Universitat Oberta de Catalunya

uoc.edu

Identification of a common stemness gene signature in breast cancer and mantle cell lymphoma

Jon Ortiz Abalia
Máster de Ciencia de Datos
Area 3

Tutor: Carles Barceló
Profesor responsable: Ferran Prados Carrasco

Entrega: 24 junio 2020


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

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Identification of a common stemness gene signature in breast cancer and mantle cell lymphoma.*
Nombre del autor:	*Jon Ortiz Abalia*
Nombre del consultor/a:	*Carles Barceló Pascual*
Nombre del PRA:	*Ferran Prados Carrasco*
Fecha de entrega (mm/aaaa):	06/2020
Titulación:	*Máster Universitario en Ciencia de Datos UOC*
Área del Trabajo Final:	*Área 3*
Idioma del trabajo:	*Inglés*
Palabras clave	Cancer Stem Cells, Cancer, Gene signature

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

El cáncer sigue siendo la segunda causa de muerte a nivel mundial. Una de las razones es la existencia de un subconjunto de células dentro del tumor con capacidad de autorrenovarse, migrar y resistir a la quimioterapia: las células madre cancerosas (*Cancer Stem Cells*, CSCs).

El objetivo del presente trabajo es estudiar si el cáncer de mama y el linfoma de células del manto comparten mecanismos moleculares relacionados con las CSCs para avanzar tanto en el conocimiento básico del cáncer como en su diagnóstico y tratamiento.

Los resultados obtenidos mediante *Gene Set Enrichment Analysis* (GSEA) muestran que todos los datos analizados (tanto de cáncer de mama como del linfoma de células del manto) están enriquecidos con genes relacionados con las CSCs, especialmente en aquellos cánceres con fenotipos invasivos, llevando a la identificación de 269 genes comúnmente enriquecidos en ambos tipos de cáncer.

El análisis de anotación funcional de los 269 genes muestra que el 10% de los genes son proteínas de unión a compuestos heterocíclicos, conocidos por ser

componentes clave de muchos de los medicamentos disponibles contra el cáncer.

El análisis con la herramienta bioinformática Kaplan-Meier Plotter muestra una correlación significativa con parámetros de supervivencia de 51 de los 53 genes más significativos de la firma genética. Este hecho se ha observado en varios tipos de cáncer analizados (mama, ovario, pulmón y estómago).

Finalmente, el poder predictivo de la firma genética ha sido evaluado mediante algoritmos de *machine learning*. Los resultados muestran una precisión en la predicción del pronóstico de los casos de cáncer mayor que la de otras firmas genéticas publicadas.

Abstract (in English, 250 words or less):

Cancer remains the second leading cause of death globally. One of the potential reasons behind this is the existence of a subset of cells within the tumour with capacity to self-renew, migrate and resist to chemotherapy: cancer stem cells (CSCs).

The objective of the present work is to study whether Breast Cancer and Mantle Cell Lymphoma share common molecular mechanisms related to the CSC machinery in order not only to advance in basic cancer knowledge but also to speed up diagnosis and provide novel and/or more effective treatment.

Results obtained using Gene Set Enrichment Analysis (GSEA) show that all BC and MCL expression datasets analysed are enriched with CSC related genes, especially in those datasets with invasive or early-onset phenotypes. Further analysis has led to the identification of a 269-CSC gene signature composed by CSC genes commonly enriched in BC and MCL.

Noticeably, functional annotation analysis of the genes included in the 269-CSC gene signature has shown that almost 10% of the genes map to heterocyclic compound binding proteins, known to be key structural components of many of the available anti-cancer drugs.

Survival analysis using Kaplan-Meier Plotter confirmed a significant correlation with survival for 51 of the genes included in the 53-CSC gene signature in various cancer types analysed (breast, ovarian, lung and gastric cancer).

Finally, the predictive power for prognosis of the gene signature was assessed using machine learning. Results showed better accuracy in predicting prognosis of cancer cases than other CSC gene signatures published.

Index

List of Figures

List of Tables

1. Introduction

1.1. Context and rationale

Despite the decline of death rates and the improvement of 5-year survival rates over the years due to improvements in screening, earlier detection and availability of new tailored treatments, cancer remains the second leading cause of death globally (Lathia et al., 2019). One of the potential reasons behind this is the existence of a subset of cells within the tumour with capacity to self-renew, migrate and resist to chemotherapy. These cells are called cancer stem cells (CSCs) in a clear analogy with their normal counterparts (Batlle et al., 2017).

Breast cancer (BC) and mantle cell lymphoma (MCL) are two types of solid cancer with high incidence and mortality rate that are suspected to share some molecular mechanisms. Although the genetic basis of tumorigenesis may vary between different cancer types, the molecular mechanisms required for metastasis are similar.

BC is the most common cancer diagnosed in women and the major cause of cancer-related mortality in women worldwide, with a 12% probability of suffering from it throughout life. The 5-year survival rate is 99% if the cancer is located only in the breast, 85% if it has spread to regional lymph nodes and only 27% if it has spread to a distant part of the body. Unfortunately, 5% of women have metastatic breast cancer when they are first diagnosed (Society, 2020).

On the other hand, MCL, considered an aggressive type of B-cell non-Hodgkin lymphoma, has the worst prognosis among blood cancers with a median overall survival of 3 to 4 years (Luanpitpong et al., 2018). First identified in 1990, MCL is difficult to diagnose and hard to cure (Roschewski, 2015).

This is part of a real biomedical research project intended to study whether BC and MCL share common molecular mechanisms and whether these are related to the CSC machinery or not. The identification of a potential a stemness gene signature common for different tumour types such as BC and MCL as well as the underlying biological processes in which those genes are involved could provide valuable information about the molecular mechanisms leading to malignancy in solid tumours. The progressive digitalization of data, the use of specific software to analyse molecular data and the use of machine learning algorithms are transforming cancer research and healthcare.

The identification of specific pathways or genes involved in the cancer stem cell machinery could be of high importance not only to advance in basic cancer knowledge but also to speed up diagnosis and provide novel and/or more effective treatments.

1.2. Objectives

i. Identify common mechanisms related to cancer stem cells in two types of solid tumours: Breast Cancer and Mantle Cell Lymphoma.

    a. Identify genes involved in CSC machinery in BC expression datasets through pathway enrichment analysis (GSEA).

    b. Identify genes involved in CSC machinery in MCL expression datasets through pathway enrichment analysis (GSEA).

    c. Generate a common CSC-gene signature for BC and MCL.

ii. Map the molecular functions and biological processes of the genes included in the CSC-gene signature

iii. Study the prognostic significance of the genes included in the CSC-gene signature

iv. Study the predictive power for prognosis of the CSC-gene signature using Machine Learning.

1.3. Project planning

The roadmap of the project is depicted in **Figure 1**. It consists in five sequential phases: 1) Definition and planning, 2) State of the art, 3) Design and implementation, 4) Preparation of document, and 5) preparation of presentation and defense.

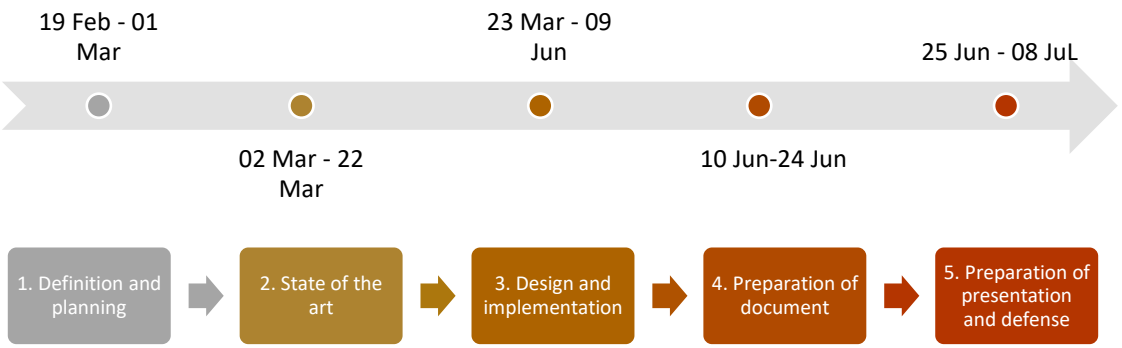


Figure 1. Roadmap of the project

1. **Definition and planning (10 days)**

2. **State of the art (20 days):**

    a. Search of bibliography for the following topics (5 -10 days):

        - Current status of Cancer Stem Cells model within cancer area (in general, Breast Cancer and Mantle Cell Lymphoma)

        - Studies comparing expression of CSCs and non-CSCs in cancer (in general, Breast Cancer and Mantle Cell Lymphoma)

        - Methodological papers: GSEA

        - Studies in which machine learning is used in cancer area (in general, Breast Cancer and Mantle Cell Lymphoma)

    b. Learn methodology for data pre-processing and data analysis (GSEA) (10-15 days)

3. **Design and implementation (76 days):**

    a. Data collection: search of CSC gene sets (gene pathway databases and bibliography) (5-10 days)

    b. Data collection: search of gene expression datasets (4 or 5): BC and MCL. (5-10 days)

    c. Data analysis: GSEA. Identification of common genes of both cancer types (BC and MCL) that are present in CSC gene sets: generation of CSC gene signatures. (10-15 days)

    d. Functional annotation analysis: functional annotation of genes included in CSC gene signatures: molecular function and biological process mappings using bioinformatic tools (Panther) (5-10 days)

    e. Correlation with prognosis: correlation analysis with prognostic parameters using bioinformatic tools (Kaplan-Meier Plotter) of the genes included in the CSC gene signatures. Comparison with GSEA results. (5-10 days)

    f. Predictive power for survival: study of the predictive power of the CSC gene signatures generated using machine learning algorithms. (5-10 days)

4. **Preparation of document (14 days)**

5. **Preparation of presentation and defense (13 days)**

The planned tasks in phase 3 ("*Design and implementation*") are summarized in

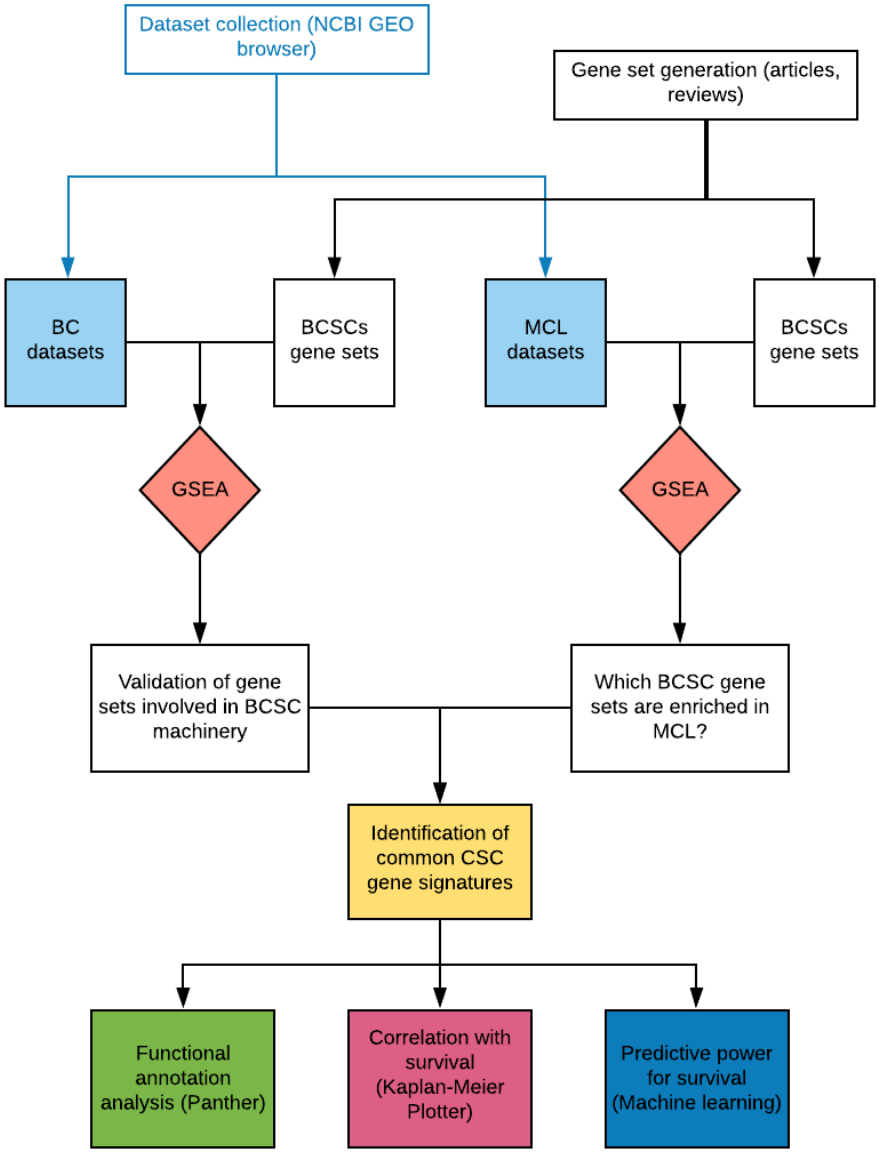


Figure 2. Design of the planned work

2. State of the art

2.1. Clinical impact of cancer stem cells (CSCs)

Multiple studies have proved that only specific cells within a tumour could initiate tumour growth. This has been confirmed by using xenograft transplantation in leukaemia and in solid tumours such as breast, brain, prostate, colon, pancreatic, ovarian, lung and skin cancer (Batlle et al., 2017) (Nassar et al., 2016).

In order to determine the clinical impact of CSCs, different approaches are being followed in clinical research. The most common strategy consists in isolating the CSCs and studying the expression of CSC markers (so called *stemness* biomarkers) to correlate it with clinical endpoints. Also, as CSCs have the intrinsic property of being resistant to chemotherapy, another strategy is to isolate CSCs and evaluate chemoresistance to current treatment regimens (Batlle et al., 2017).

A recent systematic review of 234 survival analysis extracted from 164 publications (Lathia et al., 2019) reported that high expression of CSC biomarker(s) resulted in poor overall survival (OS) and/or disease-free survival (DFS) compared with low or absence of expression in a wide group of cancer subtypes including breast cancer (BC). In general, an elevated stemness biomarker expression was found to be associated with clinicopathological parameters such as decreased tumour differentiation, increased TNM stage, vascular invasion, depth of tumour invasion, lymph node and distant metastasis.

Regarding studies where chemoresistance has been evaluated, it has been shown that treatment with oxaliplatin in colorectal cancer cell lines selectively favoured survival of dormant clones that became dominant after therapy (Kreso et al., 2012). Also, resistance to temozolomide has been detected in CSCs of mouse models of glioblastoma whereas ablation of CSCs renders this type of tumour sensitive to chemotherapy (Chen et al., 2012). In bladder cancer xenografts, chemotherapy (gemcitabine plus cisplatin) has been found to reactivate quiescent CSCs, repopulating the tumour after treatment (Kurtova et al., 2015). Resistance to chemotherapy (including cisplatin and vemurafenib) has also been detected in slow-cycling melanoma cells (Roesch et al., 2013).

2.1.1 Breast cancer stem cells (CSCs)

Breast cancer stem cells (BCSCs) were initially discovered in 2003 (Al-Hajj et al., 2003). In this study authors demonstrated that a few hundred cells were able to sustain growth when injected into mammary fat pads of non-obese diabetic severe combined immunodeficient (NOD/SCID) immunocompromised mice. Since then, many studies have confirmed the relationship between BCSCs and poor prognosis.

Another recent study revealed an association between BCSCs and relapse-free survival (RFS) in patients with early-stage breast invasive ductal carcinoma (BIDC) (Qiu et al., 2019). Levels of selected BCSCs markers (ALDH1A3, CD44+/CD24−, integrin alpha 6 (ITGA6), and protein C receptor (PROCR)) were measured using immunohistochemistry and intensity of the staining was used to determine high and low-risk groups of patients. Results showed that the proportion of patients in the low-risk group who were free of relapse at 5 years was significantly higher than that in the high-risk group.

Another approach to evaluate the clinical impact of CSCs is to study the association between the presence of mutations and prognosis. For example, in a recent work four mutations in genes known to be associated with BCSCs were studied by analysing circulating free DNA (cfDNA) extracted from plasma or serum (Liu et al., 2019). The results showed a statistically worse median time-to-metastasis (TTM) in patients with any of the four BCSC mutations.

There is growing evidence of the clinical impact of BCSCs mediated by chemoresistance mechanisms including the overexpression of ATP-binding cassette (ABC) transporters, increased ALDH activity, enhanced DNA repair mechanisms, reinforced reactive oxygen species (ROS) scavenging, cell death escape, induction of dormancy, autophagy, and possibly other resistance mechanisms that are yet to be characterized (De Angelis et al., 2019). For example, BCSCs isolated from breast cancer cell lines were found to be resistant to mitoxantrone in a mechanism thought to be mediated by ABCG2, an ABC transporter (Britton et al., 2012).

In more recent studies, they have described the resistance of BCSCs to the most commonly used agents to treat triple-negative breast cancer (TNBC): paclitaxel and doxorubicin. In one study, proliferation of BCSCs isolated from TNBC patients was inhibited after addition of sublethal doses of doxorubicin and paclitaxel, although 20-40% of cells survived the treatment. These cells, cultured in medium without chemotherapeutics, recovered gradually confirming an upregulated self-renewal capacity under chemotherapeutic stress (Li et al., 2020).

2.1.2 Mantle cell lymphoma cancer stem cells (MCL-CSCs)

In comparison to BCSCs, little is known about CSCs in MCL. In 2010, it was the first time that clonogenic cells with self-renewal capacities from MCL were isolated (Chen et al., 2010). The cells, called MCL-initiating cells (MCL-ICs), were obtained from stage 4 MCL patients. The authors observed that MCL-ICs lacked expression of the prototypic B cell surface marker CD19 and were able to recapitulate the heterogeneity of the original patient tumour upon transplantation into immunodeficient mice.

The same research group confirmed the existence of chemoresistance happening in MCL-ICs. In 2011, they observed that IC50 values were significantly higher in CD45+/CD19- MCL-ICs than in CD45+/CD19+ cells for most of the chemotherapeutic regimens tested (Jung et al., 2011). More specifically, in all patient samples, more than double the concentration of each drug agent in R-CHOP, R-CAVD, and fludarabine-based regimens were required to inhibit 50% growth of CD45+/CD19- MCL-ICs compared to CD45+/CD19+ MCL cells. Authors identified that resistance was mediated by ABCB1 transporter whose inhibition increased the sensitivity of MCL-ICs to vincristine.

In 2012, the same group described that MCL-ICs were also resistant to bortezomib as a single agent or administered within a chemotherapeutic regimen (Jung et al., 2012). Resistance to Bortezomib by MCL-ICs was again reported in 2018 by another group (Luanpitpong et al., 2018). In this study authors observed that sensitivity to Bortezomib was modulated by reactive oxygen species (ROS) and identified two key players in that modulation: the anti-apoptotic Mcl-1 and the transcription factor Zeb-1.

The association of the combination of CD45+ and CD19- with prognosis was also evaluated in another study, in which the CD45+/CD19- cell population percentage correlated with MCL prognostic index (Kim et al., 2015).

2.2. Identification of CSC biomarkers

There is controversy about the ideal methodology for reliable measurement of biomarker due to the fact that CSCs are a very rare population of cells. Moreover, there is no standardized protocols or tests for assessing presence and levels of CSC biomarkers in tumours (Lathia et al., 2019). However, multiple pathways and markers related to CSCs have been identified for a variety of cancer types.

Three main CSC signalling pathways related to self-renewal and differentiation have been identified: Notch, Wnt/beta-catenin and Hedgehog (Hh). Other important signalling pathways are the TNF-α/NF-κ-β, transforming factor-β (TGF-β), receptor tyrosine kinase RTK and Janus kinase/signal transducer and activator of transcription (JAK-STAT) pathways (Palomeras et. al, 2018).

A considerable number of CSCs markers have been identified up to date allowing the development of new therapeutic strategies to target CSCs. Those targets include tumour microenvironment, signalling pathways, stem cell differentiation, cell surface markers, apoptotic pathways, drug resistance markers and microRNAs (Prasad et al., 2019). In this recent review, an extensive list of CSC biomarkers is shown. Just to cite a few examples, CD133, a cell surface molecule, is considered a CSC marker in glioblastoma

(Chen et al., 2010) and colorectal cancer (O'Brien et al., 2007) whereas CD34 expression has been related to increased self-renewal potential in skin squamous cell carcinoma (Lapouge et al., 2012). In head and neck cancer (HNC) CD44, ALDH1, CD133, Oct3/4, Nanog and Sox2 have been considered as CSC-associated molecules (Yu et al., 2020).

2.2.1. BCSC biomarkers

More than a decade has passed since the identification of the first biomarkers associated with BCSCs, CD44+/CD24low and CD133+ (Wright et al., 2008). Along these years a considerable number of biomarkers have been added to the list. The most recent studies are summarised hereafter:

In the systematic review discussed earlier, CD44 appeared to be consistently associated with poor survival in BC. The combination of CD44+/CD24- has been associated with poor OS, DFS and/or progression-free survival (PFS) in six of the studies (Lathia et al., 2019). In another study, levels of selected BCSCs markers (ALDH1A3, CD44+/CD24−, integrin alpha 6 (ITGA6), and protein C receptor (PROCR) were measured using immunohistochemistry and intensity of the staining was used to between high and low-risk groups of patients with early-stage breast invasive ductal carcinoma (BIDC) (Qiu et al., 2019).

Using bioinformatic tools, a recent study identified 32 key genes that modulate BC stemness characteristics and, among them, 12 genes strongly correlated with BC survival: TPX2, EXO1, CCNB2, CENPA, SGO1, RAD54L, SKA1, FOXM1, PLK1, CDC20, KIF4A and SGO1 (Pei at al., 2020). Another gene, TRIP6, has been recently associated with CSC-like properties and poor prognosis in BC (Zhao et al., 2020) as well as a 20-gene stem cell signature obtained from the transcriptional profile of normal mammary stem cells (Pece et al., 2019). This gene panel was able to predict early and late recurrence in triple negative and luminal BC.

In another study, hundreds of genes that regulate BCSC fate were identified using a genome-wide RNAi screen in a breast cancer cell line (Arfaoui et al., 2019). Those genes were then integrated in a functional mapping of the CSC-related processes uncovering potential therapeutic targets. Among 15 compounds tested, mifepristone, salinomycin and JQ1 showed the best anti-BCSC activity. Regarding chemoresistance, it has been described that silencing SOX2, a gene related to pluripotency and stemness, lead to an increased chemosensitivity to paclitaxel in BCSCs isolated from TNBC patients in vitro (Mukherjee et al., 2017).

2.2.2. MCL-CSC biomarkers

Compared to BC, a small number of markers have been identified in CSCs of MCL. We have already mentioned the identification of the combination of CD45+ and CD19- markers in MCL-initiating cells isolated from patient blood samples. This combination of markers was associated first with chemoresistance (Jung et al., 2011), (Jung et al., 2012) and later with poor prognosis (Kim et al., 2015).

2.3. Data science and cancer research

In recent years, in parallel to the growing data complexity and size, the fields of bioinformatics and machine learning have seen dramatic advances. Their application in the biomedicine field is becoming ever more popular with the goal to support research and healthcare by translating patient data to successful therapies.

2.3.1 Data science and cancer stem cells

One of the most important studies using data science techniques related to the topic covered in this work is the identification of stemness features associated with oncogenic dedifferentiation (Malta et al., 2018). One-class logistic regression algorithm (OCLR) was used to extract transcriptomic and epigenetic feature sets derived from non-transformed pluripotent stem cells and their differentiated progeny. Authors used publicly available molecular profiles from normal cell types that exhibit various degrees of stemness and developed a model using One-class logistic regression algorithm (OCLR). As a result, two independent stemness indices were generated, one reflective of epigenetic features (mDNAsi) and the other of gene expression (mRNAsi). The indices were then associated with novel oncogenic pathways, somatic alterations, and microRNA and transcriptional regulatory networks. Results showed that higher indices were associated with biological processes active in cancer stem cells, with greater tumour dedifferentiation and pathology grading for the majority of the Cancer Genome Atlas (TCGA) cases.

Authors also used GSEA to compare the mRNAsi index with 16 genes sets that were associated with stemness in cancer and healthy cells in previous studies. In all cases, they found that the published stemness gene sets were significantly enriched in mRNAsi. Moreover, compounds specific to selected molecular targets and mechanisms that may eventually lead to novel treatments were identified. Using the mRNAsi index, another group identified 32 key genes that modulate BC stemness characteristics and, among them, 12 genes strongly correlated with BC survival (Pei at al., 2020).

2.3.2 GSEA and breast cancer

GSEA is one of the most important data science tools toward establishing a link between molecular features and phenotypes. It has been extensively used to study differences between tumour and normal samples or between different tumour subtypes.

In order to understand the implication of transcription factors (TFs) in breast cancer, a study was performed using 14 breast cancer gene expression datasets from the public functional genomic repository NCBI-GEO (Li et al., 2017). Among the 22 up-regulated pathways identified by GSEA, the most relevant were cell cycle, DNA replication, spliceosome, proteasomes, mismatch repair, p53 signalling pathway and nucleotide excision repair. Among the 25 down-regulated pathways, the most relevant were fatty acid metabolism, adipocytokine signalling and valine, leucine and isoleucine degradation.

In a recent study, authors used differential gene expression together with context data with the aim of identifying specific drug targets for the basal-like type BC (Parks et al., 2019). For that, they used RNA-seq data from the Breast Invasive Carcinoma (BRCA) dataset of the TCGA repository and generated a regulatory module enrichment score (RMES) using algorithms specific for gene regulatory networks such as GRNBoost2, and single sample gene set enrichment analysis (ssGSEA). Then, RMESs were used as features for ~~ML (~~Machine Learning (ML) processing using Support Vector Machine algorithm (SVM) in order to perform multiclass classification of samples. Results showed an accuracy score of 99.07% in basal-like BC classification.

GSEA has been traditionally used taking into account one single molecular feature as the score of each gene need to be a scalar. Either only one molecular feature is analysed or information coming for multiple features (i.e. DNA sequences mutations, mRNA transcripts, CNVs, single nucleotide polymorphisms or DNA methylations) is synthesized into one single score prior the enrichment analysis. In order to extend GSEA to multiplatform data, a new method called Multivariate Gene Enrichment Analysis (MGSEA) was recently developed (Tiong et al., 2019). Data from three molecular features, mRNA expression, CNV and DNA methylation, were retrieved from TCGA with the aim of finding functional categories of genes related to BC and glioblastoma subtypes. A combined gene score integrating the three molecular features was generated. Results showed that mRNA expression appeared more frequently as a dominant feature than CNV or DNA methylation in both cancer subtypes. In BC, mRNA expression was the only dominant feature in functional categories involved in cell proliferation such as cell cycle control, estrogenic response, DNA repair, MYC targets and E2F targets whereas CNV was the only dominant feature in functional categories involved in invasion and metastasis, such as cell adhesion and EMT. In glioblastoma,

mRNA expression was the only dominant feature in diverse functional features such as cell adhesion, inflammatory response, angiogenesis and EMT.

2.3.3 GSEA and mantle cell lymphoma

There is only one publication in which GSEA has been described for MCL, and is related to the study of molecular subsets of MCL defined by the IGHV mutational status and SOX11 expression (Navarro et al., 2012) . A GSEA was performed on 38 MCL patient samples, divided in mutated (M-MCL) and unmutated (U-MCL) depending on the presence of IGHV-IGHD-IGHJ rearrangement and SOX11 expression. Four specific gene sets related to normal B-cell subtypes were used for the analysis. The results showed that SOX11-positive U-MCL expressed a signature enriched in genes related to naïve B-cells whereas SOX11-negative M-MCL had a signature related to memory B-cells.

3. Methodology

3.1. Data collection

Data used in this project have been retrieved from scientific articles obtained through PubMed NCBI browser (https://www.ncbi.nlm.nih.gov/pubmed/) and UOC online library (http://biblioteca.uoc.edu/es/), from gene expression datasets obtained through Gene Expression Omnibus repository (GEO) (https://www.ncbi.nlm.nih.gov/geo/) and from gene sets obtained from the Molecular Signatures Database (MSigDB) of the Broad Institute (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp).

3.1.1 Cancer stem cell gene sets (CSC gene sets)

Gene sets involved in CSC machinery have been generated using three different methods:

1. Identification of CSC genes through revision of the bibliography listed in the **State of the art** section of this work. Five gene sets have been generated (list of genes can be found in section 5 of Appendix).

    - "*Prognosis_BC*": CSC genes related to prognosis in BC. Sources: (Lathia et al., 2019), (Qiu et al., 2019), (Liu et al., 2019), (De Angelis et al., 2019), (Pei at al., 2020), (Zhao et al., 2020), (Pece et al., 2019).

    - "*Stemness_BC*": CSC genes related to stemness in BC. Sources: (De Angelis et al., 2019), (Pei at al., 2020).

    - "*Stemness2_BC*": CSC genes related to stemness in BC. Sources: (Arfaoui et al., 2019).

    - "*Dormancy_BC*": CSC genes related to dormancy in BC. Sources (Kim et al., 2015), (De Angelis et al., 2019).

    - "*Chemorresistance_BC*": CSC genes related to chemoresistance in BC. Sources: (De Angelis et al., 2019), (Prasad et al., 2019)

2. Identification of CSC genes through assessment of differential gene expression in 4 datasets obtained from GEO NCBI browser. GEO datasets used were GSE25976, GSE43730, GSE95402 and GSE132083. GSEA was performed to identify the top 50 significantly over and under-expressed genes. The resulting 8 gene sets generated were:

- "*GSE25976_OVER*": 50 top overexpressed genes selected from GSEA performed on metastatic and non-metastatic BCSCs.

- "*GSE25976_UNDER*": 50 top underexpressed genes selected from GSEA performed on metastatic and non-metastatic BCSCs.

- "*GSE43730_OVER*": 50 top overexpressed genes selected from GSEA performed on malignant and non-malignant BC cells.

- "*GSE43730_UNDER*": 50 top underexpressed genes selected from GSEA performed on malignant and non-malignant BC cells.
- "*GSE95042_OVER*": 50 top overexpressed genes selected from GSEA performed on BCSCs and primary BC.

- "*GSE95042_UNDER*": 50 top underexpressed genes selected from GSEA performed on BCSCs and primary BC.

- "*GSE132083_OVER*": 50 top overexpressed genes selected from GSEA performed on BCSCs and non-BCSCs.

- "*GSE132083_UNDER*": 50 top underexpressed genes selected from GSEA performed on BCSCs and non-BCSCs.

3. Identification of CSC genes involved in pathways known to play a role in CSCs as reviewed in bibliography and mentioned in **State of the art** section of the present work. The list of genes involved in the following pathways were collected from the MSigDB database from the Broad Institute:

    - KEGG_HEDGEHOG_SIGNALING_PATHWAY (Ref.: M1053)
    - REACTOME_SIGNALING_BY_HIPPO (Ref.: M591)
    - KEGG_JAK_STAT_SIGNALING_PATHWAY (Ref.: M17411)
    - PID_MYC_PATHWAY (Ref.: M139)
    - KEGG_NOTCH_SIGNALING_PATHWAY (Ref.: M7946)
    - KEGG_TGF_BETA_SIGNALING_PATHWAY (Ref.: M2642)
    - TNF (Ref.: M128)
    - HALLMARK_WNT_BETA_CATENIN_SIGNALING (Ref.: M5895)

    The resulting gene sets generated were: "*Hedhehog*", "*Hippo*", "*Jak_Stat*", "*Myc*", "*Notch*", "*TGF_beta*", "*TNF*" and "*Wnt_Bcatenin*".

The **table 1** summarises the details of the 21 gene sets generated for the study.

Gene_set	Method	Num_genes
Prognosis_BC	Bibliography	62
Stemness_BC	Bibliography	53
Stemness2_BC	Bibliography	332
Dormancy_BC	Bibliography	51
Chemorresistance_BC	Bibliography	21
GSE25976_OVER	GSEA	50
GSE25976_UNDER	GSEA	50
GSE43730_OVER	GSEA	50
GSE43730_UNDER	GSEA	50
GSE95042_OVER	GSEA	50
GSE95042_UNDER	GSEA	50
GSE132083_OVER	GSEA	50
GSE132083_UNDER	GSEA	50
Hedgehog	MSigDB	56
Hippo	MSigDB	20
Jak_Stat	MSigDB	155
Myc	MSigDB	25
Notch	MSigDB	47
TGF_beta	MSigDB	85
TNF	MSigDB	46
Wnt_Bcatenin	MSigDB	42

Table 1. Description of the gene sets used in the study

3.1.2 Breast cancer expression datasets (BC gene expression datasets)

Expression datasets were collected from Gene Expression Omnibus repository (GEO). Preference was given to studies performed on clinical samples. The following GEO datasets were used in the study:

- GSE5764: Invasive breast cancer tissue. 20 Tumoral samples (10 ductal, 10 lobular) + 10 Normal samples. Array: Affymetrix Human Genome U133 Plus 2.0 Array.

- GSE6883: Breast cancer tissue. 3 Tumoral samples + 3 Normal samples. Array: Affymetrix Human Genome U133A Array.

- GSE73540: Breast cancer tissue. 3 Tumoral samples + 3 Normal samples. Array: Affymetrix Human Genome U133A Array.

- GSE92252: Breast cancer tissue. 6 Tumoral samples + 3 Normal samples. Array: NimbleGen Homo sapiens Expression Array.

- GSE71862: Breast cancer cell lines. 3 Breast cancer cell line derived from metastatic site (MCF7) + 3 Normal-like mammary epithelial cell line (MCF10A). Pre-Ranked GSEA. Array: NimbleGen Homo sapiens Expression Array.

- GSE109169: Early-onset breast cancer tissue. 25 Tumoral samples + 25 matched normal tissue. Array: Affymetrix Human Exon 1.0 ST Array.

The **table 2** summarises the details of the 6 BC datasets collected for the study.

Dataset	Type_Sample	Number_Sample	Array
GSE5764	Tissue	20 Tumoral/10 Normal	Affymetrix Human Genome U133 Plus 2.0 Array
GSE6883	Tissue	3 Tumoral/3 Normal	Affymetrix Human Genome U133A Array
GSE73540	Tissue	28 Tumoral/3 Normal	Affymetrix Human Transcriptome Array 2.0
GSE92252	Tissue	6 Tumoral/3 Normal	NimbleGen Homo sapiens Expression Array
GSE71862	Cell lines	3 BC cell line/3 Normal cell line	NimbleGen Homo sapiens Expression Array
GSE109169	Tissue	25 Tumoral/25 Normal	Affymetrix Human Exon 1.0 ST Array

Table 2. Description of the BC datasets used in the study

3.1.3 Mantle cell lymphoma expression datasets (MCL gene expression datasets)

Expression datasets were collected from Gene Expression Omnibus repository (GEO). Preference was given to studies performed on clinical samples. The following GEO datasets were used in the study:

- GSE30189 (classical form): MCL tumor cells. 6 MCL (classical form) + 4 normal mantle zone B-lymphocytes. Array**:** Illumina HumanWG-6 v3.0 expression beadchip.

- GSE30189 (aggressive form): MCL tumor cells. 7 MCL (aggressive form) + 4 normal mantle zone B-lymphocytes. Array: Illumina HumanWG-6 v3.0 expression beadchip.

- GSE45717: MCL tumor cells. 5 MCL + 8 healthy B-lymphocytes. Array: Affymetrix Human Exon 1.0 ST Array.

- GSE60023: MCL tumor cells. 3 MCL + 2 (CD19+) B-lymphocytes from healthy donor. Array: Arraystar Human LncRNA microarray V2.1.

- GSE95291: MCL tumor cells.2 MCL + 2 B-lymphocytes from healthy donor. Array: Illumina HumanHT-12 V4.0 expression beadchip.

- GSE21452: MCL tumor cells. 64 MCL with external control. Array**:** Affymetrix Human Genome U133 Plus 2.0 Array

The **table 3** summarises the details of the 6 MCL datasets collected for the study.

Dataset	Type_Sample	Number_Sample	Array
GSE30189_classical	Cells	6 Tumoral/4 Normal	Illumina HumanWG-6 v3.0 expression beadchip
GSE30189_aggresive	Cells	7 Tumoral/4 Normal	Illumina HumanWG-6 v3.0 expression beadchip
GSE45717	Cells	5 Tumoral/8 Normal	Affymetrix Human Exon 1.0 ST Array
GSE60023	Cells	3 Tumoral/2 Normal	Arraystar Human LncRNA microarray V2.1
GSE95291	Cells	2 Tumoral/2 Normal	Illumina HumanHT-12 V4.0 expression beadchip
GSE21452	Cells	64 Tumoral	Affymetrix Human Genome U133 Plus 2.0 Array

Table 3. Description of the MCL datasets used in the study

3.2. Data analysis

3.2.1 GSEA

To identify CSC biomarkers in BC and MCL we used GSEA computational method (GSEA-P Software) created by the Broad Institute and described previously (Subramanian et al 2005). GSEA 4.0.3. desktop version was installed in the computer following instructions. We analyzed above described BC and MCL gene expression datasets to determine whether these are significantly enriched by the genes present in the stemness gene signatures.

The steps followed to perform GSEA are the following:

1) Loading expression dataset (*.gct* format)

   Expression datasets were prepared using a Gene Cluster Text (GCT) format (.gct) that describes an expression dataset. GCT is convenient for analysis of matrix-compatible datasets as it allows metadata about an experiment to be stored alongside the data from the experiment. GCT files enable storing both row and column metadata. Typically, each column represents a specific experiment and each row represents features that are measured in the assay.

2) Preparation and loading of chip annotations (*.chip* format)

   The Chip description file (.chip) contains annotations about a microarray. The file typically specifies which probes map to the same genomic unit of interest. While this file is not used directly in the GSEA algorithm, it is used to annotate the output results and may also be used to collapse each probe set in the expression dataset to a single gene vector.

   In the majority of cases, information for chip annotation was found stored in the program. That was the case of the following chips: Affymetrix Human Genome U133A Array, Affymetrix Human Transcriptome Array 2.0 and Illumina

HumanHT-12 V4.0 expression beadchip. For the rest of chips, annotation with external sources was performed. That was the case of the following chips: Affymetrix Human Genome U133 Plus 2.0 Array, NimbleGen Homo sapiens Expression Array, Affymetrix Human Exon 1.0 ST Array, Illumina HumanWG-6 v3.0 expression beadchip, Arraystar Human LncRNA microarray V2.1.

3) Creation and loading of phenotype labels (*.cls* format)

   The CLS file format (.cls) defines phenotype (class or template) labels and associates each sample in the expression data with a label. Categorical labels define discrete phenotypes (e.g. normal vs tumor).

4) Loading of gene sets (*.gmt* format)

   The Gene set file format (.grp) contains a single gene set in a simple newline-delimited text format while GMT or GMX file formats are used to create multiple gene sets in the same file. In the present study, the 21 CSC gene sets were placed on a GMT format file to be loaded and processed.

5) Running analysis

   The top (over-expressed) and bottom (under-expressed) of the list in the datasets correspond to the largest differences in expression between tumoral and normal tissue. GSEA calculates the enrichment score (ES) that represents the amount to which the genes in the set are over-represented at either the top or bottom of the list. The ES is the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov–Smirnov-like statistic. The program also estimates the statistical significance (p value) of the ES by calculating a phenotypic-based permutation test to produce a null distribution for the ES and adjusts the estimated significance level to account for multiple hypothesis testing. The enrichment scores for each gene signature is normalized and a false discovery rate is calculated together with a normalized enrichment score (NES). The proportion of false positives is checked by calculating the false discovery rate (FDR).

3.2.2 Visualization analysis (Tableau)

Tableau software (desktop edition) has been used to summarize and enhance interpretability of GSEA results by creating some figures (1-3) and tables (1-12) included in this work.

3.2.3 Functional annotation analysis

Panther Classification System (Huaiyu et al., 2019) has been used to perform functional annotation analysis of the genes included in the 269-CSC gene signature. Panther combines gene function, ontology, pathways and statistical tools to enable large-scale analysis. In this work, gene list analysis has been performed to group the genes as per their molecular function and biological process. Grouping by molecular function consists on classifying the genes by the function of the protein itself or the proteins that interact directly with it at a biochemical level, whereas grouping by biological process consists on classifying the genes by the function of the protein in the context of a larger network of proteins that interact to accomplish a process at the level of the cell or organism.

A high-level analysis was first performed with the 269 genes for both, molecular function and biological process analysis. Then, a deeper analysis of the major categories found in the high-level analysis was subsequently performed.

3.2.4 Kaplan-Meier survival analysis

Correlation with prognostic parameters has been performed using a bioinformatic tool called Kaplan-Meier Plotter (Nagy et al., 2018). This tool is capable to assess the effect of genes on survival in multiple cancer types. Sources used include GEO and TCGA, among others. For this analysis, the genes included in the 53-CSC gene signature were correlated with survival in breast, ovarian, lung and gastric cancer. Only genes found to be significant were further studied. The following parameters were used:

1) Hazard Ratio: ratio of the hazard rates corresponding to high and low expression of an individual gene.

2) Median survival: length of time from either the date of diagnosis or the start of treatment for a disease, that half of the patients are still alive.

3) Logrank p-value: significance given by a hypothesis test comparing the survival distributions of high and low expression of an individual gene.

4) Expression state given by the GSEA results found for that individual gene (up: upregulated, down: downregulated)

3.2.5 Correlation analysis

Correlation analysis of the genes of the CSC-gene signature with prognostic parameters was performed using cor() function of Stats R package and plots of matrix correlations

were generated using corrplot R package. Pearson coefficient was used as the test statistic.

3.2.6 Machine learning

Random forest, a supervised classification algorithm, was built using the Caret R package. *Random forest* is a learning algorithm that generates multiple decision trees and, in case of classification, outputs the classes of the individual trees. The predicted class of the input instance is decided upon majority vote.

Model was trained as follows: data was randomly split into training and test sets using different ratios: 70%/30%, 65%/35% and 60%/40% (training/test). Cross-validation was used instead of Out-Of-Bag bootstrap method. Different values for n-fold and repeats were tested (5-fold with 3 repeats and 3-fold with no repeats). Default parameters for ntrees (value of 500), mtry (square root of the total number of variables), maxnodes (trees are grown to maximum possible) and nodesize (value of 1) were used. For the evaluation in the test set, accuracy, sensitivity and specificity were measured. Variable importance regarding the mean decrease in accuracy for each predictor was used.

4. Results

4.1. GSEA analysis

GSEA was performed on 6 BC (see section 3.1.2 for more details) and 6 MCL (see section 3.1.3 for more details) expression datasets.

4.1.1 Breast Cancer GSEA

Selected BC datasets were analyzed using GSEA in order to identify genes involved in CSC machinery that are enriched in a tumoral state, especially in a context where CSCs may have a key role such as in metastatic or invasive phenotypes, as reviewed in the introduction of the present work.

A total of 6 BC datasets were tested for gene set enrichment using 21 gene sets specially generated for being involved in CSC machinery (for more details on the gene set generation, check **Methodology** section).

The results show an enrichment of CSC genes in all datasets studied (**¡Error! No se encuentra el origen de la referencia.**). A total of 331 genes were found to be downregulated whereas 184 genes were found to be upregulated in tumoral compared to normal state. The dataset GSE5764, corresponding to invasive BC, was the one showing a higher amount of upregulated CSC genes (95 genes) whereas GSE73540, corresponding to primary BC was the dataset showing higher amount of downregulated CSC genes (130 genes).

	Dataset					
Up/Downregulated	GSE5764	GSE6883	GSE71862	GSE73540	GSE92252	GSE109169
DOWN	35	29	85	130	52	
UP	95	9	19	16		45

Table 4. GSEA for BC datasets (count of up/downregulated genes)

A deeper analysis on which specific gene sets were contributing to the enrichment in each dataset was performed (**¡Error! No se encuentra el origen de la referencia.**).

Dataset	Gene Set	
GSE5764	STEMNESS_BC	26
	PROGNOSIS_BC	25
	GSE95042_UNDER	25
	GSE95042_OVER	25
	TGF_BETA	17
	MYC	10
	HIPPO	10
GSE6883	STEMNESS_BC	29
	GSE132083_UNDER	9
GSE71862	GSE43730_UNDER	23
	GSE132083_OVER	19
	GSE95042_OVER	16
	GSE43730_OVER	14
	GSE25976_OVER	13
	DORMANCY_BC	12
	GSE25976_UNDER	10
GSE73540	STEMNESS2_BC	58
	GSE95042_OVER	35
	STEMNESS_BC	27
	PROGNOSIS_BC	21
	GSE43730_UNDER	16
GSE92252	GSE95042_OVER	32
	TNF	20
GSE109169	STEMNESS_BC	24
	PROGNOSIS_BC	18
	DORMANCY_BC	15

Up/Downregulated
DOWN
UP

Table 5. GSEA results for BC datasets

We observed that upregulated genes in GSE5764, which corresponds to an invasive BC, were distributed among 5 different CSC gene sets: “PROGNOSIS_BC”, “STEMNESS_BC”, “GSE95042_OVER”, “MYC” and “TGF_BETA”. All of them, except “TGF_BETA”, were found to be highly significant (p-value < 0.05) (**¡Error! No se encuentra el origen de la referencia.**).These results pointed to the existence of an enriched CSC phenotype in that particular dataset.

Similar results were found when analysing GSE109169 dataset, which corresponds to samples of early-onset BC. The genes that were found to be significantly upregulated came from 3 different CSC gene sets: “STEMNESS_BC”, “PROGNOSIS_BC” and “DORMANCY_BC” (***Table 5***). Moreover, the analysis of the significance suggests that samples of the dataset display important CSC features such as stemness or dormancy (**¡Error! No se encuentra el origen de la referencia.**).

In the opposite side, GSE71862 dataset, which corresponds to a breast cancer line derived from a metastatic site, had its downregulated genes distributed across 6 different CSC gene sets: “GSE43730_UNDER”, “GSE95042_OVER”, “GSE43730_OVER”, “DORMANCY_BC”, “GSE25976_UNDER”, “GSE25976_OVER”. These results show that GSE71862 doesn’t display a clear CSC phenotype as the majority of the downregulated genes correspond to gene sets generated from BCSCs (GSE95042_OVER, GSE43730_OVER) and even metastatic BCSCs (GSE25976_OVER). It’s worth mentioning that significance of 3 of the gene sets (“DORMANCY_BC”, “GSE25976_OVER” and “GSE95042_OVER”) are among the 5 lowest found across datasets being the p-values close to 0.1 (***Figure 3*****¡Error! No se encuentra el origen de la referencia.**).

Similar results of an apparent absence of a CSC phenotype was found for GSE73540 dataset, corresponding to primary BC, in which downregulated genes came from 4 distinct CSC gene sets: “STEMNESS2_BC”, “GSE95042_OVER”, “STEMNESS_BC” and “PROGNOSIS_BC”. Again, as mentioned for GSE71862 dataset, the significance of 2 of the 4 gene sets was among the 5 lowest, being the p-values close to 0.1 (***Figure 3***).

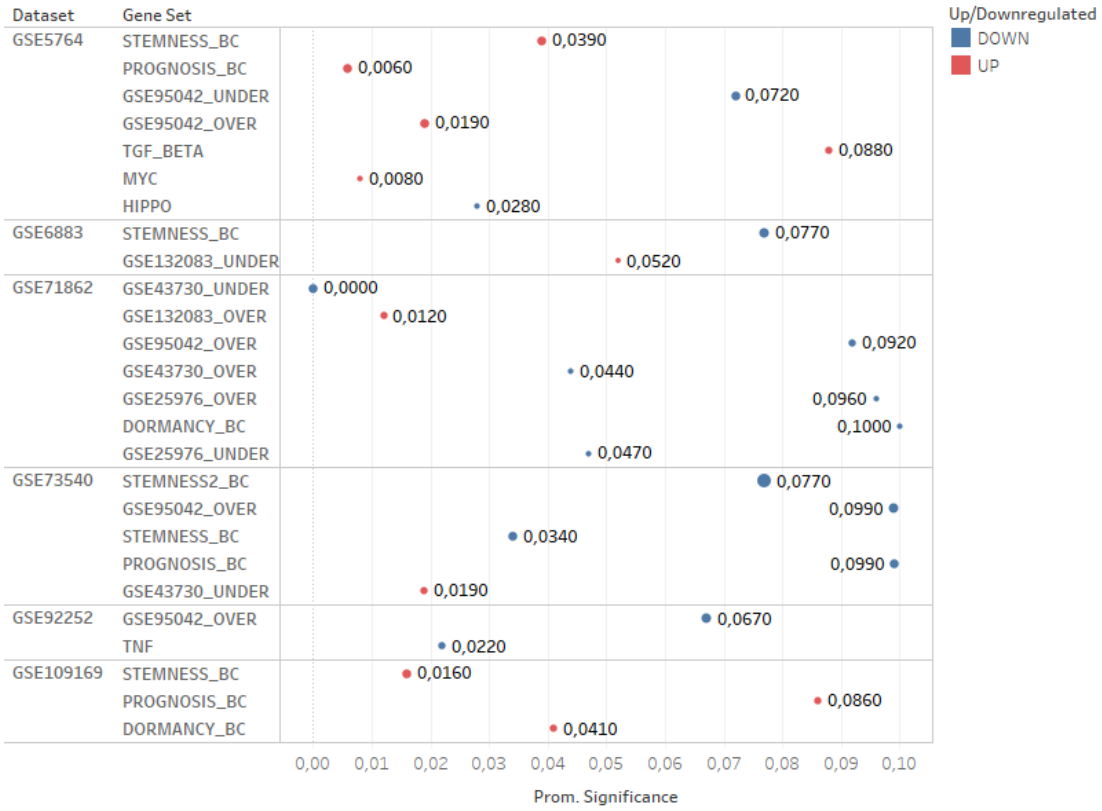


Figure 3. GSEA for BC datasets. Average significance of gene sets.

4.1.2 Mantle Cell Lymphoma GSEA

A total of 6 MCL datasets were tested for gene set enrichment.

The results show an enrichment of CSC genes in all datasets studied (***Table 6***). A total of 259 genes were found to be downregulated whereas 698 genes were found to be upregulated in tumoral compared to normal state. The dataset GSE21452, corresponding to 64 primary MCL tumors, was the one showing a higher amount of upregulated CSC genes (418 genes) whereas GSE95291, corresponding to 2 primary MCL cells was the dataset showing higher amount of downregulated CSC genes (165 genes). Except two datasets ("GSE30189_aggresive", "GSE30189_classical"), the other four displayed a higher proportion of upregulated genes compared to downregulated ones.

	Dataset					
Up/Dow..	GSE21452	GSE30189_agg	GSE30189_class	GSE45717	GSE60023	GSE95291
DOWN	58	8	28			165
UP	418			69	28	183

Table 6. GSEA for MCL datasets (count of up/downregulated genes)

A deeper analysis on which specific gene sets were contributing to the enrichment in each dataset was performed (***Table 7***).

Dataset	Gene Set	
GSE21452	STEMNESS2_BC	137
	NOTCH	42
	STEMNESS_BC	38
	GSE95042_OVER	38
	TNF	37
	PROGNOSIS_BC	34
	WNT_BCATENIN	32
	TGF_BETA	31
	DORMANCY_BC	31
	GSE25976_UNDER	30
	GSE43730_UNDER	29
	MYC	24
	HIPPO	18
GSE30189_agg	GSE132083_OVER	8
GSE30189_class	DORMANCY_BC	21
	GSE132083_OVER	7
GSE45717	STEMNESS2_BC	44
	DORMANCY_BC	13
	GSE95042_OVER	12
GSE60023	STEMNESS2_BC	13
	PROGNOSIS_BC	9
	JAK_STAT	6
GSE95291	STEMNESS2_BC	81
	HEDGEHOG	44
	GSE25976_UNDER	31
	GSE25976_OVER	31
	GSE43730_OVER	30
	STEMNESS_BC	28
	PROGNOSIS_BC	25
	TNF	23
	GSE95042_OVER	20
	GSE95042_UNDER	18
	DORMANCY_BC	18
	MYC	11

Up/Downregulated
DOWN
UP

Table 7. GSEA results for MCL datasets

We observed that upregulated genes in GSE21452 were distributed among 11 different CSC gene sets that were (ordered in descendent order as per the number of genes involved): “STEMNESS2_BC”, “NOTCH”, “STEMNESS_BC”, “GSE95042_OVER”, “TNF”, “PROGNOSIS_BC”, “WNT_BCATENIN”, “TGF_BETA”, “DORMANCY_BC”, “MYC” and “HIPPO”,. It seems that 6 out of 8 gene sets involving CSC pathways were significantly enriched in this particular dataset. All of the gene sets, except “TGF_BETA” were found to be highly significant (p-value < 0.05) (***Figure 4***). These results pointed to the existence of a particularly enriched CSC phenotype in that particular dataset.

Mixed results were obtained in dataset GSE95291, in which a similar proportion of upregulated and downregulated genes was found. Upregulated genes came from 6 gene sets: “STEMNESS_BC”, “STEMNESS2_BC”, “PROGNOSIS_BC”, “DORMANCY_BC”, “GSE25976_UNDER” and “MYC” whereas downregulated genes came from 6 gene sets: “GSE95042_OVER”, “TNF”, “GSE43730_OVER”, “GSE25976_OVER”, “HEDGEHOG” and “GSE95042_UNDER” (***Table 7***). Moreover, these results appear to be highly significant with p-values lower than 0.01. It’s worth mentioning that a group of genes found to be downregulated in the dataset were originally found to be upregulated in the BCSCs samples used to construct the corresponding gene sets (“GSE25976_OVER”, “GSE43730_OVER” and “GSE95042_OVER”).

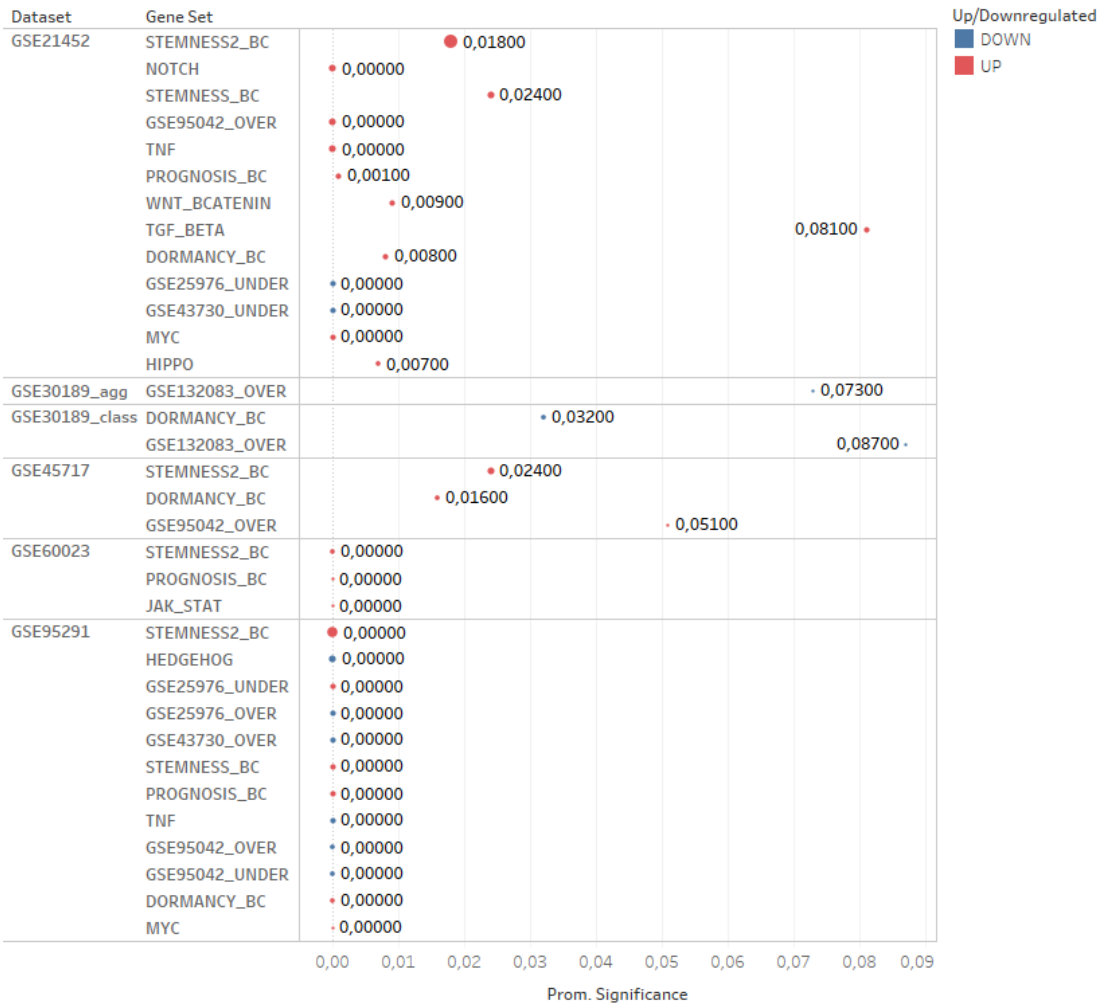


Figure 4. GSEA for MCL datasets. Average significance of gene sets.

4.1.3 Comparison of GSEA results between cancer types

A comparative analysis was performed between cancer types. The first relevant finding was the inverse proportion of up/downregulated genes found in both cancer types. While

in BC almost 62,84% of significantly enriched genes were downregulated in tumoral state, in MCL we observed the inverse: almost 74,26% of significantly enriched genes were upregulated in tumoral state (***Table** 8*, ***Table** 9*).

	Up/Downregulated	
Cancer Type	DOWN	UP
BC	266	156
MCL	252	519

Table 8. Number of up/downregulated genes comparing BC and MCL.

.

	Up/Downregulated	
Cancer Type	DOWN	UP
BC	62,84%	37,16%
MCL	25,74%	74,26%

Table 9. Percentage of up/downregulated genes comparing BC and MCL.

More granularity was added to the analysis and the proportion of genes across the different gene sets was compared (***Figure 5***). Several gene sets were identified in which the number of upregulated genes was similar when comparing cancer types. A total number of 66 and 50 genes from the gene set “STEMNESS_BC” were upregulated in MCL and BC, respectively.

Similarly, 68 and 43 genes from the gene set “PROGNOSIS_BC” were upregulated in MCL and BC, respectively. Regarding genes from “DORMANCY_BC” gene set, more than four times of genes were shown to be upregulated in MCL when compared to BC (62 and 15 genes, respectively). Also, a double number of genes from the gene set was found to be downregulated in MCL, as compared to BC (21 and 12 genes, respectively). Interestingly, genes from “STEMNESS2_BC”, “STEMNESS_BC” and “PROGNOSIS_BC” were found to be downregulated only in BC (58, 56 and 21 genes, respectively).

Regarding the gene sets generated from BCSCs, two of the ones composed by upregulated genes (“GSE43730_OVER”, “GSE25976_OVER”, respectively) were enriched both in MCL and BC, but gene expression was found to be downregulated (30 and 14 genes in “GSE43730_OVER” for MCL and BC, respectively, and 31 and 13 genes in “GSE25976_OVER” for MCL and BC, respectively). The other 2 gene sets generated form BCSCs composed by upregulated genes (“GSE95042_OVER” and “GSE132083_OVER”) lead to identification of both, upregulated an downregulated genes. Whereas for “GSE95042_OVER”, the highest proportion of downregulated genes were found for BC (83 genes compared to 20 genes for MCL), for “GSE132083_OVER” the downregulated genes (15 genes) corresponded exclusively to MCL. In the contrary, for “GSE95042_OVER” the upregulated genes were found mostly in MCL (50 genes

compared to 25 genes for BC), and for “GSE132083_OVER”, the upregulated genes were exclusive from BC (19 genes).

Regarding the gene sets composed by the downregulated counterparts, for “GSE95042_UNDER” only downregulated genes were found in both cancer types (18 and 25 genes for MCL and BC, respectively). For “GSE43730_UNDER” and “GSE25976_UNDER”, both, downregulated and upregulated genes were found, although the majority were included in the downregulated subset (30 and 10 genes in “GSE25976_UNDER” for MCL and BC, respectively; 29 and 23 genes in “GSE43730_UNDER” for MCL and BC, respectively). Regarding the upregulated genes, whereas in “GSE25976_UNDER” upregulated genes were only found for MCL (31 genes), in “GSE43730_UNDER” upregulated genes were only found in BC (16 genes). Last, in “GSE132083_UNDER”, only 9 genes were found to be enriched and corresponded to upregulated genes in BC.

All of the gene sets related to pathways involved in CSCs machinery contributed to the identification of deregulated genes in both cancer types. Among the other pathways, the ones with the highest impact in the study was “TNF”. A total number of 37 genes from the “TNF” gene set were found to be upregulated in MCL, whereas 23 and 20 genes were found to be downregulated in MCL and BC, respectively. Interestingly there were 5 of the pathways that contributed exclusively with upregulated genes: “TGF_BETA” (31 and 17 genes for MCL and BC, respectively), “MYC” (35 and 10 genes for MCL and BC, respectively), “”NOTCH” (42 genes for MCL), “WNT_BCATENIN” (32 genes for MCL) , “JAK_STAT” (6 genes for MCL). “HIPPO” contributed to both, downregulated and upregulated genes (10 genes were found to be downregulated in BC and 18 upregulated in MCL). Finally, “HEDGEHOG” was the only pathway contributing exclusively with downregulated genes (44 genes for MCL). contributing with 3, 3 and 1 upregulated genes, respectively.

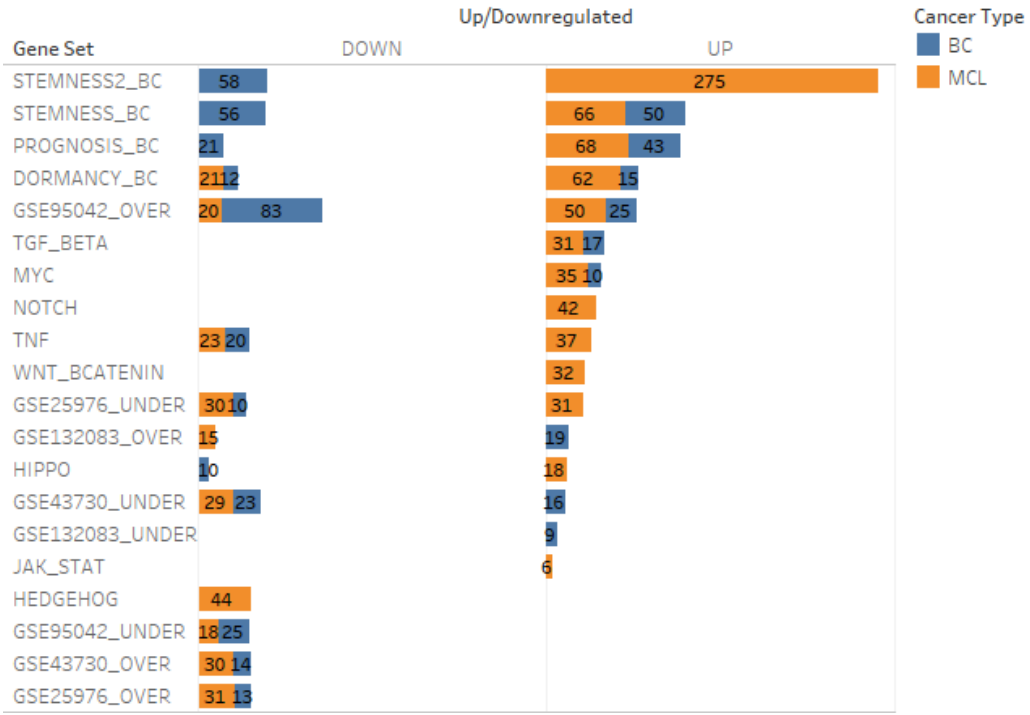


Figure 5. Number of up/downregulated genes across gene sets, comparing BC and MCL

4.1.4 Identification of common CSC gene signatures

A list of 269 common genes between cancer types has been identified (full list available in section 9.2 of Appendix).

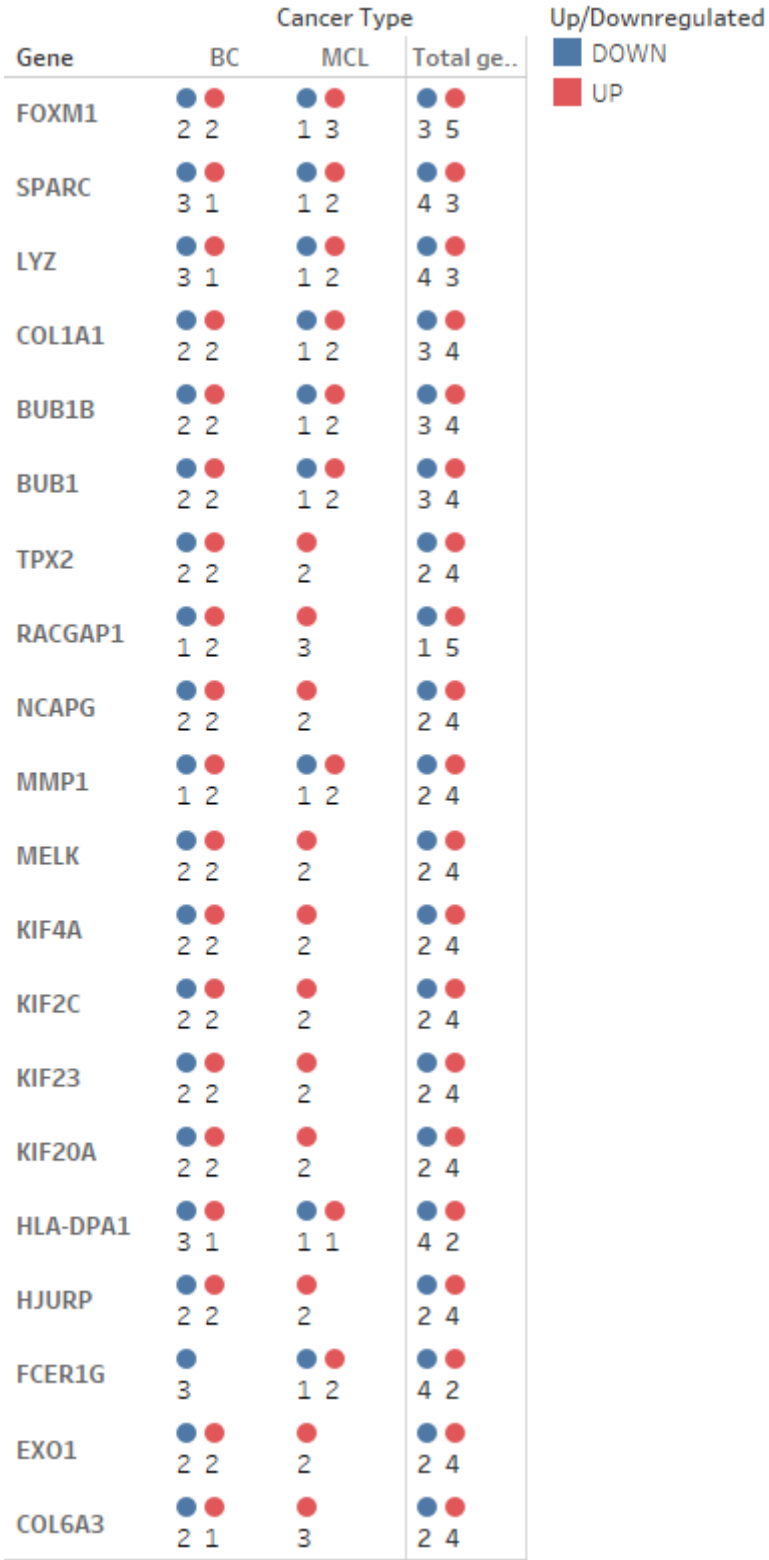
First, the list of genes was ordered by the number of total datasets in which they have been found enriched. FOXM1, forkhead box M1, a transcriptional activator involved in cell proliferation was the gene found enriched in the highest number of datasets: 8 up to 12 datasets (66,7%), 4 corresponding to BC and another 4 corresponding to MCL. Another 5 genes were found enriched in 7 datasets (4 BC and 3 MCL): SPARC, LYZ, COL1A1, BUB1B and BUB1.The top 20 are shown in ***Table 10***.

	Cancer Type		
Gene	BC	MCL	Total
FOXM1	4	4	8
SPARC	4	3	7
LYZ	4	3	7
COL1A1	4	3	7
BUB1B	4	3	7
BUB1	4	3	7
TPX2	4	2	6
RACGAP1	3	3	6
NCAPG	4	2	6
MMP1	3	3	6
MELK	4	2	6
KIF23	4	2	6
KIF20A	4	2	6
KIF4A	4	2	6
KIF2C	4	2	6
HLA-DPA1	4	2	6
HJURP	4	2	6
FCER1G	3	3	6
EXO1	4	2	6
COL6A3	3	3	6

Table 10. Top 20 common genes found enriched in GSEA

The number of datasets in which the genes were found up or downregulated comparing BC and MCL was also analyzed.

The results revealed that FOXM1 was found upregulated in 5 up to 8 datasets (2 BC and 3 MCL) and downregulated in 3 (2 BC and 1 MCL). While SPARC and LYZ were found downregulated in higher proportion (4 up to 7 datasets), COL1A1, BUB1B and BUB1 were found upregulated (4 up to 7 datasets). Comparing BC and MCL, SPARC and LYZ were mainly found downregulated in BC (3 up to 4 datasets) whereas in MCL they were found mainly upregulated (2 up to 3 datasets). The rest among the top 20 enriched genes were found more upregulated than downregulated except KIF20A and FCER1G that showed the opposite results.



Gene	BC	MCL	Total ge..
FOXM1	2 2	1 3	3 5
SPARC	3 1	1 2	4 3
LYZ	3 1	1 2	4 3
COL1A1	2 2	1 2	3 4
BUB1B	2 2	1 2	3 4
BUB1	2 2	1 2	3 4
TPX2	2 2	2	2 4
RACGAP1	1 2	3	1 5
NCAPG	2 2	2	2 4
MMP1	1 2	1 2	2 4
MELK	2 2	2	2 4
KIF4A	2 2	2	2 4
KIF2C	2 2	2	2 4
KIF23	2 2	2	2 4
KIF20A	2 2	2	2 4
HLA-DPA1	3 1	1 1	4 2
HJURP	2 2	2	2 4
FCER1G	3	1 2	4 2
EXO1	2 2	2	2 4
COL6A3	2 1	3	2 4

Table 11. Top 20 common genes found enriched in GSEA, distributed by up or downregulation.

The genes were also ordered by their significance. If the same gene was found enriched in several datasets and/or was present in different CSC gene sets, the average of the p-values was calculated and taken into account to build an "*Average p-value*". Genes were sorted in descendent order. The top 20 are shown in ***Table 12***.

Gene	Cancer Type BC	MCL	Average p-value
EPB41L4A	0,000000	0,000000	0,000000
IL1B	0,000000	0,000000	0,000000
IRF6	0,000000	0,000000	0,000000
MFAP5	0,000000	0,000000	0,000000
PI3	0,000000	0,000000	0,000000
PLD5	0,000000	0,000000	0,000000
RNF152	0,000000	0,000000	0,000000
SCEL	0,000000	0,000000	0,000000
SERPINB3	0,000000	0,000000	0,000000
SPINK7	0,000000	0,000000	0,000000
SPRR1A	0,000000	0,000000	0,000000
SPRR1B	0,000000	0,000000	0,000000
TLL1	0,000000	0,000000	0,000000
ZBED2	0,000000	0,000000	0,000000
CDK4	0,006000	0,000333	0,001750
EXOSC4	0,006000	0,000500	0,002333
ACTL6A	0,008000	0,000000	0,002667
CDKN2A	0,008000	0,000000	0,002667
PML	0,008000	0,000000	0,002667
RUVBL1	0,008000	0,000000	0,002667

Recuento definido de Dataset
1
4

Count datasets
1 3

Table 12. Top 20 genes sorted by significance (average p-value)

Most of the genes with lower significance are present in 2 datasets (one per cancer type). CDK4 is the gene with lowest significance (average p-value of 0.001750) found enriched in more than 2 datasets (3 MCL and 1 BC).

Several gene signatures were generated depending on the average p-value:

1. Gene signature with genes with average p-value lower than 0.01: composed by 53 genes (***53-CSC gene signature***).

2. Gene signature with genes with average p-value lower than 0.05: composed by 242 genes (***242-CSC gene signature***).

3. Gene signature with genes with average p-value lower than 0.1: composed by 269 genes (***269-CSC gene signature***).

Gene signatures can be found in section 9.3 of Appendix.

4.2. Functional annotation analysis

In order to further understand the biological context of the CSC genes identified in both cancer types (MCL and BC), a bioinformatic tool for functional annotation was used (Panther Classification System). The full list of 269 genes was selected for the analysis.

4.2.1 Molecular function

A first analysis regarding the molecular function was performed (**Figure 6**). The most relevant molecular functions retrieved were “binding” (73 genes) and “catalytic activity” (63 genes).

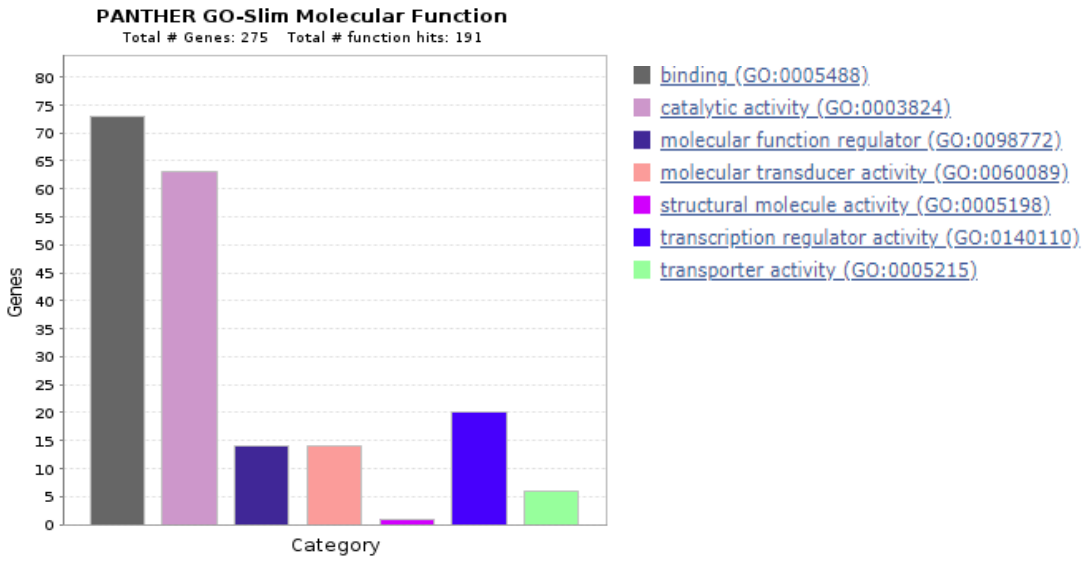


Figure 6. Functional annotation of genes in 269-CSC gene signature: molecular functions (1)

More than half of the genes that mapped to “binding” function were subgrouped into “protein binding” category (44 up to 73 genes), “heterocyclic compound binding” (23 genes) and “organic cyclic compound binding” (23 genes). These two groups contained the same genes. (**Figure 7**).

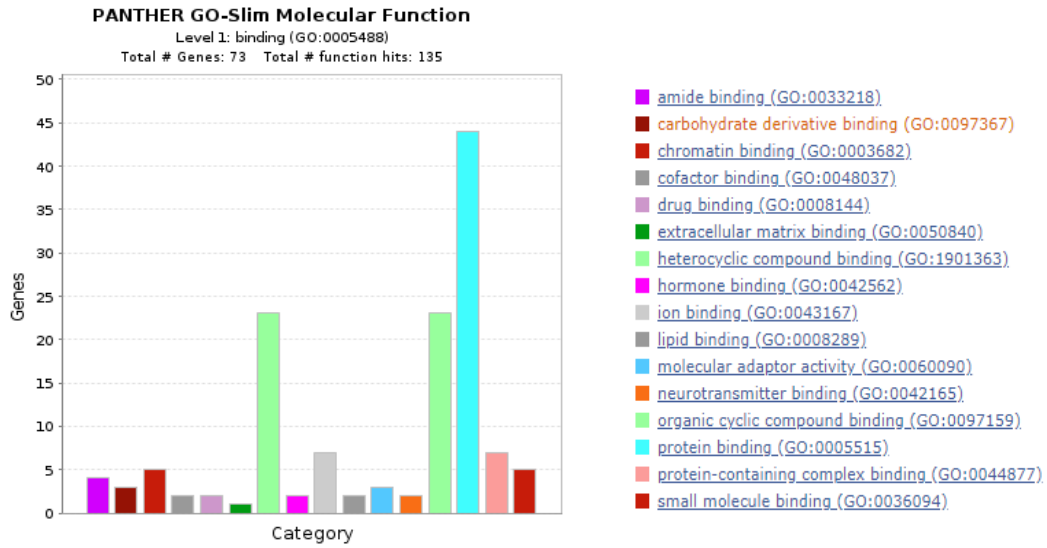


Figure 7. Functional annotation of the 73 “binding” genes.

Among the “protein binding”, the most relevant categories were “enzyme binding” (12 up to 44 genes), “cytoskeletal protein binding” (11 genes) and “signalling receptor binding” (10 genes) (**Figure 8**), whereas among the “heterocyclic compound binding” and “organic cyclic compound binding” genes, the most relevant category was “nucleic acid binding” (20 up to 23 genes) (

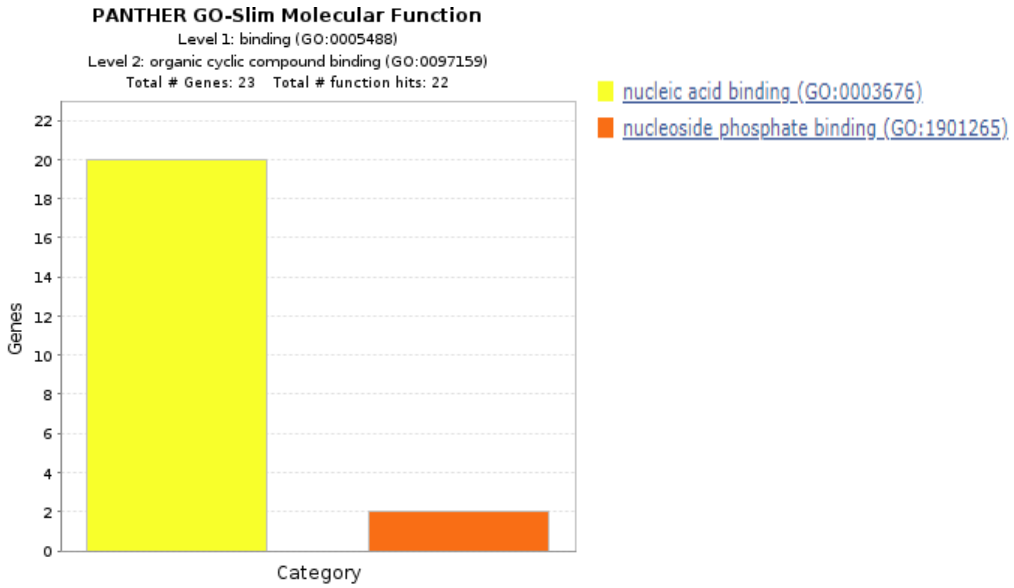


Figure 9).

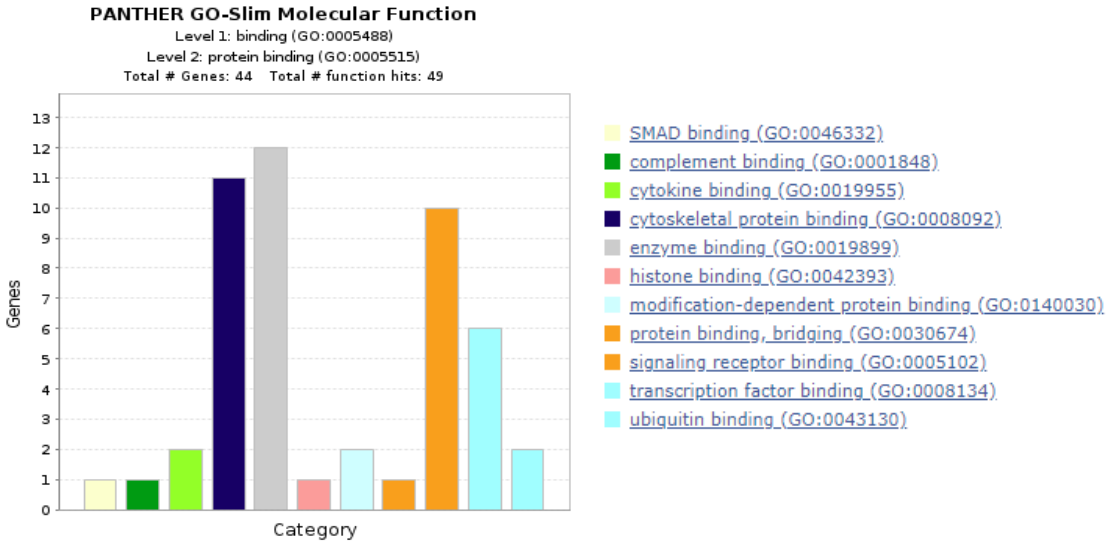


Figure 8. Functional annotation of 44 "binding protein" genes

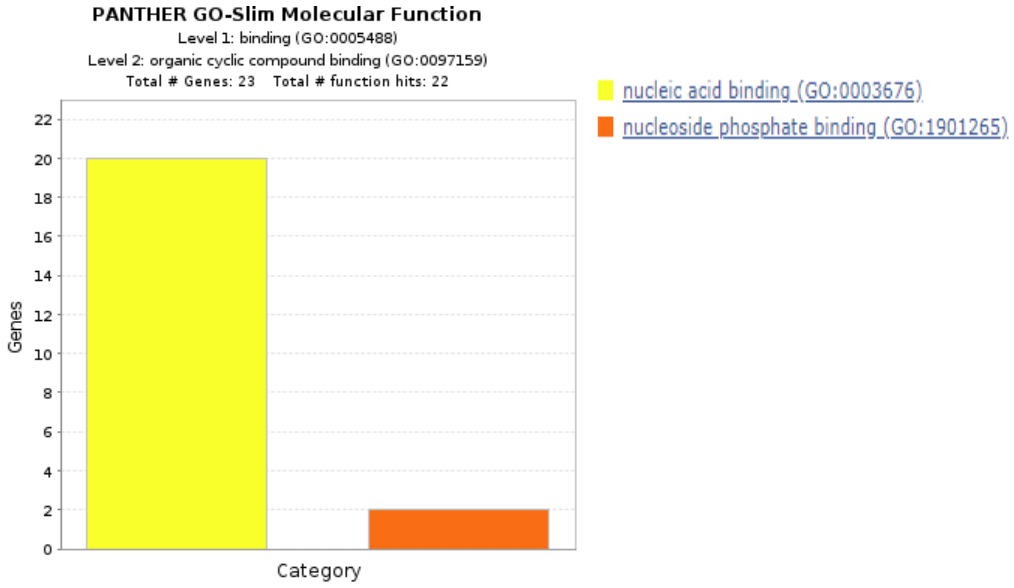


Figure 9. Functional annotation of 20 "heterocyclic compound binding" genes

4.2.2 Biological process

A second analysis focused on annotating the list of 269 genes regarding the biological processes in which they are involved was performed.

Major categories identified were “cellular process” (132 up to 269 genes), “biological regulation” (84 genes) and “metabolic process” (83 genes) (**Figure 10**).

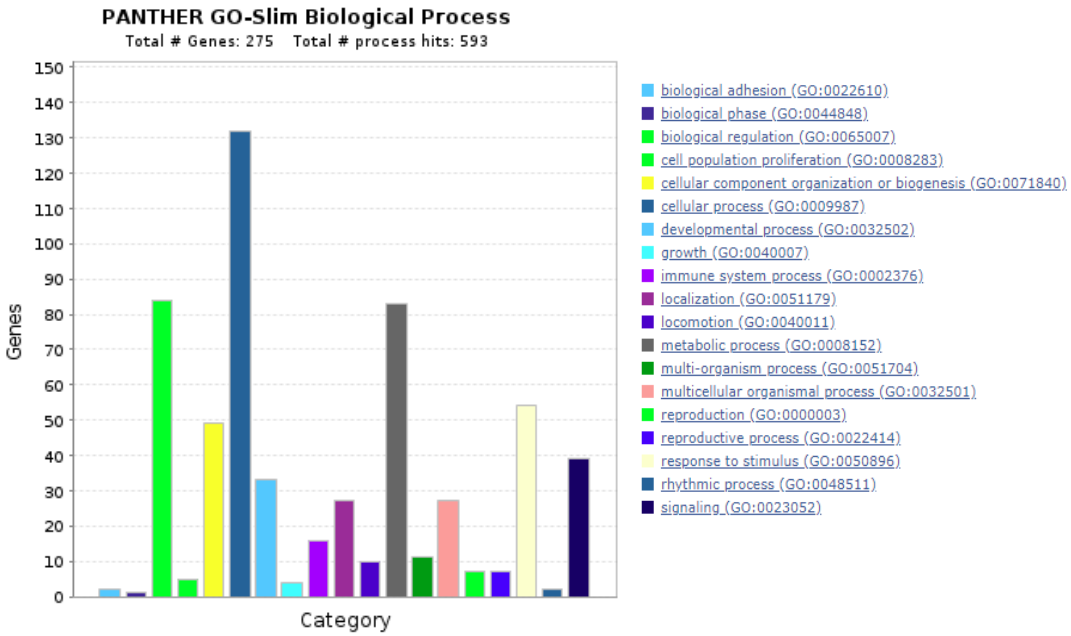


Figure 10. Functional annotation of genes in 269-CSC gene signature: biological processes

Narrowing down “cellular process” category, major subcategories found were: “cellular metabolic process” (75 up to 132 genes), “cellular component organization” (49 genes), “cellular response to stimulus” (42 genes), “cell communication” (39 genes) and “signal transduction” (38 genes) (***Figure 11***).

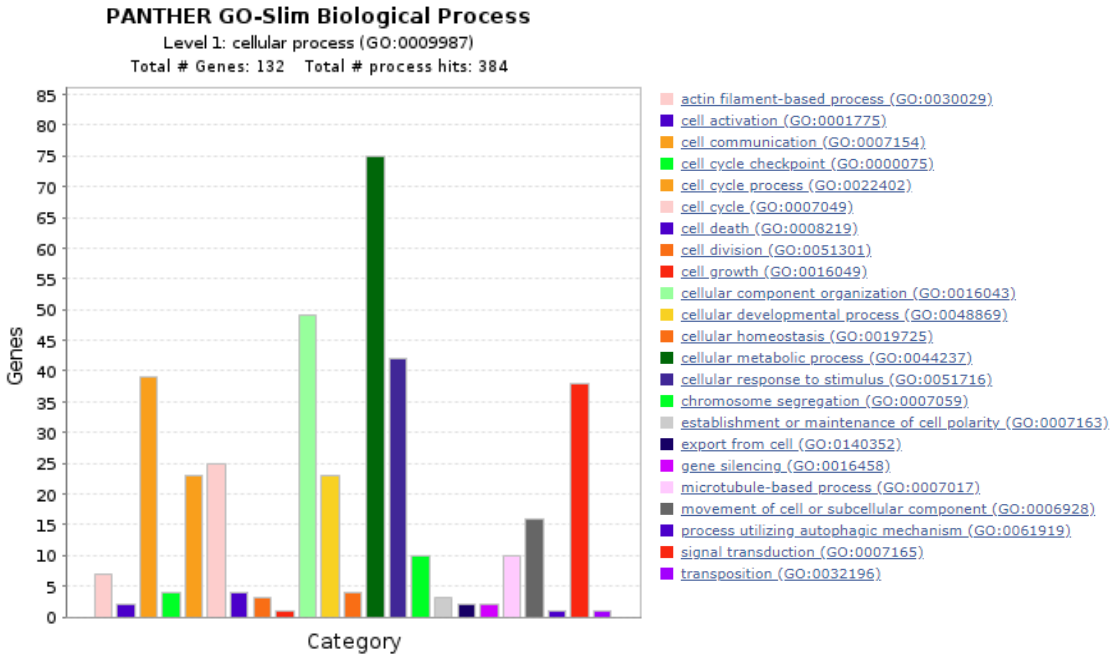


Figure 11. Functional annotation of the 132 "cellular process" genes.

Narrowing down several levels within “biological regulation” category, the major subcategories found were: “regulation of biological process” (80 up to 84 genes) → “regulation of cellular process” (78 up to 80 genes) → “regulation of cellular metabolic process” (41 up to 78 genes) and “signal transduction” (38 genes) (**Figure 12**).

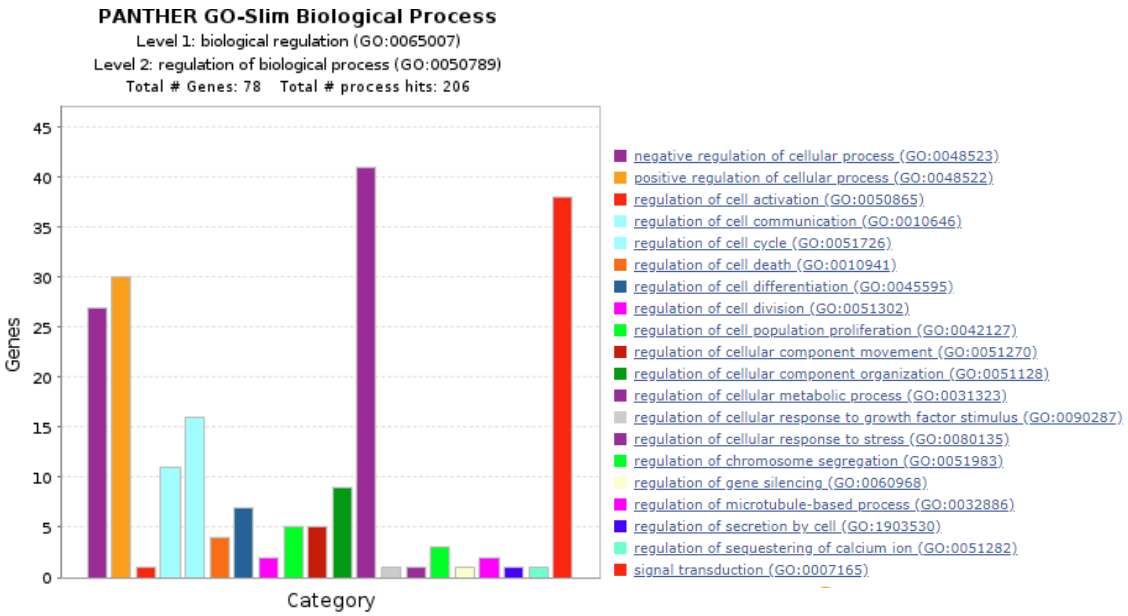


Figure 12. Functional annotation of 78 “regulation of cellular metabolic process” genes.

Last, narrowing down “metabolic process” category, major subcategories found were: “organic substance metabolic process” (79 up to 83 genes), “primary metabolic process” (77 genes), “cellular metabolic process” (75 genes) and “nitrogen compound metabolic process” (73 genes) (**Figure 13**).

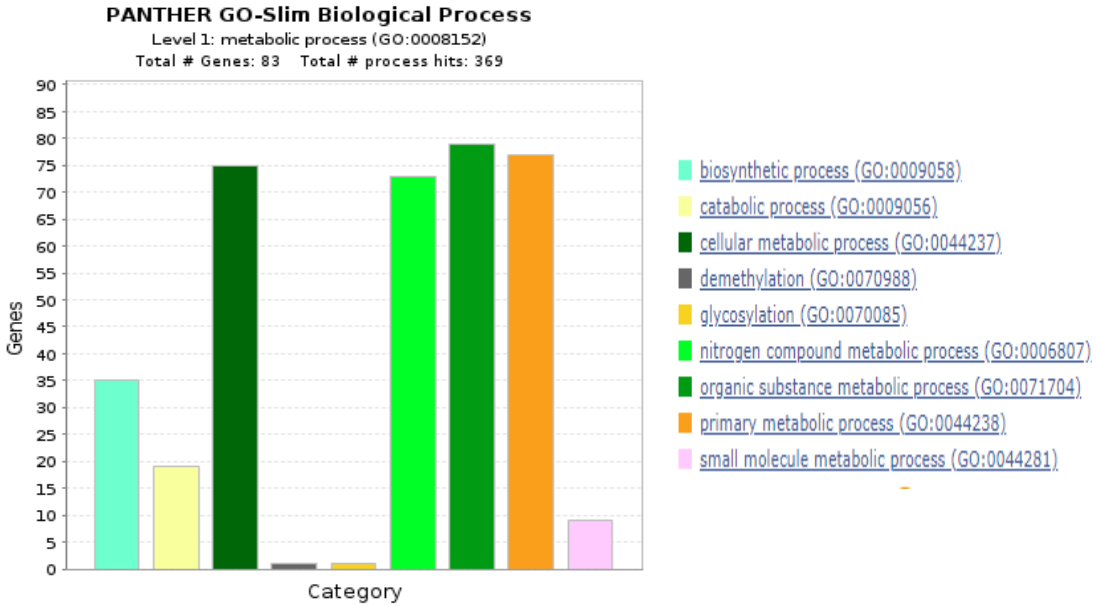


Figure 13. Functional annotation of 83 “metabolic process” genes

4.3. Correlation with prognostic parameters: Kaplan-Meier survival analysis

In order to study the prognostic significance of the CSC-gene signatures generated in cancer we followed the following approach: In silico analysis using the bioinformatic tool Kaplan-Meier Plotter. Survival analysis of the genes present in the 53-CSC gene signature has been performed using data from breast, ovarian, lung and gastric cancer studies.

I. Breast cancer cohort

Analysis has been run in data from 1764 patients. From the list of 53 genes, a total of 35 have been found significantly correlated with poor survival whereas no differences have been found for 16 of the genes. Only results for the significant genes are displayed (***Table 13***).

High expression of 13 of the genes (genes in red colour in the table) and low expression of 22 of the genes (genes in blue colour in the table) have been correlated with poor survival.

Gene	Hazard-Ratio (HR)	logrank P	Median survival in low expression cohort (months)	Median survival in high expression cohort (months)	Expression state found in our study (BC + MCL)
IL1B	0.69 (0.61-0.76)	1.4e-11***	171.43	216.66	Down (1+1)
PI3	1.12 (1.01-1.25)	0.039*	216.66	191.21	Down (1+1)
PLD5	0.82 (0.8-0.95)	0.01*	33	40	Down (1+1)
SERPINB3	0.8 (0.71-0.89)	4.6e-05***	40.53	60	Down (1+1)
SPRR1B	0.81 (0.73-0.9)	0.00016***	42.51	57.6	Down (1+1)
TLL1	0.8 (0.69-0.94)	0.0052**	32	39.95	Down (1+1)
ZBED2	0.72 (0.65-0.8)	3.9e-09***	38.4	64	Down (1+1)
CCDC80	0.81 (0.69-0.94)	0.0072**	31	40.56	Up (1+0), Down (1+1)
DENND2C	0.69 (0.59-0.8)	1.9e-06***	29	43	Down (0+1)
FAT2	0.89 (0.8-1)	0.043*	44.1	53.1	Up (1+0), Down (1+1)
FOLH1	1.21 (1.09-1.35)	0.00053***	216.66	191.21	Up (1+0), Down (0+1)
GCNT4	0.74 (0.66-0.82)	4.3e-08***	228.85	216.66	Up (1+0), Down (0+1)
NRG1	0.88 (0.79-0.98)	0.018*	43	55.2	Up (1+0), Down (0+1)
PDGFD	0.65 (0.58-0.73)	1.7e-14***	216.66	191.21	Up (1+0), Down (0+1)
PHLDA2	1.42 (1.27-1.58)	4e-10***	228.85	173.2	Up (1+1)
PPP2R5A	0.88 (0.79-0.98)	0.023*	228.85	216.66	Up (1+1)
TP63	0.67 (0.6-0.75)	1e-12***	37.2	72.2	Up (1+0), Down (0+1)
ACTL6A	1.55 (1.39-1.73)	2.3e-15***	216.66	185.16	Up (1+2)
FADD	1.55 (1.39-1.73)	3.6e-15***	228.85	184.04	Up (0+1), Down (1+1)
MAP3K7	0.75 (0.67-0.84)	3.5e-07***	40.44	60	Up (0+1), Down (1+1)
MAP4K4	1.36 (1.22-1.52)	2.4e-08***	65	36.96	Up (0+1), Down (1+1)
NFKB1	0.63 (0.57-0.71)	3.3e-16***	191.21	216.66	Up (0+1), Down (1+1)
PML	0.83 (0.75-0.93)	0.001**	44	57	Up (1+2)
PRKCZ	0.82 (0.74-0.92)	0.00052***	228.85	216.66	Up (0+1), Down (1+1)
RELA	0.83 (0.75-0.93)	0.00088***	228.85	216.66	Up (0+1), Down (1+1)
RIPK1	0.72 (0.62-0.85)	4.4e-05***	29	43	Up (0+1), Down (1+1)
RUVBL1	1.39 (1.24-1.55)	3.8e-09***	228.85	185.16	Up (1+2)
SERPINB2	0.83 (0.75-0.93)	0.0011**	43	57.3	Up (0+1), Down (2+2)
SQSTM1	1.35 (1.21-1.5)	8.7e-08***	216.66	228.85	Up (0+1), Down (1+1)
TNFAIP3	0.85 (0.77-0.95)	0.0047**	45	55	Up (0+1), Down (1+1)

CDK4	1.53 (1.37-1.71)	1.7e-14***	216.66	171.43	Up (1+3)
EXOSC4	1.39 (1.25-1.55)	3e-09***	216.66	171.43	Up (1+2)
MIEN1	1.4 (1.2-1.64)	2.2e-05***	43	30	Up (1+1)
NDUFB10	1.31 (1.12-1.53)	0.00067***	171.43	148	Up (1+1)
SKP2	1.75 (1.57-1.96)	<1e-16***	216.66	163.46	Up (1+2)

Table 13. List of 35 genes significantly correlated with poor survival (***p-value<0.001, **p-value<0.01, *p-value<0.05). In red colour, genes whose survival decrease with high expression of the gene and in blue colour, genes whose survival decreases with low expression of the gene.

The genes most significantly correlated with poor prognosis are **SKP2, NFKB1, ACTL6A, FADD, CDK4 and PDGFD. For SKP2, ACTL6A, FADD** and **CDK4** it's the high gene expression which correlates with a lower median survival whereas for NFKB1 and PDGFD, it's the low expression of the genes which is related to a poorer median survival.

These results are consistent with the findings obtained in the GSEA analysis (see section 4.1) for 4 up to 6 of these genes. GSEA analysis showed that SKP2 was upregulated both in BC (1 dataset) and MCL (2 datasets), CDK4 was upregulated both in BC (1 dataset) and MCL (3 datasets), ACTL6A was upregulated both in BC (1 dataset) and MCL (2 datasets) and NFKB1 was downregulated in both, BC (1 dataset) and MCL (1 dataset). However, for FADD and PDGFD, results showed partial consistency. FADD was only found upregulated in 1 MCL dataset whereas it was found downregulated in 1 BC and 1 MCL datasets. PDGFD was found downregulated in 1 MCL dataset but upregulated in 1 BC dataset.

II. Ovarian cancer cohort

Analysis has been run in data from 614 patients. From the list of 53 genes, a total of 31 have been found significantly correlated with poor survival whereas no differences have been found for 22 of the genes. Only results for the significant genes are displayed (***Table 14***).

High expression of 23 of the genes (genes in red colour in the table) and low expression of 8 of the genes (genes in blue colour in the table) have been correlated with poor survival.

Gene	Hazard-Ratio (HR)	logrank P	Median survival in low expression cohort (months)	Median survival in high expression cohort (months)	Expression state found in our study (BC + MCL)
IL1B	1.18 (1.03-1.37)	0.02*	21.43	19.27	Down (1+1)

MFAP5	1.32 (1.16-1.5)	2.2e-05***	23.82	18.23	Down (1+1)
PI3	1.18 (1.04-1.34)	0.013*	20.6	19	Down (1+1)
RNF152	1.46 (1.21-1.77)	7.9e-05***	19	15.13	Down (1+1)
SPRR1A	0.78 (0.68-0.9)	0.00045***	18.4	27	Down (1+1)
SPRR1B	0.79 (0.68-0.9)	0.00057***	17	21.6	Down (1+1)
TLL1	1.39 (1.15-1.68)	5e-04***	19.55	14	Down (1+1)
ZBED2	1.18 (1.03-1.36)	0.015*	21	19.23	Down (1+1)
CCDC80	1.90 (1.64-2.41)	6.4e-13***	23	11	Up (1+0), Down (1+1)
CLCA2	0.86 (0.74-0.99)	0.041*	18.79	20.56	Up (1+0), Down (1+1)
DENND2C	1.37 (1.13-1.67)	0.0017**	18.87	14	Down (0+1)
LUM	1.68 (1.46-1.94)	4.8e-13***	23	13.73	Up (1+1)
PAK2	1.19 (1.04-1.37)	0.011*	20.2	19	Up (1+1)
PDGFD	1.35 (1.19-1.53)	2.9e-06***	22.5	17.4	Up (1+0), Down (0+1)
PHLDA2	1.16 (1.01-1.32)	0.038*	22	19.23	Up (1+1)
POF1B	0.68 (0.56-0.82)	5.4e-05***	14.37	19.98	Up (1+0), Down (0+1)
TAF12	1.32 (1.16-1.52)	4.3e-05***	25	18.23	Up (1+1)
TP63	0.87 (0.76-0.99)	0.038*	18.3	23.73	Up (1+0), Down (0+1)
ACTL6A	1.34 (1.17-1.54)	3.2e-05***	23.56	18.93	Up (1+2)
CASP8	1.15 (1.01-1.3)	0.039*	22.13	19	Up (0+1), Down (1+1)
MTS1	1.25 (1.09-1.42)	0.00093***	21	19.35	Up (1+2)
MAP4K4	1.16 (1.02-1.31)	0.024*	21.43	18.86	Up (0+1), Down (1+1)
NFKB1	1.21 (1.06-1.38)	0.005**	21.13	17.9	Up (0+1), Down (1+1)
PML	1.2 (1.04-1.38)	0.013*	23.1	18.98	Up (1+2)
RIPK1	1.26 (1.04-1.53)	0.019*	19	14.53	Up (0+1), Down (1+1)
RUVBL1	1.3 (1.14-1.47)	6.5e-05***	22.24	18	Up (1+2)
SERPINB2	1.3 (1.12-1.51)	0.00049***	22	19	Up (0+1), Down (2+2)
BMI1	1.38 (1.2-1.59)	9.4e-06***	28	18.43	Up (1+1)
MIEN1	0.75 (0.62-0.91)	0.0029**	15.13	19.02	Up (1+1)
NDUFB10	0.74 (0.61-0.89)	0.0014**	15	20	Up (1+1)
SKP2	1.16 (1-1.33)	0.042*	20.56	19.8	Up (1+2)

Table 14. List of 31 genes significantly correlated with poor survival (***p-value<0.001, **p-value<0.01, *p-value<0.05). In red colour, genes whose survival decrease with high expression of the gene and in blue colour, genes whose survival decreases with low expression of the gene.

The genes most significantly correlated with poor prognosis are **CCDC80** and **LUM**. In both cases it's the high gene expression which correlates with a lower median survival.

These results are consistent with the findings obtained in the GSEA analysis (see section 4.1) for LUM. GSEA analysis showed that LUM was upregulated both in BC (1 dataset) and MCL (1 dataset), However, for CCDC80, results showed partial consistency. The

gene was only found upregulated in 1 BC dataset whereas it was found downregulated in 1 BC and 1 MCL datasets.

III. Lung cancer cohort

Analysis has been run in data from 1144 patients. From the list of 53 genes, a total of 34 have been found significantly correlated with poor survival whereas no differences have been found for 19 of the genes. Only results for the significant genes are displayed (***Table 15***).

High expression of 20 of the genes (genes in red colour in the table) and low expression of 14 of the genes (genes in blue colour in the table) have been correlated with poor survival.

Gene	Hazard-Ratio (HR)	logrank P	Median survival in low expression cohort (months)	Median survival in high expression cohort (months)	Expression state found in our study (BC + MCL)
PI3	1.33 (1.17-1.51)	1.1e-05***	79.5	54.57	Down (1+1)
PLD5	0.83 (0.7-0.98)	0.027*	69.93	88	Down (1+1)
RNF152	1.19 (1.01-1.4)	0.039*	85	71	Down (1+1)
SCEL	0.67 (0.57-0.79)	2.1e-06***	55.37	103	Down (1+1)
SERPINB3	1.23 (1.08-1.39)	0.0013**	78	63	Down (1+1)
SPRR1A	1.24 (1.09-1.41)	0.00074***	78.5	62	Down (1+1)
SPRR1B	1.42 (1.25-1.62)	4.1e-08***	86.27	53	Down (1+1)
CCDC80	0.8 (0.68-0.95)	0.0095**	65.57	88.7	Up (1+0), Down (1+1
FAT2	1.32 (1.16-1.5)	1.6e-05***	85	57	Up (1+0), Down (1+1)
GCNT4	1.17 (1.03-1.33)	0.014*	74	64.1	Up (1+0), Down (0+1)
LUM	0.77 (0.68-0.97)	3.8e-05***	59.11	80.03	Up (1+1)
PDGFD	0.76 (0.67-0.86)	1.8e-05***	52	80.9	Up (1+0), Down (0+1)
PPP2R5A	0.69 (0.61-0.79)	1.1e-08***	50	85	Up (1+1)
TAF9	1.38 (1.21-1.57)	1.1e-06***	89	54.2	Up (1+1)
TP63	1.14 (1-1.29)	0.044*	74	65.57	Up (1+0), Down (0+1)
XDH	0.71 (0.6-0.84)	5e-05***	63	93	Up (1+0), Down (1+1)
ACTL6A	1.15 (1.01-1.31)	0.03*	74	62.47	Up (1+2)
CASP8	0.74 (0.65-0.84)	3.6e-06***	57	78.5	Up (0+1), Down (1+1)
CDKN2A	1.28 (1.13-1.46)	0.00011***	84.1	57	Up (1+2)
MAP3K7	0.72 (0.64-0.82)	6.6e-07***	55	79.54	Up (0+1), Down (1+1)
MAP4K4	1.25 (1.1-1.41)	0.00062***	78.9	61.2	Up (0+1), Down (1+1)
PML	1.17 (1.03-1.33)	0.014*	77.77	63.03	Up (1+2)
PRKCZ	0.82 (0.72-0.93)	0.0023**	62.2	78	Up (0+1), Down (1+1)

RELA	0.86 (0.76-0.98)	0.024*	63	74	Up (0+1), Down (1+1)
RIPK1	0.66 (0.46-0.79)	1.5e-06***	57	104	Up (0+1), Down (1+1)
RUVBL1	1.61 (1.42-1.83)	1.9e-13***	95.5	48.8	Up (1+2)
SERPINB2	1.23 (1.08-1.4)	0.0013**	79.27	62	Up (0+1), Down (2+2)
SQSTM1	0.86 (0.76-0.98)	0.019*	63	77.6	Up (0+1), Down (1+1)
CDK4	1.51 (1.33-1.71)	2.1e-10***	85	49	Up (1+3)
EXOSC4	1.24 (1.1-1.41)	0.00073***	79.27	59	Up (1+2)
BMI1	0.72 (0.64-0.82)	4.2e-07***	59	89	Up (1+1)
MIEN1	1.24 (1.05-1.46)	0.011*	84	65.57	Up (1+1)
NDUFB10	1.2 (1.02-1.42)	0.031*	88.7	68	Up (1+1)
SKP2	1.24 (1.09-1.4)	0.00093***	78	59.53	Up (1+2)

Table 15. List of 34 genes significantly correlated with poor survival (***p-value<0.001, **p-value<0.01, *p-value<0.05). In red colour, genes whose survival decrease with high expression of the gene and in blue colour, genes whose survival decreases with low expression of the gene.

The genes most significantly correlated with poor prognosis are RUVBL1 and CDK4. In both cases it's the high gene expression which correlates with a lower median survival.

These results are consistent with the findings obtained in the GSEA analysis (see section 4.1). GSEA analysis showed that RUVBL1 was upregulated both in BC (1 dataset) and MCL (2 datasets) and that CDK4 was upregulated both in BC (1 dataset) and MCL (3 datasets).

IV. Gastric cancer cohort

Analysis has been run in data from 631 patients. From the list of 53 genes, a total of 43 have been found significantly correlated with poor survival whereas no differences have been found for 10 of the genes. Only results for the significant genes are displayed (***Table 16***).

High expression of 26 of the genes (genes in red colour in the table) and low expression of 17 of the genes (genes in blue colour in the table) have been correlated with poor survival.

Gene	Hazard-Ratio (HR)	logrank P	Median survival in low expression cohort (months)	Median survival in high expression cohort (months)	Expression state found in our study (BC + MCL)
IRF6	0.79 (0.65-0.96)	0.019*	27.6	36.4	Down (1+1)
IL1B	0.73 (0.62-0.87)	0.00045***	23.6	31.2	Down (1+1)

MFAP5	1.3 (1.1-1.54)	0.0026**	32.1	26.8	Down (1+1)
PLD5	1.81 (1.41-2.32)	2.3e-06	113.2	33.27	Down (1+1)
RNF152	1.35(1.05-1.72)	0.017*	93.2	42	Down (1+1)
SERPINB3	1.31 (1.1-1.56)	0.0021**	35.4	24.03	Down (1+1)
SPRR1A	1.45 (1.18-1.77)	0.00032***	42.07	26.5	Down (1+1)
SPRR1B	1.46 (1.21-1.76)	5.5e-05***	50.8	24.8	Down (1+1)
TLL1	1.73 (1.32-2.27)	5.8e-05***	123.8	37.93	Down (1+1)
CCDC80	1.78 (1.43-2-22)	1.9e-07***	100.8	31.3	Up (1+0), Down (1+1)
CLCA2	1.35 (1.12-1.63)	0.0018**	39.8	27	Up (1+0), Down (1+1)
DENND2C	1.76 (1.35-2-29)	2.2e-05***	107.7	36.4	Down (0+1)
FAT2	1.64 (1.37-1.97)	5.8e-08***	35.9	21.23	Up (1+0), Down (1+1)
FOLH1	0.73 (0.61-0.86)	0.00022***	21.6	35.1	Up (1+0), Down (0+1)
GCNT4	1.23 (1.04-1.46)	0.015*	33.2	26.7	Up (1+0), Down (0+1)
LUM	0.83 (0.7-0.98)	0.031*	25	34.1	Up (1+1)
NRG1	0.79 (0.67-0.94)	0.0063**	24.4	33.27	Up (1+0), Down (0+1)
PAK2	0.68 (0.57-0.82)	4.5e-05***	26.6	42	Up (1+1)
PDGFD	1.45 (1.18-1.78)	0.00033***	62	27.8	Up (1+0), Down (0+1)
PHLDA2	0.74 (0.62-0.87)	0.00038***	25.9	34.1	Up (1+1)
PPP2R5A	0.81 (0.68-0.96)	0.014*	26.7	30	Up (1+1)
TAF9	0.68 (0.58-0.81)	9.9e-06***	25.2	36.17	Up (1+1)
TAF12	0.65 (0.55-0.77)	4.3e-07***	21.6	42.07	Up (1+1)
TP63	1.46 (1.23-1.72)	1.3e-05***	39.53	23.4	Up (1+0), Down (0+1)
XDH	0.7 (0.57-0.88)	0.0015**	31.33	70.4	Up (1+0), Down (1+1)
ACTL6A	0.58 (0.47-0.7)	2.3e-08***	25.2	77.2	Up (1+2)
CASP8	0.63 (0.53-0.75)	7e-08***	24.4	39.53	Up (0+1), Down (1+1)
CDKN2A	1.73 (1.4-2.14)	2.7e-07***	70.4	25.9	Up (1+2)
FADD	0.79 ((0.66-0.93)	0.0057**	27.8	30.9	Up (0+1), Down (1+1)
MAP3K7	1.3 (1.08-1.57)	0.0056**	32.1	25.5	Up (0+1), Down (1+1)
MAP4K4	1.68 (1.4-2.01)	9.1e-09***	38.2	18.6	Up (0+1), Dow n(1+1)
PML	1.7 (1.41-2.03)	7.5e-09***	53.43	23.6	Up (1+2)
PRKCZ	1.42 (1.16-1.72)	0.00048***	45.1	26.5	Up (0+1), Down (1+1)
RELA	1.58 (1.29-1.93)	6.1e-06	57.13	26	Up (0+1), Down (1+1)
RUVBL1	0.74 (0.62-0.87)	0.00038***	26.6	35.4	Up (1+2)
SERPINB2	1.21 (1.02-1.44)	0.029*	34.1	25.5	Up (0+1), Down (2+2)
SQSTM1	1.56 (1.32-1.85)	2.3e-07***	36.4	21	Up (0+1), Down (1+1)
CDK4	1.31 (1.09-1.58)	0.0044**	31.33	22.2	Up (1+3)
EXOSC4	0.75 (0.63-0.89)	0.00098***	25.5	34.1	Up (1+2)
BMI1	0.69 (0.58-0.82)	1.8e-05***	23.6	36.4	Up (1+1)
MIEN1	**1.34 (1.07-1.68)**	0.011*	65	34.37	Up (1+1)
NDUFB10	0.78 (0.62-0.97)	0.026*	40.2	77.2	Up (1+1)
SKP2	0.73 (0.61-0.87)	0.00051***	27.4	39.8	Up (1+2)

Table 16. List of 43 genes significantly correlated with poor survival (***p-value<0.001, **p-value<0.01, *p-value<0.05). In red colour, genes whose survival decrease with high expression of the gene and in blue colour, genes whose survival decreases with low expression of the gene.

The genes most significantly correlated with poor prognosis in gastric cancer are MAP4K4 and PML. In both cases it's the high gene expression which correlates with a lower median survival.

These results are consistent with the findings obtained in the GSEA analysis (see section 4.1) for PML. GSEA analysis showed that this gene was upregulated both in BC (1

dataset) and MCL (2 datasets). However, results are partially consistent with what was found for MAP4K4. This gene was found upregulated only in 1 MCL dataset and downregulated in 1 BC and 1 MCL datasets.

In order to analyse which genes are commonly found related to survival in all cancer types studied, a Venn's diagram was performed (***Figure 14***).

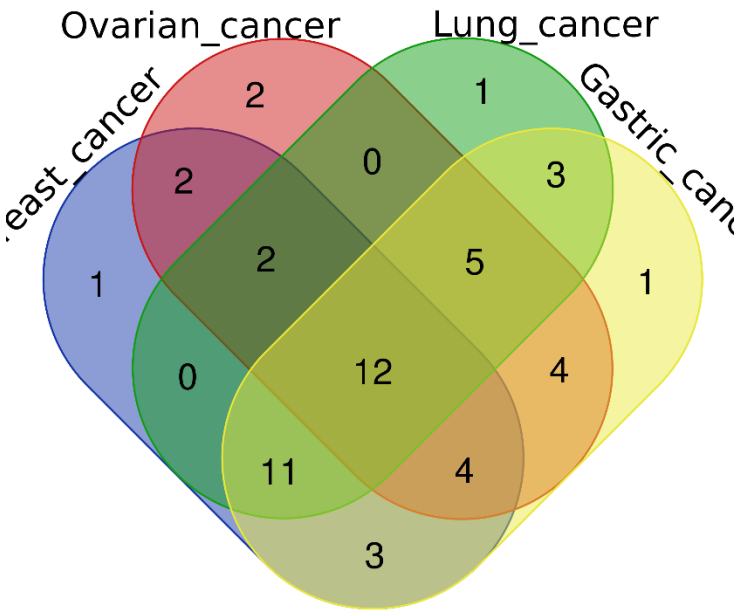


Figure 14. Analysis of genes significantly correlated with survival in 4 cancer types (Breast, Ovarian, Lung and Gastric cancer) using Venn's diagram

Names	total	elements
Breast_cancer Gastric_cancer Lung_cancer Ovarian_cancer	12	TP63 ACTL6A PML PDGFD MAP4K4 NDUFB10 CCDC80 SPRR1B SKP2 SERPINB2 RUVBL1 MIEN1
Breast_cancer Lung_cancer Ovarian_cancer	2	PI3 RIPK1
Breast_cancer Gastric_cancer Ovarian_cancer	4	PHLDA2 IL1B TLL1 DENND2C
Breast_cancer Gastric_cancer Lung_cancer	11	EXOSC4 CDK4 SERPINB3 PLD5 PPP2R5A MAP3K7 PRKCZ FAT2 RELA SQSTM1 GCNT4
Gastric_cancer Lung_cancer Ovarian_cancer	5	BMI1 SPRR1A RNF152 LUM CASP8
Breast_cancer Ovarian_cancer	2	NFKB1 ZBED2
Breast_cancer Gastric_cancer	3	FOLH1 NRG1 FADD
Gastric_cancer Ovarian_cancer	4	PAK2 MFAP5 CLCA2 TAF12
Gastric_cancer Lung_cancer	3	CDKN2A XDH TAF9
Breast_cancer	1	TNFAIP3
Ovarian_cancer	2	MTS1 POF1B
Lung_cancer	1	SCEL
Gastric_cancer	1	IRF6

Table 17. Number and name of genes found significantly correlated with survival across cancer types.

A total number of 12 genes have been found significantly correlated with survival in all 4 cancer types: TP63, ACTL6A, PML, PDGFD, MAP4K4, NDUFB10, CCDC80, SPRR1B, SKP2, SERPINB2, RUVBL1 and MIEN1. An additional number of 22 genes have been found significantly correlated with survival in at least 3 cancer types, 12 genes in at least 2 cancer types and 5 genes in specific cancer types (***Table 17***).

4.4. Evaluation of a predictive model for prognosis using machine learning

The assessment of the predictive power of the gene signature by ML required certain data specifications: 1) Cancer expression data for a high number of genes, 2) Availability of prognostic parameters, 3) Sufficient number of samples to run a ML algorithm. Data fulfilling these criteria were selected from cBioPortal, a public repository for Cancer Genomics.

The details of the data taken into account for the present work are listed here:

- **Cancer type:** Colorectal adenocarcinoma (TCGA, PanCancer Atlas)
- **Samples**: 594
- **Genes:** 20502
- **Prognostic parameters:** OS status (living/deceased), DSS status (alive or dead tumor free/dead with tumor), DFS status (disease free/recurred or progressed), PFS status (censored/progression), OS (months), PFS (months), DFS (months)
- **Demographic parameters:** age, sex
- **Clinical parameters:** stage

4.4.1 Model selection

The algorithm to assess the predictive power of the 53-CSC gene signature falls into the category of the supervised algorithms. The availability of both, numeric and categorical data, for the prognostic parameters allowed to use different ML strategies such as building a multiple linear regression model by using a numeric class parameter such as OS (months), or building a classification model by using a categorical class parameter such as OS status.

The following steps were conducted in order to take the most appropriate decision:

1. Analysis of the missing values

The results of the analysis (**Figure 15**) showed a high number of missing values (n=370) for the following prognostic parameters: DFS status, DFS (months). DSS status accounted for 24 missing values. These three parameters were discarded for further analysis.

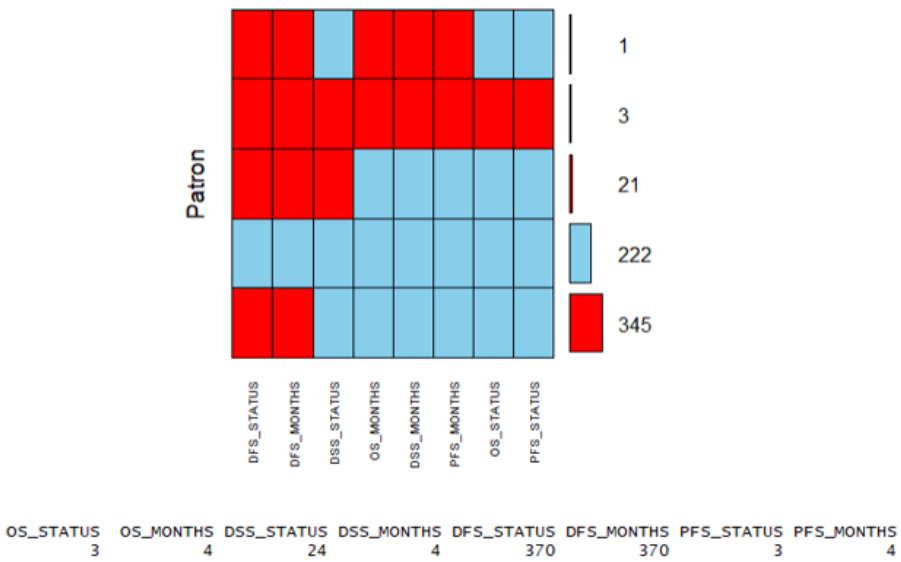


Figure 15. Analysis of missing values in prognostic variables

2. Correlation analysis (Pearson coefficient)

Next step was to perform a correlation analysis between the genes included in the 53-CSC gene signature and the following (numeric) prognostic parameters: OS (months) and PFS (months). This analysis allowed not only to select the suitable prognostic parameter for the study but also to select those genes with higher correlations for the model.

A first correlation analysis was performed to decide whether to use only the subset of "Deceased" patients (defined by the variable "OS status") or the entire dataset (20% of the dataset). In one hand, the use of the entire dataset increases the number of samples to generate the model, but in the other, a higher degree of correlation of the genes with the prognostic parameters leads to a more powerful model.

The proportion of "Deceased" and "Living" patients is the following:

```
0:LIVING 1:DECEASED
     470         119
```

The correlation analysis showed that correlations were higher when using the subset of "Deceased" patients compared to the entire dataset (results for a sample of 10 genes is shown in **Figure 16**). Moreover, results showed that the degree of correlations was higher with OS than with PFS (only results for 10 genes is shown in **Figure 16**).

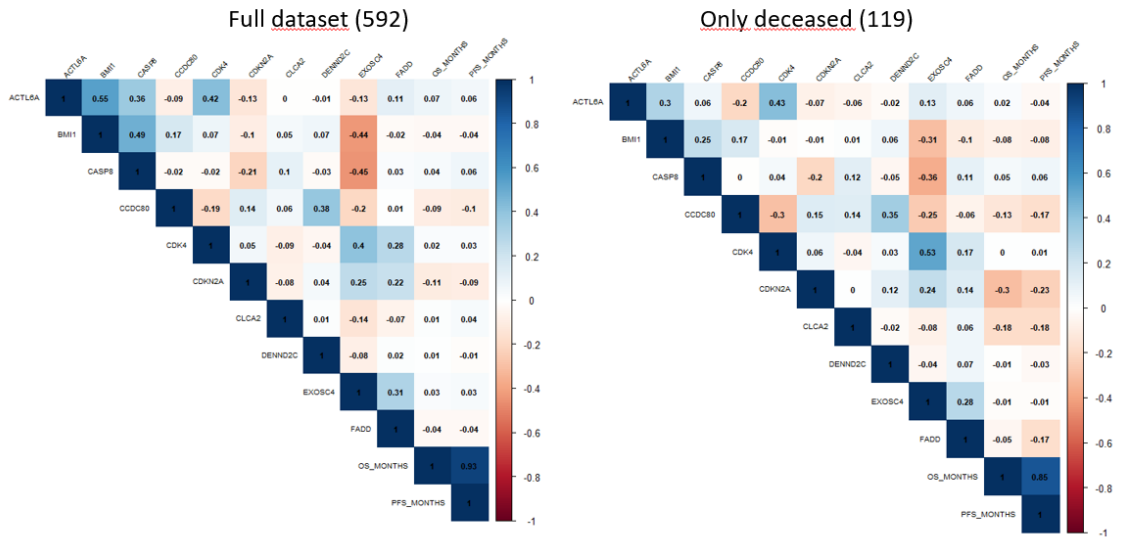


Figure 16. Correlation analysis with prognostic parameters comparing subset of "Deceased" patients with the entire dataset.

Taking into account these results, a classification algorithm (*Random forest*) was selected to generate a predictive model. For that, the 119 "Deceased" cases were split

into two classes by taking into account the median of “OS (months)”, that is 14.53 months.

- “**Early death**”: 59 cases which died before 14.53 months from diagnosis (Class 0)

- “**Late death**”: 60 cases which died after 14.53 months from diagnosis (Class 1)

4.4.2 Feature selection

The candidate genes to be included in the model resulted from the list of genes from the 53-CSC gene signature. In order to select which genes to include in the model several steps were performed:

1. Analysis of the missing values

The count of missing values for each of the 53 genes is listed below (**Table 18**):

ACTL6A	BMI1	CASP8	CCDC80	CDK4	CDKN2A	CLCA2	DENND2C	EXOSC4	FADD	FAT2	FOLH1	GCNT4	IL1B
0	0	0	0	0	0	0	0	0	0	0	0	0	0
IRF6	LUM	MAP3K7	MAP4K4	MFAP5	c17orf37	NDUFB10	NFKB1	NRG1	PAK2	PDGFD	PHLDA2	PI3	PLD5
0	0	0	0	0	0	0	0	0	0	0	0	0	39
PML	POF1B	PPP2R5A	PRKCZ	RELA	RFFL	RIPK1	RNF152	RUVBL1	SCEL	SERPINB2	SERPINB3	SKP2	SPINK7
0	0	0	0	0	0	0	0	0	0	0	0	0	39
SPRR1A	SPRR1B	SQSTM1	TAF12	TAF9	TLL1	TNFAIP3	TP63	XDH	ZBED2				
0	0	0	0	0	0	0	0	0	0				

Table 18. Count of missing values for the 53 genes of the 53-CSC gene signature

Results showed that two of the genes, SPINK7 and PLD5, accounted for 39 missing values each. As the number of cases for the analysis is small (119), these 2 genes were excluded for further analysis. Also, another gene “MIEN1” was found missing, and that was due that another name of the gene, c17orf37, was used in the original study. No missing values was found when using this other gene name.

2. Correlation analysis (Pearson coefficient)

The 51 genes (excluding SPINK7 and PLD5) were divided in 5 groups for the purpose of visualization. Results are shown in the **Appendix 9.4**. Results of the correlation analysis showed a diverse range of correlations between the 51 genes and Overall Survival ((Supplementary Figures S1-S5).

A cut-off of 15% correlation was used to select the genes for the model. Using this cut-off, a total number of 18 genes were selected: CLCA2, IL1B, MFAP5, NRG1, PDGFD, PHLDA2, RUVBL1, SERPINB2, SERPINB3, SPRR1A, SPRR1B, SQSTM1, TAF9,

TLL1, TP63, ZBED2, CDKN2A and SCEL. Gene signature can be found in section 9.3 of Appendix.

Another correlation matrix was extracted with the selected 18 genes to discard high correlations among the genes (>90%) (**Figure 17**).

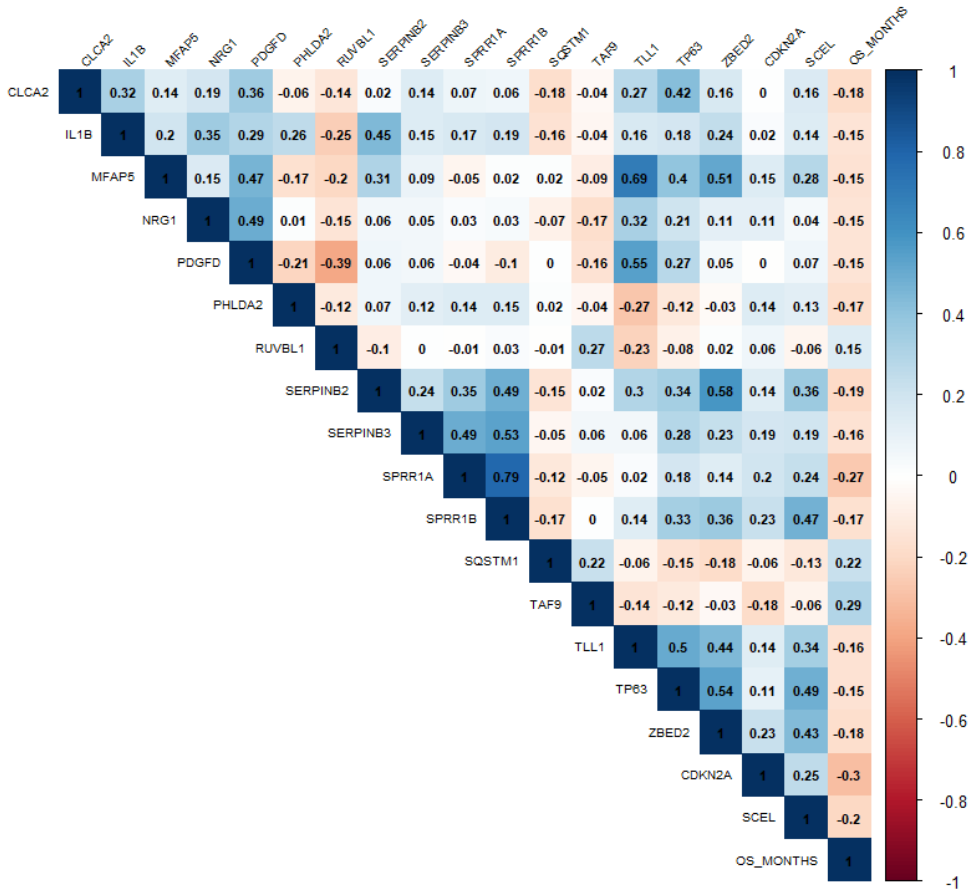


Figure 17. Correlation matrix of the 18 genes selected for the model.

The highest correlation found among the genes was 79% correlation between SPRR1A and SPRR1B, so no gene was excluded for further analysis.

Regarding the correlation with OS, 15 of the 18 genes were found negatively correlated with OS, which means that the higher the expression of those genes, the lower the overall survival. Interestingly, the other 3 genes (RUVBL1, SQSTM1 and TAF9) were found positively correlated, which means that the higher the expression of these genes, the higher the overall survival.
The highest correlations with OS were found for CDKN2A (-30%) followed by TAF9 (29%), SPRR1A (-27%) and SQSTM1 (22%).

4.4.3 Evaluation of the *Random forest* model

The following set of parameters were tested (data not shown):

1. Cross-validation with 5 folds and 3 repeats
2. Cross-validation with 3 folds and 0 repeats
3. Split into train/test following a proportion of 70%/30%
4. Split into train/test following a proportion of 65%/35%
5. Split into train/test following a proportion of 60%/40%

Best metrics were obtained using a cross-validation with 3 folds and 0 repeats, and splitting the data into 60% for the training set and 40% for the testing set. This is probably due to the small size of the dataset. A total number of 10 iterations was performed. The following metrics have been used to evaluate the model: accuracy, sensitivity and specificity.

Iteration	Accuracy	Sensitivity	Specificity
1	65.96%	62.50%	69.57%
2	51.06%	54.17%	47.83%
3	53.19%	33.33%	73.91%
4	51.06%	41.67%	60.87%
5	51.06%	66.67%	34.78%
6	55.32%	50%	60.87%
7	51.06%	41.67%	60.87%
8	46.81%	45.83%	47.83%
9	57.45%	62.50%	52.17%
10	65.96%	62.50%	69.57%
Average	**54.89%%**	**52.08%%**	**57.83%**

Table 19. Metrics corresponding to the 10 iterations of RF and their average

Results showed a low performance of the model, although surpassing 50% in all of the three metrics evaluated. Average specificity almost reached 60% (**57.83%**), however average sensitivity remained close to 50% (**52.08%).** Overall, the average accuracy of the 10 iterations was 54.89%. The variability of the metrics observed across the iterations could be due to the small size of the sample used to generate the model (119 cases). Confusion matrix of the last iteration is shown below (**Figure 18**).

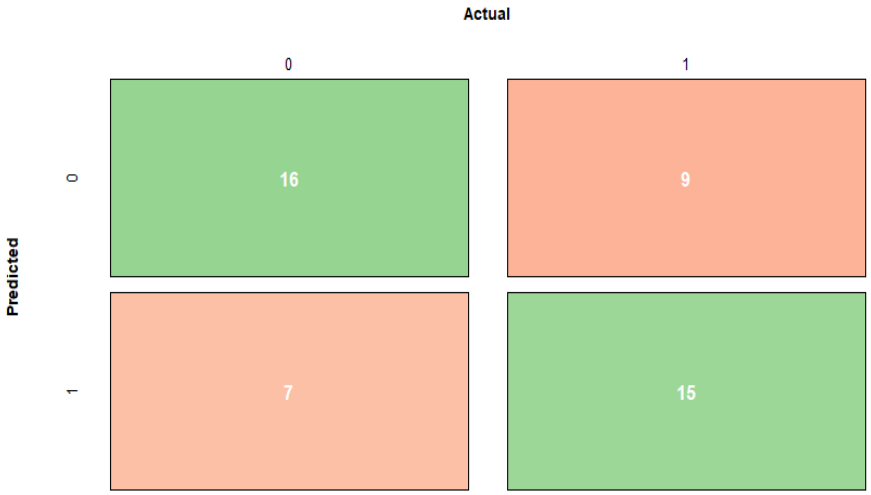


Figure 18. Confusion matrix of the 10th iteration of RF.

The relative importance of the genes was also retrieved (**Figure 19**). The genes with major impact in the model were TAF9 and SPRR1A, which is consistent with findings obtained in the correlation analysis (TAF9 was the second and SPRR1A was the third gene with highest correlation with 29% and -27%, respectively).

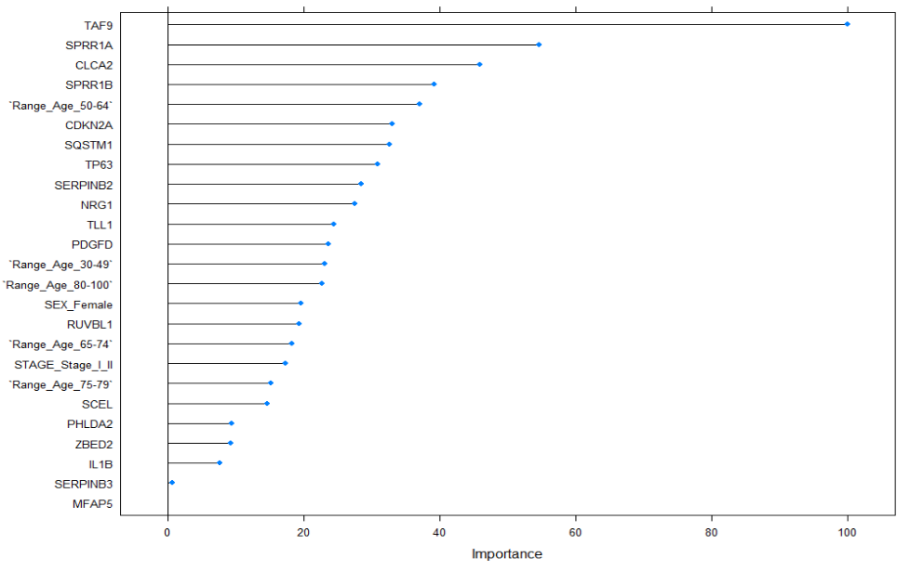


Figure 19. List of variables ranked by their relative importance in the model

4.4.4 Inclusion of demographic and clinical variables

In order to increase the performance of the model, three additional variables were included:

1. **Age** of the patient (discretized in ranges: 30-49, 50-64, 65-74, 75-89, 80-100). Ranges were selected taking into account the distribution of the Age

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
34.0    65.5    74.0    70.7    79.0    90.0
```

The frequency table of the variable “Range Age” is the following:

```
30-49 50-64 65-74 75-79 80-100
   12    18    41    23     25
```

The exploratory analysis (**Figure 20**) is shown below:

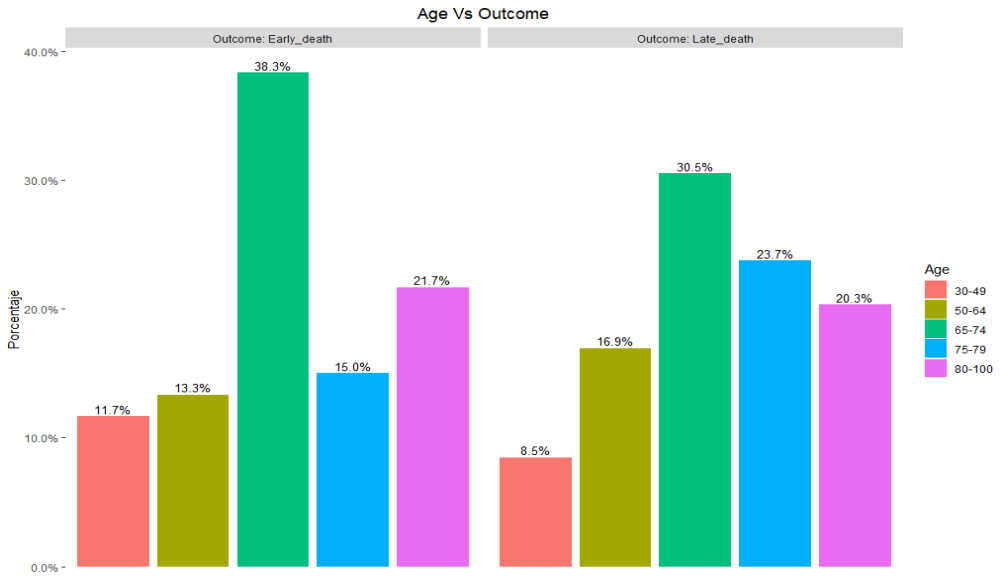


Figure 20. Exploratory analysis of variable "Range age" in early/late death groups

There are some ranges of Age (65-74 and 75-79) with noticeable differences of frequency among the groups. While it seems that there is a higher frequency of patients between 65 and 74 years that die earlier, the opposite is found for the range between 75-79 years. However, if comparing all the ranges, these differences are not statistically significant as per the results of the Chi-squared test (p-value = 0.6837).

2. **Sex** of the patient (Female, Male). The frequency table of the variable is the following:

```
Female   Male
    55     64
```

The exploratory analysis (**Figure 20**) is shown below:

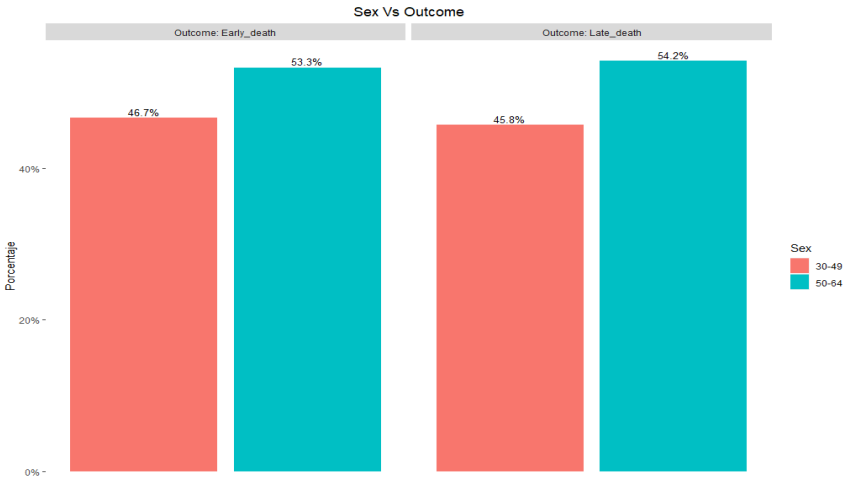


Figure 21. Exploratory analysis of variable "Sex" in early/late death groups

The analysis showed no differences of frequency among the groups. This is confirmed by the Chi-squared test (p-value = 1).

3. **Stage of the** tumour: Stage I-II, Stage III-IV. The frequency table of the variable is the following:

```
Stage_I_II Stage_III_IV
        40           74
```

The exploratory analysis (**Figure 22**) is shown below:

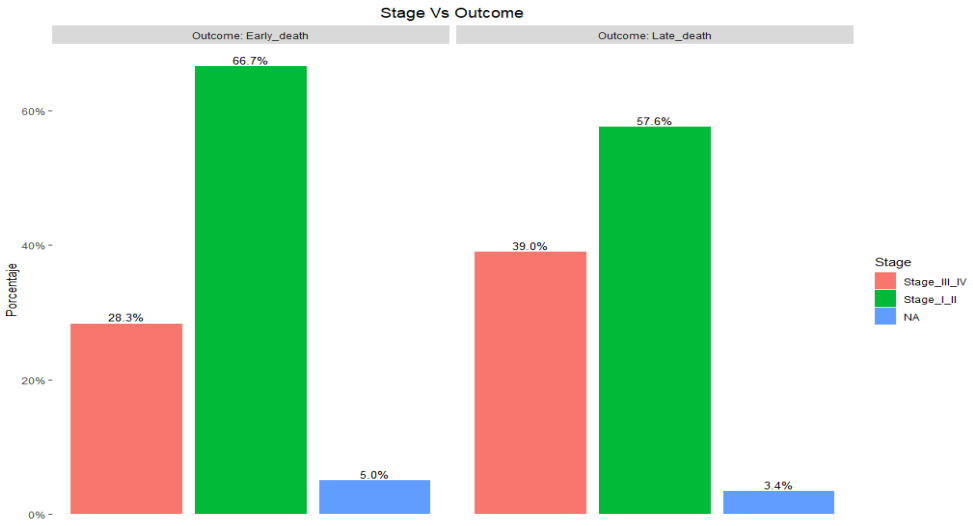


Figure 22. Exploratory analysis of variable "Stage" in early/late death groups

The analysis showed the presence of 5 missing values (3 in "*Early_death*" and 2 in "*Late_death*". It seems there are differences between groups, as expected proportionally later stages are found in the "*Early_death*" group. However, no statistically significative differences were found as per the results obtained by the Chi-squared test (p-value = 0.3265).

These results of not finding any statistically significant differences in "Range age" and "Stage" might be due to the size of the sample analyzed. Despite not finding statistically significant differences in these two variables, they were included in the model to check whether they could improve the model.

4.4.5 Evaluation of the *Random forest* model including demographical and clinical data

Results of the 10 iterations and the computed average are shown below (**Table 20**).

Iteration	Accuracy	Sensitivity	Specificity
1	61.36%	54.55%	68.18%
2	54.55%	72.73%	36.36%
3	59.09%	59.09%	59.09%
4	50%	40.91%	59.09%
5	59.09%	72.73%	45.45%
6	61.36%	68.18%	54.55%
7	65.91%	68.18%	63.64%
8	50%	45.45%	54.55%
9	59.09%	68.18%	50%
10	63.64%	59.09%	68.18%
Average	**58.41%**	**60.91%**	**55.91%**

Table 20. Metrics corresponding to the 10 iterations of RF and their average (including demographical and clinical variables)

Results including the demographical ("Range age", "Sex") and clinical data ("Stage") showed an improvement in the performance of the model: a 8% increase in sensitivity surpassing the 60% (**60.91%**) which made the accuracy increase 4% until reaching almost 60% (**58.41%**) despite a slight decrease in specificity by 2% (**55.91%**). Confusion matrix of the last iteration is shown below (**Figure 18**).

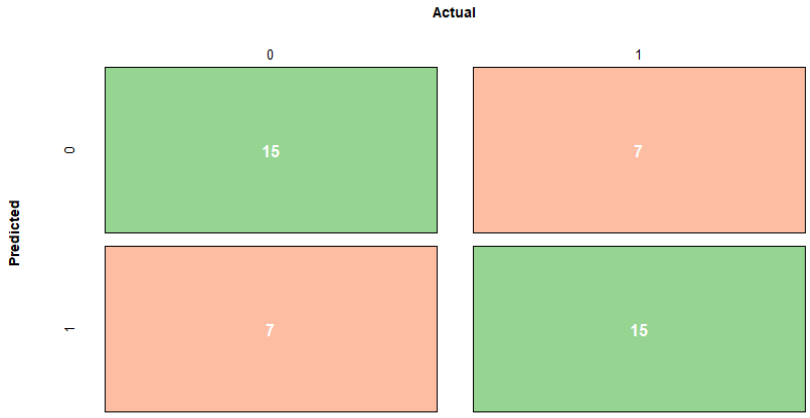


Figure 23. Confusion matrix of the 10th iteration of RF (adding demographical and clinical variables).

The relative importance of the variables was also retrieved (**Figure 24**). The variables with major impact in the model were the genes TAF9 and CDKN2A, which is again consistent with findings obtained in the correlation analysis (TAF9 was the second and CDKN2A the first gene with highest correlation with 29% and -30%, respectively). Variable "Stage" had also a significant impact on the model as it was shown to be ranked the 4th variable in order of importance.

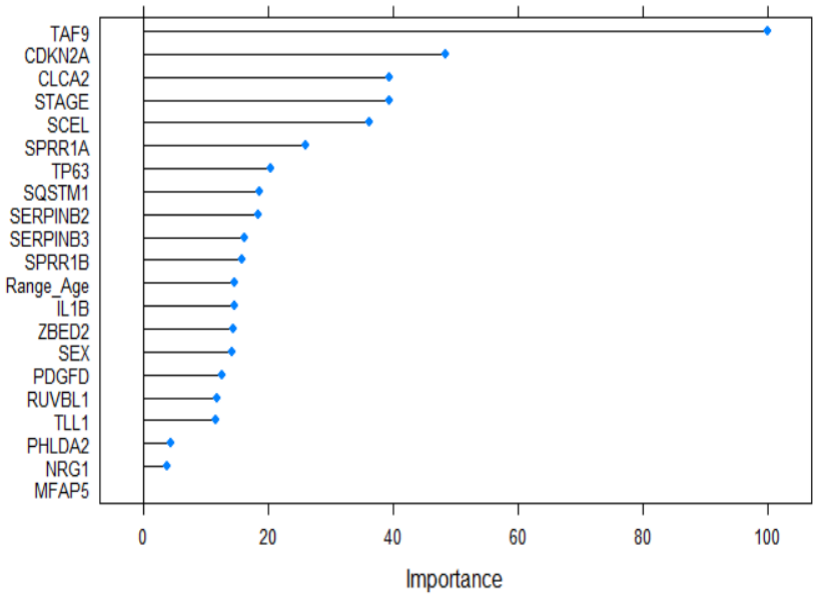


Figure 24. List of variables ranked by their relative importance in the model

4.4.6 Comparison with other CSC-gene signatures

Last, the results obtained by the "18-CSC gene signature" together with the three demographical and clinical variables were compared to other two published CSC-gene signatures (already reviewed in section 2.2.1):

1. **12-CSC gene signature** related to stemness (Pei at al., 2020)

Genes are: CCNB2, CDC20, CENPA, EXO1, FOXM1, KIF4A, PLK1, RAD54L, SGOL1, SKA1, TPX2 and TTK

2. **20-CSC gene signature** related to prognosis (Pece et al., 2019)

   Genes are: THOC4, APOBEC3B, CDK1, CENPW, EIF4EBP1, EPB41L5, EXOSC4, H2AFJ, H2AFZ, LY6E, C17orf37, MMP1, MRPS23, NDUFB10, NOL3, PHB, PHLDA2, RACGAP1, SFN and TOP2A.

A. Correlation with OS (Pearson coefficient)

Correlation analysis of the genes included in both signatures are presented below (**Figure 25**, **¡Error! No se encuentra el origen de la referencia.**).

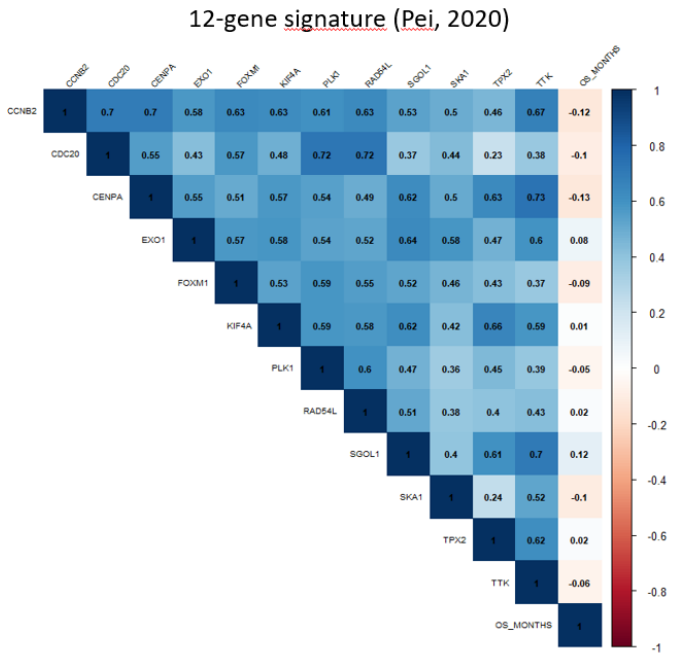


Figure 25. Correlation analysis of genes from the 12-CSC gene signature (Pei et al., 2020). Correlation with OS (in months)

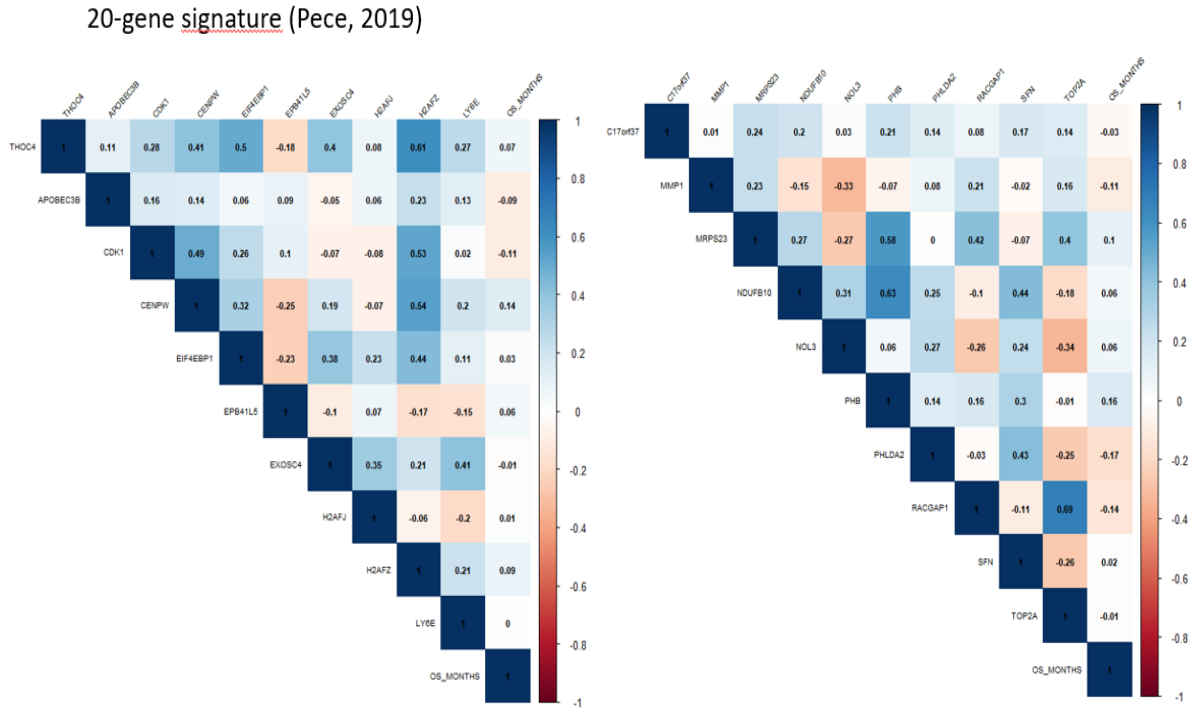


Figure 26. Correlation analysis of genes from the 20-CSC gene signature (Pece et al., 2019). Correlation with OS (in months)

The results of the analysis showed a maximum degree of correlation of -13% (gene CENPA) among all the genes included in the 12-CSC gene signature while the maximum degree of correlation among the genes included in the 20-CSC gene signature was of -17% (gene PHLDA2). These results are lower than what has been observed with the genes of the 18-CSC gene signature developed in the present work, with correlations reaching -30% (gene CDKN2A) plus 4 other genes reaching correlations over ±20% (TAF9, SPRR1A, SQSTM1 and SCEL with 29%, -27%, 22% and -20%, respectively).

B. Evaluation of the models

The three demographical and clinical variables were included in the models and 10 iterations were performed in both cases.

Results are displayed below (**Table 21**, **¡Error! No se encuentra el origen de la referencia.**):

12-gene signature (Pei, 2020)

Iteration	Accuracy	Sensitivity	Specificity
1	54.55%	50%	59.09%
2	50%	59.09%	40.91%
3	61.36%	54.55%	68.18%
4	59.09%	36.36%	81.82%
5	59.09%	59.09%	59.09%
6	50%	59.09%	40.91%
7	63.64%	54.55%	72.73%
8	54.55%	54.55%	54.55%
9	40.91%	27.27%	54.55%
10	52.27%	50%	54.55%
Average	**54.55%**	**50.46%**	**58.64%**

Table 21. Metrics corresponding to the 10 iterations of RF and their average (including demographical and clinical variables)

20-gene signature (Pece, 2019)

Iteration	Accuracy	Sensitivity	Specificity
1	45.45%	54.55%	36.36%%
2	52.27%	45.45%	59.09%
3	52.27%	63.64%	40.91%
4	43.18%	40.91%	45.45%
5	59.09%	54.55%	63.64%
6	59.09%	59.09%	59.09%
7	45.45%	54.55%	36.36%
8	59.09%	54.55%	63.64%
9	45.45%	36.36%	54.55%
10	54.55%	59.09%	50%
Average	**51.59%**	**52.27%**	**52.53%**

Table 22. Metrics corresponding to the 10 iterations of RF and their average (including demographical and clinical variables)

The relative importance of the variables was also retrieved for both models- Figures can be found in **Appendix 9.4** (Supplementary Figures S6, S7).

In summary, results showed that the 18-CSC gene signature identified in the present work had the best predictive power for OS among the 3 CSC gene signatures tested. The accuracy obtained (**58.41%**) was 4% and 7% higher than the obtained by the 12-CSC gene signature (**54.55%**) and the 20-CSC gene signature (**51.59%**), respectively. Sensitivity was also the highest among the 3 gene signatures (**60.91%**), being 10% and 8% higher that the obtained by the 12-CSC gene signature (**50.46%**) and the 20-CSC gene signature (**52.27%**), respectively. However, specificity (**55.91%**) was 3% lower than

the one obtained by the 12-CSC gene signature (**58.64%**) but 3% higher than the one obtained by the 20-CSC gene signature (**52.53%**).

	Accuracy	Sensitivity	Specificity
18-gene signature (this work)	**58.41%**	**60.91%**	**55.91%**
12-gene signature (Pei, 2020)	54.55%	50.46%	58.64%
20-gene signature (Pece, 2019)	51.59%	52.27%	52.53%

Table 23. Comparison of the metrics obtained by the 3 gene signatures tested.

5. Conclusions

The main conclusions of the present work are:

1) An enrichment in selected CSC genes has been confirmed in all 6 BC and 6 MCL gene datasets studied. The highest proportion of upregulated CSC genes in BC has been found in datasets with invasive or early onset phenotypes.

2) Differences have been found in the proportion of upregulated respect to downregulated genes in both cancer types. Whereas in MCL a 74.26% of enriched CSC genes have been found upregulated, in BC a 62.84% has been found downregulated.

3) The CSC gene sets with higher proportion of upregulated genes found enriched in both cancer types are the ones related to prognosis (*PROGNOSIS_BC*) and stemness (*STEMNESS_BC*).

4) All 8 gene sets related to CSC pathways have been found enriched at least in one of the cancer types. TGF-beta (*TGF_BETA*) and Myc (*MYC*) resulted with the highest proportion of upregulated genes found in both cancer types.

5) A total number of 269 CSC genes have been found commonly enriched in both cancer types, 53 of them with a high significance (p-value<0.01). Several CSC-gene signatures have been generated: 269-CSC gene signature (with all commonly enriched genes), 242-CSC gene signature (with genes commonly enriched with a significance <0.05) and 53-CSC gene signature (with genes commonly enriched with a significance <0.01).

6) Functional annotation analysis regarding molecular function mapped almost 10% of the 269 identified genes to heterocyclic compound binding proteins. This is important as heterocycles are key structural components of many of the anti-cancer drugs available.

7) Functional annotation analysis regarding biological processes mapped almost 30% of the 269 identified genes to nitrogen compound metabolic process. This is important as nitrogen acquisition and utilization are fundamental for cell growth and proliferation.

8) Up to 51 of the 53 genes included in the 53-CSC gene signature showed a significant correlation with survival in different cancer types using Kaplan-Meier estimator. 12 of those genes were found significantly correlated with survival in all 4 cancer types analyzed (breast, ovarian, lung and gastric cancer).

9) Up to 18 of the 53 genes included in the 53-CSC gene signature showed a correlation of more than 15% with overall survival in colorectal adenocarcinoma using Pearson coefficient, being CDKN2A, TAF9 and SPRR1A the 3 genes that showed the highest correlations (-30%, 29% and -27%, respectively).

10) The 18-CSC gene signature generated in the present work achieved 55% accuracy in predicting prognosis in colorectal adenocarcinoma. Accuracy increased to 58.41% when combined with “Age”, “Sex” and “Stage. Although far from being considered as a valid clinical classifier, the predictive power of the 18-CSC gene signature is higher than the ones displayed by other CSC-gene signatures already published.

All the objectives of the present work have been fulfilled. A common gene signature related to CSC has been identified as enriched in both, BC and MCL (with various degrees of significance). A deeper analysis of the molecular and the biological significance of those genes showed some common patterns related to cancer biology and therapeutics, such as the fact that 10% of the genes mapped to heterocyclic compound binding proteins which could be targeted by many anti-cancer drugs available. The 96% of the genes included in the 53-CSC-gene signature proved to be significantly correlated to prognosis in at least one of the cancer types studied (breast, ovarian, lung and gastric cancer) while a subset of selected 18 genes exhibited a predictive power for prognosis in cancer that was higher than other CSC gene signatures published.

The main difficulty found in the project has been the search and retrieval of valid data for accomplishing the objectives, especially for the generation of the predictive model. Many scientific articles related to cancer are published every day but few are found in which both, expression and clinical data, are available for each of the cases studied. There are public repositories in which authors of experimental and clinical studies can load their data but the majority relate to mutation and not so much to expression studies. This fact has constrained the extension of the present work as the predictive power of the gene signature has been evaluated only in a single dataset with limited size.

6. Future work

In the future it would be interesting to continue with the following research:

1. Include more BC and MCL expression datasets and repeat GSEA analysis to refine the CSC-gene signature.

2. Evaluate a second CSC-gene signature not composed by those genes with the highest significance but by those genes found enriched in the highest number of datasets (both, BC and MCL).

3. Evaluate the predictive power for prognosis of the gene signatures in a variety of cancer types.

4. Study the prognostic significance of TAF9, especially in colorectal carcinoma. This has been an unexpected finding of the present work. TAF9 (TATA-Box Binding Protein Associated Factor 9), a gene involved in transcriptional activation, has been the second gene found in the present work with the highest correlation with prognosis (29%), with a higher expression conferring longer overall survival. Whereas for CDKN2A, the gene found with the highest correlation with OS (-30%), it has already been studied its role as diagnostic ((Oh et al., 2020) and prognostic biomarker for colorectal cancer (Marcuello et al., 2019), nothing has been published concerning the potential role of TAF9 as a prognostic biomarker in colorectal cancer (no results found when searching for "*prognosis biomarker colorectal cancer TAF9*" in PubMed.gov). Only one article from 2009 (Krasnov et al., 2009) has discovered the overexpression of the gene at a protein level in colon cancer tissue but no relation to prognosis has been established so far.

7. Glossary

BC: Breast Cancer, 1

BIDC: Breast Invasive Ductal Carcinoma, 7

BRCA: Breast Invasive Carcinoma, 9

Circulating Free DNA, 5

CNVs: Copy Number Variants, 9

CSCs: Cancer Stem Cells, 1

DFS: Disease Free Survival, 7

EMT: Epithelial Mesenchymal Transition, 9

ES: Enrichment Score, 15

FDR: False Discovery Rate, 16

GCT: Gene Cluster Text, 15

IC50: Inhibiting Concentrarion 50, 6

MCL: Mantle Cell Lymphoma, 1

MCL-ICs: Mantle Cell Lymphoma - Initiating Cells, 5

MGSEA: Multivariate Gene Enrichment Analysis, 9

ML: Machine Learning, 9

NES: Normalized Enrichment Score, 16

NOD/SCID: Non-Obese Diabetic/Severe Combined Immunodeficient immunocompromised mice, 5

OCLR: One-Class Logistic Regression algorithm, 8

OS: Overall Survival, 7

RFS: Relapse-Free-Survival, 5

RMES: Regulatory Module Enrichment Score, 9

ROS: Reactive Oxygen Species, 5

SVM: Support Vector Machine, 9

TCGA: The Cancer Genome Atlas, 8

TNBC: Triple Negative Breast Cancer, 5

TTM: Time To Metastasis, 5

8. Bibliography

Al-Hajj et al., 2. (2003). Prospective identification of tumorigenic breast cancer cells. *PNAS, 100*, 3983-3988.

Arfaoui et al., 2. (2019). A genome-wide RNAi screen reveals essential therapeutic targets of breast cancer stem cells. *EMBO Mol Med, 11*.

Armstrong et al., 2. (2001). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genetics, 30*.

Batlle et al., 2. (2017). Cancer stem cells revisited. *Nat Med, 23*, 1124-1134.

Britton et al., 2. (2012). Breast cancer, side population cells and ABCG2 expression. *Cancer Lett, 323*, 97-105.

Chen et al., 2. (2010). A hierarchy of self-renewing tumor-initiating cell types in glioblastoma. *Cancer cell, 17*, 362-375.

Chen et al., 2. (2010). Prospective isolation of clonogenic mantle cell lymphoma-initiating cells. *Stem Cell Research, 5*, 212-225.

Chen et al., 2. (2012). A restricted cell population propagates glioblastoma growth after chemotherapy . *Nature, 488*, 522-526.

De Angelis et al., 2. (2019). Breast Cancer Stem Cells as Drivers of Tumor Chemoresistance, Dormancy and Relapse: New Challenges and Therapeutic Opportunities. *Cancers (Basel), 11*, 10.

Huaiyu et al., 2. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucl. Acids Res.*

Jung et al., 2. (2011). Stem-like tumor cells confer drug resistant properties to mantle cell lymphoma. *Leuk Lymphoma, 52*, 1066-1079.

Jung et al., 2. (2012). Bortezomib-resistant nuclear factor κB expression in stem-like cells in mantle cell lymphoma. *Exp Hematol, 40*, 107-118.

Kim et al., 2. (2015). A subset of CD45+/CD19 - cells in bone marrow may be associated with clinical outcomes of patients with mantle cell lymphoma. *Leuk Lymphoma, 56*, 3052-3057.

Krasnov et al., 2. (2009). Proteomic Expression Analysis of Human Colorectal Cancer: Of Soluble Overexpressed Proteins. *Mol Biol (Mosk)*, 610-615.

Kreso et al., 2. (2012). Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science, 339*, 543-548.

Kurtova et al., 2. (2015). Blocking PGE2-induced tumour repopulation abrogates bladder cancer chemoresistance. *Nature, 517*, 209-213.

Lapouge et al., 2. (2012). Skin squamous cell carcinoma propagating cells increase with tumour progression and invasiveness. *EMBO J, 31*, 4563-4575.

Lathia et al., 2. (2019). The clinical impact of cancer stem cells. *The Oncologist, 24*, 1-9.

Li et al., 2. (2017). Comprehensive tissue-specific gene set enrichment analysis and transcription factor analysis of breast cancer by integrating 14 gene expression datasets. *Oncotarget, 8*, 6775-6786.

Li et al., 2. (2020). Chemotherapeutic Stress Influences Epithelial–Mesenchymal Transition and Stemness in Cancer Stem Cells of Triple-Negative Breast Cancer. *International Journal of Molecular Sciences, 21*, 404.

Liu et al., 2. (2019). Detection of breast cancer stem cell gene mutations in circulating free DNA during the evolution of metastases. *Breast Cancer Res Treat, 178*, 251-261.

Luanpitpong et al., 2. (2018). Reactive oxygen species mediate cancer stem-like cells and determine bortezomib sensitivity via Mcl-1 and Zeb-1 in mantle cell lymphoma. *Biochim Biophys Acta Mol Basis Dis, 11*, 3739-3753.

Malta et al., 2. (2018). Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell, 173*, 338-354.

Marcuello et al., 2. (2019). Circulating biomarkers for early detection and clinical management of. *Molecular Aspects of Medicine*, 107-122.

Mukherjee et al., 2. (2017). Modulation of SOX2 expression delineates an end-point for paclitaxel-effectiveness in breast cancer stem cells. *Sci Rep, 7*, 9170.

Nagy et al., 2. (2018). Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Scientific Reports*.

Nassar et al., 2. (2016). Cancer Stem Cells: Basic Concepts and Therapeutic Implications. *Annu Rev Patol, 11*, 47-76.

O'Brien et al., 2. (2007). A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. *Nature, 445*, 106-110.

Oh et al., 2. (2020). Novel biomarkers for the diagnosis and prognosis of. *Intestinal Research*, 168-183.

Palomeras et. al, 2. (2018). Targeting Breast Cancer Stem Cells to Overcome Treatment Resistance. *Molecules, 23*, 2193.

Pece et al., 2. (2019). Identification and clinical validation of a multigene assay that interrogates the biology of cancer stem cells and predicts metastasis in breast cancer: A retrospective consecutive study. *EBioMedicine, 42*, 352-362.

Pei at al., 2. (2020). Identification of key genes controlling breast cancer stem cell characteristics via stemness indices analysis. *J Transl Med, 18*.

Prasad et al., 2. (2019). Cancer cells stemness: A doorstep to targeted therapy. *Biochim Biophys Acta Mol Basis Dis, 1866*.

Qiu et al., 2. (2019). A multiple breast cancer stem cell model to predict recurrence of T1-3, N0 breast cancer. *BMC Cancer, 19*, 729.

Roesch et al., 2. (2013). Overcoming intrinsic multidrug resistance in melanoma by blocking the mitochondrial respiratory chain of slow-cycling JARID1B(high) cells. *Cancer cell, 23*, 811-825.

Roschewski, M. (2015). Mantle Cell Lymphoma: An Evolving Therapeutic Landscape. *Federal Practitioner*, 50S-53S.

Society, A. C. (2020). *www.cancer.org.* Obtenido de https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf

Wright et al., 2. (2008). Brca1 breast tumors contain distinct CD44+/CD24- and CD133+ cells with cancer stem cell characteristics. *Breast Cancer Res, 10*.

Yu et al., 2. (2020). The molecular markers of cancer stem cells in head and neck tumors. *J Cell Pysiol, 235*, 65-73.

Zhao et al., 2. (2020). TRIP6 enhances stemness property of breast cancer cells through activation of Wnt/β-catenin. *Cancer Cell Int, 20*.

9. Appendix

9.1 List of genes included in the gene sets

- *"Prognosis_BC":*

  EIF4EBP1, LY6E, NOL3, EXOSC4, ALYREF, PHB, NDUFB10, MRPS23, CDK1, TOP2A, CENPW, APOBEC3B, H2AFJ, H2AFZ, RACGAP1, EPB41L5, PHLDA2, MIEN1, SNF, MMP1, TRIP6, SGO1, KIF4A, CDC20, PLK1, FOXM1, SKA1, RAD45L, SGO1, CENPA, CCNB2, EXO1, TPX2, MSK1, GPR77, CD10, TAZ, MLF2, RPL39, HN1L, HGA6, STMN1, PROCR, PLAG2G16, OCT4, CTNNB1, LGR5, H19, MYC, HER2, CLI1, CGI99, CDK4, CD133, MKI67, CD24, BMI1, CD44, ALDH1A3, ALDH1A1, ALDH1, ACBD3

- *"Stemness_BC":*

  BUB1B, CDCA3, DLGAP5, SGO1, FOXM1, SKA1, AURKB, BUB1, CDC20, KIF23, CDC45, ORC1, KIF18B, KIF20A, RAD54L, NCAPH, CEP55, NCAPG, NDC80, MELK, CDC25A, KIF4A, TTK, EXO1, KIF2C, CCNB2, CENPA, KIFC1, PLK1, CDCA8, HJURP, TPX2, GLI1, Adipsin, SMAD, TGF-beta, STAT3, TWIST1, SOX2, ABCG2, OTUB2, NOTCH2, SHIP, PCGF4/BMI1, SNAIL, SOX9, SOX2, INTEGRIN_BETA4, ANTXR1, CXCR4, CD49f, CD61, CD133

- *"Stemness2_BC":*

  ERCC6, RAD23B, SLC4A3, CA6, DDX59, NTHL1, GNAS, SLC6A8, EIF2B4, RBL1, PRTFDC1, RARA, DDX41, STAMBPL1, GUSB, SLC37A3, CRMP1, TMX1, NAIP, AFP, MRC1, MTA1, SLC25A18, HLA-DRB1, CNTN1, ALDH1A1, TIPARP, POMT2, RIT2, MLL, MED1, GYS2, RCAN1, FAAH, HMG20A, PPAT, HSPA14, KIF2A, APBA3, PRDX1, SCAMP3, IL36B, NPTXR, CLCF1, PTDSS2, IL6, UBN1, VARS, AARSD1, PDE6G, NFRKB, CSGALNACT2, SLC1A1, SLC9A5, MSH5, NAA15, SLC40A1, DNASE2B, SEC24A, NCOR1, NR3C1, RNASE3, HOXC4, ETV5, ACADS, PLA2G2E, STX1A, ASAH2, FKBP2, NAT8L, DPF1, MDM4, RNF146, BAZ1B, LOC729974, PCGF6, RNF43, RNF213, MEFV, FBXO31, GALK1, GLP2R, G6PC2, HACE1, BIRC7, KCNJ10, PKD1L2, KCNK12, SLC9A3, PTPLB, ITPR3, GRIN3B, GRIA3, ADORA2B, GPR152, SORCS1, LTB4R2, ICK, PACSIN1, ADCK4, PKN2, PIM1, PI4KB, C19orf34, ZNF721, C18orf51, NPVF, C11orf38, CBLN4, ANKRD39, CLEC4A, CCPG1, PSAP, C3orf37, L1TD1, CCDC56, TTC31, GLT25D1, CTDSPL2, ARD1B, SAMD9L, BFSP1, STARD3, STXBP4, DMRTC1B, KLHDC3, C16orf62, NME1-NME2, PCP2, TIMM17B, HNRNPH1, C7orf50, DOCK2, GTF2H5, ZNF543, TNNT3, ZNF135, CD52, C9orf139, GSX2, ADNP2, TMEM106C, SMAP2, C1QL4, ITFG2, MUPCDH, TMEM128, RBM4B, FAM83H, HNRPLL, BATF2, KLHDC6, C3orf44, DTD1, IGFL1, ZBTB6, ZNF706, TSHZ3, MSLNL, SAMD13, FASLG, YRDC, CEP68, C9orf164, CD70, PNO1, CLRN3, TMSL8, PLEKHA9, PRAM1, C5orf34, ZNF675, OVOS2, CCDC51, LILRA5, RIC8B, CCDC94, NENF, TSC22D2, DMKN, SHISA4, DAGLA, C1orf84, FAM23B, PELI3, ESF1, ZNF746, CIAPIN1, KIAA1755, C17orf59, C16orf58, BET1L, ZNF446, PAQR5, ZNF513, DEFA3, LASS2, SYS1, DNAJC7, OR9Q2, RPL7A, PEX16, FAM13A1, KIAA0391, DEFA4, OR8B12, ECHDC1, MARVELD2, UNC84B, CORO1B, FAM168B, LENG8, OR51G2, NMB, HPS1, SLC39A11, RPRD1A, STX11, MAD1L1, PION, SNIP1, RPL38, PSAPL1, HLA-DRB5, LOC57228, SCAP, LRRC27, LRRN4, A26B1, FBLN2, SNX32, CYGB, AP1S2, C14ORF48, MPZL1, CAV2, LRRTM1, ITM2C, HMGA2, PCDHA13, ZSCAN1, COL27A1,

C12orf67, TMEM57, IGFBP1, FAM24A, SNRNP70, USHBP1, ADCYAP1, CEP97, OR6B2, CCDC7, POU5F2, FEV, C7ORF26, CLTCL1, MAPK8IP1, LGI4, RPLP0, CSPG4, RPL3, IRX2, ZNF202, JMJD1B, TIMM8B, GPSM1, ZFP82, MRPS17, FAM36A, NCSTN, ZNF385A, FZR1, OR10Z1, OR1K1, LRBA, TBC1D16, SFTPB, NKAP, IQCA1, C10orf54, MMP23B, NECAP1, TNFRSF6B, H1FOO, FCGR1B, MFAP4, OR56B4, SYCN, C9ORF16, LRFN3, PECAM1, OR52M1, LOC100129550, RALGPS1, GGCT, ARL4C, RUNDC3A, INSL5, OMD, PSMF1, KRTAP19-4, LSM6, FSTL4, NDUFV3, SS18L2, TARP, SNX20, BCL9L, API5, TOB2, YDJC, RHOT2, ADRM1, STAP1, ZNF324, OR5D13, OR11G2, MAGEE2, FAR2, MED30, COL6A3, PCDH8, PDLIM3, TREML1, RPL18, GSTT2B, OLFM2, OR1B1, FAM180A, UNC5D, GLTSCR2, ZNF600.

- *"Dormancy_BC":*

  DYRK1A, MAPK, TK1, TIMP3, PLAT, PIK3CB, ODC1, NT5E, MMP1, JUN, IL8, IGF1R, FOXM1, FOXD1, FOSL1, ESM1, EGR1, EGFR, DTYMK, DNMT1, CKS2, CEBPG, CDKN3, BUB1B, BUB1, ATF4, ATF3, ASNS, APEX1, TPM1, TP53, THBS1, TGFB2, STAT3, SREBF1, SOX9, P4HA1, NR2F1, MMP2, IGFBP5, HIST1H2BK, GATA6, EPHA5, DDR1, CTSD, COL4A5, COL1A1, BHLHE41, AMOT, ADAM10, ACVR1.

- *"Chemorresistance_BC":*

  ABCB5, ABCG2, ABCB1, STAT3, GPR77, CD10, BRD4, GPX4, CD133, ATG7, TRAIL, NRF2, PERK, CD44, ALDH family, TWIST1, SOX2, ABCG2, NOTCH1, TSPAN8, FGF5.

- *"GSE25976_OVER":*

  CSF2RA, HOOK1, CYP1B1, MFHAS1, RNA5SP449, LCN2, KIF16B, CCL20, NAP1L2, OTULINL, OLR1, SNORD116-15, TLCD4, PLCB4, TAGLN, SLC7A7, HSPA6, PAPPA, ZNF204P, SLAMF7, PTGFRN, IL13RA2, KLF8, GALNT5, TENM2, GPR65, L3MBTL4, CTSV, CRISPLD2, SLCO1B7, DCLK1, F2RL2, MCTP2, SNORD116-20, ARHGAP28, ALPK2, OLFML1, SNORD116-19, OR2A20P, CYP24A1, UCA1, SPARC, GNGT2, MAMDC2, SNORD116-13, MPZL2, MATN2, TNFSF10, LCP1, MYO1D.

- *"GSE25976_UNDER":*
  TNFSF18, GABRA3, HSD17B2, SPANXA1, CCBE1, HHIP, ITGA10, SPANXB1, AADAC, HAS2, PPIAP47, XAGE2, RGCC, RNU6-256P, SERPINB2, RNA5SP180, TLR2, SPANXD, CHRDL1, AZGP1, MMP1, COL6A3, TIE1, COX7B2, OTOGL, PCDHB15, MRGPRX3, SEMA3A, RNU6-729P, TMPRSS15, SSX1, LAMA4, AC002316.1, MAGEA1, EHF, RNA5SP366, SIDT1, RNA5SP183, ZNF521, CDH11, PRSS2, RNA5SP30, TGIF2LX, RNU6-888P, TMEM163, NCKAP1L, JPH1, RNA5SP55, H2BC1, DSCR8.

- *"GSE43730_OVER":*

  ATP8A1, DOCK10, ABCA6, RNU6-893P, RNY1P14, RNA5SP330, RNA5SP110, SCARNA10, LAMP3, PDE7B, RNU6-23P, RN7SL153P, IFI44L, LPL, RNA5SP484, SMOX, STC1, TMEM156, PPEF1, RNA5SP219, OR5M6P, SULF1, SLC27A2, ADAM12, ANGPT1, ADGRL3, MMP16, MACC1, SNAP25, PPP1R9A, SRGN, SNORD63, SLIT2, CDH2, RNA5SP242, MSL3P1, DCLK1, TBL1X, DCN, RNA5SP494, RN7SKP35, SELENOP, SNORD13P1, CHGB, SNORD1C, CCDC102B, LINGO2, FPR1, EYA1, ERBB2.

- *“GSE43730_UNDER”:*

  RNA5SP247, SERPINB13, SERPINB2, MFAP5, SERPINB3, NRG1, SPRR1B, RNU6-1263P, RNU5A-1, MIR205, RNU6-1208P, RNU6-155P, FAT2, IRF6, RNA5SP354, CYP4F11, RNF144B, TP63, RNA5SP191, PI3, SPRR2A, PDSS1P1, PLD5, RN7SL378P, RNU6-674P, CLCA2, RNF152, EPB41L4A, ZBED2, RNY3P13, RN7SL452P, GCNT4, SCEL, OR10T1P, TLL1, TRAF3IP3, CCDC80, RNU6-597P, XDH, FOLH1, DENND2C, POF1B, RGS2, ANK3, PDGFD, RNU6-577P, RPSAP52, SPINK7, SPRR1A, IL1B, CD24P4, RNA5SP374.

- *“GSE95042_OVER”:*

  GBP2, PLAAT4, HLA-DMA, MMP9, CD163, BGN, CXCL9, SPARC, AIF1, SPP1, SFRP2, HLA-DRB4, PDGFRB, LAPTM5, GIMAP4, RNASE1, MS4A6A, LYZ, CTSK, SPARCL1, CD14, FCER1G, HLA-DQA1, THY1, POSTN, CSF1R, CD74, VWF, AEBP1, ALOX5AP, DCN, HLA-DRB6, RGS1, HLA-DRB1, COL5A2, HLA-DPA1, TAGLN, HAVCR2, HLA-DRA, IGLL1, COL6A3, TYROBP, CCL8, COL1A2, C1QC, LUM, HLA-DMB, COL1A1, C1QB, COL3A1.

- *“GSE95042_UNDER”:*

  LCN2, GDF15, CXCL1, TRIB3, ERRFI1, PSAT1, BCAR3, PSAT1P3, GOLGA8S, TFPI2, GNE, MT-ND6, LRATD2, CXCL8, HNRNPKP4, ASNS, CAV2, TCEA1, NT5E, MAL2, ADM2, FADS1, COCH, STC2, TF, SPRY2, KCNG1, LARP6, ASS1P11, ELOVL6, IRS2, VGF, OSGIN2, CXCL5, KRT18, IL6, TMEM38B, GNG12, HES1, NAMPT, LAMP3, CDK6, OPRK1, SYBU, PLSCR1, BCAT1, TPBG, TVP23C, ACSS2, FOXA1.

- *“GSE132083_OVER”:*

  ZNF883, MTAP, BSN-AS2, THRB, ESCO2, CCNE2, KANSL1, CASS4, LINC00886, ZNF544, TXNDC9, AGMAT, SMIM17, ZNF382, RARA-AS1, SMG8, DNAJC3-AS1, LINC00654, DMC1, ZNF514, HEYL, GPR19, IFIT2, C1orf145, LONRF3, EIF3EP1, TAF1C, HIST1H2BG, TEFM, ARHGDIG, HIST1H2BD, FAM13A-AS1, PLA2G4C, GLDC, ACTN2, PPP1R26-AS1, TIGD6, NES, ADHFE1, RRN3P3, EOMES, KLHL7-AS1, LGALS8-AS1, GTF2H2C, RP11-227H15.4, ZNF253, LRGUK, KIAA0825, INTS4L1.

- *“GSE132083_UNDER”:*
  CBLN3, ZMYND10, SLC30A3, MRAP2, RPL17-C18orf32, GABRB3, ST3GAL6, C6orf1, FITM2, PARD6G, LINC00623, FAM171A2, LTBP4, LENG1, WBSCR27, SOCS1, C1orf21, ZBTB7C, ABCG1, PDZD7, SKI, FAM120C, RBPMS, FHOD1, DEXI, NOP14, KRT80, NPAS1, FAM78B, NELFA, LINC00869, RPL17P6, CALCB, ARFIP2, CCDC92, LINC00958, GOLGA2P7, COL16A1, PDE4A, DCHS1, LEPREL2, BACE1, HECW1, COL8A1, EMR1, BATF3, RNF130, EBF4, DPY19L2P2, CPQ.

- *“Hedgehog”:*

  WNT8B, WNT8A, WNT7B, WNT7A, WNT6, WNT5B, WNT5A, WNT4, WNT3A, WNT3, WNT2B, WNT2, WNT16, WNT11, WNT10B, WNT10A, WNT1, SUFU, STK36, SMO, SHH, RAB23, PTCH2, PTCH1, PRKX, PRKACG, PRKACB, PRKACA, LRP2, IHH, HHIP, GSK3B, GLI3, GLI2, GLI1, GAS1, FBXW11, DHH, CSNK1G3, CSNK1G2, CSNK1G1, CSNK1E, CSNK1D, CSNK1A1L, CSNK1A1, BTRC, BMP8B, BMP8A, BMP7, BMP6, BMP5, BMP4, BMP2.

- *“Hippo”:*

YWHAE, YWHAB, YAP1, WWTR1, WWC1, TJP2, TJP1, STK4, STK3, SAV1, NPHP4, MOB1B, MOB1A, LATS2, LATS1, DVL2, CASP3, AMOTL2, AMOTL1, AMOT.

- *"Jak_Stat":*

  TYK2, TSLP, TPO, STAT6, STAT5B, STAT5A, STAT4, STAT3, STAT2, STAT1, STAM2, STAM, SPRY4, SPRY3, SPRY2, SPRY1, SPRED2, SPRED1, SOS2, SOS1, SOCS7, SOCS5, SOCS4, SOCS3, SOCS2, SOCS1, PTPN6, PTPN11, PRLR, PRL, PIM1, PIK3R5, PIK3R3, PIK3R2, PIK3R1, PIK3CG, PIK3CD, PIK3CB, PIK3CA, PIAS4, PIAS3, PIAS2, PIAS1, OSMR, OSM, MYC, MPL, LIFR, LIF, LEPR, LEP, JAK3, JAK2, JAK1, IRF9, IL9R, IL9, IL7R, IL7, IL6ST, IL6R, IL6, IL5RA, IL5, IL4R, IL4, IL3RA, IL3, IL2RG, IL2RB, IL2RA, IL26, IL24, IL23R, IL23A, IL22RA2, IL22RA1, IL22, IL21R, IL21, IL20RB, IL20RA, IL20, IL2, IL19, IL15RA, IL15, IL13RA2, IL13RA1, IL13, IL12RB2, IL12RB1, IL12B, IL12A, IL11RA, IL11, IL10RB, IL10RA, IL10, IFNW1, IFNLR1, IFNL3, IFNL2, IFNL1, IFNK, IFNGR2, IFNGR1, IFNG, IFNE, IFNB1, IFNAR2, IFNAR1, IFNA8, IFNA7, IFNA6, IFNA5, IFNA4, IFNA21, IFNA2, IFNA17, IFNA16, IFNA14, IFNA13, IFNA10, IFNA1, GRB2, GHR, GH2, GH1, EPOR, EPO, EP300, CTF1, CSH1, CSF3R, CSF3, CSF2RB, CSF2RA, CSF2, CRLF2, CREBBP, CNTFR, CNTF, CLCF1, CISH, CCND3, CCND2, CCND1, CBLC, CBLB, CBL, BCL2L1, AKT3, AKT2, AKT1.

- *"Myc":*

  ZBTB17, TRRAP, TAF9, TAF12, TAF10, SUPT7L, SUPT3H, SKP2, RUVBL2, RUVBL1, PPP2R5A, PPP2CA, PML, PIN1, PAK2, MYC, MAX, KAT5, KAT2A, HBP1, GSK3B, FBXW7, CDKN2A, AXIN1, ACTL6A.

- *"Notch":*

  SNW1, RFNG, RBPJL, RBPJ, PTCRA, PSENEN, PSEN2, PSEN1, NUMBL, NUMB, NOTCH4, NOTCH3, NOTCH2, NOTCH1, NCSTN, NCOR2, MFNG, MAML3, MAML2, MAML1, LFNG, KAT2B, KAT2A, JAG2, JAG1, HES5, HES1, HDAC2, HDAC1, EP300, DVL3, DVL2, DVL1, DTX4, DTX3L, DTX3, DTX2, DTX1, DLL4, DLL3, DLL1, CTBP2, CTBP1, CREBBP, CIR1, APH1A, ADAM17.

- *"TGF_beta":*

  ZFYVE9, ZFYVE16, TNF, THBS4, THBS3, THBS2, THBS1, TGFBR2, TGFBR1, TGFB3, TGFB2, TGFB1, TFDP1, SP1, SMURF2, SMURF1, SMAD9, SMAD7, SMAD6, SMAD5, SMAD4, SMAD3, SMAD2, SMAD1, SKP1, RPS6KB2, RPS6KB1, ROCK2, ROCK1, RHOA, RBX1, RBL2, RBL1, PPP2R1B, PPP2R1A, PPP2CB, PPP2CA, PITX2, NOG, NODAL, MYC, MAPK3, MAPK1, LTBP1, LEFTY2, LEFTY1, INHBE, INHBC, INHBB, INHBA, IFNG, ID4, ID3, ID2, ID1, GDF7, GDF6, GDF5, FST, EP300, E2F5, E2F4, DCN, CUL1, CREBBP, COMP, CHRD, CDKN2B, BMPR2, BMPR1B, BMPR1A, BMP8B, BMP8A, BMP7, BMP6, BMP5, BMP4, BMP2, AMHR2, AMH, ACVRL1, ACVR2B, ACVR2A, ACVR1C, ACVR1.

- *"TNF":*

  TXN, TRAF2, TRAF1, TRADD, TNIK, TNFRSF1B, TNFRSF1A, TNFAIP3, TNF, TAB2, TAB1, STAT1, SQSTM1, SMPD2, SMPD1, RIPK1, RFFL, RELA, RACK1, PRKCZ, PRKCI, NSMAF, NRK, NFKB1, MAP4K5, MAP4K4, MAP4K3, MAP4K2, MAP3K7, MAP3K5, MAP3K3, MAP3K1, MAP2K7, MAP2K3, MADD, IKBKG, IKBKB, FADD, CYLD, CHUK, CAV1, CASP8, BIRC3, BIRC2, BAG4, ADAM17.

- *“Wnt_Bcatenin”:*

  WNT6, WNT5B, WNT1, TP53, TCF7, SKP2, RBPJ, PTCH1, PSEN2, PPARD, NUMB, NOTCH4, NOTCH1, NKD1, NCSTN, NCOR2, MYC, MAML1, LEF1, KAT2A, JAG2, JAG1, HEY2, HEY1, HDAC5, HDAC2, HDAC11, GNAI1, FZD8, FZD1, FRAT1, DVL2, DLL1, DKK4, DKK1, CUL1, CTNNB1, CSNK1E, CCND2, AXIN2, AXIN1, ADAM17.

9.2 List of common genes (BC and MCL)

GENE	EntrezID	Gene_Name
ACBD3	64746	acyl-CoA binding domain containing 3
ACSS2	55902	acyl-CoA synthetase short chain family member 2
ACTL6A	86	actin like 6A
ADAM12	8038	ADAM metallopeptidase domain 12
ADHFE1	137872	alcohol dehydrogenase iron containing 1
ADNP2	22850	ADNP homeobox 2
ADRM1	11047	adhesion regulating molecule 1
AEBP1	165	AE binding protein 1
ALOX5AP	241	arachidonate 5-lipoxygenase activating protein
ALYREF	10189	Aly/REF export factor
AMOT	154796	angiomotin
AMOTL1	154810	angiomotin like 1
AMOTL2	51421	angiomotin like 2
ANGPT1	284	angiopoietin 1
AP1S2	8905	adaptor related protein complex 1 subunit sigma 2
APBA3	9546	amyloid beta precursor protein binding family A member 3
APOBEC3B	9582	apolipoprotein B mRNA editing enzyme catalytic subunit 3B
ASNS	440	asparagine synthetase (glutamine-hydrolyzing)
ATF3	467	activating transcription factor 3
AURKB	9212	aurora kinase B
AZGP1	563	alpha-2-glycoprotein 1, zinc-binding
BAG4	9530	BCL2 associated athanogene 4
BAZ1B	9031	bromodomain adjacent to zinc finger domain 1B
BCAR3	8412	BCAR3 adaptor protein, NSP family member
BHLHE41	79365	basic helix-loop-helix family member e41
BMI1	648	BMI1 proto-oncogene, polycomb ring finger
BUB1	699	BUB1 mitotic checkpoint serine/threonine kinase
BUB1B	701	BUB1 mitotic checkpoint serine/threonine kinase B
C16orf58	64755	chromosome 16 open reading frame 58
C1QC	714	complement C1q C chain
CASP8	841	caspase 8
CAV1	857	caveolin 1
CAV2	858	caveolin 2
CCBE1	147372	collagen and calcium binding EGF domains 1
CCDC80	151887	coiled-coil domain containing 80
CCL20	6364	C-C motif chemokine ligand 20
CCNB2	9133	cyclin B2
CCNE2	9134	cyclin E2
CD14	929	CD14 molecule
CD163	9332	CD163 molecule
CD24	100133941	CD24 molecule
CD44	960	CD44 molecule (Indian blood group)
CD52	1043	CD52 molecule
CD74	972	CD74 molecule
CDC20	991	cell division cycle 20
CDC25A	993	cell division cycle 25A
CDC45	8318	cell division cycle 45
CDCA3	83461	cell division cycle associated 3

(continued)

GENE	EntrezID	Gene_Name
CDCA8	55143	cell division cycle associated 8
CDH2	1000	cadherin 2
CDK1	983	cyclin dependent kinase 1
CDK4	1019	cyclin dependent kinase 4
CDKN2A	1029	cyclin dependent kinase inhibitor 2A
CDKN3	1033	cyclin dependent kinase inhibitor 3
CENPA	1058	centromere protein A
CEP55	55165	centrosomal protein 55
CEP97	79598	centrosomal protein 97
CKS2	1164	CDC28 protein kinase regulatory subunit 2
CLCA2	9635	chloride channel accessory 2
COL1A1	1277	collagen type I alpha 1 chain
COL1A2	1278	collagen type I alpha 2 chain
COL3A1	1281	collagen type III alpha 1 chain
COL6A3	1293	collagen type VI alpha 3 chain
CORO1B	57175	coronin 1B
CSF1R	1436	colony stimulating factor 1 receptor
CTSD	1509	cathepsin D
CTSK	1513	cathepsin K
CXCL9	4283	C-X-C motif chemokine ligand 9
CXCR4	7852	C-X-C motif chemokine receptor 4
DCN	1634	decorin
DDR1	780	discoidin domain receptor tyrosine kinase 1
DDX41	51428	DEAD-box helicase 41
DENND2C	163259	DENN domain containing 2C
DLGAP5	9787	DLG associated protein 5
DMC1	11144	DNA meiotic recombinase 1
DNMT1	1786	DNA methyltransferase 1
DOCK2	1794	dedicator of cytokinesis 2
DVL2	1856	dishevelled segment polarity protein 2
EGFR	1956	epidermal growth factor receptor
EGR1	1958	early growth response 1
EHF	26298	ETS homologous factor
EOMES	8320	eomesodermin
EPB41L4A	64097	erythrocyte membrane protein band 4.1 like 4A
ESM1	11082	endothelial cell specific molecule 1
EXO1	9156	exonuclease 1
EXOSC4	54512	exosome component 4
FAAH	2166	fatty acid amide hydrolase
FADD	8772	Fas associated via death domain
FAT2	2196	FAT atypical cadherin 2
FCER1G	2207	Fc fragment of IgE receptor Ig
FOLH1	2346	folate hydrolase 1
FOXM1	2305	forkhead box M1
FPR1	2357	formyl peptide receptor 1
GBP2	2634	guanylate binding protein 2
GCNT4	51301	glucosaminyl (N-acetyl) transferase 4

(continued)

GENE	EntrezID	Gene_Name
GGCT	79017	gamma-glutamylcyclotransferase
GIMAP4	55303	GTPase, IMAP family member 4
GLDC	2731	glycine decarboxylase
GLI1	2735	GLI family zinc finger 1
GNE	10020	glucosamine (UDP-N-acetyl)-2-epimerase/N-acetylmannosamine kinase
GUSB	2990	glucuronidase beta
HAS2	3037	hyaluronan synthase 2
HAVCR2	84868	hepatitis A virus cellular receptor 2
HES1	3280	hes family bHLH transcription factor 1
HEYL	26508	hes related family bHLH transcription factor with YRPW motif like
HIST1H2BK	NA	NA
HJURP	55355	Holliday junction recognition protein
HLA-DMA	3108	major histocompatibility complex, class II, DM alpha
HLA-DMB	3109	major histocompatibility complex, class II, DM beta
HLA-DPA1	3113	major histocompatibility complex, class II, DP alpha 1
HLA-DQA1	3117	major histocompatibility complex, class II, DQ alpha 1
HLA-DRA	3122	major histocompatibility complex, class II, DR alpha
HLA-DRB1	3123	major histocompatibility complex, class II, DR beta 1
HLA-DRB4	3126	major histocompatibility complex, class II, DR beta 4
HLA-DRB5	3127	major histocompatibility complex, class II, DR beta 5
HLA-DRB6	3128	major histocompatibility complex, class II, DR beta 6 (pseudogene)
HSD17B2	3294	hydroxysteroid 17-beta dehydrogenase 2
ID2	3398	inhibitor of DNA binding 2
IFI44L	10964	interferon induced protein 44 like
IFNG	3458	interferon gamma
IGLL1	3543	immunoglobulin lambda like polypeptide 1
IL1B	3553	interleukin 1 beta
IRF6	3664	interferon regulatory factor 6
IRS2	8660	insulin receptor substrate 2
ITPR3	3710	inositol 1,4,5-trisphosphate receptor type 3
JUN	3725	Jun proto-oncogene, AP-1 transcription factor subunit
KIF18B	146909	kinesin family member 18B
KIF20A	10112	kinesin family member 20A
KIF23	9493	kinesin family member 23
KIF2A	3796	kinesin family member 2A
KIF2C	11004	kinesin family member 2C
KIF4A	24137	kinesin family member 4A
KIFC1	3833	kinesin family member C1
KLF8	11279	Kruppel like factor 8
LAMA4	3910	laminin subunit alpha 4
LAPTM5	7805	lysosomal protein transmembrane 5
LATS2	26524	large tumor suppressor kinase 2
LCN2	3934	lipocalin 2
LCP1	3936	lymphocyte cytosolic protein 1
LINGO2	158038	leucine rich repeat and Ig domain containing 2
LPL	4023	lipoprotein lipase

(continued)

GENE	EntrezID	Gene_Name
LRBA	987	LPS responsive beige-like anchor protein
LRFN3	79414	leucine rich repeat and fibronectin type III domain containing 3
LTBP1	4052	latent transforming growth factor beta binding protein 1
LUM	4060	lumican
LY6E	4061	lymphocyte antigen 6 family member E
LYZ	4069	lysozyme
MACC1	346389	MET transcriptional regulator MACC1
MADD	8567	MAP kinase activating death domain
MAP3K7	6885	mitogen-activated protein kinase kinase kinase 7
MAP4K4	9448	mitogen-activated protein kinase kinase kinase kinase 4
MED1	5469	mediator complex subunit 1
MELK	9833	maternal embryonic leucine zipper kinase
MFAP5	8076	microfibril associated protein 5
MIEN1	84299	migration and invasion enhancer 1
MKI67	4288	marker of proliferation Ki-67
MLF2	8079	myeloid leukemia factor 2
MMP1	4312	matrix metallopeptidase 1
MMP2	4313	matrix metallopeptidase 2
MMP9	4318	matrix metallopeptidase 9
MOB1B	92597	MOB kinase activator 1B
MRPS17	51373	mitochondrial ribosomal protein S17
MRPS23	51649	mitochondrial ribosomal protein S23
MS4A6A	64231	membrane spanning 4-domains A6A
NCAPG	64151	non-SMC condensin I complex subunit G
NCAPH	23397	non-SMC condensin I complex subunit H
NCSTN	23385	nicastrin
NDC80	10403	NDC80 kinetochore complex component
NDUFB10	4716	NADH:ubiquinone oxidoreductase subunit B10
NECAP1	25977	NECAP endocytosis associated 1
NES	10763	nestin
NFKB1	4790	nuclear factor kappa B subunit 1
NR2F1	7025	nuclear receptor subfamily 2 group F member 1
NRG1	3084	neuregulin 1
P4HA1	5033	prolyl 4-hydroxylase subunit alpha 1
PAK2	5062	p21 (RAC1) activated kinase 2
PDE7B	27115	phosphodiesterase 7B
PDGFD	80310	platelet derived growth factor D
PHLDA2	7262	pleckstrin homology like domain family A member 2
PI3	5266	peptidase inhibitor 3
PIM1	5292	Pim-1 proto-oncogene, serine/threonine kinase
PLAAT4	5920	phospholipase A and acyltransferase 4
PLD5	200150	phospholipase D family member 5
PLK1	5347	polo like kinase 1
PLSCR1	5359	phospholipid scramblase 1
PML	5371	promyelocytic leukemia
POF1B	79983	POF1B actin binding protein
POMT2	29954	protein O-mannosyltransferase 2

(continued)

GENE	EntrezID	Gene_Name
PPAT	5471	phosphoribosyl pyrophosphate amidotransferase
PPP1R9A	55607	protein phosphatase 1 regulatory subunit 9A
PPP2CA	5515	protein phosphatase 2 catalytic subunit alpha
PPP2R5A	5525	protein phosphatase 2 regulatory subunit B'alpha
PRKCI	5584	protein kinase C iota
PRKCZ	5590	protein kinase C zeta
PSAP	5660	prosaposin
PSMF1	9491	proteasome inhibitor subunit 1
RACGAP1	29127	Rac GTPase activating protein 1
RARA	5914	retinoic acid receptor alpha
RELA	5970	RELA proto-oncogene, NF-kB subunit
RFFL	117584	ring finger and FYVE like domain containing E3 ubiquitin protein ligase
RGCC	28984	regulator of cell cycle
RGS1	5996	regulator of G protein signaling 1
RIPK1	8737	receptor interacting serine/threonine kinase 1
RNF152	220441	ring finger protein 152
RNF213	57674	ring finger protein 213
RUVBL1	8607	RuvB like AAA ATPase 1
SAMD9L	219285	sterile alpha motif domain containing 9 like
SAV1	60485	salvador family WW domain containing protein 1
SCAMP3	10067	secretory carrier membrane protein 3
SCEL	8796	sciellin
SEC24A	10802	SEC24 homolog A, COPII coat complex component
SEMA3A	10371	semaphorin 3A
SERPINB2	5055	serpin family B member 2
SERPINB3	6317	serpin family B member 3
SGO1	151648	shugoshin 1
SKA1	220134	spindle and kinetochore associated complex subunit 1
SKP2	6502	S-phase kinase associated protein 2
SLAMF7	57823	SLAM family member 7
SLC39A11	201266	solute carrier family 39 member 11
SLC40A1	30061	solute carrier family 40 member 1
SMAP2	64744	small ArfGAP2
SMPD1	6609	sphingomyelin phosphodiesterase 1
SMPD2	6610	sphingomyelin phosphodiesterase 2
SMURF2	64750	SMAD specific E3 ubiquitin protein ligase 2
SNX20	124460	sorting nexin 20
SOX9	6662	SRY-box transcription factor 9
SPARC	6678	secreted protein acidic and cysteine rich
SPINK7	84651	serine peptidase inhibitor, Kazal type 7 (putative)
SPRR1A	6698	small proline rich protein 1A
SPRR1B	6699	small proline rich protein 1B
SQSTM1	8878	sequestosome 1
SREBF1	6720	sterol regulatory element binding transcription factor 1
SRGN	5552	serglycin
STAP1	26228	signal transducing adaptor family member 1
STARD3	10948	StAR related lipid transfer domain containing 3

(continued)

GENE	EntrezID	Gene_Name
STAT1	6772	signal transducer and activator of transcription 1
STAT3	6774	signal transducer and activator of transcription 3
STMN1	3925	stathmin 1
SYBU	55638	syntabulin
SYS1	90196	SYS1 golgi trafficking protein
TAF12	6883	TATA-box binding protein associated factor 12
TAF9	6880	TATA-box binding protein associated factor 9
TAGLN	6876	transgelin
THY1	7070	Thy-1 cell surface antigen
TIMM17B	10245	translocase of inner mitochondrial membrane 17B
TIMP3	7078	TIMP metallopeptidase inhibitor 3
TJP2	9414	tight junction protein 2
TK1	7083	thymidine kinase 1
TLL1	7092	tolloid like 1
TMEM38B	55151	transmembrane protein 38B
TNFAIP3	7128	TNF alpha induced protein 3
TNFRSF1A	7132	TNF receptor superfamily member 1A
TNFSF10	8743	TNF superfamily member 10
TOP2A	7153	DNA topoisomerase II alpha
TP63	8626	tumor protein p63
TPX2	22974	TPX2 microtubule nucleation factor
TTK	7272	TTK protein kinase
TVP23C	201158	trans-golgi network vesicle protein 23 homolog C
TXN	7295	thioredoxin
TYROBP	7305	TYRO protein tyrosine kinase binding protein
UBN1	29855	ubinuclein 1
UCA1	652995	urothelial cancer associated 1
WWC1	23286	WW and C2 domain containing 1
XDH	7498	xanthine dehydrogenase
YAP1	10413	Yes associated protein 1
YDJC	150223	YdjC chitooligosaccharide deacetylase homolog
ZBED2	79413	zinc finger BED-type containing 2
ZNF204P	7754	zinc finger protein 204, pseudogene
ZNF385A	25946	zinc finger protein 385A

9.3 CSC gene signatures

- *53-CSC gene signature:*

  IRF6, IL1B, IRF6, MFAP5, PI3, PLD5, RNF152, SCEL, SERPINB3, SPINK7, SPRR1A, SPRR1B, TLL1, ZBED2, CCDC80, CLCA2, DENND2C, FAT2, FOLH1, GCNT4, LUM, NRG1, PAK2, PDGFD, PHLDA2, POF1B, PPP2R5A, TAF9, TAF12, TP63, XDH, ACTL6A, CASP8, CDKN2A, FADD, MAP3K7, MAP4K4, NFKB1, PML, PRKCZ, RELA, RFFL, RIPK1, RUVBL1, SERPINB2, SQSTM1, TNFAIP3, CDK4, EXOSC4, BMI1, MIEN1, NDUFB10, SKP2

- *242-CSC gene signature:*

  EPB41L4A, IL1B, IRF6, MFAP5, PI3, PLD5, RNF152, SCEL, SERPINB3, SPINK7, SPRR1A, SPRR1B, TLL1, ZBED2, ACSS2, ADAM12, AEBP1, AMOT, ANGPT1, BAG4, BCAR3, CAV1, CCDC80, CCL20, CDH2, CLCA2, COL3A1, CTSK, CXCL9, DENND2C, FAAH, FAT2, FOLH1, FPR1, GCNT4, GNE, HES1, IFI44L, IFNG, IGLL1, IRS2, KLF8, LCP1, LINGO2, LPL, LRFN3, LUM, MACC1, MADD, MMP9, NRG1, PAK2, PDE7B, PDGFD, PHLDA2, PLSCR1, POF1B, POMT2, PPP1R9A, PPP2R5A, PRKCI, RGCC, RGS1, SLAMF7, SLC39A11, SMPD1, SMPD2, SRGN, STAT1, SYBU, TAF9, TAF12, TAGLN, THY1, TMEM38B, TNFRSF1A, TNFSF10, TP63, TVP23C, TXN, UCA1, XDH, ZNF204P, ACTL6A, AZGP1, CASP8, CCBE1, CD74, CDKN2A, EHF, FADD, GBP2, HAS2, HLA-DMA, HLA-DPA1, HLA-DQA1, HLA-DRB4, HLA-DRB6, HSD17B2, LAMA4, LAPTM5, MAP3K7, MAP4K4, NFKB1, PLAAT4, PML, PRKCZ, RELA, RFFL, RIPK1, RUVBL1, SEMA3A, SERPINB2, SQSTM1, TNFAIP3, CDK4, RACGAP1, ALYREF, CD24, CDK1, EXOSC4, LY6E, MKI67, MLF2, MRPS23, STMN1, TOP2A, APOBEC3B, BMI1, CD44, MIEN1, NDUFB10, ASNS, DDR1, P4HA1, SKP2, DVL2, MMP1, CCNB2, CDC20, CENPA, KIF4A, TPX2, NCSTN, AMOTL1, AMOTL2, LATS2, MOB1B, SAV1, TJP2, WWC1, YAP1, PPP2CA, EXO1, SKA1, FOXM1, ADNP2, ADRM1, BAZ1B, C16orf58, CORO1B, DDX41, GUSB, ITPR3, MED1, MRPS17, NECAP1, PIM1, PPAT, PSAP, RNF213, SCAMP3, STARD3, TIMM17B, UBN1, COL1A1, COL6A3, HLA-DRB1, ADHFE1, CCNE2, AURKB, CDC25A, CDC45, CDCA3, CDCA8, CEP55, DLGAP5, GLI1, HJURP, KIF2C, KIF20A, KIF23, KIFC1, MELK, NCAPG, TTK, PLK1, BUB1, BUB1B, CDKN3, CKS2, TK1, DNMT1, HIST1H2BK, CD14, CSF1R, GIMAP4, LYZ, MS4A6A, CD52, CEP97, DOCK2, GGCT, LRBA, PSMF1, RARA, SAMD9L, SEC24A, SMAP2, STAP1, SYS1, YDJC, CTSD, EGR1, CD163, JUN, SREBF1, AP1S2, APBA3, HLA-DRB5, KIF2A, SLC40A1, SNX20, ZNF385A, STAT3, CAV2, CXCR4, KIF18B, NCAPH, NDC80, DCN, ESM1

- *269-CSC gene signature:*

  EPB41L4A, IL1B, IRF6, MFAP5, PI3, PLD5, RNF152, SCEL, SERPINB3, SPINK7, SPRR1A, SPRR1B, TLL1, ZBED2, ACSS2, ADAM12, AEBP1, ALOX5AP, AMOT, ANGPT1, BAG4, BCAR3, C1QC, CAV1, CCDC80, CCL20, CDH2, CLCA2, COL1A2, COL3A1, CTSK, CXCL9, DENND2C, FAAH, FAT2, FOLH1, FPR1, GCNT4, GNE, HES1, IFI44L, IFNG, IGLL1, IRS2, KLF8, LCN2, LCP1, LINGO2, LPL, LRFN3, LUM, MACC1, MADD, MMP9, NRG1, PAK2, PDE7B, PDGFD, PHLDA2, PLSCR1, POF1B, POMT2, PPP1R9A, PPP2R5A, PRKCI, RGCC, RGS1, SLAMF7, SLC39A11, SMPD1, SMPD2, SRGN, STAT1, SYBU, TAF9, TAF12, TAGLN, THY1, TMEM38B, TNFRSF1A, TNFSF10, TP63, TVP23C, TXN, UCA1, XDH, ZNF204P, ACTL6A, AZGP1, CASP8, CCBE1, CD74, CDKN2A, EHF, FADD, GBP2, HAS2, HLA-DMA, HLA-DMB, HLA-DPA1, HLA-DQA1, HLA-DRA, HLA-DRB4, HLA-DRB6, HSD17B2, LAMA4, LAPTM5, MAP3K7, MAP4K4, NFKB1, PLAAT4, PML, PRKCZ, RELA, RFFL, RIPK1, RUVBL1, SEMA3A, SERPINB2, SQSTM1, TNFAIP3, CDK4, RACGAP1, ALYREF, CD24, CDK1, EXOSC4, LY6E, MKI67, MLF2, MRPS23, STMN1, TOP2A, ACBD3, APOBEC3B, BMI1, CD44, MIEN1, NDUFB10, ASNS, DDR1, P4HA1, SKP2, DVL2, MMP1, CCNB2, CDC20, CENPA, KIF4A, TPX2,

NCSTN, AMOTL1, AMOTL2, LATS2, MOB1B, SAV1, TJP2, WWC1, YAP1, PPP2CA, BHLHE41, MMP2, TIMP3, EXO1, SKA1, FOXM1, ADNP2, ADRM1, BAZ1B, C16orf58, CORO1B, DDX41, GUSB, ITPR3, MED1, MRPS17, NECAP1, PIM1, PPAT, PSAP, RNF213, SCAMP3, STARD3, TIMM17B, UBN1, COL1A1, COL6A3, HLA-DRB1, ADHFE1, CCNE2, DMC1, EOMES, GLDC, HEYL, NES, AURKB, CDC25A, CDC45, CDCA3, CDCA8, CEP55, DLGAP5, GLI1, HJURP, KIF2C, KIF20A, KIF23, KIFC1, MELK, NCAPG, TTK, PLK1, BUB1, BUB1B, CDKN3, CKS2, TK1, DNMT1, HIST1H2BK, CD14, CSF1R, FCER1G, GIMAP4, LYZ, MS4A6A, SPARC, CD52, CEP97, DOCK2, GGCT, LRBA, PSMF1, RARA, SAMD9L, SEC24A, SMAP2, STAP1, SYS1, YDJC, CTSD, EGR1, CD163, JUN, SREBF1, AP1S2, APBA3, HLA-DRB5, KIF2A, SLC40A1, SNX20, ZNF385A, STAT3, CAV2, CXCR4, KIF18B, NCAPH, NDC80, SGO1, SOX9, TYROBP, DCN, ATF3, EGFR, ESM1, NR2F1, HAVCR2, ID2, LTBP1, SMURF2

- *18-CSC gene signature:*

CLCA2, IL1B, MFAP5, NRG1, PDGFD, PHLDA2, RUVBL1, SERPINB2, SERPINB3, SPRR1A, SPRR1B, SQSTM1, TAF9, TLL1, TP63, ZBED2, CDKN2A and SCEL

9.4 Supplementary figures

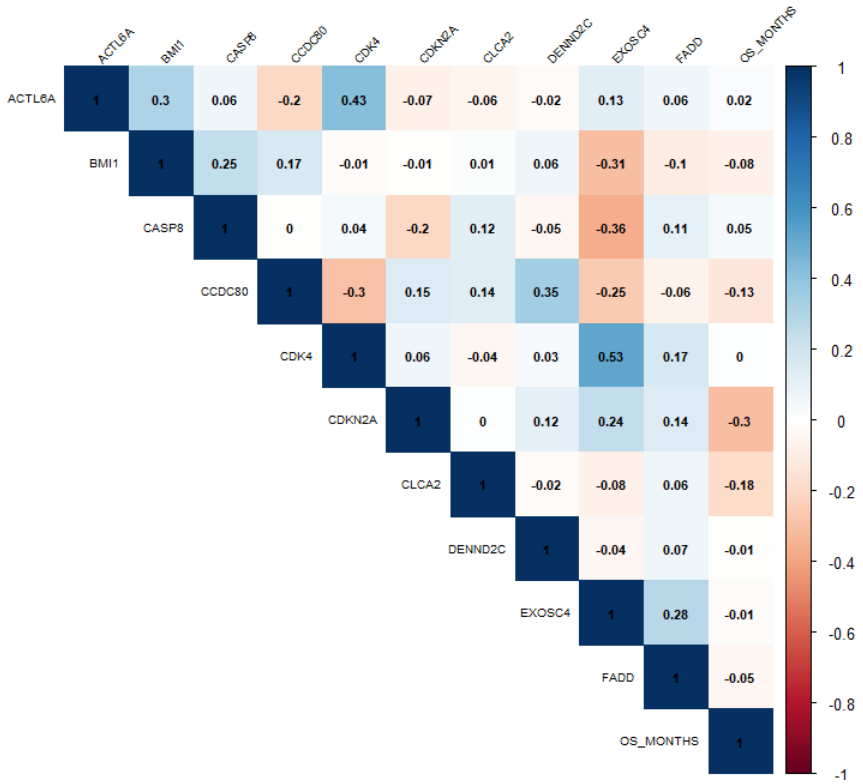


Figure S1. Correlation analysis of genes from the 53-CSC gene signature (subset 1). Correlation with OS (in months)

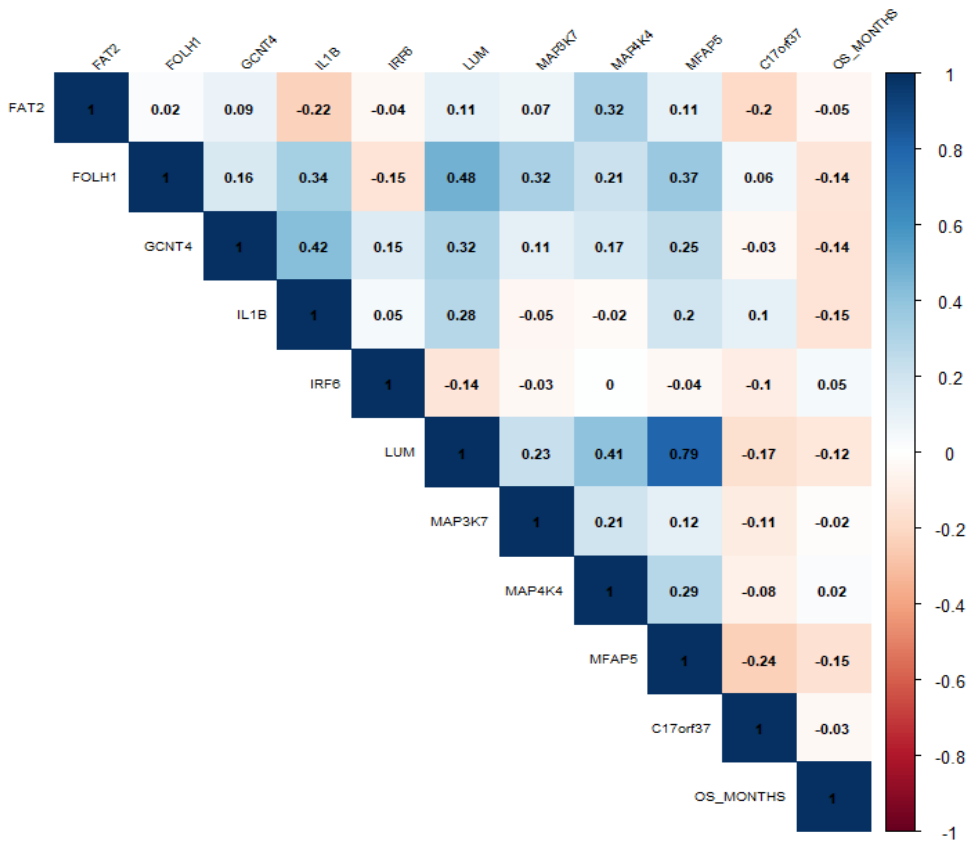


Figure S2. Correlation analysis of genes from the 53-CSC gene signature (subset 2). Correlation with OS, DSS and PFS (in months)

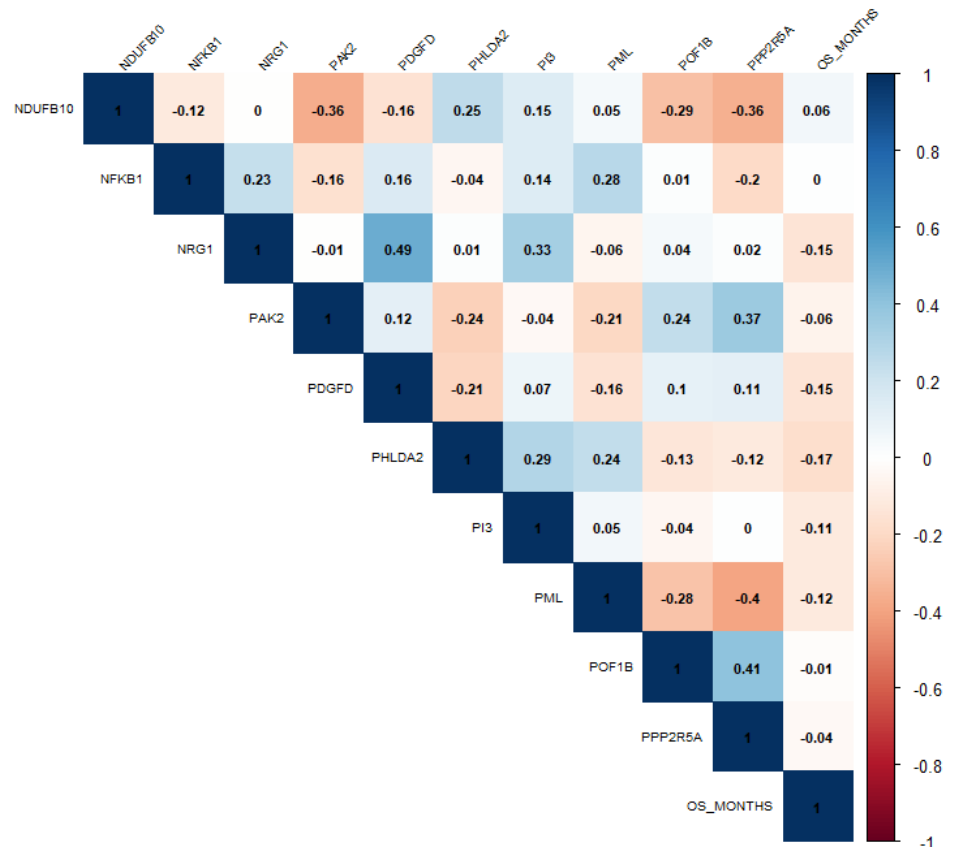


Figure S3. Correlation analysis of genes from the 53-CSC gene signature (subset 3). Correlation with OS, DSS and PFS (in months)

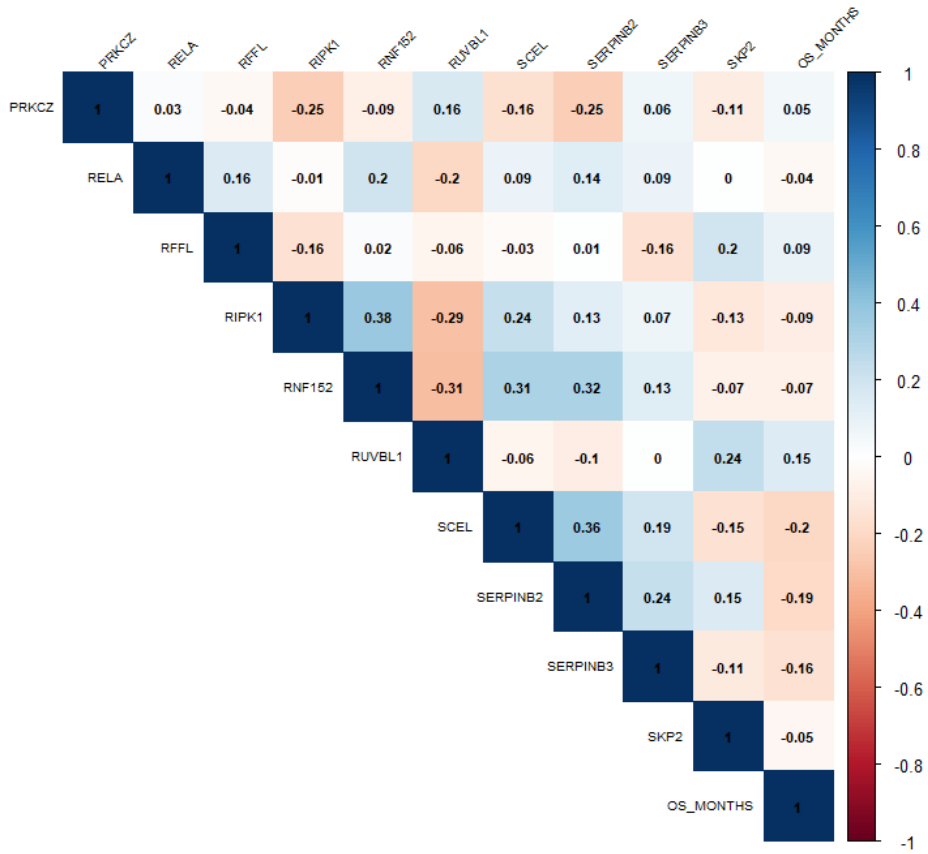


Figure S4. Correlation analysis of genes from the 53-CSC gene signature (subset 4). Correlation with OS, DSS and PFS (in months)

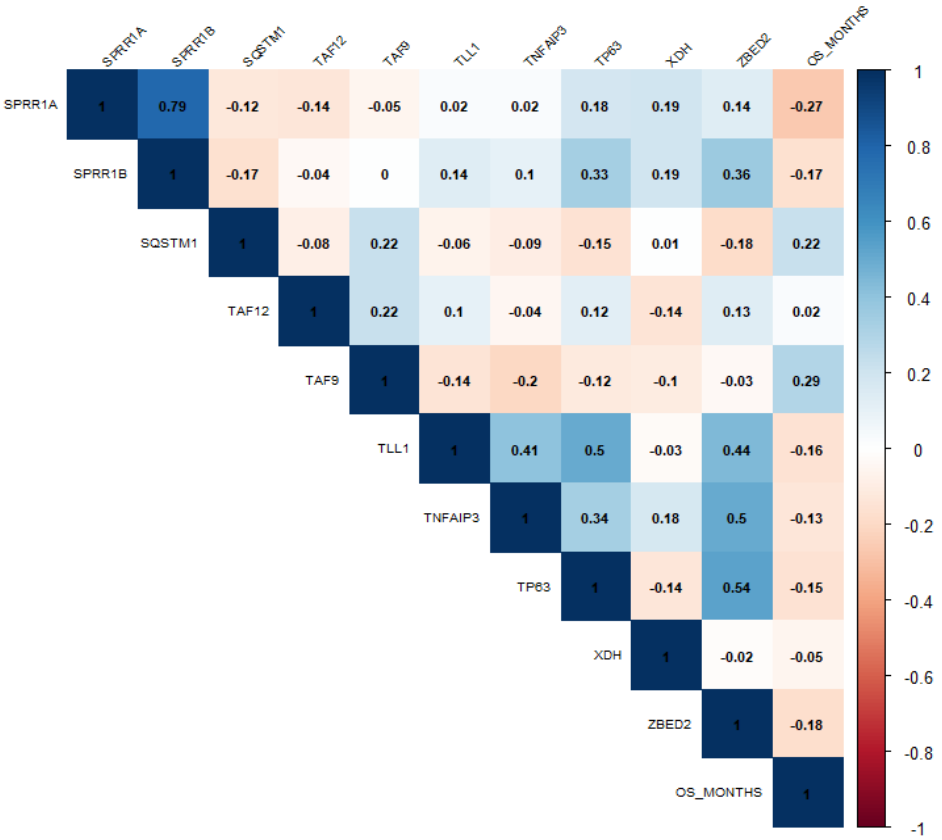


Figure S5. Correlation analysis of genes from the 53-CSC gene signature (subset 5). Correlation with OS, DSS and PFS (in months)

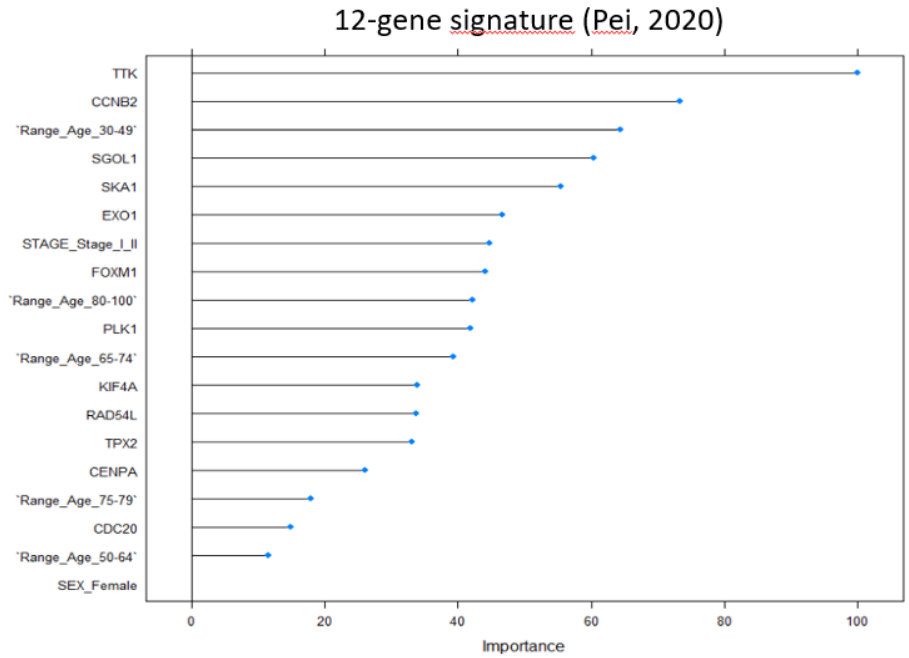


Figure S6. List of variables ranked by their relative importance in the model

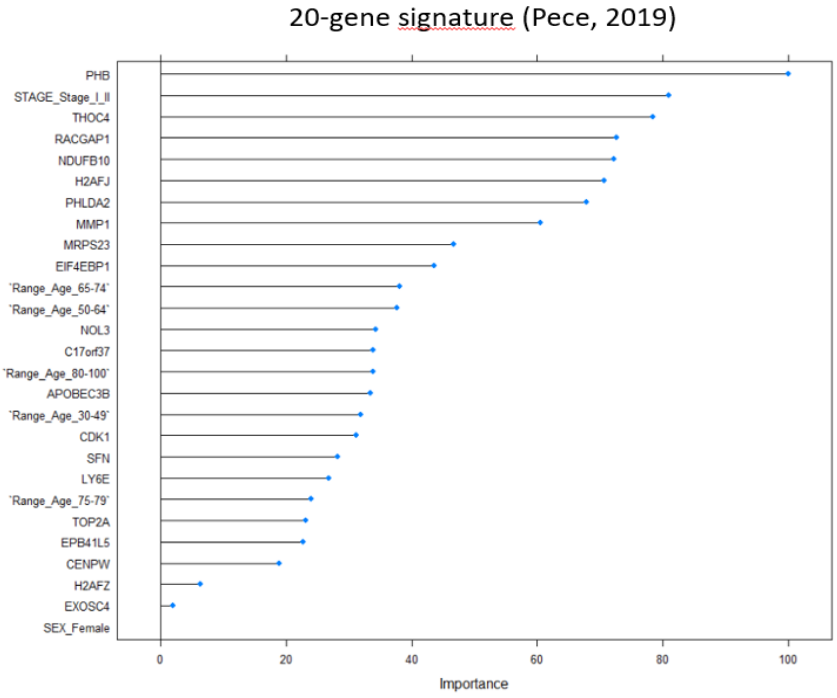


Figure S7. List of variables ranked by their relative importance in the model

9.5 Code

Code used in the project is available in the following github link:

https://github.com/jonortizabalia/Final-Master-s-thesis-UOC-2020.git