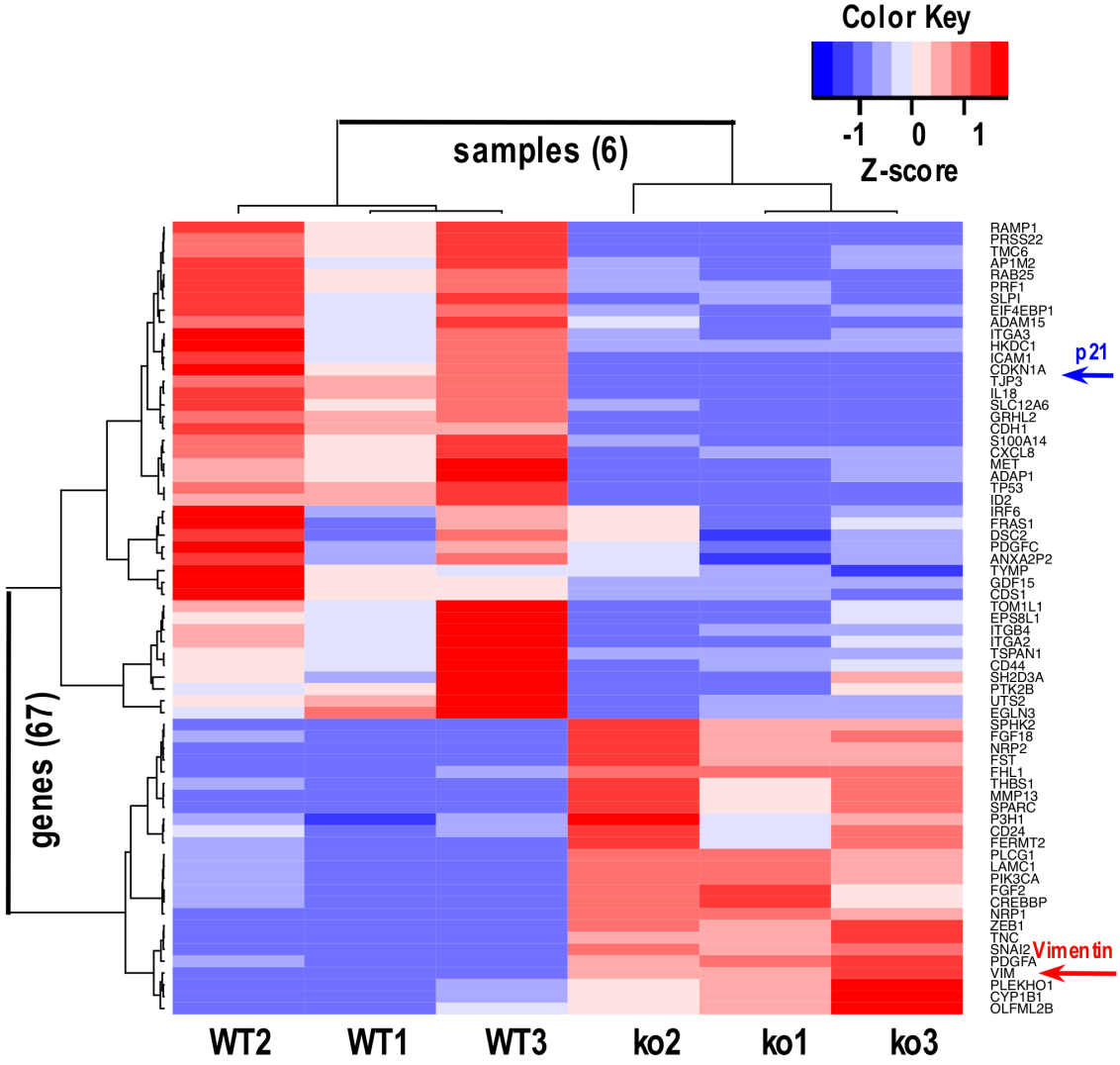


Figure 5.4: Comparison of colon cancer HTC116 cells wild type (WT) versus p21ko (KO), heatmap with the expression profile of 67 genes selected as most significant .

Colon cancer HCT116 WT cells versus p21 KO cells

Fig: 5.4 shows a heatmap with the expression profile of 67 genes selected as most significant. VIM and p21 are highlighted. In the heatmap we can see the differences between one cell line (WT) and another (KO). Next experiments will show that UP and DW genes represent the subtypes CMS1 and CMS4.

The genes marked as blue in WT cell line and as red in KO are the defined in the tables as KO and CMS4 positive markers.

The genes marked as red in WT cell line and as blue in KO are the defined in the tables as WT and CMS1 positive markers.

General function	Genes	Query List	Reference List	Enrichment p.value	Similarity	Functional Terms assigned (concurrent enrichment)
Fibroblast phenotype, cytoskeleton, Cancer	PIK3CA, PLCG1, FGF18, FGF2, PDGFA	5 (23)	50 (34208)	1.79E-10	0.7500	GO:0008543:fibroblast growth factor receptor signaling pathway (BP); Kegg:04810:Regulation of actin cytoskeleton; Kegg:05218:Melanoma
Extracellular matrix changes	THBS1, SPARC, TNC, OLFML2B, LAMC1, MMP13	6 (23)	134 (34208)	3.08E-10	0.9129	GO:0031012:extracellular matrix (CC); GO:0005576 extracellular region (CC)
Cell adhesion	THBS1, TNC, LAMC1, PDGFA, PIK3CA	5 (23)	197 (34208)	1.86E-07	0.8944	Kegg:04510:Focal adhesion
Cell contact regulation	CD44, ITGA3, ITGB4, ITGA2, ADAM15, PTK2B, RAMP1, DSC2	8 (40)	93 (34208)	1.57E-13	0.7335	GO:0009986:cell surface (CC); GO:0007160:cell-matrix adhesion (BP); Kegg:04512:ECM-receptor interaction; GO:0007229:integrin-mediated signaling pathway (BP); Kegg:05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Cancer genes	CDKN1A(p21), TP53, CDH1, MET, PDGFC, TYMP	6 (40)	63 (34208)	1.12E-10	0.7717	Kegg:05218:Melanoma; Kegg:05219:Bladder cancer
Cell adhesion, cytoskeleton	ITGA3, ITGB4, ITGA2, MET, PDGFC	5 (40)	111 (34208)	1.97E-07	0.7500	Kegg:04510:Focal adhesion; Kegg:04810:Regulation of actin cytoskeleton

Figure 5.5: Comparison of colon cancer HTC116 cells wild type (WT) versus p21ko (KO), functional enrichment analysis of the selected genes using concurrent annotation.

Colon cancer HCT116 WT cells versus p21 KO cells

Fig: 5.5. Functional enrichment analysis of the selected genes using concurrent annotation. The image shows the relationship between genes selected from WT cell line and CMS1 cell functions. Moreover, the KO genes are associated with CMS4 cell functions.

Both the red an blue genes define cell functions related with cancer. The table shows a high amount of functions related with cell adhesion, cytoskeleton, cell matrix, cell contact...

This kind of functions are strongly related with our hypothesis, indicating that the diferences in this cell lines are the same as described between CMS1 and CMS4.

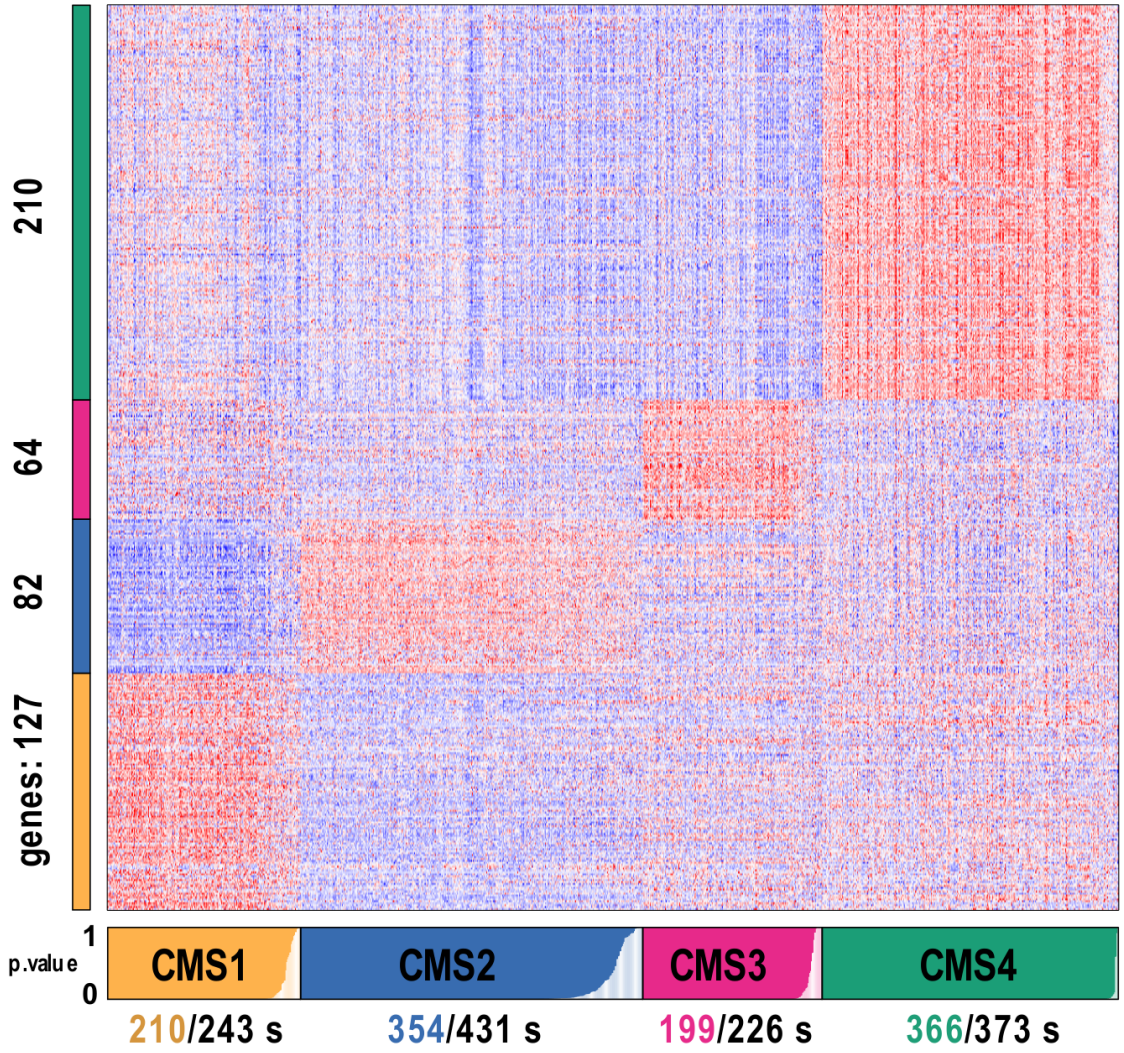


Figure 5.6: Comparative analysis of the expression profiles and CMS subtypes. Heatmap of a cohort of 1273 CRC patients according to the Consensus Molecular Subtypes (CMSs)

Fig: 5.6. CMScaller (Eide et al., 2017) and CMSclassifier have been considered to classify samples in four groups corresponding to each one of the Consensus Molecular Subtypes (CMSs). The list of genes recommended by the authors are considered by the classifiers. Both classifiers are combined to predict the CMS classes.

The heatmap shows the gene expression profiles for the genes considered by the classifiers. Each CMS subtype is characterised by the expression level of a group of genes. The numbers referenced in the class labels are:

First number is the number of samples classified by both predictors as the same class.

Second number is the samples assigned by CMScaller alone.

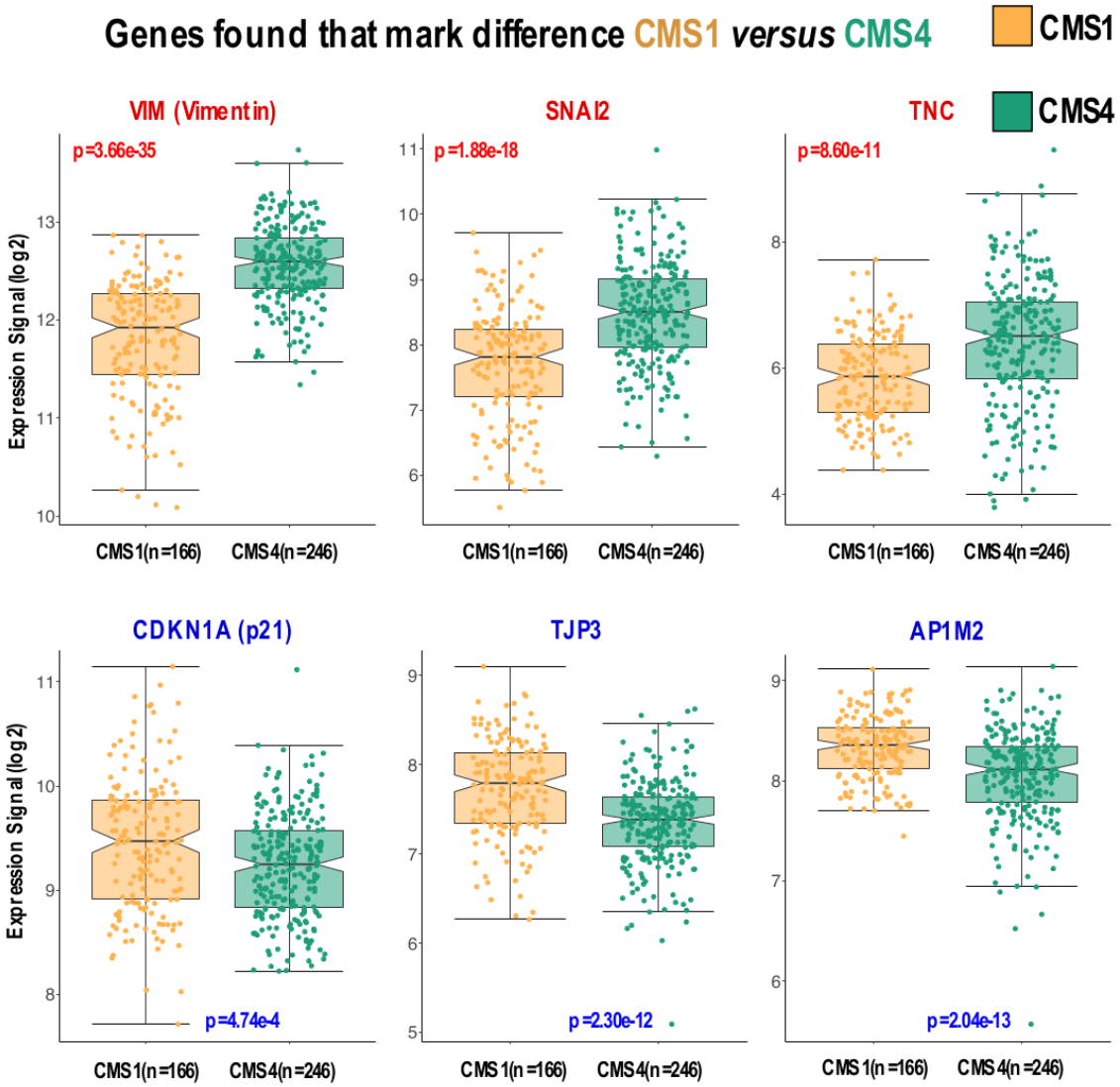


Figure 5.7: Comparative analysis of the expression profiles and CMS subtypes. Association with the gene signature identified in the analysis of HTC116-WT versus HTC116-p21ko

Fig: 5.7 shows the expression level for the top three genes UP and DOWN regulated genes from tables Tab: 5.1 and 5.2. The p-value determines if the difference betweeen the median expression levels for both boxplot (CMS1 and CMS4) is statistically significant. UP genes (name in red) are upregulated in p21 KO vs WT. DOWN genes (name in blue), the down regulated genes, like p21 are down in KO vs WT.

Therefore, the experiments suggest that genes upregulated in p21 KO are related with CMS4 and genes upregulated in WT are related with CMS1.

5.3.3 Survival analysis of CMS subtypes

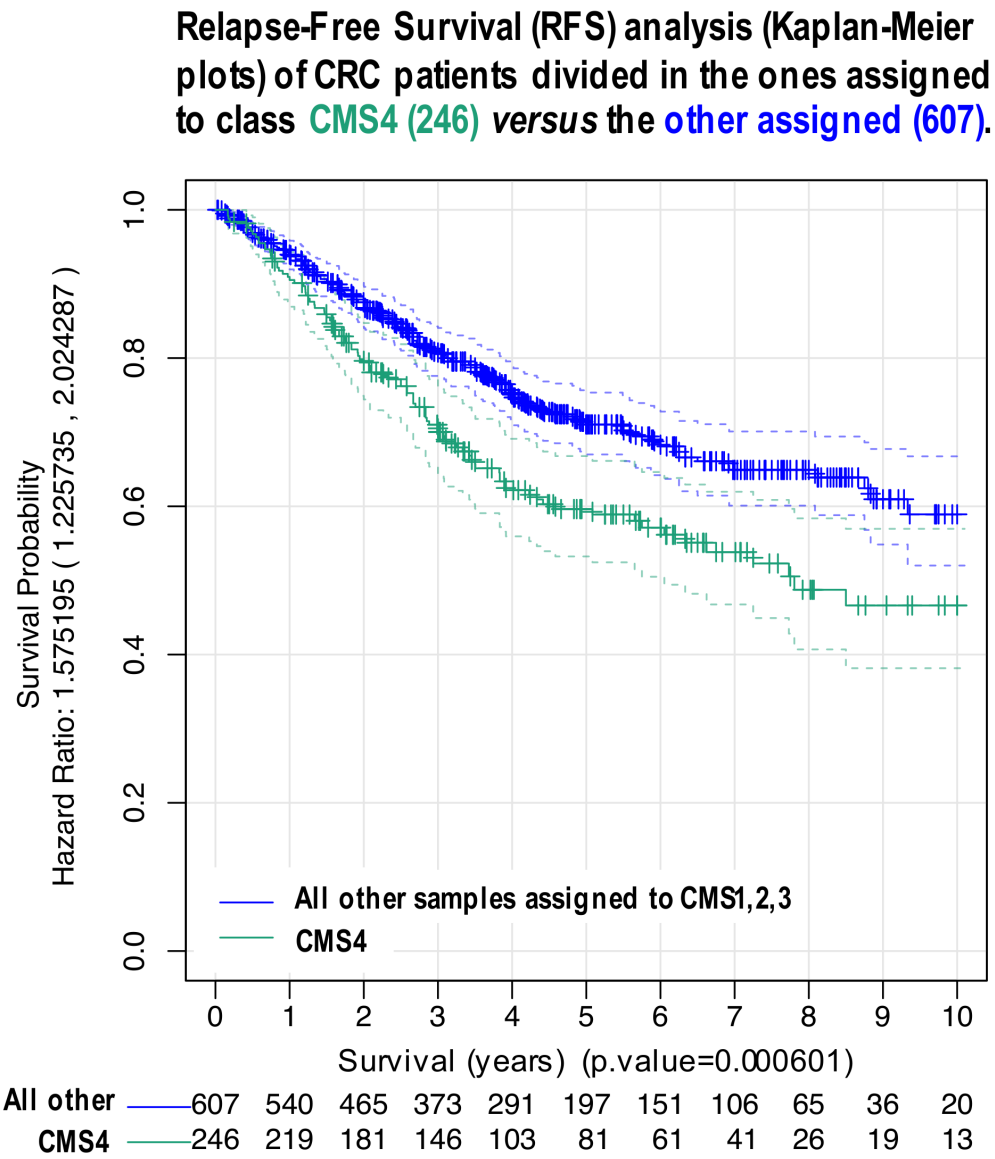
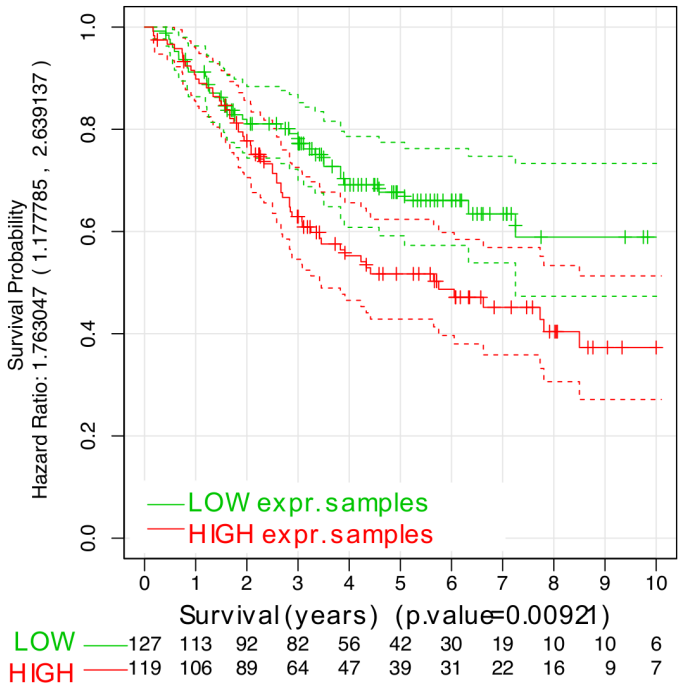


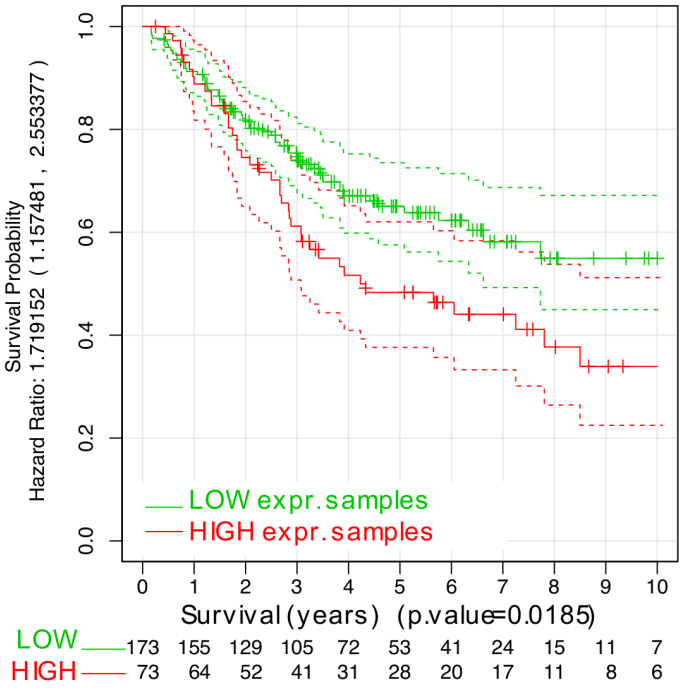
Figure 5.8: Analysis of the survival of 246 CRC samples identified as CMS4 vs the rest.

Fig: 5.8. The ensamble of CMS classifiers has been applied to categorize patients in two groups, CMS1 and CMS4. Next, the survival curves for both groups are built. The logrank test is computed to check if the difference is statistically significant. CRC samples identified as CMS4 versus the ones assigned to other CMSs (i.e. CMS1,2,3)

The Kaplan-Meier plot shows that the curves corresponding to CMS4 and the other CMS subtypes have differences as previously reported. This result has been obtained by other authors in the literature (Dienstmann et al., 2017) (Guinney et al., 2015).



(a) VIM in 246 CMS4



(b) SNAI2 in 246 CMS4

Figure 5.9: Analysis of the survival of 246 CRC samples, gene by gene.

Survival of two representative markers from our list is studied. **Fig: 5.9** shows the survival curves for the two groups obtained stratifying the patients by VIM expression level. This method is explained in detail in section 2.2.4. Red curve corresponds to high gene expression level and green with low. The higher the expression level they have, the worst is the prognosis for VIM and SNAI2 gene.

5.3.4 Functional enrichment analysis of CMS predicted subtypes

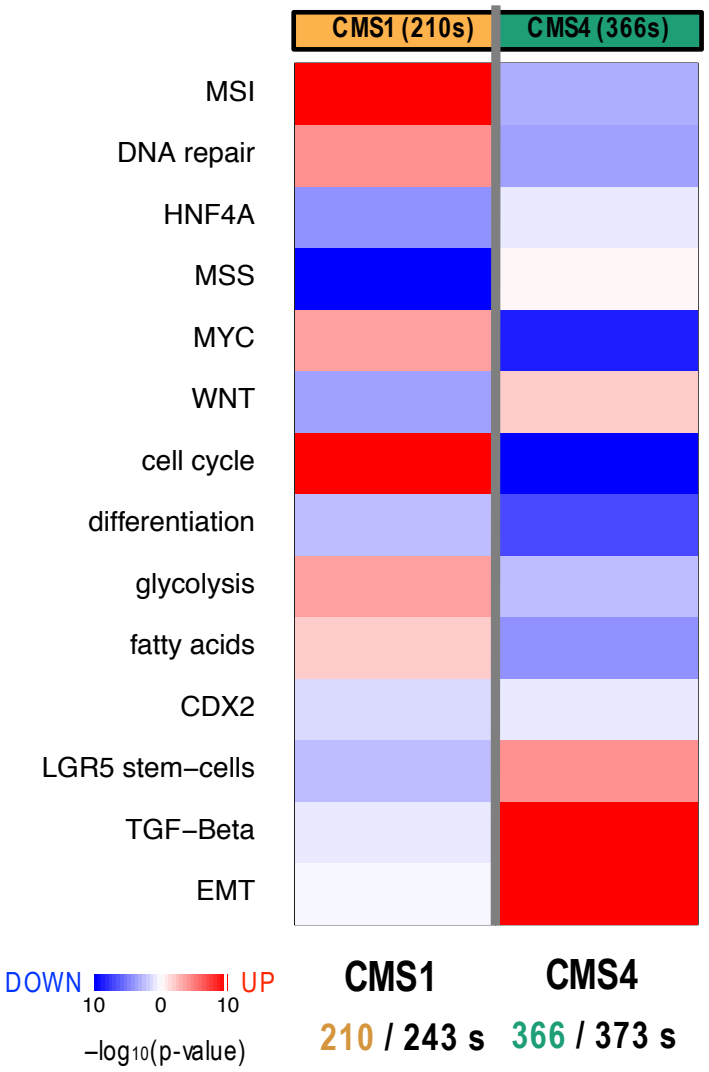


Figure 5.10: Gene-set functional enrichment analysis of CMS1 versus CMS4 based on the gene expression profile of the patients assigned to these two subtypes

CMS1 versus CMS4 classes showing the strong relationship of our genes with cancer-related celular functions. The figure shows that the relationship of CMS4 and Epithelial to Mesenchymal Transition (EMT) is high. As expected, the replication speed is slower in CMS4.

The CMS classes used are the predicted in microarray series using the consensus between both predictors, CMScaller and CMSclassifier.

5.3.5 Risk prediction considering synergistic interactions

In order to measure the capability of risk prediction and the relation with survival in cancer of our marker genes, an analysis was performed considering combinations of two genes. As explained in the previous section, a Cox regression with an interaction term was used.

We generated all possible combinations of elements taken two by two with our markers. For each combination, the resulting p-value of the cox model was compared with the individual value of each gen in the pair. If the model's value performed better than the individual one, we stored the p-value as a metric of how strong was this interaction, and the genes involved.

The final result (as an interactions table) was represented as a network in which the interactions between the top related markers may be easily observed, we filtered the p-values by the usual threshold (p val < 0.05) and fold change > 0.5.

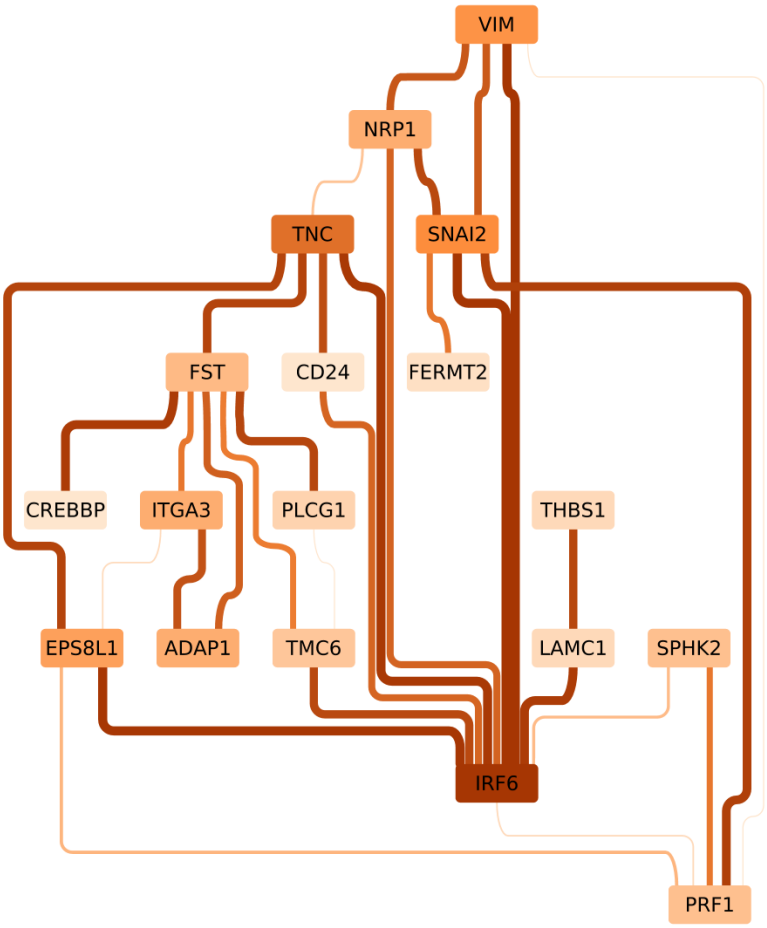


Figure 5.11: Interactions network

Symbol	Interactions
IRF6	10
FST	6
PRF1	5
SNAI2	5
TNC	5
EPS8L1	4
NRP1	4
VIM	4
ITGA3	3
TMC6	3
ADAP1	2
CD24	2
LAMC1	2
PLCG1	2
SPHK2	2
CREBBP	1
FERMT2	1
THBS1	1

Table 5.3: Filtered interactions showing the relations of each node.

The genes shown in **Fig: 5.11** are represented with a darker colour which is proportional to the number of genes related with each one of them. The width and colour of the line that links each pair defines the p-value, the darker and wider is the line, the more significant is the p-value.

For further interpretation of this result, the table Tab: 5.3 show the number of interactions for each one of the nodes in the network. As can be seen, the most interactive genes are IRF6, VIM, SNAI2, TNC...

5.4 Discussion

CMS markers and their relationship with risk

Several authors have suggested in the literature that the HTC116-WT and HTC116-p21ko cell lines are strongly related to the colon cancer subtypes CMS1 and CMS4. However, the relation between the expression patterns of the cellular cell lines and the CMS subtypes is not well understood. This relation is relevant to the development of preclinical models in cancer colon. In this chapter we have studied the relation between the expression patterns of CMS1 vs. HTC116-WT and CMS4 vs. HTC116-p21ko.

The experimental results show that there is a strong relation between the expression patterns of the p21 KO (EMT like) cell lines and CMS4. The proposed markers from these cell lines are strongly differentiated in predicted CMS groups from two validation cohorts of microarray and RNAseq, which confirms the hypothesis. Moreover, the relationship between the markers and EMT characteristics is also proven. In particular, genes that characterize the cellular cell lines are annotated with cell functions that are described as related to CMS1 and CMS4 subtypes.

We have discovered that there are several pairs of genes that synergistically interact and have strong association with patient risk. Considering those interactions helps to improve the risk prediction in cancer colon.

These synergistic interactions are candidates to be tested in the laboratory.

Finally, survival analysis of CMS subtypes confirms that CMS4 has poorer survival than other groups. This can be explained by the relation between the expression pattern of CMS4 and metastatic cell lines.

Chapter 6

Conclusions

Along the four chapters of this PhD dissertation, multiple bioinformatic methods and algorithms have been applied and also developed to address and solve specific problems in the management, analysis and interpretation of complex omic data (mainly genomic and transcriptomic) from human samples derived from cancer patients. The work has been focused in two major types of cancer: Breast Cancer (BRCA) and ColoRectal Cancer (CRC); and the omic data have been integrated and analyzed together with clinical information, mostly data about the survival and outcome of the patients. We think that our research effort has been quite successful thanks to the close collaboration of the bioinformatics and computational biology team, with several research groups, cancer biology experts, as well as with medical groups that provided direct access to the human samples and a deeper understanding of the biomedical questions that we wanted to address in each part of the work.

6.1 General Conclusions

The general conclusions of this Doctoral Thesis, are the following:

1. The proposed methods for dataset normalization and batch effect reduction should be used in order to reduce the bias when merging different sources of data. The importance of these methods is demonstrated in the study performed in *Chapter 4*. The lack of a proper normalization and standardization when merging the data from different sources is something that usually led to failure in reproducibility. The use of these methods, allowed the advance and fulfillment of the *first* and *third objectives* of this thesis. Large and homogeneous datasets of breast cancer (1024 samples) and colorectal cancer (1273 samples) with survival information were successfully generated and analyzed.

2. The methods and algorithms developed have made possible the characterisation of a group of proposed marker genes that identify the "Triple Negative" subtype of breast cancer (TNBC) in a "positive" way and that further relate them with risk and prognosis value. The success in the compilation and comparison of several large BRCA datasets with survival allowed us to do this analysis and to affirm that the *second objective* of this thesis is accomplished.

3. In the study of colorectal cancer (CRC) the relationship found between the p21 KO of HCT116 with the CMS4 subtype and HCT116 WT with the CMS1 subtype, made possible to identify genes that define these subtypes of CRC. The proposed genes were further evaluated and the relationship between these markers and survival was discovered. In the same study, a method capable of identifying genes that strongly define risk when taking into account the binary interaction between genes, was used. The relevant genes obtained will be tested *in vivo* in order to define them as reliable markers. All this study was made possible thanks to the compilation of a large colorectal integrated dataset including transcriptomic data and survival data, as well as to the use of several classifiers and predidtors that apply robust machine learning methods.

4. The proposed bioinformatic methods and strategies to discover "survival marker genes" have been applied successfully to different cancer cohorts, addressing different biological problems. Some of them have been published in JRC journals and others will be submitted soon. In particular, the methods developed and applied to transcriptomic data combined with survival data have demonstrated the capability to discover groups of genes that outperform the already proposed markers for risk prediction and patient stratification, while keeping the relationship with the most relevant clinical features.

5. The importance of robustness, every time a biostatistical analysis of omic data is done, can be one of the main conclusions of this Thesis. This has been taken into account for every algorithm designed and applied. Validation of the results presented in this dissertation was achieved using independent cohorts and datasets, distinct platforms, multiple data integration and computational cross-validation techniques.

6.2 Future work

It is clear from the results and the described above conclusions, that the first urgent work to be done as a product of this Doctoral Thesis will be to achieve the publication in good scientific journals of the three research articles that we have in preparation. These articles are presented in an Appendix below, including the title and the authors that worked in each one of them. As a whole, we think that this PhD provides a good demonstration that cooperative work is essential in current biomedical studies. We will continue in this effort by bringing our bioinformatics and computational expertise close to cancer biologists and medical oncologists.

Another, near future product that we are preparing as a direct result of our bioinformatics work in this Doctoral Thesis is the implementation of the computational methods and functions in an integrated software package written in R for Bioconductor (https://www.bioconductor.org/). This R software will be Open Access and freely available for anyone to use it.

Finally, an specific research point that we would like to achieve in the near future is the biochemical validation of the top new "survival markers" that we have discovered and proposed in this work, in particular for CRC and TNBC.

Bibliography

T. Abbas and A. Dutta. p21 in cancer: intricate networks and multiple activities. *Nat. Rev. Cancer*, 9(6):400–414, 2009.

R. Aguirre-Gamboa, H. Gomez-Rueda, E. Martinez-Ledesma, A. Martinez-Torteya, R. Chacolla-Huaringa, A. Rodriguez-Barrientos, J. G. Tamez-Pena, and V. Trevino. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS ONE*, 8(9):e74250, 2013.

S. Aibar, C. Fontanillo, C. Droste, B. Roson-Burgo, F. J. Campos-Laborie, J. M. Hernandez-Rivas, and J. De Las Rivas. Analyse multiple disease subtypes and build associated gene networks using genome-wide expression profiles. *BMC Genomics*, 16 Suppl 5:S3, 2015.

M. J. Alvarez, Y. Shen, F. M. Giorgi, A. Lachmann, B. B. Ding, B. H. Ye, and A. Califano. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.*, 48(8):838–847, 2016.

A. Aytes, A. Mitrofanova, C. Lefebvre, M. J. Alvarez, M. Castillo-Martin, T. Zheng, J. A. Eastham, A. Gopalan, K. J. Pienta, M. M. Shen, A. Califano, and C. Abate-Shen. Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer Cell*, 25(5):638–651, 2014.

D. H. Barnett, S. Sheng, T. H. Charn, A. Waheed, W. S. Sly, C. Y. Lin, E. T. Liu, and B. S. Katzenellenbogen. Estrogen receptor regulation of carbonic anhydrase XII through a distal enhancer in breast cancer. *Cancer Res.*, 68(9):3505–3515, 2008.

J. M. Bartlett, J. Bayani, A. Marshall, et al. Comparing Breast Cancer Multiparameter Tests in the OPTIMA Prelim Trial: No Test Is More Equal Than the Others. *J. Natl. Cancer Inst.*, 108(9), 2016.

P. R. Benusiglio, P. D. Pharoah, P. L. Smith, F. Lesueur, D. Conroy, R. N. Luben, G. Dew, C. Jordan, A. Dunning, D. F. Easton, and B. A. Ponder. HapMap-based study of the 17q21 ERBB2 amplicon in susceptibility to breast cancer. *Br. J. Cancer*, 95(12):1689–1695, 2006.

M. F. Bijlsma, A. Sadanandam, P. Tan, and L. Vermeulen. Molecular subtypes in cancers of the gastrointestinal tract. *Nat Rev Gastroenterol Hepatol*, 14(6): 333–342, 2017.

W. W. Bivin, O. Yergiyev, M. L. Bunker, J. F. Silverman, and U. Krishnamurti. GRB7 Expression and Correlation With HER2 Amplification in Invasive Breast Carcinoma. *Appl. Immunohistochem. Mol. Morphol.*, 25(8):553–558, 2017.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

J. Brettschneider, F. Collin, B. M. Bolstad, and T. P. Speed. Quality assessment for short oligonucleotide arrays. *Technometrics*, In press, 2007.

F. J. Campos-Laborie, A. Risueno, M. Ortiz-Estevez, B. Roson-Burgo, C. Droste, C. Fontanillo, R. Loos, J. M. Sanchez-Santos, M. W. Trotter, and J. De Las Rivas. DECO: decompose heterogeneous population cohorts for patient stratification and discovery of sample biomarkers using omic data profiling. *Bioinformatics*, 2019.

W. Che, Y. Bao, and F. Tang. Down-regulation of C35 decreased the cell viability and migration of breast ductal carcinoma cells. *PLoS ONE*, 12(8):e0183941, 2017.

H. Chen, X. Sun, W. Ge, Y. Qian, R. Bai, and S. Zheng. A seven-gene signature predicts overall survival of patients with colorectal cancer. *Oncotarget*, 8(56): 95054–95065, 2017.

X. Chen and H. Ishwaran. Random forests for genomic data analysis. *Genomics*, 99 (6):323–329, 2012.

Z. Chen, L. Ai, M. Y. Mboge, C. Tu, R. McKenna, K. D. Brown, C. D. Heldermon, and S. C. Frost. Differential expression and function of CAIX and CAXII in breast cancer: A comparison between tumorgraft models and cells. *PLoS ONE*, 13(7): e0199476, 2018.

X. Cheng, L. Wei, X. Huang, J. Zheng, M. Shao, T. Feng, J. Li, Y. Han, W. Tan, W. Tan, D. Lin, and C. Wu. Solute Carrier Family 39 Member 6 Gene Promotes Aggressiveness of Esophageal Carcinoma Cells by Increasing Intracellular Levels of Zinc, Activating Phosphatidylinositol 3-Kinase Signaling, and Up-regulating Genes That Regulate Metastasis. *Gastroenterology*, 152(8):1985–1997, 2017.

F. Chibon. Cancer gene expression signatures - the rise and fall? *Eur. J. Cancer*, 49(8):2000–2009, 2013.

M. H. Chien, T. H. Ying, Y. H. Hsieh, C. H. Lin, C. H. Shih, L. H. Wei, and S. F. Yang. Tumor-associated carbonic anhydrase XII is linked to the growth of primary oral squamous cell carcinoma and its poor prognosis. *Oral Oncol.*, 48(5): 417–423, 2012.

A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biol.*, 17:13, 2016.

I. Dagogo-Jack and A. T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol*, 15(2):81–94, 2018.

V. Das, J. Kalita, and M. Pal. Predictive and prognostic biomarkers in colorectal cancer: A systematic review of recent advances and challenges. *Biomed. Pharmacother.*, 87:8–19, 2017.

S. Davis and P. Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 14:1846–1847, 2007.

R. Dienstmann, L. Vermeulen, J. Guinney, S. Kopetz, S. Tejpar, and J. Tabernero. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat. Rev. Cancer*, 17(2):79–92, 2017.

X. Dong, Y. Huang, L. Kong, J. Li, J. Kou, L. Yin, and J. Yang. C35 is overexpressed in colorectal cancer and is associated tumor invasion and metastasis. *Biosci Trends*, 9(2):117–121, 2015.

S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.

S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*, 4(8):1184–1191, 2009.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman, New York, 1993.

M. Ehrlich. DNA hypomethylation in cancer cells. *Epigenomics*, 1(2):239–259, 2009.

P. W. Eide, J. Bruun, R. A. Lothe, and A. Sveen. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci Rep*, 7(1):16618, 2017.

L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2005.

F. A. Elian, E. Yan, and M. A. Walter. FOXC1, the new player in the cancer sandbox. *Oncotarget*, 9(8):8165–8178, 2018.

J. Ferlay, M. Colombet, I. Soerjomataram, T. Dyba, G. Randi, M. Bettio, A. Gavin, O. Visser, and F. Bray. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *Eur. J. Cancer*, 103: 356–387, 2018.

E. Fessler and J. P. Medema. Colorectal Cancer Subtypes: Developmental Origin and Microenvironmental Regulation. *Trends Cancer*, 2(9):505–518, 2016.

G. C. Fletcher, S. Patel, K. Tyson, P. J. Adam, M. Schenker, J. A. Loader, L. Daviet, P. Legrain, R. Parekh, A. L. Harris, and J. A. Terrett. hAG-2 and hAG-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene C4.4a and dystroglycan. *Br. J. Cancer*, 88(4):579–585, 2003.

P. Flicek, M. R. Amode, D. Barrell, et al. Ensembl 2014. 42(D1):D749–D755, 2014.

J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. ISSN 1548-7660.

L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004. ISSN 1367-4803.

B. George and S. Kopetz. Predictive and prognostic markers in colorectal cancer. *Curr Oncol Rep*, 13(3):206–215, 2011a.

B. George and S. Kopetz. Predictive and prognostic markers in colorectal cancer. *Curr Oncol Rep*, 13(3):206–215, 2011b.

M. Gerlinger, A. J. Rowan, S. Horswell, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.*, 366 (10):883–892, 2012.

J. Gui and H. Li. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008, 2005.

J. Guinney, R. Dienstmann, X. Wang, et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.*, 21(11):1350–1356, 2015.

S. Hannenhalli and K. H. Kaestner. The evolution of Fox genes and their role in development and disease. *Nat. Rev. Genet.*, 10(4):233–240, 2009.

A.-C. Haury, P. Gestraud, and J.-P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, 6: e28210, 2011.

J. He, Z. Zhou, M. Reed, and A. Califano. Accelerated parallel algorithm for gene network reverse engineering. *BMC Syst Biol*, 11(Suppl 4):83, 2017.

K. A. Hoadley, C. Yau, T. Hinoue, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2): 291–304, 2018.

Y. Hong, X. Q. Chen, J. Y. Li, C. Liu, N. Shen, B. B. Zhu, J. Gong, and W. Chen. Current evidence on the association between rs3757318 of C6orf97 and breast cancer risk: a meta-analysis. *Asian Pac. J. Cancer Prev.*, 15(19):8051–8055, 2014.

Y. Hoshida. Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment. *PLoS ONE*, 5(11):e15543, 2010.

S. Huang, X. Chen, J. Zheng, Y. Huang, L. Song, Y. Yin, and J. Xiong. Low SIRT3 expression contributes to tumor progression, development and poor prognosis in human pancreatic carcinoma. *Pathol. Res. Pract.*, 213(11):1419–1423, 2017.

W. Huber, V. J. Carey, R. Gentleman, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, 12(2):115–121, 2015.

T. J. Hudson, W. Anderson, A. Artez, et al. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

M. B. Jensen, A. V. Lænkholm, T. O. Nielsen, J. O. Eriksen, P. Wehn, T. Hood, N. Ram, W. Buckingham, S. Ferree, and B. Ejlertsen. The Prosigna gene expression assay and responsiveness to adjuvant cyclophosphamide-based chemotherapy in premenopausal high-risk patients with breast cancer. *Breast Cancer Res.*, 20 (1):79, 2018.

B. Y. Jiang, X. C. Zhang, J. Su, W. Meng, X. N. Yang, J. J. Yang, Q. Zhou, Z. Y. Chen, Z. H. Chen, Z. Xie, S. L. Chen, and Y. L. Wu. BCL11A overexpression predicts survival and relapse in non-small cell lung cancer and is modulated by microRNA-30a and gene amplification. *Mol. Cancer*, 12:61, 2013.

P. Jiang, Y. Li, A. Poleshko, V. Medvedeva, N. Baulina, Y. Zhang, Y. Zhou, C. M. Slater, T. Pellegrin, J. Wasserman, M. Lindy, A. Efimov, M. Daly, R. A. Katz, and X. Chen. The Protein Encoded by the CCDC170 Breast Cancer Gene Functions to Organize the Golgi-Microtubule Network. *EBioMedicine*, 22:28–43, 2017.

K. Kankava, E. Kvaratskhelia, G. Burkadze, G. Tsikhiseli, T. Azanishvili, T. Tkemaladze, and E. Abzianidze. Assessment of the value of methylation features in different tissues for preoperative identification of high-risk breast tumors. *Georgian Med News*, 1(289):143–151, 2019.

E. Katz, S. Dubois-Marshall, A. H. Sims, D. Faratian, J. Li, E. S. Smith, J. A. Quinn, M. Edward, R. R. Meehan, E. E. Evans, S. P. Langdon, and D. J. Harrison. A gene on the HER2 amplicon, C35, is an oncogene in breast cancer whose actions are prevented by inhibition of Syk. *Br. J. Cancer*, 103(3):401–410, 2010.

W. T. Khaled, S. Choon Lee, J. Stingl, et al. BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nat Commun*, 6: 5987, 2015.

E. R. King, C. S. Tung, Y. T. Tsang, Z. Zu, G. T. Lok, M. T. Deavers, A. Malpica, J. K. Wolf, K. H. Lu, M. J. Birrer, S. C. Mok, D. M. Gershenson, and K. K. Wong. The anterior gradient homolog 3 (AGR3) gene is associated with differentiation and survival in ovarian cancer. *Am. J. Surg. Pathol.*, 35(6):904–912, 2011.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, Springer, New York, 1997.

M. Kobayashi, T. Matsumoto, S. Ryuge, K. Yanagita, R. Nagashio, Y. Kawakami, N. Goshima, S. X. Jiang, M. Saegusa, A. Iyoda, Y. Satoh, N. Masuda, and Y. Sato. CAXII Is a sero-diagnostic marker for lung cancer. *PLoS ONE*, 7(3):e33952, 2012.

J. M. Kocarnik, S. Shiovitz, and A. I. Phipps. Molecular phenotypes of colorectal cancer and potential clinical applications. *Gastroenterol Rep (Oxf)*, 3(4):269–276, 2015.

M. A. Komor, L. J. Bosch, G. Bounova, et al. Consensus molecular subtype classification of colorectal adenomas. *J. Pathol.*, 246(3):266–276, 2018.

J. Kopecka, I. Campia, A. Jacobs, A. P. Frei, D. Ghigo, B. Wollscheid, and C. Riganti. Carbonic anhydrase XII is a new therapeutic target to overcome chemoresistance in cancer cells. *Oncotarget*, 6(9):6776–6793, 2015.

S. Kopetz, J. Tabernero, R. Rosenberg, Z. Q. Jiang, V. Moreno, T. Bachleitner-Hofmann, G. Lanza, L. Stork-Sloots, D. Maru, I. Simon, G. Capella, and R. Salazar. Genomic classifier ColoPrint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. *Oncologist*, 20(2):127–133, 2015.

M. Krajewska, R. Dries, A. V. Grassetti, S. Dust, Y. Gao, H. Huang, B. Sharma, D. S. Day, N. Kwiatkowski, M. Pomaville, O. Dodd, E. Chipumuro, T. Zhang, A. L. Greenleaf, G. C. Yuan, N. S. Gray, R. A. Young, M. Geyer, S. A. Gerber, and R. E. George. CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. *Nat Commun*, 10(1):1757, 2019.

P. Kupfer, R. Guthke, D. Pohlers, R. Huber, D. Koczan, and R. W. Kinne. Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. *BMC Med Genomics*, 5:23, 2012.

D. T. Lang and the CRAN team. *RCurl: General Network Client Interface for R*, 2019. URL `http://CRAN.R-project.org/package=RCurl`. R package version 1.95-4.12.

J. Lee, I. Sohn, I. G. Do, et al. Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery. *PLoS ONE*, 9(3):e90133, 2014.

S. E. Lee, E. Oh, B. Lee, Y. J. Kim, D. Y. Oh, K. Jung, J. S. Choi, J. Kim, S. J. Kim, J. W. Yang, J. An, Y. L. Oh, and Y. L. Choi. Phenylethanolamine N-methyltransferase downregulation is associated with malignant pheochromocytoma/paraganglioma. *Oncotarget*, 7(17):24141–24153, 2016.

Y. S. Lee, B. H. Kim, B. C. Kim, A. Shin, J. S. Kim, S. H. Hong, J. A. Hwang, J. A. Lee, S. Nam, S. H. Lee, J. Bhak, and J. W. Park. SLC15A2 genomic variation is associated with the extraordinary response of sorafenib treatment: whole-genome analysis in patients with hepatocellular carcinoma. *Oncotarget*, 6 (18):16449–16460, 2015.

Q. Li, A. C. Eklund, N. Juul, B. Haibe-Kains, C. T. Workman, A. L. Richardson, Z. Szallasi, and C. Swanton. Minimising immunohistochemical false negative ER classification using a complementary 23 gene expression signature of ER status. *PLoS ONE*, 5(12):e15031, Dec 2010.

X. Li, J. J. Jung, L. Nie, M. Razavian, J. Zhang, V. Samuel, and M. M. Sadeghi. The neuropilin-like protein ESDN regulates insulin signaling and sensitivity. *Am. J. Physiol. Heart Circ. Physiol.*, 310(9):H1184–1193, 2016.

Y. Li, P. W. Yan, X. E. Huang, and C. G. Li. MDR1 gene C3435T polymorphism is associated with clinical outcomes in gastric cancer patients treated with post-operative adjuvant chemotherapy. *Asian Pac. J. Cancer Prev.*, 12(9):2405–2409, 2011.

R. C. Lim, J. T. Price, and J. A. Wilce. Context-dependent role of Grb7 in HER2+ve and triple-negative breast cancer cell lines. *Breast Cancer Res. Treat.*, 143(3):593–603, 2014.

J. F. Linnekamp, X. Wang, J. P. Medema, and L. Vermeulen. Colorectal cancer heterogeneity and targeted therapy: a case for molecular disease subtypes. *Cancer Res.*, 75(2):245–249, 2015.

P. Liu, J. R. Keller, M. Ortiz, L. Tessarollo, R. A. Rachel, T. Nakamura, N. A. Jenkins, and N. G. Copeland. Bcl11a is essential for normal lymphoid development. *Nat. Immunol.*, 4(6):525–532, 2003.

R. Liu, W. Zhang, Z. Q. Liu, and H. H. Zhou. Associating transcriptional modules with colon cancer survival through weighted gene co-expression network analysis. *BMC Genomics*, 18(1):361, 2017.

V. Lopez and S. L. Kelleher. Zip6-attenuation promotes epithelial-to-mesenchymal transition in ductal breast tumor (T47D) cells. *Exp. Cell Res.*, 316(3):366–375, 2010.

L. Luo, W. Xia, M. Nie, Y. Sun, Y. Jiang, J. Zhao, S. He, and L. Xu. Association of ESR1 and C6orf97 gene polymorphism with osteoporosis in postmenopausal women. *Mol. Biol. Rep.*, 41(5):3235–3243, 2014.

A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.

U. E. Martinez-Outschoorn, M. Peiris-Pages, R. G. Pestell, F. Sotgia, and M. P. Lisanti. Cancer metabolism: a therapeutic perspective. *Nat Rev Clin Oncol*, 14 (1):11–31, 2017.

M. Masin, J. Vazquez, S. Rossi, S. Groeneveld, N. Samson, P. C. Schwalie, B. Deplancke, L. E. Frawley, J. Gouttenoire, D. Moradpour, T. G. Oliver, and E. Meylan. GLUT3 is induced during epithelial-mesenchymal transition and promotes tumor cell proliferation in non-small cell lung cancer. *Cancer Metab*, 2:11, 2014.

C. Matsui, T. Takatani-Nakase, Y. Hatano, S. Kawahara, I. Nakase, and K. Takahashi. Zinc and its transporter ZIP6 are key mediators of breast cancer cell survival under high glucose conditions. *FEBS Lett.*, 591(20):3348–3359, 2017.

C. Mbogning and P. Broet. Bagging survival tree procedure for variable selection and prediction in the presence of nonsusceptible patients. *BMC Bioinformatics*, 17(1):230, 2016.

M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2):242–253, 2010a.

M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2):242–253, 2010b.

M. N. McCall, P. N. Murakami, M. Lukk, W. Huber, and R. A. Irizarry. Assessing affymetrix genechip microarray quality. *BMC Bioinformatics*, 12(1):137, 2011.

McCarthy, D. J., Chen, Yunshun, Smyth, and G. K. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 2012.

P. Mertins, D. R. Mani, K. V. Ruggles, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605):55–62, 2016.

V. Moreno and R. Sanz-Pamplona. Altered pathways and colorectal cancer prognosis. *BMC Med*, 13:76, 2015.

I. Moy, V. Todorović, A. D. Dubash, J. S. Coon, J. B. Parker, M. Buranapramest, C. C. Huang, H. Zhao, K. J. Green, and S. E. Bulun. Estrogen-dependent sushi domain containing 3 regulates cytoskeleton organization and migration in breast cancer cells. *Oncogene*, 34(3):323–333, 2015.

D. M. Muzny, M. N. Bainbridge, and K. Chang. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

K. Naidoo, P. T. Wai, S. L. Maguire, et al. Evaluation of CDK12 Protein Expression as a Potential Novel Biomarker for DNA Damage Response-Targeted Therapies in Breast Cancer. *Mol. Cancer Ther.*, 17(1):306–315, 2018.

T. Nakamura, Y. Yamazaki, Y. Saiki, M. Moriyama, D. A. Largaespada, N. A. Jenkins, and N. G. Copeland. Evi9 encodes a novel zinc finger protein that physically interacts with BCL6, a known human B-cell proto-oncogene product. *Mol. Cell. Biol.*, 20(9):3178–3186, 2000.

NANOSTRING. Prosigna. http://www.nanostring.com/diagnostics/prosigna, 2019.

NCBI. Gene expression omnibus. http://www.cancer.gov/tcga, 2019.

A. Negroni, D. Venturelli, B. Tanno, R. Amendola, S. Ransac, V. Cesi, B. Calabretta, and G. Raschella. Neuroblastoma specific effects of DR-nm23 and its mutant forms on differentiation and apoptosis. *Cell Death Differ.*, 7(9):843–850, 2000.

M. N. Nguyen, T. G. Choi, D. T. Nguyen, et al. CRC-113 gene expression signature for predicting prognosis in patients with colorectal cancer. *Oncotarget*, 6(31): 31674–31692, 2015.

H. Pan, Z. Peng, J. Lin, X. Ren, G. Zhang, and Y. Cui. Forkhead box C1 boosts triple-negative breast cancer metastasis through activating the transcription of chemokine receptor-4. *Cancer Sci.*, 109(12):3794–3804, 2018.

C. Peters, K. Brejc, L. Belmont, A. J. Bodey, Y. Lee, M. Yu, J. Guo, R. Sakowicz, J. Hartman, and C. A. Moores. Insight into the molecular mechanism of the multitasking kinesin-8 motor. *EMBO J.*, 29(20):3437–3447, 2010.

M. Pinterić, I. I. Podgorski, S. Sobočanec, M. Popović Hadžija, M. Paradžik, A. Dekanić, M. Marinović, M. Halasz, R. Belužić, G. Davidović, A. Ambriović Ristov, and T. Balog. De novo expression of transfected sirtuin 3 enhances susceptibility of human MCF-7 breast cancer cells to hyperoxia treatment. *Free Radic. Res.*, 52(6):672–684, 2018.

L. Qu, L. Liang, J. Su, and Z. Yang. Inhibitory effect of upregulated DR-nm23 expression on invasion and metastasis in colorectal cancer. *Eur. J. Cancer Prev.*, 22(6):512–522, 2013.

P. Raman, S. Zimmerman, K. S. Rathi, L. de Torrenté, M. Sarmady, C. Wu, J. Leipzig, D. M. Taylor, A. Tozeren, and J. C. Mar. A comparison of survival analysis methods for cancer gene expression rna-sequencing data. *Cancer Genetics*, 235-236:1 – 12, 2019. ISSN 2210-7762.

M. Ramos. *curatedTCGAData: Curated Data From The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects*, 2019. R package version 1.4.3.

M. Ramos, L. Schiffer, and L. Waldron. *TCGAutils: TCGA utility functions for data management*, 2019. R package version 1.2.2.

B. Ramsey, T. Bai, A. Hanlon Newell, M. Troxell, B. Park, S. Olson, E. Keenan, and S. W. Luoh. GRB7 protein over-expression and clinical outcome in breast cancer. *Breast Cancer Res. Treat.*, 127(3):659–669, 2011.

A. Risueno, C. Fontanillo, M. E. Dinger, and J. De Las Rivas. GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs. *BMC Bioinformatics*, 11:221, 2010.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11(3):R25, 2010.

D. M. Roden and R. F. Tyndale. Genomic medicine, precision medicine, personalized medicine: what's in a name? *Clin. Pharmacol. Ther.*, 94(2):169–172, 2013.

K. S. Saini, H. A. A. Jr, O. Metzger-Filho, S. Loi, C. Sotiriou, E. de Azambuja, and M. Piccart. Beyond trastuzumab: New treatment options for her2-positive breast cancer. *The Breast*, 20, Supplement 3(0):S20 – S27, 2011. ISSN 0960-9776.

Proceedings of the 12th International Conference on Primary Therapy of Early Breast Cancer.

R. Salazar, P. Roepman, G. Capella, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J. Clin. Oncol.*, 29(1): 17–24, 2011.

A. S. Sameer. Colorectal cancer: molecular mutations and polymorphisms. *Front Oncol*, 3:114, 2013.

F. Sanchez-Vega, M. Mina, J. Armenia, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, 173(2):321–337, 2018.

R. Sandberg and O. Larsson. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*, 8:48, 2007.

R. Sanz-Pamplona, A. Berenguer, D. Cordero, S. Riccadonna, X. Sole, M. Crous-Bou, E. Guino, X. Sanjuan, S. Biondo, A. Soriano, G. Jurman, G. Capella, C. Furlanello, and V. Moreno. Clinical value of prognosis gene expression signatures in colorectal cancer: a systematic review. *PLoS ONE*, 7(11):e48877, 2012.

E. Satterwhite, T. Sonoki, T. G. Willis, L. Harder, R. Nowak, E. L. Arriola, H. Liu, H. P. Price, S. Gesk, D. Steinemann, B. Schlegelberger, D. G. Oscier, R. Siebert, P. W. Tucker, and M. J. Dyer. The BCL11 gene family: involvement of BCL11A in lymphoid malignancies. *Blood*, 98(12):3413–3420, 2001.

H. Schwender. *siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches*, 2012. R package version 1.32.0.

A. C. Society. The american cancer society medical and editorial content team. colorectal cancer stages. https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.html, 2017.

Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, 18(11):696–705, 2018.

C. K. Stein, P. Qu, J. Epstein, A. Buros, A. Rosenthal, J. Crowley, G. Morgan, and B. Barlogie. Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC Bioinformatics*, 16:63, 2015.

A. Sveen, T. H. Agesen, A. Nesbakken, G. I. Meling, T. O. Rognum, K. Liestøl, R. I. Skotheim, and R. A. Lothe. ColoGuidePro: a prognostic 7-gene expression signature for stage III colorectal cancer patients. *Clin. Cancer Res.*, 18(21):6001–6010, 2012.

Z. Sztupinszki and B. Győrffy. Colon cancer subtypes: concordance, effect on survival and selection of the most representative preclinical models. *Sci Rep*, 6:37169, 2016.

W. Tai, Z. Chen, and K. Cheng. Expression profile and functional activity of peptide transporters in prostate cancer cells. *Mol. Pharm.*, 10(2):477–487, 2013.

J. G. Tate, S. Bamford, H. C. Jubb, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.*, 47(D1):D941–D947, 2019.

D. V. F. Tauriello and E. Batlle. Targeting the Microenvironment in Advanced Colorectal Cancer. *Trends Cancer*, 2(9):495–504, 2016.

TCGA. The cancer genome atlas. http://www.ncbi.nlm.nih.gov/geo/, 2019.

T. M. Therneau. *A Package for Survival Analysis in S*, 2014. URL `http://CRAN.R-project.org/package=survival`. R package version 2.37-7.

X. Tian, X. Zhu, T. Yan, C. Yu, C. Shen, Y. Hu, J. Hong, H. Chen, and J. Y. Fang. Recurrence-associated gene signature optimizes recurrence-free survival prediction of colorectal cancer. *Mol Oncol*, 11(11):1544–1560, 2017.

R. Tibshirani. *uniCox: Univarate shrinkage prediction in the Cox model*, 2009. URL `http://CRAN.R-project.org/package=uniCox`. R package version 1.0.

J. F. Tien, A. Mazloomian, S. G. Cheng, C. S. Hughes, C. C. T. Chow, L. T. Canapi, A. Oloumi, G. Trigo-Gonzalez, A. Bashashati, J. Xu, V. C. Chang, S. P. Shah, S. Aparicio, and G. B. Morin. CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. *Nucleic Acids Res.*, 45(11): 6698–6716, 2017.

D. Tkocz, N. T. Crawford, N. E. Buckley, F. B. Berry, R. D. Kennedy, J. J. Gorski, D. P. Harkin, and P. B. Mullan. BRCA1 and GATA3 corepress FOXC1 to inhibit the pathogenesis of basal-like breast cancers. *Oncogene*, 31(32):3667–3678, 2012.

A. Trinh, C. Ladrach, H. E. Dawson, S. Ten Hoorn, P. J. K. Kuppen, M. S. Reimers, M. Koopman, C. J. A. Punt, A. Lugli, L. Vermeulen, and I. Zlobec. Tumour budding is associated with the mesenchymal colon cancer subtype and RAS/RAF mutations: a study of 1320 colorectal cancers with Consensus Molecular Subgroup (CMS) data. *Br. J. Cancer*, 119(10):1244–1251, 2018.

N. Tsao, Y. C. Yang, Y. J. Deng, and Z. F. Chang. The direct interaction of NME3 with Tip60 in DNA repair. *Biochem. J.*, 473(9):1237–1245, 2016.

V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

G. Tuteja and K. H. Kaestner. Forkhead transcription factors II. *Cell*, 131(1):192, 2007a.

G. Tuteja and K. H. Kaestner. SnapShot: forkhead transcription factors I. *Cell*, 130(6):1160, 2007b.

T. Vargas, J. Moreno-Rubio, J. Herranz, et al. ColoLipidGene: signature of lipid metabolism-related genes to predict prognosis in stage-II colon cancer patients. *Oncotarget*, 6(9):7348–7363, 2015.

J. Veeraraghavan, Y. Tan, X. X. Cao, J. A. Kim, X. Wang, G. C. Chamness, S. N. Maiti, L. J. Cooper, D. P. Edwards, A. Contreras, S. G. Hilsenbeck, E. C. Chang, R. Schiff, and X. S. Wang. Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. *Nat Commun*, 5: 4577, 2014.

D. Venet, J. E. Dumont, and V. Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.*, 7 (10):e1002240, 2011.

L. Wang, X. Xiao, D. Li, Y. Chi, P. Wei, Y. Wang, S. Ni, C. Tan, X. Zhou, and X. Du. Abnormal expression of GADD45B in human colorectal carcinoma. *J Transl Med*, 10:215, 2012.

M. A. Weniger, K. Pulford, S. Gesk, S. Ehrlich, A. H. Banham, L. Lyne, J. I. Martin-Subero, R. Siebert, M. J. Dyer, P. Moller, and T. F. Barth. Gains of the proto-oncogene BCL11A and nuclear accumulation of BCL11A(XL) protein are frequent in primary mediastinal B-cell lymphoma. *Leukemia*, 20(10):1880–1882, 2006.

G. Xu, M. Zhang, H. Zhu, and J. Xu. A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. *Gene*, 604:33–40, 2017.

C. W. Yoo, B. H. Nam, J. Y. Kim, H. J. Shin, H. Lim, S. Lee, S. K. Lee, M. C. Lim, and Y. J. Song. Carbonic anhydrase XII expression is associated with histologic grade of cervical cancer and superior radiotherapy outcome. *Radiat Oncol*, 5:101, 2010.

F. Y. Yu, Q. Xu, D. D. Wu, A. T. Lau, and Y. M. Xu. The Prognostic and Clinicopathological Roles of Sirtuin-3 in Various Cancers. *PLoS ONE*, 11(8): e0159801, 2016.

H. B. Zhao, X. F. Zhang, H. B. Wang, and M. Z. Zhang. Migration and invasion enhancer 1 (MIEN1) is overexpressed in breast cancer and is a potential new therapeutic molecular target. *Genet. Mol. Res.*, 16(1), 2017.

S. Zhao, S. S. Chen, Y. Gu, E. Z. Jiang, and Z. H. Yu. Expression and Clinical Significance of Sushi Domain- Containing Protein 3 (SUSD3) and Insulin-like Growth Factor-I Receptor (IGF-IR) in Breast Cancer. *Asian Pac. J. Cancer Prev.*, 16(18):8633–8636, 2015.

Appendix: publications in progress

Article 1 in preparation

Discovery of Breast Cancer (BRCA) survival markers associated to standard clinical markers. R package for survival analysis based on transcriptomic profiling

Santiago Bueno-Fortes[1], Manuel Martín-Merino[2#], Javier De Las Rivas[1#]

[1] Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IMBCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Científicas (CSIC) and University of Salamanca (USAL), Campus Miguel de Unamuno s/n, Salamanca, Spain

[2] Department of Computer Science, Universidad Pontificia de Salamanca (UPSA) Salamanca, Spain

[#] Equally contributed senior authors

Correspondence to:

Javier De Las Rivas
E-mail: jrivas@usal.es

Key words: bioinformatics, survival, risk prediction, feature selection, breast cancer, IHC marker genes

Running title: Discovery of Breast Cancer (BRCA) survival markers associated to standard clinical markers

Article 2 in preparation

Unravel positive markers and regulators of Triple Negative Breast Cancer (TNBC) using transcriptomic and regulatory profiling combined with survival analysis

Santiago Bueno-Fortes[1], Lauren Joyce Antrim[2], Francisco Campos Laborie[1,3], Manuel Martín-Merino[4#], Javier De Las Rivas[1#]

[1] Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IMBCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Científicas (CSIC) and University of Salamanca (USAL), Campus Miguel de Unamuno s/n, Salamanca, Spain
[2] LAC+USC Medical Center, Los Angeles, California, USA
[3] Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, UK
[4] Department of Computer Science, Universidad Pontificia de Salamanca (UPSA)

[#] Senior authors

Correspondence to:

Javier De Las Rivas
E-mail: jrivas@usal.es

Key words: bioinformatics; feature selection; breast cancer; triple negative; biomarker; methylation; survival; risk prediction

Running title: Unravel positive markers and regulators of Triple Negative subtype (TNBC)

Article 3 in preparation

CDKN1A (p21) loss triggers the mesenchymal CMS4 subtype of colorectal cancer

Santiago Bueno-Fortes[1*] , Julienne K. Muenzner[2*] , Alberto Berral[1] , Pablo Lindner[2] , Tobias Baeuerle[3] , Javier De Las Rivas[1#] , Regine Schneider-Stock[2#]

[1] Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IMBCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Científicas (CSIC) and University of Salamanca (USAL), Campus Miguel de Unamuno s/n, Salamanca, Spain

[2] Experimental Tumor Pathology, Institute of Pathology, University Hospital of the Friedrich- Alexander-University Erlangen-Nuremberg, Erlangen, Germany

[3] Preclinical Imaging Platform Erlangen (PIPE), Institute of Radiology, University Hospital Erlangen-Nuremberg, Erlangen, Germany

[*] These authors contributed equally to this work as first authors
[#] Equally contributed senior authors

Correspondence to: Regine Schneider-Stock
E-mail: regine.schneider-Stock@uk-erlangen.de

Key words: bioinformatics; cancer; colon; colorectal cancer; gene expression; gene marker; Kaplan-Meier analysis; survival; transcriptomics Kaplan-Meier analysis; Survival; Transcriptomics

Appendix: publication

Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling

Jorge Martinez-Romero[1,2*], Santiago Bueno-Fortes[1*], Manuel Martín-Merino[3#], Ana Ramirez de Molina[2#], Javier De Las Rivas[1#]

[1] Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IMBCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Científicas (CSIC) and University of Salamanca (USAL), Campus Miguel de Unamuno s/n, Salamanca, Spain

[2] Molecular Oncology and Nutritional Genomics of Cancer Group, Precision Nutrition and Cancer Program, IMDEA Food Institute (CEI UAM/CSIC), Madrid, Spain

[3] Department of Computer Science, Universidad Pontificia de Salamanca (UPSA) Salamanca; Spain

* equally contributed first authors
# senior authors

Correspondence to:

Javier De Las Rivas
E-mail: jrivas@usal.es

Key words: cancer, colorectal cancer, colon, survival, Kaplan-Meier analysis, gene marker, bioinformatics, transcriptomics, gene expression

Martinez-Romero *et al. BMC Genomics* 2018, **19**(Suppl 8):857
https://doi.org/10.1186/s12864-018-5193-9

BMC Genomics

RESEARCH Open Access

Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling

Jorge Martinez-Romero[1,2†], Santiago Bueno-Fortes[1†], Manuel Martín-Merino[1,3], Ana Ramirez de Molina[2] and Javier De Las Rivas[1*]

From Selected articles from the IV Colombian Congress on Bioinformatics and Computational Biology & VIII International Conference on Bioinformatics SoIBio 2017
Santiago de Cali, Colombia. 13-15 September 2017

Abstract

Background: Identification of biomarkers associated with the prognosis of different cancer subtypes is critical to achieve better therapeutic assistance. In colorectal cancer (CRC) the discovery of stable and consistent survival markers remains a challenge due to the high heterogeneity of this class of tumors. In this work, we identified a new set of gene markers for CRC associated to prognosis and risk using a large unified cohort of patients with transcriptomic profiles and survival information.

Results: We built an integrated dataset with 1273 human colorectal samples, which provides a homogeneous robust framework to analyse genome-wide expression and survival data. Using this dataset we identified two sets of genes that are candidate prognostic markers for CRC in stages III and IV, showing either up-regulation correlated with poor prognosis or up-regulation correlated with good prognosis. The top 10 up-regulated genes found as survival markers of poor prognosis (i.e. low survival) were: DCBLD2, PTPN14, LAMP5, TM4SF1, NPR3, LEMD1, LCA5, CSGALNACT2, SLC2A3 and GADD45B. The stability and robustness of the gene survival markers was assessed by cross-validation, and the best-ranked genes were also validated with two external independent cohorts: one of microarrays with 482 samples; another of RNA-seq with 269 samples. Up-regulation of the top genes was also proved in a comparison with normal colorectal tissue samples. Finally, the set of top 100 genes that showed overexpression correlated with low survival was used to build a CRC risk predictor applying a multivariate Cox proportional hazards regression analysis. This risk predictor yielded an optimal separation of the individual patients of the cohort according to their survival, with a *p*-value of 8.25e-14 and Hazard Ratio 2.14 (95% CI: 1.75–2.61).

(Continued on next page)

* Correspondence: jrivas@usal.es
†Jorge Martinez-Romero and Santiago Bueno-Fortes contributed equally to this work.
[1]Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Cientificas (CSIC) and University of Salamanca (USAL), Salamanca, Spain
Full list of author information is available at the end of the article

(Continued from previous page)
Conclusions: The results presented in this work provide a solid rationale for the prognostic utility of a new set of genes in CRC, demonstrating their potential to predict colorectal tumor progression and evolution towards poor survival stages. Our study does not provide a fixed gene signature for prognosis and risk prediction, but instead proposes a robust set of genes ranked according to their predictive power that can be selected for additional tests with other CRC clinical cohorts.

Keywords: Cancer, Colorectal cancer, Colon, Survival, Kaplan-Meier analysis, Gene marker, Bioinformatics, Transcriptomics, Gene Expression

Background

Colorectal cancer (CRC) is one of the most frequent tumors that causes great morbidity worldwide. It is the third most common cancer in men, the second most common cancer in women and the third leading cause of global cancer mortality (https://www.wcrf.org/). CRC is a heterogeneous disease since from one patient to another it differs in clinical presentation, molecular characteristics, and prognosis [1]. The heterogeneity of CRC increases the complexity of this tumoral pathology, making subtyping and stratification a difficult task for therapeutic decisions. In this way, personalized medicine for CRC is becoming increasingly needed, especially for targeted therapies where large variations between individual's treatment responses exist [1, 2]. In this context, the need to find robust gene markers associated with specific subtypes of CRC led us to this study. Furthermore, the specific purpose of our work was to find consistent biomolecular targets that, together to facilitate samples stratification, could be related to the prognosis of the disease using survival data.

The genomic and transcriptomic profiling of human cancer samples has been demonstrated over the last decade as an excellent way to obtain a better molecular characterization of many tumor types and subtypes. While gene expression-based CRC classifications has been heavily approached [2], little consensus in CRC standalone gene bio-marking has been achieved. In fact, several studies have identified a broad variety of gene sets as gene expression profiles for classification and categorization of this malignant disorder [3, 4]. Moreover, several transcriptomic-based tests oriented towards prognosis have also been investigated. Some examples of these are: *ColoLipidGene* [5], *ColoGuidePro* [6] or *ColoPrint* [7]; that include gene signatures associated with CRC survival in some specific biological contexts. Despite these efforts, at present there is not a clear compendium of gene markers for CRC survival and it is quite difficult to find consistency in the literature.

In the clinic, patients are classified into four CRC stages based in the anatomo-pathologycal characteristics of their tumors. It is common to use the *TNM Staging System* (where **T** stands for tumor, **N** for lymph node, and **M** for metastasis). The disease "staging" also allows grouping the patients in 4 progressive cancer stages, indicated by roman numerals: **I**, **II**, **III**, and **IV** [8]. In this way, stages I and II correspond to cases which had not shown cancer cells beyond the tumor or blood. By contrast, stages III and IV correspond to individuals in where the cancer had diseminate to the lymph system or other organs in the body. This four stage categorization represents significantly distinctive patients groups for final outcome or disease relapse, but the stages do not predict the risk of each individual patient because they are not directly associated to survival [9].

Based on the described need and potential benefits to find survival marker genes correlated with high risk and poor prognosis in CRC; we investigated global gene expression profiles of colorectal tumors and its alteration throughout stages, to identify genes that could be levered as biomarkers of survival and prognosis for CRC in late stages (i.e., III and IV). To undertake this work we performed a deep analysis on a large cohort of human samples derived from a robust integration of several datasets that had transcriptomic and clinical survival data. The integration provided a homogeneous and well-standardized meta-dataset that includes 1273 human colorectal samples. The identification of candidate markers was performed using an initial contrast between the gene expression of the subset of patients with CRC allocated by their clinical features to stages I and II versus the patients with tumors corresponding to stages III and IV. Finally, after internal and external cross-validation, the genes selected as best survival markers were used to construct a risk predictor to allow stratification of the patients with respect to their relative risk.

Results

A large dataset of CRC samples including global expression and survival data

We first built a large cohort of CRC samples collected from individuals that had clinical record with survival data times, as well as genome-wide expression profiles of their colorectal primary tumors at diagnosis (i.e. before any drug treatment). Our aim was to achieve a meta-dataset with at least 1 thousand samples and to

demonstrate a good integration of the global transcriptomic profiles of different samples sets avoiding the typical batch-effects that can alterate any unified analysis.

Table 1 presents the datasets of CRC samples that were collected to produce the integrated dataset analysed in this work. All the CRC samples included in this meta-dataset were tested for global gene expression profiling using the platform of high-density microarrays from *Affymetrix*: Human Genome U133 Plus 2.0. Using this platform, the probesets of the arrays were mapped to single genes (as indicated in Risueño et al.) [10] and, in this way, each microarray measured the expression signal of 20,079 human genes (using the mapping provided by the Chip Description File, CDF v.21 from: http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/).

As a whole, Table 1 includes 7 series that were obtained from the Gene Expression Omnibus repository (GEO, https://www.ncbi.nlm.nih.gov/geo/). These datasets included a total amount of 1352 CRC samples, but after collecting the clinical survival data and carrying out the integration and normalization protocols we finished with 1273 samples, since we filtered 79 samples that did not have survival data or did not show comparable data distributions after normalization. The phenotypic and clinical information about the final collection of 1273 samples, i.e., the available data about age, gender, survival time, location of the tumor, degree and TNM staging, presence of mutation in some cancer genes (TP53, KRAS, BRAF), etc.; is included in Additional file 1: Table S1. When information was not available for a given sample the table includes *not assigned* values (NA).

Evaluation of normalization procedures to integrate independent batches

We performed the integration and combined normalization of the CRC expression datasets using 5 different procedures. The procedures applied different normalization algorithms to provide a homogeneous signal matrix, avoiding bias due to batch effect on the global expression profile of the CRC samples. The procedures applied were: **(i)** Robust Multi-array Average (RMA) algorithm [11]; **(ii)** RMA plus Combatting Batch effects (ComBat) algorithm [12]; **(iii)** Frozen Robust Multi-array Average (fRMA) algorithm [13]; **(iv)** fRMA plus Combat; **(v)** fRMA plus scaling of the data using mean-centered expression values.

Table 1 Summary information about the series of colorectal cancer (CRC) samples that were collected to produce the integrated data set analyzed in this work

GEO dataset	Sample Source	Sample Description	Total samples in dataset	PubMed PMID	Authors and Year	Samples discarded	Samples processed
GSE14333	Royal Melbourne Hospital, Western Hospital and Peter MacCallum Cancer Center, AUSTRALIA. H Lee Moffitt Cancer Center, USA	primary colorectal cancers	290	19996206	Jorissen RN et al. (2009)	64	226
GSE17536	Moffitt Cancer Center, USA	colorectal cancer patients	177	19914252	Smith JJ et al. (2010)	0	177
GSE31595	Roskilde Hospital, DENMARK	patients with stage II and III colorectal cancer	37	–	Thorsteinsson M et al. (2011)	0	37
GSE33113	Academic Medical Center in Amsterdam, NETHERLANDS	primary tumor resections from stage II colorectal patients	90	22496204	Kemper K et al. (2012)	0	90
GSE38832	Vandervilt University Medical Center, USA	tumor samples collected from colorectal patients	122	25320007	Tripathi MK et al. (2014)	0	122
GSE39084	Toulouse Hospital, FRANCE	sporadic early onset primary colorectal carcinomas	70	25083765	Kirzin S et al. (2014)	1	69
GSE39582	Institut G. Roussy (Villejuif), Hosp. Saint Antoine (Paris), Hosp. G.Pompidou (Paris), Hosp. Hautepierre (Strasbourg), Hosp. Purpan (Toulouse), Institut P. Calmettes (Marseille), Centre Antoine Lacassagne (Nice), FRANCE	colorectal cancer samples	566	23700391	Marisa L et al. (2013)	14	552
Total number			1352				1273

All the CRC samples were tested for global gene expression profiling using high-density microarrays Human Genome U133 Plus 2.0 from *Affymetrix* (that measure the signal of 20,141 human genes). The total collection included 1352 samples, but only 1273 were finally used. A group of 79 samples were discarded because they did not have survival data or they presented anomalous data distributions with respect to the other samples of the same series

To evaluate and compare the results provided by each one of these 5 procedures we carried out several analyses. Figure 1 presents the heatmaps derived from an unsupervised clustering of the samples using in each case the expression data matrix derived from each one of the 5 procedures applied. Due to the fact that each series has a different number of samples (one with more than 500 and several other with less than 100), we did a random selection of an even number of samples for each dataset to be included in the cluster analysis: 30 samples from each one. In this way, each heatmap is composed of 210 samples (30 × 7): 30 samples from each one of the 7 datasets (identified by the ID number, GSE, from GEO). In Fig. 1 the samples of each batch are identified by a color that is indicated in the horizontal bar below the dendrograms. Each heatmap represents a different preprocessing and normalization method performed to merge the datasets in one meta-dataset. The results shown in these clustering analyses indicate that in the case of methods that gave the heatmaps A, C and E, several samples of the same color are grouped together showing that they have a common correlation profile within the global expression signature. By contrast, in the case of methods that gave the heatmaps B and D, there is a clearer shuffling of all the colors, which reflects a homogenous mix of the overall expression signal coming from different datasets.

The clustering analysis presented in the symmetric heatmaps of Fig. 1 was done using, for each sample, a vector including the expression signals along all genes and calculating with these vectors the pair-wise *Pearson* correlations between samples and the pair-wise distance matrix derived from such correlations. This approach can reveal major effects associated to the global expression signal of the samples, but it is not very sensitive to detect minor changes in a small number of genes. For this reason we applied a second approach to compare the results provided by the 5 normalization procedures in order to select the one that produces the best unification of the 7 CRC datasets, preserving a good signal to noise ratio in the expression distributions. Algorithms of dimensionality reduction, such as PCA (Principal Component Analysis), allow exploring large datasets in an accurate way to identify factors that are relevant for the variance of studied variables (in our case the expression of the genes in the unified meta-dataset of 1273 samples). Figure 2 presents the plots derived from the PCA done over the 5 expression matrices (i.e. the signal of 20,079 genes in 1273 samples) obtained with 5 different normalization approaches. These results show very clearly that the RMA method (Fig. 2a) is not good to

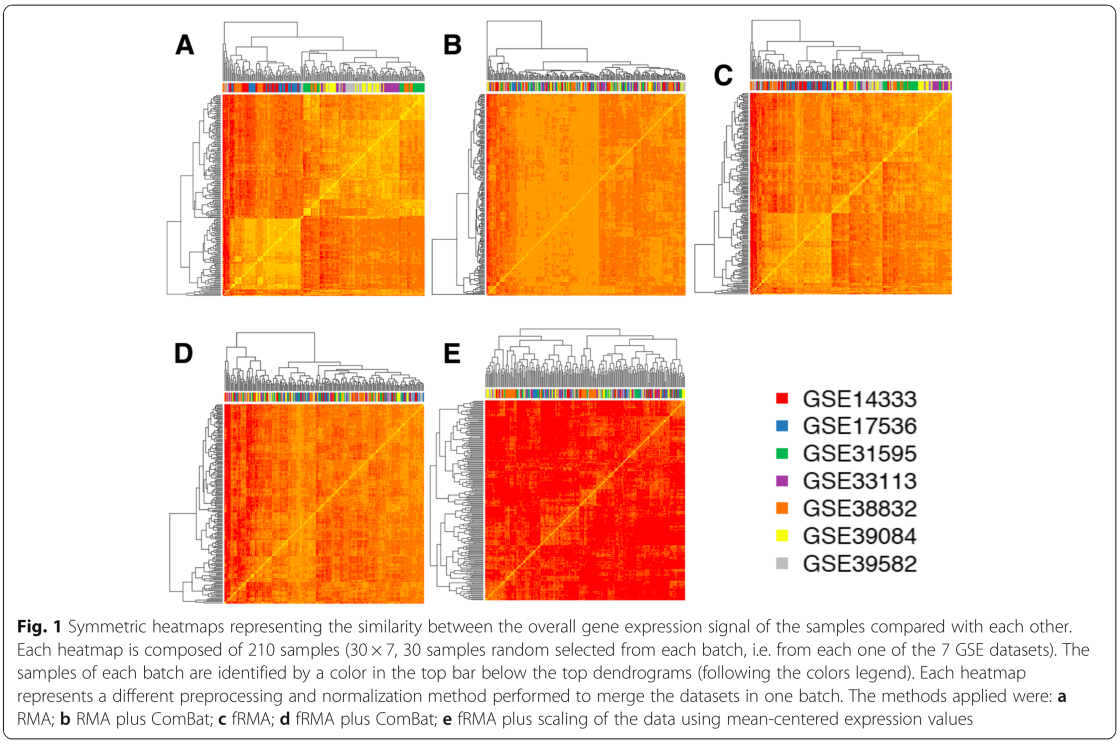


Fig. 1 Symmetric heatmaps representing the similarity between the overall gene expression signal of the samples compared with each other. Each heatmap is composed of 210 samples (30 × 7, 30 samples random selected from each batch, i.e. from each one of the 7 GSE datasets). The samples of each batch are identified by a color in the top bar below the top dendrograms (following the colors legend). Each heatmap represents a different preprocessing and normalization method performed to merge the datasets in one batch. The methods applied were: **a** RMA; **b** RMA plus ComBat; **c** fRMA; **d** fRMA plus ComBat; **e** fRMA plus scaling of the data using mean-centered expression values

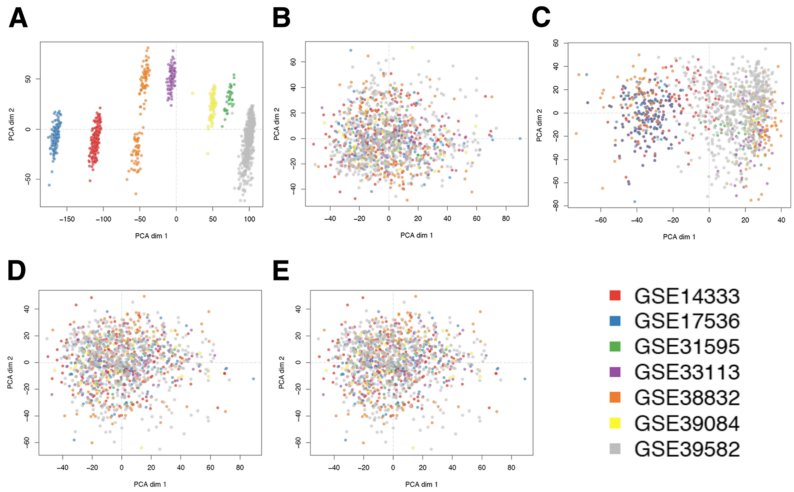


Fig. 2 Plots presenting the distribution of the 1273 samples from 7 datasets (GSEs) obtained by Principal Component Analysis (PCA) of the global gene expression profile of each sample; that converts the signal of each sample using an orthogonal transformation in linearly uncorrelated variables called principal components or dimensions. Each plot presents the values of the two main dimensions (dim 1 versus dim 2) and corresponds to the PCA results obtained using the expression data calculated with different preprocessing and normalization methods. The methods applied were: **a** RMA; **b** RMA plus ComBat; **c** fRMA; **d** fRMA plus ComBat; **e** fRMA plus scaling of the data using mean-centered expression values. The samples of each batch are identified by color dots following the colors legend

provide a proper normalization of different batches, since the samples keep a very strong signal associated to each batch. The fRMA method (Fig. 2c) neither is good, since some samples (specially the ones from the largest batch GSE39582) still keep a strong signal associated to their batch. By contrast, the analysis of the data provided by the other 3 procedures (RMA plus Combat, fRMA plus Combat and fRMA plus mean-centered scaling, Fig. 2b, d and e, respectively) showed an adequate mix of all the samples from different batches. Within these 3 procedures, the normalization is very similar keeping a good signal to noise ratio along the genes and a small signal reduction. We finally select option B, RMA plus Combat, because the heatmap in Fig. 1b showed the best mix between series and a better similarity between the samples (compared to options D or E).

As a final testing to identify the best integration and normalization procedure of the 7 CRC expression datasets, we carried out a linear regression analyses on the global expression matrix considering as predictors 7 independent dummy variables or factors. These variables correspond to the series from which each sample comes from. In this way, if these factors have a significant influence in the expression signal distributions, the linear regression analysis will show a significant p-value and correlation. The results of this analysis are presented in Table 2, that reveals again that only the data matrices produced by the methods B and D (RMA plus Combat and fRMA plus Combat, respectively) do not show a significant effect attributed to belonging to one of the series. Finally, we choose B versus D as the final procedure applied because, despite being very similar, the application of RMA plus Combat provoked less dramatic changes with respect to the raw signal expression.

Identification of genes associated to advanced CRC that mark survival differences

Once we produced a large and well-integrated meta-dataset of CRC samples, having global expression profiles and clinical survival data for all cases, we proceed to the identification of the subset of genes that suffer significant changes with colorectal tumor progression. To do this, we explored the overall expression matrix to detect the genes that showed a significant expression change when comparing CRC tumors in early stages (stages I and II) versus CRC tumors in late or advanced stages (stages III and IV). This comparison was done applying LIMMA, differential expression algorithm, and retrieving all genes that gave a significant p-value (adjusted $p < 0.05$) in either direction (i.e., genes up-regulated with the progression of the disease, in late versus early CRC stages; or genes down-regulated with the progression of the disease). Such differential expression analysis gave a subset of 2707 human genes: 2524 corresponding to protein-coding genes and the rest to non-coding genes (in this work we focused only in the protein-coding genes).

Table 2 Results of the linear regression analyses on the global expression matrix calculated for the 1273 samples from 7 datasets (GSEs) combined using 5 different preprocessing and normalization methods

FACTORS considered	Estimated coefficients	std. error	t value	*p*.value	Factor effect
(A) RMA					
Intercept	6.925	0.014	512.610	<2e-16	–
(GSE14333+) GSE17536	0.387	0.019	20.230	<2e-16	yes
GSE31595	−1.212	0.019	−63.440	<2e-16	yes
GSE33113	−0.577	0.019	−30.210	<2e-16	yes
GSE38832	−0.355	0.019	−18.570	<2e-16	yes
GSE39084	−0.978	0.019	−51.180	<2e-16	yes
GSE39582	−1.375	0.019	−71.970	<2e-16	yes
(B) RMA plus Combat					
Intercept	6.219	0.013	473.582	<2e-16	–
(GSE14333+) GSE17536	0.000	0.019	0.001	0.999	no
GSE31595	0.002	0.019	0.122	0.903	no
GSE33113	0.001	0.019	0.051	0.959	no
GSE38832	−0.001	0.019	−0.033	0.973	no
GSE39084	0.002	0.019	0.092	0.927	no
GSE39582	0.001	0.019	0.029	0.977	no
(C) fRMA					
Intercept	6.535	0.015	450.434	<2e-16	–
(GSE14333+) GSE17536	−0.011	0.021	−0.553	0.580	no so much
GSE31595	0.089	0.021	4.329	0.000	yes
GSE33113	0.071	0.021	3.455	0.001	yes
GSE38832	0.054	0.021	2.641	0.008	yes
GSE39084	0.096	0.021	4.695	0.000	yes
GSE39582	0.089	0.021	4.336	0.000	yes
(D) fRMA plus Combat					
Intercept	6.590	0.014	457.338	<2e-16	–
(GSE14333+) GSE17536	0.000	0.020	0.001	1.000	no
GSE31595	0.002	0.020	0.093	0.926	no
GSE33113	0.001	0.020	0.072	0.942	no
GSE38832	0.000	0.020	0.019	0.985	no
GSE39084	0.002	0.020	0.089	0.929	no
GSE39582	0.000	0.020	0.007	0.994	no
(E) fRMA plus mean centered					
Intercept	0.000	0.000	−1.638	0.101	–
(GSE14333+) GSE17536	0.000	0.000	1.264	0.206	yes
GSE31595	0.000	0.000	0.288	0.773	no so much
GSE33113	0.000	0.000	1.605	0.108	yes
GSE38832	0.000	0.000	1.449	0.147	yes
GSE39084	0.000	0.000	−0.076	0.940	no
GSE39582	0.000	0.000	1.395	0.163	yes

The methods applied were: **(A)** RMA; **(B)** RMA plus ComBat; **(C)** fRMA; **(D)** fRMA plus ComBat; **(E)** fRMA plus scaling of the data using mean-centered expression values. The linear regression is done to evaluate the "batch effect" (i.e. considering that the tested factors are the fact of "belonging" to a given dataset). Thus, when the *p*-value of the factors are significant (< 0.05), the "batch effect" remains on the overall expression signal. A marginal low significance was considered when *p*-values were < 0.20 in the case E

Once we had the subset of genes that can be associated to advanced or progression of CRC, we perform a second analysis on these gene candidates to find out which ones can be correlated with the survival of the corresponding patient samples based on their expression signals. To do this, we carried out Kaplan-Meier (KM) analysis of the survival times of the set of 1273 colorectal cancer samples for each one of the 2524 genes found in the previous exploration. In this analysis, the genes were ranked considering the non-parametric log-rank test that evaluates the separation between the two KM curves for two prognostic groups: one with good survival and another with poor survival. To do this, our algorithm performs for each gene multiple splits of the sample cohort in two groups, and looks for the splitting that provides the best separation between groups (i.e. the best p-value). Then, a stringent cut-off value (adjusted $p < 0.0003$) was used to select the genes that are considered significant. This allowed the identification of 429 significant genes in which the overexpression correlated with low survival, plus 336 significant genes where the repression correlated with low survival. These analyses were done in a univariate mode, considering each gene as an independent factor.

Figure 3 shows the Kaplan-Meier plots corresponding to the survival profiles of the two populations of individuals that were segregated according to the expression values of the gene tested. The 4 plots correspond to the top genes: DCBLD2 and PTPN14 with overexpression correlated to low survival; and EPHB2 and DUS1L with repression correlated to low survival. The separation of the two populations in both cases is very significant, with KM p-values < 1.0e-10 and Hazard Ratios (HR) around 2.0 for overexpression cases and around 0.45 for repression cases. These parameters were calculated using all the 1273 samples; however it was necessary to do an internal cross-validation of these results to assess how stable and reliable was the signal for each one of the selected genes.

We carried out a cross-validation of the top-200 genes selected in any of the two conditions (i.e. selected as survival markers when they were up-regulated for the cases of poor survival or when they were up-regulated for the cases of better survival). This internal cross-validation was done using for each gene a resampling strategy that randomly selected 80% of the sample 100 times (i.e. doing 100 iterations). The results corresponding to the top 100 genes are included in Additional file 2**:** Table S2, for the case of up-regulation for poor survival, and the other top 100 genes in Additional file 3: Table S3, for the case up-regulation for better survival.

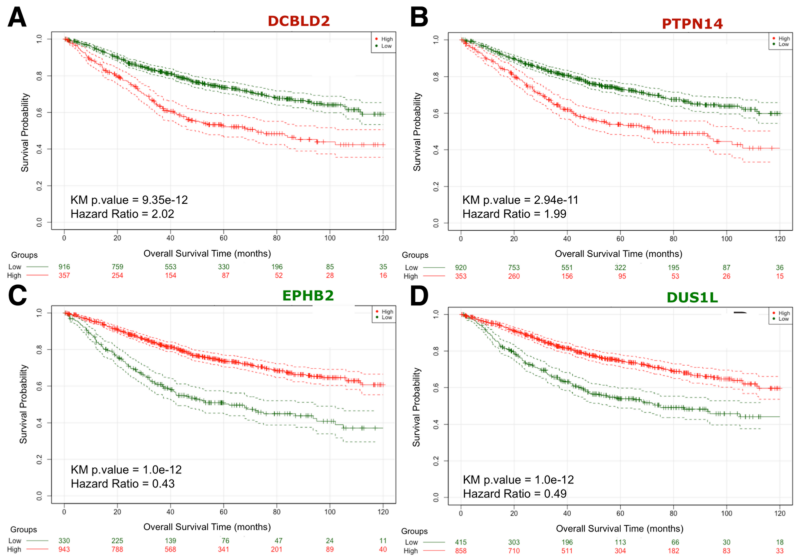


Fig. 3 Kaplan-Meier plots of the survival analysis of the set of 1273 samples from colorectal cancer (CRC) patients. The patients are separated in two groups (high in red and low in green) according to the expression profiles of 4 genes: **a** DCBLD2, **b** PTPN14, **c** EPHB2, **d** DUS1L. These genes provided the best split between patients of high and low risk based in their expression levels. In the case of genes DCBLD2 and PTPN14 (labelled in red) the over-expression is correlated with poor survival; and in the case of genes EPHB2 and DUS1L (labelled in green) the over-expression is correlated with good survival. In all cases the adjusted p.values of the analyses are very significant (as indicated inside each plot), indicating that the two populations represented by the two curves have a very clear difference in their overall survival

A short view of these data is shown in Table 3 that presents the 50 genes selected as best survival markers of CRC: the first part of the table corresponds to the top 25 genes, where up-regulation corresponds to shorter survival and higher risk (HR > 1); the second part of the table corresponds to the top 25 genes, where up-regulation corresponds to longer survival and lower risk (HR < 1). The genes were ranked by their KM *p*-values and the HR values calculated for the whole dataset (i.e. for all the 1273 samples, all-dt). As indicated, the stability and robustness of the gene survival markers was assessed via a resampling strategy with random selection of 80% of the dataset 100 times. For the final ranking of the genes included in these tables we also considered that they had to give a significant adjusted p-value in more than 80 out of 100 bootstrap iterations (i.e. N-sinf-in-100i > 80).

External validation of prognostic markers with a CRC cohort studied using RNA-seq

The analyses done so far provided a ranked collection of genes found as robust markers of survival in CRC. The consistency of the results obtained with the internal cross-validation gives strong support to the top genes found (presented in Table 3), but we had to consider the value of using other external independent CRC cohorts to corroborate these findings. As far as we could investigate we did not find other large CRC datasets (i.e., sets with more than one thousand samples) that included global gene expression data plus survival as part of the clinical characterization of samples. Despite this limitation, we look for independent datasets and found in The Cancer Genome Atlas (TCGA, http://tcga-data.nci.nih.gov/docs/publications/coadread_2012/) a well-characterized cohort of 276 colorectal carcinomas that had been studied with several genome-scale technologies (including RNA-seq gene expression profiling) and that had survival data for 269 samples [14]. We used these data to validate the top genes found as best survival markers in our previous analysis. The results indicated a good performance in more than two thirds of the genes tested. In Additional file 4: Table S4 we present the KM *p*-values and HR of the genes that were validated from the top 10 previously found: 7 genes of the top 10 for the case of up-regulation associated with poor survival (PTPN14, LAMP5, TM4SF1, LCA5, CSGALNACT2, SLC2A3 and GADD45B) and 6 genes of the top 10 previously found for the case of up-regulation associated with good survival (EPHB2, DUS1L, NUAK2, FANCC, MYB and CHDH).

External validation of prognostic markers using multivariate survival analysis

Up to now the search to find gene survival markers associated to the prognosis of CRC have been done using univariate analysis that look for the value and influence of each singular gene. The results presented provided multiple parameters to allow a proper statistical assessment and ranking of each gene survival markers proposed (Table 3). To provide extra support to these results we did another external validation using a second independent cohort of CRC samples from the platform SurvExpress [15]. The CRC dataset selected was called "Colon-Metabase-Uniformized" and it included 482 samples with overall survival data and genome-wide expression determined with *Affymetrix* microarrays. We performed several multivariate survival analyses (OS, overall survival) on this dataset using combinations of the top genes proposed in Table 3. As an example of these analyses we present the KM plot (Additional file 5: Figure S1) corresponding to the multivariate survival study done using the top 5 genes found up-regulated for poor survival (DCBLD2, PTPN14, LAMP5, TM4SF1 and NPR3). It can be seen that the combination of these genes provides a very good separation of two CRC populations: one group of high-risk, associated to the overexpression (or up-regulation) of the genes; and another group of low-risk, associated to the lower expression (or down-regulation) of these genes (Additional file 5: Figure S1). This analysis was repeated with several other combinations of the top up-regulated genes associated with poor survival (present in Table 3), resulting in similar results. For example, combining DCBLD2, LAMP5, TM4SF1, NPR3 and GADD45B the separation of the high and low-risk groups improved a bit: KM *p*-value = 2.21e-07 and HR = 2.23 (95% confidence interval, CI: 1.65–3.02). Another combination that provided very good separation was using genes DCBLD2, LAMP5, TM4SF1, NPR3 and AKAP12: KM p-value = 2.51e-10 and HR = 2.74 (95% CI: 2.00–3.74).

Gene expression profiles of CRC tumor samples versus normal colorectal samples

All the integrated datasets, so far presented in this study corresponded to CRC samples, because we want to provide genes that are disease markers present in the transformed tumor cells of the intestinal epithelium, and genes that mark the progression and aggravation of this type of cancer. In addition, we can only have survival information about patients since in healthy individuals survival time cannot be related to disease and there are not disease-associated events. Despite this obvious consideration, it is interesting to explore what would be the level of expression of the genes, that we identified as survival markers, when they are analysed in normal colorectal tissue. Exploring back on the experimental series used to create our meta-dataset of 1273 CRC samples, we found in series GSE33113 and GSE39582 a collection of 25 samples that corresponded to normal colorectal tissue. We took these samples and included then with our CRC dataset

Table 3 Genes selected as top-50 best *survival markers* of colorectal cancer (CRC)

Number	GENE ENSEMBL_ID	GENE Symbol	KM.*p*.value (all-dt)	HR (all-dt)	N-signf-in-100i (KM.p.value)	HR (mean-in-100i)	GENE HGNC_ID	GENE DESCRIPTION
1	ENSG00000057019	DCBLD2	0.0000000000	2.02	99	2.106	24627	discoidin; CUB and LCCL domain containing 2 [HGNC:24627]
2	ENSG00000152104	PTPN14	0.0000000000	1.99	99	2.082	9647	protein tyrosine phosphatase; non-receptor type 14
3	ENSG00000125869	LAMP5	0.0000000000	1.99	93	2.046	16097	lysosomal associated membrane prot.member 5 [HGNC:16097]
4	ENSG00000169908	TM4SF1	0.0000000001	1.96	93	2.031	11853	transmembrane 4 L six family member 1 [HGNC:11853]
5	ENSG00000113389	NPR3	0.0000000002	1.95	97	2.136	7945	natriuretic peptide receptor 3 [HGNC:7945]
6	ENSG00000186007	LEMD1	0.0000000003	1.95	85	1.937	18,725	LEM domain containing 1 [HGNC:18725]
7	ENSG00000135338	LCA5	0.0000000003	1.89	97	2.021	31,923	LCA5; lebercilin [HGNC:31923]
8	ENSG00000169826	CSGALNACT2	0.0000000008	1.91	92	1.974	24,292	chondroitin sulfate N-acetylgalactosaminyltransferase 2
9	ENSG00000059804	SLC2A3	0.0000000014	1.93	89	1.993	11,007	solute carrier family 2 member 3 [HGNC:11007]
10	ENSG00000099860	GADD45B	0.0000000018	1.92	97	2.074	4096	growth arrest and DNA damage inducible beta [HGNC:4096]
11	ENSG00000136155	SCEL	0.0000000018	1.88	87	1.928	10,573	sciellin [HGNC:10573]
12	ENSG00000100625	SIX4	0.0000000019	1.89	91	1.951	10,890	SIX homeobox 4 [HGNC:10890]
13	ENSG00000131016	AKAP12	0.0000000028	1.85	95	2.092	370	A-kinase anchoring protein 12 [HGNC:370]
14	ENSG00000158270	COLEC12	0.0000000028	1.84	92	1.941	16,016	collectin subfamily member 12 [HGNC:16016]
15	ENSG00000154553	PDLIM3	0.0000000047	1.84	91	1.985	20,767	PDZ and LIM domain 3 [HGNC:20767]
16	ENSG00000082781	ITGB5	0.0000000049	1.82	88	1.911	6160	integrin subunit beta 5 [HGNC:6160]
17	ENSG00000144366	GULP1	0.0000000050	1.81	88	1.911	18,649	engulfment adaptor PTB domain containing 1 [HGNC:18649]
18	ENSG00000171951	SCG2	0.0000000051	1.81	93	2.034	10,575	secretogranin II [HGNC:10575]
19	ENSG00000185567	AHNAK2	0.0000000066	1.80	87	1.896	20,125	AHNAK nucleoprotein 2 [HGNC:20125]
20	ENSG00000138061	CYP1B1	0.0000000075	1.84	85	1.884	2597	cytochrome P450 family 1 subfamily B member 1 [HGNC:2597]
21	ENSG00000184304	PRKD1	0.0000000451	1.74	87	1.872	9407	protein kinase D1 [HGNC:9407]
22	ENSG00000152583	SPARCL1	0.0000000471	1.74	85	1.863	11,220	SPARC like 1 [HGNC:11220]
23	ENSG00000147883	CDKN2B	0.0000000717	1.73	84	1.847	1788	cyclin dependent kinase inhibitor 2B [HGNC:1788]
24	ENSG00000213190	MLLT11	0.0000001989	1.70	84	1.813	16,997	myeloid/lymphoid or mixed-lineage leukemia; translocated to 11
25	ENSG00000135218	CD36	0.0000002751	1.69	85	1.891	1663	CD36 molecule [HGNC:1663]
1	ENSG00000133216	EPHB2	0.0000000000	0.43	100	0.426	3393	EPH receptor B2 [HGNC:3393]
2	ENSG00000169718	DUS1L	0.0000000000	0.49	98	0.481	30,086	dihydrouridine synthase 1 like [HGNC:30086]
3	ENSG00000163545	NUAK2	0.0000000001	0.51	96	0.495	29,558	NUAK family kinase 2 [HGNC:29558]
4	ENSG00000158169	FANCC	0.0000000002	0.51	95	0.498	3584	Fanconi anemia complementation group C [HGNC:3584]

Table 3 Genes selected as top-50 best *survival markers* of colorectal cancer (CRC) *(Continued)*

Number	GENE ENSEMBL_ID	GENE Symbol	KM.*p*.value (all-dt)	HR (all-dt)	N-signf-in-100i (KM.p.value)	HR (mean-in-100i)	GENE HGNC_ID	GENE DESCRIPTION
5	ENSG00000277972	CISD3	0.0000000002	0.51	87	0.511	27,578	CDGSH iron sulfur domain 3 [HGNC:27578]
6	ENSG00000099800	TIMM13	0.0000000003	0.53	95	0.511	11,816	translocase of inner mitochondrial membrane 13 [HGNC:11816]
7	ENSG00000116771	AGMAT	0.0000000005	0.52	95	0.515	18,407	agmatinase [HGNC:18407]
8	ENSG00000118513	MYB	0.0000000006	0.52	93	0.508	7545	MYB proto-oncogene. Transcription factor [HGNC:7545]
9	ENSG00000016391	CHDH	0.0000000006	0.53	90	0.520	24,288	choline dehydrogenase [HGNC:24288]
10	ENSG00000137460	FHDC1	0.0000000008	0.52	96	0.505	29,363	FH2 domain containing 1 [HGNC:29363]
11	ENSG00000132846	ZBED3	0.0000000009	0.52	88	0.522	20,711	zinc finger BED-type containing 3 [HGNC:20711]
12	ENSG00000162408	NOL9	0.0000000015	0.54	92	0.527	26,265	nucleolar protein 9 [HGNC:26265]
13	ENSG00000109534	GAR1	0.0000000017	0.50	99	0.479	14,264	GAR1 ribonucleoprotein [HGNC:14264]
14	ENSG00000133477	FAM83F	0.0000000019	0.54	93	0.518	25,148	family with sequence similarity 83 member F [HGNC:25148]
15	ENSG00000100348	TXN2	0.0000000036	0.53	88	0.527	17,772	thioredoxin 2 [HGNC:17772]
16	ENSG00000108479	GALK1	0.0000000036	0.55	88	0.525	4118	galactokinase 1 [HGNC:4118]
17	ENSG00000110917	MLEC	0.0000000045	0.55	96	0.476	28,973	malectin [HGNC:28973]
18	ENSG00000114738	MAPKAPK3	0.0000000048	0.55	92	0.520	6888	mitogen-activated protein kinase-activated 3 [HGNC:6888]
19	ENSG00000137752	CASP1	0.0000000180	0.56	87	0.523	1499	caspase 1 [HGNC:1499]
20	ENSG00000131844	MCCC2	0.0000000183	0.57	93	0.516	6937	methylcrotonoyl-CoA carboxylase 2 [HGNC:6937]
21	ENSG00000178409	BEND3	0.0000000193	0.55	88	0.529	23,040	BEN domain containing 3 [HGNC:23040]
22	ENSG00000114737	CISH	0.0000000216	0.55	87	0.508	1984	cytokine inducible SH2 containing protein [HGNC:1984]
23	ENSG00000011376	LARS2	0.0000000239	0.55	91	0.528	17,095	leucyl-tRNA synthetase 2; mitochondrial [HGNC:17095]
24	ENSG00000164045	CDC25A	0.0000000481	0.57	90	0.539	1725	cell division cycle 25A [HGNC:1725]
25	ENSG00000154655	L3MBTL4	0.0000000606	0.54	90	0.506	26,677	l(3)mbt-like 4 (Drosophila) [HGNC:26677]

The first part of the table corresponds to the top-25 genes where up-regulation corresponds to shorter survival and higher risk (i.e., HR > 1); the second part of the table corresponds to the top-25 genes where UP-regulation corresponds to longer survival and lower risk (HR < 1). The genes were ranked by their KM adjusted *p* values and the Hazard Ratio values calculated for the whole dataset, i.e. for all the 1273 samples (all-dt). The stability and robustness of the gene survival markers was assessed by cross-validation, applying to each gene a resampling strategy with random selection of 80% of the samples 100 times (i.e. doing 100 iterations). For the ranking we also considered that the genes had to give a significant adjusted p-value in more than 80 iterations (N-sinf-in-100i > 80)

using the same normalization protocol. After this integration, we could explore the expression level of the top up-regulated genes (identified as markers of poor survival), comparing the expression distribution on a set of cancer samples versus a set of normal tissue samples. In both cases the number of samples compared were 25, since this is the number of normal samples that we had. We did this comparison 20 times, random selecting each time a different subset of 25 cancer samples. The results were always very similar and the boxplots of the expression distributions for the top 10 genes are presented in Additional file 6: Figure S2. These results indicate that the gene markers, identified in our survival studies, are most of the times also up-regulated in CRC tumors with respect to normal colorectal tissue.

Risk predictor score based in the multivariate analysis of candidate survival markers

Finally, to obtain a more accurate evaluation of the prognostic value of all the genes selected as best candidates

(reported in Additional files 2 and 3, Table S2 and Table S3), we performed another analysis of the candidate markers using a regularized multivariate Cox proportional-hazards regression with L1 norm penalty [16], with the scope of building a multigenic "risk predictor". This analysis was done on the cohort of 1273 samples of CRC patients, using for the multivariate analysis the top 100 genes that showed up-regulation correlated with poor prognosis (i.e. overexpressed in low survival cases). The results are presented in Fig. 4 that shows a graph ordering the patients according to their risk score, from low-risk (blue) to high-risk (red), including also an intermediate region (grey) (Fig. 4a). A recursive algorithm using 10-fold cross-validation was applied to find the value of risk score. The threshold (marked with a vertical black line) is obtained by maximizing the separability between the survival curves for the resulting groups. Therefore, it allows the best splitting of the cohort in two groups. A Kaplan-Meier plot showing the separation of these two groups is also presented (Fig. 4b); dividing the population into a high risk group including 425 individuals and a low risk group including 848 individuals. As shown, the division is significant (p-value = 8.25e-14) and allows an optimal separation of individuals according to their survival. The analysis of the beta factors assigned by the regression to each of the top 100 genes, i.e. to each variable within the multivariate vector (data included in Additional file 7: Table S5), allows the identification of the genes that were the most influential factors in this risk analysis and therefore it facilitated the selection of the best "gene survival markers". As indicated in previous sections, the top 100 genes included in the construction of this multigenic risk predictor score were selected from the list of best markers found during the survival test with single genes.

Discussion

CRC is a complex disease composed of biologically and clinically diverse subtypes, which can originate in different ways provoking multiple clinical scenarios [1, 2]. This complexity causes the molecular characterization of CRC to remain deficient, with a lack of clear gene markers associated to specific CRC subtypes and to the prognosis of the disease [17–19]. In fact, current molecular phenotyping of colorectal tumors is usually linked to the

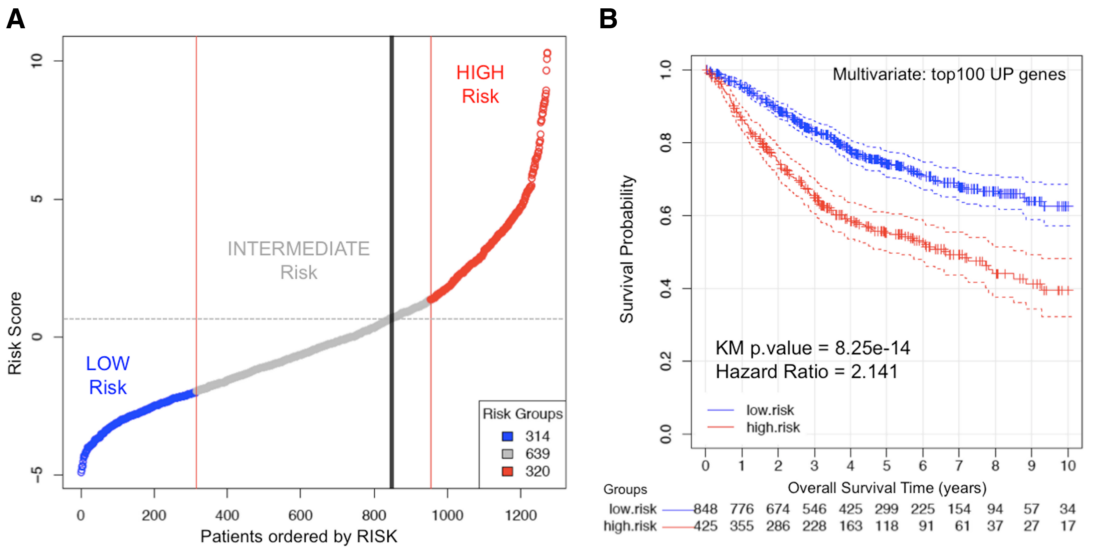


Fig. 4 Risk prediction done for the cohort of 1273 patients of CRC based in the multivariate analysis using the top 100 genes that showed up-regulation correlated with poor prognosis (i.e. overexpressed in low survival cases). **a** Plot presenting the patients according to their risk score, from Low (blue) to High (red) risk. A recursive algorithm using 10-fold cross-validation finds the value of risk score (marked with a vertical black line) that allows the best splitting of the cohort in two groups. **b** Kaplan-Meier plot showing the separation of these two groups: a high-risk group including 425 individuals (in red) and a low-risk group including 848 individuals (in blue). The analysis has been done using a multivariate Cox proportional-hazards regression. As shown, the division is very significant (p-value = 8.25e-14) and allows an optimal separation of individuals according to their survival. The analysis of the *beta* factors assigned by the regression to each of the top 100 genes (i.e. to each variable within the multivariate vector) allows the identification of the genes that are the most influential factors in this risk analysis and therefore it helps in the selection of the best "gene survival markers"

traditional determination of somatic mutations in well-known oncogenes such as KRAS and BRAF [20].

The recent advance of genomic and transcriptomic technologies applied to the study of clinical samples did open the way to obtain genome-wide expression profiles of multiple patient cohorts and correlate the expression of certain genes with different disease subtypes, disease stages and progression [21, 22]. This approach had been widely applied in cancer research in the last decade and is very powerful when the identification of marker genes is associated with survival time. The correlation between gene expression and survival is an excellent tool to investigate prognosis of the disease and to build risk predictors that will be applicable to individual patients.

The identification of molecular biomarkers with prognostic value in CRC has been a challenging task [23–26]. Molecular prognosis of colorectal tumor samples by transcriptional profiling started about 15 years ago (see review [24]), and in more recent years several specific gene signatures associated with CRC survival have been published [5–7, 27–31]. Despite these efforts, at present there is not a clear compendium of gene markers for CRC survival and it is quite difficult to find consistency in the literature [24]. A clear limitation comes from the fact that, in most of previous studies, the number of tumor samples used to select the genes that enter into the construction of the prognostic predictors is small (i.e., the size of the patient cohorts rarely it is greater than a few hundred individuals). For example, *ColoPrint* is a 18-gene signature for prognosis prediction of stage II and III CRC, that was identified using as training set tumor samples from 188 patients [7, 27]; a 113-gene expression signature for predicting prognosis in patients with CRC was built using 145 samples as dicovery set [28]; a 7-gene signature to predict overall survival of CRC patients was based in an initial training set of 67 samples [29]; a recurrence-associated CRC signature of 13 genes was developed using a screening set of 145 samples [30]; a 15-gene signature for prediction of CRC recurrence and prognosis was elaborated using for the gene selection a set of 55 patients [31]. In conclusion, we can say that as far as it is reflected in the current literature, the size of the initial training sets used to identify candidate gene markers for CRC survival is small and the overlap between the published gene signatures is very reduced and inconsistent. To address these critical problems, we constructed a large, well-standardized, integrated data set of 1273 tumor samples with survival information, which was used to identify genes that had a clear change in expression in the middle and late stages of CRC and were consistent markers of the disease-outcome and patient-risk.

With respect to the specific genes proposed as CRC survival markers, we want to underline that our study does not pretend to provide a fixed gene signature for prognosis and risk prediction, like the reported signatures of 7-genes, 15-genes or 113-genes [28, 29, 31]; but instead we propose a robust set of genes ranked according to their predictive power of CRC survival. In this way, an ordered list of 200 genes including the best survival markers is presented: 100 genes for which up-regulation marks "poor survival" and 100 genes for which up-regulation marks "good survival". We think that this approach is more useful, since it allows an open selection of different number of genes for further purposes or investigations (for example, for additional tests with other CRC clinical cohorts). In fact, we used the 100 most significant genes, up-regulated with the progression of CRC, to build the risk predictor (presented in Fig. 4); and we used the top 5 or top 10 genes of this list for the external validations with different independent datasets.

Another relevant comment is that, as reminded above, we constructed the risk predictor using the genes that showed up-regulation correlated with poor prognosis. This was done because in the selection of biomarkers it is better to use the ones that provide a positive signal (i.e. "gain-of-function" factors) than the ones that provide a negative signal. Therefore, all the gene survival markers that we proposed were detectable as overexpressed in the CRC patients with high risk. The fact that they give a positive signal will also make easier their detection by standard biomolecular protocols (PCR, ELISA, immunohistochemistry, etc).

Finally, we are investigating the biological meaning of the genes found as best predictive and prognostic markers. We are focusing our efforts in the top 10 for which up-regulation marked poor survival: DCBLD2, PTPN14, LAMP5, TM4SF1, NPR3, LEMD1, LCA5, CSGALNACT2, SLC2A3, GADD45B. The analysis of the literature reveals some relevant observations. For example, the transmembrane protein DCBLD2 (ESDN), member of a family of neuropilin-like proteins, is a novel regulator of mitotic and metabolic effects of insulin, and it modulates signal transduction through regulation of the insulin receptor interaction with its adaptor proteins [32]. The importance of insulin regulation in the function of our digestive system is clear, and this adds extra value to the proposal of DCBLD2 as a CRC survival marker. Other genes within the top rank have been recently involved in cancer progression, like the case of SLC2A3 (GLUT3) a glucose transporter that mediates glucose utilization and glycogenolysis, which is induced during epithelial-mesenchymal transition and promotes tumor cell proliferation [33]. Recent publications have also proposed the role of some other genes found as prognostic markers, like the case of LAMP5 that has been included in a multigenic assay to predict

recurrence for gastric cancer patients after surgery [34]. As a final example, GADD45B (growth arrest and DNA-damage-inducible 45 beta) is a gene that responds to environmental stresses, associated with cell growth control, apoptosis and DNA damage repair response. GADD45B overexpression has been recently correlated with shorter overall survival in colorectal carcinoma [35]. Moreover, a recent integrative analysis of multiple colon cancer gene-expression-based subtype classifiers reported that one of the three highest scoring genes included in several classifiers was GADD45B [36].

Despite all these positive findings that correspond to the biological value and the support of the genes identified as most significant markers of CRC survival, there are some possible limitations of the results, beginning with the general observation about the frequent heterogeneity of the colorectal tumors [1, 17]. In fact, it is clear from the anatomical pathology that CRC can affect quite different regions of the digestive tract: ascending colon, transverse colon, descending colon, sigmoid colon and rectum. The causal genes that drive tumors in these different regions may not be the same, and most CRC studies do not enter into a detailed separation of these regions [19]. The variability due to the different staging of the tumors is another factor that can bring limitations to any CRC study; but in this case we clearly indicated that our work searched for genes that were candidate prognostic markers for CRC in stages III and IV. A final reason for the limitations of the results may be an over-adjustment to the tested data sets. To avoid this kind of limitations, we built a large well-normalized data set with more than a thousand samples, performed a cross-validation analysis on that set, and also explored the validity of the gene markers in two other independent sets.

Conclusions

In conclusion, we consider that the results presented in this work provide strong support and a solid rationale for the prognostic value of a new set of genes in CRC and for their potential to predict colorectal tumor progression and evolution towards stages III and IV. The final proposed set of gene survival markers includes an open list of one hundred up-regulated genes, with a robust statistical estimation of the value of each one. In this way the set of genes is clearly ranked, being the top in the list the ones that provide best prognostic strength and the ones that can be introduced to build smaller predictors. In fact, our results showed that a selection of the top 5 genes applied to independent external cohorts provided very good separation of CRC samples in two distinct groups of high and low risk.

Methods

Genome-wide expression data sets

In this study, we have analysed and integrated seven data sets of CRC samples (Table 1). All data sets are available at GEO repository, corresponding to 7 series with the following accession numbers: GSE14333, GSE17536, GSE31595, GSE33113, GSE38832, GSE39084 and GSE 39582. All these series included the raw expression signal and correspond to data obtained with the microarrays expression platform: *Affymetrix GeneChip U133 Plus 2.0* for *Homo sapiens*. The phenotypic information corresponding to all these series was analysed in order to select only the samples that included information regarding: the cancer *stage* and the *Overall Survival* (OS). The samples that did not have any survival information were discarded from the study. In all cases only primary tumors samples were considered for our analysis; in this way individuals who had received preoperative chemotherapy and/or radiotherapy were also discarded.

For the external validation we used two independent datasets. A cohort of 276 colorectal carcinomas that had been studied using RNA-seq gene expression profiling, and that had survival data for 269 samples [14] (which can be found in http://tcga-data.nci.nih.gov/docs/publications/coadread_2012/). A second cohort of CRC samples from the platform SurvExpress [15]. This second dataset selected, called "Colon-Metabase-Uniformized", included 482 CRC samples with overall survival data and genome-wide expression determined with *Affymetrix* microarrays (see the website http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp).

Expression data sets exploration and integrative normalization

Previously, to make the best use of the information obtained from the microarrays, we have considered the importance to ascertain the quality of the data. To assess the validity of generated microarray information we have performed a wide variety of quality assessment methods, both in raw and pre-processed information. In this way, several explanatory data analysis were applied for the detection of problematic arrays. We used the R function *image* to create chip images of the raw intensities to discover spatial artefacts in the samples. We have also look at the distribution of probes intensities across all arrays, using the *boxplot* method available for the Affybatch class. We also applied to the samples the Normalized Unscaled Standard Error (NUSE) algorithm. This quality assessment tool requires a previous PLM fitting procedure applied on the raw expression data. We have used the function *fitPLM* provided in the *AffyPLM* package to create the *PLMset* class object used as the input in the elaboration of the NUSE analysis. After applying the referred quality assessment methods, we discarded 79 of the initial samples collected and proceed with the remaining 1273 (Table 1).

To create a table with all the phenotypic characteristics of the patients selected which involved all samples

GSM accession numbers and related clinic variables in a consistent and homogenize way, we used *getGEO* and *pData* functions from *GEOquery* package (this table is provided as Additional file 1: Table S1). We made use of regular expressions and common text manipulation R functions to solve the issue of formatting heterogenic data. Finally, we created a binary variable to label the patients and select them in a proper way during the hypothesis contrasts and statistical modeling.

Batch effect removal

Batch effect is one of the main problems when several datasets are combined to be studied together, because different batches usually add large unwanted variability to the data. To avoid this effect we tested a combination of different pre-processing and normalization algorithms: Robust Multi-array Average (RMA) algorithm [11]; Combatting Batch effects (ComBat) algorithm [12]; Frozen Robust Multi-array Average (fRMA) algorithm [13]. For the fRMA algorithm application, we constructed the frozen parameter vector using a training dataset in where we distributed randomly selected samples proportionally to each labelled group to obtain a balanced sample from the 7 batches of microarrays.

Another important issue addressed was the fact that the *Affymetrix* probe-sets included in the expression microarrays many times do not correspond to singular genes and some probes inserted in the defined probe-sets are ambiguous or inaccurate [10]. *Affymetrix* GeneChip is a popular and usefull platform for gene expression profiling, but the use of its probes and probe-sets mapping has multiple inconveniences. In fact, the probe-sets for the *Affymetrix* Human Genome U133 Plus 2.0 Array are based on UniGene database (Build 133, April 20, 2001) and considering how rapidly human genome has evolved many probes on the array are not correctly assigned. To avoid this problem, we used the updated probe alignment and gene mapping that is provided by the Chip Definition File (CDF): *hgu133plus2hsensgcdf* (downloaded from http://brainarray.mbni.med.umich.edu/).

Batch effect removal evaluation

We performed unsupervised hierarchical clustering to observe unlikely clustering based on batches in those expression value matrixes where batch effects remained after pre-processing. We used a 30-random sampling per batch, identifying each batch by a different color (Fig. 1). The batch effect was also investigated using principal components analysis (PCA) (Fig. 2). A linear regression of average gene expression on array batch per pre-processing method was the final approach fulfilled to assure removal (Table 2).

Differential expression analysis

For the identification of gene whose altered expression achieved statistical significance we used the R algorithm Linear Models for Microarrays (LIMMA package). We applied LIMMA to the expression data matrix fixing an adjusted *p*-value threshold of FDR ≤ 0.01 to select significant genes. The comparison was done separating the samples according to their clinical and pathological stage (comparing CRC stages I and II versus III and IV). In this way we found a set of 2707 candidates genes, corresponding to 2524 protein-coding genes that were tested in the survival analysis (the rest were non-coding genes). In this work we focus only on the genes that encode proteins because we wanted to find CRC survival markers that later can be tested at protein level using, for example, immunohistochemistry (IHC) analysis.

Survival analysis

Our intention in this research was to identify genes whose relative expression level affect survival and prognosis in CRC, once we had made a preselection in its behavior through stage evolution of 2524 protein-coding genes.

The first step for the survival analysis was to define for each gene two separated distributions of high and low expression along the sample dataset investigated. This separation based in expression level determined the explanatory variable. We used the *Surdiff* function in the *Survival* package to address the issue. By sorting all the samples in ascending order, we performed *Surdiff* hypothesis testing, splitting the group of samples for each gene and every sample between quantile 25% and 75% to obtain its Chi-square associated p-value. Then we selected minimum p-value to perform final group assignation of high and low expression. Once we had the two groups clearly defined, we used the *Coxph* model to obtain each associated p-value and hazard ratio (HR) from every candidate gene. In this way, the survival analysis along the two groups also allowed estimating hazard ratios (HR) or, what is the same, tried to measure how the expression, in terms of high and low relative expression for each candidate gene, altered the hazard function. Finally, for computing the time to event, the response variable in the models was the *Overall Survival* (OS) time. All the data sets that we integrated in our analyses had OS information. In some cases for some individuals, *Disease Specific Survival* (DSS) times or *Relapse Free Survival* (RFS) times were also provided with the original data, but we did not considered these time-events since we wanted to focus on OS to achieve a homogeneous analysis.

Additional files

Additional file 1: Table S1. Phenotypic and clinical information about the collection of 1273 colorectal cancer samples that has been integrated in this work. The table includes the IDs of the samples in GEO and all the

available data about age, gender, survival time, location of the tumor, degree and TNM staging, presence of mutation in some cancer genes (TP53, KRAS, BRAF), etc. When information was not available for a given sample the table includes NA (not available values). (XLSX 272 kb)

Additional file 2: Table S2. Top-100 best survival marker genes for colorectal cancer (CRC) that are up-regulated when survival is poor and the risk is higher (i.e., HR > 1). This table is an expension of the data in Table 3. The genes were ranked by their KM adjusted *p*-values and the HR values calculated for the whole dataset (i.e. for all the 1273 samples = all-dt). The stability of each survival marker gene was assessed by cross-validation (100 iterations). The table also includes the number of times that a survival marker was significant in the iterations (N-sinf-in-100i). (XLSX 73 kb)

Additional file 3: Table S3. Top-100 best survival marker genes for colorectal cancer (CRC) that are down-regulated when survival is poor and the risk is higher (i.e., HR < 1). This table is an expension of the data in Table 3. The genes were ranked by their KM adjusted p-values and the HR values calculated for the whole dataset (i.e. 1273 samples = all-dt). The stability of each survival marker gene was assessed by cross-validation (100 iterations). The table also includes the number of times that a survival marker was significant in the iterations (N-sinf-in-100i). (XLSX 70 kb)

Additional file 4: Table S4. Validation of the survival data done in an independent set of samples taken from The Cancer Genome Atlas (TCGA), that included 269 colorectal carcinomas with survival information and RNA-seq global expression profiling. The table includes the KM p-values and HR of the genes that were validated from the top-10 survival marker genes previously found presented in Table 3. Of the top-10 for the case of up-regulation associated with poor survival, 7 were validated (PTPN14, LAMP5, TM4SF1, LCA5, CSGALNACT2, SLC2A3 and GADD45B). Of the top-10 found for down-regulation associated with poor survival, 6 genes were validated (EPHB2, DUS1L, NUAK2, FANCC, MYB and CHDH). (XLSX 51 kb)

Additional file 5: Figure S1. Survival multivariate analysis of an independent set of 482 samples of CRC patients carried out considering the expression profiles of 5 genes: DCBLD2, PTPN14, LAMP5, TM4SF1 and NPR3. **(A)** Kaplan-Meier plot presenting the patients divided in two groups according their risk score: High risk (red) and Low risk (green). **(B)** Box plots showing the distributions of global expression corresponding to these 5 genes. For each gene, the dataset of 482 samples was divided in the two groups of patients indentified as High risk (red) and Low risk (green). (PDF 356 kb)

Additional file 6: Figure S2. Comparison of the distributions of the expression signal corresponding to ten genes in 25 samples from normal colorectal epithelium (green boxplots) versus 25 samples from CRC (red boxplots). The genes selected for this analysis were the top-10 best survival marker genes found up-regulated for poor prognosis (i.e. markers up-regulated when there is low CRC survival): DCBLD2, PTPN14, LAMP5, TM4SF1, NPR3, LEMD1, LCA5, CSGALNACT2, SLC2A3 and GADD45B. The tumor samples were not selected by stage (i.e. they were selected from any CRC stage: I, II, III or IV) and this comparison was done 20 times with different subsets of 25 CRC samples to check the stability of the signal. The plots of all the other comparisons were very similar to the plot here presented. (PDF 46 kb)

Additional file 7: Table S5. Beta factors assigned by regression analysis to each of the top-100 survival marker genes. These genes are taken as variables within the multivariate Kaplan-Meier survival analysis included in Fig. 4b. The factors allowed the identification of the genes that were the most influential variables in this risk analysis (i.e. the higher the better) and therefore facilitate an additional evaluation of each survival marker gene. (XLSX 62 kb)

Abbreviations

CDF: Chip definition file; CRC: Colorectal cancer; DSS: Disease specific survival; GEO: Gene expression omnibus database; GSE: GEO Series (set of sample files that together form a single experiment); HR: Hazard ratio; IHC: Immunohistochemistry; KM: Kaplan-Meier hazard ratio; LIMMA: Linear models for microarray data analysis; OS: Overall survival; PCA: Principal component analysis; RFS: Relapse free survival; RMA: Robust multi-array average algorithm; TCGA: The cancer genome atlas

Acknowledgements

We acknowledge the funding provided to *JDLR* research group by the Spanish Government with grants of the ISCiii co-funded by FEDER (references PI15/00328 and AC14/00024). We also acknowledge a PhD research grant to *SBF* (from the Program "Ayudas a la Contratación de Personal Investigador") provided by the "Junta de Castilla y Leon" (JCyL) with the support of the "Fondo Social Europeo" (FSE). The funding boards had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Funding

The publication costs for this article were funded by the research grants AC14/00024 and PI15/00328, from the Instituto de Salud Carlos III (ISCiii) co-funded by the "Fondo Europeo de Desarrollo Regional" (FEDER).

Availability of data and materials

All the data presented in this study is provided free and open to be used, included in the Supplementary Files that are quoted and described along the manuscript.

About this supplement

This article has been published as part of *BMC Genomics Volume 19 Supplement 8, 2018: Selected articles from the IV Colombian Congress on Bioinformatics and Computational Biology & VIII International Conference on Bioinformatics SoIBio 2017*. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-19-supplement-8.

Authors' contributions

JMR carried out the data collection, the databases construction and together with SBF the R progamming developments, the computational analyses and the datasets comparisons. They also contributed to write the manuscript. JDLR designed the study with the support of ARM and MMM. JDLR devised and designed the study, identified the experimental data sets used in the tests and validations, supervised the R programming, wrote the manuscript and managed the authors' collaboration. ARM and MMM also contributed to the design of the work and help in the preparation of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Ethics approval and consent to participate is "not applicable", because this work does not include samples from new patients or donors. All the information and data of human samples used in this work come from data sets already public in open repositories and corresponded to Anonymized Patient Level Data (APLD). Moreover, the Ethical Committees of our Research Centers (CiC-IBMCC and IMDEA-Food) supervised the adequate use of the data corresponding to human samples.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Publisher's Note

Author details

[1]Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Cientificas (CSIC) and University of Salamanca (USAL), Salamanca, Spain. [2]Molecular Oncology and Nutritional Genomics of Cancer Group, Precision Nutrition and Cancer Program, IMDEA Food Institute (CEI, UAM/CSIC), Madrid, Spain. [3]Department of Computer Science, Universidad Pontificia de Salamanca (UPSA), Salamanca, Spain.

Published: 11 December 2018

References

1. Linnekamp JF, Wang X, Medema JP, Vermeulen L. Colorectal cancer heterogeneity and targeted therapy: a case for molecular disease subtypes. Cancer Res. 2015;75:245–9.
2. Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nat Rev Cancer. 2017;17:79–92.
3. Liu R, Zhang W, Liu ZQ, Zhou HH. Associating transcriptional modules with colon cancer survival through weighted gene co-expression network analysis. BMC Genomics. 2017;18:361.
4. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015; 21:1350–6.
5. Vargas T, Moreno-Rubio J, Herranz J, Cejas P, Molina S, González-Vallinas M, et al. ColoLipidGene: signature of lipid metabolism-related genes to predict prognosis in stage-II colon cancer patients. Oncotarget. 2015;6:7348–63.
6. Sveen A, Ågesen TH, Nesbakken A, Meling GI, TO R, Liestøl K, et al. ColoGuidePro: a prognostic 7-gene expression signature for stage III colorectal cancer patients. Clin Cancer Res. 2012;18:6001–10.
7. Kopetz S, Tabernero J, Rosenberg R, Jiang ZQ, Moreno V, Bachleitner-Hofmann T, et al. Genomic classifier ColoPrint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. Oncologist. 2015;20:127–33.
8. The American Cancer Society medical and editorial content team. Colorectal Cancer Stages. https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.html. Accessed 06 Oct 2017.
9. Tauriello DVF, Batlle E. Targeting the microenvironment in advanced colorectal Cancer. Trends Cancer. 2016;2:495–504.
10. Risueño A, Fontanillo C, Dinger ME, De Las Rivas J. GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs. BMC Bioinformatics. 2010;11:221.
11. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe-level data. Biostatistics. 2003;4:249–64.
12. Stein CK, Qu P, Epstein J, Buros A, Rosenthal A, Crowley J, et al. Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. BMC Bioinformatics. 2015;16:63.
13. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010;11:242–53.
14. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7.
15. Aguirre-Gamboa R, Gomez-Rueda H, Martínez-Ledesma E, Martínez-Torteya A, Chacolla-Huaringa R, Rodriguez-Barrientos A, et al. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. PLoS One. 2013;8:e74250.
16. Gui J, Li H. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics. 2005;21:3001–8.
17. Sameer AS. Colorectal cancer: molecular mutations and polymorphisms. Front Oncol. 2013;3:114.
18. Fessler E, Medema JP. Colorectal Cancer subtypes: developmental origin and microenvironmental regulation. Trends Cancer. 2016;2(9):505–18.
19. Bijlsma MF, Sadanandam A, Tan P, Vermeulen L. Molecular subtypes in cancers of the gastrointestinal tract. Nat Rev Gastroenterol Hepatol. 2017; 14(6):333–42.
20. Kocarnik JM, Shiovitz S, Phipps AI. Molecular phenotypes of colorectal cancer and potential clinical applications. Gastroenterol Rep. 2015;3(4):269–76.
21. Aibar S, Fontanillo C, Droste C, Roson-Burgo B, Campos-Laborie FJ, Hernandez-Rivas JM, et al. Analyse multiple disease subtypes and build associated gene networks using genome-wide expression profiles. BMC Genomics. 2015;16(Suppl 5):S3.
22. Aibar S, Abaigar M, Campos-Laborie FJ, Sánchez-Santos JM, Hernandez-Rivas JM, De Las Rivas J. Identification of expression patterns in the progression of disease stages by integration of transcriptomic data. BMC Bioinformatics. 2016;17(Suppl 15):432.
23. Moreno V, Sanz-Pamplona R. Altered pathways and colorectal cancer prognosis. BMC Med. 2015;13:76.
24. Sanz-Pamplona R, Berenguer A, Cordero D, Riccadonna S, Solé X, Crous-Bou M, et al. Clinical value of prognosis gene expression signatures in colorectal cancer: a systematic review. PLoS One. 2012;7(11):e48877.
25. George B, Kopetz S. Predictive and prognostic markers in colorectal cancer. Curr Oncol Rep. 2011;13(3):206–15.
26. Das V, Kalita J, Pal M. Predictive and prognostic biomarkers in colorectal cancer: a systematic review of recent advances and challenges. Biomed Pharmacother. 2017;87:8–19.
27. Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol. 2011;29(1):17–24.
28. Nguyen MN, Choi TG, Nguyen DT, Kim JH, Jo YH, Shahid M, et al. CRC-113 gene expression signature for predicting prognosis in patients with colorectal cancer. Oncotarget. 2015;6(31):31674–92.
29. Chen H, Sun X, Ge W, Qian Y, Bai R, Zheng S. A seven-gene signature predicts overall survival of patients with colorectal cancer. Oncotarget. 2016; 8(56):95054–65.
30. Tian X, Zhu X, Yan T, Yu C, Shen C, Hu Y, et al. Recurrence-associated gene signature optimizes recurrence-free survival prediction of colorectal cancer. Mol Oncol. 2017;11(11):1544–60.
31. Xu G, Zhang M, Zhu H, Xu J. A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. Gene. 2017;604:33–40.
32. Li X, Jung JJ, Nie L, Razavian M, Zhang J, Samuel V, et al. The neuropilin-like protein ESDN regulates insulin signaling and sensitivity. Am J Physiol Heart Circ Physiol. 2016;310:H1184–93.
33. Masin M, Vazquez J, Rossi S, Groeneveld S, Samson N, Schwalie PC, et al. GLUT3 is induced during epithelial-mesenchymal transition and promotes tumor cell proliferation in non-small cell lung cancer. Cancer Metab. 2014;2:11.
34. Lee J, Sohn I, Do IG, Kim KM, Park SH, Park JO, et al. Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery. PLoS One. 2014;9:e90133.
35. Wang L, Xiao X, Li D, Chi Y, Wei P, Wang Y, Ni S, Tan C, Zhou X, Du X. Abnormal expression of GADD45B in human colorectal carcinoma. J Transl Med. 2012;10:215.
36. Sztupinszki Z, Győrffy B. Colon cancer subtypes: concordance, effect on survival and selection of the most representative preclinical models. Sci Rep. 2016;6:37169.

Appendix: R functions and code

Title
Risk, Survival and Marker Gene selection
Version
1.0
Author
Santiago Bueno-Fortes
Description
Risk prediction and marker gene selection for survival analysis using the Cox and Glmr models Especially useful for high-dimensional data, including microarray data, RNAseq data...
Maintainer
Santiago Bueno-Fortes
<sjbuenofortes@usal.es>
Depends
survival, rbsurv, survcomp, uniCox, scales

Functions documented:
function_km_bootstrap
function_km_groups
function_riskTenfold
function_sam
function_PredGlmnet
function_km_groups_plot
function_kmplot

function_km_bootstrap

Description

Function to compute the Kaplan-Meier curves with probability of classification for microarray or RNAseq.

Usage

```
funcion_km_bootstrap(geneExpr, time, status, geneName)
```

Arguments

geneExpr: vector containing gene expression (normalised), including the sample name in names(geneExpr).
time: vector containing the survival time for each patient/sample, including the sample name in names(time).
status: vector containing the survival status(censored or not values 0 or 1'repasar cual es cada') for each patient/sample, including the sample name in names(status).
geneName: string containing the name of the gene (string).

Details

This function computes the Kaplan-Meier curves with probability of classification using vectors for expression, time and status, generating a pdf file with the plots in Workspace Directory.

Value

A list with components:

- vector of group classification 2=high expression, 1=low expression
- vector of probability of classification
- p.value of logrank test

Examples

```
library(survival)
library(rbsurv)
library(survcomp) # risk score


# extracting mExpr, TIME and STATUS from rbsurv
data(gliomaSet)
mExpr <- exprs(gliomaSet)
mExpr <- log2(mExpr)
```

```
# take into account that the order of time AND status MUST be
   the same in expression matrix's COLUMNS
# the TIME value must be transformed to YEARS
gene<-mExpr[1,]
time <- gliomaSet$Time/365
names(time)<-colnames(mExpr)
status <- gliomaSet$Status
names(status)<-colnames(mExpr)
# the gene expression vector must be provided with the NAMES
   of each sample, which match with the NAMES of the time and
   status vectors

# km curves from expression
# we need: the expression vector, the time and status vectors,
    also the name of the gene to display it in the plots and
   pdf name
funcion_km_bootstrap(gene, time, status, "Gene Name")
```

Function

```
function_km_bootstrap <- function( genExpr, time.import,
   status.import, genName)
{
  # error control
  if(length(genExpr)!=length(time.import)){message("ERROR: the
      expr values and time vector must have the same length")}
  if(length(genExpr)!=length(status.import)){message("ERROR:
     the expr values and status vector must have the same
     length")}
  # gene name in vectors time status
  if(!identical(names(genExpr), names(time.import))){message("
     ERROR: expr values names must match time vector names")}
  if(!identical(names(genExpr), names(status.import))){message
     ("ERROR: expr values names must match status vector names
     ")}

  mSurv<-cbind(time.import, status.import)
  colnames(mSurv)<-c("time","status")
  mSurv<-as.data.frame(mSurv)
  rownames(mSurv)<-names(time.import)
  iter<-100

  library("survcomp")
  mSurv$status[mSurv$time>10]<-1
  mSurv$time[mSurv$time>10]<-10.1
  n.samples <- length(genExpr)

  for25<-round(n.samples*0.25)
  for75<-round(n.samples*0.75)
```

```
vector.exprs <- as.numeric(genExpr)
order.vector.exprs <- order( vector.exprs )


muestra<-NULL
for(i in 1:iter){
  m<-sample( 1:n.samples, size=round(n.samples), replace=
     TRUE)
  muestra[[i]] <- m
}
muestra[[1]]<-head(1:n.samples, round(n.samples))
muestra[[2]]<-tail(1:n.samples, round(n.samples))

matrixgr<- matrix(0, nrow = n.samples, ncol = iter)
rownames(matrixgr)<-names(genExpr)


for(i in 1:iter){
  if(i%%50==0){(print(i))
    print(Sys.time())}
  genExpr2<-genExpr[muestra[[i]]]
  mSurv2<-mSurv[muestra[[i]],]

  # funcion is the funcion_km_groups.txt
  g<-funcion_km_groups(genExpr2,mSurv2, genName)
  g<-g[[1]]
  names(g)<-names(genExpr2)
  matrixgr[match(names(g), rownames(matrixgr) ) ,i]<- as.
     vector(g)

}
group.assignation.vector<-NULL
group.assignation.vector[[1]] <-apply(matrixgr, 1, function(
   x) round( mean(x[x!=0]) ))
group.assignation.vector[[2]] <-group.assignation.vector[[1
   ]]
for (k in 1:n.samples) {
  group.assignation.vector[[2]][k]<-sum(matrixgr[k,]==group.
     assignation.vector[[1]][k])/sum(matrixgr[k,]!=0)
}
p.val <- NULL
print(table(is.na(group.assignation.vector[[2]])) )
print(vector.exprs[group.assignation.vector[[1]]==1])
print(vector.exprs[group.assignation.vector[[1]]==2])

log.rank.groups.surv <- survdiff( Surv( time, status) ~
   group.assignation.vector[[1]], data= mSurv )
p.val <- 1 - pchisq(log.rank.groups.surv$chisq, length(log.
   rank.groups.surv$n) - 1)
```

```
  fits1 <- survfit( Surv( time, status) ~ group.assignation.
     vector[[1]], data= mSurv )
  jj<-genName
  # library hazardR

  pdf( paste(jj,".pdf"))
  kmplot( fits1 ,xaxis.at=c(0:10), col.surv=c(3,2),  group.
     names=c('low.exp','high.exp'), xlab=paste('Survival(years
     ) and p.value: ',format(p.val,3) ) , ylab=paste('Survival
      Probability','\n Hazard Ratio:',format(1/hazardR$hazard.
     ratio,3),' (',format(1/hazardR$lower,3),', ',format(1/
     hazardR$upper,3), ')'), legend=TRUE, loc.legend='
     bottomleft', lty.ci=2, lwd.ci=1, col.ci=c(2,3), xlim=c(0,
     10)  )

  Ttest<-wilcox.test(vector.exprs[group.assignation.vector[[1
     ]]==1],vector.exprs[group.assignation.vector[[1]]==2])
  boxplot(vector.exprs[group.assignation.vector[[1]]==1],
     vector.exprs[group.assignation.vector[[1]]==2], col=c(3,2
     ), ylab="expression level", xlab="wilcox-test p.value for
      groups:", sub=format( Ttest$p.value ))
  legend("bottomright", title=jj, c(as.character(sum(group.
     assignation.vector[[1]]==1)),as.character(sum(group.
     assignation.vector[[1]]==2))), fill=c(3,2), horiz=TRUE)

  dev.off()


  group.assignation.vector[[3]]<-p.val
  group.assignation.vector
}
```

function_km_groups

Description

Function to compute the Kaplan-Meier logrank value and groups to be used in the call of funcion_km_bootstrap for microarray or RNAseq.

Usage

```
funcion_km_grupos(geneExpr, mClin, geneName)
```

Arguments

geneExpr: vector containing gene expression (normalised), including the sample name in names(geneExpr).
mClin: data frame created with time and status vectors provided to funcion_km_bootstrap.
geneName: string containing the name of the gene (string).

Details

This function computes the Kaplan-Meier logrank value and groups to be used in the call of funcion_km_bootstrap using vectors for expression and the clinical data matrix.

Value

A list with two elements:

- a vector defining the groups of high expression and low expression
- the p.value obtained by maximizing the separability of the KM curves

Examples

```
library(survival)
library(rbsurv)
library(survcomp)
library(uniCox)
library(scales)

# extracting mExpr, TIME and STATUS from rbsurv
data(gliomaSet)
mExpr <- exprs(gliomaSet)
mExpr <- log2(mExpr)
# take into account that the order of time AND status MUST be
   the same in expression matrix's COLUMNS
# the TIME value must be transformed to YEARS
```

```
gene<-mExpr[1,]
time <- gliomaSet$Time/365
names(time)<-colnames(mExpr)
status <- gliomaSet$Status
names(status)<-colnames(mExpr)
# the gene expression vector must be provided with the NAMES
   of each sample, which match with the NAMES of the time and
   status vectors

matrix.surv<-cbind(time, status)
colnames(matrix.surv)<-c("time","status")
matrix.surv<-as.data.frame(matrix.surv)
rownames(matrix.surv)<-names(time)

# 1: km curves from expression
# we need: the expression vector, the time and status vectors,
    also the name of the gene to display it in the plots and
   pdf name
funcion_km_grupos(gene, matrix.surv, "Gene Name2")
```

Function

```
funcion_km_groups <- function( genExpr, mSurv , list.genes)
{
  matrix.groups<-NULL
  n.samples <- length(genExpr)
  for25<-round(n.samples*0.25)
  for75<-round(n.samples*0.75)
  p.value.genes.ordering <-0

  for( j in list.genes )
  {
    vector.exprs <- as.numeric(genExpr)
    order.vector.exprs <- order( vector.exprs )

    group.assignation.vector <- rep(0,n.samples)
    p.val <- rep(1,n.samples)
    for( i in for25:for75 )
    {
      group1 <- order.vector.exprs[1:i ]
      group2 <- order.vector.exprs[ (i+1):n.samples ]
      group.assignation.vector[ group1 ] <- 1
      group.assignation.vector[ group2 ] <- 2
      log.rank.groups.surv <- survdiff( Surv( time, status) ~
         group.assignation.vector, data= mSurv )
      p.val[i] <- 1 - pchisq(log.rank.groups.surv$chisq,
         length(log.rank.groups.surv$n) - 1)

    }
```

```
    ordered.pval.indexes <- order(p.val)
    lowest.pvalue.index <- ordered.pval.indexes[1]
    group1 <- order.vector.exprs[1:lowest.pvalue.index ]
    group2 <- order.vector.exprs[ (lowest.pvalue.index+1):n.
       samples ]
    group.assignation.vector[ group1 ] <- 1
    group.assignation.vector[ group2 ] <- 2

    p.value.genes.ordering[j] <- p.val[ lowest.pvalue.index ]
    if(j ==1){
      matrix.groups<-group.assignation.vector
    }else{
      matrix.groups<-rbind(matrix.groups, group.assignation.
         vector)
    }
    fits1 <- survfit( Surv( time, status) ~ group.assignation.
       vector, data= mSurv )
  }
list( as.vector(matrix.groups), p.val[ lowest.pvalue.index ])
}
```

funcion_riskTenfold

Description

Function to compute the multivariate ten fold crossvalidated calculation of risk score given a group of candidate marker genes for microarray or RNAseq, the train and validation groups may be the same but it's not recommended.

Usage

```
function_riskTenfold(mExprTrain,  time, status, lambda,
   mExprValid)
```

Arguments

mExprTrain matrix containing by rows the gene expression (normalised) for each candidate, including the sample (columns) names in colnames(mExprTrain) and gene names in rownames(mExprTrain) for training.
time vector containing the survival time for each patient/sample, including the sample name in names(time).
status vector containing the survival status(censored or not) for each patient/sample, including the sample name in names(status).
mExprValid matrix containing gene expression (normalised), including the sample names in colnames(mExprValid) and gene names in rownames(mExprValid).
lambda lambda value, 0 (include all candidates) by default, with higher lambdas some of the worst candidate genes will be removed.

Details

This function computes the multivariate ten fold crossvalidated calculation of risk score.
This function generates two pdf plots:
1) riskANDpval.pdf

- page 1: the p.values, showing the local minima, and the global minimum in the ordered training risk scores, used to get the cutpoint for predicted risk scores.
- page 2: ordered TRAIN risk score groups calculated from previous analysis.
- page 3: ordered VALIDATION risk score groups calculated from previous analysis.

2) RiskTrain.pdf

- page 1: Kaplan Meier plot with two groups, the high and low risk. Training dataset.
- page 2: ordered TRAIN risk score groups calculated from previous analysis.

Examples

```
library(survival)
library(survcomp)
library(uniCox)
library(scales)

# extracting mExpr, TIME and STATUS from rbsurv

# take into account that the order of time AND status MUST be
   the same in expression matrix's COLUMNS
# the TIME value must be transformed to YEARS
time <- mSurv380$time
names(time)<-rownames(mSurv380)
status <- mSurv380$censurado
names(status)<-rownames(mSurv380)

timeValid <- mSurv644$time
names(timeValid)<-rownames(mSurv644)
statusValid <- mSurv644$censurado
names(statusValid)<-rownames(mSurv644)
# the gene expression vector must be provided with the NAMES
   of each sample, which match with the NAMES of the time and
   status vectors

geneList<-rownames(mExpr32ER)
# we need: the expression vector, the time and status vectors,
    also the name of the gene to display it in the plots and
   pdf name
mExprTrain<-mExpr380[match(geneList, rownames(mExpr380)),]
mExprValid<-mExpr644[match(geneList, rownames(mExpr644)),]
x<-funcion_riskTenfold(mExprTrain, time, status, mExprValid,
   timeValid, statusValid)
```

Function

```
funcion_riskTenfold <- function(mExpr,  time.import, status.
   import, mExprValid, timeValid, statusValid, lambda=0){
  library(survcomp)
  library(scales)
  library(uniCox)
  # error control
  if(dim(mExpr)[2]!=length(time.import)){message("ERROR: the
     matrix samples and time vector must have the same length"
     )}
  if(dim(mExpr)[2]!=length(status.import)){message("ERROR: the
      matrix samples and status vector must have the same
     length")}
  if(dim(mExprValid)[2]!=length(timeValid)){message("ERROR:
     the matrixValid samples and time vector must have the
     same length")}
```

```
if(dim(mExprValid)[2]!=length(statusValid)){message("ERROR:
    the matrixValid samples and status vector must have the
    same length")}
# gene name in vectors time status
if(!identical(colnames(mExpr), names(time.import))){message(
    "ERROR: mExpr colnames must match time vector names")}
if(!identical(colnames(mExpr), names(status.import))){
    message("ERROR: mExpr colnames must match status vector
    names")}

#defining pData matrix
mSurv<-cbind(time.import,status.import)
rownames(mSurv)<-names(time.import)
colnames(mSurv)<-c("time","status")
mSurv<-as.data.frame(mSurv)
#defining pData matrix validation
mSurvValid<-cbind(timeValid,statusValid)
rownames(mSurvValid)<-names(timeValid)
colnames(mSurvValid)<-c("time","status")
mSurvValid<-as.data.frame(mSurvValid)

folds <- cut(seq(1,length(mExpr[1,])),breaks=10,labels=FALSE
    )
# Perform 10 fold cross validation
betasMatrix<-NULL
riskValidMatrix<-NULL

riskValidDefinitive<-rep(0,length(mExprValid[1,]))
names(riskValidDefinitive)<-names(mExprValid[1,])
riskDefinitive<-rep(0,length(mExpr[1,]))
names(riskDefinitive)<-names(mExpr[1,])
for(i in 1:10){
  # Segement your data by fold using the which() function
  testIndexes <- which(folds==i,arr.ind=TRUE)
  testData <- mExpr[,testIndexes ]
  trainData <- mExpr[,-testIndexes ]
  mSurvUniCox<-mSurv[1:dim(mExpr)[2],]
  fit<-uniCox(t(trainData), mSurvUniCox[-testIndexes,1],
     mSurvUniCox[-testIndexes,2], lamlist=lambda, del.thres
     =.01, max.iter=5) #########incluir explicacion
     parametro lambda y porque lo dejamos en 0, en los
     details de la funcion
  betasMatrix<-cbind(betasMatrix, fit$beta)
  riskScore<-predict.uniCox(fit,t(testData))
  riskDefinitive[testIndexes]<-riskScore
  riskScore<-predict.uniCox(fit,t(mExprValid))
  riskValidMatrix<-cbind(riskValidMatrix,riskScore)
}
```

```
order.riskDefinitive<-riskDefinitive[order(riskDefinitive)]
#50% central trainData logrank
group.assignation.vector<-rep(0,length(riskDefinitive))
p.val<-rep(1, length(riskDefinitive))
for (j in round(0.25*length(riskDefinitive)):round(0.75*
   length(riskDefinitive))) {
  group.assignation.vector[riskDefinitive<order.
     riskDefinitive[j]]<-1
  group.assignation.vector[riskDefinitive>=order.
     riskDefinitive[j]]<-2
  log.rank.groups.surv <- survdiff( Surv( time, status) ~
     group.assignation.vector, data= mSurv[match(names(
     riskDefinitive), rownames(mSurv)),] )
  p.val[j] <- 1 - pchisq(log.rank.groups.surv$chisq, length(
     log.rank.groups.surv$n) - 1)
}
p.val30a70<-rep(1, length(riskDefinitive))
p.val30a70[round(0.3*length(riskDefinitive)):round(0.7*
   length(riskDefinitive))]<-p.val[round(0.3*length(
   riskDefinitive)):round(0.7*length(riskDefinitive))]
pvalue.ordered.indexes <- order(p.val30a70)
lowest.p.value.index <- pvalue.ordered.indexes[1]
cutPoint<-order.riskDefinitive[lowest.p.value.index]

# group association
definitiveGroups<-rep(0,length(riskDefinitive))
definitiveGroups[riskDefinitive<as.numeric(cutPoint)]<-1
definitiveGroups[riskDefinitive>=as.numeric(cutPoint)]<-2
# computing beta mean
betas<-NULL
for(i in 1:dim(mExpr)[1]){
  betas[i]<-sum(betasMatrix[i,])/10
}
# validation risk score
for(i in 1:length(riskValidMatrix[,1])){
  riskValidDefinitive[i]<-mean(riskValidMatrix[i,])
}
# risk score derivated plots:
# lines of ordered p.values, it shows how well distinguished
    is the central part of the risk curve
################################### PLOT
pdf("riskANDpval.pdf")
###################################
plot(p.val, type="l", col="red", ylim=c(0,  summary(p.val30a
   70[round(0.3*length(riskDefinitive)):round(0.7*length(
   riskDefinitive))])[3]), xlab="Patients ordered by Risk\n
   (green line: 2 groups cutpoint)", ylab="p.value")
abline(h=(0.01-min(p.val))*0.1, col=2)
minp.val<-min(which(p.val<min(p.val)+(0.5-min(p.val))*0.5))
```

```
   # dinamic p.value computing
maxp.val<-max(which(p.val<min(p.val)+(0.5-min(p.val))*0.5))
abline(v=minp.val,col=2)
abline(v=maxp.val,col=2)
abline(v=table(definitiveGroups)[1],col=3)

# risk plot cutpoints
# 1) central redline
# 2) left(p.acum>1) blue
# 3) right(p.acum>1) red
colores<-rep(0,length(p.val))
colores[1:minp.val]<-4
colores[maxp.val:length(p.val)]<-2
colores[minp.val:maxp.val]<-8
cutPointLow<-as.numeric(order.riskDefinitive)[min(which(
   colores==8))]
cutPointHigh<-as.numeric(order.riskDefinitive)[min(which(
   colores==2))]

plot(rescale(as.numeric(order.riskDefinitive), to = c(0, 100
   )), col=colores, xlab="Patients ordered by Risk\n (green
   line: 2 groups cutpoint)", ylab="Risk Score")
abline(v=minp.val,col=2)
abline(v=maxp.val,col=2)
abline(v=table(definitiveGroups)[1],col=3)
legend("bottomright", title="Risk Groups", c(as.character(
   sum(colores==4)),as.character(sum(colores==8)),as.
   character(sum(colores==2))), fill=c(4,8,2), horiz=FALSE)

colores<-rep(1.5,length(p.val))
colores[riskDefinitive<cutPointLow]<-1
colores[riskDefinitive>=cutPointHigh]<-2
groups3train<-colores
################################### validation samples
colores<-rep(8,length(riskValidDefinitive))
order.riskValidDefinitive<-riskValidDefinitive[order(
   riskValidDefinitive)]
minp.valVal<-min(which(order.riskValidDefinitive>order.
   riskDefinitive[minp.val]))
maxp.valVal<-min(which(order.riskValidDefinitive>order.
   riskDefinitive[maxp.val]))
colores[1:minp.valVal]<-4
colores[maxp.valVal:length(riskValidDefinitive)]<-2
plot(rescale(as.numeric(riskValidDefinitive), to = c(0, 100)
   )[order(rescale(as.numeric(riskValidDefinitive), to = c(0
   , 100)))], col=colores, xlab="Patients ordered by Risk\n
   (green line: 2 groups cutpoint)", ylab="Risk Score")
abline(v=minp.valVal,col=2)
abline(v=maxp.valVal,col=2)
```

```
  legend("bottomright", title="Risk Groups", c(as.character(
     sum(colores==4)),as.character(sum(colores==8)),as.
     character(sum(colores==2))), fill=c(4,8,2), horiz=FALSE)
  colores<-rep(1.5,length(riskValidDefinitive))
  colores[riskValidDefinitive<cutPointLow]<-1
  colores[riskValidDefinitive>=cutPointHigh]<-2
  groups3valid<-colores
  groupsValid<-rep(0,length(riskValidDefinitive))
  groupsValid[riskValidDefinitive>=cutPoint]<-2
  groupsValid[riskValidDefinitive<cutPoint]<-1
  abline(v=table(groupsValid)[1],col=3)
  #  END PLOT
  dev.off()
  #
  hazardR<-hazard.ratio(x=definitiveGroups, surv.time=time[
     match(colnames(mExpr),names(time))], surv.event=status[
     match(colnames(mExpr),names(status))])
  funcion_km_groups_plot_risk(mExpr,mSurv[match(colnames(mExpr
     ),rownames(mSurv)),], "RiskTrain", definitiveGroups,
     hazardR$hazard.ratio,hazardR)


  hazardR<-hazard.ratio(x=groupsValid, surv.time=time[match(
     colnames(mExprValid),names(time))], surv.event=status[
     match(colnames(mExprValid),names(status))])
  funcion_km_groups_plot_risk(mExprValid,mSurvValid[match(
     colnames(mExprValid),rownames(mSurvValid)),], "
     RiskValidation", groupsValid, hazardR$hazard.ratio,
     hazardR)

  rList<-list("cut"=c(cutPoint,minp.val,maxp.val,minp.valVal,
     maxp.valVal),"risk"=rescale(as.numeric(riskDefinitive),
     to = c(0, 100)),"betas"=betas,"riskValid"=rescale(as.
     numeric(riskValidDefinitive), to = c(0, 100)),"
     groupsTrain"=definitiveGroups,"groupsValid"=groupsValid,"
     groupsTrain3"=groups3train,"groupsValid3"=groups3valid)
  return(rList)
}
```

function_sam

Description

Function to pre-filter the dataset, obtaining the most relevant genes for a group for microarray or RNAseq provided as a frequency table, this will be used later for further analysis with glmnet function.

Usage

```
function_sam( mExpr, groups_vector, sam_lambda )
```

Arguments

mExpr matrix containing by rows the gene expression (normalised) for each candidate, including the sample (columns) names in colnames(mExpr) and gene names in rownames(mExpr) .
groups_vector vector containing the group etiquette from the phenodata that we want to use, it must be provided as a 0 and 1 vector.
sam_lambda the lambda value to be used, it depends on the FDR computed by sam, 4 by default, but can be checked with a simple iteration in order to modify and obtain more or less robust and relevant genes.

Details

This function makes by differential expression a list that filter genes separated by the provided grouping. The ideal way to evaluate the best lambda is to check the output and look if the FDR is near the desired value.

Value

A table is provided with the incidence of the data in order to filter the most robust genes related with the phenodata grouping.

Examples

```
library("siggenes")
# this function calculates gene by gene the value of
   differential expression done with the previous defined
   groups, in this case, ER+ and ER-
# reevaluated lambda should be checked
# otherwise modify the lambda and train again
sam_sER<-function_sam(mExpr380, mSurv380$er, 4)
```

Function

```
function_sam <- function( mExpr, groups_vector, sam_lambda )
{
  library("siggenes")
  # error control
  if(dim(mExpr)[2]!=length(groups_vector)){message("ERROR: the
      matrix samples and group vector must have the same
     length")}
  ## bootstrap 100 samples
  n.genes <- dim( mExpr )[1]
  n.samples <- dim(mExpr)[2]
  iter<-100

  sampl<-NULL
  list.genes<-NULL

  for(i in 1:iter){
    m<-sample( 1:n.samples, size=round(n.samples), replace=
       TRUE)#sample generation
    sampl[[i]] <- m
  }
  print("samples ok")
  print(Sys.time())
  lista<-NULL
  #
  for(i in 1:iter){
    if(i%%10==0){(print(i))
      print(Sys.time())}
    #checking iterations
    #list of 500 vectors with relevant names
    #using a restrictive lambda
    mExpr2<-mExpr[,sampl[[i]]]
    groups_vector2<-groups_vector[sampl[[i]]]
    samR<-sam(mExpr2, groups_vector2, method=d.stat, var.equal
       =FALSE)
    #extracting best genes by p.val
    list.genes<-c(list.genes, list.siggenes(samR, sam_lambda))
    #incidence as number of times it shows as significative
       value, is what it's retourned as a table
  }

  resultado<-table(list.genes)
}
```

funcion_PredGlmnet

Description

Function to predict phenodata and etiquettes for microarray or RNAseq, it is designed to be used provided a list of most relevant genes obtained when we filter the table from function_sam.

Usage

```
funcion_PredGlmnet(mExpr, vectorGroups, vectorSampleID )
```

Arguments

mExpr: matrix containing by rows the gene expression (normalised) for each candidate, including the sample (columns) names in colnames(mExpr) and gene names in rownames(mExpr) filtered by rows with function_sam output.
vectorGroups: vector containing the group etiquette from the phenodata that we want to use, it must be provided as a 0 and 1 vector.
vectorSampleID: vector containing the sample identifiers as colnames(mExpr).

Details

This function compute predictions measuring the predictive power of each gene, in order to discover how each gene correlates and is able to predict a pheno-etiquette provided to make the groups.

Value

A list is provided with information about each gene, the bootstraped iterations allows to provide a beta matrix to measure the eficiency of each gene, and also the AUC roc to measure the prediction power the beta values must be observed as: the closest to 1 in absolute value, the better.

Examples

```
library(glmnet)
library(ROCR)

# output from list of genes used to create the new matrix, in
   our example the list of 34 markers
Pred_ER<-funcion_PredGlmnet(t(mExpr34), mSurv380$er, as.
   character(colnames(mExpr34)))
```

Function

```
funcion_PredGlmnet <- function( mExpr, vectorGroups,
   vectorSampleID )
{
  # expression matrix t, with pre-selected genes
  # error control
  if(dim(mExpr)[2]!=length(vectorGroups)){message("ERROR: the
     matrix samples and group vector must have the same length
     ")}
  if(dim(mExpr)[2]!=length(vectorSampleID)){message("ERROR:
     the matrix samples and vectorSampleID vector must have
     the same length")}

  library(glmnet)
  library(ROCR)

  iterations<-100
  n.genes <- dim( mExpr )[2]
  n.samples <- dim(mExpr)[1]
  n.train<-round(n.samples*2/3)
  # 2/3 sample as training set
  sampl<-NULL
  # randomised samples
  for(i in 1:iterations){
    m<-sample( 1:n.samples, size=n.samples, replace=FALSE)
    sampl[[i]] <- m
  }
  print("samples ok")
  print(Sys.time())

  list<-NULL

  for(i in 1:iterations){
    if(i%%10==0){(print(i))
      print(Sys.time())}
    # printing iterations

    # for each time a sample is taken
    mExpr_i<-mExpr[sampl[[i]],]
    vectorGroups_i<-vectorGroups[sampl[[i]] ]
    vectorSampleID_i<-vectorSampleID[sampl[[i]] ]

    # calling predictor: training, 1:n.train with 2/3
    object_cv_glmnet_train<-cv.glmnet(x=mExpr_i[1:n.train,], y
       =vectorGroups_i[1:n.train],nfolds=5, type.measure="auc"
       , alpha=0.75, family="binomial")
    # calling predictor: predicting, (n.train+1):n.samples the
```

```
        restant 1/3
    object_cv_glmnet_coeff<-predict(object=object_cv_glmnet_
       train, newmat=mExpr_i[(n.train+1):n.samples,],type="
       coeff", s=object_cv_glmnet_train$lambda.1se)
    # AUC measuring predictive power
    object_cv_glmnet_response<-predict(object=object_cv_glmnet
       _train, newx=mExpr_i[(n.train+1):n.samples,],type="
       response", s=object_cv_glmnet_train$lambda.1se)

    x<-NULL
    # cumulative matrix of beta values picturing each probeset
        predictive power. AUC value is also stored
    x$coeff<-object_cv_glmnet_coeff[object_cv_glmnet_coeff[,1]
       !=0,]
    x$coeff<-x$coeff[2:length(x$coeff)]

    auc_prediction<-prediction(as.double(object_cv_glmnet_
       response[,1]), vectorGroups[pmatch(rownames(object_cv_
       glmnet_response), as.character(vectorSampleID))])

    x$auc<-as.numeric((performance(auc_prediction,"auc"))@y.
       values)
    list[[i]]<-x
  }
  list
}
```

function_km_groups_plot

Description

Function to plot predefined group as KM curves for microarray or RNAseq, providing gene expression and group.

Usage

```
funcion_km_groups_plot( genExpr, time.import, status.import ,
   genName, group.assignation.vector, hazardR)
```

Arguments

genExpr: expression vector.
time.import: vector containing the survival time in years.
status.import: vector containing the survival status in 0 and 1.
genName: string containing the gene name.
group.assignation.vector: vector containing the 0 or 1 group for each sample.
hazardR: a hazard.ratio object computed using the package from library survcomp.

Details

This function compute KM curves as the ones used in our functions using a provided vector with phenodata group.

Value

A Kaplan-Meier pdf plotting the groups as survival curves.

Examples

```
library("survcomp")
# computing hazard ratio with provided groups
hazardR<-hazard.ratio(x=(mSurv380$er), surv.time=(mSurv380$
   time), surv.event=(mSurv380$censurado))
# provide with GROUPS AS 0 AND 1
funcion_km_groups_plot(mExpr32ER[1,], mSurv380$time, mSurv380$
   censurado, "ENSG00000074410", mSurv380$er, hazardR
```

Function

```
funcion_km_groups_plot <- function( genExpr, time.import,
   status.import , genName, group.assignation.vector, hazardR)
```

```
{

  if(length(genExpr)!=length(time.import)){message("ERROR: the
      expr values and time vector must have the same length")}
  if(length(genExpr)!=length(status.import)){message("ERROR:
     the expr values and status vector must have the same
     length")}
  # gene name in vectors time status
  if(!identical(names(genExpr), names(time.import))){message("
     ERROR: expr values names must match time vector names")}
  if(!identical(names(genExpr), names(status.import))){message
     ("ERROR: expr values names must match status vector names
     ")}

  mSurv<-cbind(time.import, status.import)
  colnames(mSurv)<-c("time","status")
  mSurv<-as.data.frame(mSurv)
  rownames(mSurv)<-names(time.import)
  mSurv$status[mSurv$time>10]<-0
  mSurv$time[mSurv$time>10]<-10.1
  n.genes <- 1
  n.samples <- length(genExpr)
  probesets.names <- names( genExpr)
  # grant at least 25 % range to avoid relative minima
  for15<-round(n.samples*0.15)
  for85<-round(n.samples*0.85)
  # future pvals
  p.value.genes.ordering <- rep(0, n.genes)

  for( j in genName )
  {
    # for each row # OUTDATED
    vector.exprs <- as.numeric(genExpr)
    # ordered expresion
    order.vector.exprs <- order( vector.exprs )
    # making groups
    # selecting minimum pval (optimising separability between
       curves)

    fits1 <- survfit( Surv( time, status) ~ group.assignation.
       vector, data= mSurv )

    log.rank.groups.surv <- survdiff( Surv( time, status) ~
       group.assignation.vector, data= mSurv )
    p.val <- 1 - pchisq(log.rank.groups.surv$chisq, length(log
       .rank.groups.surv$n) - 1)

    jj<-genName
```

```
    pdf( paste(jj,".pdf"))
    kmplot( fits1 ,xaxis.at=c(0:10), col.surv=c("brown","#4
       DAF4A"),  group.names=c('ALL','CMS4'), xlab=paste('
       Survival(years) and p.value: ',format(p.val,3) ) , ylab
       =paste('Survival Probability','\n Hazard Ratio:',format
       (hazardR$hazard.ratio,3),' (',format(hazardR$lower,3),'
       , ',format(hazardR$upper,3), ')'), legend=TRUE, loc.
       legend='bottomleft', lty.ci=2, lwd.ci=1, col.ci=c("
       brown","#4DAF4A"), xlim=c(0,10)  )

    # expression and t test
    Ttest<-wilcox.test(vector.exprs[group.assignation.vector==
       0],vector.exprs[group.assignation.vector==1])
    boxplot(vector.exprs[group.assignation.vector==0],vector.
       exprs[group.assignation.vector==1], col=c(4,2), ylab="
       expression level", xlab="wilcox-test p.value for groups
       :", sub=format( Ttest$p.value ))
    legend("bottomright", title=jj, c(as.character(sum(group.
       assignation.vector==0)),as.character(sum(group.
       assignation.vector==1))), fill=c(4,2), horiz=TRUE)
    dev.off()


  }
  print(p.val)

}
```

function_kmplot

Description

Sub-function to be called in order to compute plot and options for all other KM plot function.

Usage

```
kmplot( fits1 ,xaxis.at=c(0:10), col.surv=c("brown","#4DAF4A")
   ,  group.names=c('ALL','CMS4'), xlab=paste('Survival(years)
    and p.value: ',format(p.val,3) ) , ylab=paste('Survival
   Probability','\n Hazard Ratio:',format(hazardR$hazard.ratio
   ,3),' (',format(hazardR$lower,3),', ',format(hazardR$upper,
   3), ')'), legend=TRUE, loc.legend='bottomleft', lty.ci=2,
   lwd.ci=1, col.ci=c("brown","#4DAF4A"), xlim=c(0,10)  )
```

Arguments

fits1: expression vector.
group.names: vector containing the survival time in years.
...: fixed parameters.

Details

This function is a sub-function that computes the KM curve plot object in order to be plotted by others KM plot function.

Function

```
kmplot <- function(km, mark=3, simple=FALSE,
                   xaxis.at=pretty(km$time), xaxis.lab=xaxis.
                      at,
                   lty.surv=1, lwd.surv=1, col.surv=1,
                   lty.ci=0, lwd.ci=1, col.ci=col.surv, #By
                      default (lty.ci=0), confidence intervals
                       are not plotted.
                   group.names=NULL, group.order=seq(length(km
                      $n)), extra.left.margin=4,
                   label.n.at.risk=FALSE, draw.lines=TRUE, cex
                      .axis=1,
                   xlab='', ylab='', main='', xlim=c(0,max(km$
                      time)), ylim=c(0,1),
                   grid=TRUE, lty.grid=1, lwd.grid=1, col.grid
                      =grey(.9),
```

```
                    legend=!is.null(km$strata), loc.legend='
                       topright', add=FALSE,
                    ... # ... is passed to par()
){
  # Version 2.5.5: 2014/5/19

  # km is the output from survfit() function in survival
     package.##############################

  # xaxis.at specifies where 'n at risk' will be computed and
     printed.
  # xaxis.lab specifies what will be printed at xaixs.at.  (
     see example)

  # If group names are long, add extra left margin by setting
     extra.left.margin to something greater than 0.

  # line specifications (lty.surv, lwd.surv, col.surv) will be
      recycled.
  # Set lty.ci to 1 if confidence intervals are needed.
  # group.names will overwrite whatever is specified in
     survfit() output.
  # group.order specifies the order of groups from top in 'n
     at risk'.  1 is at top, 2 next, and so on.

  # if add=TRUE, then par() is not refreshed.  allows multiple
      panels by
  # using, e.g., par(mfrow=c(2,2)).

  # op <- par(no.readonly = TRUE)



  ng0 <- length( km$strata ) ; ng <- max(ng0,1)
  # When only one group...
  if(ng0==0){ km$strata <- length(km$time) ; names(km$strata)
     <- 'All' ; legend <- draw.lines <- FALSE }

  lty.surv <- rep(lty.surv, ng) ; lwd.surv <- rep(lwd.surv, ng
     ) ; col.surv <- rep(col.surv, ng)
  lty.ci <- rep(lty.ci, ng) ;      lwd.ci <- rep(lwd.ci, ng)
         ; col.ci <- rep(col.ci, ng)

  ## group names and error checking
  gr <- c(km$strata)
  if( is.null(group.names) ){ group.names <- names(km$strata)
     }
  if( length(unique(group.names)) != ng ){ stop('\n','length(
     unique(group.names)) != number of groups.') }
```

```
if( suppressWarnings(any( sort(group.order) != 1:ng)) )
{ stop('\n', 'Something wrong with group.order.','\n','sort(
   group.order) must equal 1:', ng, '.') }
group.names <- gsub(' *$', '', group.names)  #to remove
   unwanted white spaces in group.names.
if(ng==1 & (group.names[1]=='group.names') ){ group.names <-
    'N at risk' ; label.n.at.risk = FALSE }

## graphic parameters
if(!add){
  par(list(oma=c(1,1,1,1), mar=c(4+ng,4+extra.left.margin,4,
     2)+.1))
  if(simple) par( mar=c(3,4,2,1)+.1 )
  par( list(...) )
}

## reformat survival estimates
dat <- data.frame(time=km$time, n.risk=km$n.risk, n.event=km
   $n.event, survival=km$surv, std.err=km$std.err,
                  lower=km$lower, upper=km$upper, group=rep(
                     group.names, gr) )
dat.list <- split(dat, f=dat$group)

## plot (but not survival curves)
plot(0,type='n', xlim=xlim, ylim=ylim, xaxt='n', yaxt='n',
   xlab='', ylab='' )
if(grid){
  par('xpd'=FALSE)
  abline(v=xaxis.at,       lty=lty.grid, lwd=lwd.grid, col=
     col.grid )
  abline(h=pretty(c(0,1)), lty=lty.grid, lwd=lwd.grid, col=
     col.grid )
}
axis( side=2, at=pretty(c(0,1)), cex.axis=cex.axis )
axis( side=1, at=xaxis.at, label=xaxis.lab, line=-0.5, tick=
   FALSE, cex.axis=cex.axis )
axis( side=1, at=xaxis.at, label=rep('',length(xaxis.at)),
   line=0, tick=TRUE )
title(xlab=xlab, line=1.5, adj=.5, ...) ; title(ylab=ylab,
   ... )

if(!simple){
  ## write group names
  group.name.pos <- (par()$usr[2]-par()$usr[1]) / -8 ;
     padding <- abs( group.name.pos / 8 )
  line.pos <- (1:ng)[order(group.order)] + 2
  mtext( group.names, side=1, line=line.pos, at=group.name.
     pos, adj=1, col=1, las=1, cex=cex.axis )
```

```
    ## draw matching lines for n at risk.
    if(draw.lines){
      par('xpd'=TRUE)
      for(i in 1:ng){
        axis(side=1, at=c(group.name.pos+padding,0-2*padding),
            labels=FALSE, line=line.pos[i]+0.6, lwd.ticks=0,
             col=col.surv[i], lty=lty.surv[i], lwd=lwd.surv[i]
                ) }
    }

    ## numbers at risk
    kms <- summary(km, times=xaxis.at) ; if(is.null(kms$strata
       )) kms$strata <- rep(1,length(kms$time) )
    d1 <- data.frame(time = kms$time, n.risk = kms$n.risk,
       strata = c(kms$strata))
    d2 <- split(d1, f=d1$strata)

    ## Right-justifying the numbers
    ndigits <- lapply(d2, function(x) nchar(x[,2]) )
    max.len <- max( sapply(ndigits, length) )
    L <- do.call('rbind', lapply(ndigits, function(z){ length(
       z) <- max.len ; z} ) )
    nd <- apply( L, 2, max, na.rm=T )
    for( i in seq(ng) ){
      this <- d2[[i]]
      w.adj <- strwidth('0', cex=cex.axis, font=par('font')) /
          2 * nd[1:nrow(this)]
      mtext( side=1, at=this$time+w.adj, text=this$n.risk,
         line=line.pos[i], cex=cex.axis, adj=1, col=1, las=1)
    }
    if(label.n.at.risk) mtext( side=1, text='N at risk', at=
       group.name.pos, line=1.5, adj=1, col=1, las=1, cex=cex.
       axis )
  } ## End of if(!simple)

  # Legend
  rlp <- group.order
  if(legend){
    bgc <- ifelse( par('bg')=='transparent', 'white', par('bg'
       ) )
    legend(x=loc.legend, legend=group.names[rlp], col=col.surv
       [rlp], lty=lty.surv[rlp], lwd=lwd.surv[rlp],
           bty='o', cex=cex.axis, bg=bgc, box.col='transparent
              ', inset=.01 )
  }

  ## draw confidence intervals
  for(i in 1:ng){
    this <- dat.list[[i]]
```

```
    x <- this$time ; L <- this$lower ; U <- this$upper ; S <-
       this$survival
    naL <- which( is.na(L) ) ; L[naL] <- L[naL-1] ; U[naL] <-
       U[naL-1]
    lines( x, L, type='s', col=col.ci[i], lty=lty.ci[i], lwd=
       lwd.ci[i] )
    lines( x, U, type='s', col=col.ci[i], lty=lty.ci[i], lwd=
       lwd.ci[i] )
  }
  # draw curves
  lines(km, conf.int=FALSE, col=col.surv, lty=lty.surv, lwd=
     lwd.surv, mark=mark, xmax=xlim[2], ymin=ylim[1])

  box(bty=par('bty'))

  # par(op)
}
```

Appendix: Resumen en castellano

Introducción

El desarrollo de tecnologías ómicas robustas (genómica, transcriptómica, proteómica, etc.) para generar y comprender las alteraciones del genoma tiene cada vez mayor impacto en la atención médica, con particular relevancia en el cáncer y la oncología. En el contexto actual de Medicina Personalizada, Medicina de Precisión y Medicina Genómica (Roden and Tyndale, 2013), la investigación moderna del cáncer debe realizarse considerando el uso adecuado de datos a gran escala, derivados de estas nuevas tecnologías. Algunas de estas tecnologías, como el perfil de expresión transcriptómico, ya se han aplicado a miles de muestras humanas (consultar la base de datos pública GEO (NCBI, 2019)), y proporcionan información sobre el nivel de expresión de todos los genes conocidos para cada individuo. Para ser útil y aplicable a la investigación médica, estos datos ómicos deben integrarse con los datos clínicos correspondientes utilizando herramientas y métodos computacionales y bioinformáticos. Lo anteriormente descrito compone el marco principal de esta tesis doctoral.

Incidencia del cáncer en Europa

Respecto al estudio en el área del cáncer, el trabajo en la presente tesis se basa en dos tipos de cáncer: el cáncer de mama (BRCA) y el cáncer colorrectal (CRC). Estos tipos de cáncer son, hoy en día, los más frecuentes en Europa, representando la mayor proporción de todos los tipos de cáncer (ecis.jrc.ec.europa.eu/). En particular, en Europa en 2018, los tipos de cáncer más comunes fueron los del seno femenino (523,000 casos), seguidos del cáncer colorrectal (500,000), el de pulmón (470,000) y el de próstata (450,000). Las cifras globales para Europa en 2018 se estimaron en 3,91 millones de nuevos casos de cáncer y 1,93 millones de muertes por cáncer (Ferlay et al., 2018). Dado que la población europea está cerca de 513 millones, teniendo cada año alrededor de 4 millones de nuevos casos y alrededor de 2 millones de muertes, el cáncer representa la segunda causa más importante de muerte en Europa. Teniendo en cuenta los números específicos para el cáncer de mama y el

cáncer colorrectal, la incidencia de estos tipos de cáncer en Europa se presenta en la Figura 6.1, que muestra el mapa coloreado según las incidencias por país.

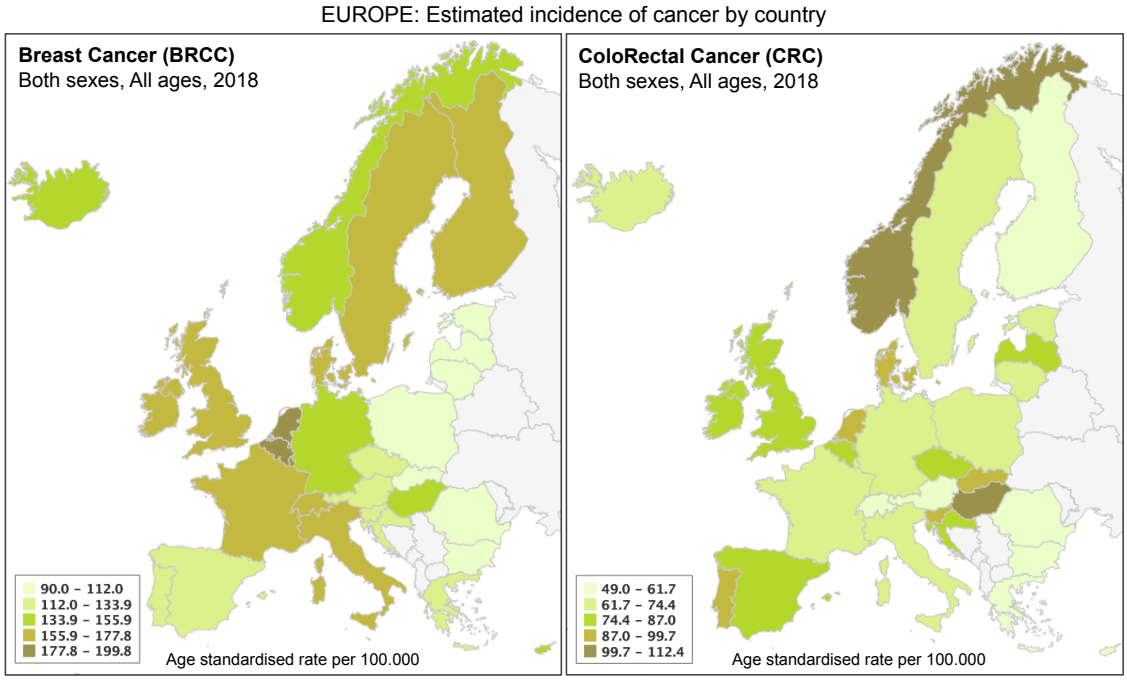


Figure 6.1: Incidencia estimada del cáncer de mama y del cáncer colorrectal en Europa por país (fuente: Sistema Europeo de Información sobre el Cáncer,European Cancer Information System, ecis.jrc.ec.europa.eu).

Cáncer: una enfermedad genómica provocada por mutaciones

Como se indicó anteriormente, la investigación actual sobre el cáncer está muy afectada por el valor y el poder de las tecnologías ómicas aplicadas al avance de la oncología médica y clínica. La aplicación de tecnologías ómicas en todo el genoma para el estudio del cáncer en los últimos 20 años ha generado una nueva comprensión de esta compleja enfermedad que ya no puede llamarse "enfermedad genética", ya que es más propiamente una "enfermedad genómica". De hecho, en las últimas dos décadas, los esfuerzos de secuenciación integral han revelado los datos genómicos de formas comunes de cáncer humano (Kankava et al., 2019). Estos estudios han revelado alrededor de 140 genes humanos que, cuando son alterados por mutaciones intragénicas, pueden promover o "conducir" la génesis del tumor y la malignidad celular. Un tumor típico contiene de dos a diez de estas mutaciones de algún "gen conductor"; Las mutaciones restantes no confieren una ventaja de crecimiento selectivo. En tumores sólidos comunes (como los derivados del colon, la mama, el pulmón, el cerebro o el páncreas), un promedio de 25 a 75 genes muestran mutaciones somáticas sutiles que podrían alterar sus productos proteicos (Gerlinger et al., 2012).

La Figura 6.2 presenta la complejidad del cáncer, que puede afectar a muchos tipos diferentes de células, tipos de tejidos y órganos en el cuerpo humano, desde niños hasta adultos. Esta complejidad se debe no solo a todos los diferentes tipos de cáncer originados en diferentes ubicaciones del cuerpo, sino también a la gran cantidad de mutaciones somáticas que se han encontrado gracias a los análisis a

escala genómica de todos los genes humanos en muchos Miles de muestras de tumores (Gerlinger et al., 2012).

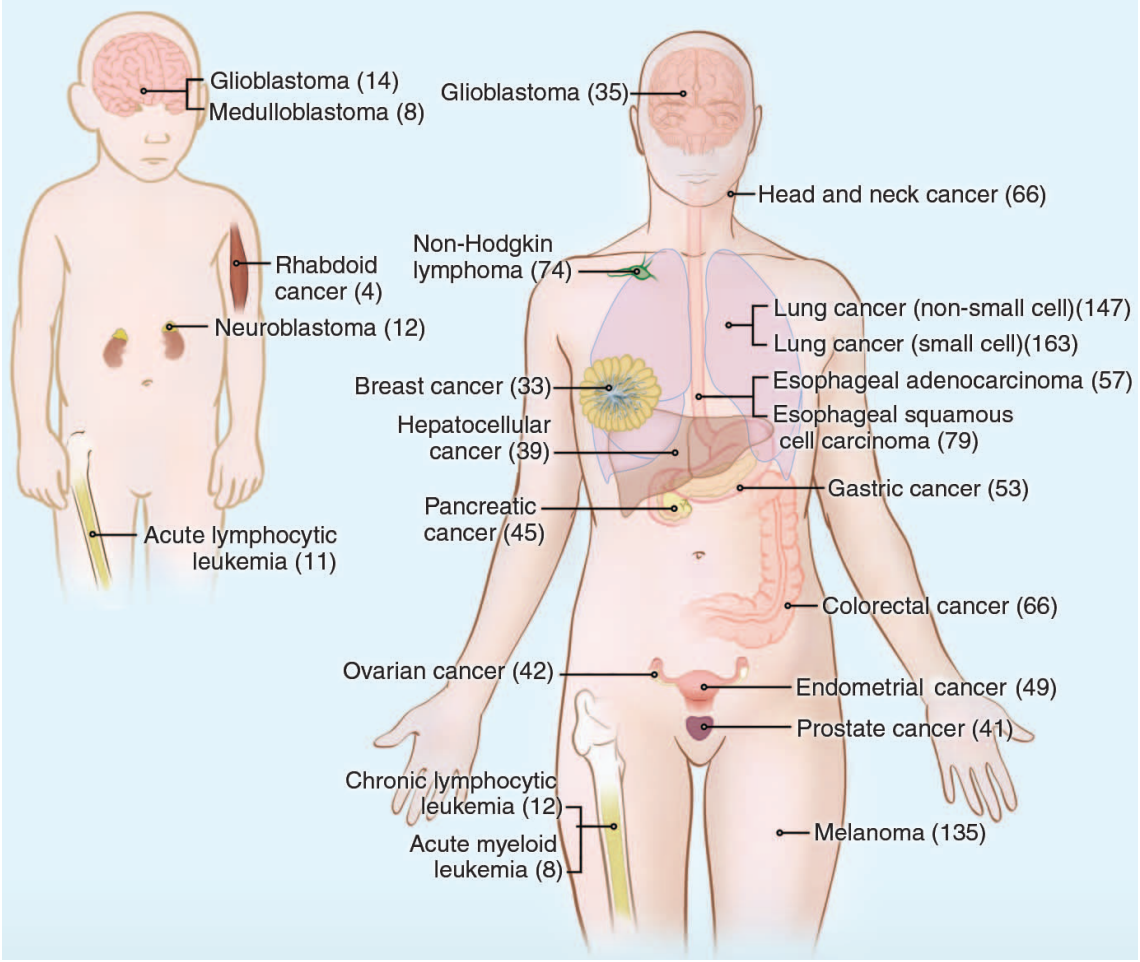


Figure 6.2: Número promedio de mutaciones somáticas en una colección representativa de casos humanos de cáncer, detectados por estudios de secuenciación de ADN en todo el genoma. Se han analizado los genomas de un grupo diverso de cáncer en adultos (derecha) y pediátricos (izquierda). Los números entre paréntesis indican el número medio de mutaciones no repetidas por tumor. (fuente: Vogelstein et al. 2013 *Science*, www.sciencemag.org).

La heterogeneidad del cáncer: un desafío para la era genómica

El cáncer es una enfermedad heterogénea con características genéticas y fenotípicas únicas que difieren entre pacientes e incluso entre regiones tumorales (Martinez-Outschoorn et al., 2017). La observación de la heterogeneidad individual en el cáncer se ha descrito muchas veces, pero se logró un gran avance cuando se comprobó la heterogeneidad intratumoral y el crecimiento de tumores evolutivos ramificados utilizando la tecnología ómica de la secuenciación de ADN espacial de todo el exoma (Gerlinger et al., 2012). Esta "heterogeneidad intratumoral" agregó un nuevo nivel de complejidad al estudio del cáncer, que ya tenía una naturaleza compleja debido a los muchos posibles orígenes tisulares de los tumores (Figura 6.3). Además, el cáncer es una enfermedad muy dinámica en la que las células tumorales proliferan y evolucionan con el tiempo, incluso en el mismo lugar; por lo que debe agregarse "heterogeneidad temporal" a "heterogeneidad espacial" (Martinez-Outschoorn et al., 2017).

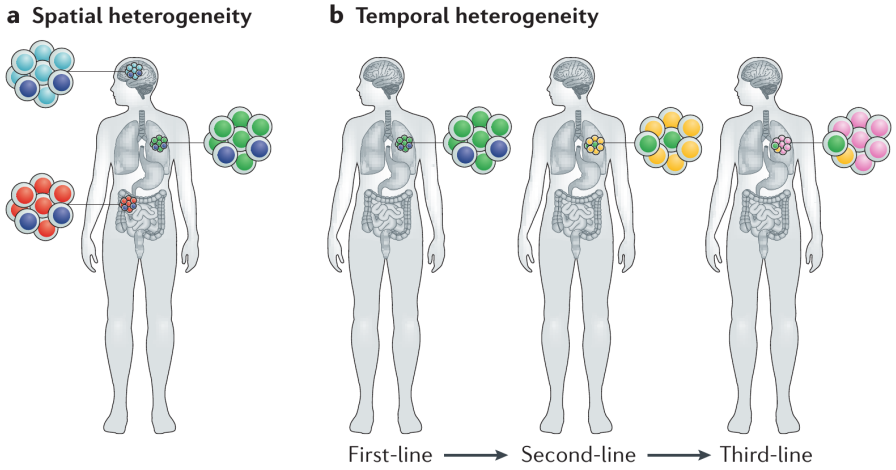


Figure 6.3: Heterogeneidad espacial **(a)**: distribución desigual de los subclones de cáncer en diferentes regiones del tumor primario o de sus metastásis. Heterogeneidad temporal **(b)**: variaciones en la composición molecular de una lesión única a lo largo del tiempo, ya sea como resultado de la progresión natural del tumor o como resultado de la exposición a presiones selectivas creadas por intervenciones clínicas. (fuente: Dagogo-Jack et al. 2018 *Nat Rev Clin Oncol*, www.nature.com/nrclinonc).

Los "pathways" del conductor oncogénico: un cambio de paradigma

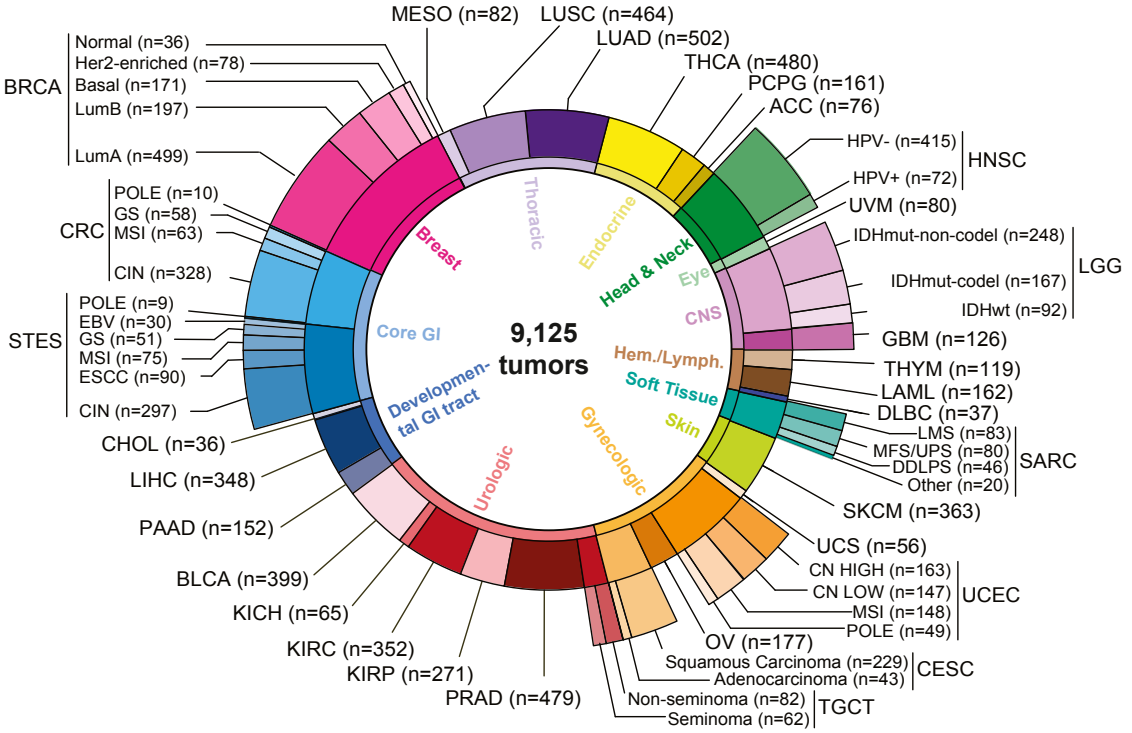


Figure 6.4: Distribución de los tipos de cáncer en TCGA, incluidos los subtipos moleculares analizados. TCGA pancancer atlas contiene 9,125 muestras de tumores. (fuente: Sánchez-Vega et al. 2018 textit Cell, www.cell.com).

Los resultados del esfuerzo mundial realizado por la aplicación de tecnologías ómicas de todo el genoma al estudio de muchos tipos y subtipos de cáncer (por ejemplo, en el Proyecto del Atlas del Genoma del Cáncer, TCGA) (Figura 6.4) proporcionan una nueva comprensión molecular más profunda de la biología del cáncer, lo que sugiere que la naturaleza de los tumores se explica mejor cuando se asocian con firmas de genes moleculares específicos y con vías biológicas específicas, en lugar de la asociación clásica con la "célula de origen". (Hoadley et al., 2018).

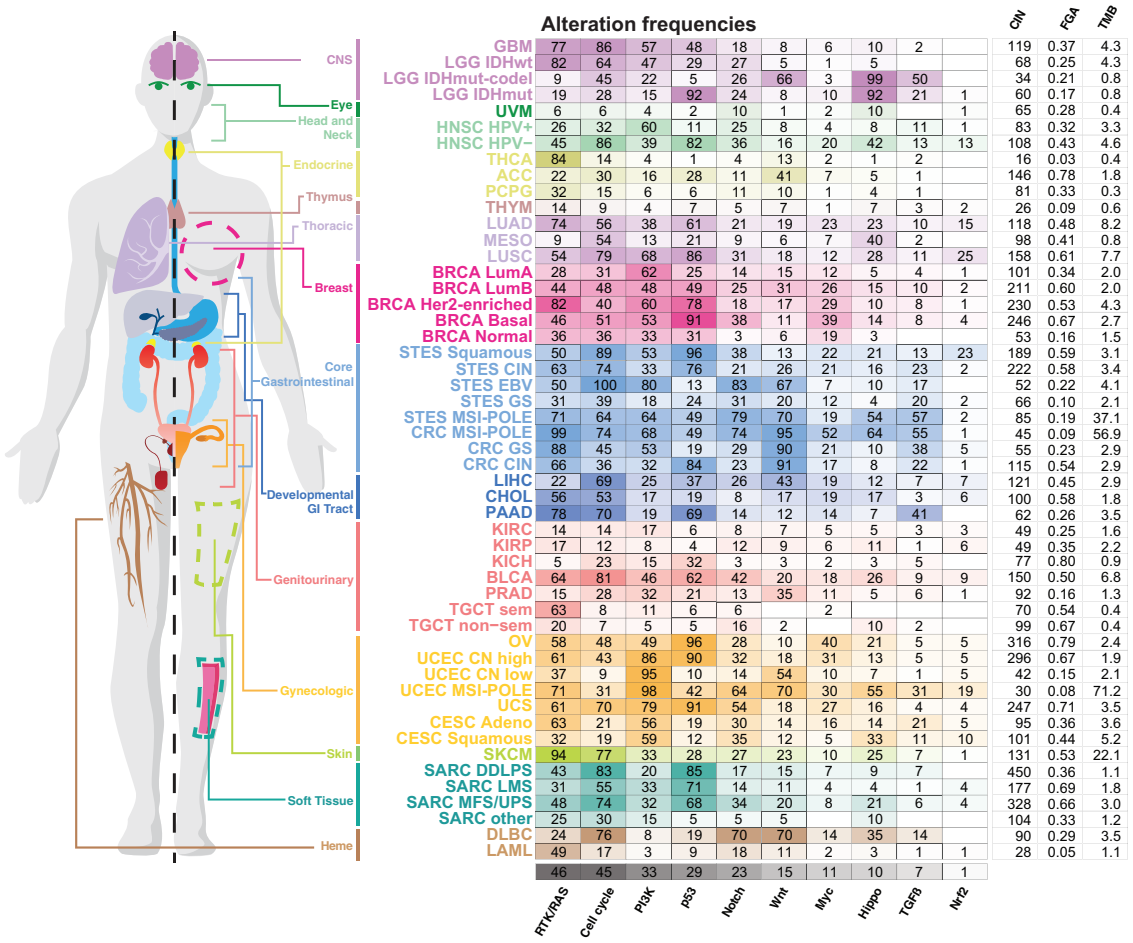


Figure 6.5: Fracción de muestras alteradas de la cohorte TCGA por vía y subtipo de tumor. Las vías se ordenan disminuyendo la frecuencia media de las alteraciones. Las intensidades de color crecientes reflejan porcentajes más altos. También se proporciona el recuento promedio de mutaciones, así como el número de segmentos no equilibrados y el genoma de la fracción alterado (dos medidas del grado de alteración del número de copias) por subtipo de cáncer. (fuente: Sánchez-Vega et al. 2018 textit Cell, www.cell.com).

Mediante el uso de mutaciones somáticas, alteraciones en el número de copias, cambios en la expresión del ARNm y modificaciones de la metilación del ADN detectadas en 9.125 tumores (perfilados por TCGA), Nik Schultz, Chris Sander y colaboradores (Sanchez-Vega et al., 2018) analizaron los mecanismos y patrones de las alteraciones somáticas, identificando diez pathways como los que tienen la mayor parte de esas alteraciones.

Hipótesis

El trabajo presentado en este doctorado se centra en el campo de la **Bioinformática y biología computacional** aplicado a la **investigación en cáncer**, con un enfoque particular en el análisis e integración de datos ómicos y datos clínicos para mejorar el descubrimiento y la identificación de nuevos biomarcadores moleculares relacionados con el pronóstico de los pacientes con cáncer.

En particular, nuestra principal hipótesis para comenzar y desarrollar nuestra investigación fue la siguiente: "**Consideramos que se debe realizar un análisis detallado de los datos de supervivencia de los pacientes con cáncer combinados con los datos transcriptómicos derivados de las biopsias de tumores de dichos pacientes. una forma muy poderosa y significativa de descubrir nuevos biomarcadores genéticos directamente relacionados con la naturaleza y el pronóstico del cáncer específico de cada paciente**".

Para probar y desarrollar esta hipótesis, trabajamos en este doctorado con muestras de los dos tipos principales de cáncer: **Cáncer de mama** (**BRCA**, es decir, carcinoma de mama invasivo) y **Cáncer coloRectal** (**CRC**). Estos tipos de cáncer son hoy en día los más frecuentes en Europa (https://ecis.jrc.ec.europa.eu/), representando en conjunto la mayor proporción de todos los tipos de cáncer.

Un primer desafío crítico para llevar a cabo este trabajo fue recopilar e integrar en conjuntos de datos uniformes una **gran cantidad de muestras de cáncer** (es decir, más de mil) que incluía **expresión de genoma completo y datos de supervivencia**. Decimos esto porque, la mayoría de los análisis de supervivencia que encontramos en la literatura se restringieron a conjuntos de datos más pequeños (es decir, a cientos de muestras, o incluso menos). Muchos investigadores en el campo no se dan cuenta de que el poder estadístico y la importancia de todos los algoritmos y métodos de supervivencia dependen, de manera crítica, del número de muestras estudiadas juntas.

Un punto final con respecto a la hipótesis es que no propusimos solo la aplicación de métodos computacionales estándar para el análisis de supervivencia, sino que deseamos desarrollar y aplicar nuevos algoritmos bioinformáticos para hacerlo. En particular, para mejorar la forma en que los datos de supervivencia se pueden integrar con los datos de expresión probados de todo el genoma para descubrir nuevos **marcadores de supervivencia**.

Objetivos

Una vez que hemos descrito la hipótesis principal de nuestra Tesis doctoral, debemos presentar los objetivos que describen de una manera más tangible el trabajo particular realizado y los desafíos específicos que enfrentamos durante los cinco años de nuestro doctorado. Los objetivos se dividen en dos grupos principales: **(i)** los primeros dos objetivos (1º y 2º) corresponden al trabajo realizado con **datos de cáncer de mama**; y **(ii)** el segundo grupo de objetivos (3º y 4º) corresponde al trabajo realizado con los datos de **ColoRectal Cancer**. Estos **cuatro objetivos** se presentan en esta disertación como **cuatro capítulos separados** después de este.

Los objetivos:

Objetivo 1.- Generación de un gran conjunto de datos homogéneos de muestras de cáncer de mama (BRCA) que incluyen datos de expresión de genoma completo y datos de supervivencia de pacientes; y el descubrimiento de marcadores de supervivencia BRCA asociados con los tres marcadores clínicos actualmente estándar (ER, PR y HER2) a través del desarrollo y la aplicación de algoritmos robustos para análisis de supervivencia basados en perfiles transcriptómicos.

Objetivo 2.- Desentrañar y descubrir marcadores genéticos positivos y reguladores del Cáncer de Mama Triple Negativo (TNBC) utilizando perfiles transcriptómicos y reguladores de genes combinados con análisis de supervivencia. Estudio realizado por comparación y contraste de TNBC con el subtipo más frecuente de cáncer de mama, el BRCA luminal.

Objetivo 3.- Generación de un gran conjunto de datos homogéneos de muestras de cáncer coloRectal (CRC) que incluyen datos de expresión de genoma completo y datos de supervivencia del paciente; y el descubrimiento de nuevos genes marcadores de supervivencia de CRC derivados de una integración robusta y un metanálisis de múltiples conjuntos de datos transcriptómicos.

Objetivo 4.- Análisis integrativo de los perfiles transcriptómicos de múltiples muestras de cáncer colorrectal (CRC) para identificar y caracterizar los cuatro subtipos moleculares de consenso (CMS1, 2, 3 y 4); la integración de estos datos transcriptómicos con los datos de supervivencia de los pacientes; y la caracterización relativa de una firma EMT asociada al KO de P21 (es decir, el gen CDKN1A KO).

Marcadores de supervivencia del cáncer de mama (BRCA) asociados a marcadores clínicos estándar y algoritmos robustos para el análisis de supervivencia basado en perfiles transcriptómicos

Motivación

El tratamiento del cáncer de mama está determinado por una clasificación estándar de tumores en cuatro grupos. La clasificación se lleva a cabo considerando principalmente tres marcadores clínicos: ER (ESR1), PR (PGR) y HER2 (Saini et al., 2011) (ERBB2 o NEU) obtenidos por inmunohistoquímica (IHC). Los marcadores definen las subclases; Luminal A, Luminal B, HER2 enriquecido y triple negativo (TNBC). Algunos marcadores complementarios, como AURKA o MKI67, se están considerando recientemente para mejorar la predicción del riesgo.

Sin embargo, los errores en la estimación de marcadores clínicos estándar son mayores que lo esperado (Li et al., 2010). Esto puede conducir a un tratamiento incorrecto del paciente. Además, los grupos obtenidos por solo tres marcadores suelen ser demasiado heterogéneos (Venet et al., 2011) (Bartlett et al., 2016) (Mertins et al., 2016). La identificación de genes relacionados con los marcadores clínicos puede ayudar a mejorar la estratificación y el tratamiento de los pacientes que proporcionan nuevos objetivos terapéuticos.

Se han desarrollado varias plataformas comerciales que consideran una firma genética multivariable. Sin embargo, la superposición entre las firmas genéticas y los grupos de riesgo es pequeña (Venet et al., 2011) (Bartlett et al., 2016) (Mertins et al., 2016). La plataforma funciona como una caja negra y las decisiones no se pueden interpretar en términos de marcadores clínicos estándar. Esto evita la aplicación en la práctica clínica. Hay una falta de investigaciones que se centren en la influencia del método de selección de características en el rendimiento y la estabilidad de la firma (Haury et al., 2011).

Además, se ha propuesto un gran número de firmas genéticas pronósticas en la literatura. El consenso entre ellos es bastante pequeño y con frecuencia son dependientes de la muestra (Mertins et al., 2016). Es decir, el algoritmo recupera un subconjunto diferente de genes con respecto al conjunto de datos considerado. Algunos autores (Ein-Dor et al., 2005), han estudiado varias firmas de genes y han llegado a la conclusión de que la estabilidad, la reproducibilidad y la solidez siguen siendo un problema difícil. Para superar esto, una validación en una serie independiente de RNAseq es otro objetivo.

En este capítulo se obtiene una firma genética multivariable robusta que se interpreta en términos de los marcadores clínicos. Esta firma genética proporciona una alternativa para desarrollar nuevos tratamientos.

Se desarrollarán varias estrategias sólidas nuevas en este capítulo, con el fin de mejorar cuatro de los análisis más comunes en bioinformática: normalización, expre-

sión diferencial, selección de características y modelos de predicción multivariante y univariante.

Materiales y métodos

GEO ID	Orig N	Final N	Surv Type	PMID	Year	Journal	Description
GSE6532	87	87	RFS, DMFS	17401012	2007	J Clin Oncol	Molecular subtypes in estrogen receptor positive breast carcinomas.
GSE12276	204	204	MFS	19421193	2009	Nature	Genes that mediate breast cancer metastasis to the brain
GSE19615	115	115	RFS, MFS	20098429	2010	Nat Med	Chemotherapy resistance and recurrence of BRCA.
GSE17907	55	39	MFS	20932292	2010	BMC Cancer	Genome profiling of ERBB2-amplified breast cancers
GSE20685	327	327	OS, MFS	21501481	2011	BMC Cancer	BRCA molecular subtypes and clinical outcomes: treatment optimisation.
GSE21653	266	252	DFS	20490655	2011	BRCA Res Treat	A gene expression signature identifies two prognostic subgroups of basal BRCA
TOTAL	1054	1024					

Table 6.1: Recopilación de series y fuentes de microarrays BRCA.

Todas las series consideradas, deben contener los siguientes metadatos: (i) Tiempo de supervivencia. (ii) Estado (si los datos están censurados o no al final del seguimiento del paciente). (iii) Medición de IHC, si es posible, para los marcadores primarios de BRCA: ER, PR y HER2.

Estas series se obtuvieron principalmente de Gene Expression Omnibus (GEO) (NCBI, 2019), utilizando las herramientas de búsqueda GEO como la función getGEO de **GEOquery** (Davis and Meltzer, 2007) para R.

Tenemos un total de 1024 muestras de Plus2, un subconjunto de 380 con valores de IHC para ER, PR y HER2. Para las otras 644 muestras, los valores de IHC faltan o están incompletos, pero todos tienen tiempo de supervivencia y valor de estado.

El conjunto de datos RNAseq se basa total o parcialmente en los datos generados por la red de investigación TCGA (TCGA, 2019). Este conjunto de datos y phenodata son proporcionados al usuario por **RCurl** (Lang and the CRAN team, 2019), **curatedTCGAData** (Ramos, 2019), y **TCGAutils** (Ramos et al., 2019) paquetes de R. Los paquetes nos permiten acceder a los recuentos sin procesar, a la matriz FPKM o RPKM ya toda la información clínica disponible. Elegimos los RPKM recomendados por TCGA.

Varios algoritmos han sido propuestos en la literatura para eliminar el efecto batch. efecto. En (Kupfer et al., 2012) **RMA + COMBAT** (**inSilicoMerging** paquete de R) se propone un método que hemo modificado para obtener nuestra normalización. Para nuestro problema se necesita una corrección personalizada del efecto batch. El hecho de que los tamaños de muestra difieran entre series, las condiciones de los experimentos y el año del estudio podría llevar a un efecto de batch más fuerte así que este apartado se desarrolla en profundidad en otro capítulo.

Descubrimiento de marcadores IHC y predicción de estado y riesgo

Como se explicó anteriormente, los errores en la determinación de marcadores estándar por inmunohistoquímica pueden tener un impacto sustancial en la salud del paciente (Venet et al., 2011). En particular, el valor de estos marcadores nos permite clasificar a los pacientes con cáncer de mama en cuatro grupos.

Para evitar este problema, se desarrolló un predictor robusto para reducir los errores en la determinación de los marcadores IHC. Aunque se han aplicado varios predictores a la estimación de marcadores IHC utilizando los perfiles de expresión génica (Bartlett et al., 2016), no son capaces de reducir los errores significativamente (Mertins et al., 2016).

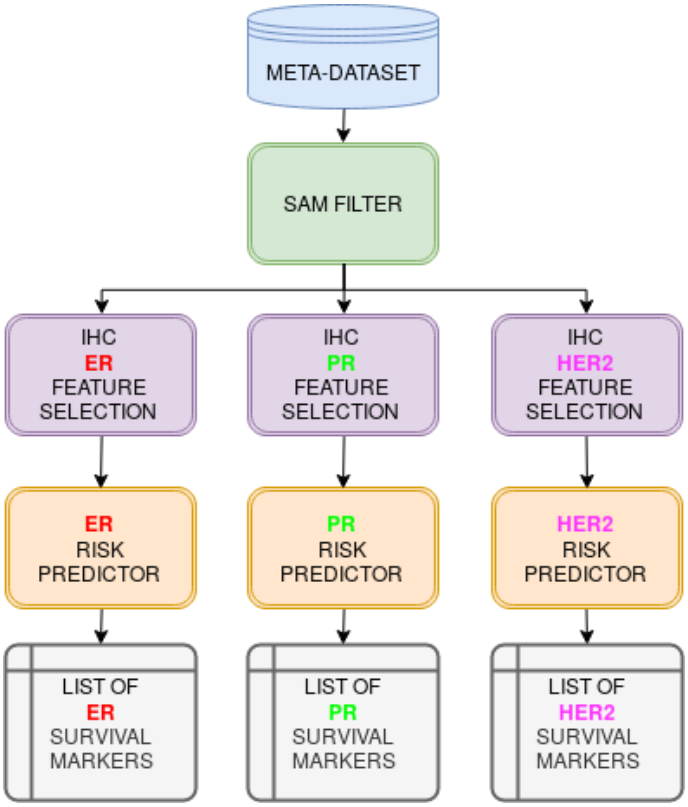


Figure 6.6: Descubrimiento de los genes marcadores IHC, algoritmo de predicción de características y riesgo.

Nuestro enfoque se basa en la idea de que los marcadores clínicos están determinados por un conjunto de pathways y genes. Por lo tanto, se selecciona un pequeño subconjunto de características que están fuertemente asociadas a marcadores clínicos. Este subconjunto de genes puede proporcionar dianas alternativas a los marcadores clínicos estándar. A continuación, se construye un conjunto de clasificadores lineales utilizando una estrategia de bagging. (Mbogning and Broet, 2016).

La figura (**Fig: 6.6**) describe el proceso. Primero, se aplica un algoritmo robusto

de SAM para filtrar genes no expresados. Luego, para cada marcador, ER, PR y HER2, un conjunto de clasificadores obtiene una lista relacionada de genes y mejora la predicción de marcadores clínicos estándar. El procedimiento incluye remuestreo y validación por AUC. A continuación, se entrenan predictores de riesgo que son capaces de evaluar de forma multivariante la capacidad de cada gen de influir en el mismo. Por último se obtienen las listas de genes marcadores. Estos genes tienen la capacidad tanto de predecir clases como riesgo. Están fuertemente relacionados con los tres marcadores clínicos.

Discusión y resultados

Las funciones desarrolladas para la selección robusta de características mediante marcadores IHC y el algoritmo redefinido de red elástica han demostrado la capacidad de predecir subclases. Las predicciones de IHC son robustas cuando se enfrentan a variaciones de muestra. Los valores altos de AUC demuestran el alto rendimiento del predictor.

El método multivariante que utiliza un grupo de genes para predecir las curvas de riesgo es capaz de reducir la "zona de incertidumbre" o la zona de riesgo medio con nuestro enfoque. La diferencia con otros métodos al evaluar los intervalos para definir alto y bajo riesgo permite buscar un mínimo que maximice la separabilidad de las curvas de Kaplan Meier. La utilidad de los mínimos relativos cuando se mide el ancho de la zona de riesgo medio es clave cuando el método crea el punto de corte para los riesgos altos y bajos predichos. Por lo tanto, la estratificación del paciente está garantizada y los pacientes asignados a un grupo no concluyente o incorrecto disminuyen.

El método logrank desarrollado para redefinir la estratificación en expresión alta o baja con la maximización de la separabilidad de las curvas pronósticas utiliza una nueva "probabilidad de pertenencia". Esta probabilidad de ser asignado a un grupo es una estrategia que elimina la dependencia de la muestra para mejorar la robustez. Por lo tanto, la posibilidad de que un paciente cambie de alto a bajo o viceversa se reduce al mínimo si se agregan o si se restan nuevas muestras.

El poder de predicción de riesgo de los genes marcadores propuestos se evalúa en comparación con dos de las listas de genes más famosas, Prosigna y Oncotype. En ambos casos, los genes marcadores propuestos son capaces de superar a las otras listas.

Por lo tanto, los genes marcadores propuestos son objetivos prometedores para la validación *in vivo*.

Para finalizar, hay una fuerte relación entre los marcadores propuestos y las funciones celulares relacionadas con el cáncer. Se ha descubierto que todos los genes tienen relación con la proliferación celular, la inestabilidad del ADN, la diferenciación y todo tipo de factores fuertemente asociados con el cáncer.

Marcadores positivos y los reguladores del cáncer de mama triple negativo (TNBC) utilizando perfiles transcriptómicos combinados con análisis de supervivencia

Motivación

El cáncer de mama (BRCA) se clasifica como Triple Negativo (TNBC); cuando no muestra una expresión significativa del receptor de estrógeno (ER) o del receptor de progesterona (PR) y no sobreexpresa el receptor 2 del factor de crecimiento epidérmico humano (HER2). La presencia de estos marcadores moleculares se realiza mediante inmunohistoquímica (IHC) e hibridación in situ con fluorescencia y se ha demostrado que tiene una variabilidad significativa entre laboratorios. Esto es problemático, ya que las muestras de HER2 + o de receptores hormonales positivos (HR++) con falsos negativos a través de estos análisis corren el riesgo de ser clasificadas como TNBC.

Por lo tanto, las muestras designadas como TNBC podrían someterse a un paso de verificación para garantizar que los tumores no se hayan clasificado incorrectamente como TNBC. La definición de TNBC es la fuente de una posible clasificación errónea, parece prudente que una clasificación adicional de TNBC se base en criterios inclusivos (es decir, usando biomarcadores positivos). Para identificar biomarcadores potenciales, se utilizó DECO (Decomposing heterogenous Cohorts using Omic data profiling) para identificar 24 genes cuya regulación al positiva caracteriza mejor los casos de TNBC. Los biomarcadores identificados se pueden usar para propósitos de clasificación y predicción.

Además, se utilizó Viper (Virtual Inference of Protein-activity by Enriched Regulon Analysis) para determinar qué factores de transcripción (TF) son diferencialmente más activos en TNBC que en HR++. Viper identificó a BCL11A y FOXC1 como conductores potenciales de TNBC. Estos TF podrían servir como objetivos potenciales para terapias dirigidas a TNBC.

Además, se analizará la relación entre los marcadores descubiertos y la supervivencia y el riesgo. Este análisis de riesgo y supervivencia proporcionará relevancia a los marcadores.

Materiales y métodos

General workflow of the study

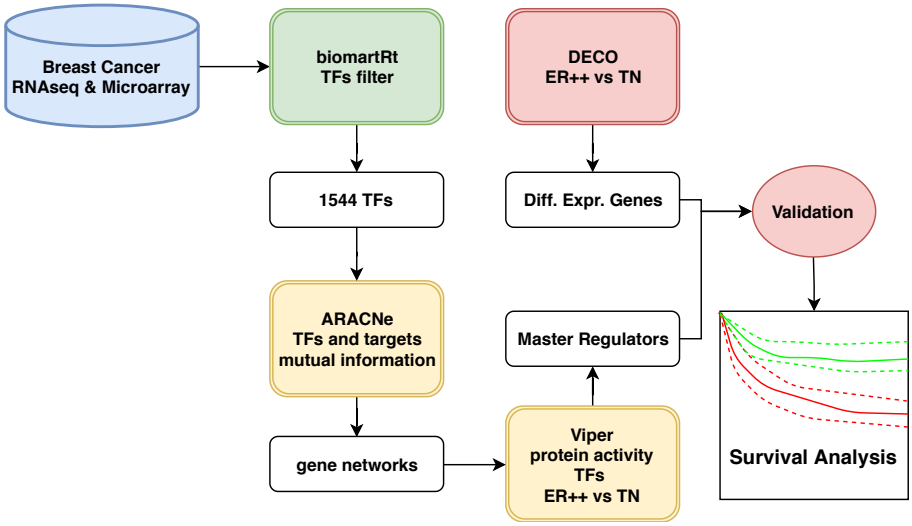


Figure 6.7: ER++ vs TNBC metodología y algoritmos aplicados.

En la **Fig: 6.7**, se describe el proceso. Primero, se usan los datos de la recopilación de microarray y series RNAseq. El primer paso es el filtro para identificar los TF descritos anteriormente, se realizó utilizando **biomartRt** (Durinck et al., 2005) (Durinck et al., 2009).

Se analizaron los TFs(factores de transcripción) usando el paquete de R **ARACNe** (Margolin et al., 2006) (He et al., 2017) y se creó una red de TFs y sus targets. A continuación. se utilizó el paquete de R **Viper** (Alvarez et al., 2016) para obtener la firma de los "master regulators".

Además, se realizó un análisis independiente de DECO para seleccionar genes equivalentes a TFs.

La relación de los marcadores con la supervivencia se verificó utilizando las herramientas desarrolladas en esta tesis.

Discusión y resultados

Las muestras de tumores BRCA son evaluadas actualmente por IHC y FISH para determinar la presencia de los tres marcadores moleculares de BRCA: ER, PR y HER2. Cuando una muestra no tiene una expresión significativa en ninguno de estos marcadores, se le asigna la etiqueta de TNBC. Por lo tanto, las muestras de HER2 + o HR++ que muestran falsos negativos a través de estos métodos de análisis están en riesgo de ser clasificadas como TNBC.

Viper se utilizó para evaluar la diferencia en la actividad de las proteínas en TNBC y HR++. Viper identificó 12 TFs como más activos en las muestras de HR++

en comparación con las muestras de TNBC. Entre los 12 TFs se encontraban ER y PR, como era de esperar para esta comparación. Estos hallazgos refuerzan la eficacia de los métodos utilizados y justifican la exploración de otros FTs obtenidos. Dado que los subtipos derivados de ER y PR se han tratado exitosamente con fármacos en HR++, parece lógico que los FTs que se identificaron como más activos en el TNBC pudieran servir como posibles objetivos para la inhibición en el TNBC. Por último, Viper identificó 34 TFs con mayor actividad en TNBC. BCL11A y FOXC1 se identificaron en ambos conjuntos de datos como los mas relevantes.

BCL11A un represor transcripcional relacionado con el BRCA

BCL11A codifica un TF de dedos de zinc que funciona como un represor transcripcional. BCL11A también es un protooncogén para los cánceres hematológicos y es un biomarcador propuesto para los tumores de células no pequeñas del pulmón (Jiang et al., 2013) (Weniger et al., 2006) (Nakamura et al., 2000). Su expresión es esencial para el desarrollo adecuado de las células B y T, y se ha encontrado en niveles bajos en el timo, la médula ósea y los ganglios linfáticos, así como en niveles altos en las células B del centro germinal y del cerebro fetal (Satterwhite et al., 2001) (Liu et al., 2003). También se ha descubierto recientemente que BCL11A es un regulador del desarrollo normal de las glándulas mamarias, necesario para el desarrollo de células madre en la mama. Sin embargo, se ha demostrado que los altos niveles de expresión de BCL11A promueven la tumorigénesis en TNBC y se correlacionan negativamente con la supervivencia (Khaled et al., 2015).

FOXC1, infraexpresado tanto en HER2 + como HR++

FOXC1 fue identificado como marcador de expresión diferencial entre HR++, HER2 + y TNBC siendo superado tan solo por ER y ERBB2. FOXC1 está definido por el aumento de su expresión en TNBC y por su disminución tanto en HER2 + como en HR++.

Los miembros de la familia FOX desempeñan diversos roles en la organogénesis, la regulación del ciclo celular y la diferenciación celular (Tuteja and Kaestner, 2007a) (Tuteja and Kaestner, 2007b). FOXC1 se ha asociado específicamente con varios tipos de cáncer, incluido el linfoma de Hodgkin, el linfoma no Hodgkin, el carcinoma hepatocelular, el cáncer de endometrio y el cáncer de mama. (Elian et al., 2018).

Los factores de transcripción como potenciales marcadores

Hemos demostrado que FOXC1 y BCL11A ejercen su influencia en el fenotipo TNBC a través de la sobreexpresión, el aumento de la actividad de otros TFs y otras relaciones sinérgicas. Estos hallazgos nos han llevado a postular que FOXC1 y BCL11A pueden ser genes que definen positivamente el TNBC. La hipótesis es que estos TFs podrían servir como objetivos viables para el tratamiento de este subtipo de cáncer. Se recomiendan como marcadores para estudios adicionales utilizando modelos murinos y líneas celulares para evaluar su papel y su uso potencial en la terapia dirigida

de TNBC.

La relación entre los marcadores TNBC descubiertos y el poder de predicción de riesgo y supervivencia es una gran adición al estudio. El valor que esta relación incorpora a los marcadores es triple. Primero, la posibilidad de calcular y predecir el riesgo para nuevos pacientes en la clínica mientras se evalúa la pertenencia a un subgrupo. Segundo, la posibilidad de ser usados como factores que definen el subgrupo BRCA triple negativo apenas conocido. Finalmente, los marcadores propuestos podrían investigarse para definir subgrupos dentro del TNBC.

En conclusión, nuestro trabajo nos ha permitido identificar biomarcadores positivos en TNBC. Estos biomarcadores podrían servir para confirmar que las muestras designadas como TNBC son verdaderamente TNBC y no HER2 + de HR++ definidos como falsos negativos a través de FISH o IHC. Además, nuestro trabajo permitió la identificación de potenciales marcadores de TNBC. Estos marcadores potenciales deben investigarse más a fondo para determinar si su actividad es, de hecho, el motor de TNBC y si estos TFs serían claros objetivos a considerar en una terapia viable en el tratamiento de TNBC.

Genes marcadores de supervivencia del cáncer de colon (CRC) derivados de la integración y el metanálisis de múltiples conjuntos de datos transcriptómicos

Motivación

El CCR (cáncer colorectal) es una enfermedad heterogénea, ya que de un paciente a otro difiere en la presentación clínica, las características moleculares y el pronóstico (Linnekamp et al., 2015). La heterogeneidad de los CCRs aumenta la complejidad de esta patología tumoral, lo que hace que los subtipos y la estratificación sean una tarea difícil.

De esta manera, la medicina personalizada para el CCR es cada vez más necesaria, especialmente en terapias dirigidas donde existen grandes variaciones entre las respuestas al tratamiento del individuo (Linnekamp et al., 2015) (Dienstmann et al., 2017). En este contexto, la necesidad de encontrar marcadores genéticos robustos asociados con subtipos específicos de CCR es lo que nos llevó a este estudio. Además, el propósito específico de nuestro trabajo fue encontrar objetivos biomoleculares consistentes para facilitar la estratificación de las muestras y que pudieran relacionarse con el pronóstico de la enfermedad utilizando datos de supervivencia.

En la clínica, los pacientes se clasifican en cuatro estadios de CCR según las características anatomopatológicas de sus tumores. Es común usar el Sistema de estadificación TNM (donde T significa tumor, N para ganglios linfáticos y M para metástasis). Esta categorización de cuatro etapas representa grupos de pacientes significativamente distintos para la recaída de la enfermedad, pero las etapas no predicen el riesgo de cada paciente individual porque no están directamente asociadas a la supervivencia (Tauriello and Batlle, 2016).

Se investigarán los perfiles globales de expresión génica de los tumores colorrectales y su alteración a lo largo de las etapas. Así se espera identificar los genes que podrían aprovecharse como biomarcadores de supervivencia y pronóstico de CCR en las últimas etapas (es decir, III y IV). Para llevar a cabo este trabajo, realizamos un análisis profundo en una gran cohorte de muestras humanas derivadas de una sólida integración de varios conjuntos de datos que tenían datos de supervivencia clínica y transcriptómica. La integración proporcionó un meta-conjunto de datos homogéneo y bien estandarizado que incluye 1273 muestras colorrectales humanas. La identificación de marcadores candidatos se realizó mediante un contraste inicial entre la expresión génica del subconjunto de pacientes con CCR asignados por sus características clínicas a los estadios I y II frente a los pacientes con tumores correspondientes a los estadios III y IV. Finalmente, después de la validación cruzada interna y externa, los genes seleccionados como mejores marcadores de supervivencia se utilizaron para construir un predictor de riesgo para permitir la estratificación de los pacientes con respecto a su riesgo relativo.

Materiales y métodos

General workflow of the study

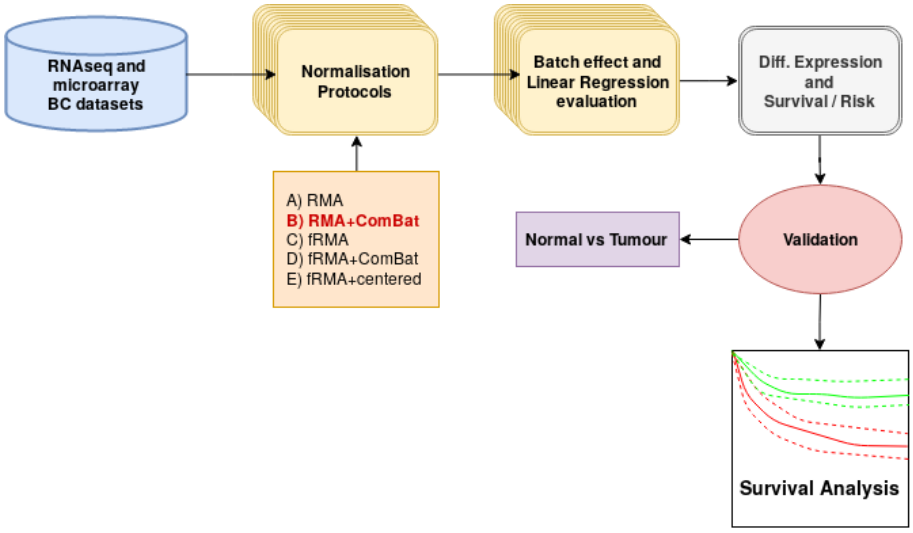


Figure 6.8: CRC metodología y algoritmos aplicados.

En la **Fig: 6.8**, se resume el proceso que se ha seguido. Primero, el conjunto de datos se fusiona y se normaliza utilizando cinco protocolos diferentes. La elección del mejor método de ejecución se explica con más detalle en el capítulo, pero depende de su capacidad para eliminar el efecto "batch". Se selecciona el método **RMA + ComBat**, utilizando este método dada su eficiencia para paliar el "batch effect".

El análisis de la expresión diferencial se realizó usando **Limma**. Este análisis proporcionó una lista de genes relevantes que se analizaron más a fondo utilizando las herramientas de supervivencia desarrolladas en esta tesis. La lista final de marcadores se valida en dos estudios independientes.

El primero, evalúa la relación multivariante de los genes con el riesgo, que no se analiza en el paso anterior. El segundo, compara muestras normales sin enfermedad y muestras de CRC. Las muestras colorrectales de la serie tienen una expresión más alta que las muestras normales para nuestros genes marcadores definidos por la sobreexpresión. Además, los genes marcadores definidos por la infraexpresión están regulados a la baja en las muestras de CRC.

Discusión y resultados

Con respecto a los genes específicos propuestos como marcadores de supervivencia de CRC, queremos subrayar que nuestro estudio no pretende proporcionar una firma genética fija para el pronóstico y la predicción de riesgo, como las firmas reportadas de 7 genes, 15 genes o 113 genes. (Nguyen et al., 2015) (Chen et al., 2017) (Xu et al., 2017) pero, en cambio, proponemos un conjunto sólido de genes clasificados de acuerdo con su poder predictivo de supervivencia CRC. De esta manera,

se presenta una lista ordenada de 200 genes que incluyen los mejores marcadores de supervivencia: 100 genes para los que la sobreexpresión está relacionada con supervivencia y 100 genes para los que la sobreexpresión no está relacionada con supervivencia. Creemos que este enfoque es más útil, ya que permite una selección abierta de diferentes números de genes para propósitos o investigaciones adicionales (por ejemplo, para pruebas adicionales con otras cohortes clínicas de CCR). De hecho, utilizamos los 100 genes más significativos, sobreexpresados en relación con la progresión en CRC, para construir el predictor de riesgo y utilizamos los 5 genes principales o los 10 principales de esta lista para las validaciones externas con diferentes conjuntos de datos independientes.

Otro comentario relevante es, que como se recordó anteriormente, construimos el predictor de riesgo utilizando los genes que mostraron sobreexpresión correlacionada con un pronóstico desfavorable. Esto se hizo porque en la selección de biomarcadores es mejor usar los que indican una señal positiva (es decir, factores que proporcionan una "ganancia de función") que los que indican una señal negativa. Por lo tanto, todos los marcadores de supervivencia de genes fueron definidos como sobreexpresados en pacientes de CRC de alto riesgo. El hecho de que estén sobreexpresados facilitará su detección mediante protocolos biomoleculares estándar (PCR, ELISA, inmunohistoquímica, etc.).

Finalmente, estamos investigando el significado biológico de los genes que etiquetados como marcadores de predicción y pronóstico. Nuestros esfuerzos se centran en los 10 principales para los cuales la sobreexpresión se relacionaba con mala supervivencia: DCBLD2, PTPN14, LAMP5, TM4SF1, NPR3, LEMD1, LCA5, CSGALNACT2, SLC2A3, GADD45B. El análisis de la literatura revela algunas observaciones interesantes. Por ejemplo, la proteína transmembrana DCBLD2 (ESDN), miembro de la familia de proteínas de tipo neuropilina, es un nuevo regulador de los efectos mitóticos y metabólicos de la insulina y modula la transducción de señales a través de la regulación de la interacción del receptor de insulina con sus proteínas adaptadoras. (Li et al., 2016). La importancia de la regulación de la insulina en la función de nuestro sistema digestivo es clara, y agrega un valor extra a la propuesta de **DCBLD2** como marcador de supervivencia de CCR.

En conclusión, consideramos que los resultados presentados en este trabajo brindan un fuerte respaldo y una sólida justificación para el valor pronóstico de un nuevo conjunto de genes en CCR y su potencial para predecir la progresión del tumor colorrectal y la evolución hacia los estadios III y IV. El conjunto final de marcadores de supervivencia incluye una lista abierta de cien genes regulados al alza, con una estimación estadística sólida del valor de cada uno. De esta manera, el conjunto de genes se clasifica claramente, siendo los primeros en la lista los que brindan la mejor fortaleza pronóstica y los que se pueden introducir para construir predictores con menor número de genes. De hecho, nuestros resultados mostraron que una selección de los 5 mejores genes aplicados a cohortes externas independientes proporcionó una muy buena separación de muestras de CCR en dos grupos distintos de alto y bajo riesgo.

Perfil transcriptómico integrativo de los subtipos moleculares de consenso del cáncer colorrectal (CRC, por sus siglas en inglés) con datos de supervivencia y caracterización de una firma de genes EMT asociada al KO de P21, CDKN1A (- / -)

Motivación

La formación de metástasis se basa en un proceso de varios pasos conocido como la cascada invasión-metastásica. Comienza con la diseminación de células cancerosas desde el tumor primario, su supervivencia en el sistema circulatorio, la extravasación y, eventualmente, la recolonización de un órgano distante, generando así un tumor secundario. Todos los pasos individuales requieren características específicas de células tumorales, que están conectadas en gran medida a la transición epitelio mesenquimal (EMT) y al fenotipo de las células madre de cáncer. Aunque es bien sabido que la progresión epitelio mesenquimal de colon normal a un carcinoma invasivo y metastásico está fuertemente asociada con el proceso de la EMT y la capacidad de las células tumorales para sobrevivir en condiciones no adherentes. Definir la diseminación metastásica en pacientes sigue siendo un foco importante de la investigación de CCR.

El inhibidor de la quinasa dependiente de ciclina (p21) es un regulador negativo tanto de la progresión del ciclo celular como de la expresión génica (Abbas and Dutta, 2009). Un paso no regulado de células a través del punto de control G1/S por infraregulación o pérdida de función de p21 podría inducir a una proliferación aberrante y, por lo tanto, a desencadenar la transformación del tumor. En CRC, se ha reportado que la regulación a la baja de la expresión de p21 se correlaciona con el desarrollo de metástasis y la escasa supervivencia del paciente (Abbas and Dutta, 2009). Por lo tanto, el silenciamiento de p21 parece ser de gran importancia para la proliferación sin restricciones de células cancerosas.

Estudios previos han reportado que existe una relación directa entre las líneas celulares HCT116 y HCT116 p21 KO en subtipos de cáncer colorrectal (CCR) definidos como subtipos de consenso molecular, CMS1 y CMS4, respectivamente. También se ha informado de que algunos genes como VIM y p21 están relacionados con la transición epitelio mesenquimal (EMT), que es el sello principal de CMS4. La diferencia en la expresión de estos genes en una línea y en la otra es la medida de la diferencia entre las subclases.

Por lo tanto, el objetivo central es demostrar si existe una relación entre las líneas celulares HCT116 y HCT116 p21 KO y los subtipos CMS1 y CMS4, respectivamente.

Para corroborar nuestra hipótesis, utilizamos los conjuntos de datos de microarrays y RNAseq de CRC para validar los resultados obtenidos del análisis de la expresión diferencial. Clasificamos las muestras en los subtipos de CMS utilizando

clasificadores ya desarrollados. Esto nos permite verificar si las observaciones biológicas en líneas celulares se corresponden con los datos obtenidos en pacientes humanos. Los genes marcadores se validarán midiendo si el patrón genético encontrado en las líneas celulares HTC166 y HTC166 KO es similar al obtenido en los subtipos CMS1 y CMS4, respectivamente.

Al mismo tiempo, se evalúa la interacción de los genes marcadores con el riesgo y la supervivencia. La diferencia ya propuesta en el resultado del análisis de supervivencia entre CMS4 y el resto de subtipos se investigará como validación.

Según (Guinney et al., 2015), el subtipo CMS4 refleja la firma del génica de las células mesenquimales junto con la señalización de TGF-β y la remodelación de la matriz. Curiosamente, el subtipo CMS4 también se correlacionó con la resistencia a fármacos y el aumento de los brotes tumorales (Trinh et al., 2018). En la clasificación por consenso de subtipo molecular de adenoma colorrectal no hubo subtipo CMS4 ya que no existe un estroma asociado a la invasión (Komor et al., 2018). Hasta ahora no se ha determinado la firma genética detallada de las células HCT116 p21 KO. Para una mejor comprensión del modelo p21 KO particularmente como un modelo preclínico para el análisis de la respuesta terapéutica, nos propusimos evaluar si el p21 KO es lo suficientemente fuerte como para cambiar el subtipo molecular de la línea celular HCT116.

Materiales y métodos

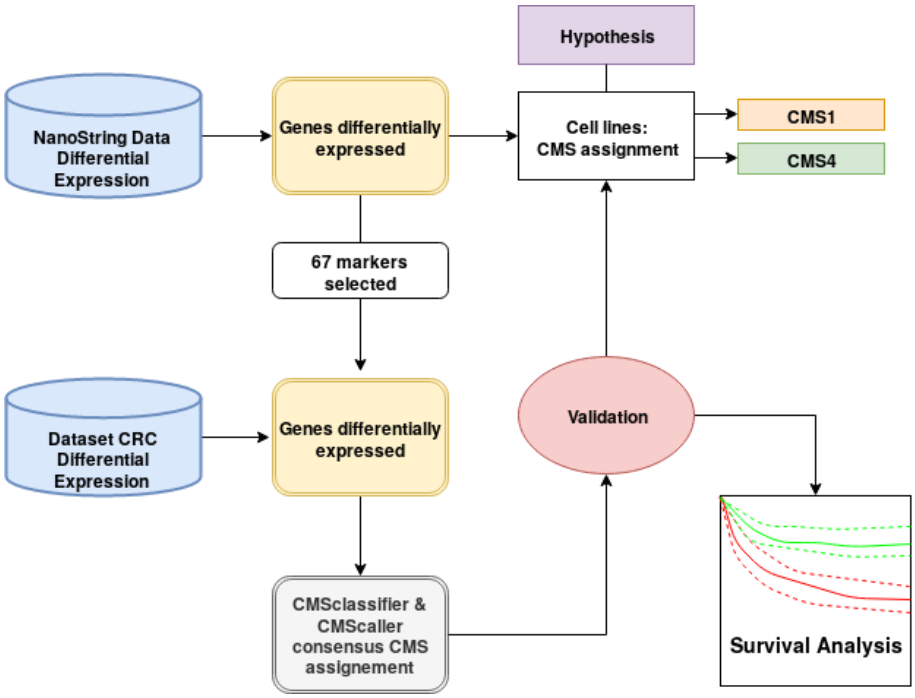


Figure 6.9: CRC metodología y algoritmos aplicados.

En la **Fig: 6.9**, se resume procedimiento. Primero, tenemos los datos de la comparación entre las líneas celulares y la serie de microarrays y RNAseq de CRC. El

primer paso fue obtener los genes principales que mostraban una expresión diferencial entre la línea celular original y el knock out (KO).

Esto nos dio 67 marcadores candidatos que estaban relacionados con los subtipos CMS1 o CMS4. Si la expresión del gen es mayor para la línea celular similar a CMS1, se define como un marcador positivo de CMS1. Al contrario, si la expresión es más alta para CMS4 en la línea celular KO (p21 KO), entonces se define como un marcador positivo de CMS4.

El estudio con datos de microarrays y RNAseq se realiza para validar estas hipótesis. Para la clasificación de pacientes en los subtipos de CMS, utilizamos el consenso entre dos predictores de clase, CMSclssifier y CMSCaller. Las herramientas desarrolladas en esta tesis se aplicaron para el posterior análisis relacionado con la supervivencia.

Discusión y resultados

Los experimentos y análisis realizados en este capítulo demuestran la relación entre el tipo WT y las líneas celulares p21 KO CMS4 (tipo EMT). Los marcadores propuestos de estas líneas celulares están fuertemente diferenciados en los grupos de CMS predichos en las dos cohortes de validación, microarrays y RNAseq, lo que confirma la hipótesis. Por otra parte, la relación entre los marcadores y las características EMT también se ha probado.

La utilidad del conjunto de herramientas, las funciones y el paquete desarrollados en esta tesis adquiere relevancia dados los resultados obtenidos. Como validación adicional, la relación ya descrita entre el grupo de CMS4 y el resto de pacientes con cáncer colorrectal se confirma mediante el análisis de supervivencia.

Finalmente, el estudio sobre cómo los pares de genes son capaces de mejorar su capacidad para definir el riesgo cuando se incluye la interacción entre ellos es un resultado considerable. Generalmente, la mayoría de los estudios que incluyen regresiones de Cox, lineales o de otro tipo no tienen en cuenta esta relación como una variable o coeficiente adicional que debería de incluirse en el modelo. Como se describe, esto puede llevar a peores resultados o interacciones no descubiertas. Esta lista de marcadores sería un objetivo ideal para futuros estudios y desarrollo de herramientas.

La relación del p21 KO con el subtipo CMS4 se valida de varias maneras, utilizando dos conjuntos de datos diferentes de CRC. La relación de los marcadores propuestos con las funciones celulares que se describen relacionadas con los subtipos CMS1 y CMS4 demuestra que los genes tienen la capacidad de identificar ambas subclases.

La validación *in vivo* de los marcadores más prometedores ya se ha propuesto como un trabajo futuro.

La aplicación de una versión modificada de la regresión multivariante de Cox, incluidas las interacciones entre términos, ha proporcionado una nueva métrica para

evaluar los marcadores. La importancia de las interacciones entre pares de genes queda demostrada al mostrar cómo se redefine la red y cómo las interacciones entre los genes son aún más prometedoras. La evaluación *in vivo* de estos genes tiene el potencial de conducir a nuevos descubrimientos debido a la relación entre ellos.

Conclusiones finales

A lo largo de los cuatro capítulos de este Ph.D. Se han propuesto algoritmos y métodos bioinformáticos para abordar los principales problemas en el análisis y la integración de datos. Las conclusiones generales de este Ph.D., se enumeran a continuación:

1. Los métodos propuestos para la normalización del conjunto de datos y la reducción del efecto de "batch" deben utilizarse para reducir el sesgo al fusionar diferentes fuentes de datos. La importancia de estos métodos se demuestra en el estudio realizado en **Capítulo 4**. La falta de una normalización y estandarización adecuadas al fusionar los datos de diferentes fuentes es algo que generalmente conduce a la falta de reproducibilidad. Usando estos métodos, se consiguen cumplir el **primero** y el **tercero** de los objetivos de esta tesis. Se generaron conjuntos de datos grandes (1024 muestras) y (1273 muestras) y homogéneos de cáncer de mama y cáncer colorrectal con datos de supervivencia.

2. Los métodos y algoritmos desarrollados han hecho posible la caracterización de un grupo de genes marcadores propuestos que identificaron el subtipo TNBC de cáncer de mama y que se relacionan con el riesgo y el valor pronóstico. El éxito en la compilación de un gran conjunto de datos BRCA con supervivencia nos permitió realizar este análisis y afirmar que se ha logrado el **segundo objetivo** de esta tesis.

3. Además, la relación entre el p21 KO del HCT116 con el subtipo CMS4 y el HCT116 WT con el CMS1 ha permitido identificar genes que definen este subtipo de cáncer colorrectal. Los genes propuestos se evaluaron más a fondo y se descubrió la relación entre los marcadores y la supervivencia. En el mismo estudio, se probó un método capaz de identificar genes que definen fuertemente el riesgo cuando se tiene en cuenta la interacción binaria. Los resultados relevantes obtenidos se probarán *in vivo* para definirlos como marcadores. Esto fue posible gracias a la compilación de un gran conjunto de datos integrado de cáncer colorrectal desarrollado previamente y que se utilizó para este estudio. Con esta contribución, se cumple el **cuarto y último objetivo** de esta tesis.

4. La importancia de la robustez, cada vez que se realiza un análisis bioestadístico, es uno de los temas principales de esta tesis. Esto se ha tenido en cuenta para cada algoritmo diseñado y aplicado. Por lo tanto, la validación mediante cohortes y conjuntos de datos independientes, incluso utilizando distintas plataformas, se ha realizado en cada paso y estudio realizado.

5. El método propuesto para descubrir genes marcadores se ha aplicado con éxito, demostrando su capacidad para descubrir grupos de genes que superan a los ya propuestos para la predicción del riesgo y la estratificación de pacientes, manteniendo la relación con características clínicas importantes. Del mismo modo, la variación robusta de Kaplan-Meier con la optimización de la probabilidad de pertenencia a grupos es capaz de superar los problemas ómicos habituales que pueden llevar a un sobreajuste.

Trabajo futuro

Los descubrimientos de varios capítulos de esta tesis ya se han publicado, pero el objetivo es terminar pronto una o dos publicaciones.

Los marcadores positivos obtenidos en el Capítulo 5 son candidatos para ser evaluados *in vivo*.

El desarrollo de un paquete de bioconductores R que incluye todas las herramientas utilizadas para realizar los estudios en esta tesis se encuentra en fase tardía.

ACKNOWLEDGEMENTS

Agradecimientos

Agradecer especialmente a mi director de tesis, de TFG, de TFM, el Dr. Javier De Las Rivas, por el trato tan especial a lo largo de estos años, por encauzarme siempre que me podía la falta de ganas y por la confianza y la preocupación que siempre tiene por todos nosotros. Dudo mucho que en mis posibles futuros jefes vaya a encontrar trato mejor.

Y no menos a mi co-director de tesis, el Dr. Manuel Martín-Merino, por la paciencia durante las MUY largas horas, desde que le conocí en sus clases de circuitos en la facultad de Informática de la UPSA en la asignatura que mejor se me daba, y por la larga trayectoria asesorándome sin descanso. Sin él no habría sido posible llegar hasta aquí.

Otra mención especial al Dr. José Manuel Sánchez, por las risas y sus consejos estadísticos.

A todo el personal del Centro de Investigación del Cáncer y la Universidad de Salamanca, administración, conserjería, limpieza, servicio técnico, mantenimiento, y compañeros en general, por todo el apoyo y ayuda prestada. Especialmente a Marga, porque sin ella me habrían quitado la beca hace años, menos mal que está en todo.

A Mónica, que nos ha alegrado los veranos y nos ha ayudado muchísimo, espero haber aprendido todo el inglés que has tratado de enseñarme durante tus "vacaciones".

Cuando pienso en los años que llevo aquí ni me acuerdo de cuantos son y me da la risa, parece que ha sido toda una vida. A lo largo del tiempo he conocido a una gran cantidad de personas increíbles, y eso que yo no soy muy de hacer amigos. Gracias al laboratorio 19, y a todas las personas que por aquí han pasado, y que soy capaz de listar porque estoy mirando la tesis de Curro: Bea, Conrad, Óscar, Sara Aibar, Katia, Elena, Luis, Andrea, Irelka, Lauren, Jesús, JORGE, Alberto, Fernando y muchos más. Ah y Curro, claro.

Los podría citar en plan ranking pero mejor no, no vaya a ser que se lo lean. Espero que Andrea pueda perdonarnos, que Bea sea feliz en Valencia (que pena que no vaya a ser mi jefa (bueno nunca se sabe)), que OJKAR no se rompa algo subiendo a piedras y no le revienten mas las ventanas del coche, a EleNA que es una chica 6/10, que espero que sigamos viéndonos mucho, a Laurene que a ver si se

estresa menos y si deja que su perro viva un poco mas tranquilo, a Jorge que fue como un padre durante el tiempo que estuvo por aquí (le debo mucho, me escribió un artículo), a nisandra no le digo nada, le tendré que aguantar todos los días en el WoW, y a Curro (que voy a decir de Curro, menos mal que empezamos a la vez, porque le debo tanto que ni sabría por donde empezar, le he echado mucho de menos este año).

Al laboratorio 17, a mi SaraO, Arturo, SaraG, José Antonio, Víctor, Sergio, Ana, María, Nico, Antonio, y a Chema y Carmen.

Gracias SaraO por los momentos tan geniales contigo. Aquí va un especial para Arturo y Elena, por los momentos tan geniales que pasé con ellos desde el principio, les debo un par de años o tres de sueldo, porque nunca me acuerdo de entregar papeles.

A mi familia. Que son los que mas hacen por mí, en las buenas y en las malas. Estoy muy orgulloso de ellos. A mi madre que no se como lo hace, y a mis hermanos, por ser además de mis hermanos mis amigos. A mi padre por ser un ejemplo de no ejemplo.

A los que me olvido, perdón porque estoy muy estresado y me queda poco tiempo para acabar todo esto.