INTERDEPARTMENTAL MASTERS PROGRAMME

INFORMATICS FOR LIFE SCIENCES



Identification of RNA Binding Proteins as candidate therapeutic targets for the treatment of cancer

Kamzolas Ioannis

Patras 2017

ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

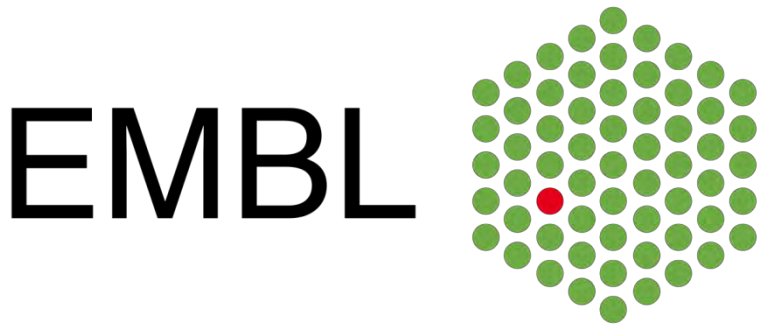
ΠΛΗΡΟΦΟΡΙΚΗ ΕΠΙΣΤΗΜΩΝ ΖΩΗΣ



Εντοπισμός πρωτεϊνών που δένουν RNA ως υποψήφιων θεραπευτικών στόχων για τον καρκίνο

Καμζόλας Ιωάννης

Πάτρα 2017



This Masters' Thesis is in collaboration with Chemical Biology Core Facility (EMBL, Heidelberg)

Principal Supervisor

Dr. Zoi Lygerou, Professor at the Medical Department of the University of Patras

Co-supervisor

Dr. Joe Lewis, Head of Chemical Biology Core Facility, EMBL Heidelberg

Examination Committee

Dr. Zoi Lygerou, Professor at the Medical Department of the University of Patras

Dr. Joe Lewis, Head of Chemical Biology Core Facility, EMBL Heidelberg

Dr. Stavros Taraviras, Professor at the Medical Department of the University of Patras

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Professor Zoi Lygerou. Being the director of studies of the Interdepartmental Postgraduate Program, she was next to me whenever I needed help and support. These two years I have learned a lot from her long experience in biology. It should also be mentioned that she is the one who initiated my research journey at EMBL, undoubtedly the best experience in my life so far.

Furthermore, I would like to thank Dr. Joe Lewis, Head of Chemical Biology Core Facility at the EMBL in Heidelberg. From the onset, his support was huge and it would not be an exaggeration to say that Dr. Lewis is my mentor. I will always appreciate him as a scientist, but also as a friend, and I will never forget his advice.

My sincere thanks also go to the member of my examination committee Professor Stavros Taraviras for his insightful comments and interest.

In addition, I want to express my gratitude to the European Molecular Biology Laboratory (EMBL) for this amazing journey into the world of research and science. I would like to thank all the members of the Chemical Biology Core Facility and especially Dr. George Reid and Dr. Iryna Charapitsa. Their support was utterly valuable, addressing to all my biological questions during this project.

Lastly, I am obliged to underline my honor having as my parents Panagiotis Kamzolas and Dimitra Panagiotakopoulos, who side by side with my partner, Anna Basta, indulged me with their unconditional love, understanding and everlasting support.

Kamzolas Ioannis,

September 2017

Abstract

RNA binding proteins (RBPs) are proteins that bind to the RNA in cells and form ribonucleoprotein complexes. It is estimated that approximately 7.5% of all protein-coding genes in the human genome code for RBPs and their aberrant expression has profound effects on cellular physiology and the pathogenesis of human disease, including cancer.

This thesis is part of a drug target discovery project. The aim of the project is to evaluate the role of RBPs in cancer prognosis and identify RBPs that could be potential drug targets in four types of cancer (lung, gastric, ovarian and breast cancer).

Our results reduce the number of RBPs that will be screened as potential novel drug targets. We start with an analysis of the effect of high level expression of known RBPs on cancer patient survival using publicly available RNA expression and patient survival data and continue with a comparison between the gene expression in normal and cancerous tissues.

More specifically, the main steps of this study are the following:

1. Identification of RBPs whose mis-expression correlates with poor survival in cancer patients, ranked by statistical significance
2. Prioritization of this hit list, integrating diverse data sources and comparing the difference in expression between normal and cancerous tissues
3. Selection of candidates to be taken forward into drug screening

Περίληψη

Οι RNA binding πρωτεΐνες (RBPs), είναι πρωτεΐνες που προσδένονται στο RNA των κυττάρων και συμμετέχουν στο σχηματισμό ριβονουκλεοπρωτεϊνικών συμπλόκων. Εκτιμάται ότι περίπου το 7.5% των γονιδίων στο ανθρώπινο γονιδίωμα κωδικοποιούν RBPs, ενώ η μη φυσιολογική έκφρασή τους έχει εμφανείς συνέπειες στην κυτταρική φυσιολογία και την παθογένεση πολλών ανθρωπίνων ασθενειών, συμπεριλαμβανομένου του καρκίνου.

Η παρούσα διπλωματική αποτελεί μέρος μιας προσπάθειας για την ανακάλυψη νέων φαρμακευτικών στόχων. Σκοπός είναι η αξιολόγηση του ρόλου που διαδραματίζουν οι RNA binding πρωτεΐνες στην πρόγνωση του καρκίνου και η αναζήτηση πιθανών RBPs που θα μπορούσαν να είναι φαρμακευτικοί στόχοι σε 4 τύπους καρκίνου (καρκίνο του πνεύμονα, του στομάχου, των ωοθηκών και του μαστού).

Αρχικά αναλύονται τα αποτελέσματα που έχει η υψηλή έκφραση όλων των γνωστών RNA binding πρωτεϊνών στην επιβίωση των καρκινοπαθών χρησιμοποιώντας δημοσίως διαθέσιμα σύνολα δεδομένων με τα επίπεδα έκφρασης του RNA σε διαφορετικούς τύπους καρκίνου και τα αντίστοιχα δεδομένα επιβίωσης. Στη συνέχεια συγκρίνεται η γονιδιακή έκφραση μεταξύ φυσιολογικών και καρκινικών ιστών, ενώ τελικό στόχο αποτελεί η μείωση του αριθμού των RBPs που θα ελεγχθούν πειραματικά.

Συνοπτικά, τα βασικά βήματα της ανάλυσης είναι τα εξής:

1. Εντοπισμός RNA binding πρωτεϊνών, των οποίων η ανώμαλη έκφραση σχετίζεται με το γρήγορο θάνατο των καρκινοπαθών (με βάση τα αποτελέσματα στατιστικής ανάλυσης)
2. Περαιτέρω φιλτράρισμα της επιτυχούς λίστας των RNA binding πρωτεϊνών χρησιμοποιώντας την πληροφορία δεδομένων όπως η διαφορά στην έκφραση μεταξύ φυσιολογικών και καρκινικών ιστών.
3. Επιλογή των κατάλληλων υποψηφίων πρωτεϊνών για drug screening.

Contents

Acknowledgements 7

Abstract 9

Περίληψη 11

Contents 13

Table of figures 15

Chapter 1: Introduction 17

- Cancer 17
- Drug Discovery and Development 18
- Gene Expression Identification Methods in Cancer 21
  - Microarrays 22
  - RNA-Sequencing 25
  - Additional Collected Information 27
- Publicly Available RNA Expression Datasets 28
  - Array Express and GEO 28
  - The Cancer Genome Atlas 30
- Survival Rates and Kaplan-Meier Plots 31
- KM-Plotter 36
- RNA-Binding Proteins 38

Chapter 2: Aim and Motivation 41

Chapter 3: Materials and Methods 43

- KM-Plotters for different Cancer Types 43
  - Lung cancer KM-Plotter 43
  - Gastric cancer KM-Plotter 48
  - Ovarian cancer KM-Plotter 51
  - Breast cancer KM-Plotter 53
- Data Collection 56
- Python Scripts 60
  - Unique genes in a list of genes 60

Mutual genes in two lists 62
Compare genes between 2 lists according to their effect in each cancer type 65
Different trends of the same gene 68
Collected data statistics 71
Chapter 4: Results 73
KM-Plotter for data collection 73
The distribution of the collected gene instances in each cancer and effect type 76
Filtering according to the effect on cancer patients' outcome 78
Filtering according to the different expression in normal and cancerous tissues 80
Exclusion of genes with instances that follow extremely different trends 83
Ranking of the final 22 RBPs 87
Chapter 5: Discussion 93
General 93
Evaluation of the results and emerging problems 95
Further research 97
Appendix 99
Script producing data statistics 99
Script for RBP final ranking according to the results of String Database 104
Chapter 6: Bibliography 107

Table of figures

Figure 1: A lung cancer cell is dividing to give two new cancer cells 17
Figure 2: A drug and a drug target. 18
Figure 3: State of the art robots participate in drug screening 19
Figure 4: A simplistic approach, with proteins depicted as large globular structures that have deep grooves 20
Figure 5: The central dogma of molecular biology 21
Figure 6: Probe intensities in a two color spotted microarray 23
Figure 7: Differences in expression between normal and cancerous cells (two colors spotted microarray) 23
Figure 8: Affymetrix DNA Microarray Platform 24
Figure 9: The steps of the RNA-Sequencing method 26
Figure 10: The web interfaces of ArrayExpress and GEO 29
Figure 11: The web interface of “The Cancer Genome Atlas” 31
Figure 12: A Kaplan-Meier plot derived from the Table 3 35
Figure 13: KM-Plotter’s homepage 36
Figure 14: KM-Plotter results for FXR1 gene 37
Figure 15: The structure of two RBPs; ADAR and Pumilio 39
Figure 16: Outline of thesis aim 42
Figure 17: The Lung KM-plotter environment with all chosen settings 44
Figure 18: Multiple probe sets available for a specific gene name 45
Figure 19: The Gastric KM-plotter environment with all chosen settings 49
Figure 20: The Ovarian KM-plotter environment with all chosen settings 52
Figure 21: The Breast KM-plotter environment with all chosen settings 54
Figure 22: The results of KM-Plotter for three different genes 57
Figure 23: The format of the excel file with all collected data. 59
Figure 24: The txt file with the genes which is given as an input to Script 1 61
Figure 25: The txt file with the unique genes derived from a set of genes 62
Figure 26: Two txt files with the compared lists 64
Figure 27: Information about the mutual genes between two gene lists 64
Figure 28: Part of the list with the lung genes 67
Figure 29: The mutual genes between lung and gastric lists 67
Figure 30: Part of the excel file with the collected data 70
Figure 31: Gene classification according to the trends followed by their gene instances 70
Figure 32: Collected data statistics 72
Figure 33: KM-Plotter results for LIN28B gene in lung cancer environment 74
Figure 34: Part of the excel file with all collected data 75
Figure 35: The distribution of the collected instances in each cancer and effect type 76
Figure 36: The distribution of the lung, gastric, ovarian and breast instances, among the three effect types 77

Figure 37: The distribution of the exponents of the p-values for each cancer's type data 78
Figure 38: The distribution of the 590 instances among the classes of negative, positive and zero 79
Figure 39: The distribution of the genes among lung and gastric 80
Figure 40: The excel file with the collected data is extended by one more column 81
Figure 41: The distribution for the exponents of the "expression p-values" 81
Figure 42: The different trends for FUS gene 84
Figure 43: The different trends for MSI2 gene 85
Figure 44: The different trends for HNRNPC gene 85
Figure 45: The different trends for LSM5 gene 85
Figure 46: The different trends for DDX18 gene 86
Figure 47: The different trends for PRKDC gene 86
Figure 48: The different trends for EIF4G1 gene 86
Figure 49: The different trends for MRPS12 gene 87
Figure 50: String database interaction sources 89
Figure 51: Protein-protein interactions according to the String database 89
Figure 52: Part of the csv file returned by the String Database with the interactions and scores between all couples of proteins 90
Figure 53: Outline of computational analysis steps and results 95

Chapter 1: Introduction

Cancer

Cancer is the most frequent cause of death in the world, alongside with cardiovascular diseases (Perez-Herrero et Fernandez-Medarde, 2015). It describes a collection of related diseases in which abnormal cells divide without control and form tumors. All tumors consist of damaged cells that normally would be destroyed by the human body in the programmed cell death (apoptosis) (Elmore, 2007). A tumor can fall into two types; malignant or benign. Almost all benign tumors are harmless as they are unable to invade nearby tissues. The cells that form a malignant cancerous tumor can spread to nearby tissues or travel far distances to other body parts (through the blood and lymph system) and form new tumors (metastasis) (Perez-Herrero et Fernandez-Medarde, 2015).

Main types of cancer are carcinoma, sarcoma, leukemia, lymphoma and multiple myeloma. For males, the most common cancer types are lung, prostate and stomach cancer, while for females are breast, colorectal, lung and cervical cancer (Torre et al., 2016). Every year more than 14.1 million new cases are diagnosed worldwide. Specifically in 2015 cancer caused about 8.8 million deaths corresponding to 15.7% of human deaths (Wang et al., 2016).

Cancer is induced by damage (or mutations) in the DNA because of environmental or inherited reasons. In non-metastatic cancers surgery could be effective for the patient, while in metastatic cancers, drugs (like chemotherapy) inhibit the uncontrolled division of cancer cells. However drugs also affect normal cells causing inevitable side effects which indicate the need to find new effective treatments (Perez-Herrero et Fernandez-Medarde, 2015).

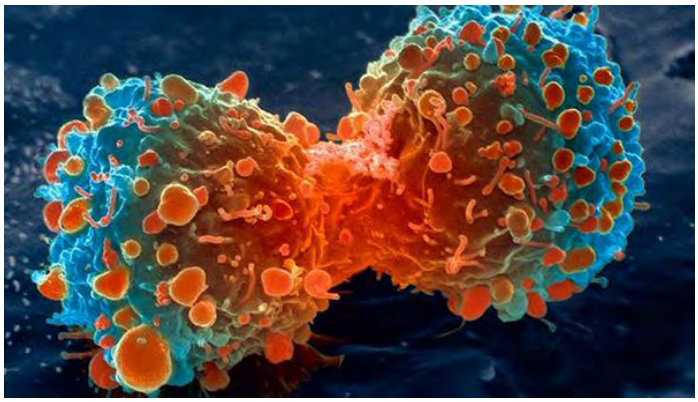


Figure 1: **A lung cancer cell is dividing to give two new cancer cells** (photo from National Institutes of Health).

Drug Discovery and Development

Drug development starts with the discovery of the most suitable targets for treating or preventing the disease (**target discovery**) (Moffat et al., 2014). Targets are usually proteins in the patient's body which are associated with a disease or proteins in microorganisms, causing a disease. The challenge is to identify which proteins are relevant and more importantly to confirm their role in a disease (Eder et Herling, 2015).



Figure 2: **A drug and a drug target.** Most of the times, a drug target is a protein. Proteins are large biomolecules made up of aminoacids. Long strings of aminoacids fold in a particular way, determining the shape of the protein (photo from chimera software tool).

Over the last years, pharmaceutical companies focus on understanding cellural networks of proteins or pathways (Moffat et al., 2014). A single protein may transmit messages to several other proteins, often in multiple pathways affecting their function. Knowing how these pathways work and interact, helps to identify the most appropriate target for a developing drug. The pathway approach allows a better understanding of the mechanisms that induce a disease. This knowledge and the will to answer important medical questions, outlines the priorities in target discovery (Eder et Herling, 2015).

Drug discovery is the process in which a lead compound is identified and a new candidate medication is discovered (Moffat et al., 2014). One of the methods used for the identification of potential drugs is called **screening**. In screening, large chemical libraries are tested for a particular biological activity. The most recent version of screening is **high-throuput screening** in which a researcher has the ability to conduct quickly, millions of chemical or pharmacological tests using robots, sensitive detectors

and data processing software (Figure 3) (Everett, 2015). Active compounds are identified for the next step of drug design.

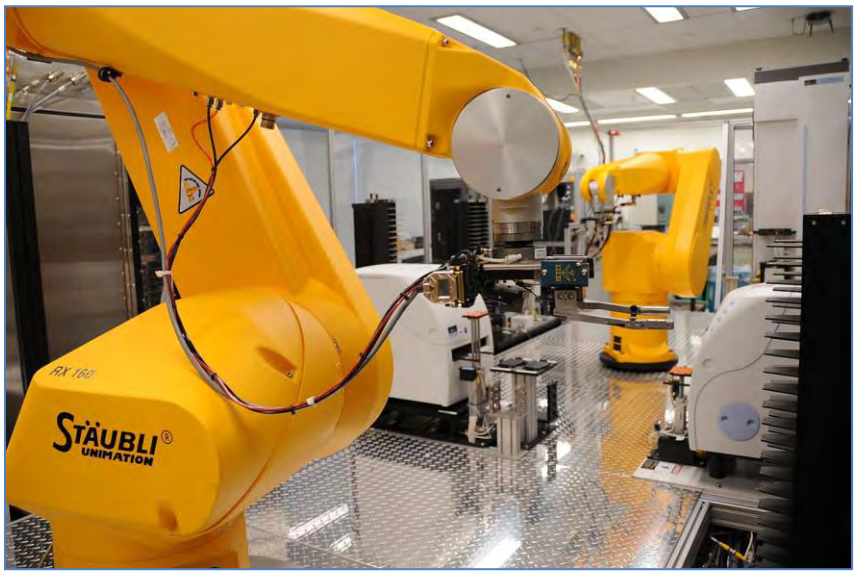


Figure 3: **State of the art robots participate in drug screening.** A robot takes assay plates from incubators and transfers them to another robot (shown in the background) or puts them at compound transfer stations (photo from National Human Genome Research Institute).

The point of screening is to find chemical compounds that bind to the identified target. If a compound modulates the target in a way that is expected to alter the disease, there is a **hit** (Everett, 2015). Subsequently, the compound will be refined to improve its safety and effectiveness and eventually will become a drug candidate (Figure 4).

Discovering and introducing a new drug may take many years of research and clinical development efforts (Ciocola et al., 2014). The process to bring a new pharmaceutical drug to the market is called **drug development** and can be divided into pre-clinical and clinical development (Moffat et al., 2014). In former kind, research starts before clinical trials and the main point is to assess biological activity and safety profile of the candidate compound (Ciocola et al., 2014). In the late pre-clinical stage further experiments are conducted on the drug candidate to ensure its safety for patients and ensure the required pharmacokinetic properties (such as appropriate absorption and metabolism by the human body). These experiments are executed with extraordinary diligence to minimize any risks to human test subjects.

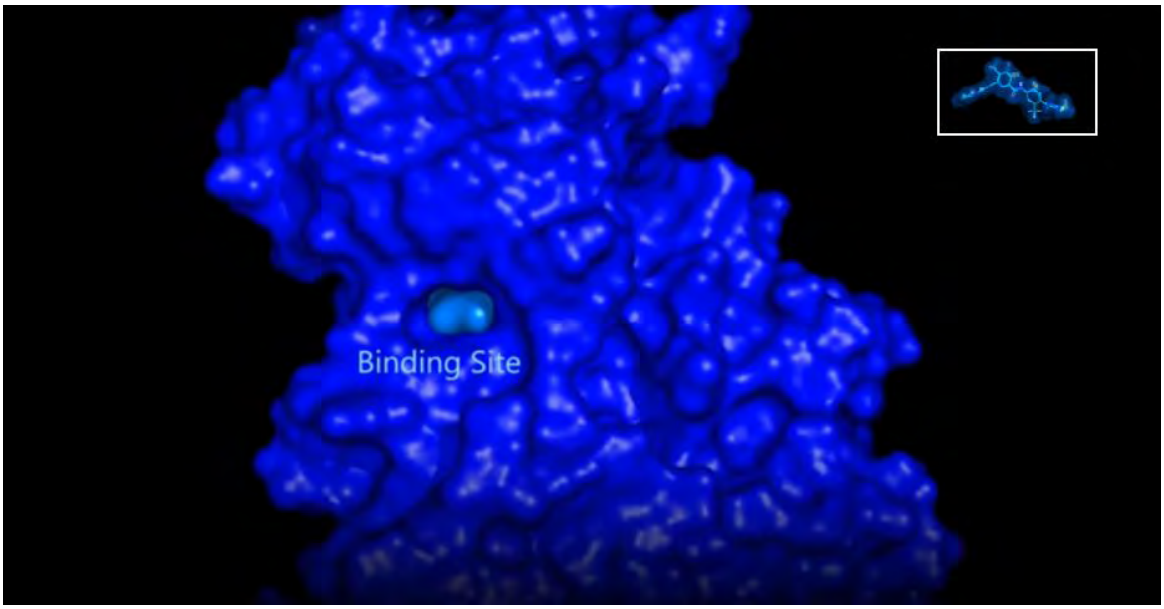


Figure 4: **A simplistic approach, with proteins depicted as large globular structures that have deep grooves.** This protein has a buried binding site which is good for "druggability". A small drug molecule (shown in the frame at the right) can be found or designed to fill this binding site. A drug is small compared to the protein and the better the fit of the ligand to the protein binding site, the more potent the drug can be. This can be simulated by a "key in a lock" image to explain ligand-protein binding (photo from chimera software tool).

Pre-clinical stage is followed by **clinical development**, in which observations and experiments (called clinical trials) take place. Clinical trials consist of 3 (or sometimes 4) phases in total, focused on evaluating drug safety and effectiveness or efficacy (Ciociola et al., 2014).

- **Phase 1 trials:**
  Safety and dosing are determined in healthy volunteers.

- **Phase 2 trials:**
  Further safety and efficacy are determined in small numbers of patients (about 100 – 250 patients) who have the disease. During this phase the optimal dose is evaluated and the possible side effects are checked (they could be different in patients compared to the healthy volunteers participated in phase 1).

- **Phase 3 trials:**
  Safety and efficacy are determined in large number of patients with the specific disease (1000 - 3000 patients). If safety and efficacy are substantiated in an adequate extent, the possibility for ceasing the clinical trials emerges at this point.

- **Phase 4 trials:**
  Sometimes, FDA (Food and Drug Administration) forces for one more phase. Hence, in this phase additional data are gathered, that were not collected in phase 3. They include efficacy, safety, risk information and pharmaco-economic data.

During clinical trials most compounds fail because of their levels of toxicity or because they don't have the required effect on the disease as was indicated in clinical trials (Moffat et al., 2014).

Gene Expression Identification Methods in Cancer

According to the central dogma of biology, the instructions to make mRNAs derive from genes (Crick, 1958). Gene expression describes the process of using gene information for the synthesis of gene products. More often the final gene products are proteins but they can also be functional RNAs.

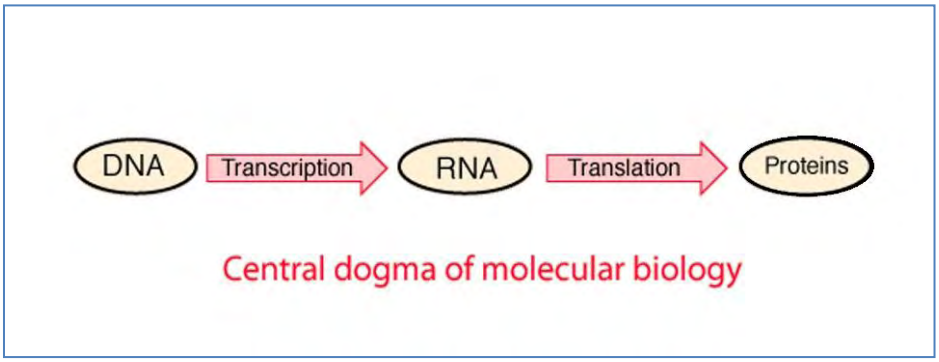


Figure 5: **The central dogma of molecular biology.** It describes the flow of the genetic information within a biological system and was first described by Francis Crick in 1958. The DNA is transcribed into mRNA and mRNA is translated into proteins (Crick, 1958).

In any given moment, different genes are activated or deactivated, meaning that only a fraction of their total number participates to make mRNA. Whether a gene will be activated is influenced by its local environment or by different experimental conditions. For instance, a cancer cell expresses different genes than a normal cell and different levels of an mRNA sequence means an altered need for the respective protein (Kim et

al., 2010). Gene expression identification refers to the expression measurement of thousands of genes simultaneously and answers important questions about the functions of the cell (Kim et al., 2010).

Several methods can be used to identify gene expression. However, there is not a best method for every occasion (Kim et al., 2010). Factors such as the available genome information for the species of interest, the expertise in analyzing the data and the available amount of money, determine the most suitable method each time. Different methods should be used for the discovery of novel genes or for transcripts likely to be expressed at very low or very high levels.

In the current chapter are explained two methods (Zhang et al., 2015): **DNA microarrays** and **RNA-sequencing**. Both methods demonstrate a high run to run reproducibility with their gene expression profiles being highly correlated. However, noise and biases are prominent when measuring very low expression values.

Microarrays

In **DNA Microarrays method**, predefined DNA fragments (probes) are spotted to a solid surface (made of glass or plastic) and form a microarray gene chip (Rogojina et al., 2010). DNA deriving from the sample of interest (cells or tissues) is labeled with fluorophores and hybridized to the DNA probes on the microarray. The intensity of the fluorescent at each spot, determines the analyzed expression pattern of the gene which corresponds to this probe (Wang et Xi, 2013).

Several microarray platforms are available for gene expression profiling. Among their major differences is the number of target groups hybridized to each array, the size of the probes and the technology used to spot them on the microarray chip, the used method to extract RNA from the samples of interest and the labeling with fluorophores of the targeted DNA (Rogojina et al., 2010). Main types of DNA microarrays are the **two colors spotted microarrays** and the **Affymetrix microarrays** (Hardiman, 2004).

In the two-color-spotted microarrays, probes are cDNA fragments or oligonucleotides (Rogojina et al., 2010). Two target groups are labeled with two different fluorescent dyes and when the targets are hybridized to the probes, a laser (which recognizes the two fluorescent dyes) scans the strength at each spot (Woo et al., 2004). This type of microarray constitutes a quite affordable method with the probes being created beforehand and then spotted on the microarray chip (Rogojina et al., 2010). The length of fragments usually varies between 500 and 1000 nucleotides, while the length of oligonucleotides between 20 and 100 nucleotides.

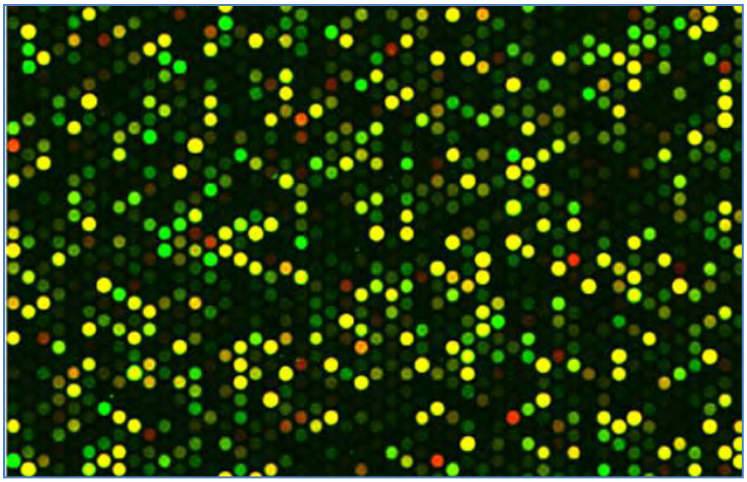


Figure 6: **Probe intensities in a two color spotted microarray.** In total the chip has 44000 probes and each probe corresponds to a specific mRNA (image from European Brain Research Institute).

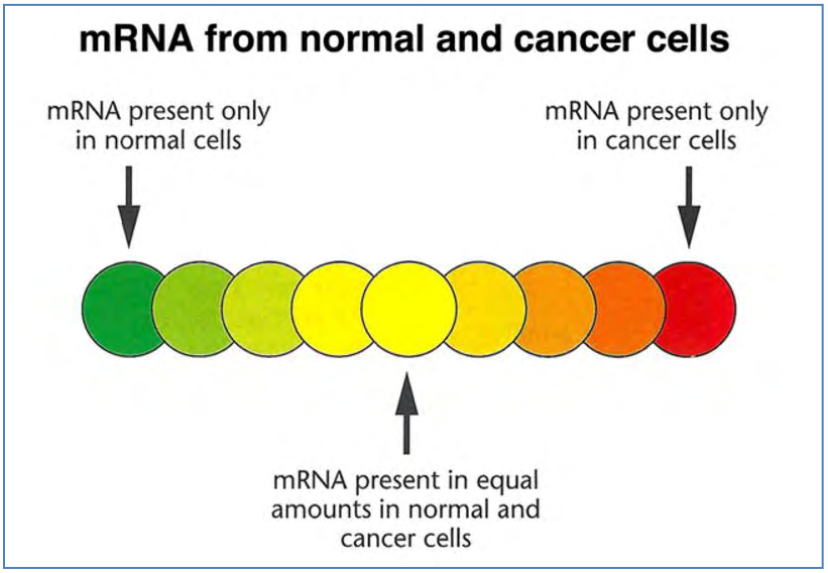


Figure 7: **Differences in expression between normal and cancerous cells (two colors spotted microarray).** There is a full range of colors between red and green which illustrates the relative differences in expression. The green color indicates that mRNA is present only in normal cells while the red color that mRNA is present only in tumourous cells. All the intermediate colors mean that the mRNA is expressed more in normal or cancerous cells respectively.

Affymetrix belong to the one-color microarrays, with only one target group hybridized to each array (Woo et al., 2004). Their absolute expression values can be compared to the results of another sample which uses the same probes and the experimental reproducibility is assured. Each transcript is represented by multiple probes (usually between 11 and 20) (Rogojina et al., 2010). An Affymetrix probe is made of a pair of oligonucleotides; a PM (Perfect Match) and a MM (Mismatch). The oligonucleotides typically consist of 25 bases, with both PM and MM being identical except for a nucleotide in the middle of the sequence. Mismatch oligonucleotides are negative controls, detecting noise and hybridization mistakes (Woo et al., 2004).



Figure 8: **Affymetrix DNA Microarray Platform**. The specific product line is called GeneChip (photo from www.affymetrix.com).

In Affymetrix microarrays, compared to the two-color-spotted microarrays, less RNA for each target is used. In the spotted microarrays two different fluorescents (Cy5 and Cy3) are added before hybridization and the probes suffer from biases, as the Cy5 molecule is heavier and its attachment to the target is less facilitated. In Affymetrix microarrays only one fluorescent dye is used, which is attached to the target after hybridization (Rogojina et al., 2010).

A possible shortcoming in Affymetrix microarrays is that a gene may correspond to multiple probes. These probes may differ significantly, presenting even contradictory results. Developed score methods evaluate the probe sets and Affymetrix microarray users can use all available probe sets or only the best one according to the score

method. In the last case, there could be a one-to-one match for genes and probe sets (Li et al., 2011).

In microarray technologies, an emerging problem is the prominent mistakes in measurements (Draghici, 2006). These methods of gene expression profiling suffer in measurement precision, albeit powerful, reliable and highly used. Specifically in Affymetrix microarrays due to mistakes in manufacturing procedures, many spotted probes are not the intended ones and they do not correspond to the appropriate transcripts (Mecham 2004; Knight, 2001; Forman-Kay, 1999). A gene can have many variants and the accuracy of the results could be ensured only if the identities of the genes and variants are well known (Lee et Roy, 2004; Boue et al., 2003). Microarrays are based on hybridization of already existing probes and are incapable to detect novel genes or transcripts. The hybridization limits their sensitivity, rendering impossible the detection of very similar sequences like isoforms (Wang et Xi, 2013). However, they still remain an inexpensive gene expression identification method (Zhang et al., 2015).

RNA-Sequencing

RNA-Sequencing reveals the presence and quantity of mRNA at a specific moment (Wei et al., 2014). It can detect gene expression differentiations between different groups or different expressions for the same group over time. Compared to microarrays, RNA-Sequencing is preferred because the cost is lower and there is no bias in probe selection (Zhang et al., 2015). It provides a comprehensive view of the transcriptome as it is independent of any previous sequence knowledge. Instead, every single transcript (known or unknown) in the sample is sequenced and therefore is possible to identify novel genes and transcripts as well as structural variations (Wei et al., 2014).

Compared to microarrays which measure probe intensities and present poor reproducibility for runs in different platforms, RNA-sequencing quantifies discrete digital read counts aligned to a particular sequence and can be reanalyzed (Draghici, 2006). Due to the sequencing of each individual transcript, the technology is inherently more sensitive and superior in detecting transcripts with low abundance and different isoforms (Zhang et al., 2015). RNA-Sequencing is a new technology and although the data output is potentially much richer in information the analysis is much more complex (Wang et al., 2009). Specialized computing infrastructure is required to take advantage of the initial biological information in these data (Zhang et al., 2015).



Figure 9: **The steps of the RNA-Sequencing method.** This method is used to sequence cDNA copies of all the mRNA in a sample. A very large number of short sequencing reads are aligned to the genome to find the level of expression for each gene. Parts of the reads align to non-contiguous segments of the genome (junction reads) while other reads completely confine to an exon (exonic reads). Junction reads are convenient for the identification of alternative splicing sites. At the end of this figure is a diagram showing the RNA expression level in each nucleotide position (Wang et al., 2009).

The best strategy may be a combination of technologies that maximizes the effectiveness, overcoming the limitations. For example, RNA sequencing provides a comprehensive view of the transcriptome, independently of any sequence knowledge beforehand and is able to detect structural variations. However, there are difficulties in storing the data, the cost is relatively higher and specialized infrastructure is required compared to microarrays. Custom microarrays might be effective tools to capture comprehensive sequencing information and to systematically profile and compare gene expression based on sequencing findings promptly, reproducibly and cost effectively (Zhang et al., 2015).

Microarrays	RNA-Sequencing
+ Robust, reliable method, proven over decades of use + High throughput method with thousands of samples analyzed per month + Can be easily automated + Straightforward data analysis + Lower cost	+ Provides a comprehensive view of the transcriptome + Independent of any prior sequence knowledge + Can detect structural variations such as gene fusions and alternative splicing events + Provides a real digital solution (absolute versus relative abundance)
– Depends on prior sequence knowledge – Structural variations detection is impossible – Isoforms detection is impossible – Hybridization and sample labeling has biases – Relative quantitation method	– New technology to most researchers – Data storage is more challenging – Analysis is more complicated (as there is not a standard protocol) – Specialized computing infrastructure and suitably trained personnel are required

Table 1: **Advantages and shortcomings of DNA Microarrays and RNA-Sequencing** (Zhang, 2015)

Additionally Collected Information

Beyond gene expression, additional information is collected including mutations, amplifications, clinical status and survival.

- Mutations:
  Mutations are referred to a permanent change of a nucleotide sequence in the genome. There is a variety in mutations as there can be a change from a single base pair to huge segments that include more than one gene. They are classified into two types. Firstly, hereditary mutations are inherited from one of the parents and are present during the whole life of the individual. Secondly, acquired (or somatic) mutations are not inherited and emerge during one's life because of environmental reasons (sun radiation, smoking etc.). Mutations are important as they are connected with an increased risk of cancer development (Loeb et al., 2003; Ionov et al., 1993).

- Amplifications:
  Amplification (or gene duplication) is any kind of duplication in a DNA region that contains a gene. They can be detected through the use of microarrays or RNA-Seq platforms. It could be seen as one of many ways of gene overexpression. In

terms of their role in cancer, if an oncogene is duplicated, there are increased possibilities for cancer development (Zhang, 2003).

- Clinical Status:
  Clinical Status is used to describe a patient's condition in a hospital. Patient's clinical status may be referred to their current state (for example it could be positive or serious) or to a prognosis (patient's condition could be improved, getting worse or being stable) (Rich et al., 2010)

- Survival:
  Survival deals with data on time and refers to the ability of a patient to survive over a specified period of time (Rich et al., 2010).

Publicly Available RNA Expression Datasets

Publicly available RNA Expression datasets contain information about **transcriptomics data** derived from either microarrays or RNA-seq experiments (Parkinson et al., 2011; Barrett et al., 2013). Having submitted their data to RNA Expression Databases, scientists can have organized their records, improve the reproducibility of their studies, and finally serve the goal of science generating more knowledge.

Array Express and GEO

Examples of databases where researchers can submit their functional genomics data are **Array Express** and **GEO (Gene Expression Omnibus)**. Both databases first curate the data and then store them so that they are publicly accessible. Array Express is created and supported by EMBL-EBI (European Bioinformatics Institute) while GEO by NCBI (National Center for Biotechnology Information) (Parkinson et al., 2011; Barrett et al., 2013).

Array Express started in 2002, and at first contained only microarray data. However, now its database information is enriched by RNA-seq data too (Parkinson et al., 2011). On the other hand, GEO was created earlier (in 2000) and today consists of data derived from a variety of experimental techniques (microarray data, RNA-seq data and high-throughput data) (Barrett et al., 2013). At the time of writing Array Express hosted 2.223.195 assays from 70.255 experiments in total while the required memory space to save the archived data was 45.65 TB (https://www.ebi.ac.uk/arrayexpress/). By the

same time, GEO hosted 2.148.356 samples from 87.309 different experiments (https://www.ncbi.nlm.nih.gov/gds).

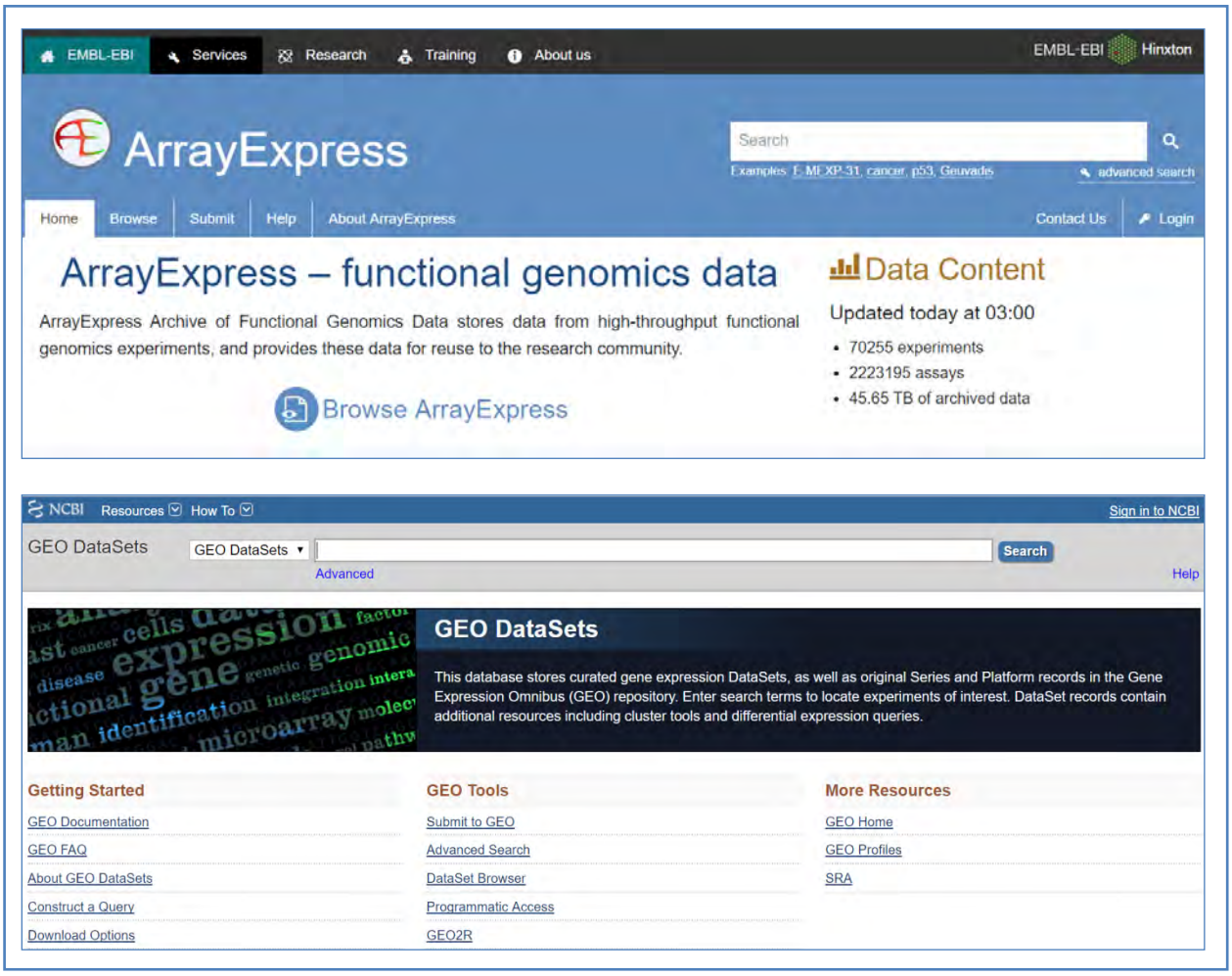


Figure 10: **The web interfaces of ArrayExpress and GEO**. Both databases provide information for specific gene expression data via an advanced search tool. Users have access to ArrayExpress and GEO by the following links: **https://www.ebi.ac.uk/arrayexpress/** and **https://www.ncbi.nlm.nih.gov/gds**

As part of the collaboration between the three biotechnological centers worldwide (**NCBI** in the USA, **EMBL** in Europe and **DDBJ** in Asia) every day the available information is shared among them. The point is the integration and the easiest access to knowledge (Barrett et al., 2012). Part of the information in Array Express derives from GEO, thus Array Express contains data derived from RNA-Seq experiments, too (not only microarray data) (Parkinson et al., 2011).

Both databases, except for RNA expression data contain meta-data. Examples of meta-data consist of the description of each experiment, the protocols and the analysis

techniques in each experiment, the sample annotation and information about the author of the publication (Barrett et al., 2012). Especially for Next Generation Sequencing experiments, there are specifications for each library used as all of them are in accordance with the MIAME protocol (Brazma et al., 2001). MIAME protocol is an international standard for expression data based on microarray experiments, which declares the minimal information required for the experiment to be described. The aim of this protocol is the homogeneity and clearness of the followed procedures in each experiment.

The Cancer Genome Atlas

In 2005 NCI (National Cancer Institute) and NHGRI (National Human Genome Research Institute) collaborated and started a project called **The Cancer Genome Atlas (TCGA)** (Tomczak et al., 2015). The goal was to organize genetic mutations that were connected to cancer and collect expression data for different cancer types into a database. TCGA data derive from more than 10.000 samples from 33 different cancer types (Lee, 2016).

For the development of TCGA and the analysis of patient samples, they were used different techniques like gene expression identification techniques, copy number variation profiling and microRNA profiling (Cancer Genome Atlas Research Network, 2013). For the submission of new information into TCGA, the procedure is strict; the required biospecimens, like blood or tissue, are taken by authorized patients suffering from cancer. Then, the quality of the samples is checked and the data are led to HTS (High-Throughput-Screening). TCGA's data are publicly available, accompanied by a set of tools which provide the research community continuing access to the generated information (Tomczak et al., 2015).

Apart from RNA expression data, **The Cancer Genome Atlas** contains a plethora of other cancer data too. By August 2017, they had been submitted almost 2 Petabytes of continually increasing genomic information, assisting the researcher's efforts to prevent, diagnose and treat the different cancer types.

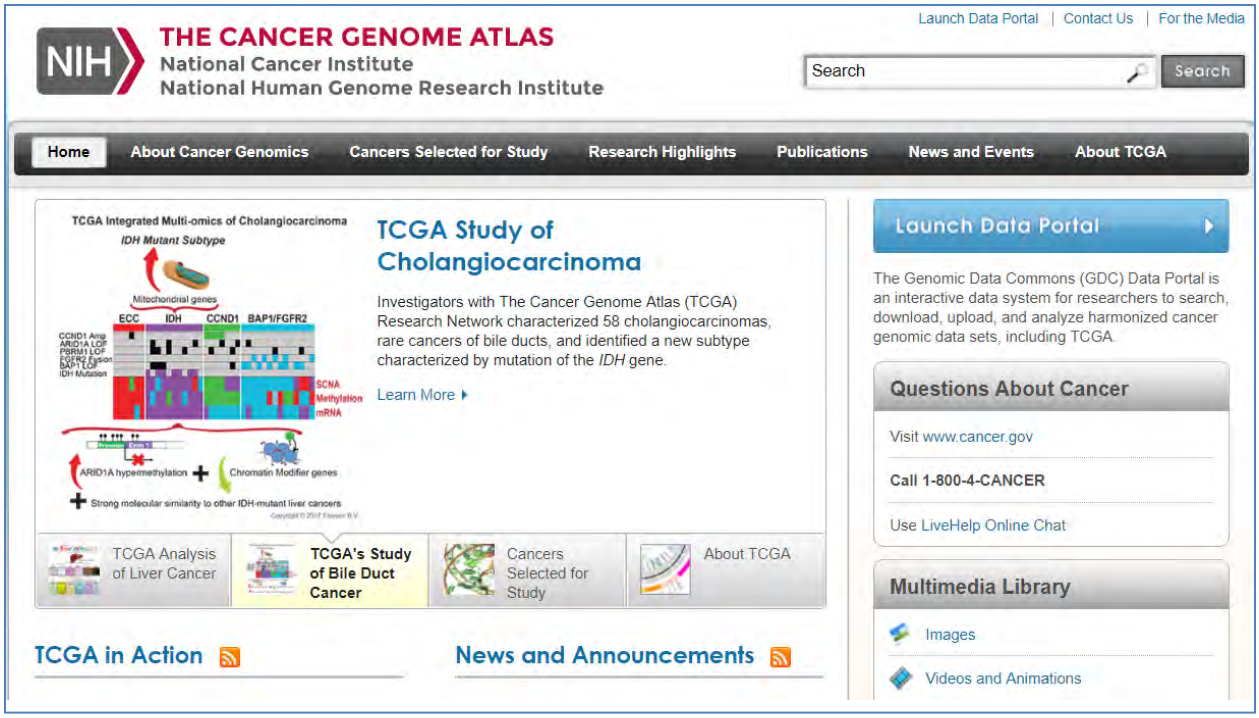


Figure 11: **The web interface of "The Cancer Genome Atlas".** By the following link https://cancergenome.nih.gov/, the user has access to almost 2 Petabytes of information about cancer data, including RNA Expression Data derived from microarray and RNA-seq experiments.

Survival Rates and Kaplan-Meier Plots

Survival analysis deals with data on time until an event occurs. This event describes any change in state or a relapse of a disease (most of the times is death). It is assumed that each subject "suffers" only once from the event in question but in more complicated survival models this restriction could be removed. The biggest complication associated with survival analysis is censoring (Rich et al., 2010). This means that not everyone will necessarily have the event of interest. For instance, if a cancer patient participates in the research, he could not be followed long enough for the event to be observed. In a standard period of time, he may quit the study and no one knows his accurate time of death.

Important part of survival analysis is **survival rates**. Survival rates describe the proportion of people (participating in a study) who are still alive after diagnosis for a particular period of time. The different types of survival rates are (Rich et al., 2010):

- ***Overall Survival***
  The time that has passed from the start of the study or the date of the diagnosis, with the patients diagnosed with the disease remaining alive.

- ***Relapse Free Survival***
  It is also known as “Disease Free Survival” and is referred to the time that passes after the primary treatment for a cancer and the patient remains alive without any symptoms of that cancer. Relapse Free Survival is an indicator to check the effectiveness of new treatments.

- ***Event Free Survival***
  In contrast to the previous types of survival rates, Event Free Survival describes a possibility and not a length of time. After a treatment designed to prevent a group of events, Event Free Survival is the possibility for the patients to have these specific events.

- ***Progression Free Survival***
  Progression Free Survival refers to the time that has passed from the first time that a treatment was applied, until the patient dies or the disease progresses. It is another survival rate to evaluate a treatment on a disease. Specifically in oncology, it refers to the length of time that a patient has a tumor.

- ***Median (Overall) Survival***
  Median Survival expresses the time after which half of the population participating in the study have met the event of interest. The advantage is that the calculation of this type of survival rate requires only half of the population to meet the end of the trial, while the respective calculation of the mean would require all subjects to meet the end of the trial.

- ***Five-year Survival***
  Five-year Survival measures survival at 5 years after treatment

A prominent way to depict survival rates are **Kaplan-Meier plots** (or KM-Plots or Kaplan-Meier curves) (Kaplan et Meier, 1958). They were first published by Edward Kaplan and Paul Meir as an effective way to deal with incomplete observations. Kaplan-Meier plot’s role is even more important when there are patients who are “censored” from the study (some patients may quit before the study meets an end-point).

In medical research, these plots are used to measure the length of time that patients are still alive after a treatment or until a disease progresses. However, they are also used in other fields and in general they measure the length of time until a specific event occurs. For example, a Kaplan-Meier plot could depict the length of time that passes since

people remain unemployed until they find another job (the event of interest is that they find again another job) (Meyer et al., 1990)

Table 2, Table 3 and Figure 12 explain the above mentioned in a simple example (Rich et al., 2010). Table 2 has all the variables to prepare the survival analysis. Therefore for each subject participating in the study the following data are collected:

1. The serial time
   The time passing from the enrolment of the subject into the study until an event of interest or censoring occurs.
2. The status at the end of serial time
   At the end of serial time, the subject could be censored or have met the event of interest.
3. The study group each subject belongs to
   In Kaplan-Meier plots, are compared the survival rates between two different study groups. Each subject falls into the first group or the second one.

The end of the study is at 5 years. From the first group, one subject survived until the end point, while in the second group there were no remaining members after 3.5 years. The mean survival time for Group 1 was 3.25 years (with a 3-year median survival) and for Group 2 was 1.75 years (with 1 year median survival).

SUBJECT	SERIAL TIME (years)	STATUS AT SERIAL TIME (1=event / 0=cencored)	Group (1 or 2)
B	1	1	1
E	2	1	1
F	3	1	1
A	4	1	1
D	4.5	1	1
C	5	0	1
U	0.5	1	2
Z	0.75	1	2
W	1	1	2
V	1.5	0	2
X	2	1	2
Y	3.5	1	2

Table 2: **Patient survival data.** There are 12 patients (subjects), in ascending order according to their serial times. The status at the end of serial time will be 1 if patient died or 0 if patient quit the study before meeting the end-point. The group number for each patient (subject) is shown in the last column (Rich et al., 2010)

GROUP 1			
Event Time (years)	**No. of Events**	**No. at Risk**	**Probability**
1	1	6	0.833
2	1	5	0.667
3	1	4	0.500
4	1	3	0.333
4.5	1	2	0.167

GROUP 2			
Event Time (years)	**No. of Events**	**No. at Risk**	**Probability**
0.5	1	6	0.833
0.75	1	5	0.667
1	1	4	0.500
2	1	2	0.250
3.5	1	1	0.000

Table 3: **Kaplan-Meier results from Table 2.** In the first column is the time (in years) that has passed until an event occurs. When an event occurs, the number of subjects being at risk reduces (column 3) as well as the respective probability of a patient being alive (column 4) (Rich et al., 2010)

Different programming languages have included KM-Plots as functions or packages. For instance in R they are included in the "*survival*" package while in python they are available with the "*lifelines*" package (Ding-Geng et al., 2014). Finally, Matlab has a function called "*ecdf"* with two arguments ("*function*" and "*survivor*") which can calculate and plot them respectively. Using one of those programming languages Table 2 leads to the results shown in Table 3.

Table 3 is spit into the two study groups, with each event being in a different line. An event occurs when a patient dies and not when the subject is censored from the study. For Group 1, after the first event one patient dies, and the total number of 6 subjects at risk is reduced to 5. The probability of surviving past one year is 5/6=0.833. After the second event one more patient dies, the number at risk reduces to 4 and the respective probability is 4/6=0.667. Respectively are calculated the survival probabilities for both groups.

In Figure 12, each horizontal line illustrates the length of time that the survival probability remains steady and is interrupted by an event. The vertical lines show the differences in probabilities after an event has occurred. Both groups have a 100% probability of surviving 6 months. Slightly before the end of the first year, this percentage remains the same for Group 1, while for Group 2 has reduced to about

66.7%. At the end of the study there are no possibilities for a subject from Group 2 to be alive. For Group 1 members, this possibility is almost 20%.

In survival analysis there may be censored subjects. In Figure 12 there are two censored subjects which are depicted by dots. In Group 1, the censored subject is observed at the end of the study, while in Group 2 censoring is observed in the middle of the study. A key assumption is that censoring is uninformative. When a patient drops out of the study is ignored whether he died from the disease the next moment or after many years. Thus the results are caused to be optimistically biased and the probability in Figure 12 overestimates actual survival (Rich et al., 2010).

After observing the two KM-Plots there could be an illusory conclusion about the survival differences of the two groups. Although this difference is obvious, after a log rank-test the corresponding P-value is 0.80 indicating that there is no statistical significance (Rich et al., 2010). The explanation for the statistical insignificance is the small number of subjects that participate in our hypothetical study (Feinstein, 1985).

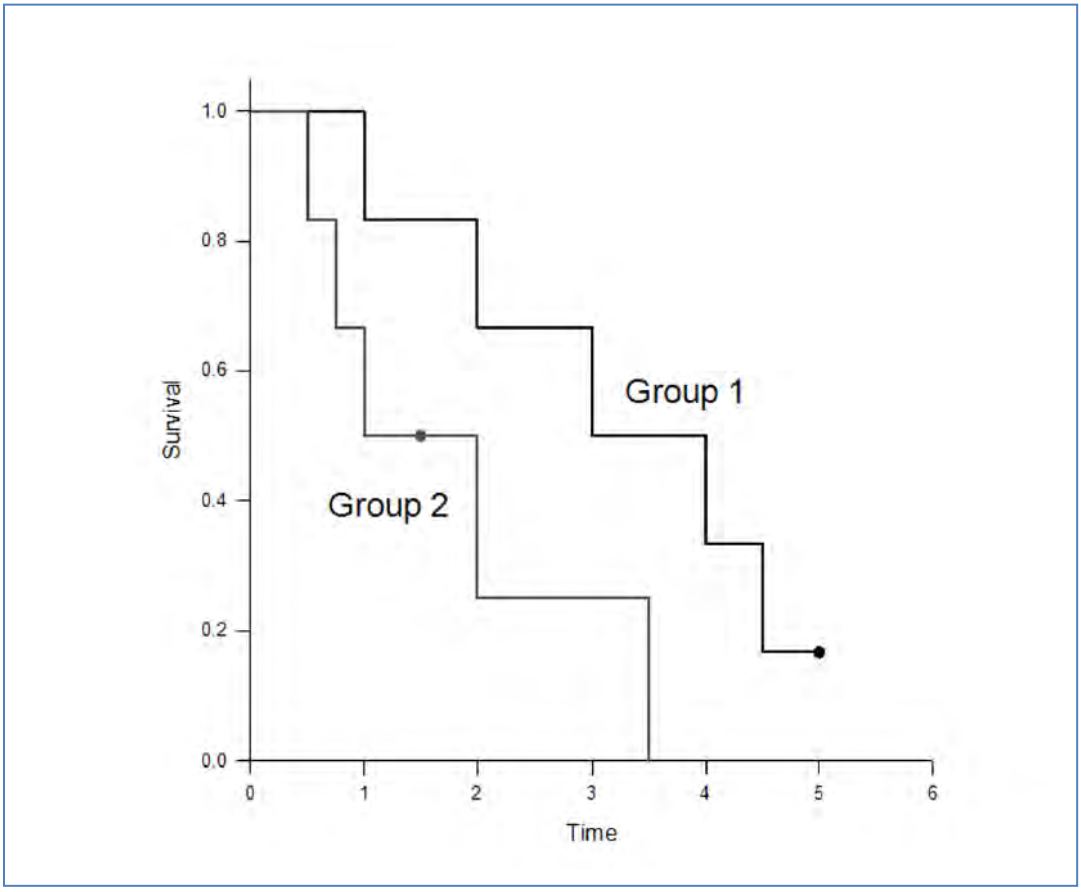


Figure 12: **A Kaplan-Meier plot derived from the Table 3.** The X-axis represents time (in years) while the Y-axis indicates the probability of a patient being alive at a specific time. There are 2 plots in the graph, one for each group of participants in the study (Rich et al., 2010).

KM-Plotter

The identification of biomarkers related to survival is vital for the comprehension of the molecular mechanisms in different types of cancer. KM-Plotter is an online tool, able to assess the effect of high gene expression on cancer patients' survival based on publically available microarray datasets (Gyorffy et al., 2013). In total, data derive from 10461 cancer samples from lung, gastric, ovarian or breast cancer patients.

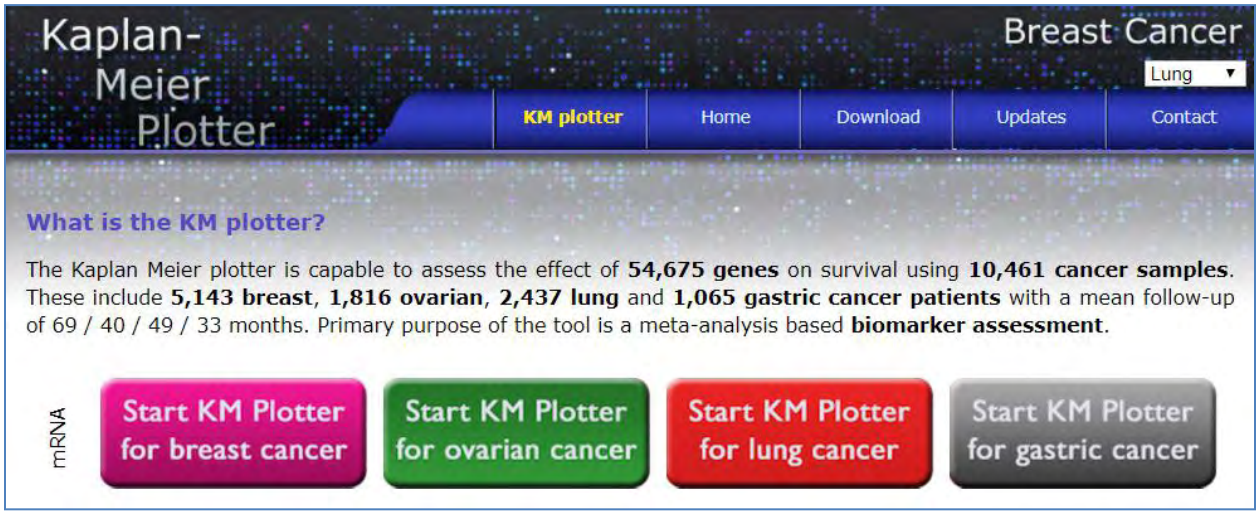
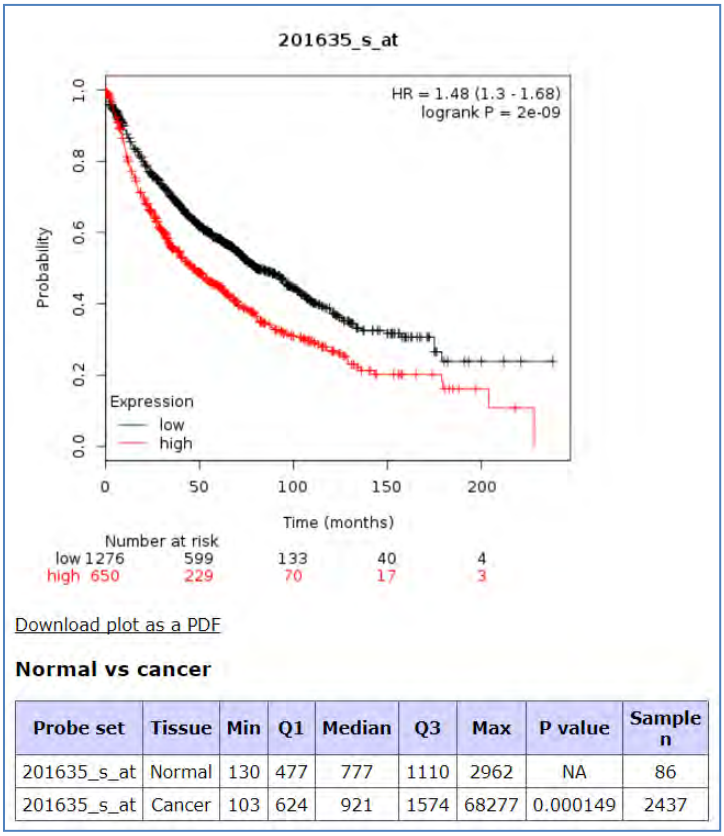


Figure 13: **KM-Plotter's homepage**. There are four different KM-Plotter environments for breast, ovarian, lung and gastric cancer. Using 10461 cancer samples in total, the tool is able to assess the effect of 54675 genes on cancer patients' survival. KM-Plotter is accessible via the following link: www.kmplot.com

KM-Plotter is based on a background database with data deriving from TCGA, EGA and GEO repositories (GEO datasets include only Affymetrix microarray data) (Gyorffy et al., 2012; Gyorffy et al., 2013; Szasz et al., 2016). Survival and gene expression data are integrated into the system and the samples are divided into two groups, according to the (median, lower or upper quartile) expression levels of the selected gene. A Kaplan-Meier plot compares the two patient groups and an estimated logrank p-value expresses the statistical significance (Gyorffy et al., 2012). A low p-value rejects the null hypothesis that there is no survival difference amongst these two groups of cancer patients with high or low gene expression.

Among KM-Plotter's results could be multiple plots due to the concordance of multiple probes available for each gene. As an example, for a random gene called FXR1 the results include five different plots (one of those is depicted in Figure 14). Patients are

divided into two groups according to the gene's expression levels. The groups with high and low expression levels are shown with red and black color respectively. In this Kaplan-Meier plot, red line is below the black one, thus there is a negative effect on lung cancer patients' survival when FXR1 gene is over-expressed.



Download plot as a PDF

Normal vs cancer

Probe set	Tissue	Min	Q1	Median	Q3	Max	P value	Sample n
201635_s_at	Normal	130	477	777	1110	2962	NA	86
201635_s_at	Cancer	103	624	921	1574	68277	0.000149	2437

Figure 14: **KM-Plotter results for FXR1 gene**. Among the results is a Kaplan-Meier plot showing the effect of high and low expression levels and a p-value (logrank P = 2e-09) calculating the statistical significance of high versus low level expression of FXR1 on lung cancer patients' overall survival. Below the plot (in the table) are the expression levels of FXR1 in normal and timorous tissues and the statistical significance (p-value = 0.000149) of this difference in lung cancer datasets.

RNA-Binding Proteins

RNA binding proteins (RBPs) are proteins (more often present in the nucleus or the cytoplasm) that bind to the RNA in cells to form ribonucleoprotein complexes (Lunde et al., 2007). By August 2017, in human organism there were 1542 experimentally validated RBPs, representing 7.5% of the total protein-coding genes (Pereira et al., 2017). RBPs are regulators of the RNA's translation process, involve in RNA synthesis, participate in cellular function, transport and localization and are vital in gene expression's post-transcriptional regulation (Dreyfuss et al., 2002).

Most RBPs consist only of a few simple domains (an RNA-binding domain is an aminoacid sequence that detects RNA binding sites). Among these structural motifs are RRM (RNA recognition motif), zinc fingers, S1 domain and K-homology domain (Pereira et al., 2017). A bond between a single domain and RNA is weak. By exploiting the multiple weak interactions from each individual bond, RBPs are able to form stronger bonds with RNA, binding longer segments of nucleic acids (Lunde et al., 2007). While each domain recognizes just a specific RNA, the diversity requirements are accomplished by arranging these sequences in different ways. Multiple repeats of the same domains are responsible for the structure and function of the RBP (Lunde et al., 2007). A characteristic example of RBP with multiple repeats of the same domain is human Pumilio protein. While a single domain is able to recognize only one nucleotide, eight combined repeats of domains lead to a formation that recognizes a much longer nucleotide sequence (Figure 15b)

Another factor that determines the function of RBPs is the length of the linker between two domains. A longer linker recognizes diverse stretches of nucleic acids and aids RBPs to recognize a plethora of RNA targets (Lunde et al., 2007). In RBPs are observed an increased number of intrinsically disordered regions (protein regions that deprive of a defined structure) which enhances RBP's flexibility and adaptation to different nucleotide sequences and structures (Pereira et al., 2017). However, even though their role is so important, RNA-Binding proteins have not been systematically studied (Hogan et al., 2008).

RNA transcripts are first formed in the nucleus and then instantly bind to RBPs, forming a ribonucleoprotein complex with specific characteristics and roles. RBPs are as diverse as possible to bind a variety of RNAs with different sequences (not only mRNAs but also a variety of non-coding RNAs) (Matera et al., 2017). Every RBP is flexible and able to bind different RNAs, controlling the metabolism of many transcripts. Their aberrant expression has profound effects on cellular physiology and the pathogenesis of human disease (Lukong et al., 2008). Multiple studies indicate that for many RBPs the altered expression level noticed in cancerous compared to normal tissues, is connected with the development of many cancer types (Pereira et al., 2017).

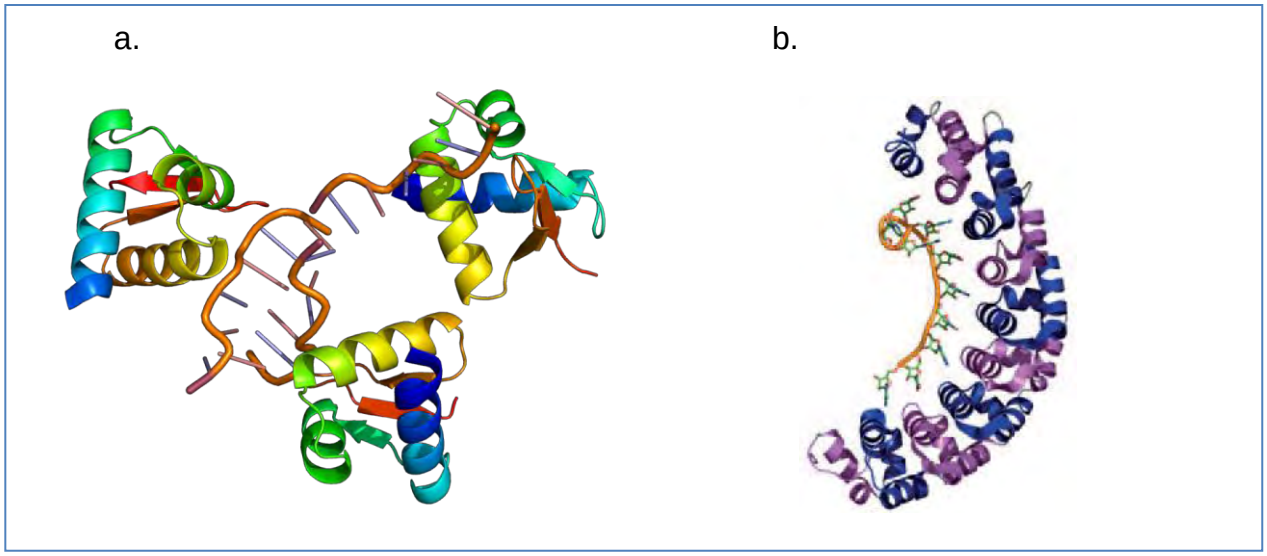


Figure 15: **The structure of two RBPs; ADAR and Pumilio. a.** The structure of ADAR, an RNA-Binding protein which edits mRNA. ADAR operates after transcription and changes nucleotides in mRNA's sequence. Specifically, it converts the nucleoside adenosine to inosine via the catalysis of an enzymatic reaction. After this process, RNA sequence is different than the sequence encoded by the DNA (Glisovic et al., 2008). **b.** The structure of human Pumilio protein. While a single domain recognizes only one nucleotide, eight repeats in Pumilio's structure recognize a larger RNA sequence. The RNA's backbone is colored orange, while the repeats in RBP are colored blue and purple respectively (Lunde et al., 2007).

Chapter 2: Aim and Motivation

Identifying new therapeutic targets for the treatment of cancer is of paramount importance. RNA binding proteins (RBPs), central regulators of gene expression, constitute ideal candidate targets which have been poorly exploited so far. As part of a drug target discovery project, this thesis aims to reduce the total number of RBPs that will be taken forward into drug screening for novel anti-tumour compounds, based on the results of a computational analysis.

A list of 1334 candidate RBPs was compiled by the Lewis laboratory based on available datasets of RNA binding proteins encoded by the human genome. The aim of this thesis is to use available data on the expression levels of the corresponding genes and cancer survival data in order to restrict this list to the most promising candidates, to be taken forward to drug screening.

Specifically the planned steps in the analysis are:

1. Collect available information on the effect of the expression levels of each one of the candidate RBPs on cancer patient survival. To that end, an online tool called KM-Plotter (Gyorffy et al., 2013) will be used to calculate the statistical significance (p-value) of high versus low-level expression of each RBP in cancer patient overall survival, using data from hundreds of patients with four different human cancers: lung, gastric, ovarian and breast cancer.
2. Collect information on the expression levels of each RBP in normal and cancerous tissues and the statistical significance (p-value) of this difference in the same four cancer datasets.
3. Progressively apply filters to narrow down the candidate RBPs, based on negative or positive effect of expression levels on patient survival and the corresponding p-values, the difference in expression levels of RBPs in tumour versus normal cells, the concordance of multiple probes available for each gene and the differential effects in different types of cancer to produce a concise list of the most promising candidate molecules
4. Perform network analysis on candidate RBPs. To that end, the STRING database of known and predicted protein-protein interactions will be used. RBPs constituting a hub in the network are estimated as more likely to be successful drug targets.

The computational analysis pipeline to be followed is outlined below. Starting with 1334 RBPs, the aim of this thesis is to restrict the RBPs list, keeping less than 30 RBPs. That is the final list of RBPs that will be taken forward into drug screening.

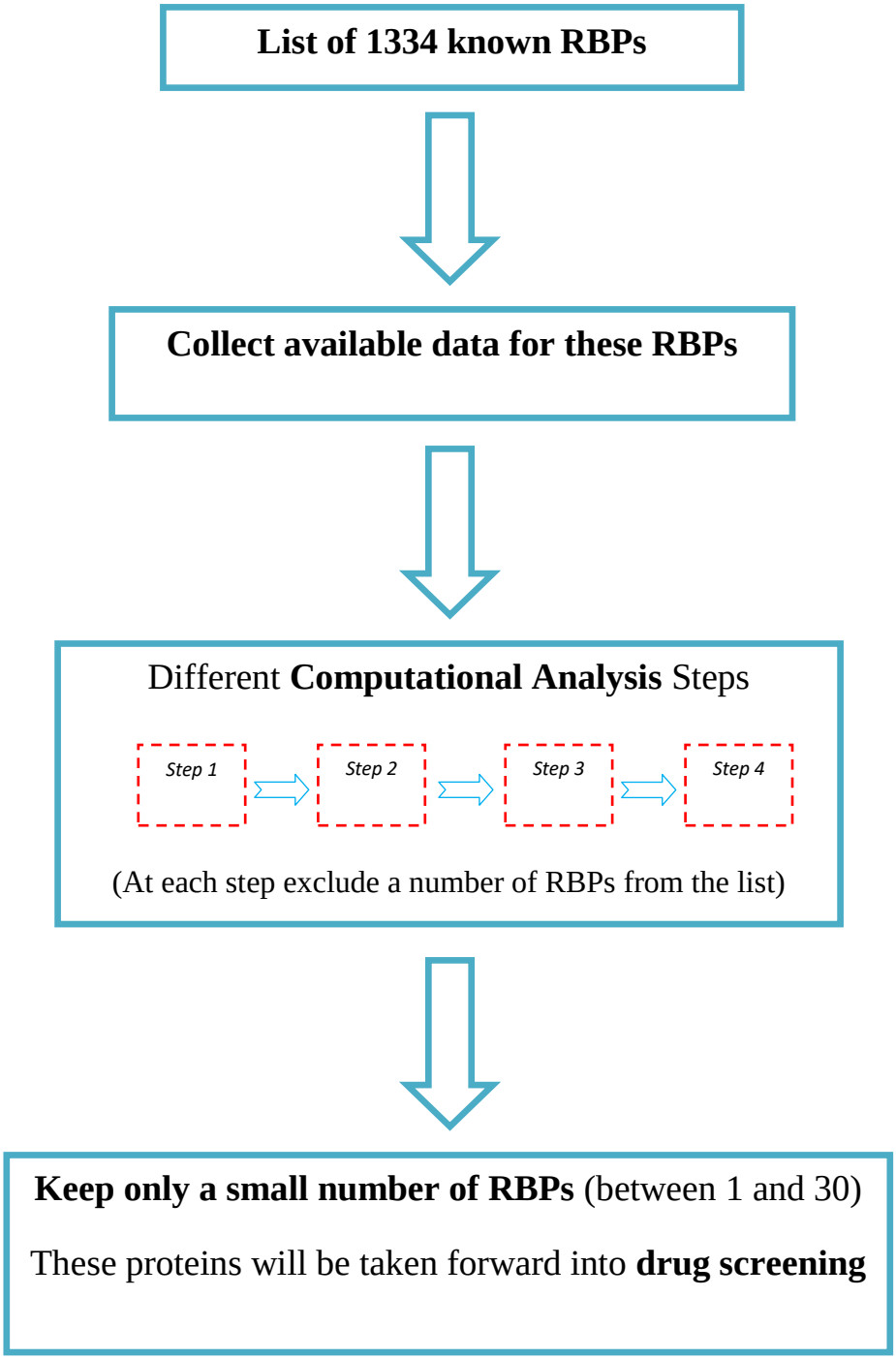


Figure 16: **Outline of thesis aim**

Chapter 3: Materials and Methods

This thesis reduces the total number of RBPs that will be taken forward to drug screening for novel anti-tumour compounds, based on the results of a computational analysis. A list of 1334 RBPs was compiled by the Lewis laboratory based on available datasets of RNA binding proteins encoded by the human genome.

For the collection of available data was used an online tool called KM-Plotter (Gyorffy et al., 2013), accessible via the following link: www.kmplot.com. The collected information was on the effect of the expression levels of each one of the candidate RBPs on cancer patients' survival and the expression levels of each RBP in normal and cancerous tissues. KM-Plotter has one basic environment for each type of cancer (lung, gastric, ovarian, breast). In the compiled list there are 1334 RBPs, thus data retrieved for all RBPs and for each one of the 4 cancer types will be collected. Using 10461 cancer samples in total, KM-Plotter is able to assess the effect of 54675 genes on cancer patients' survival.

KM-Plotters for different Cancer Types

Lung cancer KM-Plotter

For the development of the Lung Cancer database, they were used microarray datasets from 3 different repositories (Gyorffy et al., 2013):

1. **caBIG (Cancer Biomedical Informatics) (Cimino et al., 2007)**
2. **GEO (Gene Expression Omnibus) (Clough et al., 2016)**
3. **TCGA (The Cancer Genome Atlas) (Tomczak et al., 2015)**

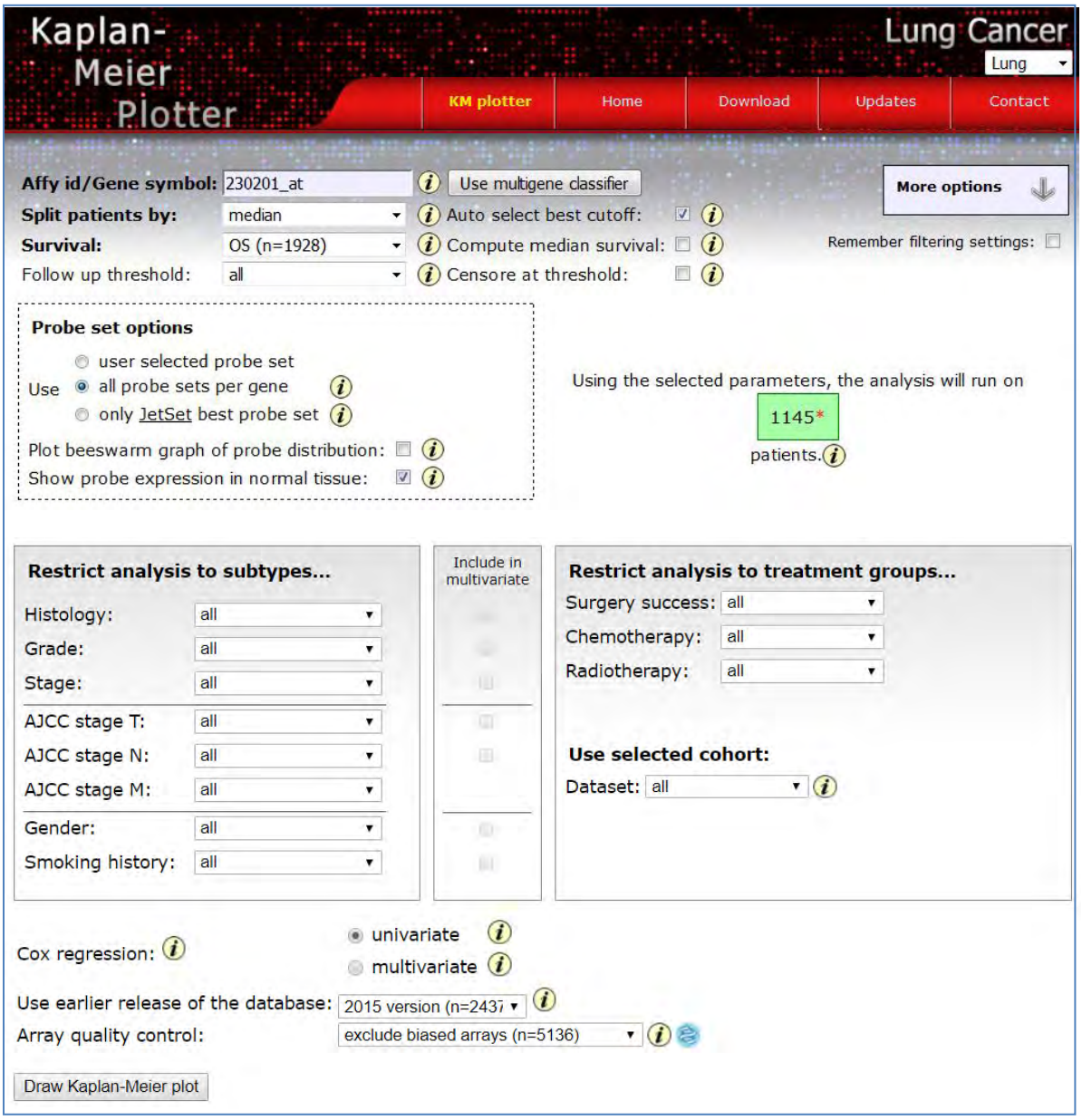


Figure 17: **The Lung KM-plotter environment with all chosen settings.** A random gene called FXR1 is used as an example. The Affy id symbol represented in KM-Plotter for this gene is 230201_at.

These datasets included information for the expression of genes and the survival rates of cancer patients. In total, the collected data corresponded to **2437 lung cancer patients**.

In Figure 17 there is an example, for a random gene in lung cancer. The gene is called “FXR1” and the Affy id symbol used in KM-Plotter is “230201_at”. It is obvious that the

collection of data for all the other genes (corresponding to the rest RPBs of the compiled list) follows the same procedure.

The selected settings for the Lung KM-Plotter are the following (The same settings are used for the other 3 cancer types too):

- Affy id / Gene symbol:
  On **Affymetrix microarrays** a gene can be detected by multiple probe sets with measurements that sometimes are contradictory. To avoid this inconvenience a scoring method (called JetSet) has been applied and assesses each probe set in different terms (sensitivity, specificity, coverage) (Li et al., 2011)
  In KM-Plotter the color code corresponds to the JetSet status of the current gene, with the Green Color representing the best probe set, while the Yellow and Red Color showing worse probe sets for a gene. The recommended JetSet probe set is the **green** one (Figure 18). For all the genes of the compiled list that were searched on this tool, a green probe set was available.

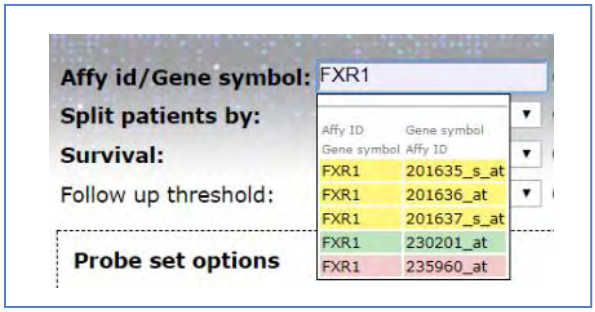


Figure 18: **Multiple probe sets available for a specific gene name.** The recommended JetSet probe set is the green one. After repeating the procedure for the total number of 1334 genes, for all genes was available a green JetSet probe set. However, for the results of this thesis we used all available probes, independently of their color.

- Split patients by: **median**
  The alternatives are lower quartile, lower tertile, upper tertile and upper quartile. Median is the default setting.

- Survival: **OS**
  OS (Overall Survival) represents the period of time that the patients continue to be alive, starting from the date of diagnosis or the beginning of treatment.

- Follow up threshold: **all**
  To generate the Kaplan-Meier plot all available patients are included.

- Auto select best cutoff: **YES**
  All the percentiles between the lower and upper quartiles are computed and the best performing threshold is used as a cutoff.

- Compute median survival: **NO**
  Median survival will be computed only in case both cohorts reach a median survival. Upper quartile survival will be computed in the remaining cases

- Censore at threshold: **NO**
  Patients surviving over the selected “Follow up threshold” are censored (It doesn’t change our results, because at the “Follow up threshold” we selected “all”)

- Probe set options: **all probe sets per gene**
  The analysis will run on all available probe sets representing the selected gene. In case only JetSet best probe set was used, the optimal probe set for each gene would be selected by JetSet.

- Show probe expression in normal tissue: **YES**
  The expression table will be shown for the selected probe in normal and cancerous tissue.

Using FXR1 gene as an example and with the selected parameters (as mentioned above), the analysis runs on 1145 patients. The color of the box depicts the robustness of the sample number. In this case it is **green**, meaning more than 500 samples, thus the **analysis is highly reliable**. Blue, yellow and red, would mean a neutral (200-500 samples), a preliminary (50-200 samples) or an explorative analysis (less than 50 samples) respectively.

The rest of the settings will have the default values. If the users want to apply more settings it is possible to restrict analysis according to histology (adenocarcinoma or squamus cell carcinoma), grade (I, II or III), cancer stage (1, 2, 3 or 4), gender or smoking history (using only those never smoked or excluding them). The analysis could be restricted to treatment groups that used radiotherapy or chemotherapy and if the surgery was successful (it is successful if no cancer cells are seen at the outer edge of the lung).

For the analysis all available datasets are included. Also, not all patients have all clinical information, so the multivariate analysis uses fewer patients than the univariate.

Finally, the database release that is used is the earliest one (2015 version, including 2437 patients with available clinical data). About the array quality control, the default settings are used, meaning that biased arrays will be excluded.

Collected Characteristics	Different Values	Number of Patients
Histology	Adenocarcinoma	866
	Squamous Cell Carcinoma	675
Grade	I (poor)	202
	II (moderate)	310
	III (well)	77
Stage	1	652
	2	320
	3	70
	4	4
AJCC stage T	1	475
	2	686
	3	99
	4	48
AJCC stage N	0	863
	1	296
	2	113
AJCC stage M	0	818
	1	10
Gender	Male	1387
	Female	818
Smoking history	Exclude those never smoked	970
	Only those never smoked	247
Surgery success	Only surgical margins negative	730
Chemotherapy	No	317
	Yes	178
Radiotherapy	No	276
	Yes	73
Entire Lung Database (2015 version): 2437 Patients		

Table 4: **Collected Patient Characteristics of the datasets included in the analysis for the lung cancer** (Gyorffy et al., 2013)

The collected clinical characteristics are the following:

- Stage of cancer (Describes if the cancer has spread to other tissues than the lung)

- Stage **T** (Describes the size of the primary **tumor** and if there is a nearby tissue metastasis)
- Stage **N** (Describes whether nearby lymph **nodes** are involved)
- Stage **M** (Describes if there is **metastasis** in distant tissues)
- Histology (Lung cancer can be grouped into histological categories, like adenocarcinoma or squamous cell carcinoma) (Maggiore et al., 2004)
- Grade (In the collected clinical characteristics, cancer is categorized as poor, moderate or well. In "well" grade, metastasis is a common phenomenon, while in "poor", cells are less likely to spread)
- Gender (The patient is possible to be a male or a female)
- Smoking History (For each lung cancer patient, there is information about smoking history)
- Surgery success (After the surgery, it is examined if lung is clear or if there are any cancer cells. In case no cancer cells are seen at the outer edge of the lung, the surgical margin is characterized as negative. In the positive case, more surgery is needed to remove the remaining cancer cells)
- Chemotherapy (The patient could have used chemotherapy treatment or not)
- Radiotherapy (The patient could have used radiotherapy treatment or not)

The collected clinical characteristics for the development of (Lung Cancer) KM-Plotter are summarized in table 4.

Gastric cancer KM-Plotter

The database for the Gastric cancer (Szasz et al., 2016) retrieved microarray data from 4 different sources (3 of them are research centers which have also published their results in GEO):

1. **Max Delbruck Center for Molecular Medicine center in Berlin (Forster et al., 2011)**
2. **National Cancer Institute in Bethesda (Kim et al., 2012)**
3. **Peter MacCallum Cancer Center in Melbourne (Busuttil et al., 2014)**
4. **GEO (Gene Expression Omnibus) (Clough et al., 2016)**

In total, data were collected from **1065 gastric cancer patients**.

For the Gastric Cancer we follow the same procedure as we did for the Lung Cancer. Again, the gene “FXR1” with the Affy id symbol “230201_at” is used as an example (Figure 19).

This time, for the FXR1 gene the analysis runs on 631 patients. Similarly with the lung cancer case, the green box color means that there are more than 500 samples (the analysis is highly reliable). For each gene in the compiled list, this box was always green, meaning more than 500 samples and a robustness of the sample number.

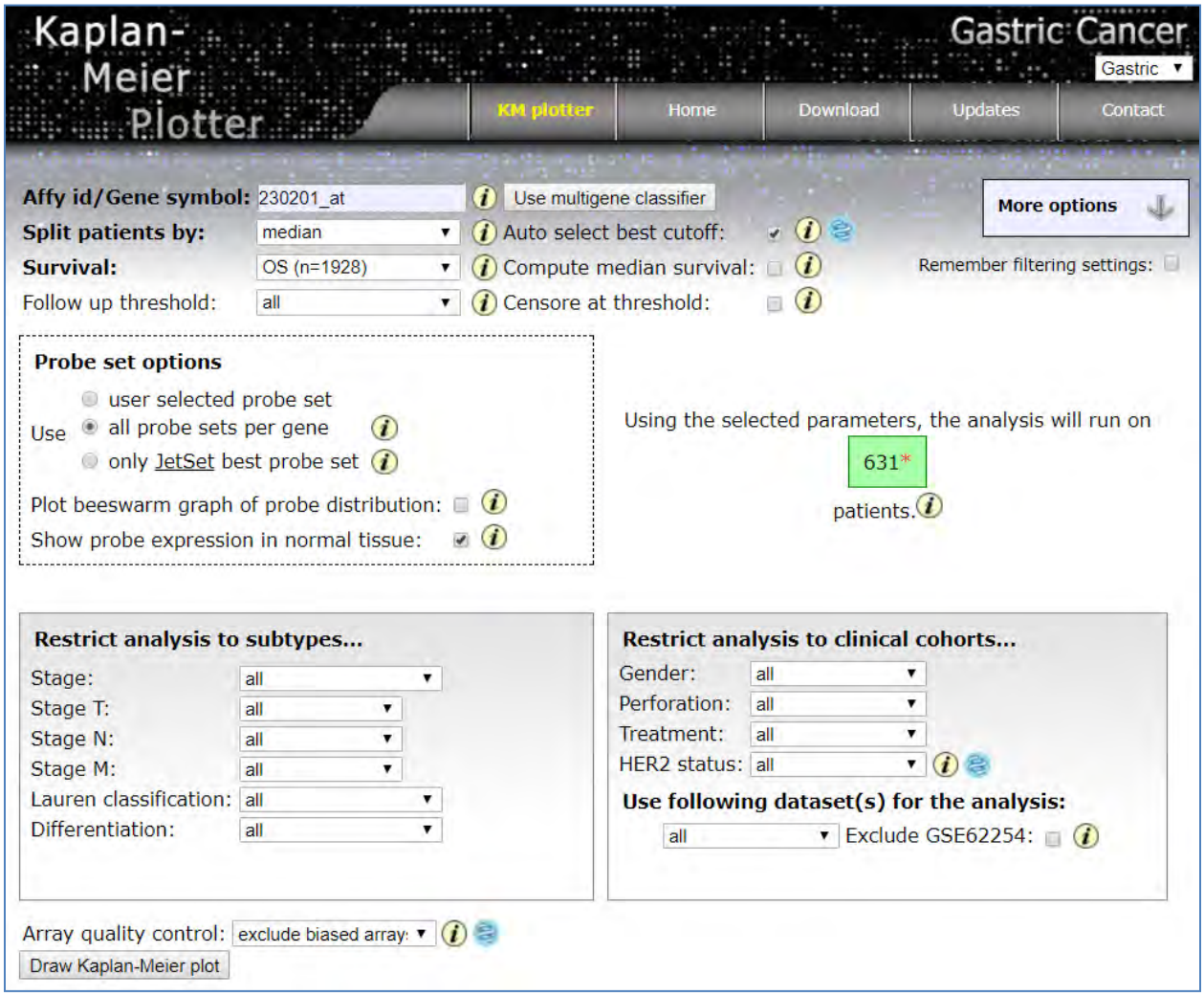


Figure 19: The Gastric KM-plotter environment with all chosen settings. A random gene called FXR1 is used as an example. The Affy id symbol represented in KM-Plotter for this gene is 230201_at.

For the rest of the settings, the default values are used. However, for gastric cancer the selected settings (same with the collected clinical characteristics) are different. For example in lung cancer there were collected characteristics like “Smoking History” which is important for patient outcome in lung cancer but less trivial in gastric cancer.

Collected Characteristics	Different Values	Number of Patients
Stage	1	652
	2	320
	3	70
	4	4
Stage T	1	14
	2	253
	3	208
	4	39
Stage N	0	76
	1	232
	2	129
	3	76
Stage M	0	459
	1	58
Lauren classification	Intestinal	336
	Diffused	248
	Mixed	33
Differentiation	Poorly differentiated	166
	Moderated differentiated	67
	Well differentiated	32
Gender	Male	1387
	Female	818
Perforation	Yes	4
	No	169
Treatment	Surgery alone	393
	5 FU based adjuvant	158
	Other adjuvant	80
HER2 status	Positive	425
	Negative	641
Entire Gastric Database: 1065 Patients		

Table 5: **Collected Patient Characteristics of the datasets included in the analysis for the gastric cancer** (Marcell Szasz et al., 2016)

The collected clinical characteristics are summarized in table 5 and are the following:

- Stage of cancer (Describes if a cancer is affecting other parts of the body)
- Stage **T** (Describes the size of the primary **tumor** and if there is a nearby tissue metastasis)
- Stage **N** (Describes whether nearby lymph **nodes** are involved)
- Stage **M** (Describes if there is **metastasis** in distant tissues)
- Lauren classification (Divides adenocarcinoma of the stomach into 2 main types; intestinal type and diffused type)
- Differentiation (Adenocarcinoma presents three degrees of differentiation; poor, moderate, well)
- Gender (A patient can be male or female)
- Perforation (A hole in part of the wall of stomach or small intestine) (Roviello et al., 2006)
- Treatment (In the treatment can take part only surgery or an adjuvant too)
- HER2 status (HER2 is a gene that is often overexpressed in breast cancer) (Kroese et al., 2007)

Ovarian cancer KM-Plotter

For the ovarian cancer database (Gyorffy et al., 2012), two repositories contributed their data:

- **GEO (Gene Expression Omnibus) (Clough et al., 2016)**
- **TCGA (The Cancer Genome Atlas) (Tomczak et al., 2015)**

In total, microarray data and survival information from **1816 ovarian cancer patients** are used. As an example, for the gene used in Lung and Gastric KM-Plotter (FXR1 gene) the analysis runs on 655 patients (Figure 20). For all RBPs there were more than 500 samples (this number is a threshold, indicating highly reliable analysis). In addition, the database release for ovarian KM-Plotter was the earliest one (2017 version with all data available, collected from 1816 patients)

The collected characteristics (which are also part of the settings in case a user wants to restrict the analysis to subtypes) for the ovarian cancer patients are the following (summarized in table 6):

- Stage of cancer (Describes if a cancer has spread to other tissues than the ovaries)

- Histology (Ovarian cancer can be grouped into histological categories, like endometrioid or serous) (Hauptmann et al., 2014)
- Grade (Cancer can be categorized as grade 1, 2, 3 or 4. or as “high” or “low” grade. In grade 4 metastasis is a common phenomenon, while in grade 1 cells are less likely to spread)
- TP53 mutation (Mutations in TP53 are associated with a variety of cancers) (Olivier et al., 2010)
- Debulk (The reduction of as much of the volume of a tumor as possible is achieved by surgical removal. The tumor removal can be optimal or suboptimal in our case)
- Chemotherapy (For patients undergone to chemotherapy, there is information about the patient’s number who used specific drugs)

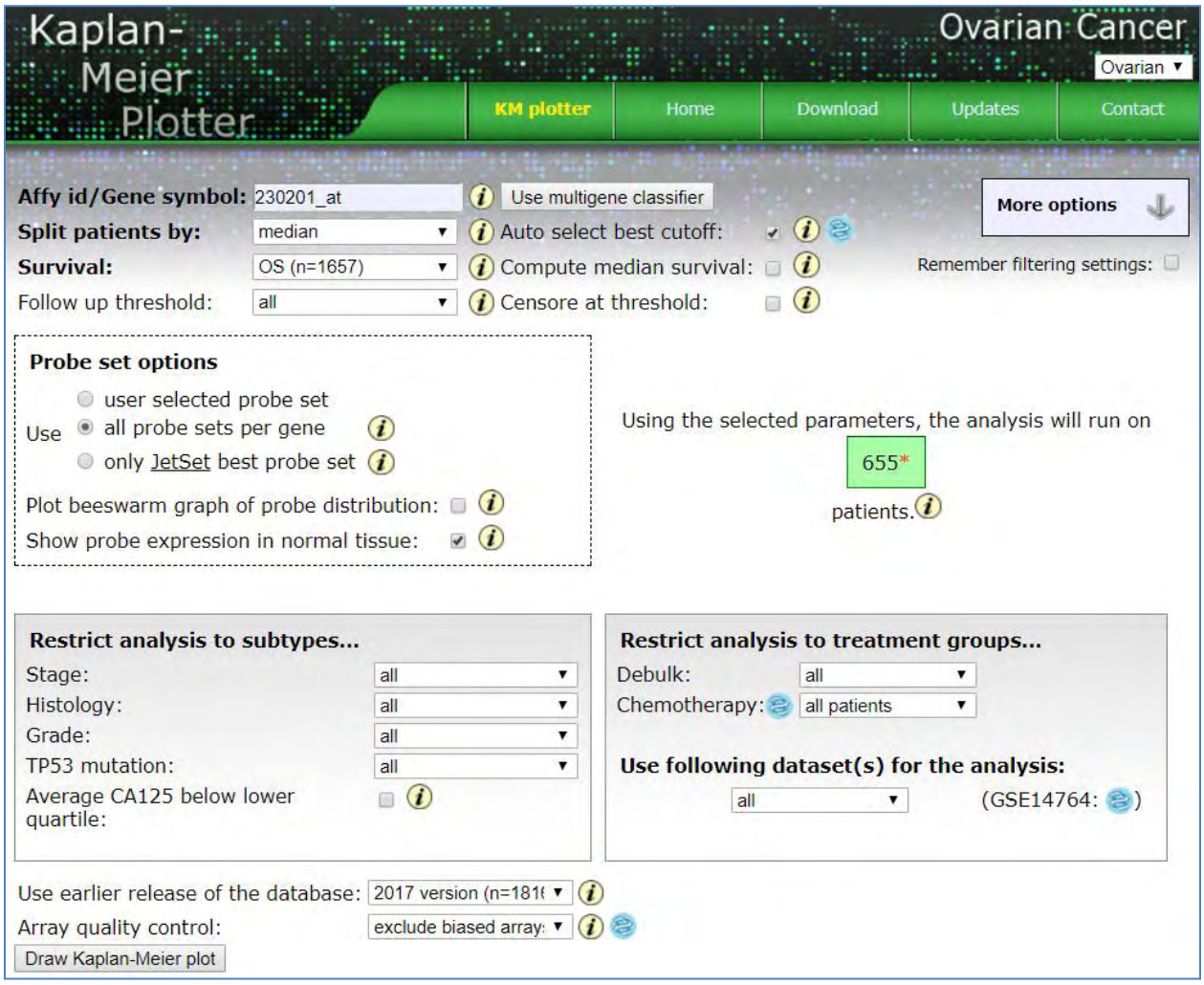


Figure 20: **The Ovarian KM-plotter environment with all chosen settings.** FXR1 gene is used as an example. The Affy id symbol represented in KM-Plotter for this gene is 230201_at.

Collected Characteristics	Different Values	Number of Patients
Stage	1	107
	2	72
	3	1079
	4	189
Histology	Endometrioid	62
	Serous	1232
Grade	1	56
	2	325
	3	1024
	4	21
TP53 mutation	Mutated	516
	Wild type	102
Debulk	Optimal	802
	Suboptimal	536
Chemotherapy	Contains platin	1438
	Contains Taxol	821
	Contains platin + Taxol	804
	Contains Avastin	50
	Contains Docetaxel	108
	Contains Gemcitabine	135
	Contains Paclitaxel	248
	Contains Topotecan	119
Entire Ovarian Database: 1816 Patients		

Table 6: **Collected Patient Characteristics of the datasets included in the analysis for the ovarian cancer** (Gyorffy et al., 2012)

Breast cancer KM-Plotter

For the development of Kaplan-Meier Plotter in breast cancer (Gyorffy et al., 2010), are used in total data from **5143 breast cancer patients**. Survival information and gene expression microarray data were downloaded from **GEO (Gene Expression Omnibus)** (Clough et al., 2016).

Specifically for the example-gene FXR1, the analysis runs on 1764 patients (Figure 21). A bigger number of the used breast cancer data, indicates a higher reliability in the analysis compared to the previous 3 cancer types.

Our analysis uses the earliest version of the database (2017 version) with data from 5143 breast cancer patients.

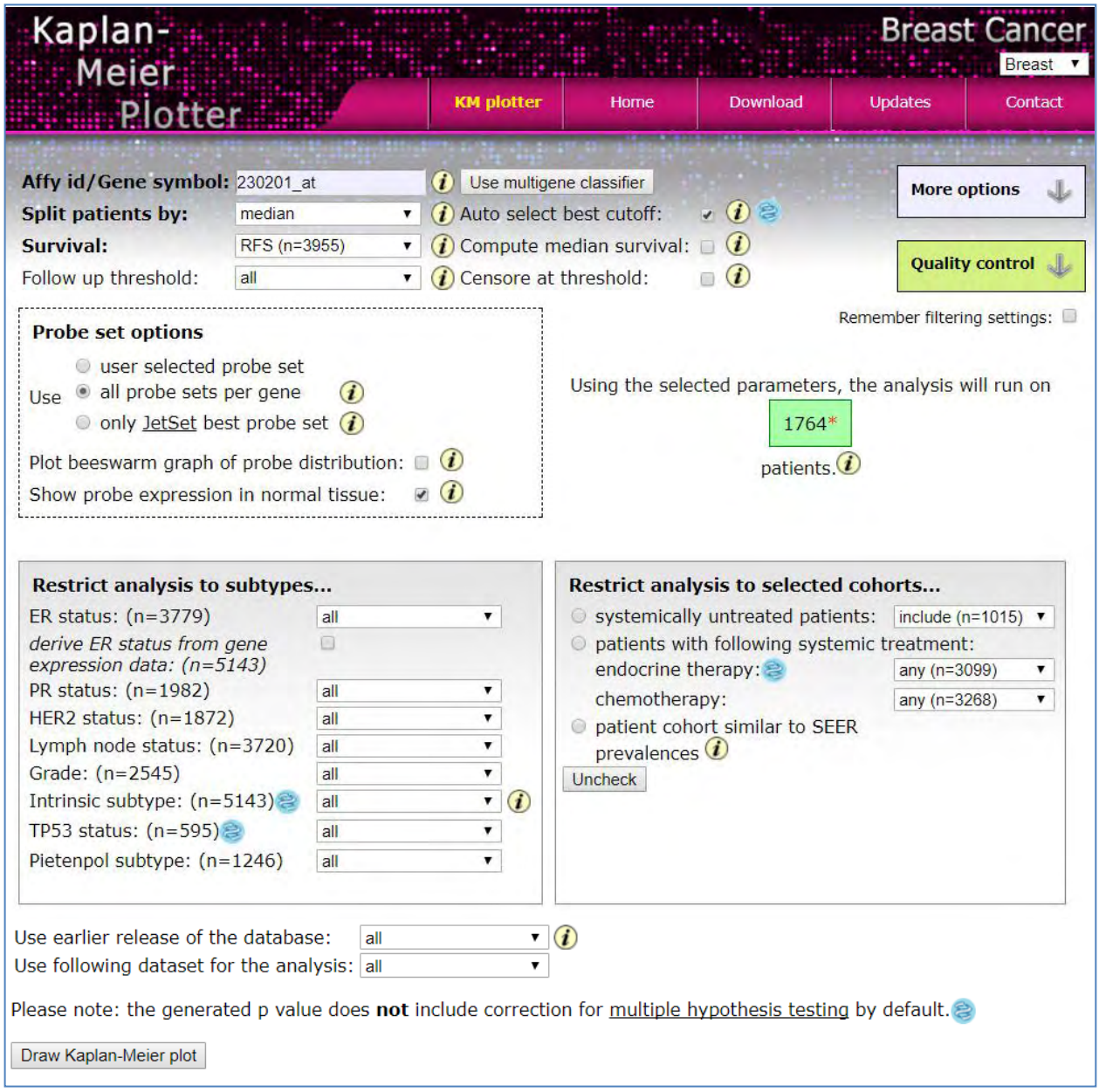


Figure 21: **The Breast KM-plotter environment with all chosen settings.** A random gene called FXR1 is used as an example. The Affy id symbol represented in KM-Plotter for this gene is 230201_at.

Collected Characteristics	Different Values	Number of Patients
ER status	ER positive	2565
	ER negative	1214
PR status	PR positive	954
	PR negative	1028
HER2 status	HER2 positive	416
	HER2 negative	1456
Lymph node status	Lymph node positive	1459
	Lymph node negative	2259
Grade	1	378
	2	1077
	3	1090
Intrinsic subtype	Basal	879
	Luminal a	2504
	Luminal b	1425
	HER2+	335
TP53 status	Mutated	232
	Wild type	363
Pietenpol subtype	Basal-like 1	239
	Basal-like 2	97
	Immunomodulatory	290
	Mesenchymal	229
	Mesenchymal stem-like	115
	Luminal androgen receptor	276
Endocrine Therapy	Include	1381
	Exclude	1718
	Tamoxifen only	1043
Chemotherapy	Adjuvant	678
	Neoadjuvant	884
Entire Breast Database: 5143 Patients		

Table 7: **Collected Patient Characteristics of the datasets included in the analysis for the breast cancer** (Gyorffy et al., 2010)

The collected characteristics for the breast cancer patients are:

- ER status (Cancer development is fueled when the hormone estrogen is present. Based on whether cancers have this hormone or not, are characterized respectively as positive or negative)
- PR status (Cancer development is fueled when the hormone progesterone is present. Based on whether cancers have this hormone or not, are characterized as positive or negative)
- HER2 status (HER2 is a gene that is often overexpressed in breast cancer) (Kroese et al., 2007)

- Lymph node status (Breast cancer is possible to metastasize to different parts of the body through the lymph systems. If the lymph nodes are clear of cancer, they are characterized as negative, otherwise as positive)
- Grade (In this case, cancer is categorized as grade 1, 2, 3. In grade 3 metastasis is a more common phenomenon than in grade 1 in which cells are less likely to spread)
- Intrinsic subtype (Is referred to several distinct breast cancer subtypes. Luminal A and B are the most common. Other types are the HER2+ and the Basal-like tumors) (Dai et al., 2015)
- TP53 status (Mutations in TP53 are associated with a variety of cancers. The status of the gene can be “mutated” or “wild type”. By “wild type” is meant the typical form of the gene, without being mutated) (Olivier et al., 2015)
- Endocrin therapy (Some breast cancer patients underwent endocrine therapy. Information is also collected about the number of them who have used only tamoxifen, a medication to prevent breast cancer)
- Chemotherapy (Adjuvant therapy is an administration given after the initial therapy for better results while Neoadjuvant therapy is administrated before the main therapy to reduce the size of the tumor)

The collected patient’s characteristics are summarized in table 7.

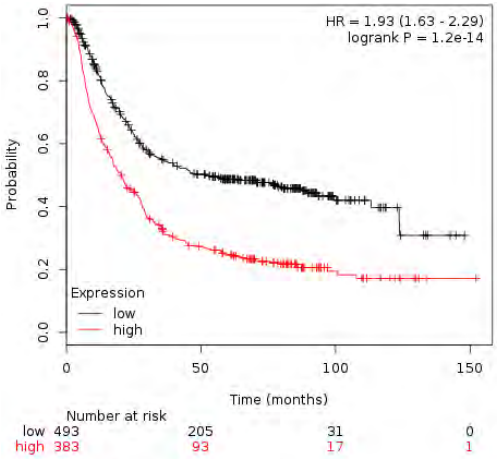
Data Collection

In the previous sections, the settings used in KM-Plotter were described in detail. In Figure 22 are presented the KM-Plotter’s results for three different genes (LUC7L, LARP4 and CPEB2, respectively). A gene in KM-Plotter has multiple available probes, with each probe corresponding to a different plot.

The following three cases are possible:

1. High gene expression has a negative effect on cancer patient outcome (Figure 22a)
2. High gene expression has a zero effect on cancer patient outcome (Figure 22b)
3. High gene expression has a positive effect on cancer patient outcome (Figure 22c)

a) LUC7L

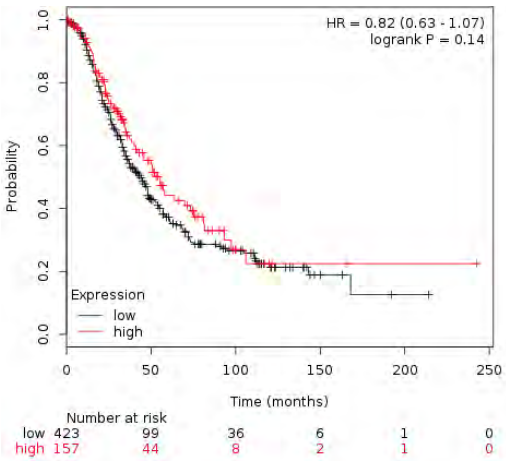


Download plot as a PDF

Normal vs cancer

Probe set	Tissue	Min	Q1	Median	Q3	Max	P value	Normal n
220143_x_at	Normal	58	201	278	364	766	NA	57
220143_x_at	Cancer	57	222	312	433	1558	0.0876	NA

b) LARP4

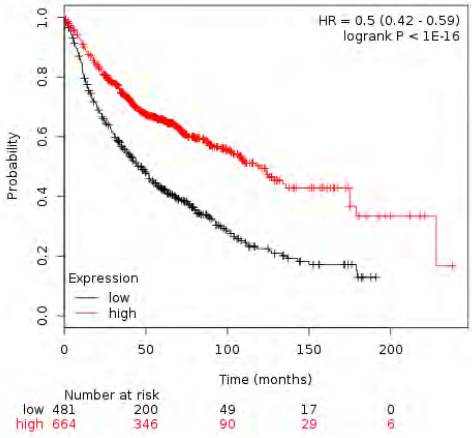


Download plot as a PDF

Normal vs cancer

Probe set	Tissue	Min	Q1	Median	Q3	Max	P value	Normal n
238960_s_at	Normal	52	120	177	177	195	NA	5
238960_s_at	Cancer	4	67	132	202	875	0.703	NA

c) CPEB2



Download plot as a PDF

Normal vs cancer

Probe set	Tissue	Min	Q1	Median	Q3	Max	P value	Normal n
226939_at	Normal	603	1068	1400	2047	9765	NA	86
226939_at	Cancer	35	499	759	1176	5828	2.03E-21	NA

Figure 22: **The results of KM-Plotter for three different genes.** a. High expression of LUC7L gene has a negative effect on cancer patients' outcome. b. High expression of LARP4 gene has a zero effect on cancer patients' outcome. c. High expression of CPEB2 has a positive effect on cancer patients' outcome.

For each probe of the genes in the compiled list, data are collected in a different row of an excel file. Each row of the excel file contains the following data:

- **Tissue**
  The KM-Plotter results derive from microarray data of four different types of cancer (lung, gastric, ovarian or breast)

- **Number of probes**
  The selected gene is represented by one or more probes. Thus, there could be more than one plots (and more than one lines in the excel file) for a gene. Although they correspond to the same gene the collected data of different probes have different values.

- **Effect on cancer patients outcome**
  A highly expressed gene in cancer could have a negative (Figure 22a), a zero (Figure 22b) or a positive effect (Figure 22c) on cancer patients' survival. This effect is explained by the red and black lines in each plot. The black line depicts the survival rates for patients with low expression of the specific gene (respectively, the red one for patients with high gene expression). The effect type was assigned arbitrarily by eye, observing the results of KM-Plotter. In case there are intersections between the red and black line after five months from the beginning of the study, the effect is characterized as **zero** (Figure 22b). In case the red line is above the black one (after the five month's period) patients with high gene expression in this cancer type have better survival rates and the effect is **positive** (Figure 22c). Otherwise, if the red line is below (Figure 22a) the effect is **negative**, meaning worse survival rates for cancer patients with a high expression of this specific gene.

- **Logrank P-value**
  In Figures 22a, 22b and 22c, inside the box of the plots there is a "logrank P-value". The KM-Plotter developers used the "survival" package and the "survplot" function in R (statistical language) to plot the Kaplan-Meier curve and calculate the logrank P-value. This value assesses if red and black curves differ significantly. A low P-value means rejection of the null hypothesis that there is no survival difference amongst these two groups of cancer patients with high or low gene expression.

- **Expression in normal tissue**
  In Figure 22, beyond the survival plots, there is information about the expression in normal and cancerous tissues (for this specific probe set). In the column

"Ex.Norm" of the excel file is collected the median number of transcripts derived from the microarray data (in normal tissues).

- **Expression in cancerous tissue**
  Similarly with expression in normal tissue, the median expression (median number of transcripts) for all patients with tumourous tissues is collected in the column "Ex. Tum" of the excel file.

- **Expression P-value**
  In the column "Ex. P-value" (Figure 23) is collected a value which assesses if the expressions in normal and the expression in cancerous tissues differ significantly. A low P-value means rejection of the null hypothesis that there is no difference between them.

	A	B	C	D	E	F	G	H
1	RBP	Tissue	# probes	Risk	KM p-value	Ex. Norm	Ex. Tum	Ex. P-value
2	A1CF	gastric	2	neg	6.30E-04	333	451	4.77E-02
3	A1CF	gastric	2	neg	5.40E-02	81	83	8.86E-01
4	A1CF	lung	2	neg	1.00E-04	109	73	6.86E-01
5	A1CF	lung	2	neg	6.05E-02	16	13	3.63E-01
6	A1CF	ovarian	1	pos	1.40E-03	87	89	5.51E-01
7	A1CF	breast	1	zero	1.10E-01	14	16	2.12E-01
8	AATF	gastric	1	neg	1.47E-02	673	1199	7.50E-08
9	AATF	lung	1	neg	1.00E-03	729	905	1.68E-11
10	AATF	ovarian	1	zero	6.86E-01	701	777	4.44E-01
11	ABCF1	lung	1	neg	4.10E-03	1417	1388	3.85E-01
12	ABCF1	ovarian	1	zero	1.54E-01	798	1418	8.89E-03
13	ABT1	lung	1	neg	1.00E-04	231	482	1.88E-31
14	ABT1	gastric	1	pos	1.00E-03	257	401	9.72E-11
15	ABT1	ovarian	1	pos	1.80E-03	517	417	3.67E-01
16	ABT1	breast	1	zero	7.50E-02	280	412	2.06E-11
17	ACAA2	gastric	2	neg	6.60E-04	113	75	1.19E-04
18	ACAA2	lung	2	neg	5.10E-06	40	58	6.82E-04
19	ACAA2	ovarian	2	neg	4.30E-04	517	714	1.13E-01
20	ACAA2	gastric	2	pos	8.00E-10	892	1204	2.23E-06
21	ACAA2	lung	2	pos	1.10E-07	961	675	8.69E-11
22	ACAA2	ovarian	2	zero	2.74E-02	100	44	3.44E-02

Figure 23: **The format of the excel file with all collected data.**

Data are collected in an excel file that has the format of Figure 23. In total there are the following eight columns:

A. The name of the “gene-RNA binding protein”
B. The name of the specific KM-Plotter used to collect the data of this line. The different KM-Plotters are for lung, gastric, ovarian and breast cancer patients
C. The number of probes (there are genes with multiple probes available)
D. The risk (high gene expression could have a negative, a positive or a zero effect on cancer patients outcome)
E. The corresponding (risk) p-value.
F. The medium expression level in normal tissue
G. The medium expression level in cancerous tissue
H. The corresponding (expression) p-value

Python Scripts

In this thesis the terms “gene” and “RNA-Binding Proteins” (or “RBPs”) are used equivalently, to describe an RBP from the compiled list. In the following paragraphs, are presented some basic scripts written in Python (programming language) which are vital for the analysis and retrieving of basic information about the collected data.

Unique genes in a list of genes

In many parts of the analysis is required to detect from a list of genes how many of those are unique. In KM-Plotter online tool, there are genes corresponding to multiple probes. Each probe is represented by a Kaplan-Meier plot, with the available data collected in a different line of an excel file. Script 1, detects the genes in a given list and returns a set in which each gene exists just one time.

```
l=open("C:/Users/~/All Instances.txt","r")
k=l.readline()
genes=[]
while(k!=''):
    genes.append(k)
    k=l.readline()

l.close()

different_genes=[]
for i in genes:
    if i not in different_genes:
        different_genes.append(i)



g=open("C:/Users/~/Different RBPs.txt","w")

for i in different_genes:
    g.write(i)

g.close()
```

Script 1: **Detection of the unique genes in a list of genes.** The script reads one by one the lines of a txt file which contains a set of genes. When a gene is detected for the first time, this gene name will be added to a new list. At the end of the program, this list is written in a new txt file (with the name: Different RBPs.txt).

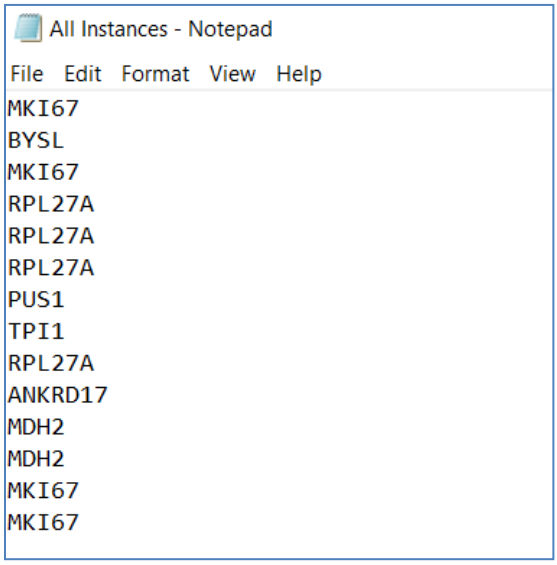


Figure 24: **The txt file with the genes which is given as an input to Script 1.** In total there are 14 genes in this list. The output of the script for this gene list is depicted in Figure 25.

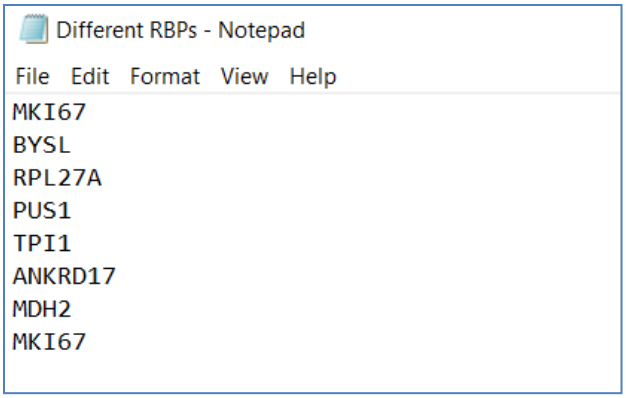


Figure 25: **The txt file with the unique genes derived from a set of genes.** There are 8 unique genes in total derived from the list of genes in Figure 21. These genes are the output of Script 1.

Script 1 takes as an input a txt file with the name "**All Instances.txt**" (Figure 24). The file consists of a list with genes and is saved in the following computer's path: **"C:/Users/~/".** In fact, this path is extended but for sake of brevity is used the symbol "**~**" to shorten the path's length. The outcome of the program is a restricted list in which each gene exists only once. The list is saved under the name "**Different RBPs.txt"**, in the same path with the input file (Figure 25). The input list of genes contains in total **14** genes, but as the output list indicates, the different (unique) genes are only **8**.

Mutual genes in two lists

The Python Script of this paragraph detects for 2 lists of genes, the gene names that exist in both lists. For instance, given that two lists of genes have a special role in lung and gastric cancer respectively, an obvious question is to detect the mutual genes in both cancer types.

```
l=open("C:/Users/~/list1.txt","r")
k=l.readline()
list1=[]
while(k!=''):
    list1.append(k)
    k=l.readline()

l.close()


l=open("C:/Users/~/list2.txt","r")
k=l.readline()
list2=[]
while(k!=''):
    list2.append(k)
    k=l.readline()

l.close()



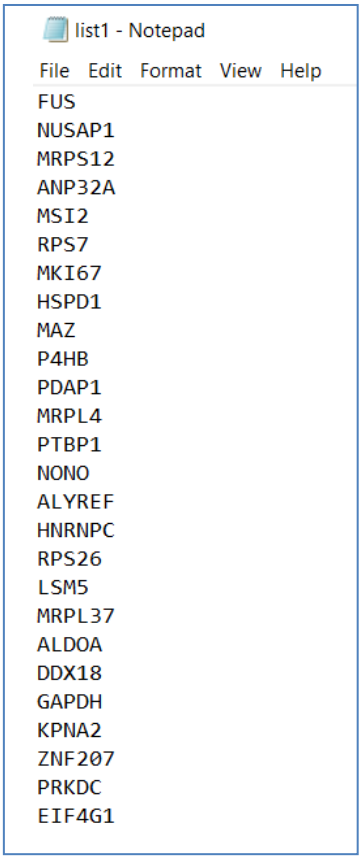
common_genes=[]

for i in list1:
    for j in list2:
        if i==j and i not in common_genes:
            common_genes.append(i)

w=open("C:/Users/~/common_genes.txt","w")
w.write('There are ' + str(len(list1)) + ' genes in the List 1\n')
w.write('There are ' + str(len(list2)) + ' genes in the List 2\n')
w.write('_____\n')
w.write('The common genes are:\n')
for i in common_genes:
    w.write(i)
```

Script 2: **Detection of common genes between two gene lists.** The script reads one by one the lines of two txt files and when it finds a gene that exists in both files, it writes it in a new list (if it is not already written to that list). The genes of the new list are written in a txt file ("common_genes.txt").

Script 2 takes as an input two lists of genes, saved in the files "**list1.txt**" and "**list2.txt**". In the example files in Figure 26, the two input lists consist of 26 and 14 genes respectively. The output of Script 2 is the list of four mutual genes shown in Figure 27, with the file name "**common_genes.txt**", saved in the same path with the input files.



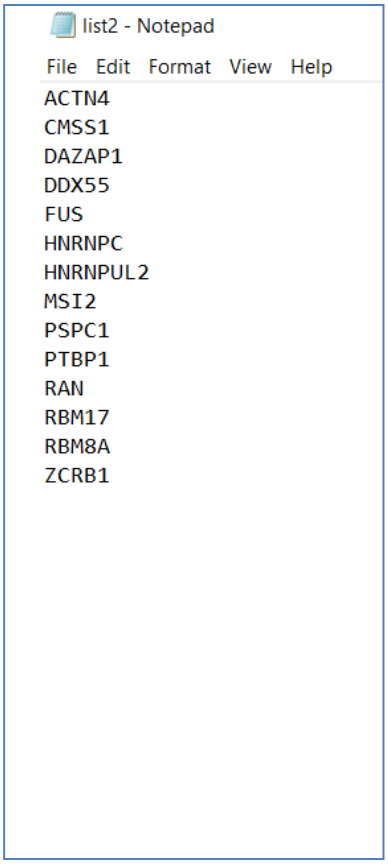


Figure 26: **Two txt files with the compared lists.** The first txt file is called "list1.txt" and contains 26 genes in total. The second one is the file "list2.txt" and contains 14 genes in total. The mutual genes are given in the file "common_genes.txt" in Figure 27.

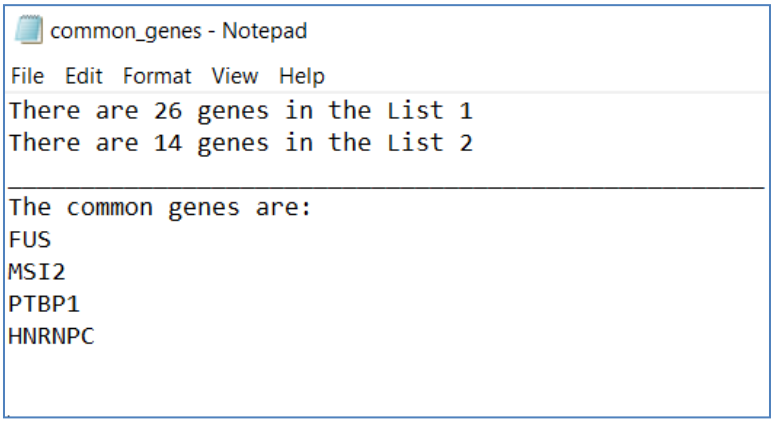


Figure 27: **Information about the mutual genes between two gene lists.** Information is about the total number of genes in each list, as well as the names of the mutual (common) genes among the two lists.

Compare genes between 2 lists according to their effect in each cancer type

This script is similar to the previous one. However, apart from the detection of mutual genes between two cancer types this script is based on the comparison among their role in both cancer types.

In Figure 28 is shown part of a list with genes having a special role in lung cancer (the file is called "**lung.txt**"). This role is indicated in the second column and refers to the positive (pos) or negative (neg) cancer patients' outcome when the gene is overexpressed. Similarly, the file "**gastric.txt**" contains another list for the respective genes in gastric cancer.

Script's 3 inputs are the lung and gastric lists and the outcome are the calculations in Figure 29. The possible outcomes for the effect on lung and gastric cancer patients are the following:

1. Negative effect in lung – Negative effect in gastric
2. Negative effect in lung – Positive effect in gastric
3. Positive effect in lung – Negative effect in gastric
4. Positive effect in lung – Positive effect in gastric

ACTN4, MAZ, PPIA, PTBP1 and RPL35A have a negative effect on cancer patients' outcome in both cancer types, while BCLAF1, G3BP2, MATR3 and PPIG have a positive effect. HNRNPC has a negative effect on lung cancer patients and a positive effect on gastric cancer patients. ADD1, CSRP1, CTNNA1, DDX17, MAP4 and RBMS3 have a positive effect on lung and a negative on gastric cancer patients' outcome (Figure 29).

```
l=open("C:/Users/~/lung.txt","r")
k=l.readline()
lung=[]
while(k!=''):
    p=k.split('\t')
    p[1]=p[1][0:3]
    lung.append(p)
    k=l.readline()

l.close()


g=open("C:/Users/~/gastric.txt","r")
k=g.readline()
gastric=[]
while(k!=''):
    p=k.split('\t')
    p[1]=p[1][0:3]
    gastric.append(p)
    k=g.readline()

g.close()


neg_neg=[] #negative lung - negative gastric
neg_pos=[] #negative lung - positive gastric
pos_neg=[] #positive lung - negative gastric
pos_pos=[] #positive lung - positive gastric


for i in gastric:
    for j in lung:
        if i[0]==j[0]:
            if (i[1]==j[1]=='neg'):
                neg_neg.append(i[0])
            elif i[1]=='pos' and j[1]=='neg':
                neg_pos.append(i[0])
            elif i[1]=='neg' and j[1]=='pos':
                pos_neg.append(i[0])
            elif (i[1]== j[1]=='pos'):
                pos_pos.append(i[0])

print ('\nNegative lung - Negative gastric')
for i in neg_neg:
    print(i)

print ('\nNegative lung - Positive gastric')
for i in neg_pos:
    print(i)

print ('\nPositive lung - Negative gastric')
for i in pos_neg:
    print(i)

print ('\nPositive lung - Positive gastric')
for i in pos_pos:
    print(i)
```

Script 3: **Comparison of two gene lists according to the effect on lung and gastric cancer patients' outcome.** The high gene expression could have a positive or negative effect. The script takes as an input two txt files which contain the gene names and the effect type. The results are depicted in Figure 29.

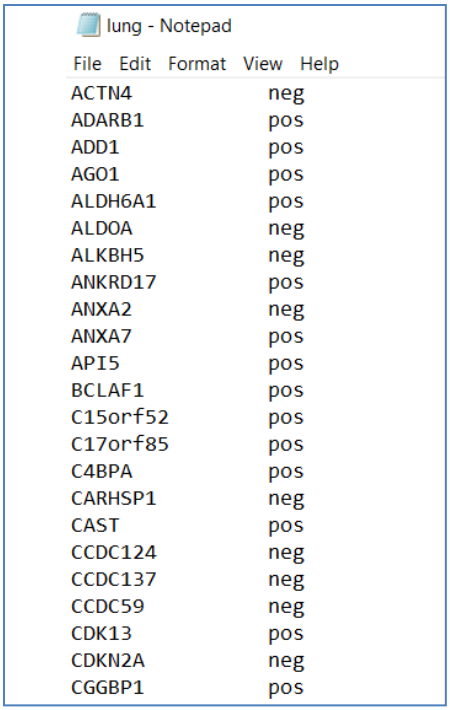


Figure 28: **Part of the list with the lung genes.** The file is composed of 2 columns. In the first column are the names of the genes and in the second one, the effect of high gene's expression on cancer patients' outcome (negative or positive). Similar is the list with the gastric genes.

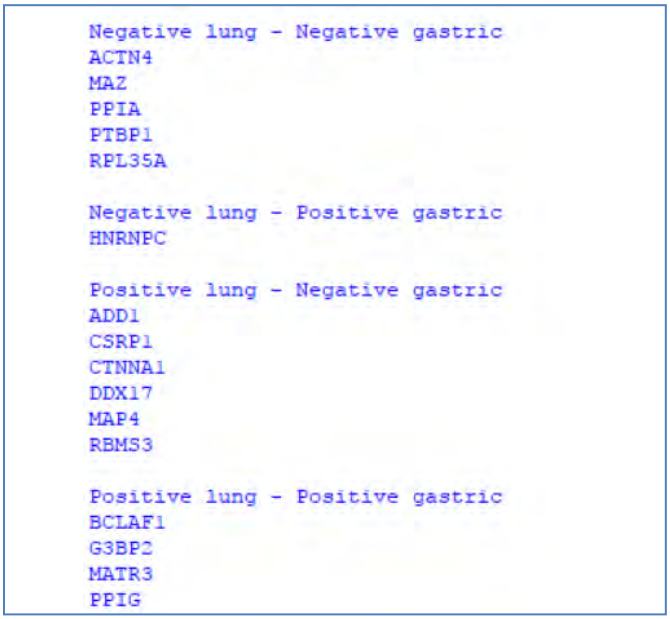


Figure 29: **The mutual genes between lung and gastric lists.** The results are in accordance to the effect on cancer patients' outcome (for each cancer type). There are genes with a negative or a positive effect when overexpressed.

Different trends of the same gene

According to the results of KM-Plotter online tool, there is a concordance of multiple available probes for most genes. Thus, many genes in the collected data correspond to more than one instance. The high or low gene expression is connected with a positive, a negative or a zero effect in cancer patients' survival. This effect is indicated by each instance separately. While some genes present the same trend for all their instances, others present extremely different trends. Script 4 detects the gene instances which show similar or different trends.

```
l=open("C:/Users/~/Data text.txt","r")
k=l.readline()
instances=[]
effect=[]
while(k!=''):
    p=k.split('\t')
    instances.append(p[0])
    effect.append(p[5])
    k=l.readline()

l.close()

genes=[]
neg=[]
pos=[]
total=[]


for i in range(len(instances)):
    if instances[i] not in genes:
        genes.append(instances[i])
        neg.append(int(0))
        pos.append(int(0))
        total.append(int(0))
    if effect[i]=='neg':
        neg[-1]+=1
    elif effect[i]=='pos':
        pos[-1]+=1
    total[-1]+=1

bad_genes=[]
bad_neg=[]
bad_pos=[]
for i in range(len(genes)):
    if neg[i]!=0 and pos[i]!=0:
        bad_genes.append(genes[i])
        bad_neg.append(neg[i])
        bad_pos.append(pos[i])
```

```
w=open("C:/Users/~/Information.txt","w")

w.write('Negative Genes\tNeg\tPos\tTotal\n--------------------\n')
for i in range(len(genes)):
    if neg[i]>0 and pos[i]==0:
        if len(genes[i])>7:
            w.write(str(genes[i]) + '\t' + str(neg[i]) + \
                    '\t' + str(pos[i]) + '\t' + str(total[i]) + '\n')
        else:
            w.write(str(genes[i]) + '\t\t' + str(neg[i]) +\
                    '\t' + str(pos[i]) + '\t' + str(total[i]) + '\n')



w.write('\n\n\n\nPositive Genes\tNeg\tPos\tTotal\n-----------\n')
for i in range(len(genes)):
    if neg[i]==0 and pos[i]>0:
        if len(genes[i])>7:
            w.write(str(genes[i]) + '\t' + str(neg[i]) + '\t' +\
                    str(pos[i]) + '\t' + str(total[i]) + '\n')
        else:
            w.write(str(genes[i]) + '\t\t' + str(neg[i]) + '\t' +\
                    str(pos[i]) + '\t' + str(total[i]) + '\n')



w.write('\n\n\n\n"Bad Genes"\tNeg\tPos\tTotal\n--------------\n')
for i in range(len(bad_genes)):
    if len(bad_genes[i])>7:
        w.write(str(bad_genes[i]) + '\t' + str(bad_neg[i]) + '\t' +\
                str(bad_pos[i]) + '\t' + str(bad_neg[i]+ bad_pos[i]) + '\n')
    else:
        w.write(str(bad_genes[i]) + '\t\t' + str(bad_neg[i]) + '\t' +\
                str(bad_pos[i]) + '\t' + str(bad_neg[i]+ bad_pos[i]) + '\n')
```

Script 4: **Characterization of the different trends for multiple instances of the same gene.** There are gene instances showing the same trends and others with extremely different trends.

	A	B	C	D	E	F	G	H
1	RBP	Tissue	# probes	Risk	KM p-value	Ex. Norm	Ex. Tum	Ex. P-value
2	NUSAP1	lung	2	neg	1.00E-16	57	350	2.66E-21
3	NUSAP1	lung	2	neg	1.00E-16	405	1058	1.58E-09
4	FUS	lung	3	neg	1.10E-16	45	460	4.15E-13
5	FUS	lung	3	neg	6.40E-03	158	88	3.88E-13
6	FUS	lung	3	pos	1.60E-07	2401	1898	1.76E-05
7	MRPS12	lung	4	neg	4.00E-15	38	108	2.27E-16
8	MRPS12	lung	4	neg	7.20E-14	209	1201	2.49E-24
9	MRPS12	lung	4	neg	5.90E-13	47	81	1.48E-14
10	MRPS12	lung	4	neg	5.20E-06	79	61	1.50E-01
11	ANP32A	lung	2	neg	8.40E-13	76	391	4.90E-13
12	ANP32A	lung	2	zero	9.60E-02	17	12	1.04E-02
13	MSI2	lung	8	neg	8.70E-15	101	377	7.26E-25
14	MSI2	lung	8	neg	7.50E-09	121	210	1.51E-15
15	MSI2	lung	8	neg	8.70E-09	40	88	1.35E-15
16	MSI2	lung	8	neg	3.50E-06	66	130	5.83E-26
17	MSI2	lung	8	neg	8.20E-06	29	36	4.81E-02
18	MSI2	lung	8	neg	1.30E-03	115	121	2.14E-01
19	MSI2	lung	8	neg	4.30E-03	576	1034	1.53E-14
20	MSI2	lung	8	pos	9.50E-04	477	862	5.26E-16

Figure 30: **Part of the excel file with the collected data.** This file contains 26 different genes and 71 instances in total and is given as an input to Script 4.

Figure 30 shows part of the excel file with the collected data. The file contains 71 gene instances from 26 different genes (only 19 instances are shown in the figure), which are given as an input to Script 4. The output is a file with the genes separated into the following three categories:

1. Negative Genes: all the gene instances have a negative effect
2. Positive Genes: all the gene instances have a positive effect
3. "Bad Genes": some of the gene instances have a negative effect, while others have a positive effect on cancer patients' outcome.

Instances with a zero effect are ignored. The results after the execution of Script 4 for the input data (shown in Figure 30) are given in Figure 31:

Negative Genes	Neg	Pos	Total
NUSAP1	2	0	2
MRPS12	4	0	4
ANP32A	1	0	1
RPS7	2	0	2
MKI67	4	0	4
HSPD1	1	0	1
MAZ	2	0	2
P4HB	1	0	1
PDAP1	2	0	2
MRPL4	3	0	3
PTBP1	4	0	4
ALYREF	2	0	2
RPS26	1	0	1
MRPL37	1	0	1
ALDOA	2	0	2
GAPDH	3	0	3
KPNA2	2	0	2
ZNF207	1	0	1

Positive Genes	Neg	Pos	Total

"Bad Genes"	Neg	Pos	Total
FUS	2	1	3
MSI2	7	1	8
NONO	2	1	3
HNRNPC	4	2	6
LSM5	1	2	3
DDX18	2	2	4
PRKDC	1	1	2
EIF4G1	1	1	2

Figure 31: **Classification of the genes according to the trends followed by their gene instances.** Genes are classified into 3 different categories, depending on the effect of their instances on cancer patients' outcome.

There are 18 genes in total that fall into the first category (all instances express the same trend). The number of negative instances for the same gene range between 1 and 4 (shown at the left part of Figure 31). The right part of the figure shows that there are no "Positive Genes", while there are 8 genes that have both positive and negative

instances. Even though only one instance is different and all the others follow the same trend, the gene will be classified into the last category. For example, gene MSI2 has 7 negative instances and only 1 positive, therefore is characterized as “bad gene”.

One aspect for the development of this script that could be important is the statistical significance (indicated in column E of the excel file in Figure 30) with a higher p-value for an instance meaning a lower statistical significance. Independently of a high or low p-value for an instance, in the current script the contribution to the final result is the same.

Collected data statistics

Calculation of statistics about the data is vital in many parts of the analysis. A Python Script takes as an input a txt file (similar to the file depicted in Figure 30) and generates information about the data. The output is printed on the screen as well as in a different txt file which contains all important information.

For an input file with all the collected data, the results after the script execution are given in Figure 32. For sake of brevity, information about the breast cancer data are not displayed in the figure, but their form is similar to the other cancer types. In the left part, is displayed the number of collected data (all instances) in total and separately for each cancer type and the number of positive, negative and zero instances. The total number of instances is 7036, with 1996 lung instances, 1963 gastric instances and 1998 ovarian instances. There are 2510, 2633 and 1893 gene instances respectively, expressing a negative, positive and zero effect on cancer patients’ outcome.

The 7036 gene instances correspond to 993 different genes, due to the multiple probes for many genes in KM-plotter online tool. The Python script detects the number of different genes in each cancer type as well as the number of common (mutual) genes between different cancer types. For sake of brevity only part of the list with the different genes in lung and gastric is displayed. In the right part of Figure 32 is shown the list of genes in ovarian cancer and the lists with the common genes between different cancer types (lung-gastric, lung-ovarian and gastric-ovarian). Information about breast cancer is not displayed and common genes between lung-breast, gastric-breast and ovarian-breast are not given in Figure 32. The Python script is given by the code in the Appendix.

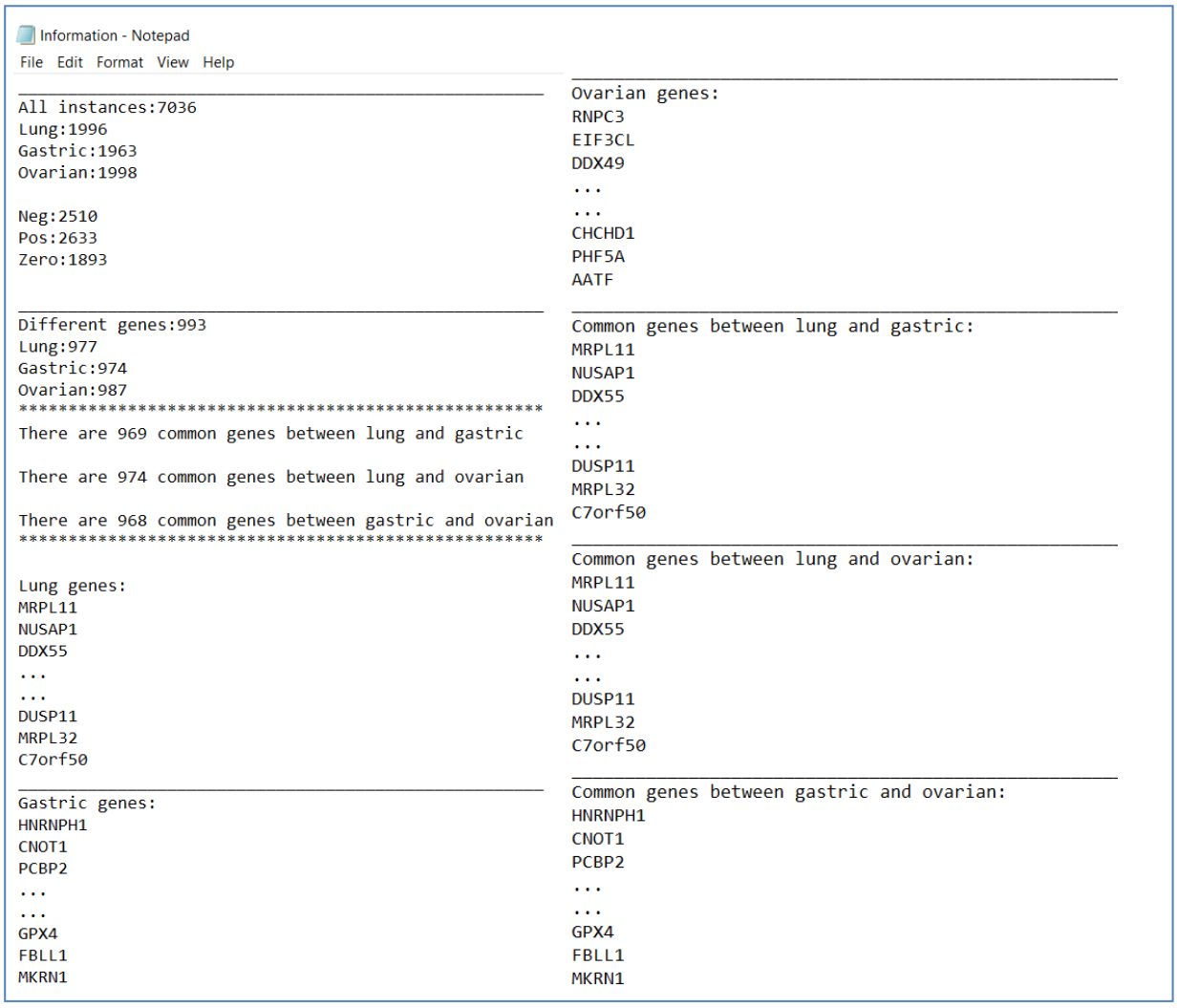


Figure 32: **Collected data statistics.** The Python Script takes as an input a txt file with the collected data and generates information and statistics about them.

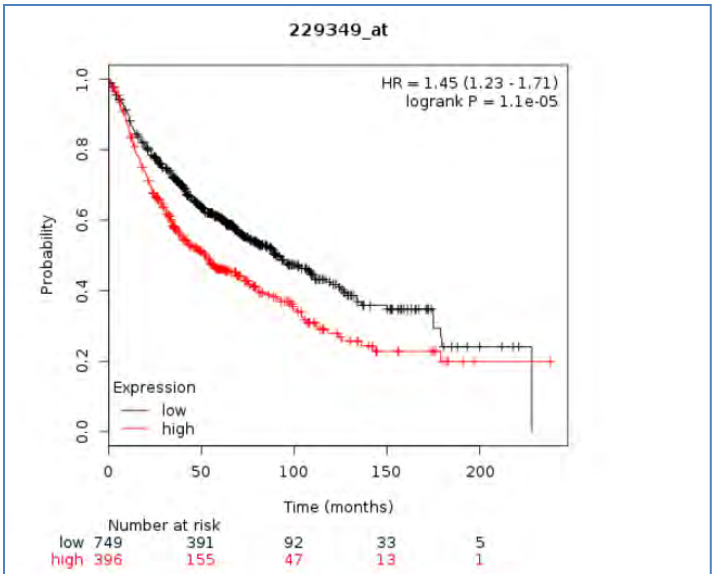
Chapter 4: Results

A list of human RNA Binding Proteins (RBPs) was compiled by the Lewis laboratory based on available datasets of RNA binding proteins encoded by the human genome and contains 1334 different candidates in total. The purpose of this thesis is to reduce the total number of candidates that will be taken forward to drug screening for novel anti-tumour compounds. The reduction will be based on the results of a computational analysis.

KM-Plotter for data collection

KM-Plotter is an online tool which assesses the effect of high gene expression on cancer patients' survival (Gyorffy et al., 2013). Using KM-Plotter all available information about the RBPs was collected from patients suffering from lung, gastric, ovarian or breast cancer. Four KM-Plotter environments correspond to each cancer type respectively. The datasets which constitute KM-Plotter derive from caBIG (Cancer Biomedical Informatics; Cimino et al., 2007), TCGA (The Cancer Genome Atlas; Tomczak et al., 2015) and GEO (Gene Expression Omnibus; Clough et al., 2016). A gene name from the compiled list, is given as an input in KM-Plotter's environment of each cancer type, with the same procedure repeated for all genes of the compiled list (gene names and RBP names from the compiled list are used equivalently). The output is different Kaplan-Meier plots with available data on the expression levels of the corresponding genes and cancer survival data.

Datasets retrieved by GEO include only Affymetrix microarray datasets (Gyorffy et al., 2010, 2012, 2013; Szasz et al., 2016). In Affymetrix microarrays, multiple probes are available for a gene, leading to multiple plots for each gene in KM-Plotter results. For each probe of the genes in the compiled list, a plethora of data are collected in a different row of an excel file. The collected data are the tissue name, the number of probes available for each gene, the effect of high gene expression on cancer patients' survival and the expression levels of each RBP in normal and cancerous tissues (with the corresponding statistical significances).



Download plot as a PDF

Normal vs cancer

Probe set	Tissue	Min	Q1	Median	Q3	Max	P value	Sample n
229349_at	Normal	2	10	35	59	122	NA	86
229349_at	Cancer	1	10	20	45	3563	0.0527	2437

Figure 33: **KM-Plotter results for LIN28B gene in lung cancer environment.** Among LIN28B results was only one plot because there is only one probe for this gene in KM-Plotter. The red and black lines depict patient's survival when the gene is over and under-expressed respectively. The red line is below the black one, showing a negative effect on lung cancer patients' survival when LIN28B is over-expressed.

The **tissue** name could be lung, gastric, ovarian or breast, as the same procedure is repeated for each one of the four KM-Plotter environments. A specific probe's data (saved in a line of the excel file) may derive from one of the four cancer types. The **number of probes** describes the number of different plots for a specific gene in KM-Plotter, as a gene could be represented by multiple probes.

Among the output data are the **effect of high gene expression (Risk)** and the **statistical significance (KM p-value or Risk p-value)** of high versus low-level expression of each RBP in cancer patients' overall survival. A High gene expression could have a negative, a positive or a zero effect on cancer patients' survival. During the collection of the data, the effect type was assigned arbitrarily by eye, observing the results of KM-Plotter. In case there are intersections between high and low gene expression curves in Kaplan-Meier plots after five months from the beginning of the study, the effect is characterized as zero. Intersections after 50 months are ignored, with a possible positive or negative effect despite the intersections. In case high gene expression's plot is above low gene expression's plot (after the five month's period) the

effect is characterized as positive while in the last case of being below, there is a negative effect on cancer patients' outcome. A low P-value means rejection of the null hypothesis that there is no survival difference amongst these two groups of cancer patients with high or low gene expression.

Collected information on the **expression levels of each RBP in normal (Ex. Norm) and cancerous (Ex. Tum) tissues** contain the median number of transcripts derived from the microarray data in normal and cancerous tissues respectively. A **p-value (Ex. p-value)** on this difference in the same four cancer datasets assesses if the expression in normal and the expression in cancerous tissues differ significantly. A low P-value means rejection of the null hypothesis that there is no difference among them.

	A	B	C	D	E	F	G	H
1	RBP	Tissue	# probes	Risk	KM p-value	Ex. Norm	Ex. Tum	Ex. P-value
2	LARP1B	breast	5	neg	1.60E-01	74	104	1.74E-04
3	LARP1B	breast	5	pos	5.70E-04	103	169	1.16E-12
4	LARP1B	breast	5	pos	2.40E-03	112	169	2.23E-05
5	LARP1B	breast	5	zero	1.30E-02	61	86	2.13E-03
6	LARP1B	breast	5	zero	9.90E-02	270	261	2.87E-01
7	LARP1B	gastric	5	neg	8.50E-05	140	72	6.11E-07
8	LARP1B	gastric	5	pos	5.40E-08	340	436	4.16E-05
9	LARP1B	gastric	5	pos	2.40E-05	308	169	2.55E-07
10	LARP1B	gastric	5	pos	9.20E-04	101	87	3.25E-02
11	LARP1B	gastric	5	pos	9.50E-04	279	182	3.53E-07
12	LARP1B	lung	5	pos	1.10E-16	269	110	9.18E-08
13	LARP1B	lung	5	pos	1.10E-10	163	135	7.07E-03
14	LARP1B	lung	5	pos	8.40E-06	430	307	3.89E-06
15	LARP1B	lung	5	pos	6.80E-03	111	68	2.30E-05
16	LARP1B	lung	5	pos	8.30E-02	178	65	3.96E-12
17	LARP1B	ovarian	5	neg	2.00E-02	165	87	7.96E-02
18	LARP1B	ovarian	5	pos	1.10E-03	210	158	1.49E-01
19	LARP1B	ovarian	5	pos	8.10E-02	204	230	5.98E-01
20	LARP1B	ovarian	5	zero	3.00E-01	138	160	9.58E-01
21	LARP1B	ovarian	5	zero	4.10E-01	77	68	6.50E-01
22	LIN28B	breast	1	zero	2.30E-02	12	17	4.55E-02
23	LIN28B	gastric	1	neg	2.90E-03	37	20	4.76E-03
24	LIN28B	lung	1	neg	1.10E-05	35	20	5.27E-02
25	LIN28B	ovarian	1	neg	2.40E-07	31	27	9.86E-01

Figure 34: **Part of the excel file with all collected data.** LARP1B had 5 probes in KM-Plotter results for each one of the four cancer environments (lung, gastric, ovarian, breast). The total number of collected instances for this gene was 20. For LIN28B, 4 collected instances correspond to one instance for each cancer environment.

The distribution of the collected gene instances in each cancer and effect type

KM-Plotter's data derive from Affymetrix microarrays, with a concordance of multiple probes available for the same gene or a probe set ID mapping to several genes. From the total number of 1334 RBPs, 993 were detected in KM-Plotter online tool (with one or more probes corresponding to each gene), contributing gene expression and survival information to the collected data. The rest 341 were either undetected by the tool or corresponded to probe sets which mapped to several genes. Due to the multiple probes available for each gene, the total number of collected instances was 7036.

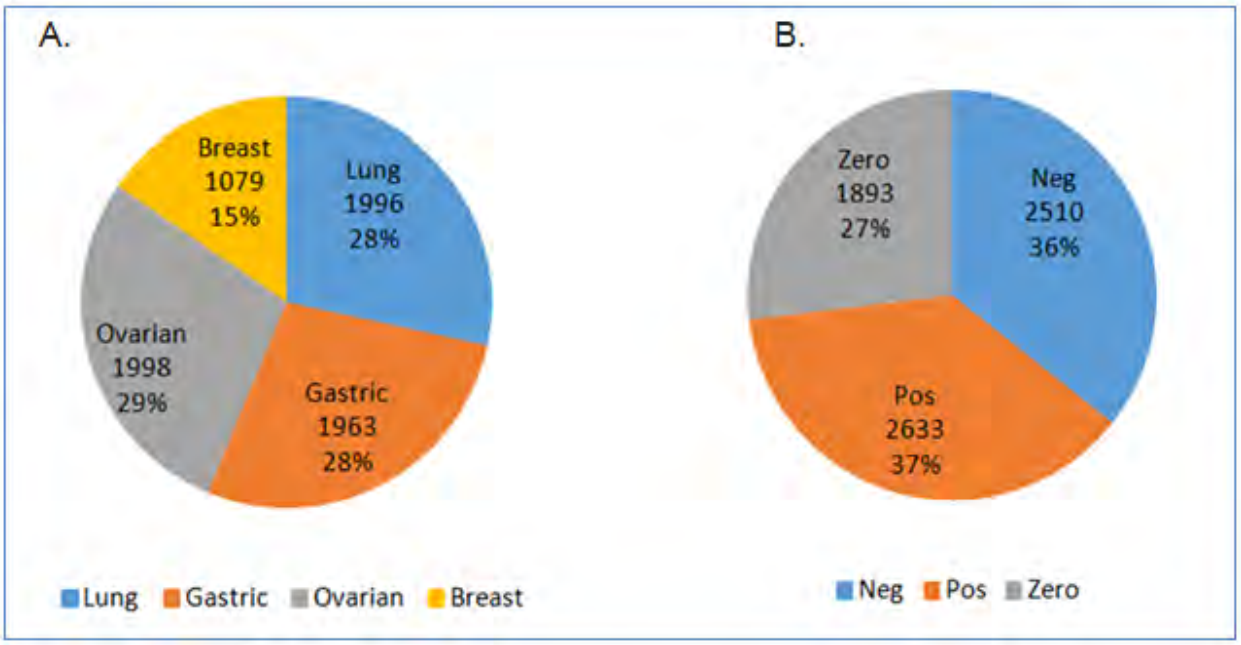


Figure 35: **The distribution of the collected instances in each cancer type and each effect type.** A) In total there are 7036 collected gene instances distributed among lung, gastric, ovarian and breast cancer. The lung, gastric and ovarian instances are the majority compared to the number of breast cancer instances which are less. B) All these gene instances are distributed among negative, positive and zero. Most of them show a negative or a positive trend, with the number of instances which have a zero effect being more restricted.

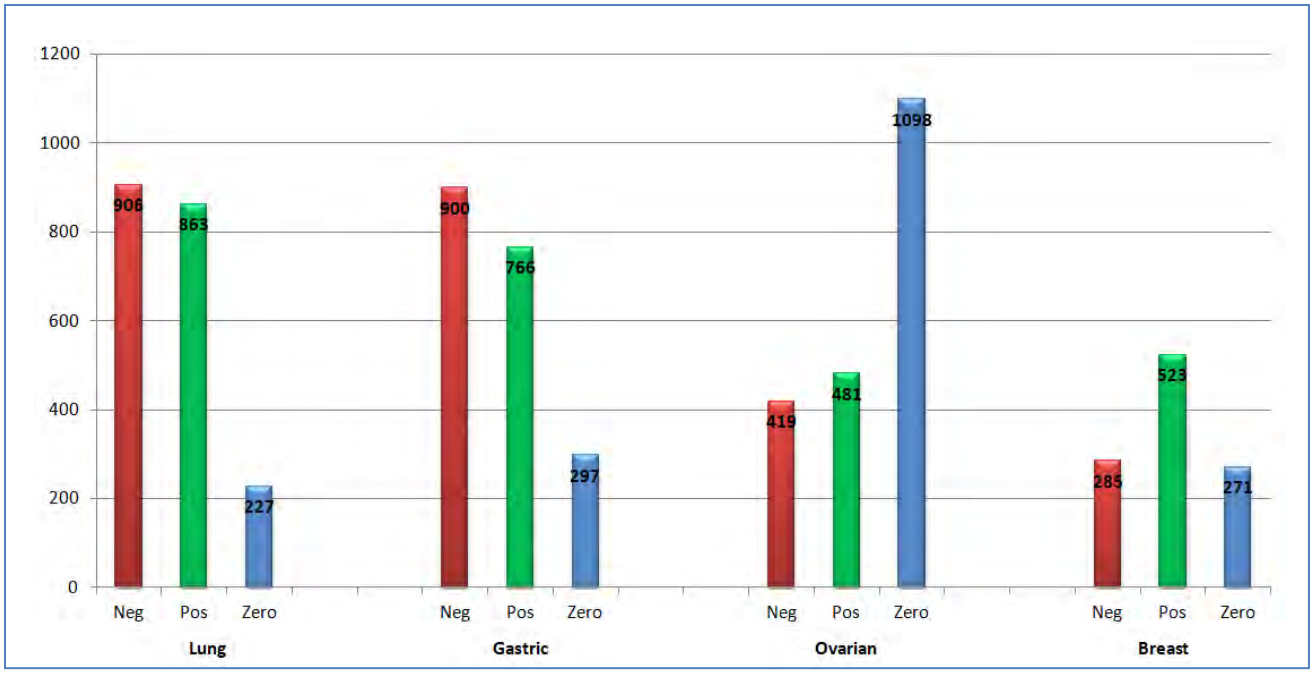


Figure 36: **The distribution of the lung, gastric, ovarian and breast instances, among the three effect types.** In total there are 1996 lung instances, 1963 gastric instances, 1998 ovarian instances and 1079 breast instances which are distributed among the classes of negative, positive and zero.

Figure 35, shows the distribution of the collected gene instances among each cancer (lung, gastric, ovarian, breast) and effect type (negative, positive, zero). In KM-Plotter online tool, there were genes that returned a different number of probes for the different cancer types. This explains the small number of breast instances, compared to the lung, gastric and ovarian instances which are the majority. Specifically they were collected 1996 lung instances, 1963 gastric instances, 1998 ovarian instances and 1079 breast instances.

For the lung and gastric cancers, most instances show a positive or negative trend, while for the ovarian cancer the majority has a zero effect on ovarian cancer patients' outcome. For breast cancer data, most instances show a positive trend (Figure 36). In the next paragraphs, are followed the steps of a computational analysis to reduce the number of data, until only a small number of candidate RBPs remains for drug screening. The distribution of the data in each effect and cancer type, explains why after each step of the computational analysis remain these specific data.

Filtering according to the effect on cancer patients' outcome

Among the collected data is a p-value (called "Risk p-value") which assesses in the Kaplan-Meir plots whether the curves for the two patient groups (with high or low gene expression) differ significantly. A low P-value means rejection of the null hypothesis that there is no survival difference amongst these two groups and expresses the statistical significance of each instance on cancer patients' outcome effect.

The number of total collected gene instances is 7036. In this step of the computational analysis, a filtering according to the effect of high gene expression on cancer patients' outcome (negative, positive or zero effect) reduces the number of instances. Figure 37, shows the distribution of the exponents of the p-values for the data of each cancer type. The most statistically significant data are observed for the lung and gastric data. For the ovarian and breast cancer, the majority of data have p-values greater than 1E-08.

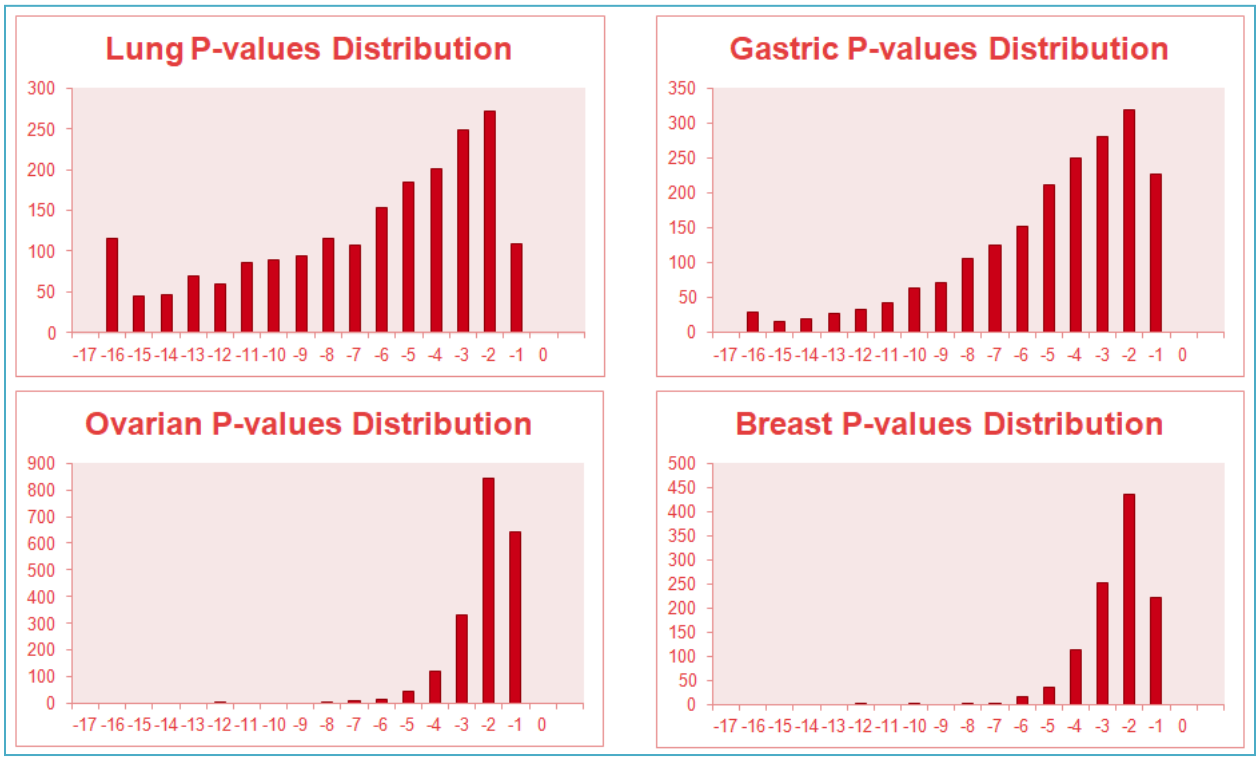


Figure 37: **The distribution of the exponents of the p-values for each cancer's type data.** The x-axis shows the exponents and the y-axis shows the absolute number of data.

The data that will be excluded are all of those with p-values greater than 1E-10. This filtering leaves **591 instances** which represent 8% of the total data. Among them, there

was also a breast instance that (being only one) was unable to contribute information and was excluded too. The most statistically significant data are for lung and gastric cancer data and after this step of the analysis all ovarian and breast cancer instances have been excluded. The total number of remaining instances is **590**.

From the 590 instances, 320 (54%) had a negative and 270 (46%) had a positive effect on cancer patients' outcome (Figure 38). Zero effect instances have been excluded as a high statistical significance means that the curves for the two patient groups (high or low gene expression) in KM-Plots differ significantly.

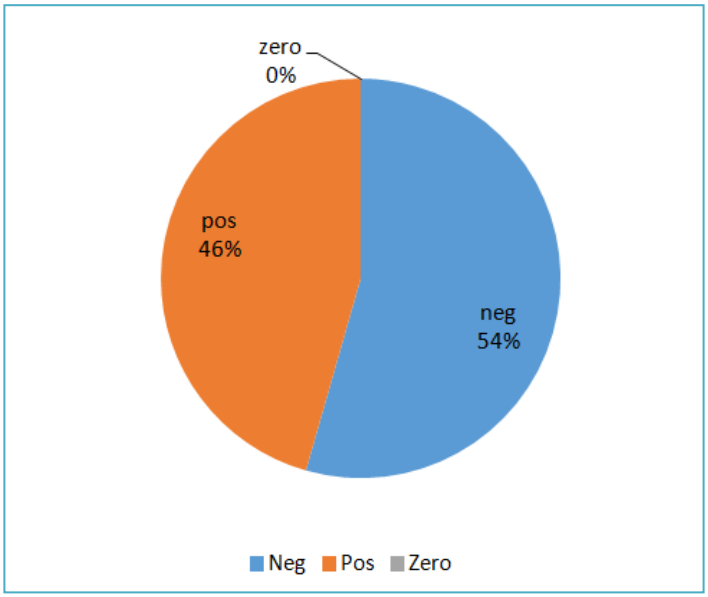


Figure 38: **The distribution of the 590 instances between the classes of negative, positive and zero.** There are 320 (54%) negative instances and 270 (46%) positive instances. Positive instances will be excluded from the next steps of the analysis, as genes with instances showing a negative effect on cancer patients' survival, are estimated as more possible drug targets.

Instances which have a negative effect are more possible drug candidates, thus the positive instances are excluded from the next parts of the analysis. A Python script described in Chapter 3, calculates the distribution of the remaining instances among the classes of negative, positive and zero, as well as the different number of genes that these instances correspond to. From the **320 negative instances**, 212 are lung instances and 108 are gastric instances. The number of different genes is 241, from which 171 are lung genes and 88 are gastric genes (18 genes are mutual between lung and gastric) (Figure 39). This is a high number of genes, meaning that more computational analysis steps should be applied to restrict even more the list of the corresponding RBPs which will be taken to drug screening.

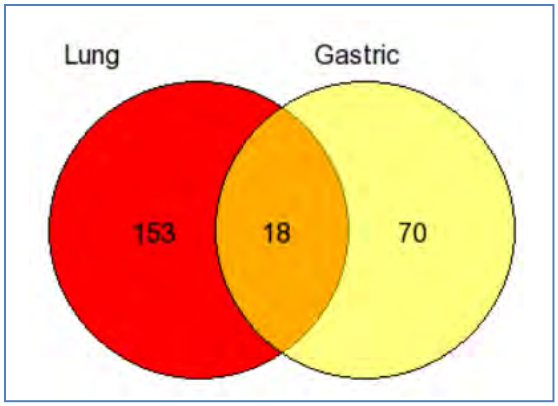


Figure 39: **The distribution of the genes among lung and gastric**. The lung genes are 171 (153+18) and the gastric genes are 88 (70+18). 18 genes are overlapping between lung and gastric.

Filtering according to the different expression in normal and cancerous tissues

Additional collected characteristics are the mean gene expression in cancerous and normal tissues and a respective p-value (called “Ex. P-value”). “Ex. P-value” assesses if the expressions in normal and cancerous tissues differ significantly. A low “Ex. P-value” means rejection of the null hypothesis that there is no difference between them.

At the end of the previous step of the analysis, there were 320 remaining gene instances. The next idea is to reduce even more the number of instances (restricting the list with the RBPs that will be taken forward into drug screening), based on the difference in expression between normal and cancerous tissues.

The excel file which contains the collected data is extended by one more column (Figure 40). Each line of the column has the respective exponent from the “Ex. P-value” field. However, according to the difference in expression between normal and cancerous tissues is used a positive or negative sign. In case of overexpression in cancerous tissues is used the positive sign, while in case of overexpression in normal tissues is used the negative sign. For example, in the first line of this excel file (Figure 40), the “Ex. P-value” is **4.13E-09**. The exponent is **9** and because the expression in normal tissues is greater than the expression in tumourous tissues (**598 > 403**) the sign will be **“minus”**. Thus, in the last column of the excel file is added the value **-9**. As another example, in the second line is added the value 2 (positive value), meaning that the expression in cancer is greater (**1623 > 1019**).

	A	B	C	D	E	F	G	H	I
1	RBP	Tissue	# probes	Risk	KM p-value	Ex. Norm	Ex. Tum	Ex. P-value	Ex. P-value (plus or minus)
2	ACIN1	gastric	1	neg	1.20E-13	598	403	4.13E-09	-9
3	ACTN1	gastric	3	neg	8.80E-12	1019	1623	1.08E-02	2
4	ACTN4	gastric	1	neg	1.00E-16	2881	2349	3.19E-02	-2
5	ACTN4	lung	1	neg	1.40E-13	1843	1923	1.39E-01	1
6	ADARB1	gastric	1	neg	1.20E-12	223	224	1.75E+00	1
7	ADD1	gastric	3	neg	1.10E-16	717	425	2.23E-05	-5
8	ADD1	gastric	3	neg	3.40E-13	1180	1108	4.01E-02	-2
9	AKAP8L	gastric	3	neg	1.40E-12	293	307	2.87E-01	1
10	ALDOA	lung	2	neg	6.70E-16	6167	12076	3.71E-12	12
11	ALDOA	lung	2	neg	7.80E-16	5905	12721	1.27E-13	13
12	ALKBH5	lung	3	neg	5.60E-16	443	542	4.59E-01	1
13	ALYREF	lung	2	neg	2.80E-13	408	1009	2.17E-14	14
14	ANP32A	lung	2	neg	8.40E-13	76	391	4.90E-13	13
15	ANXA2	lung	2	neg	1.60E-15	16693	15402	1.40E-04	-4
16	ANXA2	lung	2	neg	1.90E-15	16432	15910	4.10E-03	-3
17	ARF1	lung	3	neg	5.00E-11	44	52	5.01E-01	1
18	ATP5C1	gastric	2	neg	2.70E-14	303	341	6.16E-01	1
19	ATXN2	gastric	1	neg	3.00E-15	629	434	1.41E-11	-11
20	ATXN2L	gastric	2	neg	1.90E-13	136	97	1.26E-03	-3

Figure 40: **The excel file with the collected data is extended by one more column.** This column contains the respective exponent from the “Ex. P-value” field, using a “plus” sign in case of overexpression in cancerous tissues or a “minus” sign in case of overexpression in normal tissues.

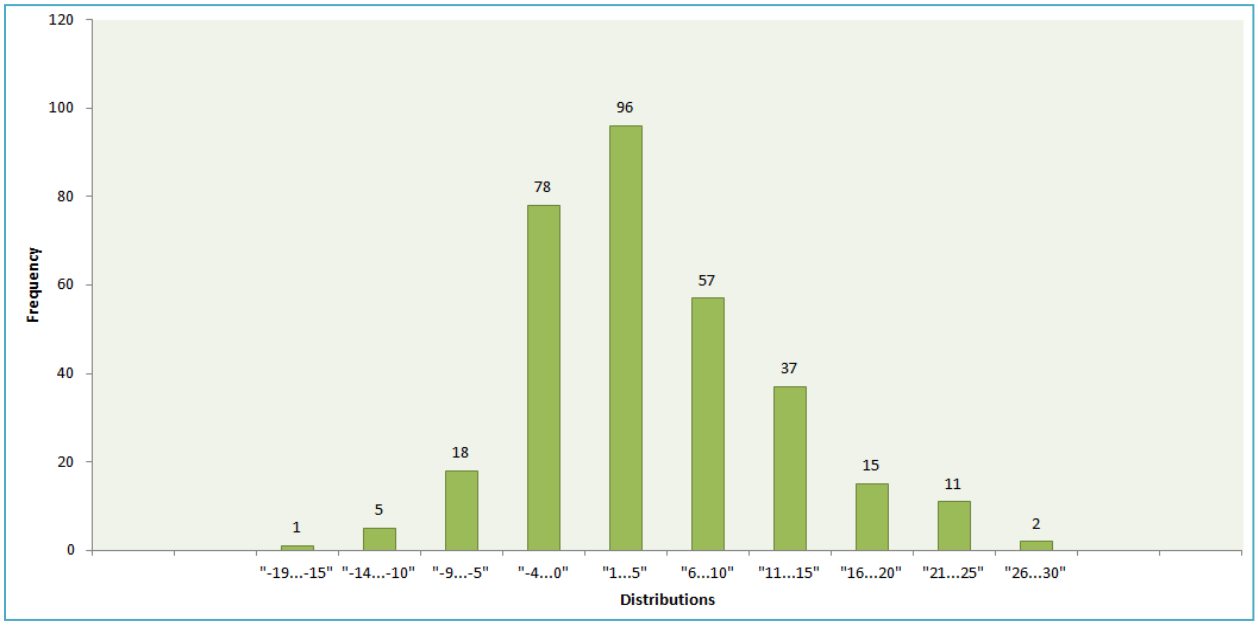


Figure 41: The **distribution for the exponents of the “expression p-values”.** There are 218 gene instances for which the expression p-value exponents are marked with the “plus” sign, meaning that for these instances the expression in tumourous tissues is higher than the expression in normal tissues. Respectively, the 102 exponents marked with the “minus” sign, mean that these gene instances are overexpressed in normal tissues.

The distribution for the values of the last column is shown in Figure 41. One reason that was used the strategy with the “plus-minus trick” is that the graph gives information about the total number of instances that have a higher expression in normal or cancerous tissues respectively. The majority of instances have been distributed to exponents greater than zero, meaning that there are more instances which are overexpressed in cancerous tissues.

The aim of this chapter is to reduce the number of the RBPs and keep only a small number of them which will be taken forward into drug screening. At this step of the analysis, the least statistically significant data (according to the expression p-value) will be excluded. Trying many different thresholds, it was decided that only data with p-value exponents greater than 5 or less than -4 will continue. The data that will be excluded are the following:

- Data with higher expression in tumourous tissues and p-values greater than 1E-05
- Data with higher expression in normal tissues and p-values greater than 1E-04

The number of the remaining instances is **146** (about 2% of the 7036 instances that were available at the beginning). Even this small number of instances (compared to the 7036 initial instances) will be huge for the corresponding RBPs that continue for drug screening. Therefore, even more gene instances have to be excluded.

The next filtering step is to keep only instances which are overexpressed in tumourous tissues. After checking different thresholds, it was decided that a good ratio would be to continue only with instances which have at least double expression in tumourous compared to normal tissues. The decision about thresholds comes after observing the figures. Although the point is to reduce the number of instances, an adequate number of data has to be available for the next steps. The number of the remaining instances at the end of this step is **30**.

The final goal is a list of RBPs that could be potential drug targets in cancer. At this point, the 30 remaining instances which show an over-expression in tumourous tissues correspond to less than 30 different genes. Specifically, they correspond to 27 different genes (2% of the total number of the original genes) from which 26 are lung genes and 1 is gastric gene. The **26 lung genes**, that when over-expressed have a negative effect on cancer patients’ outcome and are over-expressed in cancerous tissues (compared to normal ones) are the following:

1. FUS
2. NUSAP1
3. MRPS12
4. ANP32A
5. MSI2
6. RPS7
7. MKI67
8. HSPD1
9. MAZ
10. P4HB
11. PDAP1
12. MRPL4
13. PTBP1
14. NONO
15. ALYREF
16. HNRNPC
17. RPS26
18. LSM5
19. MRPL37
20. ALDOA
21. DDX18
22. GAPDH
23. KPNA2
24. ZNF207
25. PRKDC
26. EIF4G

Exclusion of genes with instances that follow extremely different trends

At the beginning of the current chapter, there were 1334 RBPs-genes. The point is to restrict the final list with candidate RBPs that will proceed to drug screening. After the previous steps of the analysis, the list has been restricted and contains 26 genes. At this point the multiple instances of the same genes will be checked whether they follow similar trends. The trends that will be checked are the following:

1. The effect of high gene expression on cancer patients' outcome (a high gene expression could have a negative, a positive or a zero effect)
2. The comparison between the expression in normal and cancerous tissues (some gene instances have the same expression in normal and cancerous tissues while others are over-expressed in normal or tumourous tissues respectively)

In the final list, most of the 26 genes show the same trend for all instances. However, genes for which their multiple instances follow an extremely different trend are unreliable and will be excluded, restricting even more the final list.

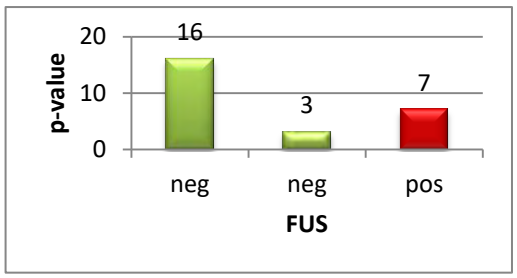
For **18 genes**, all instances show a negative effect on patient's outcome and are over-expressed in lung cancer. However, there are **8 genes** with at least one instance showing a positive effect or not being over-expressed in lung cancer. In Figures 42-49 the trends of these 8 genes are explained in detail. In each couple of figures, the first figure describes all the instances for a specific gene that present a positive or negative effect on patient's outcome. The vertical axis gives information about the p-values of

each instance. For example, the first gene (Figure 42a) has 3 instances. Two negatives with p-values 1.1E-16 and 6.4E-03, and one positive with p-value 1.6E-07. On the vertical axis the values 16, 3 and 7, correspond to the (unsigned) exponents of the respective p-values. The instances with a negative effect on cancer patients' outcome are depicted with green color and the instances with a positive effect are depicted with red color.

The second figure in each couple of figures, describes the difference in expression between normal and cancerous tissues. The Y-axis has the ratio between expression in cancerous and normal tissues. A ratio less than one means that cancerous tissues have a higher gene expression than normal ones and as this value approaches zero the difference in expression is even higher. For the 3 instances of the FUS gene (Figure 42b), these values are 0.1, 1.8 and 1.2 respectively, meaning that the first instance shows over-expression in lung cancer (green color), while the other two instances show under-expression (red color). The X-axis has the expression p-value for each instance.

As explained in Figures 42-49, four more genes will be excluded from the list due to extreme differences in trends between the multiple instances of the same gene. The excluded genes are FUS, HNRNPC, DDX18 and EIF4G1. Finally, in the new restricted list will be **22 genes**.

a.



b.

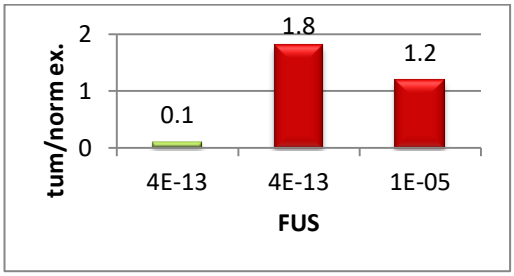


Figure 42: **The different trends for FUS gene.** This gene has 3 instances. a) Two of them have a negative effect, while the third one has a positive effect. b) The first instance shows a trend of over-expression in tumourous tissues. The other two instances show a different trend, with the second one being expressed almost twice in normal tissues. The corresponding p-values are quit low, meaning that the difference in trends is significant. **FUS** will be excluded from the final list.

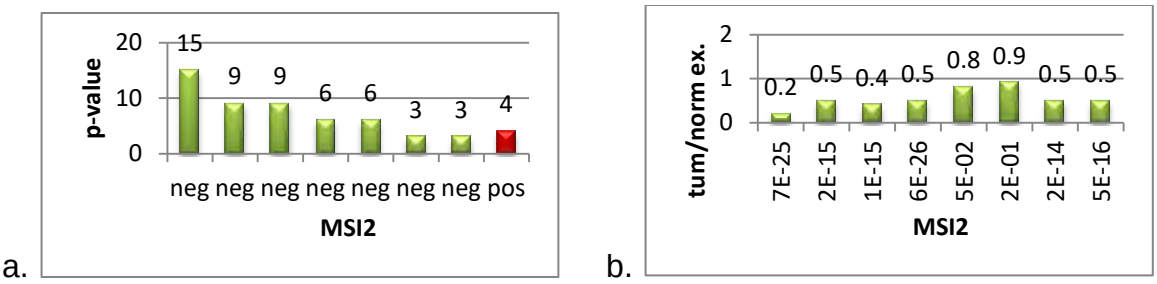


Figure 43: **The different trends for MSI2 gene.** There are 8 instances in total. a) Seven of them have a negative effect on cancer patients' outcome, while the last one has a positive effect. However, the p-value of the positive one is quite high (9.5E-04). That could be a result of inaccuracy and mistakes in microarray data. b) All instances show the same trend in the difference in expression between cancerous and normal tissues.

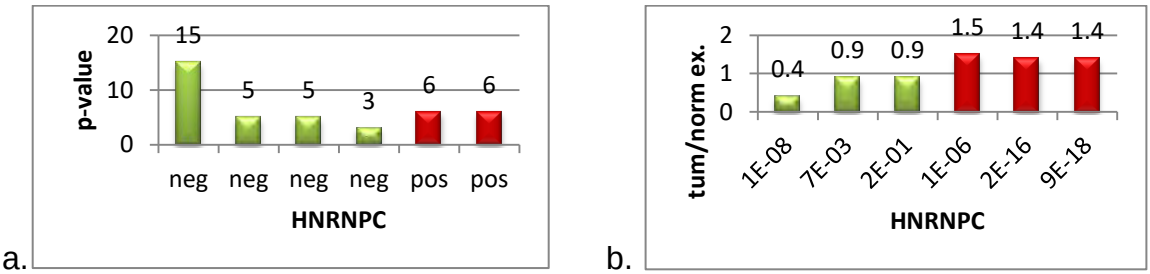


Figure 44: **The different trends for HNRNPC gene.** It has 6 instances in total. a) Four of them are negatives and two are positives. Without the first instance, the positive ones are the most statistically significant. B) Half instances are over-expressed and half are under-expressed in cancer, with the under-expressed presenting low p-values. **HNRNPC** will be excluded from the final list.

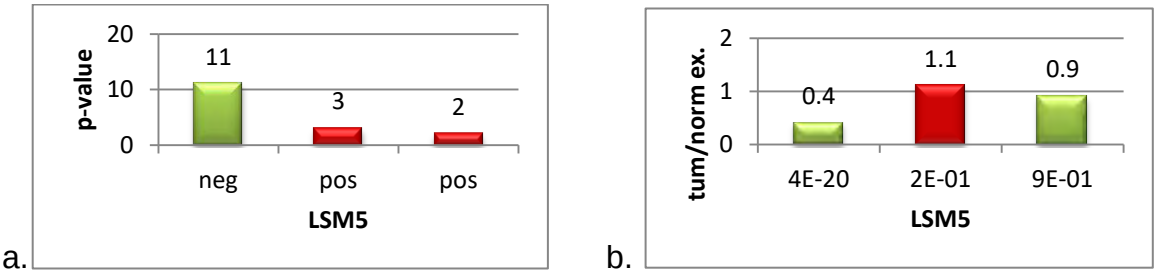
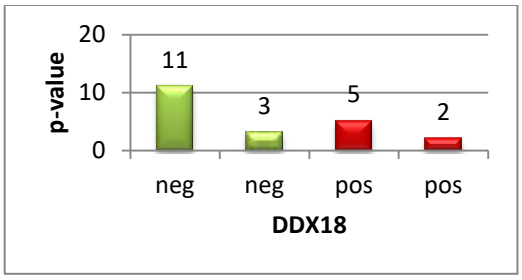


Figure 45: **The different trends for LSM5 gene.** There are 3 instances for this gene. a) The first instance is negative with a low p-value (1.5E-11). The other two, even positives, are less statistically significant with p-values 2.1E-03 and 1.2E-02 respectively. b) The first instance is over-expressed in cancer, while the second and third have almost the same expression in normal and tumourous tissues. However, with p-values 2E-01 and 9E-01 respectively, the difference in this trend can be ignored.

a.



b.

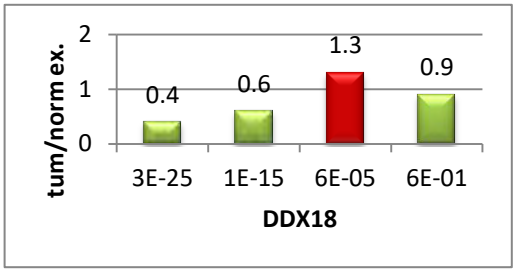
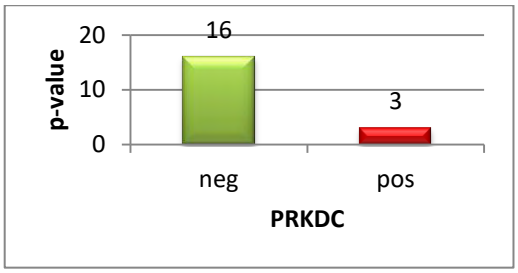


Figure 46: **The different trends for DDX18 gene.** This gene has 4 instances in total. a) Half instances present a negative effect and half a positive effect. b) The first two instances are over-expressed in cancer. The third one is under-expressed (with a relatively low p-value meaning statistical significance), while the last one presents almost the same expression between normal and cancer. Due to these differences, **DDX18** will be excluded from the list of the RBPs that will be taken forward into drug screening.

a.



b.

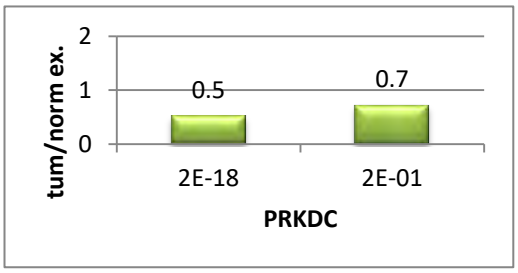
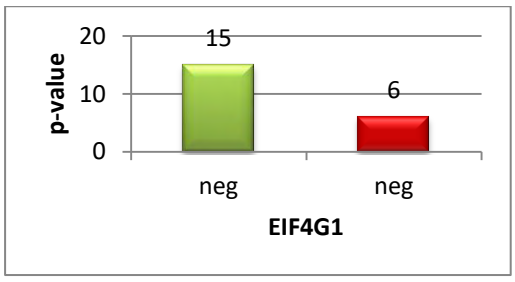


Figure 47: **The different trends for PRKDC gene.** There are 2 instances in total. a) The second instance is positive, but with a p-value equal to 5.8E-03 the difference in this trend can be ignored. b) Both instances present an over-expression in lung tumors.

a.



b.

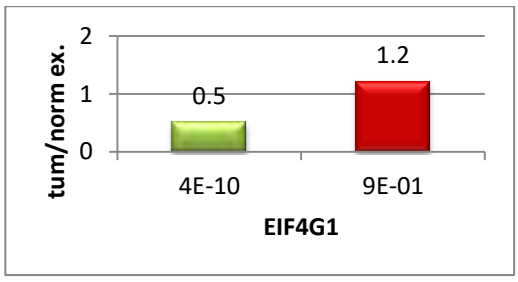
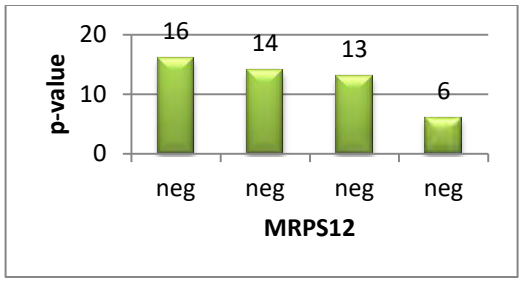


Figure 48: **The different trends for EIF4G1 gene.** a) EIF4G1 has two instances showing a different trend. b) The second instance shows under-expression in normal tissues and even though the corresponding p-value is high, **EIF4G1** will be excluded from the final list.

a.



b.

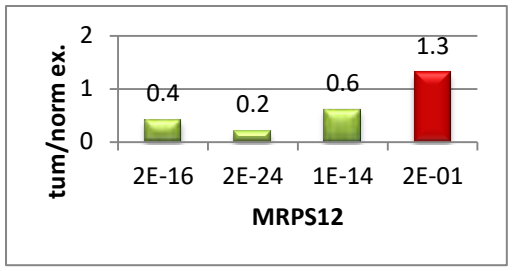


Figure 49: **The different trends for MRPS12 gene.** In total there are 4 instances. a) All instances show the same trend. b) The first three instances are over-expressed in cancer. Although the last one has a higher expression in cancerous tumors it is clear that its p-value (2E-01) is too high compared to the other p-values (2E-16, 2E-24 and 1E-14, respectively). This could be a result of pure microarray data quality. **MRPS12** remains in the list with the rest candidates, proceeding to drug screening.

Ranking of the final 22 RBPs

The previous steps of the analysis led to a list of 22 RBPs that will proceed to drug screening. A ranking of the final list gives priority to candidate RBPs that should be screened first. The ranking follows the results of STRING database for the 171 genes which have a negative effect on lung cancer patients' outcome (Figure 39).

String database (Mering et al., 2003) contains information about interactions between proteins. By protein-protein interactions (PPIs) are meant the physical contacts between two or more proteins. Proteins rarely act alone and it is a common phenomenon to be regulated by other proteins. The rationale behind the ranking is to find the protein-protein interactions between all the statistically significant RBPs (genes) that have a negative effect on lung cancer patients' survival. The hub proteins of the PPI network could be important for the disease (a hub is a small number of highly connected proteins in PPIs) (Szklarczyk et al., 2015). Thus, priority is given to these highly connected proteins and the ranking for the proteins that will be taken forward into drug screening will be done according to their number of interactions in the protein interactions network.

In Figure 51 is depicted the protein-protein interactions network according to the String database, for the 171 genes which when overexpressed have a negative effect on lung cancer patients' outcome. The network statistics indicate that this network has significantly more interactions than expected. This means that a random set of proteins of similar size is expected to have fewer interactions than the proteins of our network (the proteins are at least partially biologically connected).

In the database settings, it is possible to include or exclude specific types of interactions (Szklarczyk et al., 2017). The different sources of interactions are the following:

- Experiments
  They include only experimentally verified interactions, as imported from primary sources. Examples of this type of interactions are co-purification, co-crystallization, Yeast2Hybrid and Genetic interactions.

- Databases
  They include interactions from curated databases. Examples are signal transduction pathways, protein complexes and known metabolic pathways.

- Text-mining
  It includes proteins that are frequently mentioned together in scientific publications.

- Neighborhood
  This type of interactions includes groups of genes that are frequently detected in each other's genomic neighborhoods.

- Gene Fusion
  It is referred to interactions between genes that are sometimes fused into single open reading frames.

- Co-occurrence
  It includes gene families which show similarities in their occurrence patterns across genomes.

- Co-expression
  In this type of interactions are contained proteins whose genes seem to be correlated in expression, across a large number of experiments.

The selected sources of interactions are only the experimental interactions and the interactions derived from curated databases (Figure 50). The rest of the interaction sources are excluded, because they depend only on hypotheses, without clear evidence that there are physical contacts between the proteins.

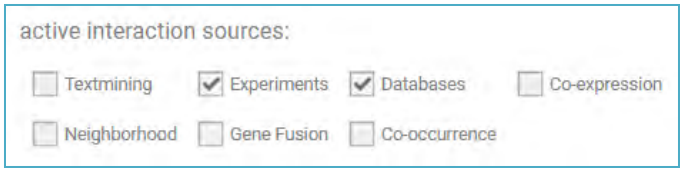


Figure 50: **String database interaction sources.** Only the experimental and database interactions are selected. For the other interaction types there is no clear evidence for physical contacts between the proteins (image from https://string-db.org/).

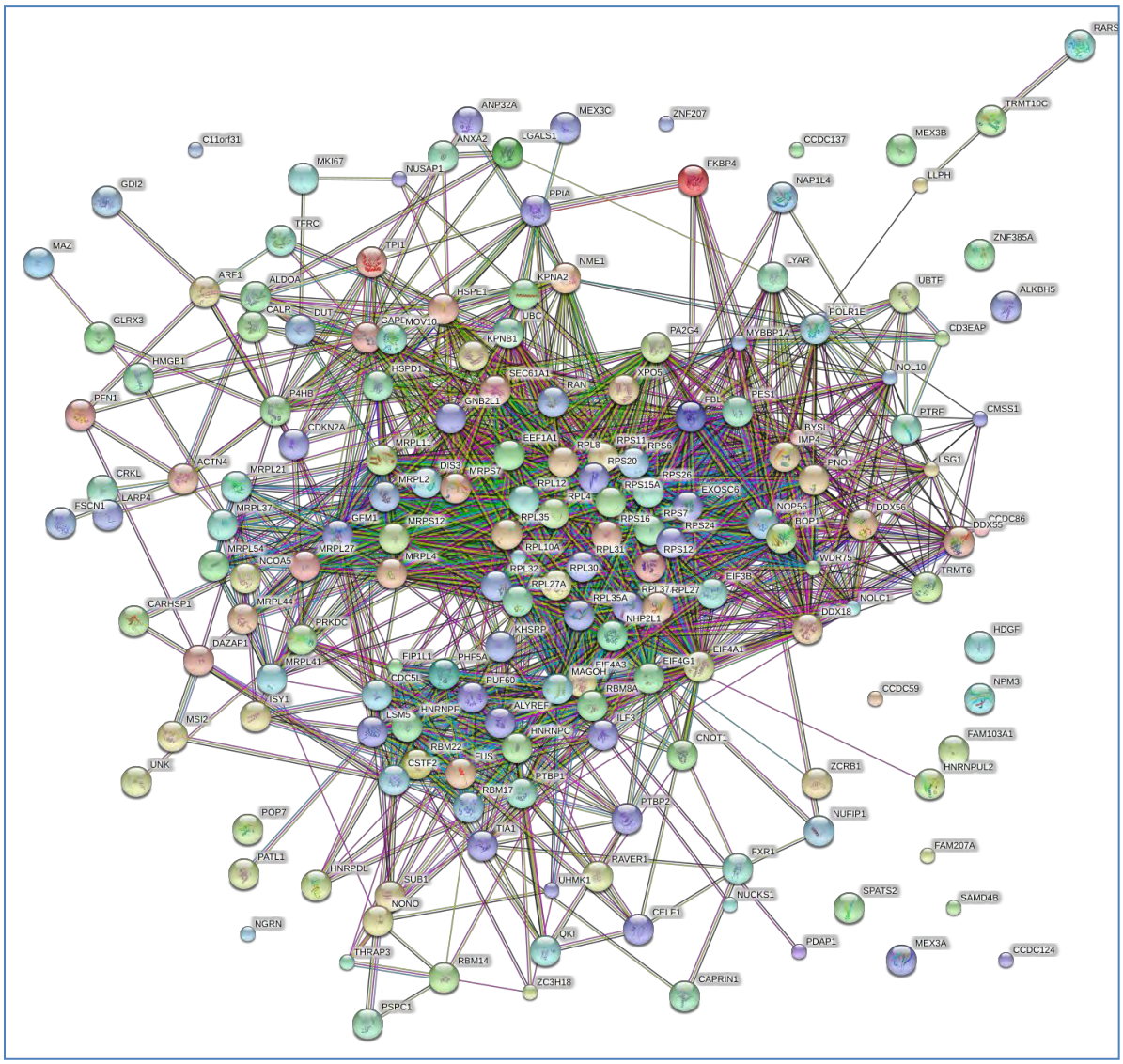


Figure 51: **Protein-protein interactions according to the String database**. In this protein network, there are proteins which have no connections (or have only a small number of connections) and proteins which are highly connected (hub proteins). Hub proteins could be important for the disease of interest (lung cancer) (image from https://string-db.org/).

In the results given by String Database, is included a csv file (Figure 52) with the interactions between the proteins of the network ranked by a combined score. The combined score is in the last column of the file. The first two columns show the names of the interacting proteins. The rest columns provide information about the individual scores, derived from the other sources of interactions. In Figure 52 only two columns have non-zero values, representing the scores from experimental and curated databases (8th and 9th column) respectively. When the individual score is zero, there are no interactions for the couple of proteins according to this source. For sake of brevity, Figure 52 shows only part of the interactions.

Having collected the data about the combined scores, the next step is to detect the highly connected proteins and give a score to each protein separately. In the csv file, each protein exists more than once. The idea is to create a new file with the ranked list of proteins (every protein existing only once). The ranking will be based on a total score.

RPS7	RPS11	0	0	0	0	0	0.998	0.900	0	0.999
RPS20	RPS11	0	0	0	0	0	0.989	0.924	0	0.999
RPS6	RPS26	0	0	0	0	0	0.997	0.900	0	0.999
RPS11	RPS16	0	0	0	0	0	0.998	0.924	0	0.999
RPS16	RPS12	0	0	0	0	0	0.998	0.900	0	0.999
...				...						...
...				...						...
...				...						...
HNRNPC	ZC3H18	0	0	0	0	0	0.417	0	0	0.416
NHP2L1	TRMT6	0	0	0	0	0	0.417	0	0	0.416
EEF1A1	XPO5	0	0	0	0	0	0.410	0	0	0.410
CDC5L	QKI	0	0	0	0	0	0.407	0	0	0.407
PES1	KPNB1	0	0	0	0	0	0.404	0	0	0.404
CSTF2	SUB1	0	0	0	0	0	0.401	0	0	0.401

Figure 52: **Part of the csv file returned by the String Database with the interactions and scores between all couples of proteins.** In the first two columns are the names of the proteins and in the last one a combined score for their interaction. The other columns have the individual scores derived from the different sources of interactions (Experiments, Databases, Text-mining, Neighborhood, Gene Fusion, Co-occurrence and Co-expression). Especially the two non-zero columns contain the individual scores from the experimental and database interactions. A zero value means that there are no interactions according to this specific source.

The ranking is explained in a simple example in which the interactions between proteins A, B, C and D (according to STRING Database) are the following:

#node1	node2	combined score
A	B	0.8
A	D	0.7
A	C	0.5
C	D	0.2

Protein A, participates in 3 interactions, protein B in 1 interaction, and proteins C and D in 2 interactions. The score will be the summary of the “combined score” for each protein interaction.

A Python script (given in the Appendix) returns a txt file with the ranking. The ranking follows the descending order of the Score and not the number of interactions for each protein. For instance, even though protein B has fewer interactions than protein C, protein B is given a better score and is ranked before C.

Genes	Interactions	Score
A	3	2.0
D	2	0.9
B	1	0.8
C	2	0.7

After this simple example, the same is applied to the list of 171 genes which have a negative effect on lung cancer patients’ outcome. The script returns 171 genes (RBPs), with the number of interactions and the score respectively. Keeping only the results for the 22 lung genes from the last step of the analysis, the final list with the candidate RBPs is the following:

	Genes - RBPs	Interactions	Score
1)	RPS7	45	41.08
2)	RPS26	39	36.40
3)	MRPS12	24	19.87
4)	MRPL4	25	19.82
5)	PTBP1	22	19.01
6)	ALYREF	17	15.13
7)	LSM5	16	14.30
8)	MRPL37	11	10.55
9)	ALDOA	7	4.70
10)	GAPDH	5	3.38
11)	P4HB	5	3.25
12)	HSPD1	5	3.14
13)	NONO	4	2.41
14)	KPNA2	2	1.92
15)	PRKDC	3	1.62
16)	MSI2	2	1.47
17)	ANP32A	2	1.42
18)	NUSAP1	1	0.61
19)	PDAP1	1	0.42
20)	MKI67	0	0
21)	MAZ	0	0
22)	ZNF207	0	0

According to the computational analysis results, the higher the interaction's score, the most possible for the RBP to be a successful drug candidate. The first RBP to be screened is RPS7, which participates in 45 interactions and is the highest ranked amongst the RBP candidates. The rest proteins from the list will be screened in respect to their given score. For three RBPs in the list (MKI67, MAZ, and ZNF207) there are no observed interactions. Thus the given score for them is zero and they will be screened last. After the analysis and the ranking procedure, the final 22 RNA-Binding Proteins are the most promising candidates to proceed into drug screening.

Chapter 5: Discussion

General

The identification of new therapeutic targets is a high priority task in cancer research. RNA binding proteins (RBPs) have poorly exploited so far and could constitute possible targets, as they are central regulators of gene expression and are highly associated with increased cancer aggressiveness (Pereira et al., 2016). According to different studies, RBPs expression is altered in cancer cells compared to normal cells, for several cancer types (Hopkins et al., 2016; Qian et al, 2015; King et al., 2011; Pereira et al., 2016).

The present thesis, based on a computational analysis, intended on reducing the total number of RBPs that will be taken forward into drug screening for novel anti-tumour compounds. Starting with a list of 1334 candidate RBPs (compiled by the Lewis laboratory) and using available data on the expression levels of the corresponding genes and cancer survival data, the list was restricted to the most promising candidates (22 final RBPs).

An online tool called KM-Plotter (Gyorffy et al., 2013) contributed already collected data from four background databases. The datasets which constitute them, derive from caBIG (Cancer Biomedical Informatics; Cimino et al., 2007), TCGA (The Cancer Genome Atlas; Tomczak et al., 2015) and GEO (Gene Expression Omnibus; Clough et al., 2016). All available information derived from lung, gastric, ovarian and breast cancer patients, with only Affymetrix microarrays data being included (Gyorffy et al., 2010, 2012, 2013; Szasz et al., 2016).

A protein rarely acts alone and it is a common phenomenon to be regulated by other proteins. In a protein-protein interactions network (PPI network), a small number of highly connected proteins (called hub proteins) may reveal a major role for the disease of interest (Szklarczyk et al., 2015). In the present thesis the STRING database results were used for known and predicted protein-protein interactions. In the STRING network participated only RBPs which have a negative effect on cancer patients' outcome when overexpressed. The RBPs constituting a hub in the network are estimated as more likely to be successful drug targets.

In summary, 993 RBPs from the compiled list were detected in KM-Plotter and the collected data corresponded to the effect of each candidate RBP on cancer patients' survival and their expression levels in normal and cancerous tissues. Applying filters, the list of candidate RBPs progressively was narrowed down based on negative or positive effect of expression levels on patient survival, the difference in expression levels of RBPs in tumourous versus normal tissues, the corresponding p-values ("Risk p-values" and "Expression p-values" respectively), and the concordance of multiple

probes available for each gene. Performing network analysis on candidate RBPs, a concise list of the most promising 22 candidate molecules was ranked to be taken forward into drug screening.

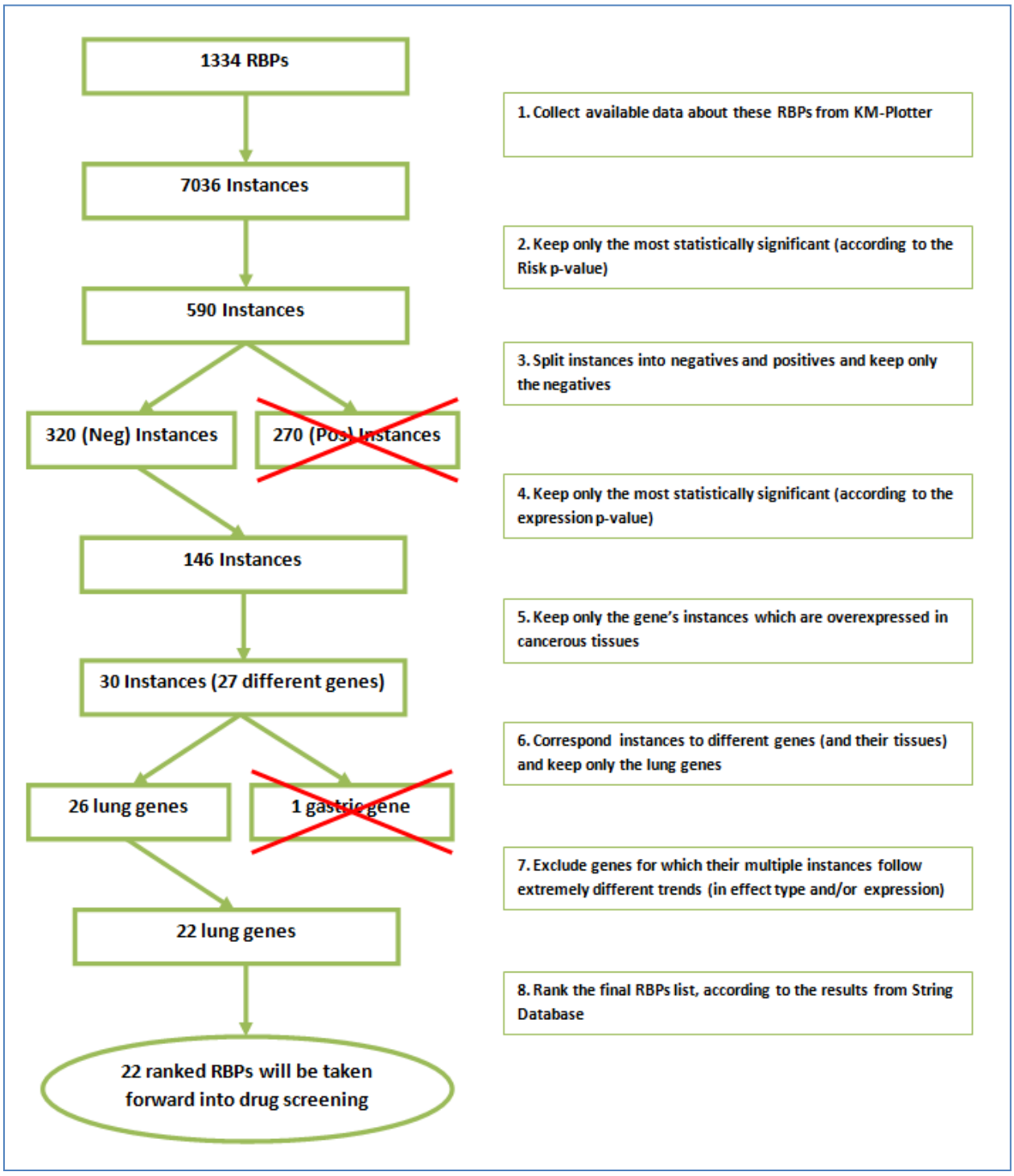


Figure 53: **Outline of computational analysis steps and results. 1:** From the total number of 1334 RBPs, 993 were detected in KM-Plotter online tool. Due to the concordance of multiple probes available for each gene, among KM-Plotter's results were multiple plots (7036 total collected instances). **2-5:** Applied filters progressively narrow down the number of instances to 30. **6:** These instances correspond to 27 different genes which have a negative effect on cancer patients' outcome when overexpressed and are overexpressed in cancerous tissues. Among the 27 genes, there is a gene deriving from gastric cancer patients and is excluded. **7:** The rest 26 genes (which correspond to lung cancer patients) are narrowed down by 4 more genes. For these genes, the multiple probes show extremely different trends in effect type or expression. **8:** The list with the final 22 RBPs is ranked according to the results of STRING database. All RBPs participating in the STRING network, when overexpressed have a negative effect on lung cancer patients and are overexpressed in cancerous tissues compared to the normal ones. From the 22 final RBPs, all of those which constitute a hub in the network are estimated as more possible to be successful drug targets.

Evaluation of the results and emerging problems

In this thesis, starting with a compiled list of 1334 RBPs, we searched them one-by-one in KM-Plotter online tool (Gyorffy et al., 2013). KM-Plotter's data derive from Affymetrix microarrays, with a concordance of multiple probes available for the same gene or a probe set ID mapping to several genes. From the total number of 1334 RBPs, 993 were detected in KM-Plotter online tool having one or more probes corresponding to each gene. The rest 341 RBPs were unable to contribute information to the gathered data, due to their correspondence to probe sets which mapped to several genes or due to them being undetected during the KM-Plotter's search.

The results in KM-Plotter derive from a background database which contains information from different Affymetrix microarray platforms (Gyorffy et al., 2010, 2012, 2013; Szasz et al., 2016). An emerging problem is the many issues in DNA microarray measurements (Draghici, 2006). Even though a highly used method of gene expression profiling, microarrays suffer in measurement precision and in Affymetrix microarrays many probes do not correspond to the appropriate transcripts (Mecham 2004). A gene may have many variants and the accuracy of the results could be ensured only if the identities of the genes and variants are well known (Lee et Roy, 2004; Boue et al., 2003). Finally, due to mistakes many spotted probes to the microarray chip, are not the intended ones (Knight, 2001; Forman-Kay, 1999).

In gene expression profiling with Affymetrix microarrays, each transcript is represented by multiple probes (Rogojina et al., 2010). Thus, for many genes the collected data correspond to more than one probe. An intense problem emerges when the multiple

probes of the same gene follow contradictory trends. In the steps of the analysis followed in the present thesis, the list of RBPs was narrowed down. At the point the list had only 26 RBPs, the different probe results of the corresponding genes were compared. Out of the 26 RBPs, 18 RBPs (70%) showed the same trend (negative effect on lung cancer patients' survival and higher gene expression in cancerous compared to normal tissues). The rest 30% of the RBPs (8 RBPs) corresponded at least to one probe which either had a positive effect on lung cancer patients' outcome or were under-expressed in lung cancer (four RBPs showed extremely different trends and were omitted during the final filtering step).

To evaluate the magnitude of the problem with the multiple probes for the same gene and the extremely different trends, from the initial list of RBPs the results of Script 4 in Chapter 3 were exploited. From the total number of 7036 collected instances, 1996 were retrieved from KM-Plotter's lung cancer environment, 1368 of those being the most statistically significant in terms of the "Risk p-value" (Risk p-value<1E-04). The 1368 instances corresponded to 768 different genes. While 45% of this set of genes showed a negative, 36% showed a positive effect on lung cancer patients' survival. For the remaining 19%, the results were ambiguously distributed among the negative and positive effect on the survival of lung cancer patients.

In this study, after the computational analysis and the narrowing down of the compiled list of RBPs, the remaining 22 RBPs will be taken forward into drug screening for novel anti-tumour compounds in lung cancer. Overexpression of these final RBPs, indicate poor survival in lung cancer patients and their expression in lung cancer cells is higher compared to that of lung cells from healthy patients. An interesting part of this project could be to evaluate the behavior of these 22 RBPs in gastric, ovarian and breast cancer. An RBP which follows a specific trend in one cancer type could either follow the same or an extremely different trend in the other three cancer types. This discovery could have a great scientific significance for a better understanding of the cancer mechanisms. To support the final results, literature on these RBPs suggests involvement of their overexpression in poor survival of patients suffering from various cancer types (Na et al., 2017; Liang et al., 2015; Li et al., 2013; Zheng et al., 2010; Gluz et al., 2008; Gousias et al., 2012; Mu et al, 2013; Zong et al, 2016; Chen et al., 2015).

Further research

The alteration in RBPs expression is highly associated with tumorigenesis (Pereira et al., 2016). In the current thesis a list of human RBPs was compiled by the Lewis laboratory and contained 1334 RBPs in total. However this number is a portion of the total experimentally validated RBPs. By August 2017, the total number of validated human RBPs was 1542 (Gertsberger et al., 2014; Pereira 2016). These numbers are really promising and it is possible that the next years even more human RBPs will be validated experimentally. In a possible future study, the same analysis could be followed for a renewed compiled list with more candidates.

Searching in KM-Plotter's online tool, only a fraction of the original RPB's list was detected, due to many RBPs failing to contribute data. KM-Plotter exploits a background database with data deriving from caBIG (Cimino et al., 2007), TCGA (Tomczak et al., 2015) and GEO (Clough et al., 2016). Instead of using KM-Plotter and its embedded data collection method (specific keywords search only), in a future study, gene expression and survival data about these RBPs could be gathered directly from the three aforementioned repositories. Subsequently, a similar procedure to the one followed by KM-Plotter developers could calculate the Kaplan-Meier plots and the p-values respectively, without any need to alter the computational analysis process (Gyorffy et al., 2010, 2012, 2013; Szasz et al., 2016).

An alternative approach in gathering data procedure would be to avoid the collection of multiple probes for the same gene, but obtaining the best probe only. A developed scoring method called Jetset, evaluates the probe sets in terms of specificity, coverage and robustness (Li et al., 2011). From the multiple Affymetrix microarray probes corresponding to each gene, only one probe set (assessed as the best by the scoring method) will be used, solving the problem of multiple sets and providing a one-to-one correspondence among genes and probe sets. Contrary to using "all probe sets per gene" setting in KM-Plotter's environment, it would be scientifically interesting to use the "only Jetset best probe per gene" setting.

Data types, such as sequence or copy number variations, could be used to contribute additional information, providing a higher accuracy in the followed computational analysis. In this study, a computational analysis was conducted based on each gene's behavior separately. The cancer patients' outcome was cross-checked according to each gene's expression. However, the survival of a patient could be influenced not only by upward expression trends of individual genes, but from under-expression of other genes simultaneously (Chen et al., 2015; Liang et al., 2015;).

Appendix

Script producing data statistics

```
l=open("C:/Users/~/Data text.txt","r")
k=l.readline()
all_instances=[]
while(k!=''):
    p=k.split('\t')
    #p[4]=p[4][0:1]
    all_instances.append(p)
    k=l.readline()

l.close()

number_of_instances=len(all_instances)
number_of_lung=0
number_of_gastric=0
number_of_ovarian=0
number_of_breast=0
number_of_neg=0
number_of_pos=0
number_of_zero=0

single_genes=[]
lung_single_genes=[]
lung_single_genes_info=[]
gastric_single_genes=[]
gastric_single_genes_info=[]
ovarian_single_genes=[]
ovarian_single_genes_info=[]
breast_single_genes=[]
breast_single_genes_info=[]


for i in all_instances:
    if i[3]=='lung':
        number_of_lung+=1
        if i[0] not in lung_single_genes:
            lung_single_genes.append(i[0])
            lung_single_genes_info.append(i)
    elif i[3]=='gastric':
        number_of_gastric+=1
        if i[0] not in gastric_single_genes:
            gastric_single_genes.append(i[0])
            gastric_single_genes_info.append(i)
    elif i[3]=='ovarian':
        number_of_ovarian+=1
        if i[0] not in ovarian_single_genes:
            ovarian_single_genes.append(i[0])
            ovarian_single_genes_info.append(i)
```

```
    elif i[3]=='breast':
        number_of_breast+=1
        if i[0] not in breast_single_genes:
            breast_single_genes.append(i[0])
            breast_single_genes_info.append(i)

    if i[0] not in single_genes:
        single_genes.append(i[0])

    if i[5]=='neg':
        number_of_neg+=1
    elif i[5]=='pos':
        number_of_pos+=1
    elif i[5]=='zero':
        number_of_zero+=1


w=open("C:/Users/~/Information.txt","w")

w.write('_____________________________________________________\nAll instances:'+\
        str(number_of_instances)+'\nLung:'+str(number_of_lung)+'\nGastric:'+\
        str(number_of_gastric)+' \nOvarian:'+str(number_of_ovarian)+' \nBreast:'+\
        str(number_of_breast)+' \n')

w.write('\nNeg:'+str(number_of_neg)+' \nPos:'+str(number_of_pos)+' \nZero:'+\
        str(number_of_zero)+' \n\n')

w.write('_____________________________________________________\nDifferent genes:'+\
        str(len(single_genes))+'\nLung:'+str(len(lung_single_genes))+'\nGastric:'+\
        str(len(gastric_single_genes))+' \nOvarian:'+str(len(ovarian_single_genes))+\
        ' \nBreast:'+str(len(breast_single_genes))+ '\n')

print('All instances:'+str(number_of_instances)+'\nLung:'+str(number_of_lung)+\
      '\nGastric:'+str(number_of_gastric)+' \nOvarian:'+str(number_of_ovarian)+\
      ' \nBreast:'+str(number_of_breast)+' \n')

print('Neg:'+str(number_of_neg)+' \nPos:'+str(number_of_pos)+' \nZero:'+\
      str(number_of_zero)+' \n\n')

print('Different genes:'+str(len(single_genes))+'\nLung:'+str(len(lung_single_genes))+\
      '\nGastric:'+str(len(gastric_single_genes))+' \nOvarian:'+\
      str(len(ovarian_single_genes))+' \nBreast:'+str(len(breast_single_genes)))



common_lung_gastric=[]
common_gastric_ovarian=[]
common_lung_ovarian=[]
common_lung_breast=[]
common_gastric_breast=[]
common_ovarian_breast=[]
```

```
for i in lung_single_genes:
    for j in gastric_single_genes:
        if i==j:
            common_lung_gastric.append(i)

    for k in ovarian_single_genes:
        if i ==k:
            common_lung_ovarian.append(i)

    for m in breast_single_genes:
        if i ==m:
            common_lung_breast.append(i)

for i in gastric_single_genes:
    for j in ovarian_single_genes:
        if i==j:
            common_gastric_ovarian.append(i)

    for m in breast_single_genes:
        if i ==m:
            common_gastric_breast.append(i)

for i in ovarian_single_genes:
    for j in breast_single_genes:
        if i==j:
            common_ovarian_breast.append(i)



w.write('*************************************************************')
if common_lung_gastric:
    w.write('\nThere are '+ str(len(common_lung_gastric))+\
            ' common genes between lung and gastric\n')
    print('***There are '+ str(len(common_lung_gastric))+\
          ' common genes between lung and gastric\n')
if common_lung_ovarian:
    w.write('\nThere are '+ str(len(common_lung_ovarian))+\
            ' common genes between lung and ovarian\n')
    print('***There are '+ str(len(common_lung_ovarian))+\
          ' common genes between lung and ovarian\n')
```

```
if common_lung_breast:
    w.write('\nThere are '+ str(len(common_lung_breast))+\
            ' common genes between lung and breast\n')
    print('***There are '+ str(len(common_lung_breast))+\
          ' common genes between lung and breast\n')
if common_gastric_ovarian:
    w.write('\nThere are '+ str(len(common_gastric_ovarian))+\
            ' common genes between gastric and ovarian\n')
    print('***There are '+ str(len(common_gastric_ovarian))+\
          ' common genes between gastric and ovarian\n')
if common_gastric_breast:
    w.write('\nThere are '+ str(len(common_gastric_breast))+\
            ' common genes between gastric and breast\n')
    print('***There are '+ str(len(common_gastric_breast))+\
          ' common genes between gastric and breast\n')
if common_ovarian_breast:
    w.write('\nThere are '+ str(len(common_ovarian_breast))+\
            ' common genes between ovarian and breast\n')
    print('***There are '+ str(len(common_ovarian_breast))+\
          ' common genes between ovarian and breast\n')
w.write('**********************************************************\n\n')

if lung_single_genes:
    w.write('Lung genes:\n')
    for i in lung_single_genes:
        w.write(str(i) + '\n' )
    w.write('_____________________________________________________\n')
if gastric_single_genes:
    w.write('Gastric genes:\n')
    for i in gastric_single_genes:
        w.write(str(i) + '\n' )
    w.write('_____________________________________________________\n')
if ovarian_single_genes:
    w.write('Ovarian genes:\n')
    for i in ovarian_single_genes:
        w.write(str(i) + '\n' )
    w.write('_____________________________________________________\n')
if breast_single_genes:
    w.write('Breast genes:\n')
    for i in breast_single_genes:
        w.write(str(i) + '\n' )
    w.write('_____________________________________________________\n')
```

```
if common_lung_gastric:
    w.write('Common genes between lung and gastric:\n')
    for i in common_lung_gastric:
        w.write(str(i) + '\n')
    w.write('_____\n')
if common_lung_ovarian:
    w.write('Common genes between lung and ovarian:\n')
    for i in common_lung_ovarian:
        w.write(str(i) + '\n')
    w.write('_____\n')
if common_lung_breast:
    w.write('Common genes between lung and breast:\n')
    for i in common_lung_breast:
        w.write(str(i) + '\n')
    w.write('_____\n')
if common_gastric_ovarian:
    w.write('Common genes between gastric and ovarian:\n')
    for i in common_gastric_ovarian:
        w.write(str(i) + '\n')
    w.write('_____\n')
if common_gastric_breast:
    w.write('Common genes between gastric and breast:\n')
    for i in common_gastric_breast:
        w.write(str(i) + '\n')
    w.write('_____\n')
if common_ovarian_breast:
    w.write('Common genes between ovarian and breast:\n')
    for i in common_ovarian_breast:
        w.write(str(i) + '\n')
    w.write('_____\n')

w.close()
```

Script 5: **Script producing data statistics**

Script for RBP final ranking according to the results of String Database

```
l=open("C:/Users/~/interactions & scores.txt","r")
k=l.readline()
k=l.readline()
genes=[]
while(k!=''):
    genes.append(k)
    k=l.readline()

l.close()

different_genes=[]
gene_score=[]
times=[]
for i in genes:
    g=i.split('\t')
    if g[0] not in different_genes:
        different_genes.append(g[0])
        gene_score.append(float(g[2]))
        times.append(1)
    else:
        position=different_genes.index(g[0])
        gene_score[position]+=float(g[2])
        times[position]+=1

    if g[1] not in different_genes:
        different_genes.append(g[1])
        gene_score.append(float(g[2]))
        times.append(1)

    else:
        position=different_genes.index(g[1])
        gene_score[position]+=float(g[2])
        times[position]+=1


for i in range (len(different_genes)):
    print(different_genes[i] + '\t'+ str(times[i]) +\
          '\t'+ str(round(gene_score[i],2)))


yx=list(zip(gene_score, different_genes))
yx.sort()
different_genes=[x for y,x in sorted(list(zip(gene_score, different_genes)))]
different_genes.reverse()
```

```
yx=list(zip(gene_score, times))
yx.sort()
times=[x for y,x in sorted(list(zip(gene_score, times)))]
times.reverse()

gene_score.sort()
gene_score.reverse()

w=open("C:/Users/~/interacting_genes.txt","w")

w.write("Genes" + "\tInteractions" + "\tConfidence\n----------------------\n")
for i in range (len(different_genes)):
    w.write(different_genes[i] + '\t\t'+ str(times[i]) + '\t'+\
            str(round(gene_score[i],2)) +'\n' )

w.close()
```

Script 6: **Script for RBP final ranking according to the results of string database**

Chapter 6: Bibliography

Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt K, Resenchuk S, Tatusova T, et al., **BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata**, Nucleic Acids Res., 40:D57–D63, **(2012)**

Barrett T, Wilhite S, Ledoux P, Evangelista C, Kim I, Tomashevsky M, Marshall K, Phillippy K, Sherman P, Holko M, Yefanov A, Lee H, Zhang N, Robertson C, Serova N, Davis S & Soboleva A, **NCBI GEO: archive for functional genomics data sets—update**, Nucleic Acids Res., 41(Database issue):D991-5, **(2013)**, doi: 10.1093/nar/gks1193

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball C, Causton H, et al., **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data**, Nat. Genet.,29:365–371, **(2001)**

Boue S, Letunic I, Bork P, **Alternative splicing and evolution**, Bioessays, 25(11):1031–4, **(2003)**, doi:10.1002/bies.10371

Busuttil R, George J, Tothill R, Ioculano K, Kowalczyk A, Mitchell C, Lade S, Tan P, Haviv I & Boussioutas A, **A signature predicting poor prognosis in gastric and ovarian cancer represents a coordinated macrophage and stromal response**. Clin Cancer Res, 20:2761–2772, **(2014)**

Cancer Genome Atlas Research Network, Weinstein J, Collisson E, Mills G, Shaw K, Ozenberger B, Ellrott K, Shmulevich I, Sander C & Stuart J, **The Cancer Genome Atlas Pan-Cancer analysis project**, Nat Genet., 45(10):1113-20, **(2013)**, doi: 10.1038/ng.2764

Chen D & Peace K, **Clinical Trial Data Analysis Using R**, CRC Press, pp. 99–108, **(2014)**

Chen L, Yang L, Qiao F, Hu X, Li S, Yao L, Yang X & Shao Z, **High Levels of Nucleolar Spindle-Associated Protein and Reduced Levels of BRCA1 Expression Predict Poor Prognosis in Triple-Negative Breast Cancer**, PLoS One, 10(10):e0140572, **(2015)**, doi: 10.1371/journal.pone.0140572.

Cimino J, Hayamizu T, Bodenreider O, Davis B, Stafford G & Ringwald M, **The caBIG terminology review process**, J Biomed Inform, 42(3):571-80, **(2009)**, doi: 10.1016/j.jbi.2008.12.003

Ciociola A, Cohen L & Kulkarni P, **How drugs are developed and approved by the FDA: current process and future directions**, Am J Gastroenterol, 109 (5): 620–3, **(2014)**

Clough E & Barrett T, **The Gene Expression Omnibus Database**, Methods Mol Biol.,1418:93-110, **(2016)**, doi: 10.1007/978-1-4939-3578-9_5

Crick F, **On Protein Synthesis. In F.K. Sanders. Symposia of the Society for Experimental Biology, Number XII: The Biological Replication of Macromolecules**, Cambridge University Press, pp. 138–163, **(1958)**

Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J & Shi B, **Breast cancer intrinsic subtype classification, clinical use and future trends**, Am J Cancer Res., 5(10):2929-43, **(2015)**

Daley G, **Stem cells and the evolving notion of cellular identity**, Philos Trans R Soc Lond B Biol Sci, 370(1680):20140376., **(2015)**, doi: 10.1098/rstb.2014.0376

Draghici S, Khatri P, Eklund A & Szallasi Z, **Reliability and reproducibility issues in DNA microarray measurements**, Trends Genet., 22(2): 101–109, **(2006)**

Dreyfuss G, Kim N & Kataoka N, **Messenger-RNA-binding proteins and the messages they carry**, Nature Rev. Mol. Cell Biol. 3, 195–205, **(2002)**

Eder J & Herrling P, **Trends in Modern Drug Discovery**, Handb Exp Pharmacol.,232:3-22., **(2016),** doi: 10.1007/164_2015_20

Elmore S, **Apoptosis: A Review of Programmed Cell Death**, Toxicol Pathol., 35(4):495–516., **(2007)**, doi: 10.1080/01926230701320337

Everett J, **Academic drug discovery: current status and prospects**, Expert Opin Drug Discov., 10(9):937-44, **(2015)**, doi: 10.1517/17460441.2015.1059816

Feinstein A, **The architecture of clinical research**, Philadelphia: W. B. Saunders Co; Clinical Epidemilogy, p. 226-227, **(1985)**

Forman-Kay JD, **The 'dynamics' in the thermodynamics of binding**. Nature Structural Biology, 6:1086–1087, **(1999)**, doi:10.1038/70008

Forster S, Gretschel S, Jons T, Yashiro M & Kemmner W, **THBS4, a novel stromal molecule of diffuse-type gastric adenocarcinomas, identified by transcriptome-wide expression profiling**, Mod Pathol., 24:1390–1403, **(2011)**

Gerstberger S, Hafner M & Tuschl T, **A census of human RNA-binding proteins**, Nature Reviews Genetics, 15, 829–845, **(2014)**, doi:10.1038/nrg3813

Glisovic T, Bachorik J L, Yong J & Dreyfuss G, **RNA-binding proteins and post-transcriptional gene regulation**, FEBS Letters Elsevier, 582 (14): 1977–1986, **(2008)**

Gluz O, Wild P, Meiler R, Diallo-Danebrock R, Ting E, Mohrmann S, Schuett G, Dahl E, Fuchs T, Herr A, Gaumann A, Frick M, Poremba C, Nitz U & Hartmann A, **Nuclear karyopherin alpha2 expression predicts poorsurvival in patients with advanced breast cancer irrespective of treatment intensity**, Int J Cancer, 123(6):1433-8, **(2015)**, doi: 10.1002/ijc.23628.

Gousias K, Becker A, Simon M & Niehusman P, **Nuclear karyopherin a2: a novel biomarker for infiltrative astrocytomas**, J Neurooncol, 109(3):545-53, **(2012)**, doi: 10.1007/s11060-012-0924-2.

Gyorffy B, Lanczky A, Eklund A, Denkert C, Budczies J, Li Q & Szallasi Z, **An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients**, Breast Cancer Res Treat, 123:725–731, **(2010)**, doi:10.1007/s10549-009-0674-9

Gyorffy B, Lanczky A & Szallasi Z, **Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients**, Endocrine-Related Cancer, 19 197-208, **(2012)**

Gyorffy B, Surowiak P, Budczies J & Lanczky A, **Online Survival Analysis Software to Assess the Prognostic Value of Biomarkers Using Transcriptomic Data in Non-Small-Cell Lung Cancer**, PLoS ONE 8(12): e82241, **(2013)**, doi:10.1371/journal.pone.0082241

Hauptmann S, du Bois A, Meinhold-Herlein I, Pfisterer J & Avril S, **Histological grading of epithelial ovarian cancer. Review and recommendation**, Pathologe, 35(5):497-503, **(2014)**, doi: 10.1007/s00292-014-1948-5

Hardiman G, **Microarray platforms–comparisons and contrasts**, Pharmacogenomics, 5:487–502, **(2004)**

Hogan J, Riordan P, Herschlag D & Brown O, **Diverse RNA-Binding Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System**, PLOS Biology. Public Library of Science, 6(10): e255, **(2008)**, PMC 2573929 , PMID 18959479, doi:10.1371/journal.pbio.0060255

Hopkins T, Mura M, Al-Ashtal H, Lahr R, Abd-Latip N, Sweeney K, Lu H, Weir J, El-Bahrawy M & Blagden S, **The RNA-binding protein LARP1 is a post-transcriptional regulator of survival and tumorigenesis in ovarian cancer**, Nucleic Acids Res. 44, 1227-1246, **(2016),** doi: 10.1093/nar/gkv1515

Ionov Y, Peinado M, Malkhosyan S, Shibata D & Perucho M, **Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis**, Nature, 363(6429): 558–61, **(1993)**

Jason R, Gail N, Randal P, Courtney V, Phil O, Nussenbaum B & Wang E, **A practical guide to understanding Kaplan-Meier Curves**, Otolaryngol Head Neck Surg, 143(3): 331–336, **(2010)**, doi:10.1016/j.otohns.2010.05.007

Kaplan L & Meier P, **Nonparametric estimation from incomplete observations**, J. Amer. Statist. Assn., 53 (282): 457–481, **(1958),** doi:10.2307/2281868

Kim H, Choi I, Kim C, Kim H, Oshima A, Yamada Y, Arao T, Nishio K, Michalowski A & Green J, **Three-gene predictor of clinical outcome for gastric cancer patients treated with chemotherapy**, Pharmacogenomics J, 12:119–127, (**2012**)

Kim K, Zakharkin S & Allison D, **Expectations, validity, and reality in gene expression profiling**, J Clin Epidemiol., 63(9):950-9, **(2010)**, doi: 10.1016/j.jclinepi.2010.02.018

King C, Cuatrecasas M, Castells A, Sepulveda A, Lee J & Rustgi A, **Lin28b promotes colon cancer progression and metastasis**, Cancer Res.; 71(12): 4260–4268., **(2011)**, doi:10.1158/0008-5472.CAN-10-4637

Knight J, **When the chips are down**, Nature, 410:860–861, **(2001)**, doi:10.1038/35073680

Kroese M, Zimmern R & Pinder S, **HER2 status in breast cancer-an example of pharmacogenetic testing**, J R Soc Med, 100(7):326-9, **(2007)**

Lee C, Roy M, **Analysis of alternative splicing with microarrays: successes and challenges**, Genome Biol., 5(7):231, **(2004)**, doi:10.1186/gb-2004-5-7-231. http://dx.doi.org/10.1186/gb-2004-5-7-231

Liang W, Yang C, Peng J, Qian Y & Wang Z, **The Expression of HSPD1, SCUBE3, CXCL14 and Its Relations with the Prognosis in Osteosarcoma**, , 73(3):763-8, **(2015)**, doi: 10.1007/s12013-015-0579-7.

Li C, Ji L, Ding Z, Zhang Q & Huang G, Overexpression of KPNA2 correlates with poor prognosis in patients with gastric adenocarcinoma**, Tumour Biol, 34(2):1021-6,** (2013)**, doi: 10.1007/s13277-012-0641-7**

Li Q, Birkbak N, Gyorffy B, Szallasi Z & Eklund A, Jetset: selecting the optimal microarray probe set to represent a gene**, BMC Bioinformatics,** (2011)**, 12:474, doi: 10.1186/1471-2105-12-474**

Loeb L, Loeb K & Anderson I, **Multiple mutations and cancer**, Proc Natl Acad Sci U S A, 100(3): 776–781, **(2003)**, doi:10.1073/pnas.0334858100

Lukong K, Chang K, Khandjian E & Richard S, **RNA-binding proteins in human genetic disease**, Trends Genet. 24, 416–425 **(2008)**

Lunde M, Moore C & Varani G, **RNA-binding proteins: Modular design for efficient function**, Nature Reviews Molecular Cell Biology, 8(6): 479–90, **(2007)**, PMID 17473849, doi:10.1038/nrm2178

Maggiore C, Mulè A, Fadda G, Rossi E, Lauriola L, Vecchio F & Capelli A, **Histological classification of lung cancer**, Rays, 29(4):353-5, **(2004)**

Matera G, Terns R & Terns M, **Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs**, Nature Reviews Molecular Cell Biology, Nature Publishing Group, 8(3): 209–220, **(2007)**, PMID 17318225, doi:10.1038/nrm2124

Meyer Bruce D, **Unemployment Insurance and Unemployment Spells**, Econometrica, 58 (4): 757–782, **(1990),** doi:10.2307/2938349

Moffat J, Rudolph J & Bailey D, **Phenotypic screening in cancer drug discovery - past, present and future**, Nat Rev Drug Discov, 13(8):588-602, **(2014)**, doi: 10.1038/nrd4366

Mu Q, Wang Y, Chen B, Qian W, Meng H, Tong H, Chen F, Ma Q, Ni W, Chen S & Jin J, **High expression of Musashi-2 indicates poor prognosis in adult B-cell acute lymphoblastic leukemia**, Leuk Res, 37(8):922-7, **(2013)**, doi: 10.1016/j.leukres.2013.05.012

Na N, Li H, Xu C, Miao B, Hong L, Huang Z & Jiang Q, High expression of Aldolase A predicts poor survival in patients with clear-cell renal cell carcinoma**, Ther Clin Risk Manag, 13:279-285,** (2017)**, doi: 10.2147/TCRM.S123199**

Olivier M, Hollstein M & Hainaut P, TP53 mutations in human cancers: origins, consequences, and clinical use**, Cold Spring Harb Perspect Biol., 2(1):a001008,** (2010)**, doi:10.1101/cshperspect.a001008**

Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N & Brazma A, **ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments**, Nucleic Acids Res, 39(Database issue):D1002-4, **(2011)**, doi: 10.1093/nar/gkq1040.

Pereira B, Billaud M & Almeida R, **RNA-Binding Proteins in Cancer: Old Players and New Actors**, Trends in Cancer, Vol. 3, No. 7, **(2017)**, http://dx.doi.org/10.1016/j.trecan.2017.05.003

Pérez-Herrero E & Fernández-Medarde A, **Advanced targeted therapies in cancer: Drug nanocarriers, the future of chemotherapy**, Eur J Pharm Biopharm. ,93:52-79., **(2015)** doi: 10.1016/j.ejpb.2015.03.018

Qian J, Hassanein M, Hoeksema M, Harris B, Zou Y, Chen H, Lu P, Eisenberg R, Wang J, Espinosa A, Ji X, Harris F, Rahman J & Massion P, **The RNA binding protein FXR1 is a new driver in the 3q26-29 amplicon and predicts poor prognosis in human cancers**, PNAS, vol. 112, no. 11, 3469–3474, **(2015)**

Roviello F, Rossi S, Marrelli D, De Manzoni G, Pedrazzani C, Morgagni P & Corso G, **Perforated gastric carcinoma: a report of 10 cases and review of the literature**, World journal of surgical oncology, 4:19, **(2006)**, doi: 10.1186/1477-7819-4-19

Siegel R, Miller K & Jemal A, **Cancer Statistics, 2015**, CA Cancer J Clin., 65(1):5-29, **(2015)** doi: 10.3322/caac.21254

Szasz M, Lanczky A, Nagy A, Forster S, Hark K, Green J, Boussioutas A, Busuttil R, Szabo A & Gyorffy B, **Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients**, Oncotarget, Advance Publications, **(2016)**

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou K, Kuhn M, Bork P, Jensen L & von Mering C, **STRING v10: protein-protein interaction networks, integrated over the tree of life**, Nucleic Acids Res, 43 (Database issue): D447–52, **(2015)**, PMC 4383874, PMID 25352553, doi:10.1093/nar/gku1003

Szklarczyk D, Morris J, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva N, Roth A, Bork P, Jensen L & von Mering C, **The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible**, Nucleic Acids Res, 45, **(2017)**

Tomczak K, Czerwińska P & Wiznerowicz M, **The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge**, Contemp Oncol (Pozn), 19(1A):A68-77, **(2015)**, doi: 10.5114/wo.2014.47136

Torre L, Siegel R, Ward E, Jemal A, **Global Cancer Incidence and Mortality Rates and Trends--An Update, Cancer Epidemiol Biomarkers Prev.**, 25(1):16-27., **(2016)** doi:10.1158/1055-9965.EPI-15-0578

Varlinskaya E, Vetter-O'Hagen C & Spear L, **Puberty and gonadal hormones: role in adolescent-typical behavioral alterations**, Horm Behav., 64(2):343-9, **(2013)**, doi: 10.1016/j.yhbeh.2012.11.012

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P & Snel B., **STRING: a database of predicted functional associations between proteins**, Nucleic Acids Res, 1;31(1):258-61, **(2003)**

Wang B & Xi Y, **Challenges for MicroRNA Microarray Data Analysis**, Microarrays (Basel), 2(2), **(2013)**, doi: 10.3390/microarrays2020034

Wang H, Naghavi M, Allen C, Barber R, Bhutta Z, Lopez A & Murray C, **Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015**. Lancet., 388(10053):1459-1544, **(2016)**, doi: 10.1016/S0140-6736(16)31012-1

Wang Z, Gerstein M & Snyder M, **RNA-Seq: a revolutionary tool for transcriptomics**, Nature Genetics Review, 10(1):57-63, **(2009)**, doi:10.1038/nrg2484

Wei I, Shi Y, Jiang H, Kumar-Sinha C & Chinnaiyan A, **RNA-Seq accurately identifies cancer biomarker signatures to distinguish tissue of origin**, Neoplasia, 16(11):918-27, **(2014)** doi: 10.1016/j.neo.2014.09.007

Zhang J, **Evolution by gene duplication: an update**, Trends in Ecology & Evolution, 18(6): 292–8, **(2003)**

Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, Thierry-Mieg D, Wang J, Fischer M, **Comparison of RNA-seq and microarray-based models for clinical endpoint prediction**, Genome Biol.,16:133, **(2015)**, doi: 10.1186/s13059-015-0694-1

Zheng M, Tang L, Huang L, Ding H, Liao W, Zeng M & Wang H, **Overexpression of karyopherin-2 in epithelial ovarian cancer and correlation with poor prognosis**, Obstet Gynecol, 116(4):884-91, (2010), doi: 10.1097/AOG.0b013e3181f104ce.

Zong Z, Zhou T, Rao L, Jiang Z, Li Y, Hou Z, Yang B, Han F & Chen S, **Musashi2 as a novel predictive biomarker for liver metastasis and poor prognosis in colorectal cancer**, Cancer Med, 5(4):623-30, **(2016)**, doi: 10.1002/cam4.624.