Màster Universitari
Anàlisi de Dades Òmiques /
Omics Data Analysis
FACULTAT DE CIÈNCIES I TECNOLOGIA
UVIC | UVIC·UCC

Master of Science in Omics Data Analysis

Master Thesis

Bioinformatics approach to identify genes whose tumour expression shows a dual association with patient outcome

By

Arnau Soler Costa

Supervisor: Dr. Miquel Àngel Pujana, Breast Cancer Group Leader, IDIBELL

Co-supervisor: Roderic Espín, Breast Cancer Bioinformatics PhD Student, IDIBELL

Academic tutor: Dr. Mireia Olivella, Structural Bioinformatics Researcher and Chair of Master in Omics Data Analysis, UVIC - UCC

Department of Bioscience

University of Vic – Central University of Catalonia

12/09/2022

Bioinformatics approach to identify genes whose tumour expression shows a dual association with patient outcome

Arnau Soler Costa

ABSTRACT

The expression and/or function of many human genes has been associated with cancer progression. Typically, these studies identify overexpression or underexpression of a given gene as associated with patient survival and/or disease progression. However, a substantial minority of genes have an impact on cancer progression that is non-monotonic; that is, their effect direction and magnitude change on time, disease status, applied therapy and/or on other disease factors. Systematic identification of this type of genes can provide new fundamental insights on cancer drivers. Here, we have developed an approach to identify this class of genes, allowing us to initially define their functions and molecular interactions. This appraoch is applicable to any cancer type and has the promise to reveal unappreciated cancer gene drivers.

Keywords: gene expression, cancer, biphasic mode of action, tumour suppressors, oncogene, disease stage

1 INTRODUCTION

Cancer is a disease in which cells in the body multiply uncontrollably and spread to other parts of the body. Under normal conditions, human cells form and multiply to form new cells as the body needs them. When cells age or become damaged, they die, and new cells replace them.

Sometimes the process does not follow this order and abnormal or damaged cells form and multiply when they should not. These cells may form tumours, which are lumps of tissue. Tumours are either cancerous (malignant) or non-cancerous (benign). Cancerous tumours spread (or invade) other parts of the body and form tumours, a process called metastasis. It is possible for cancer to start in any part of the human body, giving rise to many different types of cancer.

Cancer is a genetic disease. Changes in the genes that control how cells function, how they form and multiply, cause cancer. Genetic changes that contribute to cancer usually affect two main types of genes: proto-oncogene and tumour suppressor gene. These changes are sometimes called "onco-initiators" and could end up as “oncogenes” or “tumour suppressor genes” (TSG**) (National Cancer Institute, 2021)**. These two types of genes are called commonly named “drivers” **(Stratton, 2009)**.

Scientists had been measuring the gene expression of these drivers, using high throughput genomic technologies, to study its interaction with cancer.

Gene expression **(Zhong Wang, 2009)** studies are common in medical research as they offer information about the association of genes and the phenotype or variable of interest.

The expression and/or function of many human genes has been associated with cancer outcome. Typically, these studies identify over-expression or under-expression of a given gene as associated with risk and/or outcome and there are

also other studies that analyse gene expression over time.

The analysis of a particular change through time are normally done using time-to-event (TTE) data or to be more specific, survival data (SD). Survival data is unique because the outcome of interest is not only whether or not an event occurred, but also when that event occurred **(Columbia Public Health, 2022)**.

Analyse survival cancer data kept scientists thinking that gene expression does not change over time. This means that genes only can/could act as oncogenes or TSG, and this remains constant over time.

However, as cells, some genes have been associated to have a dormant or quiescent type of behaviour. These, are genes that could be in a dormant or quiescent state and then changing its state to "active", causing an effect such as a relapse.

This type of genes appears to show a dual mode of action, showing an effect in one direction that changes with time. For example, the transforming growth factor-beta (TGF-β), that is a pleiotropic cytokine with the capability to act as tumour suppressor or tumour promoter depending on the cellular context **(Lebrun, 2012)**. In other words, some gene products may act as tumour suppressors or oncogenes depending on disease stage or other variables. Which we called them Biphasic Genes. This type of complexity may arise from the interaction of the function of a given gene with treatment and/or tumour features. This is because gene expression is typically measured at a basal time point prior to therapy and this expression value is assessed for associations with outcome across time.

These fundamental genes, their functional roles, and impact on cancer biology remain largely unknown. Identifying these genes and understanding their interactions can provide fundamental insight to improve cancer risk and prognosis estimation.

1.2 Methods

Survival data are generally described and modelled in terms of two related probabilities, namely survival and hazard. The survival probability (which is also called the survivor function) $S(t)$ and is the probability that an individual survives from the time origin (e.g., death) to a specified future time t. It is fundamental to a survival analysis because survival probabilities for different values of t provide crucial summary information from survival data. These values directly describe the survival experience of a study cohort.

The hazard is usually denoted by $h(t)$ or $\lambda(t)$ and is the probability that an individual who is under observation at a time t has an event at that time. Put another way, it represents the instantaneous event rate for an individual who has already survived to time t **(T G Clark, Survival Analysis Part I, Basic concepts and first analyses, 2003)**.

Traditional methods of logistic and linear regression are not suited to be able to include both the event and time aspects as the outcome in the model, so some other methods are usually used to analyse survival data. The most popular ones are the visualization of Kaplan-Meier (KM) plots or the Log-rank (LR) test. But to describe the effect of categorical or quantitative variables, such as gene expression, on survival

data, the Cox Proportional-Hazards (CPH) regression model (Cox regression) is one of the most common methods used. The Cox model is essentially a regression model commonly used in medical research for investigating the association between the survival time of patients and one or more predictor variables **(Cox D. , 1972)**.

The purpose of the model is to evaluate simultaneously the effect of several factors on survival, defining the hazard level as a dependent variable which is being explained by the time-related component (so called baseline hazard) and covariates-related component **(Borucka, 2014)**.

The Cox model is expressed by the *hazard function* ($h(t)$ or $\lambda(t)$) and can be interpreted as the risk of dying at time **(T G Clark, Survival Analysis Part I: Basic concepts and first analyses, 2003)** *t*.

$$h(t) = h_0(t) \times \exp\left(b_1 x_1 + b_2 x_2 + \ldots + b_p x_p\right)$$

where,

$-\ t$ represents the survival time.

$-\ h(t)$ is the hazard function determined by a set of p covariates ($x_1 x_2$, ..., x_p).

- The coefficients (b_1, b_2, ..., b_p) measure the impact (e.g., the effect size) of covariates.

- The term h_0 is called the baseline hazard. It corresponds to the value of the hazard if all the x_i are equal to zero (the quantity $\exp(0)$ equals 1). The 't' in $h(t)$ reminds us that the hazard may vary over time.

The quantities $\exp(b_i)$ are called hazard ratios (HR). A value of b_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases.

Put another way, a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

In summary,

HR = 1: No effect

HR < 1: Reduction in the hazard

HR > 1: Increase in hazard

The Cox Proportional Hazards model makes several assumptions. Thus, it is important to assess whether a fitted Cox regression model adequately describes the data **(STHDA, 2022)**.

There are three types of diagnostics for the Cox model:

- Testing the Proportional Hazards (PH) assumption.
- Examining influential observations (or outliers).
- Detecting nonlinearity in relationship between the log hazard and the covariates.

In order to check these model assumptions, *Residuals* method are used. The common residuals for the Cox model include:

- *Schoenfeld residuals* to check the proportional hazards assumption
- *Martingale residual* to assess nonlinearity
- *Deviance residual* (symmetric transformation of the *Martinguale residuals*), to examine influential observations.

The key assumption of the Cox model is that the hazard curves for the groups of observations (or patients) should be proportional and cannot cross **(Figure 1)**.

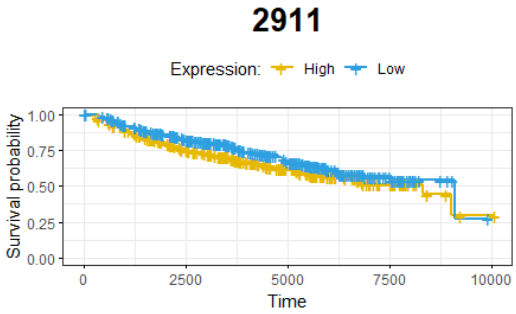

Figure 1. KM plot of 2911 gene. KM plot showing a non-crossing survival curves that has been always expected to be found in the gene expression data. This KM plot is from a gene with a Entrez Gene ID 2911 (GRM1) from the METABRIC dataset.

Consider two patients k and k' that differ in their x-values. The corresponding hazard function can be simply written as follow

Hazard function for the patient k:

$$h_k(t) = h_0(t)e^{\sum_{i=1}^{n}\beta x}$$

Hazard function for the patient k':

$$h_{k'}(t) = h_0(t)e^{\sum_{i=1}^{n}\beta x'}$$

The hazard ratio for these two patients is independent of time t:

$$\frac{h_h(t)}{h_{k'}} = \frac{h_0(t)e^{\sum_{i=1}^{n}\beta x}}{h_0(t)e^{\sum_{i=1}^{n}\beta x'}} = \frac{e^{\sum_{i=1}^{n}\beta x}}{e^{\sum_{i=1}^{n}\beta x'}}$$

Consequently, the Cox model is a proportional hazards model: the risk of the event in any group is a constant multiple of the risk in any other group. This assumption implies that, as mentioned above, the risk curves of the groups must be proportional and cannot cross **(Cox D. , 1972)**.

In other words, if an individual has a risk of death at some initial time point that is twice as great as that of another individual, then at all subsequent time points the risk of death remains twice as great.

The Cox model assumption has been applied in many cancer studies **(Cox D. R., 1972)**, studying the hazard ratio in quantifying the between-group difference in survival analysis and the PH **(Uno, 2014)** but in fact the proportionality over time is violated on multiple occasions **(Lin, 2022)**.

Since on the premise that some genes have been observed that seem to show a dual mode of action with patient outcome, showing an effect in one direction that changes over time, the aim of this study is to identify them with the help of the Cox model and KM curves to check whether there are genes that do not respect the proportional hazards assumption and therefore the hazard curves of the groups should be non-proportional and/or crossed.

There are a few ways to check this assumption, for example with the *Schoenfeld residuals* as mentioned before but, it may end up showing genes with non-proportional hazards but not necessary to have crossed hazards curves and it can end up being very tedious working with high dimensional data.

Therefore, in this project it has been implemented a bioinformatic approach to identify the genes that show this a dual mode of action with patient outcome studying the KM risk curves of the expression data of the genes and then study the biology behind.

2 GOALS

- By analysing human cancer gene expression data, identify genes whose expression reduce or increase risk of cancer relapse/death during a period of patient follow-up, but subsequently change their effect direction during subsequent follow-up.
- Integrating diverse data sources, to understand this novel class of cancer genes biologically and functionally.

3 MATERIALS

3.1 RNA-Sequencing datasets

In this study, we used two RNA-Sequencing datasets (RNA-Seq data) of human breast cancer already preprocessed and normalized. One containing the gene expression of the BRCA-TCGA project that was obtained from The Cancer Genome Atlas (TCGA) program, in the GDC Data portal **(National Cancer Institute, 2022)** and then selecting the gene signature Luminal A (LumA), which is a subtype of breast cancer categorized by the gene expression assay PAM50, that was used to categorize breast tumours into intrinsic subtypes characterized by the high expression of estrogen receptors (ERs) and progesterone receptors (PRs). These cancers tend to be of a lower grade and have a better prognosis than the other subtypes **(Le-PetrossM.D.R, 2012)**. The final dataset contains 233 samples **(Table 1, Supplementary Material)** and 15,748 genes with PFI times up to 8,000 days. The other dataset is from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) **(Bernard Pereira, 2016)** and was obtained from the cBioportal **(cBioPortal for Cancer Genomics, 2022)** that contained samples from human Breast Cancer patients and selecting the PAM50 gene signatures, also the LumA. The final dataset contains 679 samples **(Table 2, Supplementary Material)** and 18,492 genes with RFS times up to more less 10,000 days.

In both datasets, after LumA selection, only female patients associated with human breast cancer remained.

For each dataset there were a metadata file which contains phenotypic data and other covariates (including survival data), needed for the analyses.

In this study, the BRCA-TCGA and Breast Cancer METABRIC datasets and its metadata files were used to identify genes with dual association with patient outcome.

4 METHODS

All the methods throughout the study were performed in R software (ver. 4.1.2) **(R-project, 2022)** with *survival*, *survminer*, *Ensembl Biomart* and core *Bioconductor* packages.

4.1 Data processing

Before all the analyses we had to merge the RNA-Seq data with the phenotypic data, filter the samples and genes that are needed, looking for NAs between other things and finally, selecting the LumA samples and the necessary covariates. After this, it was performed a visual and descriptive analysis check of the data to see and understand how the data is.

At the end, we ended up with a dataset having all the samples in rows with 2 first columns containing the survival covariates such as Event (status: death or not) and Time (days) variables. Then, in columns, we also have all the genes.

4.2 Analysis

For this study and to identify the genes with a whose expression shows a dual association with patient outcome, we created a universal algorithm, creating a function called *BIPHASIC_GENES*. A final pipeline was also created to achieve the best results in an optimal way for all kind of dataset.

4.3 Biphasic Genes' function

To identify the Biphasic Genes, we first used the TCGA data to develop all the functions and after, we check the final algorithms with the METABRIC data.

We started looking for genes that do not respect the PH assumption with the function *cox.zph*, in the *survival* package identifying 1,574 genes of 15,748 in the TCGA data. Doing that we found genes that do not respect the PH assumption but looking at the KM plots these genes do not necessarily show this dual association with the cancer risk/outcome.

To overcome this, we had to go more deeply to the problem and after testing for the PH assumption, we thought to look for different HR at early time values and late time values between low and high gene expression. With this approach we could only find/identify 1,405 genes of 15,748 in the TCGA data but almost none had the KM curves shape as they should be.

The next move was to directly compute the values of the KM plot with the *survfit* and *Surv* functions. Studying the values of the differences in survival between the high and low gene expressions patients and looking where the sign of each difference change, from positive to negative and the other way round. We identified 1,142 genes of 15,748 in the TCGA data. This last approach let us identify some genes that its KM curves show the real shape of this dual association with patient outcome and leads us to reach the final algorithm.

The final function, named Biphasic Genes' function, is a variation of the last one. Adding the summatory of the survival differences between high and low gene expressions patients, finding an actual area between the gene expression survival curves on one side of the plot and another area but with the opposite sign in the other side of the plot. The computation of the areas is performed by the *density* and *approxfun* functions and then it is checked the sign of each area. At the end, we have implemented this final algorithm as a function adding as arguments the time values that you want to avoid, such as noise coordinates (at the start or end of the plot), and the time window you want the intersection to be. With this function we were able to find 861 Biphasic Genes of 15,748 for the TCGA dataset and 249 Biphasic Genes of 18,492 for the METABRIC dataset (see Results). Unless the function works well, it may also give some false positive genes.

As mentioned, to a better performance it is necessary to avoid the "time noise" that have the datasets in initial times and end times due to interference or miss of information.

Finally, having this last algorithm and implementing other functions also created by us, we produced a pipeline, represented in the **Figure 2,** that summarizes the full methodology from the initial data to a final list of Biphasic Genes.

4.4 Bioinformatics pipeline

4.4.1 Functions

To find the Biphasic Genes with the optimal way, a first step would be to know which time window fits your data the best to find the genes. To do this we created a function called *DIFF_SURVIVAL*, to compute the survival difference between low and high expressions of each gene to finally plot a heatmap and see the optimal time window.

The second (and optional) step and after the computation of the survival differences, we created another function, called *EXPAND_DATA*, to expand the data and have a better performance in the heatmap (in some cases this step could be an option).

The last step before the algorithm, is to plot a heatmap and look for the perfect time for the time window of the final function. It goes without saying that is a high time-consuming process.

The final step is to run the Biphasic Genes' function, that is the actual function created with the only purpose of identifying this class of genes. The algorithm consists in a function, called *BIPHASIC_GENES*, that with the initial data and the values for the time window founded in the heatmap, allows us to identify the genes with this dual association with the patient outcome.

After all this steps to find the mentioned Biphasic Genes, and to also identify the biology behind these genes, we performed a Gene Ontology (GO) enrichment analysis.

4.4.2 GO enrichment analysis

We used the *clusterProfiler* package to perform a GO **(Gene Ontology Consortium, 2022)** enrichment analysis for the biological processes (BP), the molecular functions (MF) and the cellular components (CC) of genes whose expression shows a dual association with patient outcome.

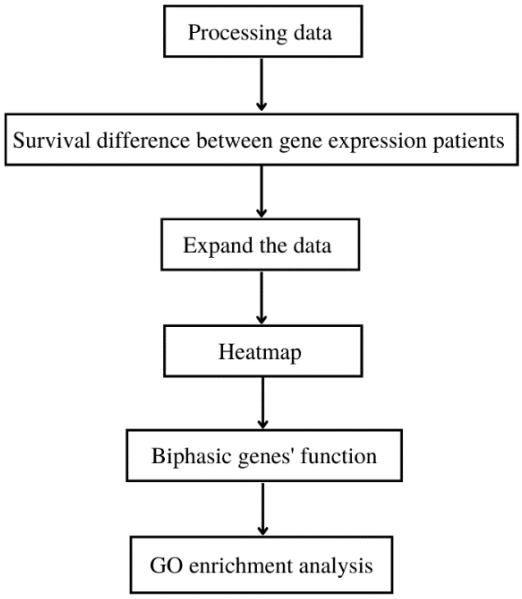

Figure 2. Proposed bioinformatics pipeline. Workflow of the full steps, using all functions and algorithms created, to identify genes whose expression shows a dual association with patient outcome, followed by a GO enrichment analysis.

All the proposed methods and scripts are in the Supplementary Material.

5 RESULTS

All the results showed are obtained using the pipeline described in the Methods part.

5.1 TCGA dataset

5.1.1 Identification of Biphasic Genes

To find the Biphasic Genes in the TCGA dataset it was first performed a prepossessing step to transform and adapt the data to the input needed in the functions. Then it is possible to run the *DIFF_SURVIVAL* function to compute the difference between survivals in each gene. With the survival differences matrix, we plotted the heatmap **(Figure 3)**. In this dataset it was not necessary to expand the data with the *EXPAND_DATA* function to see better the heatmap. We use a minimum time limit of 30 days to avoid interferences in initial times and a maximum time limit of more less 1,460 days because of the missing information of the dataset among this time.

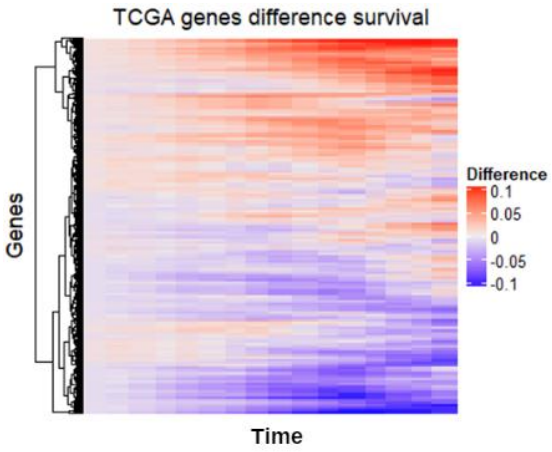

Figure 3. Heatmap of the survival differences between high and low expressions of each gene. Notice that the blue colours are for the negative differences and the red colours are for the positive ones. The x-axis is the time, and the y-axis are all the genes in the dataset.

Thanks to this heatmap we can extrapolated the time that more genes change its colour from red to blue or other way round. Meaning that the survival differences of the gene changed, therefore the KM curves should be crossed.

Looking at the heatmap we could see that around the middle of the x-axis we found the most genes that change its colour, so 730 days is the time value chose to search for intersections of the KM curves in the next step, the *BIPHASIC_GENES*.

Finally, with all the parameters set, the *BIPHASIC_GENES* function identified 861 genes whose expression shows a dual association with patient outcome from a total of 15,748 genes.

To prove if the identified genes are showing the mentioned association, we plotted some of its KM curves **(Figure 4)**.

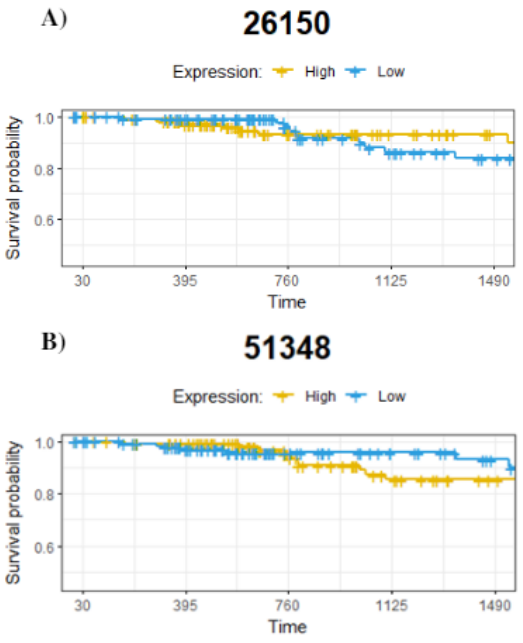

Figure 4. KM plots genes showing crossed survival curves of genes found with the *BIPHASIC_GENES* function. A) KM plot of gene 26150. B) KM plot of gene 51348. The blue curves are showing the survival probability of patients with low expression of the gene among time. The yellow curves are showing the survival probability of patients with high expression of the gene among time. The x-axis is the survival probability of the patients, and the y-axis is the time in days.

KM plot **Figure 4A** shows crossing survival curves of the gene 26150 (RIBC2), which is one of the genes

found with the *BIPHASIC_GENES* function on the TCGA dataset, with a low gene expression patients with more survival probability compared to high expression patients at early stages, crossing at more less 770 days, and then showing a less survival probability at late stages compared to the high gene expression patients curve. From the identified 861 Biphasic Genes, 509 genes have this pattern.

KM plot **Figure 4B** shows crossing survival curves of the gene 51348 (KLRF1), which is one of the genes found with the *BIPHASIC_GENES* function also on the TCGA dataset, with a high gene expression patients with more survival probability compared to the low gene expression patients at early stages, crossing around 760 days, and then showing a less survival probability at late stages compared to the low gene expression patients curve. From the identified 861 Biphasic Genes, 352 genes have this pattern.

5.1.2 GO enrichment analysis

The GO enrichment analysis of BP in the 509 genes, that have more survival probability in the low expression patients than the high expression patients in early stages, revealed enrichment in double-strand break repair via nonhomologous end joining biological process **(Figure 5, Supplementary Material)**. For the CC enrichments analysis of the same genes, revealed enrichment in peptidase, protease, and endopeptidase complexes **(Figure 6, Supplementary Material)**. In the MF enrichment analysis, no enrichment had been found.

In the genes that have more survival probability in high expression genes patients than low genes expression patients in early stages, we could not find any enrichment.

5.2 METABRIC dataset

5.2.1 Identification of Biphasic Genes

To find the Biphasic Genes in the METABRIC dataset it was also first performed a prepossessing step to transform and adapt the data to the input needed in the functions. Then we run the *DIFF_SURVIVAL* function to compute the difference between survivals in each gene. With the survival differences matrix, we plotted the heatmap **(Figure 7)**. In this dataset it was necessary to expand the data with the *EXPAND_DATA* function to see better the heatmap. We use a minimum time limit of 180 days to avoid interferences at initial times and a maximum time limit of more less 5,475 days because of the missing information of the dataset among this time.

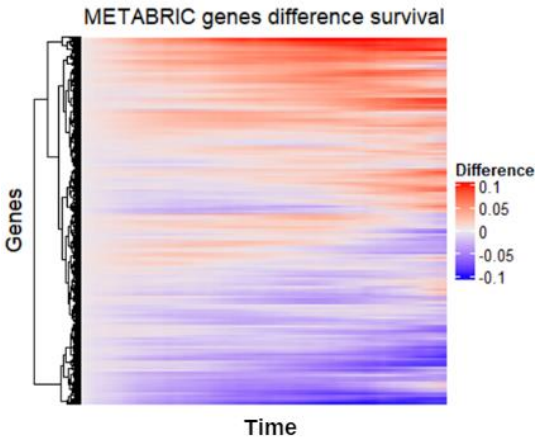

Figure 7. Heatmap of the survival differences between high and low expressions of each gene. Notice that the blue colours are for the negative differences and the red colours are for the positive ones. The x-axis is the time, and the y-axis are all the genes in the dataset.

Thanks to this heatmap we can extrapolated the time that more genes change its colour from red to blue or other way round. Meaning that the survival differences of the gene changed,

therefore the KM curves should be crossed.

Looking at the heatmap we could see that around the ¾ of the x-axis we found the most genes that change its colour, so 3,650 days is the time value chose to search for intersections of the KM curves in the next step, the *BIPHASIC_GENES* function.

Finally, with all the parameters set, the *BIPHASIC_GENES* function identified 249 genes whose expression shows a dual association with patient outcome from a total of 18,492 genes.

To prove if the identified genes are showing the mentioned association, we plotted some of its KM curves **(Figure 8)**.

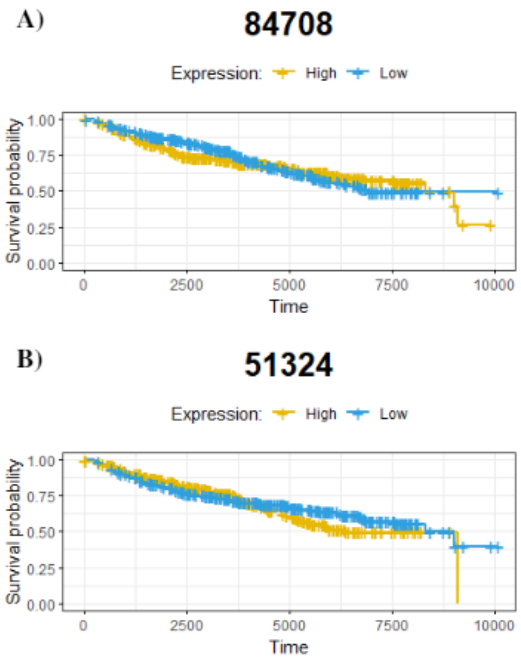

Figure 8. KM plots genes showing crossed survival curves of genes found with the *BIPHASIC_GENES* function. A) KM plot of gene 84708. B) KM plot of gene 51324. The blue curves are showing the survival probability of patients with low expression of the gene among time. The yellow curves are showing the survival probability of patients with high expression of the gene among time. The x-axis represents the survival probability of the patients, and the y-axis is the time in days.

KM plot **Figure 8A** shows crossing survival curves of the gene 84708 (LNX1), which is one of the genes found with the *BIPHASIC_GENES* function on the METABRIC dataset, with a low gene expression patients with more survival probability compared to high expression patients at early times, crossing at more less 4,200 days, and then showing a less survival probability at late times compared to the high gene expression patients curve. From the identified 249 Biphasic Genes, 120 genes follow this pattern.

KM plot **Figure 8B** shows crossing survival curves of the gene 51324 (SPG21), which is one of the genes found with the *BIPHASIC_GENES* function also on the TCGA dataset, with a high gene expression patients with more survival probability compared to the low gene expression patients at early times, crossing around 3,950 days, and then showing a less survival probability at late times compared to the low gene expression patients curve. From the identified 249 Biphasic Genes, 129 genes follow this pattern.

5.2.2 GO enrichment analysis

The GO enrichment analysis for BP of 120 genes that have more survival probability in the low expression patients than the high expression patients in early times, revealed enrichment in neurotransmitter transport, regulation of neurotransmitter levels, and regulation of neurotransmitter transport biological processes **(Figure 9, Supplementary Material)**. The MF enrichment analysis of the same genes revealed enrichment in metalloexopeptidase activity, exopeptidase activity, ADP binding, kinesin binding, carboxypeptidase activity, mitogen-activated protein kinase-kinase binding, and dipeptidase

activity **(Figure 10, Supplementary Material)**. In the CC enrichment analysis, no enrichment had been found.

In the 129 genes that have more survival probability in high expression genes patients than low genes expression patients in early times. The BP enrichment revealed immune response-regulating signaling pathway, response to molecule of bacterial origin, regulation of inflammatory response, cellular response to lipopolysaccharide, cellular response to molecule of bacterial origin, myeloid leukocyte activation, regulation of leukocyte mediated immunity, tumour necrosis factor-mediated signaling pathway, positive regulation of phospholipase activity, and regulation of B cell receptor signaling pathway **(Figure 11, Supplementary Material)**. The CC analysis revealed an enrichment in specific granule, nuclear exosome (RNase complex), and cytoplasmic exosome (RNase complex) **(Figure 12, Supplementary Material)**. The MF analysis revealed enrichment in hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amidines, a tumour necrosis factor-activated receptor activity, and cysteine-type endopeptidase activity involved in apoptotic signaling pathway **(Figure 13, Supplementary Material)**.

6 DISCUSSION

Cancer is a complex and multifaceted disease. Fundamentally, it is a disease of the genome, initiated by mutations in DNA that activate oncogenes and inactivate tumour suppressors, as well as dysregulation of the epigenome, which coordinates normal gene expression **(Alyna Katti, 2022)**. Understanding how genomic changes, cellular adaptations and changes to the microenvironment drive the initiation, progression and therapeutic response of individual cancers is crucial for developing more effective treatment options and improving outcomes for the millions diagnosed with cancer each year.

Being cancer a genetic disease, thousands of studies have been made studying the relationship between cancer and genes, analysing the gene expression of it. Normally, genes expression of many human genes has been associated with cancer outcome, typically, only in one direction. In other words, genes that show over-expression or under-expression associated with patient outcome or acting only like an oncogene or TSG. Moreover, some gene products may act as oncogenes or TSG depending on disease stage or other variables, what we call Biphasic Genes.

Here, we performed an extensive search of these class of genes using RNA-Seq breast cancer data from TCGA and METABRIC projects and selecting the LumA samples. To search for this class of genes, an algorithm was developed using RStudio and studying the KM risk curves of each gene.

For the TCGA data we found 861 Biphasic Genes from a total of 15,748 genes. From these genes, 509 have a KM plot showing the low gene expression patients with more survival probability compared to high expression patients at early stages, crossing at more less ~2 years, and then showing a less survival probability at late stages compared to the high gene expression patients' curve, such as **Figure 4A.** This means that at early stages the gene could be having an oncogene state and after ~2 years the gene is switched to act as TSG. Because

in early stages having a high gene expression is related to less survival of the patients and in late stages, having a high gene expression of it, is related to having more survival.

On the other hand, from the identified 861 Biphasic Genes of the TCGA project, 352 genes have a KM plot showing the other way round, high gene expression patients with more survival probability compared to high expression patients at early stages, crossing at more less 770 days, and then showing a higher survival probability at late stages compared to the low gene expression patients' curve, such as **Figure 4B**. This describes that at early stages the gene may act as a TSG and after ~2 years the gene is switched to act as an oncogene. Therefore, in early stages having a high gene expression is related to more survival of the patients and in late stages, having a high gene expression of it, is related to having less survival.

After performing the GO enrichment analysis, the genes acting as an oncogene and after ~2 years switch to act as a TSG, have been associated to peptidase, proteosome, and endopeptidase complexes, but it could not be linked with a relevant biological process that could define the biology behind them.

The same procedure was performed for the METABRIC data, where we found 249 Biphasic Genes from a total of 18,492 genes. From these genes, 120 have a KM plot showing the low gene expression patients with more survival probability compared to high expression patients at early stages, crossing at more less ~10 years, and then showing a less survival probability at late stages compared to the high gene expression patients' curve, such as **Figure 8A.** This means that at early stages the gene could be having an oncogene state and after ~10 years the gene is switched to act as TSG. Because in early stages having a high gene expression is related to less survival of the patients and in late stages, having a high gene expression of it, is related to having more survival.

On the other hand, from the identified 249 Biphasic Genes of the METABRIC project, 129 genes have a KM plot showing the other way round, high gene expression patients with more survival probability compared to high expression patients at early stages, crossing at more less ~10 years, and then showing a higher survival probability at late stages compared to the low gene expression patients' curve, such as **Figure 8B**. This describes that at early stages the gene may act as a TSG and after ~10 years the gene is switched to act as an oncogene. Therefore, in early stages having a high gene expression is related to more survival of the patients and in late stages, having a high gene expression of it, is related to having less survival.

The GO enrichment analysis of the 120 genes acting as an oncogene and after ~10 years switch to act as a TSG, have been associated to neurotransmitters and mainly peptide activities, such as exopeptidase activity, but it could not be linked with a relevant biological process that could define the biology behind them directly related to cancer.

For the 129 genes acting as a TSGs at early stages and switching to oncogenes at late stages, we could see a clear relation to the immune system and apoptosis signaling pathway.

As the immune system is closely related to cancer **(David Loose, 2009)**, and the inactivation of programmed cell death, or apoptosis, is central to the

development of cancer **(J Martin Brown, 2005)**, we could say that these genes could have some real impact on patient output and have this dual role association. Acting as TSG at early stages of cancer and as oncogene in late stages of cancer (such as in this case) or other way round.

Genes acting as TSG at early stages and as oncogene in late stages have been associated with cancer relapse **(Rueda, 2019)** and with the dormancy of cells **(Santos-de-Frutos, 2021)** which behaviour has been described as serious impediment to cancer treatment and therefore the prognostic of patients because of the mechanisms involved are poorly understood.

Therefore, with this work and algorithm, we could reach an accurate way to identify this class of genes, the Biphasic Genes.

6.1 Future directions

In order to improve the algorithm and the pipeline for a better identification and more accurate and universal way to find the Biphasic Genes, we studied the different distributions of the gene set to try to find a more easy and reliable way to identify them.

Using a set of already developed functions (*Mfuzz* package, **(Lokesh Kumar, 2007)**, we identified 4 different distributions in the METABRIC data set **(Figure 14, Supplementary Material)**, that could be identified in any data set.

So, the next step is to work in the development of a new algorithm, more sophisticated and accurate, to identify the Biphasic Genes using Machine Learning. This method allows the identification of this class of genes and all the different distributions that we could find. Therefore, we could find and study new genes, and its biology, related to cancer.

7 CONCLUSIONS

Biphasic Genes' function is able to identify genes whose expression shows a dual association with patient outcome. We could find 861 Biphasic Genes of 15,748 genes in the TCGA dataset and 249 Biphasic Genes of 18,492 for the METABRIC dataset. In both datasets the Biphasic Genes' function found genes acting as oncogene at early stages of cancer and as TSG at late stages of cancer and the other way round, so we have therefore fulfilled the initial objective of the project.

On the other hand, we could find some enrichment in the immune system of part of the Biphasic Genes, so we can link this class of genes to a crucial biological process in the field of diseases.

This work opens new insights to study biphasic genes and their interaction with cancer, offering a tool to find this class of genes and study the biology behind them. We are also looking to develop a new, more global, and accurate algorithm, with a different method, to improve its functionality, power and possibilities.

ACKNOWLEDGMENTS

This project would not have been possible without the support of many people. Many thanks to my supervisor, Dr. Miquel Àngel Pujana, for his guidance throughout this project. Very special thanks to my group mates, my co-supervisor Roderic Espín and Sandra Baiges, who offered all their advice and

support. Also, thanks to all the people of the ProCURE programme who have made my stay at IDIBELL very pleasant.

Thanks to the University of Vic and all its professors, in particularly my academic tutor, Dr. Mireia Olivella, for bring me all the needed knowledge to complete this project.

And finally, thanks to my parents and especially Andrea who endured this long process with me, always offering support and love.

REFERENCES

Alyna Katti, B. J. (2022). CRISPR in cancer biology and therapy. *Nature Reviews Cancer*, 259-279.

Bernard Pereira, S.-F. C. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications*, 7-11479.

Borucka, J. (2014). Extensions of Cox Model. *Ekonometria*.

cBioPortal for Cancer Genomics. (2022). *Breast Cancer* . Retrieved from (METABRIC, Nature 2012 & Nat Commun 2016): https://www.cbioportal.org/study/summary?id=brca_metabric

Columbia Public Health. (2022, Jun 7). *Population Health Methods*. Retrieved from Time-To-Event Data Analysis: https://www.publichealth.columbia.edu/research/population-health-methods/time-event-data-analysis

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 187-220.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 187,202.

David Loose, C. V. (2009). The immune system and cancer. *Cancer Biother Radiopharm*, 76-369.

Gene Ontology Consortium. (2022). *THE GENE ONTOLOGY RESOURCE*. Retrieved from http://geneontology.org/

J Martin Brown, L. D. (2005). The role of apoptosis in cancer development and treatment response. *Nature Reviews. Cancer*, 7-231.

Lebrun, J.-J. (2012). The Dual Role of TGFβ in Human Cancer: From Tumor Suppression to Cancer Metastasis. *ISRN Molecular Biology*.

Le-PetrossM.D.R, H. (2012). *Oncologic Imaging: A Multidisciplinary Approach.* Chapter 27 - Breast Cancer.

Lin, T. (2022). Incidence and impact of proportional hazards violations in phase 3 cancer clinical trials. *Journal of Clinical Oncology*.

Lokesh Kumar, M. E. (2007). Mfuzz: a software package for soft clustering of microarray data. *Bioinformation*, 5-7.

National Cancer Institute. (2021, May 5). *NIH*. Retrieved from What is cancer?: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

National Cancer Institute. (2022, May 31). *GDC Data Portal*. Retrieved from https://portal.gdc.cancer.gov/exploration?filters=%7B%22content%22%3A%5B%7B%22content%22%3A%7B%22field%22%3A%22cases.primary_site%22%2C%22value%22%3A%5B%22breast%22%5D%7D%2C%22op%22%3A%22in%22%7D%2C%7B%22content%22%3A%7B%22field%22%3A%22cases.diagnoses.ti

R-project. (2022). *The R Project for Statistical Computing*. Retrieved from https://www.r-project.org/

Rueda, O. M. (2019). Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature*, 399-404.

Santos-de-Frutos, K. (2021). When dormancy fuels tumour relapse. *Communications Biology*, 747.

STHDA. (2022). *Statistical tools for high-throughput data analysis*. Retrieved from Cox Model Assumptions:

http://www.sthda.com/english/wiki/cox-model-assumptions

Stratton, M. R. (2009). The cancer genome. *Nature*, 719–724.

T G Clark, M. J. (2003). Survival Analysis Part I, Basic concepts and first analyses. *British Journal of Cancer*, 232-238.

T G Clark, M. J. (2003). Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer*, 232-238.

Uno, H. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol*.

Zhong Wang, M. G. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review. Genetics*, 57-63.