

**POLITECNICO DI MILANO**

School of Industrial and Information Engineering

Master of Science in Computer Science and Engineering  
(Bioinformatics and E-health)

Department of Electronics, Information and Bioengineering



# **PREDICTIVE MODELING OF GENE EXPRESSION REGULATION IN OVARIAN CANCER**

Supervisor: Prof. Marco Masseroli

Co-Supervisors: Dr. Maddalena Fratelli (Pharmacogenomic Unit @ 'Mario Negri' Institute)  
Prof. Matteo Matteucci

Master Graduation Thesis by:

Chiara Regondi

ID: 852458

Academic Year 2017 – 2018



*<< 24 hours. 1 440 minutes. 86 400 seconds.*

*That's all it takes to change your life. [...]*

*One single day can fill us with more possibilities that we could imagine. >>*

Owen

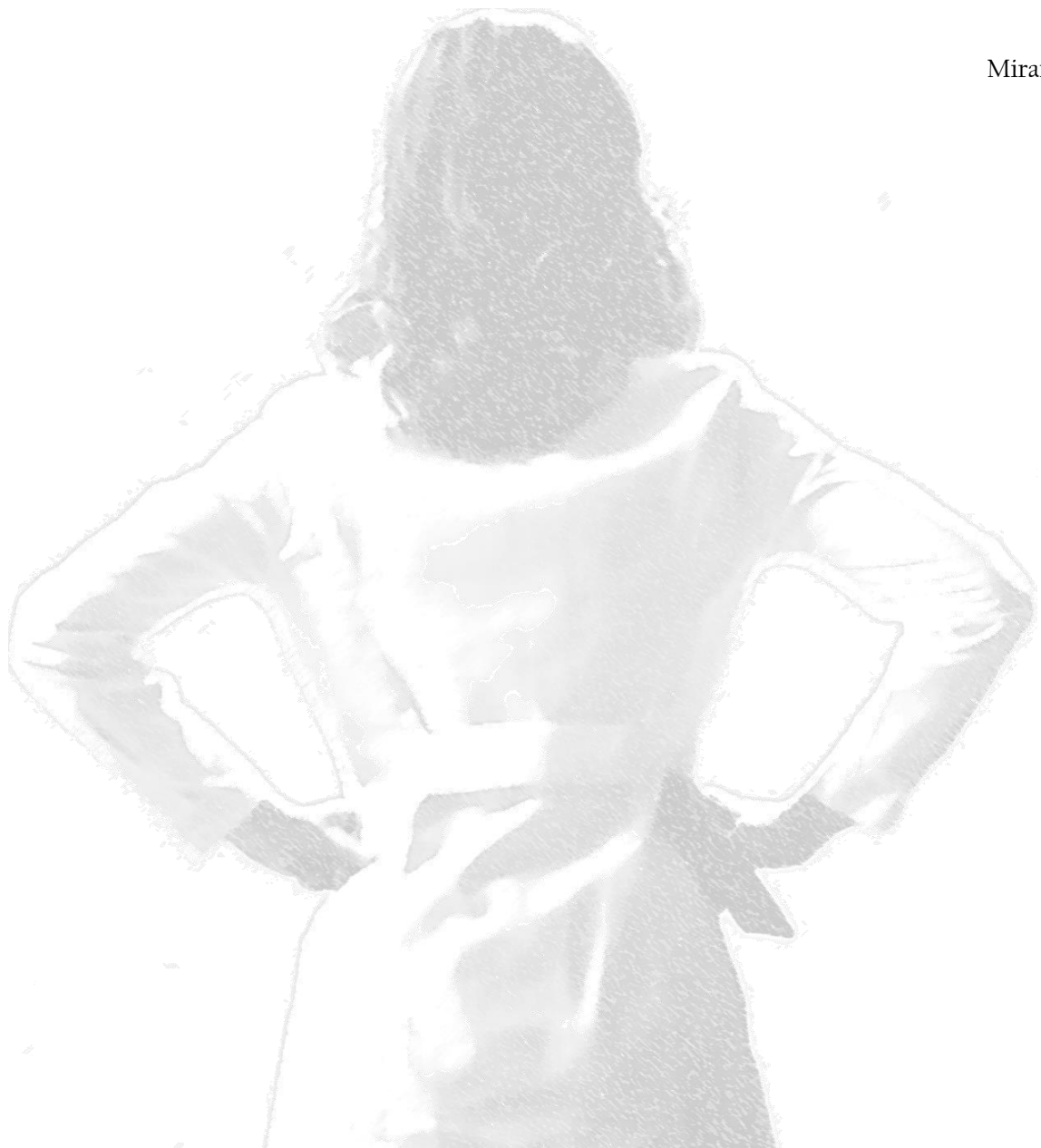




## To my parents

*« You know as well as I do it's not about what you look like,  
or your job, or how successful you are.  
It's about having people in your life that you love and who love you.  
That's all that matters. »*

Miranda



---

*« There's a [scientific study](#) that shows that if you stand like this, in superhero pose, for just five minutes before a job interview or a big presentation or a really hard task, you will not only feel more confident, you will perform measurably better. »*

Amelia



*« Today's the day my life begins. [...].  
Today I become a grown up.  
Today I become accountable to someone other than myself and my parents.  
Accountable for more than my grades.  
Today, I become accountable to the world.  
To the future.  
To all the possibilities that life has to offer.  
Starting today, my job is to show up wide eyed and willing and ready.  
For what, I don't know.  
For anything. For everything.  
To take on life.  
To take on love.  
To take on the responsibility and possibility.  
Today, my friends, our lives begin. And I for one can't wait. »*

Becca



# Abstract

Genomics is the study of all the elements composing the genomic material within an organism. The genes and their role on the expression of human traits are one of the main focuses of modern medical research. Alteration of gene expression and of its regulation is associated with disease, including cancer.

Gene expression is the process by which information encoded in a gene is interpreted and used for synthesizing a functional gene product. Regulation of gene expression is quite a complex process, involving multiple participating factors with a different impact, such as specific regulatory genes that encode for transcription factors, and epigenetic modifications like DNA methylation.

This thesis aims at inferring gene regulation networks in ovarian cancer patients, by building a predictive model for the regulation of the expression of specific target genes belonging to relevant pathways for the ovarian tumor, on the basis of their methylation and expression values, and of the expression of genes encoding for transcription factors with binding sites located in the target gene promoters.

From a computational standpoint, multiple linear regression models are built for each gene of interest, according to an incremental approach that progressively analyze all the potential regulatory features of interest.

Results are validated using other relevant and already known computational methods and a set of samples extracted from basal-like breast cancer, a tumor biomolecularly equivalent to ovarian cancer.

Thus, with this project it is possible to describe ovarian cancer related gene regulation systems, by identifying not only the main biological relationships between a gene and its already known regulators, but also additional possible associations, which may unveil still unknown and potentially interesting biological connections.



# Sommario

La genomica è lo studio di tutti gli elementi che costituiscono il materiale genetico di un organismo. I geni e il ruolo che hanno sull'espressione dei tratti umani rappresentano uno dei principali punti su cui si concentra la ricerca medica moderna. Proprio l'alterazione dell'espressione di un gene e il suo processo di regolazione risultano associati a malattie, tra cui il cancro.

L'espressione di un gene è il processo attraverso cui l'informazione contenuta nel gene viene interpretata e convertita in una macromolecola funzionale. La regolazione dell'espressione genica è un processo piuttosto complesso che dipende da numerosi fattori, ciascuno avente un impatto regolativo diverso: tra questi vi sono specifici geni regolatori che codificano fattori di trascrizione, ma anche alterazioni epigenetiche, come ad esempio la metilazione del DNA.

Questa tesi si propone di inferire reti per la regolazione dell'espressione genica in pazienti con tumore dell'ovaio, costruendo un modello predittivo per la regolazione dell'espressione di specifici geni target, appartenenti a pathways rilevanti per il tumore ovarico, sulla base dei loro valori di espressione e metilazione, e dell'espressione di altri geni che codificano fattori di trascrizione aventi siti di binding all'interno dei promotori dei geni di interesse.

Dal punto di vista computazionale, per ogni gene di interesse si costruiscono vari modelli di regressione lineare, sulla base di un approccio incrementale che analizza progressivamente tutti i potenziali regolatori di interesse.

I risultati sono infine validati, prima sfruttando altri già noti e validi metodi computazionali, poi su un gruppo di pazienti con cancro al seno di tipo basale, un tumore biomolecolarmente equivalente al tumore dell'ovaio.

Con questo progetto è quindi possibile descrivere al meglio i sistemi di regolazione di geni rilevanti per il tumore dell'ovaio, identificando non solo le relazioni biologiche tra un gene e i suoi già noti regolatori, ma anche nuove associazioni in grado di svelare correlazioni biologiche ancora sconosciute e potenzialmente interessanti.





# Ringraziamenti

*« Gratitude, appreciation, giving thanks. No matter what words you use, they all mean the same thing. Happy. We're supposed to be happy. Grateful for friends, family. Happy just to be alive. Whether we like it or not. Maybe we're not supposed to be happy, maybe gratitude has nothing to do with joy. Maybe being grateful means recognizing what you have for what it is, appreciating small victories. Admiring the struggle it takes simply to be human. Maybe we're thankful for the familiar things we know. And maybe we're thankful for the things we'll never know. At the end of the day, the fact that we have the courage to still be standing is reason enough to celebrate. »*

Meredith

Prima di tutto vorrei ringraziare il mio relatore, il Professor Marco Masseroli, che mi ha sempre seguita, consigliata e sostenuta in questo lungo percorso. È stato davvero un piacere lavorare al suo fianco e sono molto contenta di aver intrapreso questa strada con lui.

Ringrazio tantissimo la Dott.ssa Maddalena Fratelli dell'Istituto di Ricerca "Mario Negri" di Milano, che ho avuto il privilegio di conoscere all'inizio di questo progetto di tesi e con la quale ho avuto la possibilità di collaborare in questi mesi. È stato davvero un onore partecipare in prima persona ad un importante progetto per uno dei più prestigiosi istituti di ricerca d'Italia. Sono stata subito colpita da un tema che in qualche modo si lega ad esperienze personali, dalle quali ho ricevuto la spinta ad approfondire la mia conoscenza in campo biologico e genomico e a dedicarmi alla parte computazionale di questo progetto, che ha ancora possibili molteplici sviluppi futuri, con l'obiettivo finale di aggiungere un altro mattoncino nella costruzione delle cure dei tumori.

Grazie anche al Professor Matteo Matteucci per il suo supporto e i suoi preziosi consigli.

Un ringraziamento va anche alle mie colleghe e amiche Lucia, Matilde e Patrizia, con le quali ho condiviso sin dal primo giorno del primo anno l'impegnativo e bellissimo percorso universitario. Quante ne abbiamo passate insieme! Lezioni, progetti e lunghe giornate di studio, ma anche pause caffè, shopping e tanto tanto divertimento. Grazie anche a Massimo, il mio collega bioinformatico, ma soprattutto un amico, con il quale ho potuto condividere non solo l'iter della laurea magistrale, ma anche la mia immancabile ansia prima di ogni esame.

Grazie a Cesca, Fra e Marco, gli amici di una vita, con i quali ho avuto il piacere di condividere ogni momento di questi cinque anni, dopo altri cinque anni di Marie Curie: eccoci ancora qua, sempre noi, sempre uniti e con un altro traguardo raggiunto assieme!

Ma il più grande ringraziamento va ai miei genitori Elisabetta e Pietro, a mio fratello Enrico e a Giogìò, perché mi hanno supportata in ogni mia decisione, mi hanno consigliata, mi hanno aiutata ad alzarmi quando sono caduta, mi hanno sostenuta quando pensavo di non farcela da sola e perché per me ci sono sempre stati, ogni giorno, incondizionatamente. Senza di loro tutto questo non sarebbe stato possibile.



# Contents

|   |           |
|---|-----------|
| Abstract  | I         |
| Sommario  | III       |
| Ringraziamenti  | V         |
| List of Figures   | XI        |
| List of Tables  | XIII      |
| <b>1 Introduction</b>                                       | <b>1</b>  |
| <b>2 Background</b>   | <b>3</b>  |
| 2.1 Genomics, genome and genes                              | 3         |
| 2.2 Gene expression and its regulation                      | 5         |
| 2.3 Measurement of gene expression: RNA-sequencing          | 6         |
| 2.4 Transcription factors                                   | 8         |
| 2.5 ChIP-sequencing   | 8         |
| 2.6 Epigenetics and DNA methylation                         | 10        |
| <b>3 Goals</b>  | <b>13</b> |
| <b>4 Materials and Tools</b>                                | <b>17</b> |
| 4.1 Ovarian cancer relevant pathways and their genes        | 17        |
| 4.2 Breast cancer PAM50 data samples                        | 21        |
| 4.3 Human gene nomenclature and human transcription factors | 23        |
| 4.3.1 UniProt   | 23        |
| 4.3.2 ENCODE  | 24        |
| 4.3.3 HUGO Gene Nomenclature Committee                      | 24        |
| 4.4 Transcription factor dataset from ENCODE                | 25        |
| 4.5 Methylation and gene expression datasets from TCGA      | 25        |
| 4.6 Genomic annotations from GENCODE                        | 27        |
| 4.7 Computational tools                                     | 28        |
| 4.7.1 GenoMetric Query Language                             | 28        |
| 4.7.2 Python libraries                                      | 30        |
| 4.7.2.1 PyGMQL  | 31        |

|          |         |   |           |
|----------|---------|---|-----------|
|          | 4.7.2.2 | Pandas  | 32        |
|          | 4.7.2.3 | Statsmodel  | 32        |
|          | 4.7.2.4 | Scikit-learn  | 33        |
|          | 4.7.2.5 | Mlxtend   | 33        |
|          | 4.7.2.6 | NetworkX  | 34        |
|          | 4.7.3   | Cytoscape   | 34        |
|          | 4.7.4   | ARACNe  | 34        |
| <b>5</b> |         | <b>Data extraction and manipulation</b>                             | <b>35</b> |
|          | 5.1     | Genes – transcription factors mapping                               | 35        |
|          | 5.2     | Selection of transcription factors                                  | 36        |
|          | 5.3     | Identification of candidate regulatory genes                        | 41        |
|          | 5.4     | Extraction of methylation and gene expression values                | 44        |
|          | 5.4.1   | Ovarian cancer  | 44        |
|          | 5.4.2   | Breast cancer   | 49        |
|          | 5.5     | Data matrix construction  | 52        |
|          | 5.5.1   | M1: genes belonging to the same pathway of the model gene           | 52        |
|          | 5.5.2   | M2: M1 + candidate regulatory genes of the model gene               | 52        |
|          | 5.5.3   | M3: M2 + candidate regulatory genes of the model gene pathway genes | 53        |
|          | 5.5.4   | M4: M3 + genes belonging to other pathways                          | 53        |
|          | 5.5.5   | M5: M4 + candidate regulatory genes of other pathway genes          | 54        |
| <b>6</b> |         | <b>Data analysis</b>  | <b>55</b> |
|          | 6.1     | Feature/gene selection  | 57        |
|          | 6.2     | Linear regression of individual genes on ovarian tumor samples      | 63        |
| <b>7</b> |         | <b>Results and Discussion</b>                                       | <b>69</b> |
|          | 7.1     | Analysis of regression results and networks generation              | 69        |
|          | 7.1.1   | DNA_REPAIR pathway  | 71        |
|          | 7.1.2   | STEM_CELLS pathway  | 77        |
|          | 7.1.3   | GLUCOSE_METABOLISM pathway  | 81        |
|          | 7.2     | Validation  | 88        |
|          | 7.2.1   | Computational validation  | 88        |
|          | 7.2.1.1 | Comparison with the ARACNe processing                               | 90        |
|          | 7.2.1.2 | Models application on basal-like breast cancer data                 | 92        |
|          | 7.2.1.3 | Linear regression of individual genes on breast cancer samples      | 94        |
|          | 7.2.2   | Biological validation   | 96        |
|          | 7.2.2.1 | PCR data  | 96        |
|          | 7.2.2.2 | Literature-confirmed results  | 96        |

|          |   |            |
|----------|---|------------|
| <b>8</b> | <b>Conclusions</b>  | <b>99</b>  |
| <b>9</b> | <b>Future developments</b>  | <b>101</b> |
|          | <b>Bibliography and Webliography</b>  | <b>103</b> |
|          | <b>Appendix A: Python scripts flowcharts</b>                                  | <b>109</b> |
|          | <b>Appendix B : Genetic expression networks from linear regression models</b> | <b>123</b> |



# List of Figures

|      |   |    |
|------|---|----|
| 2.1  | DNA double-helix structure. Taken from [5].   | 4  |
| 2.2  | Full DNA structure. Taken from [6].   | 4  |
| 2.3  | Gene structure. Taken from [2].   | 5  |
| 2.4  | Gene Expression. Taken from [3].  | 6  |
| 2.5  | DNA transcription process. Taken from [9].  | 6  |
| 2.6  | Summary of RNA-Seq. Taken from [11].  | 7  |
| 2.7  | TFs and their potential effects on the regulation of gene activity. Taken from [3].   | 8  |
| 2.8  | Peak calling in correspondence of a TF located in the gene promoter. Taken from [3].  | 9  |
| 2.9  | ChIP-sequencing workflow. Taken from [3].   | 9  |
| 2.10 | Comparison between euchromatin and heterochromatin. Taken from [14].  | 10 |
| 2.11 | Regulation of gene activity: DNA methylation. Taken from [3].   | 11 |
| 2.12 | Global changes in DNA methylation in both normal and cancer cells. In normal cells, CpG islands in active promoters are not methylated, thus allowing transcriptional activation. CpG islands within coding regions are often methylated. Reverse patterns are observed in cancer cells. Taken from [16]. | 12 |
| 3.1  | Sample network visualizing linear regression results.   | 15 |
| 4.1  | UniProt structure. Taken from [23].   | 23 |
| 4.2  | A graphical representation of the ENCODE information. Taken from [25].  | 24 |
| 4.3  | TCGA statistics. Taken from [28].   | 26 |
| 4.4  | How a TCGA aliquot barcode can be broken down into its components. Taken from [29].   | 27 |
| 4.5  | Graphical representation of genomic region coordinates. Taken from [3].   | 29 |
| 4.6  | Excerpt of the GMQL Web interface. Taken form [43].   | 30 |
| 4.7  | A GMQL query to be executed directly on the system through the Web interface (a) and the corresponding query written in Python, to be executed remotely through the PyGMQL library (b).   | 31 |
| 4.8  | Example of how data are organized in Pandas. A Pandas table is commonly called <i>Dataframe</i> .   | 33 |
| 5.1  | Diagram of the GMQL query for the extraction of TFs binding to human genes promoters.   | 37 |
| 5.2  | COVER(1,ANY) example.   | 39 |
| 5.3  | Description of the genomic coordinates system and computation of gene promoters.  | 39 |
| 5.4  | Example of the GMQL MAP operation between two datasets containing one single data sample each.  | 40 |
| 5.5  | Python materialization of the RES dataframe and structure of the GeneTF_df dataframe.   | 40 |
| 5.6  | Main loop used for extracting each target gene list of TFs from the initial dataset.  | 41 |
| 5.7  | Main loop used for identifying candidate regulatory genes for each gene of interest, starting from the list of TFs previously extracted.  | 41 |
| 5.8  | Flowchart for the extraction of candidate regulatory genes of the genes of interest.  | 43 |

|      |   |    |
|------|---|----|
| 5.9  | Diagram of the GMQL query for the extraction of OV tumor data.  | 45 |
| 5.10 | Example of methylation sites (red circles) considered in the analysis.  | 46 |
| 5.11 | Diagram of the GMQL query for the extraction of methyl_areas.   | 47 |
| 5.12 | Diagram of the GMQL query for the extraction of BRCA tumor data.  | 50 |
| 5.13 | Python script for converting BRCA PAM50 Basal set of samples into a GMQL dataset.                                     | 51 |
| 6.1  | Initialization step in the Python scripts for performing features selection and linear regression.                    | 56 |
| 6.2  | Full set of operations performed on each gene of interest during the course of the project.                           | 57 |
| 6.3  | Feature selection for matrix M2 (a), M3 (b) and M5 (c).   | 60 |
| 6.4  | Feature selection process, repeated five times for each target gene and for each selected data matrix.                | 61 |
| 6.5  | Flowchart of the feature selection script.  | 62 |
| 6.6  | Z-score normalization (or standard normalization) in scikit-learn.  | 63 |
| 6.7  | Ordinary least squares regression in Statsmodel.  | 64 |
| 6.8  | Linear regression summary statistics in Statsmodel.   | 64 |
| 6.9  | Example of linear regression for matrix M3 of gene TKT and identification of relevant features.                       | 65 |
| 6.10 | Flowchart of the linear regression script.  | 66 |
| 7.1  | Expression networks from linear regression models M3 (a) and M5 (b) of the DSB subclass.                              | 76 |
| 7.2  | Expression networks from linear regression models M3 (a) and M5 (b) of the Cancer Therapeutic Targets subclass.       | 82 |
| 7.3  | Expression networks from linear regression models M3 (a) and M5 (b) of the Regulation of Glucose Metabolism subclass. | 87 |
| 7.4  | Example of different mutual information thresholds in ARACNe.   | 89 |
| 7.5  | Validation of ovarian cancer regression models based on the ARACNe processing.  | 91 |
| 7.6  | Ovarian cancer regression models application on basal-like breast cancer data: complete workflow.                     | 93 |
| 7.7  | ERCC1 – ERCC2 correlation in Triple Negative breast cancer and in the breast cancer Luminal subtype.                  | 97 |



# List of Tables

|      |  |    |
|------|--|----|
| 4.1  | List of genes of interest in the STEM CELLS pathway.   | 18 |
| 4.2  | List of genes of interest in the GLUCOSE_METABOLISM pathway.   | 19 |
| 4.3  | List of genes of interest in the DNA_REPAIR pathway.   | 20 |
| 4.4  | A subset of samples contained in the BRCA_PAM50 file.  | 22 |
| 5.1  | Excerpt of the final genes - TFs mapping table (GenesMapping.xlsx).  | 36 |
| 5.2  | Excerpt of the summary table with final results of TFs and candidate regulatory genes extraction phases (Full TFs-RegulatoryGenes SUMMARY Table.xlsx). | 42 |
| 5.3  | Excerpt of the final tables containing data about DNA methylation and gene expression for each TCGA aliquot under analysis.                            | 48 |
| 5.4  | Structure of the data matrixes for the data analysis process.  | 53 |
| 6.1  | Data analysis: execution times.  | 67 |
| 7.1  | DNA_REPAIR genes with either M3 or M5 model score higher than the 0.6 threshold.   | 72 |
| 7.2  | Model M5 best DNA_REPAIR genes and their features.   | 74 |
| 7.3  | STEM_CELLS genes with either M3 or M5 model score higher than the 0.6 threshold.   | 78 |
| 7.4  | Model M5 best STEM_CELLS genes and their features.   | 80 |
| 7.5  | GLUCOSE_METABOLISM genes with either M3 or M5 model score higher than the 0.6 threshold.   | 83 |
| 7.6  | Model M5 best GLUCOSE_METABOLISM genes and their features.   | 85 |
| 7.7  | Number of relevant features selected for each genetic pathway: target genes in the pathway (a) and regulatory genes (b).                               | 88 |
| 7.8  | Validation of ovarian cancer regression models based on the ARACNe processing: comparison results.   | 91 |
| 7.9  | Validation of ovarian cancer regression models based on the ARACNe processing: detailed results for a sample set of genes.                             | 92 |
| 7.10 | Ovarian cancer regression models application on basal-like breast cancer data: excerpt of the results comparison for model M5.                         | 94 |
| 7.11 | Ovarian - Breast cancer regression models comparison results.  | 95 |
| 7.12 | Ovarian - Breast cancer regression models comparison: detailed results for a sample set of genes.  | 95 |
| 7.13 | Correlation among the expression of the genes studied in the ovarian PDXs and comparison with ovarian M2 model.  | 97 |



# 1. Introduction

*<< Sometimes, the key to making progress is to recognize how to take that very first step.  
Then you start your journey. You hope for the best and you stick with it, day in, day out. >>*

Meredith

Carcinogenesis is the process that leads to the formation of cancer, where normal cells are transformed into cancer cells. This process is strictly dependent not only on gene mutations, but also on alterations in the gene expression, i.e., the activity of genes.

Some proteins, known as transcription factors (TFs), together with other multiple genetic and epigenetic influences, such as DNA methylation, are fundamental factors in regulating this activity. As a consequence, a misregulation of these TFs or of other regulatory elements may lead to the acquisition of specific tumor-related properties.

However, the way in which the gene expression is regulated and controlled within tumors is complex and related knowledge is limited. An example is represented by an extremely interesting tumor that is unfortunately the most common cause of death in women with gynecologic malignancies, the *Ovarian Serous Cystadenocarcinoma* (OV), a particularly aggressive type of epithelial ovarian cancer.

This thesis focuses on this specific tumor because of its still limited knowledge and a very poor related prognosis, with the objective of deeply understanding how regulation process works, in order to increase knowledge and hopefully improve cancer therapies.

In this project we analyze and describe the regulation systems of a set of tumor-specific target genes, using heterogeneous data that comprise both DNA methylation and candidate TFs, which may be responsible of defining tumor-specific gene expression profiles that ease tumor development and progression.

This experimental work develops an analytic, statistical and computational method that mainly aims at inferencing gene expression regulation in cancers, integrating heterogenous information from multiple sources.

In chapters 2, 3 and 4 we better explain the context in which this thesis is developed, defining the main biological concepts that are necessary for a correct interpretation and deep understanding of this work, carefully illustrating the main goals of the project and describing the materials and the tools used to perform the analysis. Next, chapters 5, 6 and 7 comprise the detailed description of the actual work, i.e., the data extraction process and the data analysis, as well as the discussion and validation of the results. Finally, In chapter 8 and 9 we present the main conclusions of the work and its related future potential developments.



## 2. Background

« *Biology determines much of the way we live. From the moment we are born, we know how to breathe and eat.*

*As we grow older, new instincts kick in. We become territorial. We learn to compete. We seek shelter. [...]*

*Biology says that we are who we are from birth. That our DNA is set in stone. »*

Meredith

### 2.1 Genomics, genome and genes

Genomics [1, 2, 3] is a branch of molecular biology that is related to the structure, function and evolution of genomes, whose main goal is the study of all the elements composing the genetic material within an organism. The genome is commonly defined as the total amount of hereditary information characterizing an organism and it is exactly identical in all its cells, with the exception of germ cells.

The main chemical structure that is studied in genomics is the DNA (Deoxyribonucleic Acid) [4], where the hereditary information is encoded: DNA is a molecule containing all the information needed for the growth, development, functioning and reproduction and, in general, for supporting the life of all the living organisms.

DNA is the biggest macromolecule in the cell and it is contained in every cell of the organism. It is a polymer composed by four different types of monomers, called *nucleotides*: each nucleotide is characterized by a sugar (i.e., deoxyribose) with five atoms of carbon, linked to a phosphate group and a nitrogenous base (Adenine, Timine, Cytosine or Guanin).

DNA, however, doesn't appear as a single chain of nucleotides: it has a specific 3D structure composed by two twisting anti-parallel (i.e., that run in opposite directions) and complementary strands, paired together thanks to specific bonds between nitrogenous bases of different nucleotides composing the strands (A bonds with T, while C bonds with G). In particular, the two nucleotide chains roll-up together in a right-handed coil, such that the nucleotide bases are arranged in the internal part of the helix, while the sugar-phosphate backbone constitute the external part (*Figure 2.1*).

The information is encoded in the DNA by the order in which these bases are located in the molecule. Then, it wraps twice around a group of eight histones (small nuclear proteins) to form a specific structural unit, called *nucleosome*. Nucleosomes are packed together to form *chromatin* that, in turn, generates *chromosomes* within the nucleus of the eukaryotic cells, as illustrated in *Figure 2.2*.

Chromosomes are usually paired together and the total number of chromosomes is peculiar of each species (e.g., the humans have 23 pairs of chromosomes).

In order to simply make these biological concepts clear, a metaphor can be used, comparing the human genome to the instructions stored in a cookbook: just as a cookbook gives the instructions needed to prepare a range of meals, the human genome contains all the instructions that are needed to make the full range of human cell types, including muscle cells and neurons. In case of the human organisms,

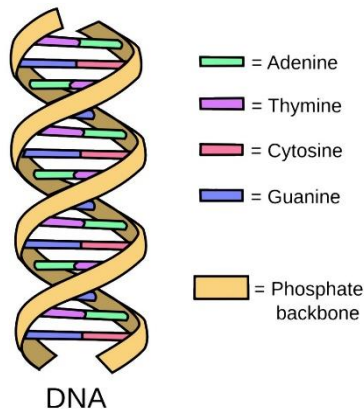


Figure 2.1: DNA double-helix structure. Taken from [5].

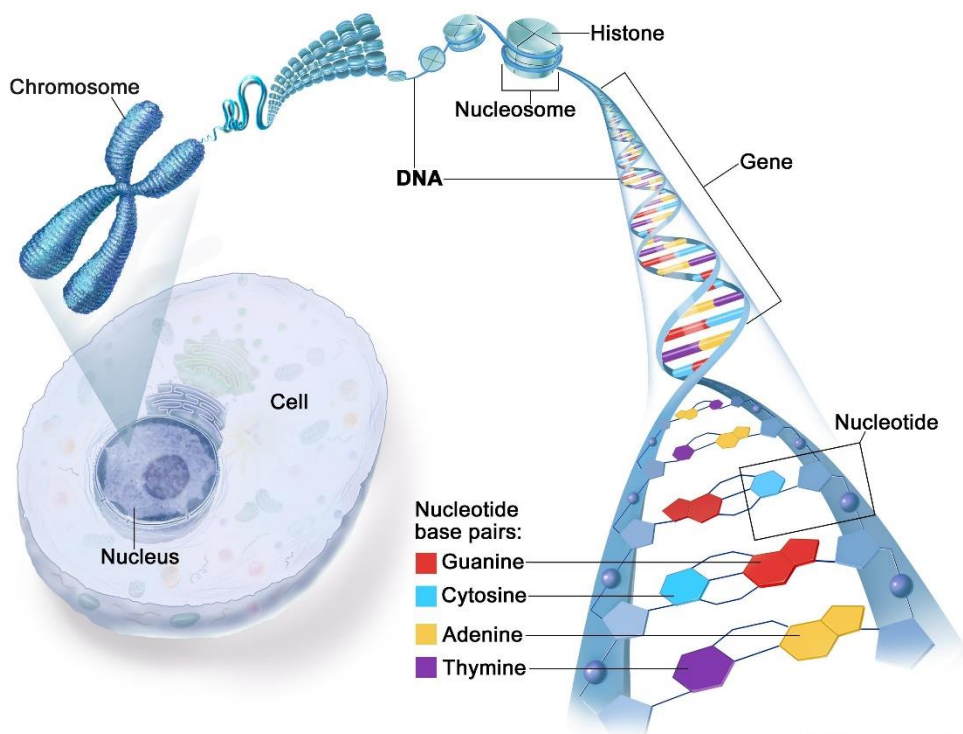


Figure 2.2: Full DNA structure. Taken from [6].

the book (i.e., the genome) contains 23 chapters (i.e., chromosomes) and each chapter contains from 48 to 250 million letters (A,C,G or T) without spaces. In total, the book contains over 3.2 billion letters and approximately 20,000 different recipes (i.e., the genes).

A gene (Figure 2.3) is a segment of DNA, i.e., a specific region of the genome, that carries the information used for synthesizing one or more proteins. Proteins are large molecules carrying out a lot of different functions within the organisms and they are one of the most important molecular structures for living beings. The genes and their role on the expression of human traits are one of the main focuses of modern medical research, mainly because the alteration of the gene expression and of its regulation has been proved to be associated with disease, including cancer.

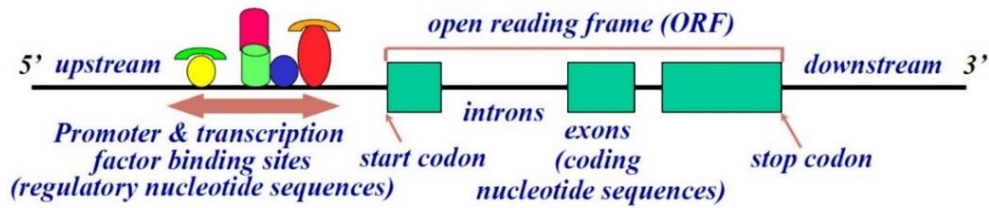


Figure 2.3: Gene structure. Taken from [2].

The study of tumors is one of the main branches of genomics, which is nowadays mainly taking advantages of the new possibilities provided by new advanced digital technologies regarding Big Data, Artificial Intelligence and machine learning algorithms, and *Next Generation Sequencing* (NGS) [7]. Contributing to the development of these technologies it is the *Human Genome Project*, an international scientific research project carried out from 1990 to 2003 with the goal of determining the sequence of nucleotide base pairs that make up human DNA and of identifying and mapping all the genes of the human genome, from both a physical and a functional standpoint.

The project was a great success and it led to the development of the new sequencing paradigm of NGS, which is changing both biological research and medical practice, thanks to a set of new technologies enabling a high-throughput, high-precision, time-limited and low-cost sequencing process, making the DNA mapping a standardized process.

## 2.2. Gene expression and its regulation

Gene expression is the process by which information encoded in a gene is interpreted and used for synthesizing a functional gene product [8]. A schematic representation of this whole process is shown in *Figure 2.4*. The gene expression process comprises two main steps:

1. TRANSCRIPTION, where a particular segment of DNA is converted into a segment of mRNA by a specific enzyme, called *RNA polymerase* (*Figure 2.5*);
2. TRANSLATION, where the mRNA is translated to produce an amino acid chain that then folds into an active protein performing its functions in the cell.

Since not all the genes are active (i.e., expressed) at the same time or in the same cells, the process of the gene expression regulation allows the cell to decide which groups of genes to express according to the specific context it lives in, in order to increase or decrease the production of gene products, such as proteins or RNA, and thus to respond to different needs and external requirements. If a gene produces mRNA, it is said to be “ON” (i.e., expressed), otherwise it is “OFF”.

However, the regulation of the gene expression is quite a complex process, because it involves multiple participating factors having a different impact on the regulation process itself, such as transcription factors or DNA methylation, which are explained in the next paragraphs.

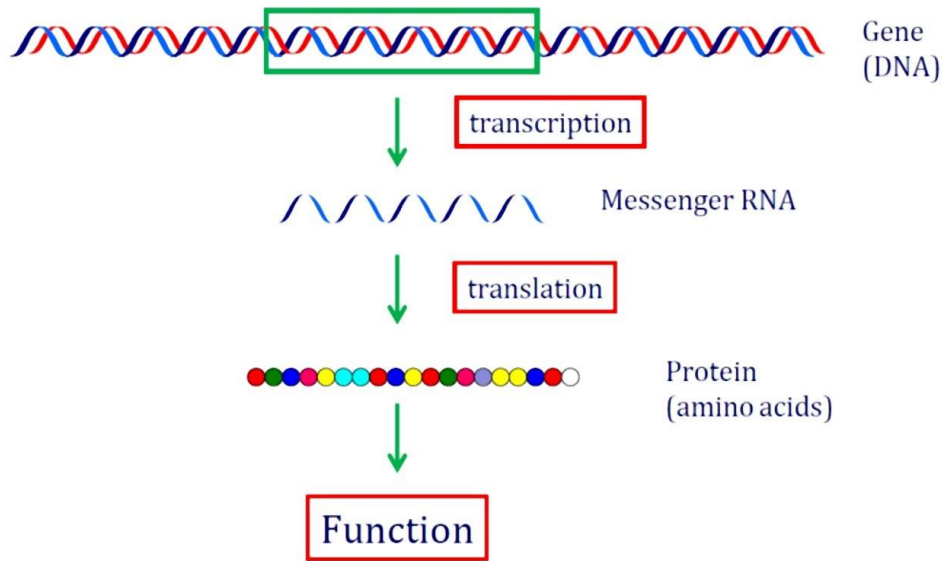


Figure 2.4: Gene Expression. Taken from [3].

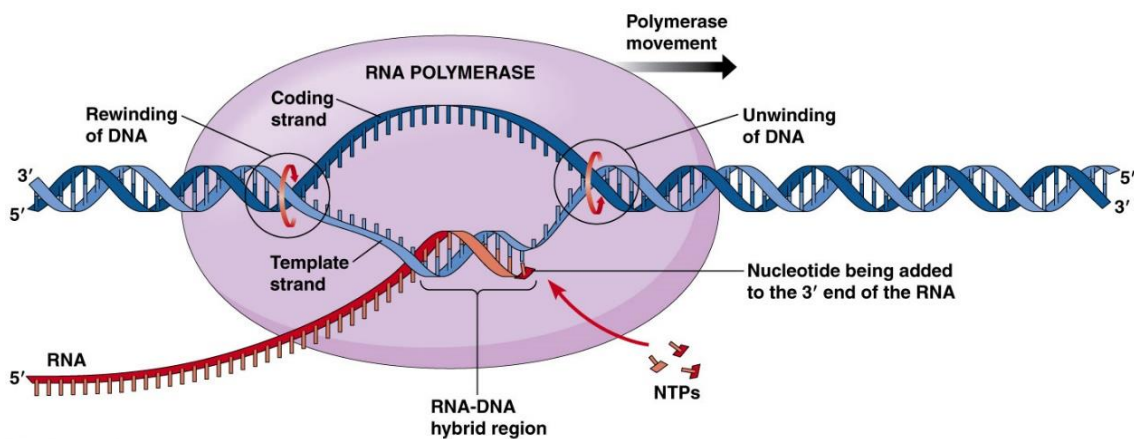


Figure 2.5: DNA transcription process. Taken from [9].

### 2.3. Measurement of gene expression: RNA-sequencing

Measuring the gene expression means quantifying the level at which a particular gene is expressed within a specific cell, tissue or organism, i.e., measuring the gene activity in particular conditions.

Besides traditional DNA microarray analysis, which uses DNA spots attached to a solid surface to measure the expression levels of large numbers of genes simultaneously, NGS uses the *RNA-sequencing* process to measure gene expression [10].

RNA-sequencing allows to quantify the amount of RNA that is present in a biological sample at a given time, by analyzing all the gene transcripts. This is the usual process to analyze the continuously



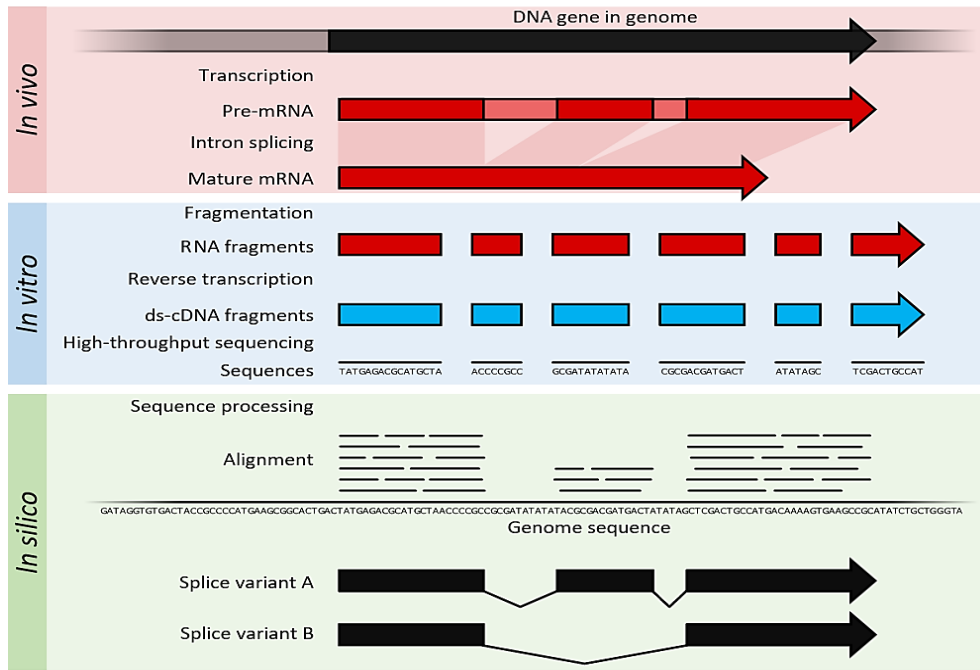


Figure 2.6: Summary of RNA-seq. Taken from [11].

changing cellular transcriptome (i.e., the set of all RNA molecules contained in a single cell of an organism), allowing to look at all the transcripts resulting from alternative splicing, genetic mutation and changes in the gene expression over time.

Figure 2.6 illustrates the RNA-seq process: mRNA is extracted from the organism, fragmented and copied into stable complementary DNA (cDNA), which is sequenced using high-throughput and short-read sequencing methods. These sequences can then be aligned to a reference genome sequence to reconstruct the genome regions that were being transcribed. Basically in RNA-sequencing, the RNA is fragmented, DNA is synthesized complementary to RNA fragments and then it is amplified to form a cluster that is finally sequenced.

A “read” is the sequence of a cluster obtained at the end of the sequencing. More precisely, a read can be defined as an inferred sequence of base pairs corresponding to all or to a part of a DNA fragment. We can say that when a set of fragments, derived from DNA fragmentation, is sequenced, then it produces a set of reads.

So, RNA-seq analysis quantifies protein-coding genes expression on the basis of the reads aligned to each genes. The common unit of measurement for the amount of the gene expression as a result of the RNA-seq process is the *Fragments Per Kilobase per Million reads* (FPKM), computed after aligning the reads to the reference genome and quantifying the mapped reads:

$$FPKM = \frac{RC_g \cdot 10^9}{RC_{pc} \cdot L}$$

where:  $RC_g$  = number of reads mapped to the gene

$RC_{pc}$  = total number of reads mapped to all protein-coding genes

$L$  = gene length (in base pairs), calculated as the sum of the length of all the exons in the gene (i.e., the actual regions in the gene encoding information)

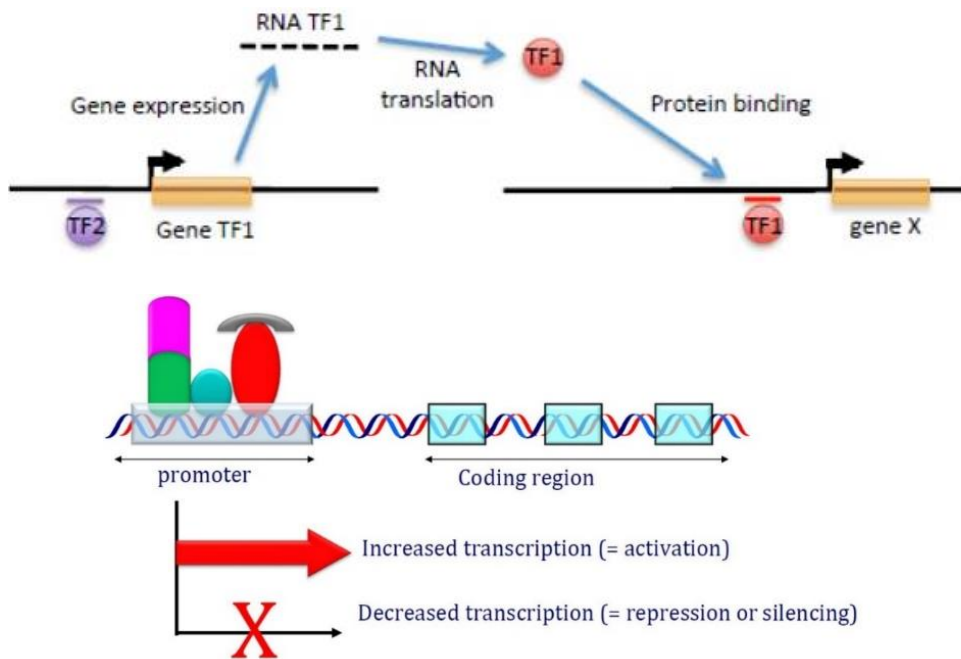


Figure 2.7: TFs and their potential effects on the regulation of gene activity. Taken from [3].

## 2.4. Transcription factors

A Transcription Factor (TF) [12] is a specific protein that binds to DNA in a specific region of a promoter which has the ability to control the transcription process and, as a consequence, it may have a main role in regulating the level of gene expression, by turning on and off specific genes (Figure 2.7).

A promoter is the region of DNA initiating the transcription of a particular gene, which is conventionally defined as that area around the gene *Transcription Start Site* (TSS), with genomic coordinates from -2k to +1k base pairs from the TSS itself.

Genes that encode for transcription factors can be referred to as “regulatory genes”, because they may be directly involved in the regulation of the expression of other genes.

One of the main techniques used for mapping and identifying all the transcription factors binding sites is the ChIP-sequencing, which is the one used in this work.

## 2.5. ChIP-sequencing

ChIP-sequencing is a method used to analyze protein interaction with the DNA. Due to its ability to rapidly decode millions of DNA fragments simultaneously, with high efficiency and relatively low cost, it is nowadays the most popular and commonly used ChIP variation method.

ChIP-seq is a very powerful technique which combines the traditional *chromatin immunoprecipitation* (ChIP) techniques for investigating the protein-DNA interaction in the cell, with parallel DNA sequencing, in order to identify genome-wide binding sites of the DNA-associated proteins. In particular, ChIP-seq is primarily used to determine how transcription factors and other proteins interact with DNA to regulate the gene expression [13].

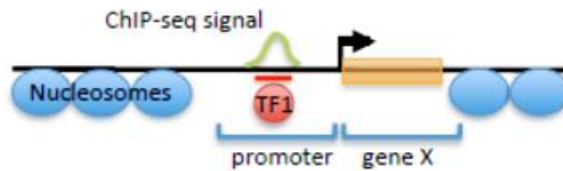


Figure 2.8: Peak calling in correspondence of a TF located in the gene promoter. Taken from [3].

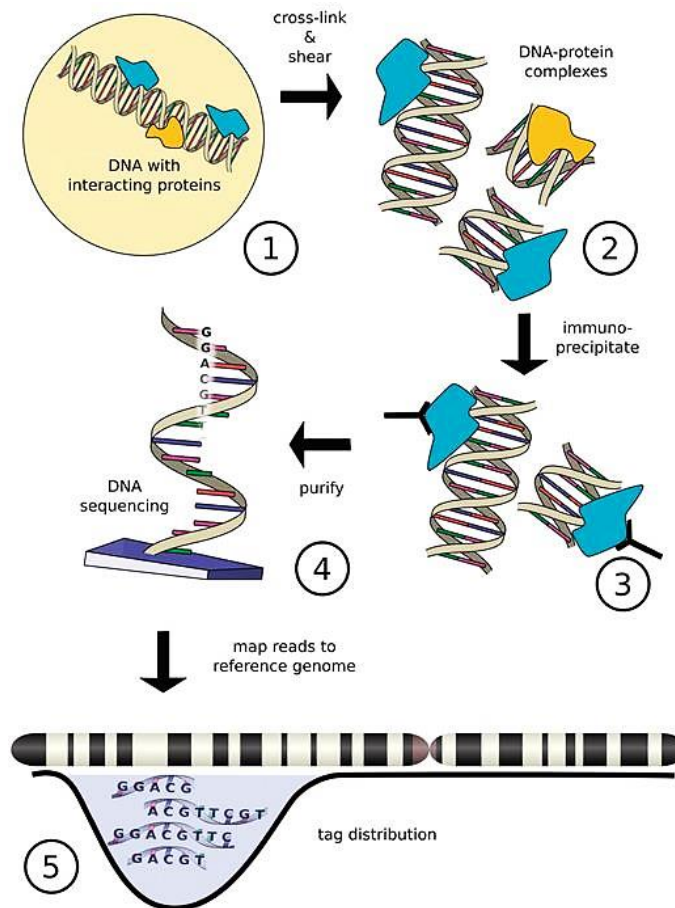


Figure 2.9: ChIP-seq workflow. Taken from [3].

The computational method used by ChIP-seq to identify these binding sites is the so called *peak calling*. Its role is to infer the actual binding loci from the positional distribution of tags, i.e., sequenced DNA fragments mapped onto a reference genome sequence. These areas of the genome, enriched with aligned reads following a ChIP-seq experiment and identified by the peak calling method, are the ones where a protein interacts with the DNA.

Figure 2.8 shows the peak in correspondence of a transcription factor.

Clearly, different data types have different peak shapes. In particular, ChIP-seq analysis algorithms are specific for identifying two possible types of enrichment: *BROAD* peaks (regions of signal enrichments, i.e., histone modifications that cover entire gene bodies) or *NARROW* peaks (peaks of signal enrichments, i.e., a transcription factor bound to an enhancer). Figure 2.9 displays the complete workflow of a ChIP-seq analysis.

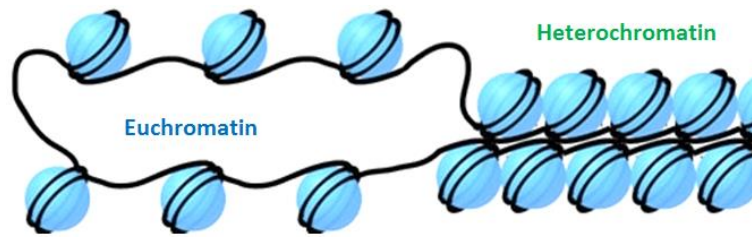


Figure 2.10: Comparison between euchromatin and heterochromatin. Taken from [14].

## 2.6. Epigenetics and DNA methylation

In addition to transcription factors, the gene activity may be also determined by epigenetic factors, leading to epigenetic modifications.

Epigenetics is the study of heritable changes in gene expression that do not involve any structural alteration of the underlying DNA sequence; basically, epigenetic factors are responsible for changes in the phenotype without a related change in the genotype.

The epigenome regulates the expression of the genome by deciding which genes to activate, in which cell and in which context, according to external stimuli coming from environment, physical activity, lifestyle, stress and other conditions generating a genetic expression signal.

Genetic mutations directly alter the DNA changing its nucleotide sequence: in some cases, changing the genetic code means changing the final gene products, and so the results of transcription and translation steps.

Epigenetic mutations do not alter the DNA from a structural standpoint: they do not change its sequence, they simply influence it, either promoting or suppressing transcription.

The remaining part of the paragraph explains how this works in deeper details. DNA wraps around nucleosomes to form chromatin, a complex of macromolecules, consisting of DNA, protein and RNA. Chromatin has a key function in the process of gene expression: in particular, the so called “euchromatin” (a lightly packed and relaxed form of chromatin) promotes the binding between DNA and transcription factors and facilitates transcription, while “heterochromatin” (a tightly packed and condensed form of chromatin), because of its “closed” structure, is barely accessible to polymerases and therefore hard to be transcribed (a comparison between the two types of chromatin is reported in *Figure 2.10*).

Epigenetic mutations are able to change the chromatin structure, either generating euchromatin and promoting gene transcription, or producing heterochromatin and repressing gene transcription. One of the main epigenetic mechanisms regulating gene expression is the DNA methylation, detailed in *Figure 2.11*.

DNA methylation [15] is a biochemical process which consists in the addition of a methyl group ( $\text{CH}_3$ ) to a nitrogenous base of the DNA: in particular, this binding occurs in cytosines of CpG sites, those regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence of bases.

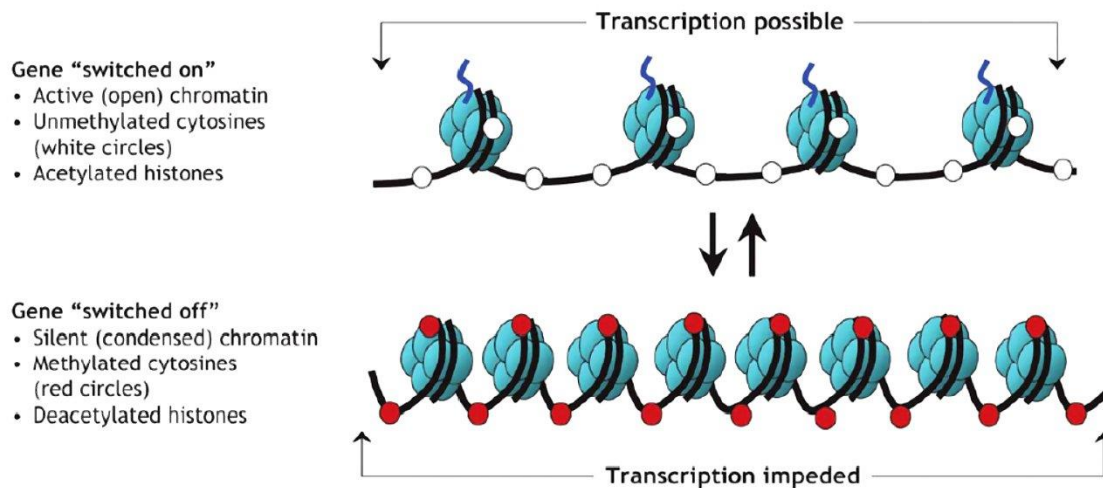


Figure 2.11: Regulation of gene activity: DNA methylation. Taken from [3].

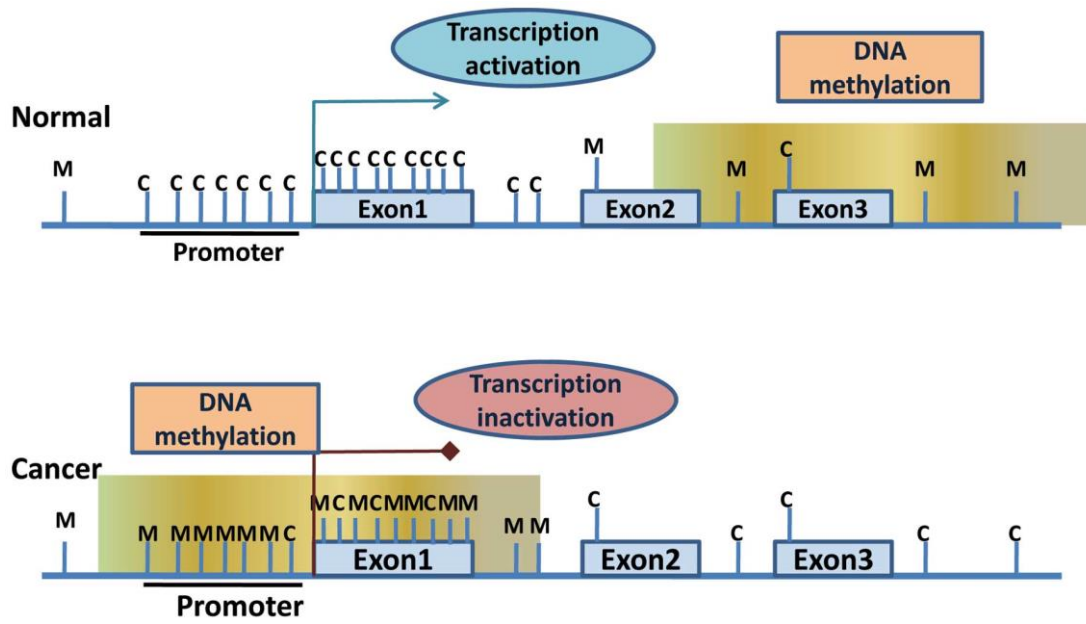
DNA methylation is one of the main events responsible for carcinogenesis [16]. Figure 2.12 illustrates DNA methylation effects in both normal and cancer cells.

CpG sites are involved in the methylation process and regions with a high frequency of CpG sites are called "CpG Islands". These islands are usually not methylated and they are mainly located near the transcription start site (TSS) or inside the gene itself. Their methylation may have different effects on the gene expression: if methylation occurs in the gene promoter, then the gene expression is surely reduced (i.e., hypermethylation is strongly related to down-regulation of the gene), while if it involves islands located inside the gene, then transcription may be promoted and the gene expression potentially increases.

However, this is only one side of the story, since methylation is not the only feature acting on the gene expression regulation. Methylation may either increase or decrease gene expression, depending on the location of methylated cytosines, although this expression-methylation correlation may not be found.

As an example, the discovery of reduced levels of DNA methylation in tumors is strongly associated with the identification of hypermethylated regions in tumor suppressor genes. A tumor suppressor gene is a gene that protects a cell from potential uncontrolled growth leading to cancer; when this gene mutates, with a loss or a reduction of its functionality, the cell is exposed and it can progress to cancer, usually in combination with other genetic changes. This repression of tumor suppressor genes can happen not only because of mutations, but also through DNA methylation.

Thus, DNA methylation causing carcinogenesis can act either directly in the tumor-related gene promoters, or indirectly by suppressing inhibition of oncogenes (i.e., those genes having the potential to cause cancer).



*Figure 2.12:* Global changes in DNA methylation in both normal and cancer cells. In normal cells, CpG islands in active promoters are not methylated, thus allowing transcriptional activation. CpG islands within coding regions are often methylated. Reverse patterns are observed in cancer cells. Taken from [16].

### 3. Goals

« Make a plan. Set a goal. Work toward it. »

Meredith

This thesis deals with the analysis and regulation of the behavior of specific human genes within cancer patients, investigating the biological relationships which hold among each other and the effect heterogeneous regulatory elements have on their expression.

In particular, this project is focused on one of the most common tumors in women worldwide, for which still a poor prognosis exists: the ovarian cancer, specifically the *Ovarian Serous Cystadenocarcinoma* (OV), as said, a particularly aggressive type of epithelial ovarian cancer. As for any other type of cancer, there are some specific human genes that are mainly related to the ovarian tumor: their mutations and their activity (i.e., expression) are strongly involved in the origin process and the gradual development of this tumor.

We identified 3 relevant pathways for the ovarian tumor, i.e., 3 groups of genes of interest with related functional behavior (whose functions may depend on each other), that are proved to be crucial for the study of ovarian cancer:

- ✓ a set of genes particularly relevant to cancer stem cells, i.e., those cells that are able to differentiate into other types of cells and divide to produce more of the same type of stem cells (STEM\_CELLS pathway);
- ✓ another set of genes involved in the glucose metabolism, i.e., the process by which simple sugars are produced, processed and used to produce energy in the organism (GLUCOSE\_METABOLISM pathway);
- ✓ a third set containing genes involved in DNA repair mechanisms, i.e., those processes by which a cell is able to identify and correct damages to the DNA molecules of its genome (DNA\_REPAIR pathway).

This thesis on inference of gene expression regulation and gene regulation networks in ovarian cancer patients has been conceived as a collaboration between Politecnico di Milano and “Mario Negri” Institute. This work examines the gene expression regulation process, using data on genes whose expression was measured at first by *The Cancer Genome Atlas* consortium (TCGA) through RNA-sequencing techniques.

The objective is building a predictive and possibly explicative model for the regulation of gene expression of the considered genes of interest belonging to the relevant pathways for the OV tumor, on



the basis of their methylation and expression values, as well as of the expression of their candidate regulatory genes, i.e., those genes that encode for transcription factors (TFs) having binding sites located in the promoter regions of the genes of interest.

Among all the existing factors affecting gene expression, we analyze only a specific subset of regulatory elements and post-evaluate which are relevant at the single gene level. So, within the ovarian cancer scenario, we focus on a limited set of target genes, integrating heterogenous information from other genes expression, DNA methylation and TFs binding sites with data measured on OV tumor patients.

The matter is understanding the relationships between the activity of each target gene and the genes belonging either to the same pathway or to the other related pathways, and the relationships between all such target genes and their candidate regulatory genes: this may lead to identify potential common regulators along each pathway, or frequent regulators with a key role in the regulation systems of the genes of interest, eventually predicting their possible oncogenic role. Whenever a correlation exists, an assessment of the potential influence that the gene methylation may have on its expression is also made.

Thus, the aim is not comparing different genes, but instead it is identifying the correlations among them and understanding all the relations between each gene and its related biological processes. This finally leads to building a large and explicative gene expression network that displays the main biological relationships between a gene and its already known regulators, along with other possible associations, which may unveil still unknown aspects of the impact of transcriptional regulation on tumor progression and either known or putative other interesting biological connections.

From a computational standpoint, the main goal is performing a thorough and as accurate as possible data analysis on samples coming from ovarian tumor patients. According to the hypotheses defined at the beginning of the project, this thesis deals with a feature selection problem in linear regression models. In particular, a linear regression model is built for each single gene of interest, assuming its expression can be predicted as a linear combination of its methylation and the expression values of the other genes. The analysis of all these potential regulators (i.e., *features*) allows to define all the existing statistically and especially biologically significant connections among these genes and the impacts they have on each other.

A sample network representing what we expect from the results of the linear regression is shown in *Figure 3.1*: the nodes represent the genes of interest and their relevant features selected by the linear regression model. Each gene is connected to its features through directed incoming edges; so, by retrieving one gene incoming edges, it is possible to extract all the relevant features participating in the regulation of its expression. In addition, each edge is labelled with the estimated linear regression coefficient assigned to the related feature in the model, quantifying the either positive (red edge) or negative (grey edge) effect the feature has in the model gene regulation system.



■ Color representing the pathway the gene belongs to  
■ Color representing the type of feature  
(e.g. gene of the same pathway, gene of another pathway, candidate regulatory gene, methylation of the current gene, ...)

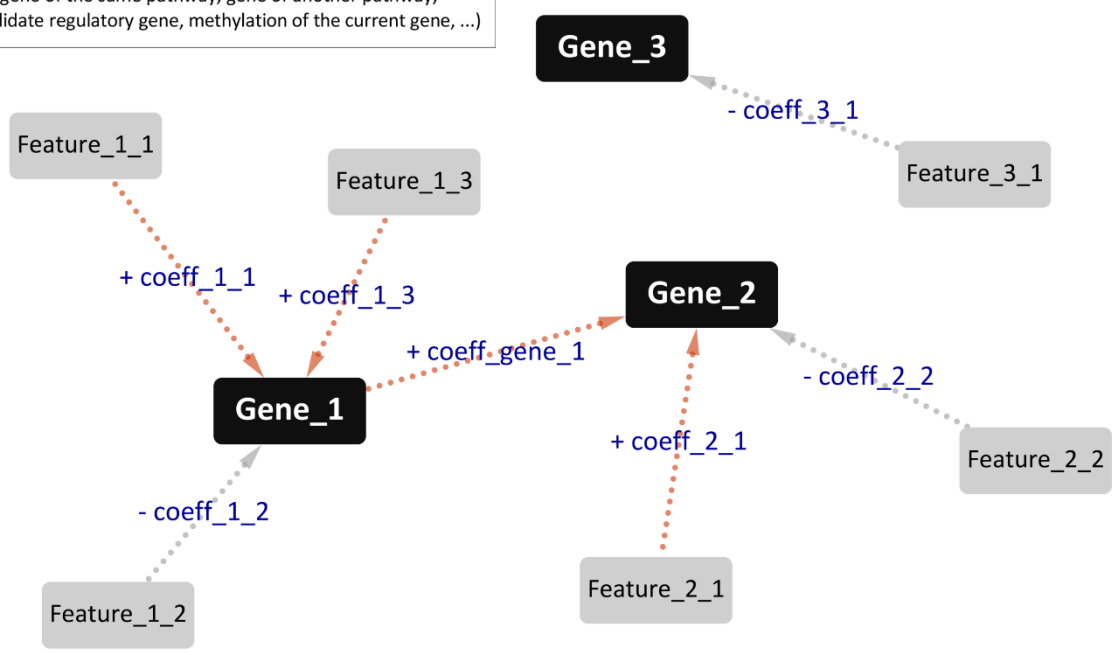


Figure 3.1: Sample network visualizing linear regression results.



## 4. Materials and Tools

« This is your starting line. This is your arena. How well you play... that's up to you. »

Richard

### 4.1 Ovarian cancer relevant pathways and their genes

The starting point of this thesis is the list of genes of interest to be analyzed: 177 genes, each one associated with its own official name (i.e., *Gene Symbol*), its numerical ID (i.e., *Entrez Gene ID*) and the genomic pathways it belongs to. As already reported in previous studies [17, 18, 19, 20], the 73 genes of STEM\_CELLS (*Table 4.1*), the 84 genes of GLUCOSE\_METABOLISM (*Table 4.2*) and the 20 genes of DNA\_REPAIR (*Table 4.3*) pathways are extremely important for ovarian cancer:

➤ tumor stem cells represent one of the main tumor relapsing mechanisms, regulating tumor's either sensitive or resistant response to Taxol and Platinum chemotherapeutic drugs; in particular, most stem cells pathway genes allow to predict cancer cells response to cisplatin and how that specific tumor may react to a chemotherapeutic treatment [17];

➤ among the major metabolic alterations of cancer cells, there is the so called *Warburg effect*, which causes an enhanced glycolysis under aerobic conditions and an increased glucose uptake via over-expression of glucose transporters. This led to *in vitro* experiments, either in the presence or in the absence of glucose, for an in-depth study of the putative association between patients response to chemotherapy and the metabolic properties of their cancer cells.

In particular, it has been proved [18] that there is a strict correlation between glucose addiction and cancer platinum-based therapy and that the resistance of some cancer cells to this type of therapy may be associated to an alteration in their metabolic profile. So, an *in vitro* analysis of cancer cells behavior in the absence of glucose or a measurement of their level of glucose addiction turns out to be a good parameter to predict patients sensitivity to platinum-based therapies;

➤ genomic instability due to DNA damages is one of the main causes of different types of cancers, including ovarian tumor. Cells are able to face these damages and preserve DNA integrity thanks to a complex network of pathways, known as *DNA Damage Response* (DDR).

The role that DNA integrity and genomic stability have within tumors is very peculiar, since on one hand, deactivation of DDR may easily lead to carcinogenesis, but, on the other hand, defects in this response may also be positive elements, making cancer more sensitive to the therapy [19].

For example, the role and the therapeutic potential of gene CDK12 is particularly important [20]: CDK12 is mainly involved in breast cancers, but its expression and related alteration have a fundamental role in different types of tumors, particularly within the ovarian cancer, making it a new candidate tumor suppressor gene in ovarian carcinoma.

Table 4.1: List of genes of interest in the STEM\_CELLS pathway.

| GENE_SYMBOL | ENTREZ_GENE_ID | PATHWAY    |        |        |            |
|-------------|----------------|------------|--------|--------|------------|
| ABCB5       | 340273         | STEM_CELLS | ITGA2  | 3673   | STEM_CELLS |
| ABCG2       | 9429           | STEM_CELLS | ITGA4  | 3676   | STEM_CELLS |
| ALCAM       | 214            | STEM_CELLS | ITGA6  | 3655   | STEM_CELLS |
| ALDH1A1     | 216            | STEM_CELLS | ITGB1  | 3688   | STEM_CELLS |
| ATM         | 472            | STEM_CELLS | JAG1   | 182    | STEM_CELLS |
| ATXN1       | 6310           | STEM_CELLS | JAK2   | 3717   | STEM_CELLS |
| AXL         | 558            | STEM_CELLS | KIT    | 3815   | STEM_CELLS |
| BMI1        | 648            | STEM_CELLS | KITLG  | 4254   | STEM_CELLS |
| BMP7        | 655            | STEM_CELLS | KLF17  | 128209 | STEM_CELLS |
| CD24        | 100133941      | STEM_CELLS | KLF4   | 9314   | STEM_CELLS |
| CD34        | 947            | STEM_CELLS | LATS1  | 9113   | STEM_CELLS |
| CD38        | 952            | STEM_CELLS | LIN28A | 79727  | STEM_CELLS |
| CD44        | 960            | STEM_CELLS | LIN28B | 389421 | STEM_CELLS |
| CHEK1       | 1111           | STEM_CELLS | MAML1  | 9794   | STEM_CELLS |
| DACH1       | 1602           | STEM_CELLS | MERTK  | 10461  | STEM_CELLS |
| DDR1        | 780            | STEM_CELLS | MS4A1  | 931    | STEM_CELLS |
| DKK1        | 22943          | STEM_CELLS | MUC1   | 4582   | STEM_CELLS |
| DLL1        | 28514          | STEM_CELLS | MYC    | 4609   | STEM_CELLS |
| DLL4        | 54567          | STEM_CELLS | MYCN   | 4613   | STEM_CELLS |
| DNMT1       | 1786           | STEM_CELLS | NANOG  | 79923  | STEM_CELLS |
| EGF         | 1950           | STEM_CELLS | NFKB1  | 4790   | STEM_CELLS |
| ENG         | 2022           | STEM_CELLS | NOS2   | 4843   | STEM_CELLS |
| EPCAM       | 4072           | STEM_CELLS | NOTCH1 | 4851   | STEM_CELLS |
| ERBB2       | 2064           | STEM_CELLS | NOTCH2 | 4853   | STEM_CELLS |
| ETFA        | 2108           | STEM_CELLS | PECAM1 | 5175   | STEM_CELLS |
| FGFR2       | 2263           | STEM_CELLS | PLAT   | 5327   | STEM_CELLS |
| FLOT2       | 2319           | STEM_CELLS | PLAUR  | 5329   | STEM_CELLS |
| FOXA2       | 3170           | STEM_CELLS | POU5F1 | 5460   | STEM_CELLS |
| FOXP1       | 27086          | STEM_CELLS | PROM1  | 8842   | STEM_CELLS |
| FZD7        | 8324           | STEM_CELLS | PTCH1  | 5727   | STEM_CELLS |
| GATA3       | 2625           | STEM_CELLS | PTPRC  | 5788   | STEM_CELLS |
| GSK3B       | 2932           | STEM_CELLS | SAV1   | 60485  | STEM_CELLS |
| HDAC1       | 3065           | STEM_CELLS | SIRT1  | 23411  | STEM_CELLS |
| ID1         | 3397           | STEM_CELLS | SMO    | 6608   | STEM_CELLS |
| IKBKB       | 3551           | STEM_CELLS | SNAI1  | 6615   | STEM_CELLS |
| CXCL8       | 3576           | STEM_CELLS | SOX2   | 6657   | STEM_CELLS |
|             |                |            | STAT3  | 6774   | STEM_CELLS |

Table 4.2: List of genes of interest in the GLUCOSE\_METABOLISM pathway.

| GENE_SYMBOL | ENTREZ_GENE_ID | PATHWAY            |         |        |                    |
|-------------|----------------|--------------------|---------|--------|--------------------|
| ACLY        | 47             | GLUCOSE_METABOLISM | PCK1    | 5105   | GLUCOSE_METABOLISM |
| ACO1        | 48             | GLUCOSE_METABOLISM | PCK2    | 5106   | GLUCOSE_METABOLISM |
| ACO2        | 50             | GLUCOSE_METABOLISM | PDHA1   | 5160   | GLUCOSE_METABOLISM |
| AGL         | 178            | GLUCOSE_METABOLISM | PDHB    | 5162   | GLUCOSE_METABOLISM |
| ALDOA       | 226            | GLUCOSE_METABOLISM | PDK1    | 5163   | GLUCOSE_METABOLISM |
| ALDOB       | 229            | GLUCOSE_METABOLISM | PDK2    | 5164   | GLUCOSE_METABOLISM |
| ALDOC       | 230            | GLUCOSE_METABOLISM | PDK3    | 5165   | GLUCOSE_METABOLISM |
| BPGM        | 669            | GLUCOSE_METABOLISM | PDK4    | 5166   | GLUCOSE_METABOLISM |
| CS          | 1431           | GLUCOSE_METABOLISM | PDP2    | 57546  | GLUCOSE_METABOLISM |
| DLAT        | 1737           | GLUCOSE_METABOLISM | PDPR    | 55066  | GLUCOSE_METABOLISM |
| DLD         | 1738           | GLUCOSE_METABOLISM | PFKL    | 5211   | GLUCOSE_METABOLISM |
| DLST        | 1743           | GLUCOSE_METABOLISM | PGAM2   | 5224   | GLUCOSE_METABOLISM |
| ENO1        | 2023           | GLUCOSE_METABOLISM | PGK1    | 5230   | GLUCOSE_METABOLISM |
| ENO2        | 2026           | GLUCOSE_METABOLISM | PGK2    | 5232   | GLUCOSE_METABOLISM |
| ENO3        | 2027           | GLUCOSE_METABOLISM | PGLS    | 25796  | GLUCOSE_METABOLISM |
| FBP1        | 2203           | GLUCOSE_METABOLISM | PGM1    | 5236   | GLUCOSE_METABOLISM |
| FBP2        | 8789           | GLUCOSE_METABOLISM | PGM2    | 55276  | GLUCOSE_METABOLISM |
| FH          | 2271           | GLUCOSE_METABOLISM | PGM3    | 5238   | GLUCOSE_METABOLISM |
| G6PC        | 2538           | GLUCOSE_METABOLISM | PHKA1   | 5255   | GLUCOSE_METABOLISM |
| G6PC3       | 92579          | GLUCOSE_METABOLISM | PHKB    | 5257   | GLUCOSE_METABOLISM |
| G6PD        | 2539           | GLUCOSE_METABOLISM | PHKG1   | 5260   | GLUCOSE_METABOLISM |
| GALM        | 130589         | GLUCOSE_METABOLISM | PHKG2   | 5261   | GLUCOSE_METABOLISM |
| GBE1        | 2632           | GLUCOSE_METABOLISM | PKLR    | 5313   | GLUCOSE_METABOLISM |
| GCK         | 2645           | GLUCOSE_METABOLISM | PRPS1   | 5631   | GLUCOSE_METABOLISM |
| GPI         | 2821           | GLUCOSE_METABOLISM | PRPS1L1 | 221823 | GLUCOSE_METABOLISM |
| GSK3A       | 2931           | GLUCOSE_METABOLISM | PRPS2   | 5634   | GLUCOSE_METABOLISM |
| GSK3B       | 2932           | GLUCOSE_METABOLISM | PYGL    | 5836   | GLUCOSE_METABOLISM |
| GYS1        | 2997           | GLUCOSE_METABOLISM | PYGM    | 5837   | GLUCOSE_METABOLISM |
| GYS2        | 2998           | GLUCOSE_METABOLISM | RBKS    | 64080  | GLUCOSE_METABOLISM |
| H6PD        | 9563           | GLUCOSE_METABOLISM | RPE     | 6120   | GLUCOSE_METABOLISM |
| HK2         | 3099           | GLUCOSE_METABOLISM | RPIA    | 22934  | GLUCOSE_METABOLISM |
| HK3         | 3101           | GLUCOSE_METABOLISM | SDHA    | 6389   | GLUCOSE_METABOLISM |
| IDH1        | 3417           | GLUCOSE_METABOLISM | SDHB    | 6390   | GLUCOSE_METABOLISM |
| IDH2        | 3418           | GLUCOSE_METABOLISM | SDHC    | 6391   | GLUCOSE_METABOLISM |
| IDH3A       | 3419           | GLUCOSE_METABOLISM | SDHD    | 6392   | GLUCOSE_METABOLISM |
| IDH3B       | 3420           | GLUCOSE_METABOLISM | SUCLA2  | 8803   | GLUCOSE_METABOLISM |
| IDH3G       | 3421           | GLUCOSE_METABOLISM | SUCLG1  | 8802   | GLUCOSE_METABOLISM |
| MDH1        | 4190           | GLUCOSE_METABOLISM | SUCLG2  | 8801   | GLUCOSE_METABOLISM |
| MDH1B       | 130752         | GLUCOSE_METABOLISM | TALDO1  | 6888   | GLUCOSE_METABOLISM |
| MDH2        | 4191           | GLUCOSE_METABOLISM | TKT     | 7086   | GLUCOSE_METABOLISM |
| OGDH        | 4967           | GLUCOSE_METABOLISM | TPI1    | 7167   | GLUCOSE_METABOLISM |
| PC          | 5091           | GLUCOSE_METABOLISM | UGP2    | 7360   | GLUCOSE_METABOLISM |

Table 4.3: List of genes of interest in the DNA\_REPAIR pathway.

| GENE_SYMBOL | ENTREZ_GENE_ID | PATHWAY    |         |       |            |
|-------------|----------------|------------|---------|-------|------------|
| BRCA1       | 672            | DNA_REPAIR | PARP1   | 142   | DNA_REPAIR |
| CDK12       | 51755          | DNA_REPAIR | POLB    | 5423  | DNA_REPAIR |
| ERCC1       | 2067           | DNA_REPAIR | POLE    | 5426  | DNA_REPAIR |
| FANCA       | 2175           | DNA_REPAIR | POLQ    | 10721 | DNA_REPAIR |
| FANCC       | 2176           | DNA_REPAIR | RAD51   | 5888  | DNA_REPAIR |
| FANCD2      | 2177           | DNA_REPAIR | TP53BP1 | 7158  | DNA_REPAIR |
| FANCF       | 2188           | DNA_REPAIR | XPA     | 7507  | DNA_REPAIR |
| MLH1        | 4292           | DNA_REPAIR | ERCC2   | 2068  | DNA_REPAIR |
| OGG1        | 4968           | DNA_REPAIR | ERCC4   | 2072  | DNA_REPAIR |
| PALB2       | 79728          | DNA_REPAIR | ERCC5   | 2073  | DNA_REPAIR |

In addition, a more detailed sub-classification is provided, defined according to the specific biological function these genes are involved in within their pathway. Genes that are not present in this classification are grouped together into an additional user-defined subclass for each pathway, called “*Unclassified*”.

These subclasses are used to generate gene expression networks from the results of the data analysis, in order to investigate unveiled relationships between genes with similar biological functions. The names of the subclasses and the number of genes they contain are reported below. A gene carrying out multiple functions can be contained in multiple subclasses.

- STEM\_CELLS (73)
  - ✓ AKT & PI3 Kinase – mTOR Signaling (2)
  - ✓ Asymmetric Division (4)
  - ✓ Cancer Stem Cells Markers (23)
  - ✓ Cancer Therapeutic Targets (17)
  - ✓ Cell Migration & Metastasis (7)
  - ✓ Cell Proliferation (5)
  - ✓ Hedgehog Signaling (2)
  - ✓ Hippo Signaling (3)
  - ✓ Loss of Stemness (6)
  - ✓ Notch Signaling (6)
  - ✓ Pluripotency (6)
  - ✓ Self-Renewal (3)
  - ✓ STAT-NFκB Signaling (3)
  - ✓ WNT Signaling (2)
  - ✓ Unclassified (0)
  
- GLUCOSE\_METABOLISM (84)
  - ✓ Gluconeogenesis (7)
  - ✓ Glycogen Degradation (6)

- ✓ Glycogen Synthesis (4)
- ✓ Glycolysis (12)
- ✓ Pentose Phosphate Pathway (11)
- ✓ Regulation of Glucose Metabolism (6)
- ✓ Regulation of Glycogen Metabolism (6)
- ✓ Tricarboxylic Acid Cycle (29)
- ✓ Unclassified (6)
  
- DNA\_REPAIR (20)
  - ✓ BER (3)
  - ✓ DSB (8)
  - ✓ NER (5)
  - ✓ Gene CDK12 Group (1)
  - ✓ Gene MLH1 Group (1)
  - ✓ Unclassified (2)

The total number of genes of interest is 177: 176 distinct, plus gene 2932 [GSK3B] present in both STEM\_CELLS and GLUCOSE\_METABOLISM. This duplicated gene can be treated as a unique gene during the data extraction process, since its methylation level or expression value in each patient is always the same, regardless of the pathway. However, during the analysis it is managed separately, according to the pathway under analysis: the fact that this gene belongs to different pathways means that it is involved in multiple biological functions and, as a consequence, the regulation of its expression is different according to the pathway it participates in.

## 4.2 Breast cancer PAM50 data samples

This second file comprises the list of PAM50 data samples, each identified by the ID of the patient it refers to, extracted at “Mario Negri” from the complete set of breast cancer (BRCA) data samples.

PAM50 is a test that helps classifying the different types of breast cancers into a set of clinically significant molecular subtypes that are important for the management of the disease, by performing a *Real Time Quantitative PCR* (RT-qPCR) analysis: in general, *Polymerase Chain Reaction* (PCR) is a molecular biology technique used to amplify and detect DNA and RNA sequences, comprised in RT-qPCR, which is commonly used to detect, characterize and quantify nucleic acids.

The PAM50 test consists in measuring the expression of a group of 50 classifier genes and five control genes to identify the intrinsic subtypes known as *Luminal A*, *Luminal B*, *HER2-enriched* and *Basal-like*. This file provides 1043 BRCA patient data samples with the following attributes:

- PAM50 represents the BRCA subtypes classification (i.e., LumA, LumB, Her2, Basal);

Table 4.4: A subset of samples contained in the BRCA\_PAM50 file.

| ID           | PAM50  | ER       | PR       | HER2            | TRIPLE | TNBCTYPE |
|--------------|--------|----------|----------|-----------------|--------|----------|
| TCGA-A1-A0SD | LumA   | Positive | Positive | Negative        | nonTN  | -        |
| TCGA-A1-A0SI | LumB   | Positive | Positive | Negative        | nonTN  | -        |
| TCGA-A1-A0SK | Basal  | Negative | Negative | Negative        | TN     | M        |
| TCGA-A1-A0SM | LumB   | Positive | Negative | Positive        | nonTN  | -        |
| TCGA-A1-A0SN | LumB   | Positive | Positive | Positive        | nonTN  | -        |
| TCGA-A1-A0SP | Basal  | Negative | Negative | Negative        | TN     | UNS      |
| TCGA-A1-A0SQ | LumA   | Positive | Positive | Negative        | nonTN  | -        |
| TCGA-A2-A04N | LumA   | Positive | Positive | [Not Evaluated] | nonTN  | -        |
| TCGA-A2-A04U | Basal  | Negative | Negative | Negative        | TN     | M        |
| TCGA-A2-A0CM | Basal  | Negative | Negative | Negative        | TN     | IM       |
| TCGA-A2-A0D0 | Basal  | Negative | Negative | Negative        | TN     | UNS      |
| TCGA-A2-A0D2 | Basal  | Negative | Negative | Negative        | TN     | UNS      |
| TCGA-A2-A0SX | Basal  | Negative | Negative | Negative        | TN     | UNS      |
| TCGA-A2-A0T0 | Basal  | Negative | Negative | Negative        | TN     | M        |
| TCGA-A2-A0YE | Basal  | Negative | Negative | Negative        | TN     | M        |
| TCGA-A8-A09X | Her2   | Negative | Negative | Negative        | TN     | LAR      |
| TCGA-AC-A2BK | Basal  | Negative | Negative | Negative        | TN     | M        |
| TCGA-AC-A2QH | Basal  | Negative | Negative | Negative        | TN     | UNS      |
| TCGA-AC-A2QJ | Basal  | Negative | Negative | Negative        | TN     | BL2      |
| TCGA-AC-A6IW | Basal  | Negative | Negative | Negative        | TN     | IM       |
| TCGA-AC-A7VC | Basal  | Negative | Negative | Negative        | TN     | MSL      |
| TCGA-AN-A04D | Basal  | Negative | Negative | Negative        | TN     | M        |
| TCGA-B6-A3ZX | Normal | Negative | Negative | Negative        | TN     | IM       |

- ER, PR and HER2 respectively denote the status of the *estrogen receptors*, the *progesterone receptors* and the *human epidermal growth factor receptor-2*, which represent three predictive biomarkers in breast pathology. If breast cancers are either “ER-positive” or “PR-positive”, it means the cancer cells grow in response to hormones estrogen and progesterone, respectively.

Tumors that are ER/PR-positive are much more likely to respond to hormone therapy than ER/PR-negative tumors, which can help preventing a return of the disease by blocking the effects of estrogen. “HER2-positive” breast cancers, instead, are characterized by cells that produce too much of a protein known as HER2, making these cancers much more aggressive and fast-growing.

If these markers are all negative, the tumors are classified as *Triple-Negative Breast Cancers*, (TN BRCA) because they do not have estrogen and progesterone receptors and do not overexpress the HER2 protein, which makes them more difficult to treat, usually requiring the combination of different therapies;

- TRIPLE indicates whether the sample is classified as a triple negative cancer or not;
- TNBCTYPE represents different subtypes of triple negative cancers.

Table 4.4 shows how this file is organized.



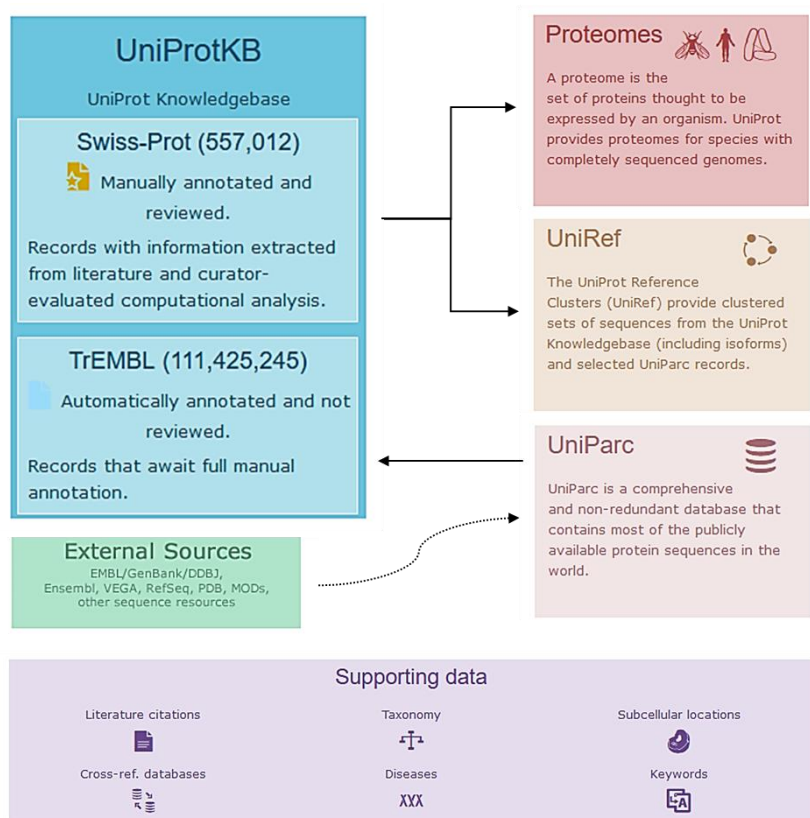


Figure 4.1: UniProt structure. Taken from [23].

Since no other ovarian cancer data samples are available and the whole set of BRCA data samples is highly heterogeneous, we use PAM50 information during the validation process for selecting and analyzing only a BRCA subtype that is biomolecularly equivalent to OV, the *Basal-like Breast Cancer* (PAM50 = Basal), which has proved to be genetically similar to ovarian tumor [21]: they share similar genetic origins and features, which may allow to treat these two difficult-to-treat cancer types with the same therapies.

## 4.3 Human gene nomenclature and human transcription factors

Some data come from the main biological and genomic resources that are open and available on the Web, providing material on both genes and proteins. In particular, we use this information to build an integrated and comprehensive spreadsheet containing the list of human genes with all their names, IDs and, if so, the transcription factors they encode: this becomes a very powerful tool for this project, specifically during the process of extraction and manipulation of data of interest.

### 4.3.1 UniProt

*UniProt* [22, 23] is a high-quality, non-redundant and freely accessible repository of protein sequences and functional information (*Figure 4.1* pictures its structure). It is the world most comprehensive catalog

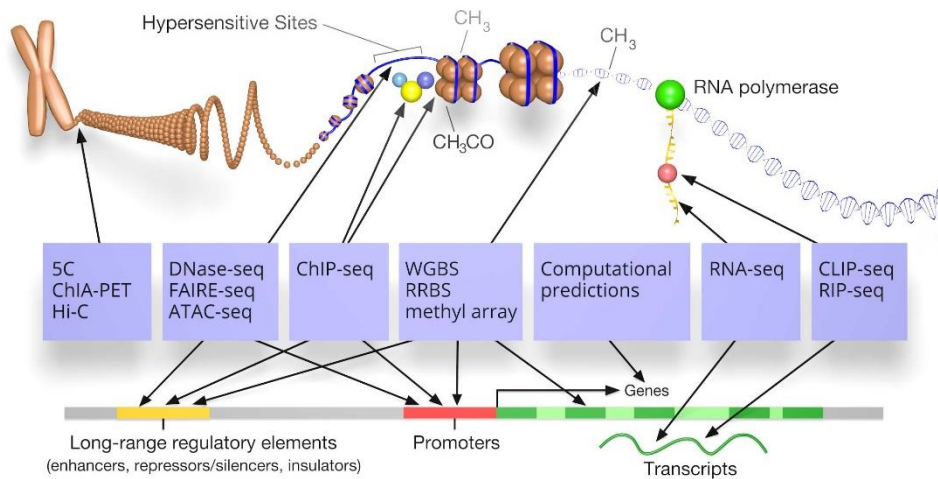


Figure 4.2: A graphical representation of the ENCODE information. Taken from [25].

of information on proteins, created by joining the information contained in [Swiss-Prot](#) (a manually curated and annotated protein sequence databank), [TrEMBL](#) (a computer-annotated protein sequence database) and [PIR](#) (a complete, integrated, cross-referenced and public resource of functional annotated protein sequences).

### 4.3.2 ENCODE

The *Encyclopedia of DNA Elements* (ENCODE) [24, 25] is a comprehensive resource of the main functional elements in the human genome, such as the elements acting at the protein and RNA levels or the regulatory elements controlling cells and circumstances in which a gene is active.

Only 1.5% of DNA in the human genome, in fact, is composed by protein-coding genes (around 20,000 in total): the primary goal of the ENCODE project is to determine the role of all the remaining components of the genome. The different types of information contained in ENCODE are represented in *Figure 4.2*.

### 4.3.3 HUGO Gene Nomenclature Committee

The *HUGO Gene Nomenclature Committee* (HGNC) [26] is a committee of the Human Genome Organization that sets the standards for human gene nomenclature. It is responsible for approving unique symbols and names for human loci, including, for example, protein-coding genes, ncRNA genes or pseudogenes, to allow unambiguous scientific communication, according to the following main naming guidelines:

- *gene symbols* (i.e., gene name abbreviations) must be unique;
- symbols only contain Latin letters and Arabic numerals;
- symbols do not contain punctuation or letter “G” for gene;
- symbols do not contain any reference to the species they belong to.

## 4.4 Transcription factor dataset from ENCODE

The first main dataset used for the extraction of data of interest is ENCODE, for selecting all the human transcription factors having binding sites located within the promoters of genes of interest.

ENCODE contains a huge amount of both original and processed data with different quality levels and often characterized by noise. In particular, there are multiple processing levels defined according to different filtering qualitative parameters, either with a higher number of enriched binding regions, but with a lower quality (i.e., lower confidence that regions are actually correct), or with fewer but more reliable and higher-quality regions.

This work uses *NARROW (Point-Source) conservative thresholded idr Peaks* regions data from ENCODE (updated in November 2017), obtained through ChIP-seq experiments regarding the localization of human transcription factor binding sites. These filters allow to select high quality data and to execute a reliable data analysis.

The cell lines are instead selected on the basis of the specific biological relevance to the study: since ENCODE does not contain data of the same exact tissue where OV tumor patients data come from, MCF7 cell line is selected, as the most significant biological and molecular item to our goal. However, since MCF7 available data are limited, we include also cell line K562 in the analysis, because of its biological similarity with MCF7 and its huge amount of available data.

It is important to underline that these are simply candidate transcription factors and they are then weighted according to the expression of their encoding genes in the different specific patients.

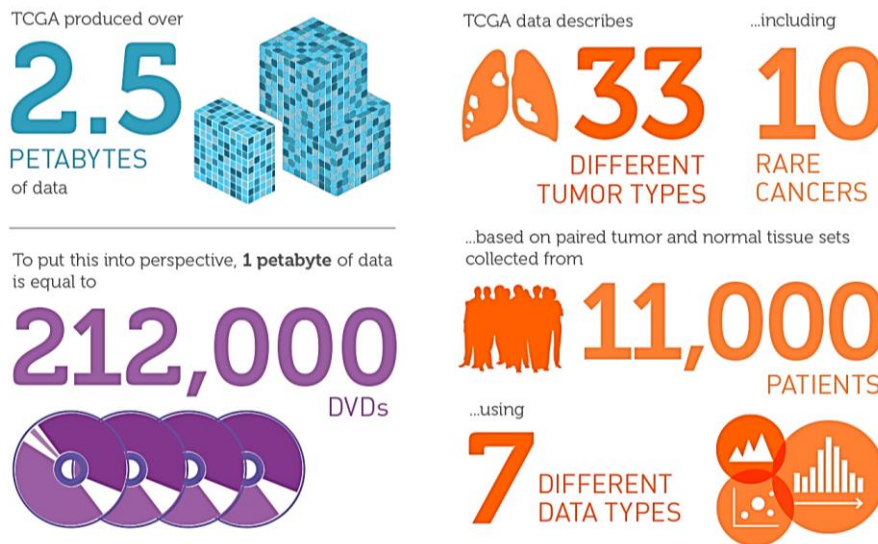
The complete data extraction process is described in [Chapter 5](#).

## 4.5 Methylation and gene expression datasets from TCGA

*The Cancer Genome Atlas* (TCGA) [27, 28] is a catalogue of all the genetic mutations that are responsible for cancer. The TCGA project has generated comprehensive and multi-dimensional maps of the key genomic changes in 33 types of cancer ([Figure 4.3](#)). TCGA mission is accelerating the understanding of the molecular basis of cancer through the application of genome analysis technologies, such as large-scale genome sequencing, with the objective of improving everyone's ability to diagnose, treat and prevent cancer.

We use TCGA for extracting methylation and gene expression data related to OV and BRCA tumors for assembly GRCh38 (i.e., the *Genome Reference Consortium Human Build 38*). In particular, two main TCGA datasets are used to extract this information:

- the TCGA methylation dataset, containing DNA methylation sites for all the different types of tumors provided by TCGA, with their methylation values expressed through the so called *beta\_values* (values within a [0-1] range), having a unique value for each methylation site (i.e., methylated cytosine base). Methylation data from two different sequencing platforms are provided (*Illumina HumanMethylation 27* and *Illumina HumanMethylation 450*), which differ



### TCGA RESULTS & FINDINGS



#### MOLECULAR BASIS OF CANCER

Improved our understanding of the genomic underpinnings of cancer

For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.



#### TUMOR SUBTYPES

Revolutionized how cancer is classified

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.



#### THERAPEUTIC TARGETS

Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development

TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

Figure 4.3: TCGA statistics. Taken from [28].

for the number of methylation probes that are used for the measurement (27k and 450k probes respectively). As for tumors of interest, the platform with the highest amount of data available is selected (i.e., 27k for OV and 450k for BRCA);

- the TCGA gene expression dataset, containing human gene expression values in all the different types of tumors provided by TCGA, derived from RNA-seq techniques and expressed in *Fragments Per Kilobase per Million reads* (FPKM).

Each data sample corresponds to a specific patient, whose cancer information has been collected through specific measurements, observations and analyses from TCGA.

Each patient is identified by a unique identifier, called *TCGA Aliquot Barcode* (Figure 4.4), composed by a collection of attributes identifying specific TCGA data elements [29, 30]:

- **project**, the name of the project data belongs to (i.e. 'TCGA', in this case);
- **TSS**, the tissue source site;

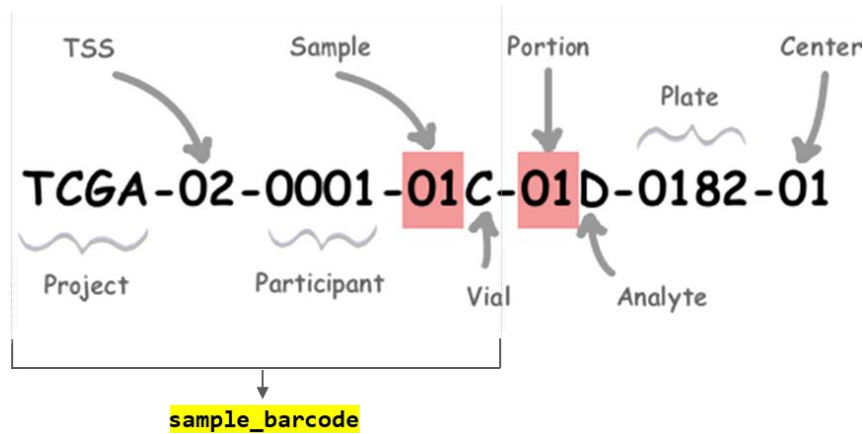


Figure 4.4: How a TCGA aliquot barcode can be broken down into its components. Taken from [29].

- **participant**, a numerical ID that identifies the patient participating in the study;
- **sample**, a numerical code identifying the type of sample (01-09 codes mean “tumor” types, 10-19 codes mean “normal”, 20-29 codes mean “control”);
- **vial**, the order of a sample in the sequence of samples considered (possible values from A to Z);
- **portion**, the order of a portion in a sequence of 100-120 mg sample portions (possible values from 01 to 99, where 01 is the first portion of the sample);
- **analyte**, the molecular type of the component for the analysis (for example, D indicates a DNA sample, while R indicates an RNA sample);
- **plate**, a 4-digit alphanumeric value that specifies the order of a plate in a sequence of 96-well plates (for example, 0182 indicates the 182<sup>nd</sup> plate);
- **center**, the sequencing or characterization center that will receive the aliquot for the analysis. Each center is identified by a specific code.

In this work we simply refer to the first part of this barcode (i.e., the *Sample Barcode*) as “TCGA Aliquot”, which is the identifier used to recognize and classify the patients in the TCGA methylation and expression data of interest.

## 4.6 Genomic annotations from GENCODE

GENCODE is a scientific project producing high-quality reference gene annotations and experimental validation for human and mouse genomes [31, 32, 33]. It aims at building an encyclopedia of genes and genes variants, by identifying all gene features in the human and mouse genome, using a combination of computational analysis and manual annotation, in order to create a comprehensive set of annotations,

including genes, transcripts, exons, protein-coding and non-coding loci, as well as variants coming from alternative splicing and pseudogenes.

Over time, different versions of the annotations have been released as a response to the progress of genetic researches and their related findings: in particular, this thesis is based on version 22 of the GENCODE genomic annotations, released in 2015 [34], as the same annotation file is also used by TCGA for processing methylation and expression data of interest [35]: in this way consistency with the TCGA data is preserved, minimizing all the potential inconsistencies that may occur using a different reference. Starting from this annotation file, we build two datasets containing the following types of data:

- data about the localization of human gene promoters in assembly GRCh38, used in combination with TFs data from ENCODE;
- data about regions of the genes in which the methylation sites of interest are located, i.e., from 4000 bases upstream to 1000 bases downstream around the gene TSSs, used in combination with the TCGA methylation dataset.

The complete datasets creation process is described in *Chapter 5*.

## 4.7 Computational tools

From a computer science standpoint, four main computational tools are used for performing the data extraction, analysis and validation processes.

### 4.7.1 GenoMetric Query Language

The GenoMetric Query Language (GMQL) [36] is a declarative language used to perform queries on big genomic data, structured according to the Genomic Data Model (GDM) [37, 38, 39, 40, 41].

The GDM is a formal framework used for representing in a uniform way genomic data with different formats. This model is mainly based on the notions of **datasets**, defined as collections of samples, and **samples**, representing different genomic data.

Each sample consists of two main parts:

- **Region data**, describing the physical coordinates of the genome areas and their features, encoded as specific fields having different values for each region;
- **Metadata**: descriptive attributes of a sample, describing its biological, clinical and experimental properties.

Regions are data format independent and provide an interoperability framework for comparing data on mutations, expression or regulation; while metadata are system independent and provide an interoperability framework for comparing samples based upon their biological aspects.



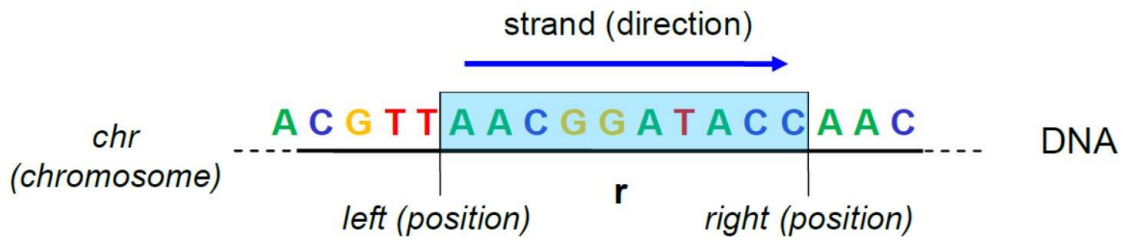


Figure 4.5: Graphical representation of genomic region coordinates. Taken from [3].

In GDM, metadata are modeled as attribute-value pairs, while region data have to follow some rules defining their structure in order to be comparable among different datasets: these rules are encoded in a *schema*. Different datasets have different schemas, but each schema must include at least the following information: the chromosome, the start and end positions of the region and the strand, which represents the direction of DNA reading.

Formally, the atomic unit of a GDM dataset is the genomic region, a portion of the genome defined by a set of coordinates (represented in Figure 4.5)

$$c = \langle \text{chr}, \text{left}, \text{right}, \text{strand} \rangle$$

and a set of features

$$f = \langle \text{feature}_1, \text{feature}_2, \dots, \text{feature}_N \rangle$$

The concatenation of these two sets of values creates the region:  $r = \langle c, f \rangle$ .

The order of the coordinates and the features is fixed in all the regions of a dataset and it is dictated by the schema and each one of these fields is typed (e.g., we have respectively *string*, *integer*, *integer*, *char* for the coordinates). Metadata, instead, are arbitrary attribute-value pairs  $\langle a, v \rangle$ .

A set of regions, i.e., a **sample**, is defined as:

$$s = \langle \text{id}, \{r_1, r_2, \dots\}, \{m_1, m_2, \dots\} \rangle$$

Each sample is identified by a specific *id*, which is unique in the whole dataset, and has multiple regions  $r_i$  and multiple attribute-value pairs of metadata  $m_i$ .

A dataset is therefore a set of samples with the same region schema.

Having genomic data represented according to the Genomic Data Model, it is possible to query them using the suitably designed and easy-to-learn query language mentioned here above: GMQL, a high-level and declarative query language inspired by the classical languages for database management, such as SQL. It uses conventional algebraic operations (e.g., selection, projection, join) together with domain-specific operations targeting bioinformatics applications.

A GMQL query is defined as a sequence of statements (i.e., algebraic operations) with the following syntax:

$$\langle \text{variable} \rangle = \text{operation}(\langle \text{parameters} \rangle) \langle \text{variables} \rangle$$

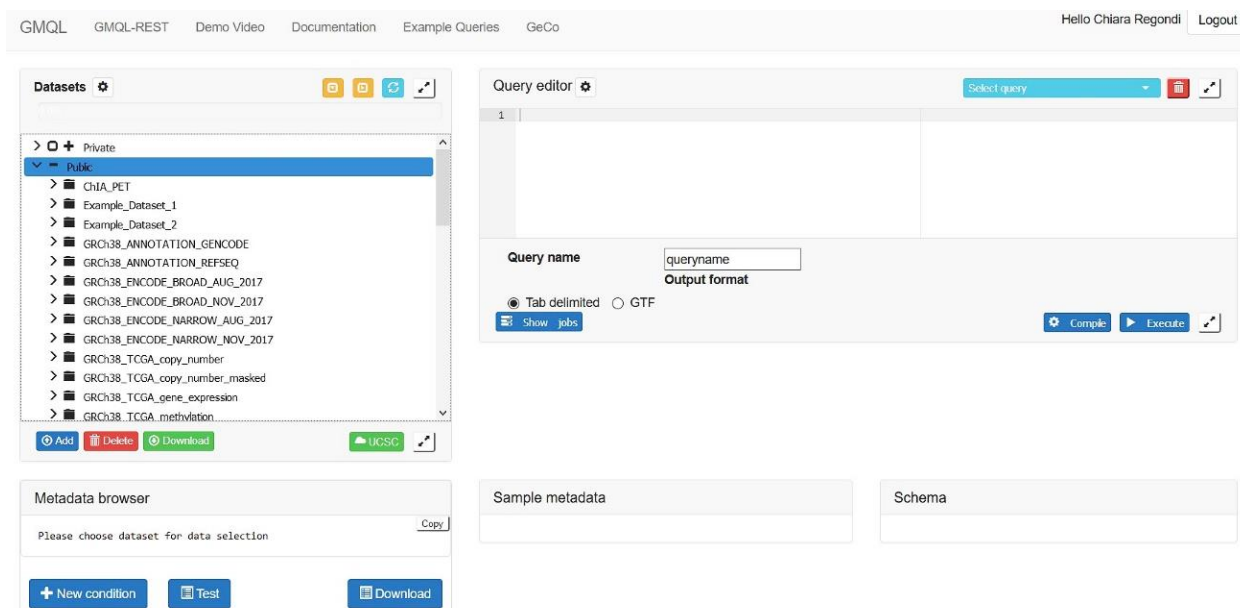


Figure 4.6: Excerpt of the GMQL Web interface. Taken from [43].

where `<variable>` is a GDM dataset and `operation(...)` specifies the procedure to be executed, with all its required and optional parameters. Operations can be either unary or binary, according to the input dataset, and they all return one dataset as a result. A complete description of all the GMQL operations can be found at [42].

The GMQL system [43] can be used online through a specific Web interface (Figure 4.6), to be accessed as a guest user or as an authenticated user, after setting up a personal account.

We use this system to extract all the necessary biological and genomic data provided by big consortia (ENCODE, TCGA, GENCODE or RefSeq) due to the following:

- it is an integrated resource that provides a comprehensive and public set of datasets, comprising all the biological and genomic data extracted from their original sources;
- it allows to easily filter and download data of interest simply by compiling and executing user-defined SQL-like queries, following the syntax of the GMQL language;
- data are well-organized and easier to use than the ones in the original web portals, since they are cleaned, processed and modelled according to the pre-defined GDM framework, which provides both resulting genome areas and additional descriptive attributes, into a straightforward table-like format.


## 4.7.2 Python libraries

Most computations are performed using the Python programming language [44].

Python is an interpreted language executing direct instructions, without the need of compiling the scripts into machine language instructions, and it allows to easily execute a lot of complex tasks, thanks to the availability of standard libraries and of a large number of resources.



a)

```
Query editor  query_name  
1 # Extract narrows peaks from ChIP-seq experiments  
2 PEAKS_DATASET = SELECT(assay == "ChIP-seq" AND output_type == "peaks") GRCh38_ENCODE_NARROW_NOV_2017;  
3  
4 # Group the data samples by cell line  
5 PEAKS_BY_CELL_DATASET = GROUP(biosample_term_name) PEAKS_DATASET;  
6  
7 # Create the final dataset and save it into the GMQL system to download it locally or make it usable in other GMQL queries  
8 MATERIALIZE PEAKS_BY_CELL_DATASET INTO FINAL_OUTPUT_DATASET;  
9
```

b)

```
# Import the library  
import gmql as gl  
  
# Connect remotely to the GMQL server  
gl.set_remote_address('http://gmql.eu/gmql-rest/')  
gl.login()  
gl.set_mode('remote')  
  
# Load the TCGA dataset to be used in the query  
encode_narrow_dataset = gl.load_from_remote(remote_name='GRCh38_ENCODE_NARROW_NOV_2017', owner='public')  
  
# Extract narrows peaks from ChIP-seq experiments  
peaks_dataset = encode_narrow_dataset.meta_select((encode_narrow_dataset['assay'] == 'ChIP-seq') &  
                                                  (encode_narrow_dataset['output_type'] == 'peaks'))  
  
# Group the data samples by cell line  
peaks_by_cell_dataset = peaks_dataset.group(meta=['biosample_term_name'])  
  
# Create the final dataset and save it.  
# The result dataset is loaded into a GDataframe, an object containing two pandas dataframes,  
# one for the region data and one for the metadata  
final_output_dataset = peaks_by_cell_dataset.materialize('./MaterializeResults')  
  
# Get the two pandas dataframes in order to use them for following manipulations  
final_output_df_regions = final_output_dataset.regions  
final_output_df_metadata = final_output_dataset.metadata
```

Figure 4.7: A GMQL query to be executed directly on the system through the Web interface (a) and the corresponding query written in Python, to be executed remotely through the PyGMQL library (b).

Python has been conveniently selected as it offers a wide set of functions for statistical modeling and machine learning analysis and it is the only programming language integrated with GMQL. Here Python is used through *Anaconda*, an open source distribution for large-scale data processing and scientific computing and the most convenient framework for Python data science and machine learning, including at the installation more than 250 popular data science packages. The main libraries used for the computations are detailed in the following paragraphs.

#### 4.7.2.1 PyGMQL

*PyGMQL* [45, 46] is a Python module that enables the user to execute GMQL queries and perform all the available operations on genomic data in a scalable way and directly from Python.

PyGMQL translates the GMQL paradigm to the interactive and script-oriented Python environment, enabling the integration of genomic data with classical Python packages for machine learning and data science. *Figure 4.7* shows an example of a GMQL query (a) translated according to the syntax of PyGMQL (b), which can be remotely executed from Python on the GMQL server.

This library is used to extract all the data needed for this project, in order to have them already encoded as specific Python data structures (i.e., *pandas* dataframes), ready to be used and manipulated.

#### 4.7.2.2 Pandas

*Pandas* [47] is a Python package for data manipulation and analysis. It provides specific data structures and operations for manipulating tables and time series, with relational or labelled data easy to work with. This is the library mostly used in this work, because it easily allows to import and export any kind of data in a tabular format (*Figure 4.8*), from Excel spreadsheets to tab-delimited or comma-separated files.

#### 4.7.2.3 Statsmodel

*Statsmodel* [48] is a Python module for data analysis, data science and statistics, providing classes and functions for exploring data, estimating statistical models and performing statistical tests.

This library allows to build linear regression models for each gene of interest during the data analysis phase: a linear model is fitted to the input dataset by adjusting a set of unknown parameters, so that the dependent variable  $Y_i$ , in each observation  $i$ , is a linear combination of these parameters:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \epsilon_i$$

where:

$X_{ij}$  : predictors, the  $i$ -th observation (i.e., the different patient samples identified by different TCGA aliquots) on the  $j$ -th independent variable (i.e., target gene regulatory features) ( $i \in [1, n], j \in [1, p]$ )

$Y_i$  : target variable (i.e., the expression of the target gene)

$\beta_1, \beta_2, \dots, \beta_p$  : regression coefficients, parameters

(they quantify the effect of the features on the output: the higher is the estimated coefficient, the higher is the contribution of the feature on the prediction of the target variable)

$\epsilon_i$  : error term, observation noise

The most common estimation method, used through the *Statsmodel* library, is the *ordinary least squares*, where regression parameters are estimated such that the *Residual Sum of Squares* is minimized:

$$RSS = e_1^2 + e_2^2 + \dots + e_n^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \dots + (y_n - \hat{y}_n)^2$$

where:

$e_i = Y_i - \hat{Y}_i$  is the *residual*, i.e., the difference between the value of the dependent variable predicted by the model ( $\hat{Y}_i$ ) and the true value of the dependent variable ( $Y_i$ ).

Once the regression model is constructed, *Statsmodel* evaluates its accuracy and establishes the accuracy of the fit (i.e., how well the defined statistical model fits the given set of observations),

|             | Column_1   | Column_2   | Column_3   | Column_4   |
|-------------|------------|------------|------------|------------|
| row_index_1 | value(1,1) | value(1,2) | value(1,3) | value(1,4) |
| row_index_2 | value(2,1) | value(2,2) | value(2,3) | value(2,4) |
| row_index_3 | value(3,1) | value(3,2) | value(3,3) | value(3,4) |
| row_index_4 | value(4,1) | value(4,2) | value(4,3) | value(4,4) |
| row_index_5 | value(5,1) | value(5,2) | value(5,3) | value(5,4) |

Figure 4.8: Example of how data are organized in Pandas. A Pandas table is commonly called *Dataframe*.

computing the *R-squared* ( $R^2$ ), ranging between 0 (poor) and 1 (excellent):

$$R^2 = 1 - (RSS/TSS)$$

where:

$$TSS = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_p - \bar{y})^2 \text{ is the } Total\ Sum\ of\ Squares.$$

An  $R^2$  score close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.

In addition, the library assesses the quality of the model also focusing on the *Adjusted  $R^2$* , an extension of the original  $R^2$ , which takes into consideration the number of independent variables (i.e. features) used to fit the model, adjusting for the number of explanatory variables in the model relative to the number of data points:

$$Adj\_R^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1}$$

where:

$p$  = number of predictors in the model

$n$  = number of observations

The complete data analysis is described in [Chapter 6](#).

#### 4.7.2.4 Scikit-learn

*Scikit-learn* [49] is a Python library providing various classification, regression and clustering algorithms for performing machine learning operations and it is used in this project for normalizing input data of the linear regression process, using the *StandardScaler* class in the preprocessing module [50], consistently helping to compare results across models .

#### 4.7.2.5 Mlxtend

*Mlxtend* [51] is a Python library that extends traditional machine learning packages, providing useful tools for data science tasks. In particular, this library is used to perform the feature selection process

before fitting the actual linear regression model, thanks to the Sequential Feature Selector (SFS) component in the `feature_selection` module [52], implementing *Forward Feature Selection*.

#### 4.7.2.6 NetworkX

*NetworkX* [53] is a Python library for creating, manipulating and studying graphs and networks, used for graphically visualize linear regression results as networks.

#### 4.7.3 Cytoscape

*Cytoscape* [54] is a bioinformatics software platform for creating, manipulating, analyzing and visualizing molecular interaction networks, integrating them with gene expression profiles and other types of data. This software is used for importing and comprehensively visualizing the networks of the linear regression models, previously created in Python through the *NetworkX* library, and for performing computational analysis through ARACNe.

#### 4.7.4 ARACNe

*ARACNe* [55, 56] is a computational algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, identifying all direct transcriptional interactions. It can be used either as a standalone software or as a Cytoscape plugin [57], in order to directly visualize the analysis results within the Cytoscape platform. Here this algorithm is used from Cytoscape during the validation phase, in order to computationally validate the results of the data analysis. The complete validation process is described in [paragraph 7.2 of Chapter 7](#).

## 5. Data Extraction and Manipulation

«Now, pick up your scalpels, place them below the xiphoid process, press firmly ... no regrets, and let's begin! »

Meredith

This chapter gives the details about all the data used for the project and how they are extracted and manipulated.

### 5.1 Genes – transcription factors mapping

As comprehensive information on human genes is available online through multiple heterogeneous sources, it may be useful to have it integrated into a unique spreadsheet to be queried whenever required during the data extraction process and their manipulation.

Combining known information about genes and transcription factors from the main biological and genomic Web resources, we create a comprehensive table (a restricted set of rows is reported as an example in *Table 5.1*) which maps 41,343 genes (downloaded from the HGNC website) with all their names and, when available, their identifiers, and in case of protein-coding genes, with the transcription factors they encode (lists of TFs with their encoding genes are taken from ENCODE and UniProt websites, updated in April 2018):

- the Gene Symbol is the official current abbreviated name of the gene, approved by the HGNC and publicly available, used to uniquely identify all human genes;
- the Entrez Gene ID is the unique numerical identifier of the gene at *NCBI Gene*, a databank integrating descriptive information about genetic loci, including official nomenclature, synonyms, phenotypes, map locations and additional related external resources;
- the Ensembl Gene ID is the alphanumeric identifier of the gene within *Ensembl*, a genome browser for the retrieval of genomic information;
- the HGNC ID is a unique ID for the gene provided by HGNC;
- the RefSeq ID is the reference sequence identifier provided by *NCBI RefSeq*, a comprehensive and integrated collection of sequences, including genomic DNA, transcripts and protein products, whose identifiers are designed to provide a stable reference for gene identification and characterization, mutation analysis or expression studies;
- Transcription Factors are the proteins encoded by the genes; this means that those specific genes contain the genetic information useful for synthesizing those proteins. It is possible to find some

Table 5.1: Excerpt of the final genes – transcription factors mapping table (GenesMapping.xlsx).

| GENE_SYMBOL | ENTREZ_GENE_ID | ENSEMBL_GENE_ID | HGNC_ID | RefSeq_ID    | Transcription Factors                           | UniProt_ID |
|-------------|----------------|-----------------|---------|--------------|---|------------|
| ACLY        | 47             | ENSG00000131473 | 115     | NM_001096    | ACLY-human                                      | P53396     |
| AES         | 166            | ENSG00000104964 | 307     | NM_001130    | AES-human, eGFP-AES-human                       | Q08117     |
| BCRP4       | 616            | ENSG00000215456 | 1017    | NG_000002    |   |            |
| BRCA1       | 672            | ENSG00000012048 | 1100    | NM_007294    | BRCA1-human, eGFP-BRCA1-human, FLAG-BRCA1-human | P38398     |
| BRCA2       | 675            | ENSG00000139618 | 1101    | NM_000059    | BRCA2-human                                     | P51587     |
| BRCA3       | 60500          |                 | 18617   |              |   |            |
| CDK12       | 51755          | ENSG00000167258 | 24224   | NM_015083    | CDK12-human                                     | Q9NYV4     |
| CHEK1       | 1111           | ENSG00000149554 | 1925    | NM_001114121 | CHK1-human                                      | O14757     |
| CRYZL2P     | 730102         | ENSG00000242193 | 52164   | NR_037167    |   |            |
| DLL1        | 28514          | ENSG00000198719 | 2908    | NM_005618    | DLL1-human                                      | O00548     |
| FAM239C     | 107987330      | ENSG00000205662 | 53416   | XR_001755780 |   |            |
| HDAC1       | 3065           | ENSG00000116478 | 4852    | NM_004964    | HDAC1-human, eGFP-HDAC1-human                   | Q13547     |
| HTC1        | 3341           |                 | 5277    |              |   |            |
| MLH1        | 4292           | ENSG00000076242 | 7127    | NM_000249    | MLH1-human                                      | P40692     |
| MLLT1       | 4298           | ENSG00000130382 | 7134    | NM_005934    | ENL-human, eGFP-MLLT1-human, MLLT1-human        | Q03111     |
| MYC         | 4609           | ENSG00000136997 | 7553    | NM_001354870 | MYC-human, eGFP-MYC-human                       | P01106     |
| PTPRC       | 5788           | ENSG00000081237 | 9666    | NM_001267798 | PTPRC-human                                     | P08575     |
| RAD51       | 5888           | ENSG00000051180 | 9817    | NM_001164269 | RAD51-human, eGFP-RAD51-human                   | Q06609     |
| SDHD        | 6392           | ENSG00000204370 | 10683   | NM_001276503 | DHSD-human                                      | O14521     |
| USP17L9P    | 391627         | ENSG00000251694 | 12615   | NR_046416    |   |            |

TFs marked with an additional prefix, indicating they come from different experiments: either *eGFP*- or *FLAG*- are used, i.e., two protein tags that can be attached to native proteins to improve data quantification. Even if a TF and its tag-labelled version represent the same transcription factor, they are here considered to be distinct TFs, allowing to recognize whether a gene encodes a TF with or without an additional marking protein attached to it;

- the UniProt ID is the unique alphanumeric identifier of the transcription factor within the UniProt database.

## 5.2 Selection of transcription factors

Starting from the ENCODE transcription factors dataset and the gene promoters dataset derived from GENCODE annotations, for each gene of interest we extract the list of TFs binding to its promoters.

Data are selected through a GMQL query, which is reported below and illustrated in the block diagram of *Figure 5.1*, executed remotely from Python on the GMQL server, using the PyGMQL library (for the sake of clearness, the query is here reported according to the original GMQL syntax):

```
# Extraction of Transcription Factors:
```

```
# Extract NARROW conservative thresholded idr peaks regions from ENCODE ChIP-seq experiments,
# for assembly GRCh38 and cell lines K562 and MCF7, removing low quality data samples.
NARROW = SELECT(assay == "ChIP-seq" AND output_type == "conservative idr thresholded peaks" AND
(biosample_term_name == "K562" OR biosample_term_name == "MCF-7") AND
assembly == "GRCh38" AND project == "ENCODE" AND file_status == "released" AND
(NOT(audit_error == "extremely low read depth" OR audit_error == "extremely low read
length" OR audit_warning == "insufficient read depth" OR audit_not_compliant ==
"insufficient read depth" OR audit_not_compliant == "insufficient replicate
concordance" OR audit_not_compliant == "missing input control" OR audit_not_compliant
```

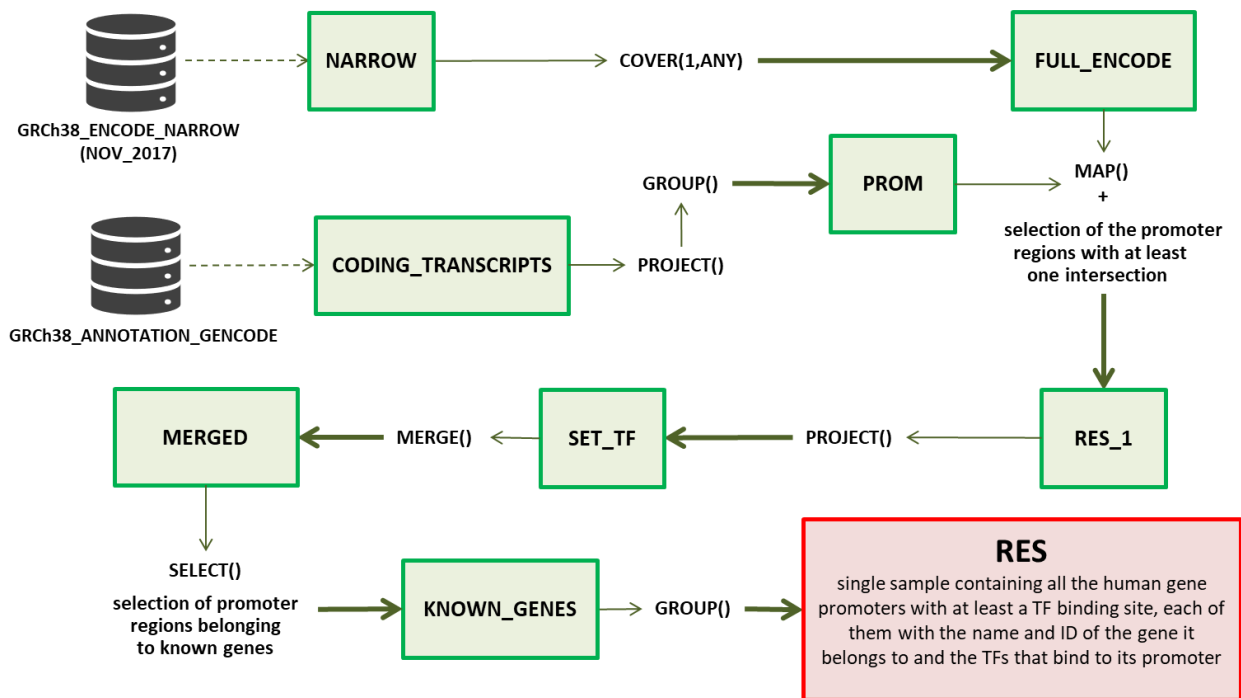


Figure 5.1: Diagram of the GMQL query for the extraction of TFs binding to human genes promoters.

```

== "severe bottlenecking" OR audit_not_compliant == "unreplicated experiment"))
GRCh38_ENCODE_NARROW_NOV_2017;

# Merge all the possible replicas of the same TF combining them in a single sample:
# the FULL_ENCODE dataset contains a data sample for each transcription factor extracted.
FULL_ENCODE = COVER(1, ANY; groupby: experiment_target) NARROW;

# Create the dataset of promoters: PROM dataset contains all the promoters of all the genes
# listed in the GENCODE version 22 annotation file. Each promoter is computed as the region
# going from 2000 bases upstream to 1000 bases downstream from the Transcription Start Site (TSS), i.e.,
# the first base of each transcript. Clearly multiple promoters correspond to multiple TSSs, which in turn
# correspond to multiple transcripts. Only known protein-coding gene transcripts are considered.

# Extract the known protein-coding transcripts from GRCh38 GENCODE annotation (version 22).
CODING_TRANSCRIPTS = SELECT(release_version == "22" AND annotation_type == "transcript";
    region: transcript_type == 'protein_coding' AND
    (tag == 'basic' OR tag == 'CCDS')) GRCh38_ANNOTATION_GENCODE;

# Compute for each transcript its promoter region.
PROM_REG = PROJECT(gene_id, gene_name, entrez_gene_id;
    region_update: start AS start - 2000, stop AS start + 1000) CODING_TRANSCRIPTS;

# Remove potential duplicated promoters (this happens when two transcripts with a different
# length, but belonging to the same gene, have the same TSS).
PROM = GROUP(region_keys: gene_name; region_aggregates: ensembl_id AS BAGD(gene_id),
    gene AS BAGD(gene_name), entrez_id AS BAGD(entrez_gene_id)) PROM_REG;

# Extract the transcription factors that overlap with at least one promoter region.
RES_0 = MAP(count_name: count_prom_TF) PROM FULL_ENCODE;
RES_1 = SELECT(region: count_prom_TF > 0) RES_0;

# Encode, for each promoter region, the TF that binds to it into a region attribute.
SET_TF = PROJECT(region_update: TF AS META(FULL_ENCODE.experiment_target, STRING)) RES_1;

```

```

# Merge all the samples into a dataset with a single sample, containing all the promoter
# regions associated with their binding TFs and remove regions belonging to unknown genes.
MERGED = MERGE() SET_TF;
KNOWN_GENES = SELECT(region: NOT(entrez_id == '')) MERGED;

# Group the regions by name, setting in the region attribute 'TFs' the list of transcription factors
# binding to that gene promoters.
RES = GROUP(region_keys: gene; region_aggregates: ensembl_gene_id AS BAGD(ensembl_id),
            gene_symbol AS BAGD(gene), entrez_gene_id AS BAGD(entrez_id),
            TFs AS BAGD(TF)) KNOWN_GENES;

# The final dataset is created and saved in the system.
MATERIALIZED RES INTO RES;

```

The first step of the query selects transcription factors in cell lines K562 or MCF7 relative to NARROW data resulting from ChIP-seq experiments, retrieving only higher quality regions (conservative `idr` thresholded peaks parameter) and removing low quality samples (the negative audit parameters).

Since multiple replicated samples, relative to different experiments, are possible for each TF, all the replicas are merged, using the GMQL operation `COVER(1, ANY; groupby: experiment_target)`: for each TF, all possible replicas are combined into a single sample: the regions of this samples are defined as the ones with at least one base in one of the experiments. In this case, the `COVER` operation considers all those areas defined by a minimum of one overlapping region in the input samples, up to any amount of overlapping regions. An example is reported in *Figure 5.2*.

In general, `COVER` parameters (`minACC`, `maxACC`) can vary and they can be customized according to the user's needs: `minACC` and `maxACC` respectively define the minimum and the maximum number of overlapping regions to be considered during `COVER` execution (i.e., the minimum and the maximum number of replicas where a region has to be contained). As a result, our `FULL_ENCODE` dataset contains as many samples as the number of transcription factors in the considered cell lines ( $n_{TF} = 276$ ).

In the second step of the query, the human gene promoters dataset is built (`PROM`,  $n_{PROM} = 43529$ ). Since no GENCODE annotations are initially present in the GMQL system, we implement a specific and parametric Python script for converting any type of GENCODE genomic annotation file into a set of GMQL data samples according to the GDM framework, and we finally upload them on the GMQL system. Specifically, versions 10 and 19 for assembly HG19 and versions 22, 24 and 27 for assembly GRCh38 are now publicly available in the system for all the end users, within the two datasets `HG19_ANNOTATIONS_GENCODE` and `GRCh38_ANNOTATIONS_GENCODE`.

Starting from all known protein-coding transcripts of annotated genes, the `PROJECT` operation is able to automatically manage the strand and correctly compute the promoter regions around each gene TSSs, i.e., those regions going from 2000 bases upstream to 1000 bases downstream from the TSSs (multiple gene transcripts mean multiple TSSs, and multiple TSSs mean multiple promoters):

- if `strand == '+'`, then the coordinates of the promoters are `[TSS - 2000, TSS + 1000]`
- if `strand == '-'`, then the coordinates of the promoters are `[TSS + 2000, TSS - 1000]`

*Figure 5.3* shows a representation of the genomic coordinate system used for identifying regions of the genome and illustrates how gene promoters are computed according to the DNA strand.



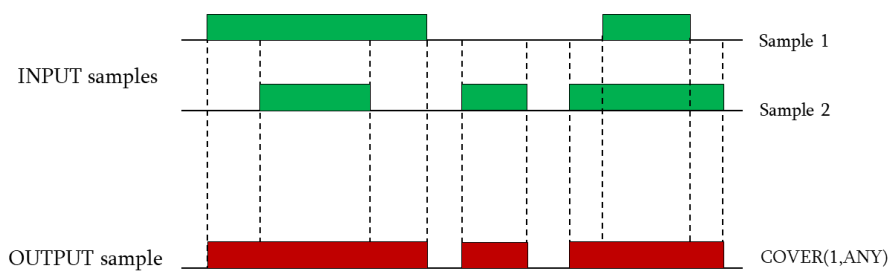


Figure 5.2: COVER(1,ANY) example.

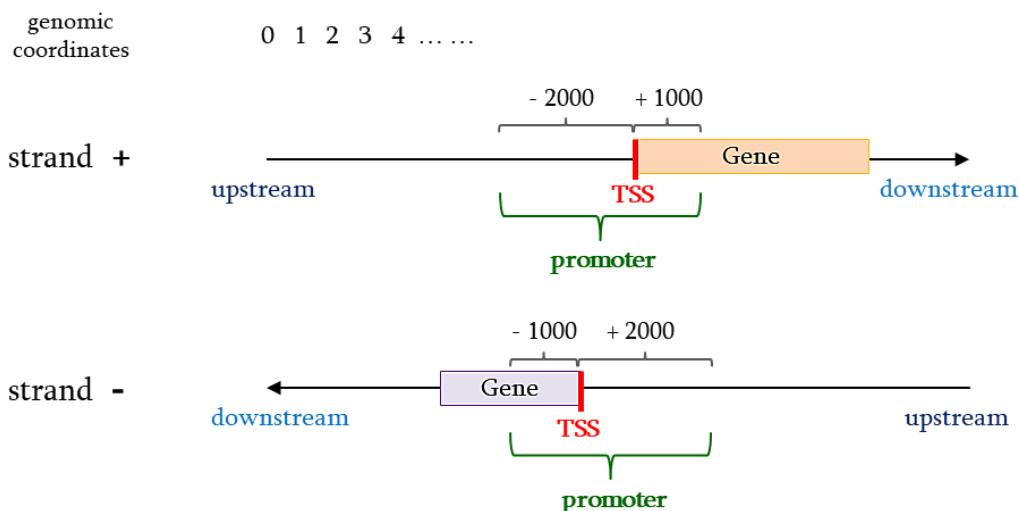


Figure 5.3: Description of the genomic coordinates system and computation of gene promoters.

The last step of the query compares TFs data and promoter regions, with the objective of extracting the transcription factors with binding sites overlapping with at least one promoter. This is done using the MAP operation (explained in Figure 5.4), which allows to map all the regions of an *experiment* dataset over the regions of a *reference* dataset, by automatically counting the number of experiment regions intersecting a certain reference region. In this case, TFs are mapped on promoters: for each TF sample, this operation counts the values of TF regions intersecting with a promoter region, for any promoter of each sample in the PROM dataset, generating dataset RES\_0 with  $n_{TF}$  samples, each one containing  $n_{PROM}$  regions. Finally, only TFs with at least one intersection are selected (RES\_1).

At the end, the TF binding to each promoter region is encoded as a region attribute through the PROJECT operation and the regions of the resulting dataset are merged into a single sample. Promoters are grouped by gene, collecting the list of its associated TFs in a single region attribute, and the dataset is finally “materialized” (i.e., created and saved).

Thanks to the PyGMQL library, the output dataset, containing all gene promoters, together with their associated TFs, is directly stored into a specific data structure (*res\_df*), called *GDataFrame*, which

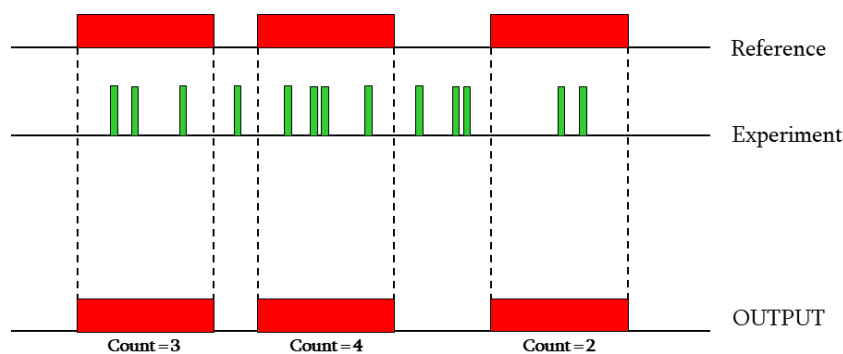


Figure 5.4: Example of the GMQL MAP operation between two datasets containing one single data sample each.

```
# Materialize the query results into a GDataframe
res_Gdf = res.materialize('./MaterializeResults')

# Extract the regions dataframe, where each row corresponds to a promoter region and each column to a region attribute
GeneTF_df = res_gdf.regs

# Test the dataframe showing its first rows
GeneTF_df.head()
```

| CHROM | LEFT  | RIGHT     | STRAND    | GENE_SYMBOL | ENSEMBL_GENE_ID | ENTREZ_GENE_ID     | TFs   |   |
|-------|-------|-----------|-----------|-------------|-----------------|--------------------|-------|---|
| 0     | chr1  | 155688000 | 155691000 | -           | YY1AP1          | ENSG00000163374.18 | 55249 | [AGO1-human, ARID1B-human, ATF7-human, BCLAF1-... |
| 1     | chr16 | 81493348  | 81496348  | +           | CMIP            | ENSG00000153815.15 | 80790 | [ARID1B-human, ARNT-human, ATF3-human, ATF7-hu... |
| 2     | chr8  | 144825558 | 144828558 | +           | ZNF7            | ENSG00000147789.14 | 7553  | [ATF7-human, C11orf30-human, CBX1-human, CC2D1... |
| 3     | chr11 | 46345571  | 46348571  | +           | DGKZ            | ENSG00000149091.14 | 8525  | [ARID1B-human, BCOR-human, BHLHE40-human, CHD1... |
| 4     | chr4  | 2417663   | 2420663   | -           | ZFYVE28         | ENSG00000159733.12 | 57732 | [ATF7-human, BHLHE40-human, BMI1-human, C11orf... |

Figure 5.5: Python materialization of the RES dataframe and structure of the GeneTF\_df dataframe.

comprises two pandas dataframes, one for the regions (GeneTF\_df) and one for the metadata. Figure 5.5 shows the lines of code used for “materializing” the query and extracting the region dataframe.

In order to extract the lists of TFs, we simply perform an iteration along the region dataframe and select, for each gene of interest, the row of the dataframe associated to it, retrieving its list of TFs from the proper column. We store the results into a Python dictionary, a data structure mapping a set of *keys* (i.e., the target genes) to a set of *values* (i.e., the list of transcription factors, with two additional parameters storing the gene ID and its pathways), with the following structure:

```
{GENE_SYMBOL_1: [TF1_1, TF1_2, ..., [ENTREZ_GENE_ID_1],[PATHWAY1_1, ...]],
  GENE_SYMBOL_2: [TF2_1, TF2_2, ..., [ENTREZ_GENE_ID_2],[PATHWAY2_1, ...]], ...,
  ...
  GENE_SYMBOL_176: [TF176_1, TF176_2, ..., [ENTREZ_GENE_ID_176],[PATHWAY176_1, ...]]}
```

Figure 5.6 displays the Python code implemented for extracting the complete list of transcription factors associated to each gene of interest: iterating along the GeneTF\_df dataframe, regions belonging to target genes are identified, their TFs are retrieved and finally stored in the dictionary. Since each gene may have more than one promoter bound to the same TF, we consider only distinct values for each transcription factors, ensuring that all TFs of interest are retrieved, with no useless duplicates.

```

# Select from the GeneTF_df only the rows with Gene Symbols of target genes of interest
for index, row in GeneTF_df.iterrows(): # iterate along the whole dataframe
    # get the current row GENE_SYMBOL
    i = row['GENE_SYMBOL']
    for value in Symbols: # check if the current gene is contained in the list of genes of interest
        if i == value: # if there's correspondence
            TrFa_list = row.TFs # extract the list of TFs
            for t in TrFa_list:
                # since each gene can have more than one promoter bound by the same TF,
                # only distinct values for each transcription factor should be inserted in the dictionary
                if t not in dict_GeneTF[i]:
                    # add the transcription factor to the list of values corresponding to that gene
                    dict_GeneTF[i].append(t)

```

Figure 5.6: Main loop used for extracting each target gene list of TFs from the initial dataset.

```

# Get the TFs of each target gene and extract the names of the genes encoding them from the mapping dataframe
for key, value in dict_GeneTF.items():
    TFs = value[:-2] # the TFs are all the elements of the value list, except for the last two
    for tf in TFs:
        # for each TF, search in the mapping dataframe for the name of the encoding gene
        if tf in mapping_df_TFs:
            # get the name (GENE_SYMBOL) of the gene encoding the transcription factor "tf"
            gene_name = TF_Gene_df.loc[TF_Gene_df['TF_NAME'] == tf, 'GENE_SYMBOL'].iloc[0]
            # add the regulatory gene in correspondence of the proper gene in the dictionary
            dict_RegulGenes[key].append(gene_name)
        # in case the transcription factor considered is not mapped in the dataframe,
        # then the name of its encoding gene is unknown ('n/a')
    else: dict_RegulGenes[key].append('n/a')

```

Figure 5.7: Main loop used for identifying candidate regulatory genes for each gene of interest, starting from the list of TFs previously extracted.

### 5.3 Identification of candidate regulatory genes

Rather than the transcription factors themselves, we are actually interested in the set of candidate regulatory genes of the genes of interest, i.e., the set of genes synthesizing these proteins.

Regulatory genes are extracted by integrating the dictionary of TFs with the genes-transcription factors mapping table previously created. *Figure 5.7* displays the Python code implemented for identifying the set of candidate regulatory genes: for all genes, the implemented script looks up each protein stored in its TFs list and retrieves the name of the corresponding encoding gene from the mapping table.

To stay consistent with the procedure performed for the selection of TFs and save this data into a convenient structure, we store candidate regulators into another Python dictionary as follows:

```

{GENE_SYMBOL_1: [REGULATORY_GENE_SYMBOL1,1, REGULATORY_GENE_SYMBOL1,2, ...],
  GENE_SYMBOL_2: [REGULATORY_GENE_SYMBOL2,1, REGULATORY_GENE_SYMBOL2,2, ...], ... ,
  ...
  GENE_SYMBOL_176: [REGULATORY_GENE_SYMBOL176,1, REGULATORY_GENE_SYMBOL176,2, ...]}

```

Finally, we summarize all information extracted so far into a comprehensive table (its structure is illustrated in *Table 5.2*): for each gene of interest, identified by its own *Gene Symbol*, there is a progressively enumerated list of all the TFs binding to its promoters, where each TF is associated with its encoding gene, in turn reported with both its own *Gene Symbol* and *Entrez Gene ID*. The table is

Table 5.2: Excerpt of the summary table with final results of TFs and candidate regulatory genes extraction phases (Full TFs-RegulatoryGenes SUMMARY Table.xlsx).

| GENE_SYMBOL | # | Transcription Factors | Regulatory Genes | Entrez_Gene_IDs | ENTREZ_GENE_ID | PATHWAYS           | #TFs | #RegulatoryGenes (distinct) |
|-------------|---|-----------------------|------------------|-----------------|----------------|--------------------|------|-----------------------------|
| ABCB5       | 1 | ATF2-human            | ATF2             | 1386            | 340273         | STEM_CELLS         | 4    | 4                           |
|             | 2 | CEBPB-human           | CEBPB            | 1051            |                |                    |      |                             |
|             | 3 | CTCF-human            | CTCF             | 10664           |                |                    |      |                             |
|             | 4 | eGFP-CEBPG-human      | CEBPG            | 1054            |                |                    |      |                             |
| PGK2        | 1 | DDX20-human           | DDX20            | 11218           | 5232           | GLUCOSE_METABOLISM | 8    | 7                           |
|             | 2 | eGFP-PBX2-human       | PBX2             | 5089            |                |                    |      |                             |
|             | 3 | FLAG-PBX2-human       | PBX2             | 5089            |                |                    |      |                             |
|             | 4 | FOS-human             | FOS              | 2353            |                |                    |      |                             |
|             | 5 | PKNOX1-human          | PKNOX1           | 5316            |                |                    |      |                             |
|             | 6 | RFX1-human            | RFX1             | 5989            |                |                    |      |                             |
|             | 7 | RFX5-human            | RFX5             | 5993            |                |                    |      |                             |
|             | 8 | ZBTB33-human          | ZBTB33           | 10009           |                |                    |      |                             |
| BRCA1       | 1 | ARID1B-human          | ARID1B           | 57492           | 672            | DNA_REPAIR         | 132  | 126                         |
|             | 2 | ARNT-human            | ARNT             | 405             |                |                    |      |                             |
|             | 3 | ATF2-human            | ATF2             | 1386            |                |                    |      |                             |
|             | 4 | ATF3-human            | ATF3             | 467             |                |                    |      |                             |
|             | 5 | ATF7-human            | ATF7             | 11016           |                |                    |      |                             |
|             | 6 | BCLAF1-human          | BCLAF1           | 9774            |                |                    |      |                             |
|             | 7 | ...                   | ...              | ...             |                |                    |      |                             |

completed by the target gene *Entrez Gene ID* and *pathways*, plus the total number of its associated TFs and distinct encoding genes (since a gene may encode for multiple TFs, the number of distinct candidate regulatory genes is generally smaller, or equal, to the number of distinct TFs). This chart is very useful for the biologists, because it allows them to easily keep track of those candidate regulatory genes encoding for multiple TFs and to identify all the different gene-TF associations, discerning among TFs attached to a protein tag from the ones that are not.

Figure 5.8 shows the complete flowchart of the Python script used for the extraction of candidate regulatory genes ("Extraction\_RegulatoryGenes.py"). A flowchart for each Python script is then reported in [Appendix A](#).

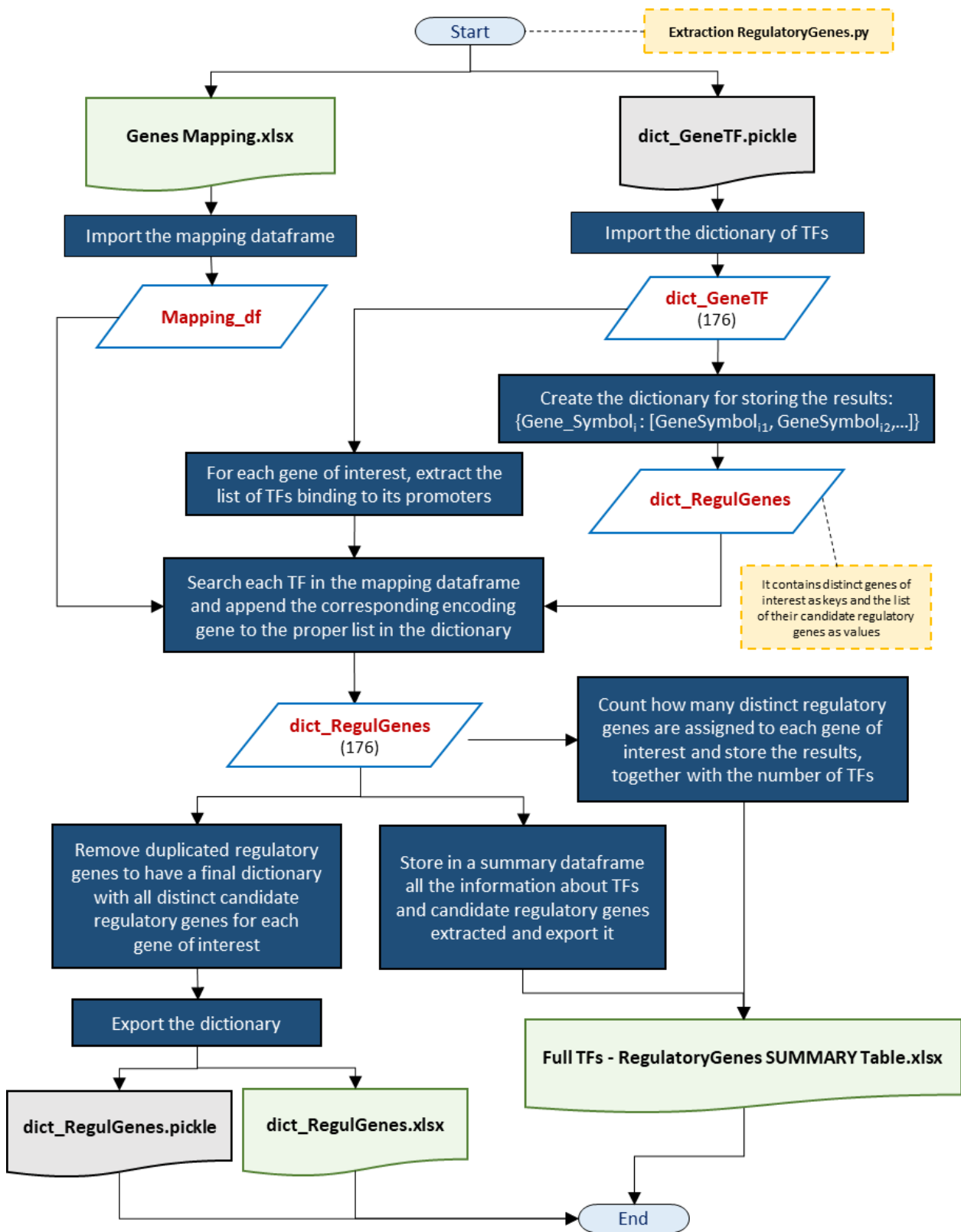


Figure 5.8: Flowchart for the extraction of candidate regulatory genes of the genes of interest.

## 5.4 Extraction of methylation and gene expression values

The next step consists in extracting target gene methylation and gene expression values, in both ovarian and breast cancers, retrieving them from the GRCh38\_TCGA\_methylation and GRCh38\_TCGA\_gene\_expression datasets in the GMQL system.

Measures on ovarian cancer are collected for the analysis, while the subset of breast cancer information is later used for validating the ovarian cancer data analysis results.

### 5.4.1 Ovarian cancer

Tumor-specific patients data for ovarian cancer is the most interesting dataset for our analysis. We collect these data by filtering the original TCGA datasets with the following query (illustrated in the block diagram of *Figure 5.9*), remotely executed from Python on the GMQL system, using the PyGMQL library:

```
# Extract methylation and expression information for OV tumor:

# Initial datasets:
# Extract all the samples for tumor OV and platform 27k and exclude, if present, either normal or
# metastatic samples, retrieving only primary or recurrent appearance tumor samples.
ALL_OV_METHYL = SELECT(manually_curated_cases_disease_type == "Ovarian Serous
    Cystadenocarcinoma" AND manually_curated_platform == "Illumina Human Methylation 27" AND
    (biospecimen_bio_sample_type == "Primary Tumor" OR biospecimen_bio_sample_type ==
    "Recurrent Tumor") AND clinical_shared_history_of_neoadjuvant_treatment == "No")
    GRCh38_TCGA_methylation;

# Extract all the gene expression data for tumor OV and exclude, if present, either normal or metastatic
# samples, retrieving only primary or recurrent appearance tumor samples.
ALL_OV_EXPR = SELECT(manually_curated_cases_disease_type == "Ovarian Serous
    Cystadenocarcinoma" AND (biospecimen_bio_sample_type == "Primary Tumor" OR
    biospecimen_bio_sample_type == "Recurrent Tumor") AND
    clinical_shared_history_of_neoadjuvant_treatment == "No") GRCh38_TCGA_gene_expression;

# Keep only those region attributes that are useful for the scope of the project, in order to lighten the
# computational time complexity of the query, and select the common samples by extracting only those aliquots
# for which both methylation and expression values are available.

# Gene Expression (OV):
OV_EXPR_0 = PROJECT(ensembl_gene_id, entrez_gene_id, gene_symbol, fpkm) ALL_OV_EXPR;
OV_EXPR = SELECT(semijoin: biospecimen_bio_bcr_sample_barcode IN ALL_OV_METHYL) OV_EXPR_0;
# The final dataset is created and saved in the system.
MATERIALIZED OV_EXPR INTO OV_EXPR;

# Methylation (OV):
OV_METHYL_0 = PROJECT(beta_value) ALL_OV_METHYL;
OV_METHYL = SELECT(semijoin: biospecimen_bio_bcr_sample_barcode IN ALL_OV_EXPR) OV_METHYL_0;
# The final dataset is created and saved in the system.
MATERIALIZED OV_METHYL INTO OV_METHYL;

# These output datasets are then used in Python to extract methylation sites located in regions -4000/+1000
# around the TSSs of the genes of interest and to extract target genes and their candidate regulatory genes
# expression values.
```

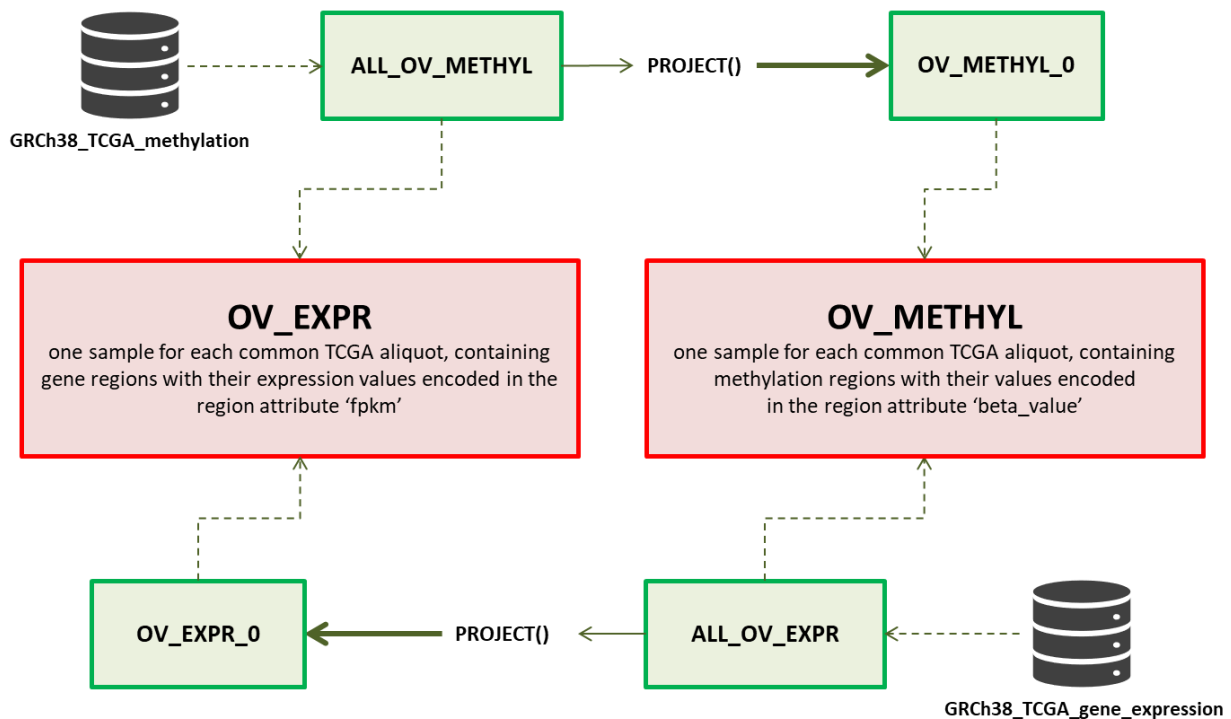


Figure 5.9: Diagram of the GMQL query for the extraction of OV tumor data.

Specifically, we focus on primary tumor samples that were not previously treated with any neo-adjuvant treatment and, if present, samples coming from patients with recurrent tumor, excluding both metastatic tumor and normal samples, which may alter the results of our analysis.

In order to minimize the number of missing values selected from TCGA, which would represent only additional noise for the analysis, we select only common samples having both methylation and expression values in the two initial TCGA datasets. This filtering is executed by setting the `semijoin` parameter in the `SELECT` operation: it executes the selection operation on the basis of the existence of one or more metadata attributes and the matching of their values with the ones associated with at least one sample in another dataset.

The output datasets are finally “materialized” using PyGMQL into two GDataframes (`methy1_Gdf` and `expr_Gdf`), so that their corresponding region (`methy1_df_regs` and `expr_df_regs`) and metadata (`methy1_df_meta` and `expr_df_meta`) dataframes are immediately available in the Python environment for completing the extraction procedure.

As for methylation data, we are only interested in methylation sites falling within target genes promoters or in a slightly wider area, because, as previously mentioned, the methylated promoter mostly influences the gene expression regulation, by considerably reducing the expression (methylation sites localized in other regions of the genes are not relevant for this project). This is the reason why we only retrieve methylations located in genomic areas going from 4000 bases upstream to 1000 bases downstream from the TSSs of the genes of interest. *Figure 5.10* illustrates an example of methylation sites of interest (represented as red circles).

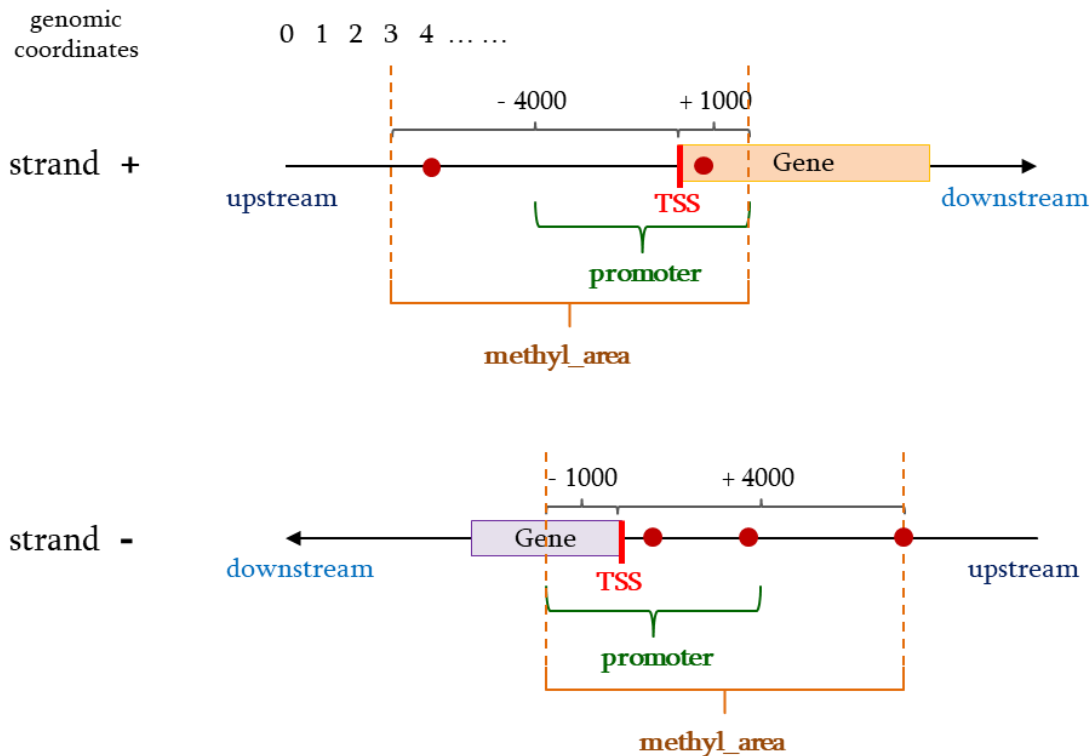


Figure 5.10: Example of methylation sites (red circles) considered in the analysis.

These areas, identified as *methyl\_areas* for the sake of convenience, are extracted through the following GMQL query (illustrated in the block diagram of Figure 5.11):

```
# Computation of methyl_areas:

# Extract the known protein-coding transcripts from GRCh38 GENCODE annotation (version 22).
CODING_TRANSCRIPTS = SELECT(release_version == "22" AND annotation_type == "transcript";
    region: transcript_type == 'protein_coding' AND
    (tag == 'basic' OR tag == 'CCDS')) GRCh38_ANNOTATION_GENCODE;

# Compute the 'methyl_areas' for each transcript of each annotated gene
METHYL_AREAS_REG = PROJECT(gene_id, gene_name, entrez_gene_id; region_update: start AS
    start - 4000, stop AS start + 1000) CODING_TRANSCRIPTS;

# Remove potential duplicated methyl_areas (this happens when two transcripts with a different
# length, but belonging to the same gene, have the same TSS).
GENCODE_GRCh38_METHYL_AREAS = GROUP(region_keys: gene_name; region_aggregates: ensembl_gene_id
    AS BAGD(gene_id), gene_symbol AS BAGD(gene_name), entrez_id
    AS BAGD(entrez_gene_id)) METHYL_AREAS_REG;

# The final dataset is created and saved in the system.
MATERIALIZED GENCODE_GRCh38_METHYL_AREAS INTO GENCODE_GRCh38_METHYL_AREAS;
```

Iterating along the GENCODE\_GRCh38\_METHYL\_AREAS dataset (once filtered for the target genes) and combining it with the methyl\_df\_regs dataframe, for each gene of interest and for all its TCGA aliquot, we select all the *beta\_values* quantifying the methylation level of the CpG sites in its *methyl\_areas*. If a gene is associated with more than one *beta\_value* in the same aliquot, its mean value is computed, in





Figure 5.11: Diagram of the GMQL query for the extraction of *methyl\_areas*.

order to have one single value approximating the absolute level of methylation for every target gene in each aliquot.

We use a Python script quite similar to the one written for the methylation, in order to extract gene expression values of target genes and of their candidate regulatory genes, in each TCGA aliquot, by iteration along the `expr_df_regs` dataframe.

Finally, extracted methylation and expression values are stored and exported in three distinct data matrices, with the set of TCGA aliquots (372 OV tumor patients) as rows and either the full list of distinct genes of interest (176) or candidate regulatory genes (249) as columns. Additional information, such as the gene IDs, the data samples names, the tumor label and the patient ID, is also added, for any future needs whatsoever. *Table 5.3* shows an excerpt of the final tables containing TCGA data for all the ovarian cancer patients under observation.

In the end, there are only 12 genes of interest with no methylation values in TCGA:

- gene 5091 [PC] has indeed null methylation values (i.e. *mean\_beta\_value* = 0) when measured in the patients of interest; however, for the sake of simplicity, these null regions are removed from the TCGA data uploaded to the GMQL system. Thus, they are actually zero when it comes to the data analysis;
- genes 100133941 [CD24], 54567 [DLL4], 3065 [HDAC1], 3099 [HK2], 3717 [JAK2], 389421 [LIN28B], 79923 [NANOG], 4853 [NOTCH2], 55066 [PDPR], 10721 [POLQ] and 8801 [SUCLG2] have no methylation probes at all in the TCGA data. This means that their methylation is not measured in the TCGA tumor patients and so their *beta\_values* are not available (i.e. *NaN*).

Information on gene expression is complete for all the target genes and their candidate regulators, with the only exception of regulatory gene 56947 [EMSY], that has no available values in the TCGA samples.

Therefore, before carrying out the complete data analysis, we discard features corresponding to the expression of EMSY and to the methylation the 11 genes listed before.

Table 5.3: Excerpt of the final tables containing data about DNA methylation and gene expression for each TCGA aliquot under analysis.

a) Target genes methylation (MethylationValues.xlsx).

|                  | ABC5       | ALDH1A1    | BMI1       | BMP7       | CD24      | CHEK1      | PC   | ... | Sample_ID   | Tumor | Patient_ID |
|------------------|------------|------------|------------|------------|-----------|------------|------|-----|-------------|-------|------------|
| ENTREZ_GENE_ID   | 340273     | 216        | 648        | 655        | 100133941 | 1111       | 5091 | ... |             |       |            |
| TCGA-36-1569-01A | 0.88004863 | 0.08121697 | 0.01768676 | 0.01620633 |           | 0.03973334 | 0    | ... | S_00000.gdm | OV    | 1569       |
| TCGA-36-1568-01A | 0.80510806 | 0.06154271 | 0.01616918 | 0.01563498 |           | 0.05202452 | 0    | ... | S_00001.gdm | OV    | 1568       |
| TCGA-04-1332-01A | 0.85770674 | 0.17624142 | 0.01054901 | 0.01294314 |           | 0.06774251 | 0    | ... | S_00006.gdm | OV    | 1332       |
| TCGA-23-1809-01A | 0.89217838 | 0.0920498  | 0.01157163 | 0.00792197 |           | 0.0517871  | 0    | ... | S_00007.gdm | OV    | 1809       |
| TCGA-23-1030-01A | 0.88117329 | 0.13621047 | 0.00891415 | 0.01029035 |           | 0.05390399 | 0    | ... | S_00010.gdm | OV    | 1030       |
| TCGA-61-1998-01A | 0.91447765 | 0.1339391  | 0.0295399  | 0.01213231 |           | 0.06598633 | 0    | ... | S_00011.gdm | OV    | 1998       |
| TCGA-20-1686-01A | 0.67479441 | 0.23423499 | 0.01265658 | 0.09237621 |           | 0.04534391 | 0    | ... | S_00012.gdm | OV    | 1686       |
| TCGA-13-0727-01A | 0.85419375 | 0.09855298 | 0.00980201 | 0.01306752 |           | 0.08505444 | 0    | ... | S_00014.gdm | OV    | 0727       |
| TCGA-25-1312-01A | 0.892138   | 0.07964036 | 0.01159257 | 0.00707971 |           | 0.04597742 | 0    | ... | S_00016.gdm | OV    | 1312       |
| TCGA-13-0905-01B | 0.91930939 | 0.06060005 | 0.00851126 | 0.01182553 |           | 0.04274153 | 0    | ... | S_00017.gdm | OV    | 0905       |
| TCGA-09-1666-01A | 0.79116528 | 0          | 0.02504361 | 0.01464563 |           | 0.11030122 | 0    | ... | S_00018.gdm | OV    | 1666       |
| TCGA-24-1546-01A | 0.84589247 | 0.04815934 | 0.00998333 | 0.01502781 |           | 0.04566988 | 0    | ... | S_00019.gdm | OV    | 1546       |
| TCGA-29-2414-01A | 0.86408993 | 0.10254702 | 0.01723743 | 0.00926446 |           | 0.05604113 | 0    | ... | S_00020.gdm | OV    | 2414       |
| TCGA-36-1576-01A | 0.87646228 | 0.09154352 | 0.01151242 | 0.0283414  |           | 0.04300342 | 0    | ... | S_00022.gdm | OV    | 1576       |
| TCGA-24-1435-01A | 0.56241766 | 0.22121593 | 0.02533332 | 0.02246621 |           | 0.12276458 | 0    | ... | S_00023.gdm | OV    | 1435       |
| TCGA-24-1416-01A | 0.79059439 | 0.20998184 | 0.01735423 | 0.0180661  |           | 0.10238849 | 0    | ... | S_00024.gdm | OV    | 1416       |
| TCGA-09-1665-01B | 0.76770618 | 0.23475299 | 0.01425861 | 0.00747348 |           | 0.10529026 | 0    | ... | S_00025.gdm | OV    | 1665       |
| TCGA-24-2254-01A | 0.7842761  | 0.09594699 | 0.02247552 | 0.02142992 |           | 0.06503955 | 0    | ... | S_00027.gdm | OV    | 2254       |
| TCGA-13-1487-01A | 0.70895062 | 0.31496744 | 0.01737893 | 0.02254837 |           | 0.16830201 | 0    | ... | S_00029.gdm | OV    | 1487       |
| TCGA-24-1417-01A | 0.59887862 | 0.29289423 | 0.02089882 | 0.02645135 |           | 0.08981947 | 0    | ... | S_00031.gdm | OV    | 1417       |
| ...              | ...        | ...        | ...        | ...        | ...       | ...        | ...  | ... | ...         | ...   | ...        |

b) Target genes expression (GeneExpression - InterestGenes.xlsx).

|                  | ABC5     | ALDH1A1   | BMI1      | BMP7       | CD24       | CHEK1     | PC       | ... | Sample_ID   | Tumor | Patient_ID |
|------------------|----------|-----------|-----------|------------|------------|-----------|----------|-----|-------------|-------|------------|
| ENTREZ_GENE_ID   | 340273   | 216       | 648       | 655        | 100133941  | 1111      | 5091     | ... |             |       |            |
| TCGA-04-1364-01A | 0.02291  | 1.998453  | 13.651514 | 1.692225   | 299.441479 | 5.987776  | 2.356133 | ... | S_00000.gdm | OV    | 1364       |
| TCGA-61-2002-01A | 0.009571 | 3.692858  | 3.418337  | 1.150803   | 246.010375 | 10.449857 | 2.711215 | ... | S_00001.gdm | OV    | 2002       |
| TCGA-57-1586-01A | 0.027514 | 7.151855  | 5.012184  | 11.090795  | 293.09535  | 1.927998  | 2.078006 | ... | S_00002.gdm | OV    | 1586       |
| TCGA-61-2097-01A | 0        | 5.631069  | 12.604977 | 9.842409   | 220.883809 | 1.374333  | 1.571335 | ... | S_00003.gdm | OV    | 2097       |
| TCGA-23-1122-01A | 0.005616 | 6.071856  | 7.816883  | 35.919864  | 724.546495 | 3.107495  | 3.780404 | ... | S_00004.gdm | OV    | 1122       |
| TCGA-24-1470-01A | 0.001497 | 3.764909  | 6.917762  | 0.365766   | 58.36363   | 6.007744  | 4.047326 | ... | S_00005.gdm | OV    | 1470       |
| TCGA-24-1552-01A | 0.019144 | 2.273179  | 8.278887  | 0.074336   | 262.532741 | 3.962442  | 3.144839 | ... | S_00006.gdm | OV    | 1552       |
| TCGA-29-1776-01A | 0.010634 | 3.068088  | 10.420312 | 31.205439  | 192.025111 | 5.976915  | 6.872311 | ... | S_00007.gdm | OV    | 1776       |
| TCGA-29-1784-01A | 0        | 11.938886 | 14.020578 | 20.217522  | 140.167166 | 4.794879  | 6.11142  | ... | S_00008.gdm | OV    | 1784       |
| TCGA-36-1576-01A | 0.036987 | 8.189365  | 9.612057  | 24.826643  | 156.883801 | 5.176216  | 3.59262  | ... | S_00009.gdm | OV    | 1576       |
| TCGA-13-1505-01A | 0.018376 | 15.178873 | 13.891931 | 10.134051  | 153.286014 | 3.20198   | 2.519232 | ... | S_00010.gdm | OV    | 1505       |
| TCGA-10-0933-01A | 0.026474 | 15.251822 | 24.441218 | 0.563663   | 201.296503 | 5.289196  | 0.744259 | ... | S_00011.gdm | OV    | 0933       |
| TCGA-61-1738-01A | 0.018394 | 1.600411  | 14.519727 | 3.303383   | 2532.3069  | 2.232936  | 4.318866 | ... | S_00012.gdm | OV    | 1738       |
| TCGA-04-1347-01A | 0.309587 | 1.660109  | 17.785528 | 124.095932 | 2105.55512 | 2.395861  | 1.829276 | ... | S_00013.gdm | OV    | 1347       |
| TCGA-25-1635-01A | 0.010763 | 11.575235 | 11.162415 | 14.910754  | 242.402065 | 2.430764  | 7.127303 | ... | S_00014.gdm | OV    | 1635       |
| TCGA-10-0927-01A | 0.071264 | 9.067047  | 11.251924 | 56.938212  | 84.78876   | 1.969504  | 0.714678 | ... | S_00015.gdm | OV    | 0927       |
| TCGA-25-1313-01A | 0.001689 | 18.468666 | 6.783292  | 9.307512   | 252.538749 | 1.361469  | 2.706795 | ... | S_00016.gdm | OV    | 1313       |
| TCGA-61-2009-01A | 0.014768 | 9.08691   | 12.069058 | 14.713253  | 408.593683 | 1.592084  | 1.46462  | ... | S_00017.gdm | OV    | 2009       |
| TCGA-30-1862-01A | 0.025827 | 5.635403  | 17.888518 | 0.268158   | 362.058013 | 1.937804  | 7.763464 | ... | S_00018.gdm | OV    | 1862       |
| TCGA-23-1029-01B | 0.02423  | 11.374037 | 15.53838  | 50.445761  | 258.662538 | 7.309248  | 4.198416 | ... | S_00019.gdm | OV    | 1029       |
| ...              | ...      | ...       | ...       | ...        | ...        | ...       | ...      | ... | ...         | ...   | ...        |

c) Candidate regulatory genes expression (GeneExpression - RegulatoryGenes.xlsx).

|                  | ATF2      | CEBPB      | CTCF      | KLF1     | ZNF687    | DPF2      | EMSY  | ... | Sample_ID   | Tumor | Patient_ID |
|------------------|-----------|------------|-----------|----------|-----------|-----------|-------|-----|-------------|-------|------------|
| ENTREZ_GENE_ID   | 1386      | 1051       | 10664     | 10661    | 57592     | 5977      | 56946 | ... |             |       |            |
| TCGA-04-1364-01A | 8.626486  | 105.782551 | 15.050922 | 0.461137 | 14.433112 | 11.065619 |       | ... | S_00000.gdm | OV    | 1364       |
| TCGA-61-2002-01A | 10.510457 | 86.270449  | 15.799646 | 0.159999 | 12.744001 | 8.77075   |       | ... | S_00001.gdm | OV    | 2002       |
| TCGA-57-1586-01A | 8.066589  | 146.126198 | 9.986646  | 0.125702 | 70.810176 | 12.384597 |       | ... | S_00002.gdm | OV    | 1586       |
| TCGA-61-2097-01A | 18.936805 | 192.852163 | 13.376466 | 0.301939 | 14.503313 | 13.294428 |       | ... | S_00003.gdm | OV    | 2097       |
| TCGA-23-1122-01A | 18.481601 | 83.243196  | 19.868367 | 0.114293 | 15.13596  | 10.013334 |       | ... | S_00004.gdm | OV    | 1122       |
| TCGA-24-1470-01A | 9.212276  | 166.113877 | 15.333887 | 0.235032 | 14.31     | 13.807878 |       | ... | S_00005.gdm | OV    | 1470       |
| TCGA-24-1552-01A | 7.118054  | 59.968358  | 8.150597  | 0.121433 | 14.270525 | 8.459142  |       | ... | S_00006.gdm | OV    | 1552       |
| TCGA-29-1776-01A | 9.092053  | 87.816039  | 17.732396 | 0.111295 | 18.145805 | 10.755706 |       | ... | S_00007.gdm | OV    | 1776       |
| TCGA-29-1784-01A | 7.928575  | 192.260107 | 10.800649 | 0.264238 | 12.41611  | 10.790263 |       | ... | S_00008.gdm | OV    | 1784       |
| TCGA-36-1576-01A | 10.710793 | 114.813848 | 15.709174 | 0.112656 | 13.677253 | 16.29588  |       | ... | S_00009.gdm | OV    | 1576       |
| TCGA-13-1505-01A | 11.883956 | 104.204626 | 14.238084 | 0.094978 | 16.798894 | 9.449083  |       | ... | S_00010.gdm | OV    | 1505       |
| TCGA-10-0933-01A | 14.37373  | 201.370561 | 15.095133 | 0.059207 | 12.924882 | 7.885083  |       | ... | S_00011.gdm | OV    | 0933       |
| TCGA-61-1738-01A | 9.829785  | 148.037349 | 9.589454  | 0.166376 | 23.475641 | 8.953291  |       | ... | S_00012.gdm | OV    | 1738       |
| TCGA-04-1347-01A | 4.878672  | 68.739427  | 7.515916  | 0.579071 | 16.849624 | 7.005869  |       | ... | S_00013.gdm | OV    | 1347       |
| TCGA-25-1635-01A | 7.031113  | 148.267617 | 18.382765 | 0.107284 | 15.533219 | 6.822243  |       | ... | S_00014.gdm | OV    | 1635       |
| TCGA-10-0927-01A | 9.249128  | 77.996766  | 9.751757  | 0.872373 | 11.117764 | 8.397714  |       | ... | S_00015.gdm | OV    | 0927       |
| TCGA-25-1313-01A | 12.61316  | 84.866241  | 14.553219 | 0.127686 | 17.153735 | 11.769521 |       | ... | S_00016.gdm | OV    | 1313       |
| TCGA-61-2009-01A | 9.888627  | 200.908539 | 11.994709 | 0.203942 | 8.784982  | 8.981544  |       | ... | S_00017.gdm | OV    | 2009       |
| TCGA-30-1862-01A | 15.701506 | 149.562462 | 16.890046 | 0.242594 | 18.532845 | 14.126315 |       | ... | S_00018.gdm | OV    | 1862       |
| TCGA-23-1029-01B | 6.680773  | 48.931378  | 12.503842 | 0.198906 | 16.608442 | 12.840061 |       | ... | S_00019.gdm | OV    | 1029       |
| ...              | ...       | ...        | ...       | ...      | ...       | ...       | ...   | ... | ...         | ...   | ...        |

## 5.4.2 Breast cancer

Finally we use BRCA tumor for the biological and computational validation of the OV tumor analysis outcomes. As already stated in the previous chapter, only a limited set of samples related to a biomolecularly specific subtype of breast cancer, i.e., the *Basal-like Breast Cancer*, is considered.

The data extraction process uses a GMQL query having a similar structure of the one used for the OV tumor data. Its block diagram is reported in *Figure 5.12*. The only difference is that no *PAM50 Basal* tag or identifier is encoded in TCGA metadata, making impossible to recognize the tumor samples of interest without providing more external information. For this reason, we implement a Python script for converting “Basal” tumor samples in the PAM50 data table (*Table 4.4* in *Chapter 4*), according to the GDM framework, into a GMQL-like dataset to be uploaded to the GMQL system (BRCA\_PAM50\_BASAL dataset). The complete code of this script is reported in *Figure 5.13*.

Using both the TCGA aliquot sample barcode and the ID of the patient identifying PAM50 samples, we can select basal-like breast cancer samples having both methylation and expression data available in TCGA. The GMQL query we used is reported below:

```
# Extract methylation and expression information for BRCA tumor:

# Initial datasets:
# Extract all the samples for tumor BRCA and platform 450k and exclude, if present, either normal or
# metastatic samples, retrieving only primary or recurrent appearance tumor samples.
ALL_BRCA_METHYL = SELECT(manually_curated_cases_disease_type == " Breast Invasive Carcinoma"
    AND manually_curated_platform == " Illumina Human Methylation 450" AND (biospecimen__
    bio_sample_type == "Primary Tumor" OR biospecimen_bio_sample_type == "Recurrent
    Tumor") AND clinical_shared_history_of_neoadjuvant_treatment == "No")
GRCh38_TCGA_methylation;
```

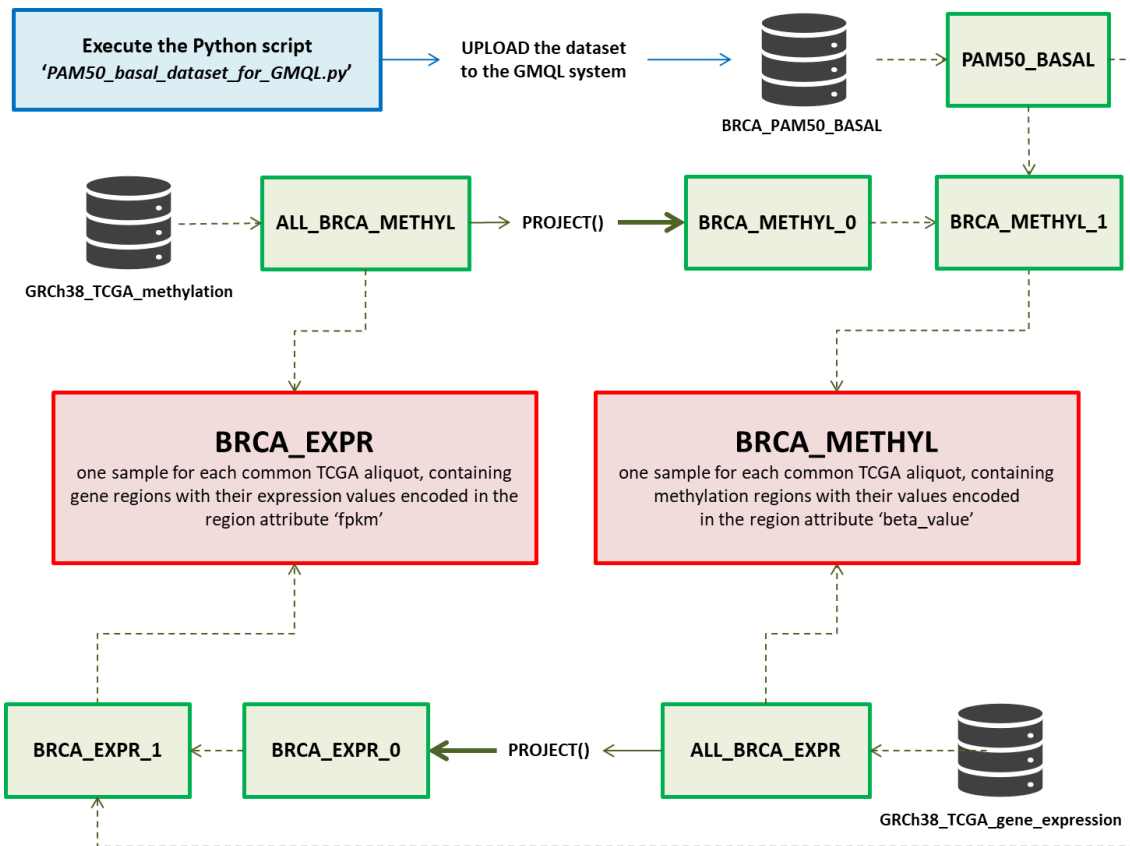


Figure 5.12: Diagram of the GMQL query for the extraction of BRCA tumor data.

```

# Extract all gene expression data for tumor BRCA and exclude, if present, either normal or metastatic
# samples, retrieving only primary or recurrent appearance tumor samples.
ALL_BRCA_EXPR = SELECT(manually_curated_cases_disease_type == "Breast Invasive Carcinoma" AND
    (biospecimen_bio_sample_type == "Primary Tumor" OR biospecimen_bio_sample_type ==
    "Recurrent Tumor") AND clinical_shared_history_of_neoadjuvant_treatment == "No")
    GRCh38_TCGA_gene_expression;

# Extract all the 'basal' PAM50 samples for tumor BRCA.
PAM50_BASAL = SELECT() BRCA_PAM50_BASAL;

# Keep only those region attributes that are useful for the scope of the project, in order to lighten the
# computational time complexity of the query, and select the common samples by extracting only those aliquots
# for which both methylation and expression values are available.
# Gene Expression (BRCA):
BRCA_EXPR_0 = PROJECT(ensembl_gene_id, entrez_gene_id, gene_symbol, fpkm) ALL_BRCA_EXPR;
BRCA_EXPR_1 = SELECT(semijoin: biospecimen_shared_bcr_patient_barcode IN PAM50_BASAL)
    BRCA_EXPR_0;
BRCA_EXPR = SELECT(semijoin: biospecimen_bio_bcr_sample_barcode IN ALL_BRCA_METHYL)
    BRCA_EXPR_1;

# The final datasets are created and saved in the system.
MATERIALIZE BRCA_EXPR INTO BRCA_EXPR;

# Methylation (BRCA):
BRCA_METHYL_0 = PROJECT(beta_value) ALL_BRCA_METHYL;
BRCA_METHYL_1 = SELECT(semijoin: biospecimen_shared_bcr_patient_barcode IN PAM50_BASAL)
    BRCA_METHYL_0;
BRCA_METHYL = SELECT(semijoin: biospecimen_bio_bcr_sample_barcode IN ALL_BRCA_EXPR)
    BRCA_METHYL_1;

# The final dataset is created and saved in the system.
MATERIALIZE BRCA_METHYL INTO BRCA_METHYL;

```

```

''' PAM50_basal_dataset_for_GMQL.py '''

# Generate a GMQL dataset for the PAM50 Basal-like BRCA samples, creating one sample for each one of them.

# Import libraries
import pandas as pd

# Import the data table containing the PAM50 samples (1043)
brca_pam50_df = pd.read_excel('./BRCA_PAM50.xlsx', sheetname='Foglio1', header=0,
                             converters={'ID':str, 'PAM50':str, 'ER':str, 'PR':str, 'HER2':str, 'TRIPLE':str, 'TNBCTYPE':str})

# Extract the 'basal' samples of interest (174)
brca_pam50_basal_df = brca_pam50_df.loc[brca_pam50_df['PAM50'] == 'Basal'].copy()
brca_pam50_basal_df.index = range(len(brca_pam50_basal_df))

# METADATA:
# Create a dataframe for defining the metadata attributes to include in the GMQL samples
Metadata_df = pd.DataFrame(index = ['biospecimen_shared_bcr_patient_barcode', 'pam50', 'er', 'pr', 'her2',
                                   'triple', 'tnbctype'], columns = ['attribute_value'])

# Create one metadata sample (.meta) for each BRCA PAM50 'basal' tumor sample
for index, row in brca_pam50_basal_df.iterrows():
    patient = row['ID']
    pam50_type = row['PAM50']
    er_val = row['ER']
    pr_val = row['PR']
    her2_val = row['HER2']
    triple_neg = row['TRIPLE']
    tnbctype_val = row['TNBCTYPE']

    Metadata_df.set_value('biospecimen_shared_bcr_patient_barcode', 'attribute_value', patient)
    Metadata_df.set_value('pam50', 'attribute_value', pam50_type)
    Metadata_df.set_value('er', 'attribute_value', er_val)
    Metadata_df.set_value('pr', 'attribute_value', pr_val)
    Metadata_df.set_value('her2', 'attribute_value', her2_val)
    Metadata_df.set_value('triple', 'attribute_value', triple_neg)
    Metadata_df.set_value('tnbctype', 'attribute_value', tnbctype_val)

# export the dataframe as a .meta file
# (i.e., a tab-delimited file with an (attribute_name, value) pair in each row of the file)
Metadata_df.to_csv('./files/S_'+str(index+1)+'.gdm.meta', sep='\t', header=False)

# REGIONS:
# Even if only information encoded in metadata samples is needed, the GDM framework requires that each
# metadata sample has an associated region data sample, in order to actually constitute a GMQL dataset.
# for this reason, a dummy region data sample is created, for each of the metadata samples defined,
# in order to correctly import the full dataset in the GMQL system.
Region_df = pd.DataFrame(index = [0], columns = ['chrom', 'left', 'right', 'strand', 'tumor'])

# Create one region data sample (.gdm) for each BRCA PAM50 'basal' sample,
# setting up the four mandatory region attributes (chromosome, genomic coordinates and strand)
for index, row in brca_pam50_basal_df.iterrows():
    Region_df.set_value(0, 'chrom', 'chr0')
    Region_df.set_value(0, 'left', 0)
    Region_df.set_value(0, 'right', 0)
    Region_df.set_value(0, 'strand', '')
    Region_df.set_value(0, 'tumor', 'BRCA')

# export the dataframe as a .gdm file
# (i.e., a tab-delimited file where each row identifies a region and its attributes are tab-separated)
Region_df.to_csv('./files/S_'+str(index+1)+'.gdm', sep='\t', header=False, index=False)

```

Figure 5.13: Python script for converting BRCA PAM50 Basal set of samples into a GMQL dataset.

## 5.5 Data matrix construction

We have to organize the selected data to let them be used as input for the data analysis, such that we can analyze the single regulation system of each gene of interest, by identifying the role of its possible regulatory features quantified as for their impact on the target gene expression, whatever the feature is, i.e., the gene methylation, the expression of a gene in the same pathway or in another pathway and the expression of genes encoding for transcription factors.

Therefore, we adopt an additive approach to build five different data matrices for each gene of interest (that we call M1, M2, M3, M4 and M5), including TCGA data extracted in the previous phases. We can keep track of each gene regulation system step by step, according to the different types of biological features. This means that, starting from the gene expression and methylation value of the model gene, the columns of each gene data matrix are gradually incremented, according to specific pre-defined rules. *Table 5.4* shows the structure of the five set of features used during the data analysis.

By gradually adding new features, i.e., genes that may positively or negatively influence the regulation of the model gene expression, we can broaden the set of regulation hypotheses, until we reach an accurate prediction of the expression of the model gene.

Gene 2932 [GSK3B] belongs to both STEM\_CELLS and GLUCOSE\_METABOLISM pathways: since its regulation can be different depending on the pathway it participates in, and thus to its biological functions, it must be considered twice in the analysis, with two data matrices of each type. In total, 885 data matrices (177 genes \* 5 set of features) with 372 rows (i.e., OV patients samples) and up to 419 columns (i.e., features, potential candidate regulators) are built. Their content is detailed in the following paragraphs.

### 5.5.1 M1: genes belonging to the same pathway of the model gene

Matrix M1 contains the expression of the model gene (measured in *fpkm*), its methylation (expressed as the mean of its *beta\_values*) and the expression of the genes belonging to the model gene pathway.

Therefore, the number of possible features of this matrix (its columns, excluding the target) is different, depending on the pathway the model gene belongs to: 20 features for DNA\_REPAIR genes, 73 features for STEM\_CELLS genes and 84 for GLUCOSE\_METABOLISM genes.

### 5.5.2 M2: M1 + candidate regulatory genes of the model gene

Matrix M2 adds the expression of all the candidate regulatory genes of the model gene to matrix M1.

The number of columns of this matrix clearly depends on the number of candidate regulatory genes which each gene of interest is associated to: the number of features is approximately in the range of 67-194 for DNA\_REPAIR genes, 73-208 for STEM\_CELLS genes and 84-258 for GLUCOSE\_METABOLISM genes.

Each new feature is added avoiding repetition with respect to M1: this means that if a candidate regulator of the model gene is already present in M1 as a gene of the model gene pathway, it is discarded.

Table 5.4: Structure of the data matrixes for the data analysis process.

| M5                                     |  |                          |  |  |                            |  |  |
|--|--|--------------------------|--|--|----------------------------|--|--|
| M4                                     |  |                          |  |  |                            |  |  |
| M3                                     |  |                          |  |  |                            |  |  |
| M2                                     |  |                          |  |  |                            |  |  |
| M1                                     |  |                          |  |  |                            |  |  |
| TARGET<br>(regression output variable) | FEATURES<br>(regression input variables) |                          |  |  |                            |  |  |
| Model Gene<br>EXPRESSION               | Model Gene<br>METHYLATION                | SAME<br>PATHWAY<br>GENES | Model gene<br>CANDIDATE<br>REGULATORY<br>GENES | Same pathway<br>CANDIDATE<br>REGULATORY<br>GENES | OTHER<br>PATHWAYS<br>GENES | Other Pathways<br>CANDIDATE<br>REGULATORY<br>GENES |  |
| TCGA_Aliquot_1                         |  |                          |  |  |                            |  |  |
| TCGA_Aliquot_2                         |  |                          |  |  |                            |  |  |
| TCGA_Aliquot_3                         |  |                          |  |  |                            |  |  |
| ...                                    |  |                          |  |  |                            |  |  |
| TCGA_Aliquot_N                         |  |                          |  |  |                            |  |  |

### 5.5.3 M3: M2 + candidate regulatory genes of the model gene pathway genes

Matrix M3 adds the expression of the candidate regulatory genes of all the genes in the model gene pathway to matrix M2. The number of columns of this matrix also depends on the number of these new candidate regulatory genes: 246 for DNA\_REPAIR genes, 310 for STEM\_CELLS genes and 324 for GLUCOSE\_METABOLISM genes.

This number is intuitively the same for all genes belonging to the same pathway, because all the genes of the pathway and all their candidate regulatory genes are examined and added at least in one of the three iterations. Furthermore, the amount of new features to be examined from this step on is exactly the same for all the genes in each pathway, so the number of columns of the next two matrixes remains identical for all the genes belonging to the same pathway.

Also in this case, new features are added avoiding repetition: if a candidate regulatory gene of some gene in the pathway is already present in M2 (i.e. it is either a gene of the model gene pathway or a candidate regulatory gene of the model gene), it is discarded.

### 5.5.4 M4: M3 + genes belonging to other pathways

Matrix M4 adds the expression of the genes belonging to the other pathways with respect to the model gene considered to matrix M3, avoiding repetitions.

However, biologists say that predicting STEM\_CELLS and DNA\_REPAIR genes expression on the basis of genes belonging to GLUCOSE\_METABOLISM pathway is confusing, because the functions they perform are quite different and poorly interrelated; including glucose metabolism involved genes within the regulation systems of either stem cells or DNA-repair genes would be misleading for the analysis. For this reason, we build M4 matrixes of genes belonging to DNA\_REPAIR and STEM\_CELLS pathways by adding only features from STEM\_CELLS and DNA\_REPAIR, respectively, while M4

matrixes of genes belonging to GLUCOSE\_METABOLISM pathway contain both DNA\_REPAIR and STEM\_CELLS potential regulators.

As already stated before, the number of columns of matrix M4 is identical for all the genes belonging to the same pathway: in particular, they are in total 314 for DNA\_REPAIR genes, 329 for STEM\_CELLS genes and 410 for GLUCOSE\_METABOLISM genes.

### **5.5.5 M5: M4 + candidate regulatory genes of other pathway genes**

Finally, matrix M5 adds the expression of the candidate regulatory genes of all the genes belonging to the other pathways to matrix M4, avoiding repetitions.

The number of columns of this matrix is intuitively the same for genes belonging either to the DNA\_REPAIR or to the STEM\_CELLS pathway (332), while it is higher (419) for GLUCOSE\_METABOLISM genes, since both other pathways are evaluated for identifying gene expression regulators.



## 6. Data Analysis

*<< We've all heard the buzz words: streamline, optimize, integrate, adapt.*

*Everyday someone comes up with a new strategy or tool or technology to increase our efficiency. [...]*

*To really be efficient, you have to eliminate what doesn't work. You have to figure out what is important and hold on tight to the things that matter the most.>>*

Meredith

This is the most important phase of the project: processing all the data arranged in the previous phases through a suitable machine learning algorithm [58, 59], in order to investigate the behavior of the target genes to meet our goals.

According to our regulation hypotheses, we build a linear regression model for each target gene and for each set of features (i.e., data matrices from M1 to M5).

The five matrices are constructed according to pre-defined rules we have established at the beginning of the project. Progressively broadening the set of potential regulatory features associated with each target gene in a well-defined order is just matter of convenience, first to keep track of the regulatory influence of each different type of factors on the target gene expression and secondly to design a computationally sustainable data analysis procedure. However, there are no specific biological rules stating that, for each target gene, genes in the same pathway have to be analyzed before its candidate regulatory genes. Therefore, these two set of features are given as inputs to our regression model all at once, by considering the whole M2 (that includes M1) as the starting point of the data analysis. The same applies for M5 (that includes M4), following the analysis on M3.

In a word, we perform only three regression processes for each gene of interest, using data in matrices M2 (genes in the same pathway and model gene candidate regulators), M3 (including also candidate regulatory genes of the model gene pathway) and M5 (including also genes in the other pathways and their candidate regulators). So, we overcome potential issues related to the order in which features are added to the models: in fact, adding to the regression model genes in the same pathway first and then candidate regulators of the model gene produces different results than considering the features in the opposite order (i.e., candidate regulators first, followed by genes in the same pathway).

The full data analysis is divided into two main processes, executed per pathway (DNA\_REPAIR, STEM\_CELLS and GLUCOSE\_METABOLISM) and per set of features (M2, M3 and M5): for each target gene in the pathway, we first perform a feature selection on the whole set of its possible regulatory features, and then we execute the linear regression algorithm.

Feature selection identifies a subset of the original predictors (i.e., the set of potential regulatory features) believed to be related to the response (i.e., the expression of the target gene) and allows to fit the regression models by reducing their sets of inputs and, as a consequence, to apply least squares on the reduced set of variables.

```

# Set parameters representing the current pathway and data matrix to process
pathway = 'DNA_REPAIR' # possible values: {DNA_REPAIR, STEM_CELLS, GLUCOSE_METABOLISM}
model = '5' # possible values: {2, 3, 5}

# Extract from the set of 177 target genes the Gene Symbols of all the genes belonging to the current pathway
SYMs_current_pathway = []
for index, row in TargetGenes_df.iterrows():
    sym = row['GENE_SYMBOL'] # get the name of the gene
    path = row['PATHWAY'] # get the pathway
    if path == pathway:
        SYMs_current_pathway.append(sym)

# FEATURE SELECTION
# for each gene of interest in the current pathway, according to data stored in the matrix considered...
for current_gene in SYMs_current_pathway:

    # Import the data matrix corresponding to the current gene
    gene_ID = TargetGenes_df.loc[TargetGenes_df['GENE_SYMBOL'] == current_gene, 'ENTREZ_GENE_ID'].iloc[0]
    data_matrix_df = pd.read_excel('./Gene '+gene_ID+' ['+current_gene+']+' ('+pathway+') - Model v'+model+'.xlsx',
                                  sheetname='Sheet1',header=0)

    # Execute the Forward FEATURE SELECTION procedure
    .....

# LINEAR REGRESSION
# for each gene of interest in the current pathway, according to data stored in the matrix considered...
for current_gene in SYMs_current_pathway:

    # Import the subset of features selected
    # [...]

    # DATA STANDARDIZATION: normalize data values according to the standard Z-score normalization
    # [...]

    # Execute the Ordinary Least Squares LINEAR REGRESSION

```

Figure 6.1: Initialization step in the Python scripts for performing feature selection and linear regression.

We adopt this approach to overcome computational issues deriving from handling more than 500 models, from a set of 372 observations (i.e., the patient-specific ovarian tumor samples) and from up to 419 potential features. Besides reducing computational complexity, we also get a better statistical accuracy: otherwise, too many predictors would lead to an overfitting of the model, causing no longer unique least squares coefficient estimate which would drastically increase the variance, making the linear method inaccurate.

Therefore, as matter of computational efficiency, data analysis on each gene is split in three steps, one for each set of features, and each step comprises two main processes: a preliminary feature selection for removing non-significant features and for reducing the set of inputs of the regression model, and the linear model fitting, which keeps only relevant features selected in the previous step for each next step. Considering the whole set of possible features, in fact, would be completely unreasonable from a computational standpoint, uselessly causing an execution taking days to complete.

Both feature selection and the subsequent linear regression are executed on each target gene and for each of the three sets of features, using two parametric Python scripts executed per pathway: at the beginning of each script, it is enough to initialize two specific parameters defining the pathway and the number of the model to be processed. *Figure 6.1* shows the Python code used for initializing these parameters.

As shown in *Figure 6.2*, so far we have described the set of operations for extracting and handling data for each gene; the following paragraphs detail the feature selection and the linear regression procedure used for the data analysis.

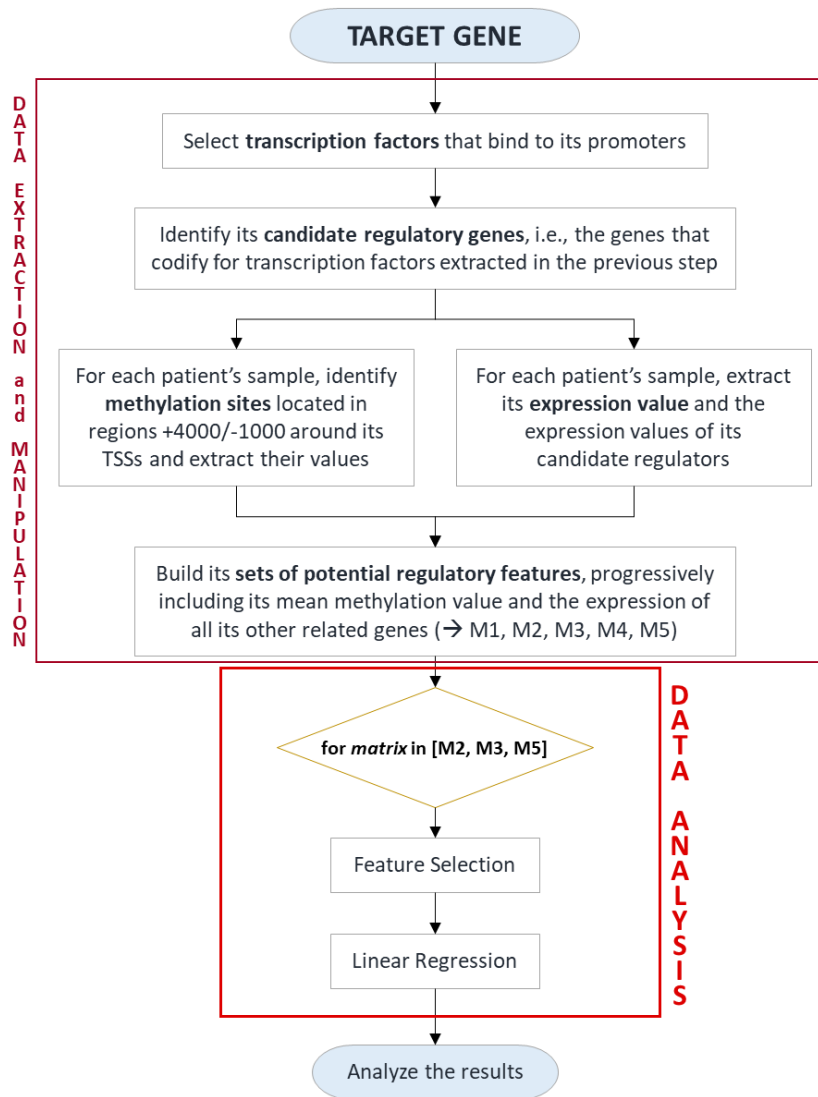


Figure 6.2: Full set of operations performed on each gene of interest during the course of the project.

## 6.1 Feature / gene selection

The first step for each target gene consists in selecting the best subset of features for that gene, according to a specific metric. In fact, it would be useless to keep inputs that are known to be non-significant for the regulation of the model gene expression beforehand. In addition, this feature selection process allows to perform a more accurate analysis for each gene in a reasonable time, reducing the dimensions of the set of features with respect to the set of observations.

The Python script implementing the feature selection process iterates along the list of target genes belonging to the pathway selected for the analysis and for each of such genes executes the following operations:

- a. the current gene data matrix is imported (M2, M3 or M5) and all its empty columns are removed, since they surely do not contribute to the regression model (e.g., columns with *NaN* methylation

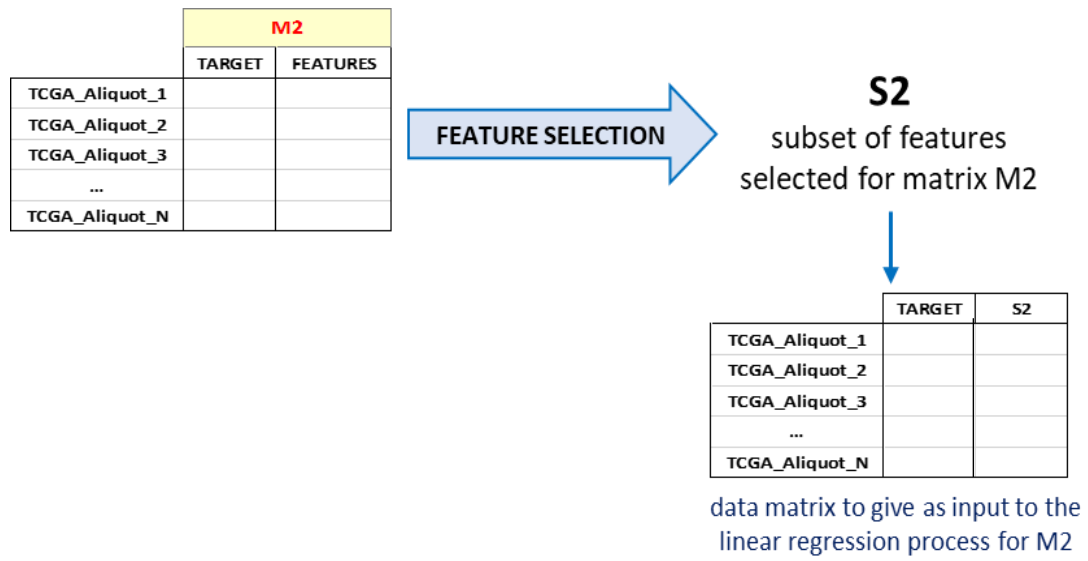
values, corresponding to genes for which no probes are measured in TCGA, or genes with missing expression values in TCGA);

- b. if the matrix selected is either M3 or M5, all those features belonging to the previous set of features of the same gene (M2 for M3 and M3 for M5) which are not selected by the feature selection process before, are removed. We assume those same features cannot be selected in the current feature selection process, since they have already been discarded once. This implementation also contributes to design a more scalable process, always keeping a limited number of considered features at each step of the data analysis. The features retrieved from the previous model, along with the new features in the current matrix, are involved in the new feature selection process; this means that one or more features selected in the previous model may potentially be discarded as a result of the current feature selection, because some of the new added features better explain the output variable;
- c. at this point, for each input matrix, the feature selection is performed five times, as follows: in order to reduce the bias, we randomly split the set of TCGA aliquots into five, possibly equal, groups of samples which are used to create five different testing sets. This partition is indeed performed only once at the beginning of the data analysis, and then the same five subsets of aliquots are used for processing all the genes. Therefore, we execute the feature selection five times, once for each generated testing set (using the remaining aliquots as training set), according to a *k-fold cross-validation* process, setting  $k=5$ . Finally, the intersection of the five sets of extracted features is computed to obtain the final array of selected features for the current gene in the current matrix. *Figure 6.3* summarizes the complete feature selection procedure.

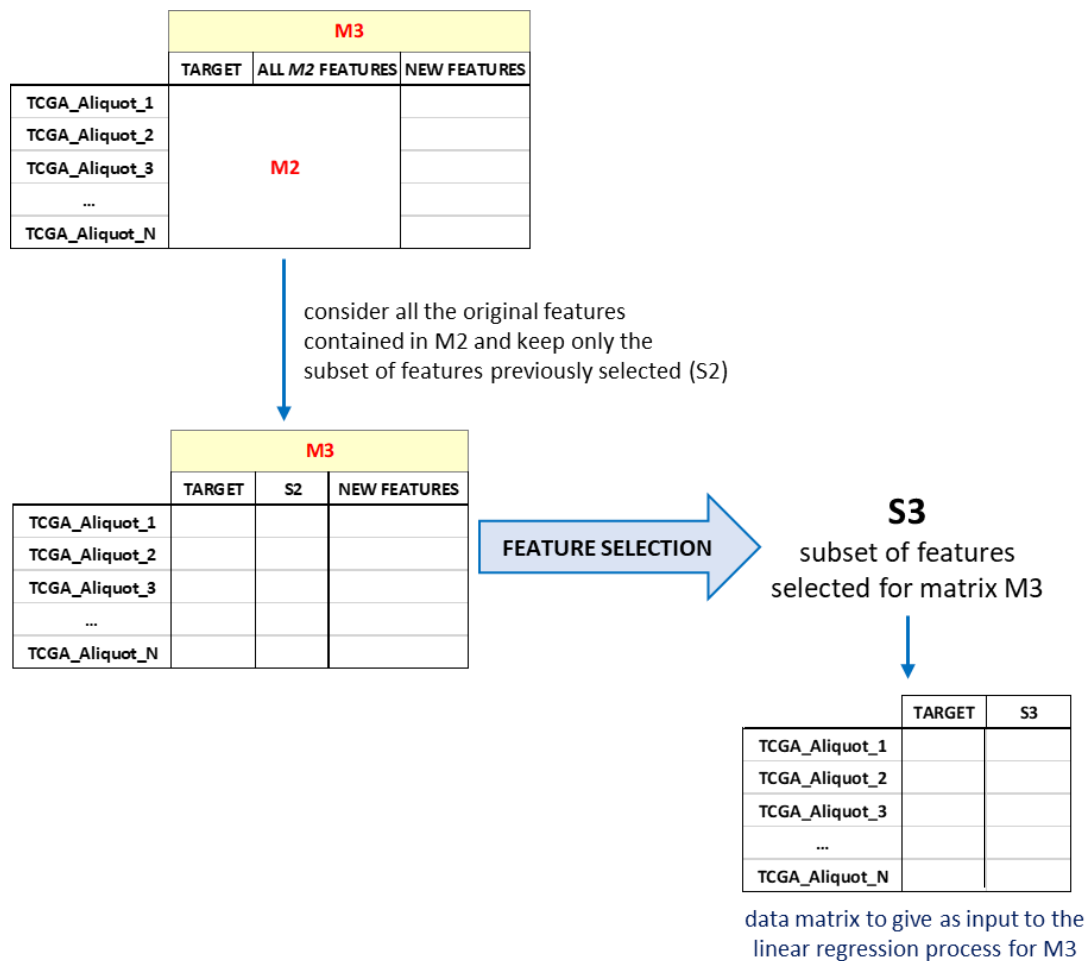
Here, feature selection follows the *forward feature selection* paradigm that analyzes the performance of all the possible sets of features, starting from a single feature and progressively adding the others one at a time, finally returning the subset with the best cross-validation performance. This means that if  $n$  is the number of considered features,  $n$  is also the number of the analyzed sets of features, with a dimension going from 1 (the first subset) to  $n$  (the last subset).

Different criteria can be used to decide which subset of features and how many features are best to be extracted: in this case, the `k_features` parameter of the selector is set to 'best', which means that the returned subset of the initial features is the one that minimizes the mean of cross-validation scores.

a) Selection of features in M2, containing the methylation of the model gene, the expression of the genes in the currently analyzed pathway and the expression of candidate regulatory genes of the model gene.



b) Selection of features in M3, adding to the previous matrix the expression of the candidate regulatory genes of all the other genes in the currently analyzed pathway.



c) Selection of features in M5, adding to the previous matrix the expression of the genes in the other pathways and of their candidate regulatory genes.

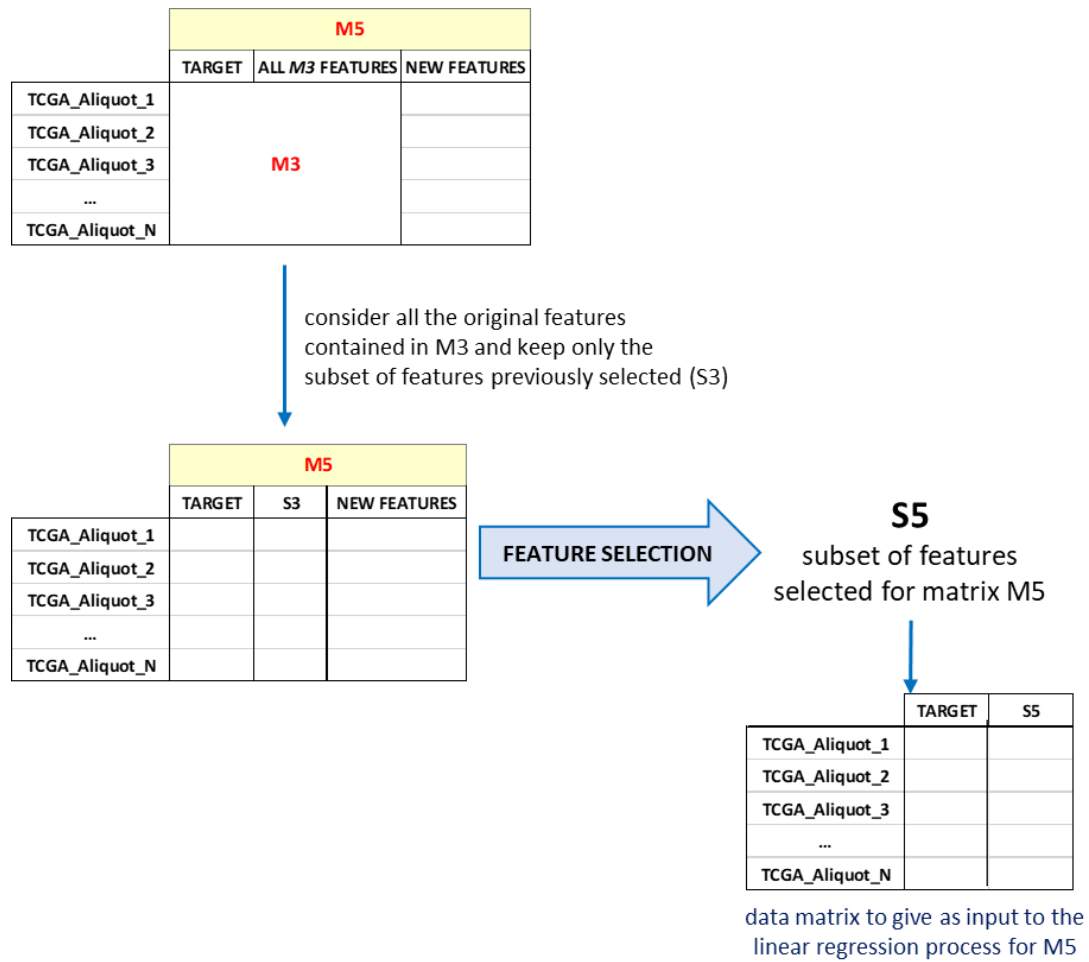


Figure 6.3: Feature selection for matrix M2 (a), M3 (b) and M5 (c).

Figure 6.4 shows the piece of code in the feature selection Python script where the feature selector is used, while in Figure 6.5 we report the complete flowchart displaying the detailed operations of the “Feature\_selection.py” Python script.

```

# APPLY FEATURE SELECTION

# Sequential Feature Selector performs forward feature selection.
# The function used is the following:
# SFS(estimator, k_features, forward, floating, scoring, cv, n_jobs, ...)
# where estimator = scikit-learn classifier or regressor
# k_features = number of features to select (int or tuple with a min and max value).
# SFS will consider return any feature combination between min and max
# that scored highest in cross-validation.
# If 'best' is provided, the feature selector will return the feature subset with the best
# cross-validation performance. If 'parsimonious' is provided as an argument,
# the smallest feature subset that is within one standard error of the cross-validation
# performance will be selected.
# forward = forward selection if true, backward selection otherwise
# floating = allows to implement SFFS or SBFS
# scoring = scoring metric
# (accuracy, f1, precision, recall, roc_auc) for classifiers
# ('mean_absolute_error', 'neg_mean_squared_error', 'median_absolute_error', 'r2') for regressors
# cv = cross-validation generator (default: 5)
# n_jobs = number of CPUs to use for evaluating different feature subsets in parallel ('-1' means 'all CPUs')

# Define the linear regression object
lr = LinearRegression()

# Count the total number of features
tot_N_features = int(X_train.shape[1])

# Perform feature selection
sfs = SFS(lr, k_features='best', forward=True, floating=False, scoring='neg_mean_squared_error', cv=5)

# Learn model from training data
sfs = sfs.fit(X_train, y_train)

# Get all the details of the forward fits:
# 'get_metric_dict(confidence_interval=0.95)' returns a dictionary, where dictionary keys are the number
# of iterations (number of feature subsets) and where the value for each key is a second dictionary.
# The keys of this second dictionary are:
# 'feature_idx': tuple of the indices of the feature subset
# 'cv_scores': list with individual CV scores
# 'avg_score': average of CV scores
# 'std_dev': standard deviation of the CV score average
# 'std_err': standard error of the CV score average
# 'ci_bound': confidence interval bound of the CV score average (around the computed cross-validation scores)
# and they each have a different value in each iteration.
# So, the general structure of this dictionary is the following:
# {Iteration_1 : {feature_idx: tuple_of_values, cv_scores: list_of_values, avg_score: value,...},
# Iteration_2 : {feature_idx: tuple_of_values, cv_scores: list_of_values, avg_score: value,...},
# Iteration_3 : {feature_idx: tuple_of_values, cv_scores: list_of_values, avg_score: value,...}, ...}
result_dict = sfs.get_metric_dict()

# Compute the mean of cross-validation scores
mean_cv_scores = []
for i in np.arange(1, tot_N_features+1): # values are generated within the interval [start, stop),
# including start but excluding stop
# since cv_scores are negative numbers in the previous dictionary, I have to add a '-' to compute the mean
mean_cv_scores.append(-np.mean(result_dict[i]['cv_scores']))

# Get the number of features selected, which corresponds to the number of features selected
# in correspondence of the minimum of the mean cross-validation scores
idx = np.argmin(mean_cv_scores)+1

# Get the features indexes for the best forward fit and convert them to list
feature_idx = result_dict[idx]['feature_idx']
selected_features_indexes = list(feature_idx)

# Extract the names of these features (columns of the model gene matrix, except for the target)
X_df = model_gene_df.drop(['EXPRESSION ('+current_gene+')'],1)
X_df_columns = list(X_df.columns)
columns_selected = []
for index in selected_features_indexes:
columns_selected.append(X_df_columns[index])

```

Figure 6.4: Feature selection process, repeated five times for each target gene and for each selected data matrix.

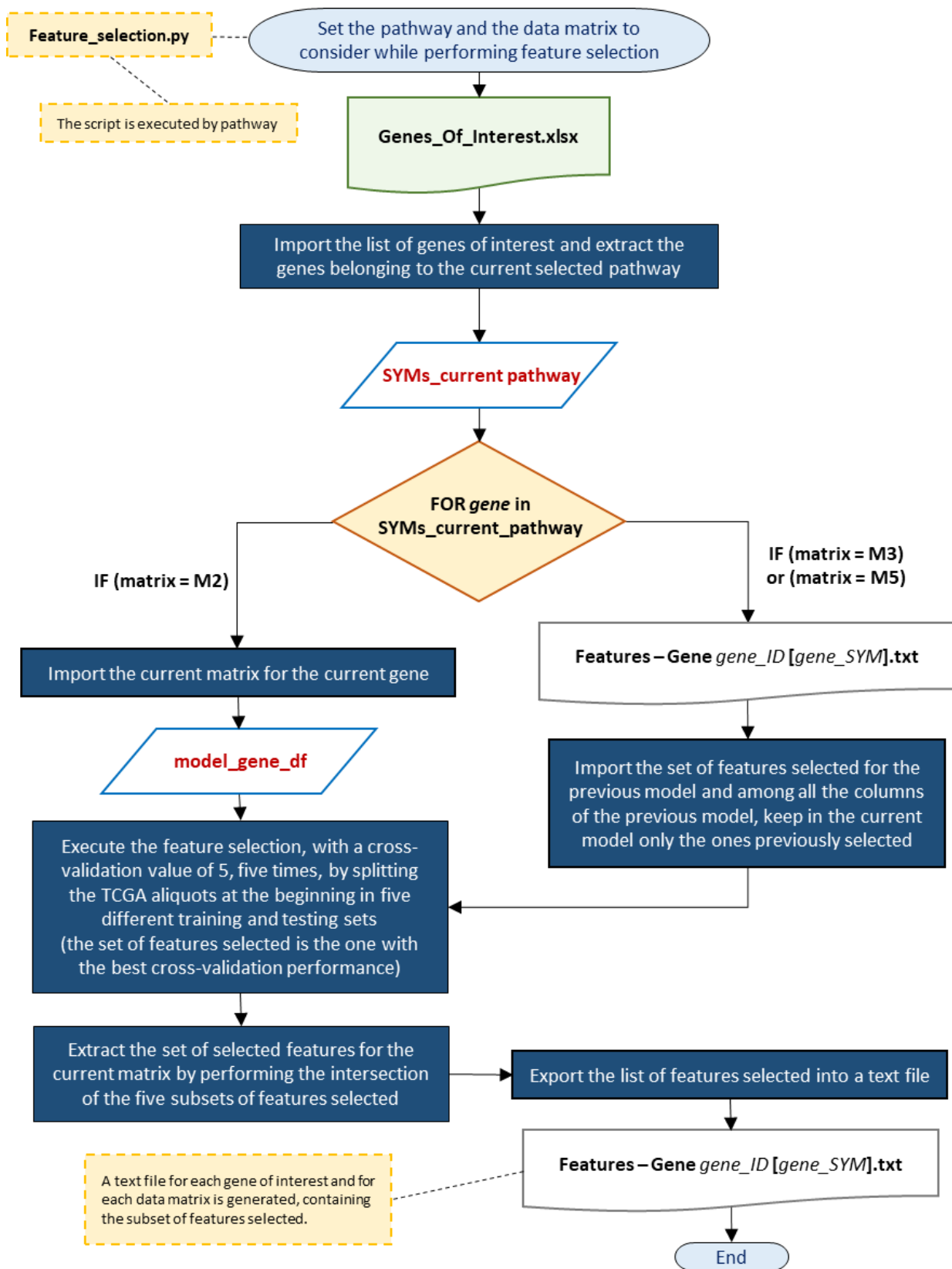


Figure 6.5: Flowchart of the feature selection script.



## 6.2 Linear regression of individual genes on ovarian tumor samples

The second step of our data analysis consists in fitting a linear model on each individual target gene and its data matrices, starting from the set of features selected in the previous step.

Just as what done for the feature selection, we implement a Python script that iterates along the list of target genes belonging to the pathway under analysis, fitting a linear model as specified in the data matrix parameter at the beginning of the script. Here are the details:

- a. data are normalized to allow results comparisons; assessing the regulation of the model gene expression starts from heterogenous data, hence normalization is needed for consistently comparing results both in individual models and across models.

The *Z-score normalization* is applied to convert each variable into a variable with a “standard” distribution, with  $\text{mean\_value}=0$  and  $\text{variance}=1$ , using the following formula:

$$\text{Normalized}(v) = \frac{v - \text{mean\_value}}{\text{std\_dev}}$$

This is achieved by using the Standard Scaler from the `sklearn.preprocessing` model (*Figure 6.6* shows how this function works). This normalization process allows comparing the regression coefficients assigned to the different features within the same model and establishing which one has the highest impact on the output variable;

- b. we build the linear model using function `OLS` from `statsmodels.regression.linear_model` module to fit a simple ordinary least squares model. *Figure 6.7* shows the script implemented for the ordinary least squares regression.

This Python library conveniently presents the results of the regression procedure to the user: besides computing the regression coefficients for each input feature, it also automatically calculates  $R^2$ , Adjusted  $R^2$ , confidence intervals and other quality-related parameters, to let the user immediately assessing the quality of the fit and the accuracy of the model. An example of this summary statistics is reported in *Figure 6.8*.

```
# DATA STANDARDIZATION
# Define the scaler:
# MinMaxScaler() normalizes values between 0 and 1
# StandardScaler() performs Z-score normalization
from sklearn import preprocessing
scaler = preprocessing.StandardScaler()

# If the data are encoded in a pandas dataframe, convert it into a numpy array
matrix_to_normalize = data_matrix_df.values

# Normalize and, eventually, convert back to pandas dataframe
matrix_scaled = scaler.fit_transform(matrix_to_normalize)
data_matrix_df_std = pd.DataFrame(matrix_scaled, index=data_matrix_df.index, columns=data_matrix_df.columns)
```

*Figure 6.6:* Z-score normalization (or standard normalization) in *scikit-learn*.

```

# Filter the initial matrix of data extracting only the columns selected by feature selection
cols_to_extract = ['EXPRESSION ('+current_gene+')']+features_selected
model_gene_df_filtered = model_gene_df[cols_to_extract].copy()

# PERFORM LINEAR REGRESSION
import statsmodels.api as sm

# Define the features (predictors X) and the target (label y)
X = model_gene_df_filtered.drop(['EXPRESSION ('+current_gene+')'],1) # all the columns, except for the target
y = model_gene_df_filtered['EXPRESSION ('+current_gene+')'] # the expression of the current gene

# Add an intercept to the model
X = sm.add_constant(X)

# Define the linear regression object and fit the model
lr_model = sm.OLS(y, X).fit()

# Make predictions
y_predicted = lr_model.predict(X)

# Extract the summary statistics
lr_summary = lr_model.summary()

# Get the Adjusted R-squared
r2 = lr_model.rsquared_adj

# Get the coefficients of the linear regression and the intercept of the model
coeff = lr_model.params

# Get the standard errors
std_err = lr_model.bse

# Compute and get the confidence intervals for the model coefficients (default: 95%)
# and compare them to the ones automatically computed by OLS, in order to double-check
# the correctness of the procedure
CI_df = lr_model.conf_int()

```

Figure 6.7: Ordinary least squares regression in Statsmodel.

**EXPRESSION of the TARGET GENE**

**QUALITY of the FIT**

OLS Regression Results

|                   |                    |                     |           |
|-------------------|--------------------|---------------------|-----------|
| Dep. Variable:    | EXPRESSION (PTPRC) | R-squared:          | 0.826     |
| Model:            | OLS                | Adj. R-squared:     | 0.823     |
| Method:           | Least Squares      | F-statistic:        | 247.4     |
| Date:             | Fri, 08 Jun 2018   | Prob (F-statistic): | 3.79e-134 |
| Time:             | 09:17:16           | Log-Likelihood:     | -202.27   |
| No. Observations: | 372                | AIC:                | 420.5     |
| Df Residuals:     | 364                | BIC:                | 451.9     |
| Df Model:         | 7                  |                     |           |
| Covariance Type:  | nonrobust          |                     |           |

|       | coef      | std err | t       | P> t  | [0.025 | 0.975] |
|-------|-----------|---------|---------|-------|--------|--------|
| const | 6.765e-17 | 0.022   | 3.1e-15 | 1.000 | -0.043 | 0.043  |
| AXL   | 0.0729    | 0.029   | 2.513   | 0.012 | 0.016  | 0.130  |
| CD38  | 0.1394    | 0.026   | 5.330   | 0.000 | 0.088  | 0.191  |
| CD44  | 0.1217    | 0.024   | 5.046   | 0.000 | 0.074  | 0.169  |
| ELF4  | -0.1252   | 0.025   | -5.033  | 0.000 | -0.174 | -0.076 |
| ERCC2 | -0.1169   | 0.023   | -4.989  | 0.000 | -0.163 | -0.071 |
| IKZF1 | 0.5857    | 0.037   | 15.671  | 0.000 | 0.512  | 0.659  |
| ITGA4 | 0.2392    | 0.037   | 6.530   | 0.000 | 0.167  | 0.311  |

**FEATURES**                      **REGRESSION COEFFICIENTS**                      **CONFIDENCE INTERVALS**

Figure 6.8: Linear regression summary statistics in Statsmodel.

For each gene and in each model, we extract and evaluate only the relevant features for the regulation of its expression, according to the values of the confidence intervals: relevant features are the ones whose regression coefficients have a confidence interval not containing 0 (“zero”).

A confidence interval represents an interval estimate which is supposed to contain the true value of an unknown parameter. It has an associated confidence level (usually 95%) quantifying the probability that the parameter (i.e., the estimated regression coefficient) lies in the interval: so, if a feature is very unlikely to be zero (i.e., the confidence interval does not contain 0), then it is relevant for the model. This means that the actual set of relevant features involved in the regulation of the gene expression is either smaller or equal to the set of features selected by the initial feature selection process and used as input for the regression model (*Figure 6.9* shows how this identification of relevant features works, by considering gene TKT as an example).

Therefore, it may happen that a feature  $f$  selected in M3 is not relevant in the results of the corresponding model, while it becomes significant for M5. This is possible because  $f$  is selected in M3 and so it also participates in the feature selection process of M5: the relevance of a selected feature is evaluated as a result of the linear model fitting.

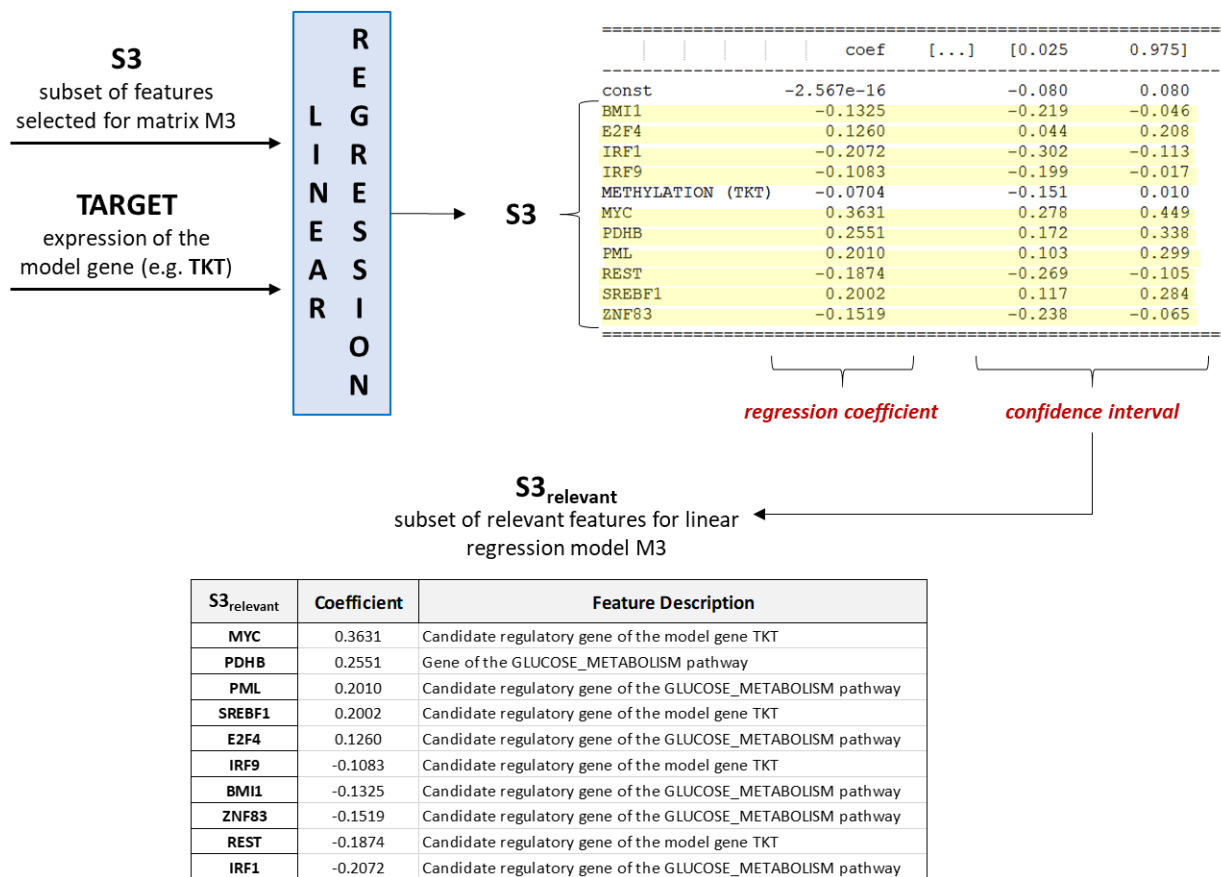


Figure 6.9: Example of linear regression for matrix M3 of gene TKT and identification of relevant features.

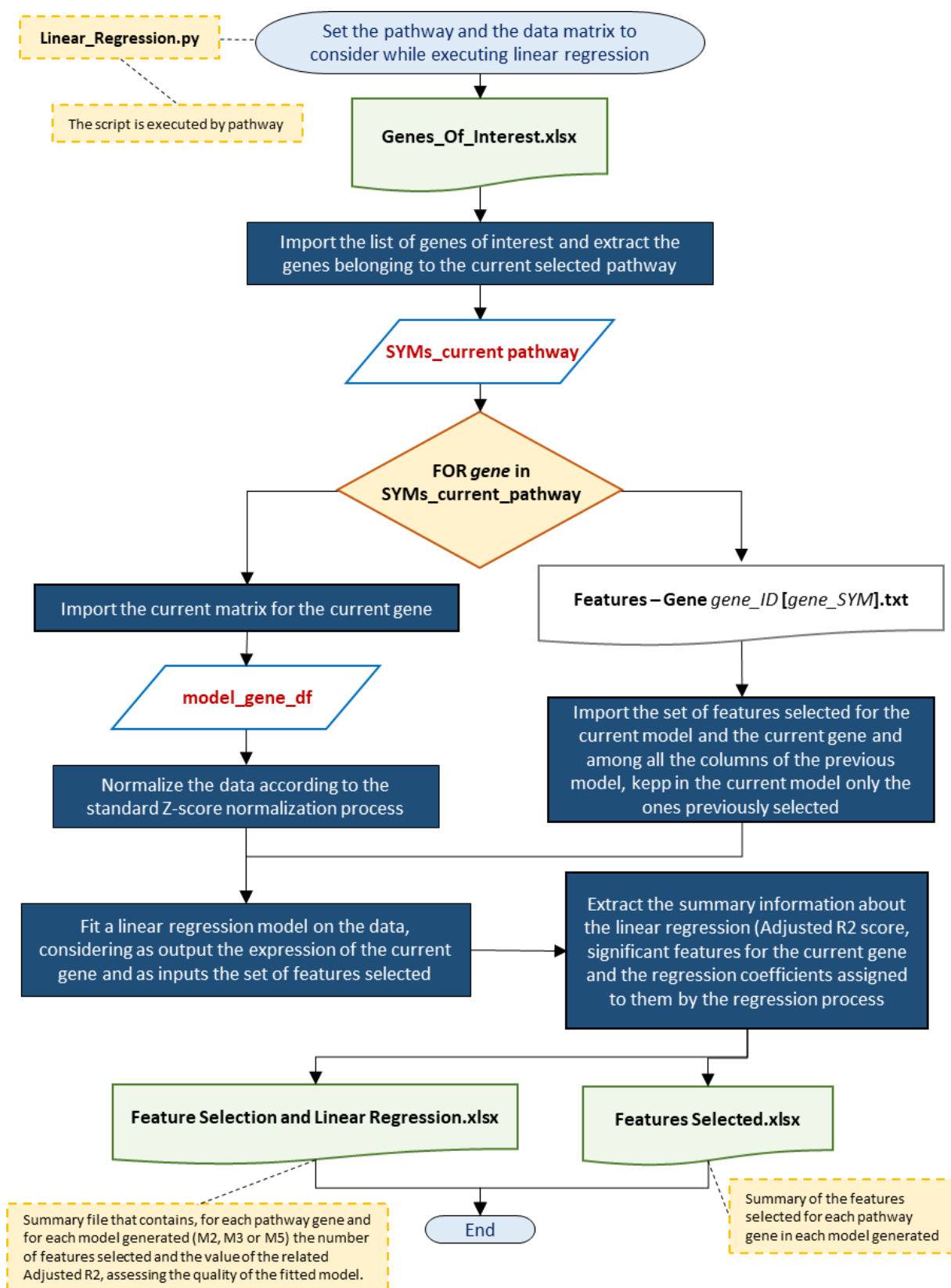


Figure 6.10: Flowchart of the linear regression script.

The flowchart in Figure 6.10 shows the detailed operations of the “Linear\_regression.py” Python script.

Table 6.1: Data analysis: execution times.

| <b>OV Tumor<br/>Data Analysis</b> | <b>M2</b>               | <b>M3</b>               | <b>M5</b>              |
|-----------------------------------|-------------------------|-------------------------|------------------------|
| <b>DNA_REPAIR</b>                 | 2 hours and 30 minutes  | 2 hours                 | 50 minutes             |
| <b>STEM_CELLS</b>                 | 12 hours and 30 minutes | 17 hours and 30 minutes | 20 minutes             |
| <b>GLUCOSE_METABOLISM</b>         | 21 hours and 40 minutes | 17 hours                | 6 hours and 30 minutes |

Table 6.1 shows the details about the execution times required for the data analysis, executed on a 4 cores CPU machine, 2.80 GHz, 16GB RAM.



## 7. Results and Discussion

*<< Knowing is better than wondering. Waking is better than sleeping.  
It can change your perspective, color your thinking. >>*

Meredith & Cristina

In this chapter the main results of the data analysis are presented and discussed, along with their interpretation and visualization as network graphs, and their validation on breast cancer data.

### 7.1 Analysis of regression results and networks generation

As specified in the previous chapter, we adopt a “computationally feasible” approach: we are not interested in all possible relationships between the target genes and their features, but we care about the associations that most efficiently contribute to the prediction; however, some existing specific aspect may be discarded during the analysis, because less relevant than other features. Our results are limited by design to the best-predicting sets of features, leaving out potential regulators with important biological functions, but with a lower predicting power.

Among the three models built for each gene of interest, we specifically focus on results of model M5, because this is the most comprehensive model, considering all the possible features that may participate in the regulation of the target gene expression.

However, some genes may have a more accurate model for M3 rather than for M5 (i.e.,  $R2_{M3} > R2_{M5}$ ): this is due to the fact that its behavior is better explained by genes within their own pathway and genes encoding for transcription factors binding to them, rather than genes involved in the other pathways. Therefore, the evaluation and the interpretation of the final results are focused on both M3 and M5 models, in order to identify the set of features mainly correlated to each target gene.

The effect of each feature on the target gene expression is quantified by its regression coefficient: a positive regression coefficient means that the regulatory element up-regulates the target gene (i.e., it contributes positively to the regulation of its activity, increasing its expression), while a negative regression coefficient means that the regulatory element down-regulates the target gene (i.e., it contributes negatively to the regulation of its activity, suppressing its expression).

The score assigned to each model is of great importance, too: we use the *Adjusted R<sup>2</sup>* for assessing the quality of the fit, because it is an unbiased estimator that takes the number of features used in the model fitting into consideration, on the basis of the sample size and the number of the estimated coefficients. Adjusted R<sup>2</sup> is always smaller than the R<sup>2</sup>, but this difference is usually very small, unless we try to estimate too many coefficients from too small a sample in presence of too much noise.

In general, upon analyzing the score of a regression model we can assess if the model actually meets the initial objectives and it is adequate to the requirements. As available data are highly heterogeneous,

we cannot expect that the target genes have values close to 1. Moreover, several other mechanisms of gene expression regulation are possible, that we did not address in this project, such as miRNAs, lncRNAs, other putative TFs not found in the ChIP-seq experiments from ENCODE, transcription factors subcellular localization and post-translational modifications: the fact that our  $R^2$  scores are usually lower than 1 reflects the lack in our models of these other regulatory elements, since this project only focuses on a specific subset of factors regulating the expression of target genes.

Assessing how good or bad the score is for a regression model is a difficult matter, because it is strictly related to the objectives of the analysis and how the dependent variable is defined. In particular, this score highly depends on whether the main objective for the linear regression is predicting the response variable or it is describing the relationship between the predictors and the target. The main objective of this project is clearly the second one: analyzing the regulatory system of each target gene, upon assessing which are the features that are more correlated to its expression, quantifying their impact on the output and describing how changes in these predictors relate to changes in the expression of the target gene.

In this case, lower scores are not particularly bad, because the relevance of a feature on the output is not related to the value of  $R^2$ . If the results of the analysis indicate that one-unit increase in the input is associated with an average of 0.5 increase in the output (i.e., the regression coefficient assigned to that feature is 0.5), this interpretation is correct regardless whether the  $R^2$  value is 0.4 or it is 0.9. However, if the main goal is to produce precise predictions, then the  $R^2$  becomes a concern, because lower values of  $R^2$  indicate a poor fit and a higher error.

We evaluate all the relationships between the predictors and the target and focus on the most relevant features, by visualizing the results of our analysis as a set of different networks: for each pathway, genes and the revealed correlations are graphically represented in networks, defined by grouping the genes according to their function-specific classification (defined in *Chapter 4* at the end of paragraph 4.1).

Relevant features are ordered according to their regression coefficients, which quantify the effect of each feature on the target gene expression: the higher is the estimated coefficient, the higher is the contribution of that feature on the gene expression and its relevance in the target gene regulation system. If gene methylation is selected as a relevant feature, its regression coefficient must correctly reflect the theory about methylation: gene hypermethylation is associated with a suppression of its expression, which means that the corresponding coefficient is negative.

We also evaluate which models better behave in terms of prediction of the output variable (i.e., models with the best linear fit), highlighting for each pathway the genes with a “good” value for either model M3 or model M5, setting a threshold of 0.6 (i.e., models with Adjusted  $R^2 \geq 0.6$ ). These best fitting genes form the set of regulations to be validated by the biologists during the experimental laboratory analysis following this work.

The next paragraphs detail the results obtained for each pathway, highlighting genes with a better linear model fit and their estimated regulatory elements, genes correlating with their own methylation,



which contributes to suppressing their expression, and more frequent regulators selected along the pathway, which may have a more relevant biological role within the whole genetic pathway.

### 7.1.1 DNA\_REPAIR pathway

The genes involved in the DNA repair mechanisms overall showing the best linear fit in the regression models are 9: BRCA1, ERCC1, ERCC2, FANCC, FANCD2, POLB, POLE, POLQ and TP53BP1 (*Table 7.1*).

In general, models accuracy increases while adding new features: this gradual development proves that adding features increasingly means progressively broadening the set of regulation hypotheses, until reaching a point where there is a set of features allowing an accurate prediction of the expression of the model gene. However, it is possible that some genes (BRCA1, ERCC2, POLE, TP53BP1) better behave in either the first (M2) or the second model (M3), rather than the last one (M5), showing their regulation systems mainly depend on the activity of genes in the same pathway and on their regulatory genes.

Model M2: the expression of each target gene is regulated by genes in the DNA\_REPAIR pathway and genes encoding for transcription factors binding to the target gene promoters.

The Adj. R<sup>2</sup> score is higher than 0.6 only for 3 genes of the pathway (POLE, TP53BP1 and FANCD2) and it reaches a maximum value of around 0.67 for gene POLE. The only features discarded a priori are the ones corresponding to missing values in TCGA, i.e., in this case only methylation of POLQ.

For 6 genes of the pathway, the gene methylation is selected as one of the relevant features involved in the regulation of their expression:

- BRCA1, with a regression coefficient of -0.3940;
- CDK12, with a regression coefficient of -0.1907;
- ERCC1, with a regression coefficient of -0.1214;
- FANCF, with a regression coefficient of -0.2287;
- ERCC4, with a regression coefficient of -0.2574;
- ERCC5, with a regression coefficient of -0.1430.

There are 4 most frequent regulators, selected as relevant regulatory features for 4 DNA-repair genes:

- POLQ (gene of the DNA\_REPAIR pathway) is selected as:
  - 1<sup>st</sup> (out of 6 features) for FANCD2 with a regression coefficient of 0.5649;
  - 2<sup>nd</sup> (out of 4 features) for BRCA1 with a regression coefficient of 0.3057;
  - 5<sup>th</sup> (out of 5 features) for MLH1 with a regression coefficient of -0.1146;
  - 10<sup>th</sup> (out of 10 features) for OGG1 with a regression coefficient of -0.3507;
- POLE (gene of the DNA\_REPAIR pathway) is selected as:
  - 1<sup>st</sup> (out of 5 features) for FANCA with a regression coefficient of 0.5217;
  - 2<sup>nd</sup> (out of 5 features) for FANCC with a regression coefficient of 0.2999;
  - 2<sup>nd</sup> (out of 8 features) for FANCF with a regression coefficient of 0.2506;
  - 6<sup>th</sup> (out of 6 features) for XPA with a regression coefficient of -0.2565;

Table 7.1: DNA\_REPAIR genes with either M3 or M5 model score higher than the 0.6 threshold.

|                | M2   |        |                       |             | M3   |        |                       |             | M5   |        |                       |             |
|----------------|------|--------|-----------------------|-------------|------|--------|-----------------------|-------------|------|--------|-----------------------|-------------|
|                | R2   | Adj.R2 | Most Relevant Feature | Coefficient | R2   | Adj.R2 | Most Relevant Feature | Coefficient | R2   | Adj.R2 | Most Relevant Feature | Coefficient |
| <b>POLB</b>    | 0.20 | 0.19   | HDAC2                 | 0.3399      | 0.55 | 0.54   | THAP1                 | 0.7135      | 0.69 | 0.68   | THAP1                 | 0.5420      |
| <b>FANCC</b>   | 0.58 | 0.57   | XPA                   | 0.3814      | 0.65 | 0.64   | XPA                   | 0.3988      | 0.68 | 0.67   | XPA                   | 0.3552      |
| <b>POLQ</b>    | 0.49 | 0.49   | FANCD2                | 0.6733      | 0.63 | 0.62   | FANCD2                | 0.4926      | 0.67 | 0.66   | FANCD2                | 0.4779      |
| <b>TP53BP1</b> | 0.68 | 0.67   | ZSCAN29               | 0.4986      | 0.64 | 0.63   | ZSCAN29               | 0.4772      | 0.66 | 0.65   | ZSCAN29               | 0.4646      |
| <b>FANCD2</b>  | 0.61 | 0.60   | POLQ                  | 0.5649      | 0.63 | 0.62   | POLQ                  | 0.5386      | 0.63 | 0.63   | POLQ                  | 0.5799      |
| <b>ERCC2</b>   | 0.52 | 0.51   | ERCC1                 | 0.5923      | 0.64 | 0.63   | ERCC1                 | 0.5457      | 0.64 | 0.63   | ERCC1                 | 0.5332      |
| <b>POLE</b>    | 0.68 | 0.68   | FANCA                 | 0.3224      | 0.58 | 0.57   | FANCA                 | 0.4439      | 0.63 | 0.62   | FANCA                 | 0.3124      |
| <b>ERCC1</b>   | 0.57 | 0.55   | ERCC2                 | 0.5905      | 0.60 | 0.59   | ERCC2                 | 0.6395      | 0.61 | 0.60   | ERCC2                 | 0.5420      |
| <b>BRCA1</b>   | 0.55 | 0.54   | FANCC                 | 0.3077      | 0.63 | 0.62   | FANCC                 | 0.2643      | 0.60 | 0.59   | FANCC                 | 0.2336      |

- FANCD2 (gene of the DNA\_REPAIR pathway) is selected as:
  - 1<sup>st</sup> (out of 5 features) for MLH1 with a regression coefficient of 0.4074;
  - 1<sup>st</sup> (out of 10 features) for OGG1 with a regression coefficient of 0.4532;
  - 1<sup>st</sup> (out of 4 features) for POLQ with a regression coefficient of 0.6733;
  - 1<sup>st</sup> (out of 8 features) for RAD51 with a regression coefficient of 0.4551;
- HCFC1 (regulatory gene of the DNA\_REPAIR pathway) is selected as:
  - 1<sup>st</sup> (out of 6 features) for PARP1 with a regression coefficient of 0.3347;
  - 2<sup>nd</sup> (out of 10 features) for POLE with a regression coefficient of 0.1900;
  - 2<sup>nd</sup> (out of 9 features) for ERCC2 with a regression coefficient of 0.3464;
  - 11<sup>th</sup> (out of 11 features) for PALB2 with a regression coefficient of -0.1770.

Model M3: the expression of each target gene is regulated by genes in the DNA\_REPAIR pathway and genes encoding for transcription factors binding to the target gene promoters or to the promoters of other genes in this pathway.

The Adj. R<sup>2</sup> score is higher than 0.6 for 6 genes of the pathway (FANCC, TP53BP1, ERCC2, FANCD2, POLQ and BRCA1) and it reaches a maximum value of around 0.64 for gene FANCC. The only features discarded a priori are the ones corresponding to missing values in TCGA, i.e., the methylation of POLQ and the expression of the regulatory gene EMSY.

For 5 genes of the pathway the gene methylation is selected as one of the relevant features involved in the regulation of their expression:

- BRCA1, with a regression coefficient of -0.3442;
- ERCC1, with a regression coefficient of -0.1347;
- FANCF, with a regression coefficient of -0.1314;
- ERCC4, with a regression coefficient of -0.1817;
- ERCC5, with a regression coefficient of -0.1630.

Differently from the previous model, gene CDK12 loses its methylation from the set of relevant features, because more relevant regulators are found among the set of candidate regulatory genes of the

pathway: in particular its methylation, along with the contribution of its regulatory gene TAF15 and the gene BRCA1 of the current DNA\_REPAIR pathway, is improved by selected regulatory genes SUZ12, BCLAF1, RUNX1 and ZNF217.

There are 4 most frequent regulators, selected as significant regulatory features for 4 DNA-repair genes:

- POLQ (gene of the DNA\_REPAIR pathway) is selected as:
  - 1<sup>st</sup> (out of 7 features) for FANCD2 with a regression coefficient of 0.5386;
  - 2<sup>nd</sup> (out of 8 features) for BRCA1 with a regression coefficient of 0.2603;
  - 7<sup>th</sup> (out of 9 features) for MLH1 with a regression coefficient of -0.2060;
  - 9<sup>th</sup> (out of 9 features) for OGG1 with a regression coefficient of -0.4165;
  
- FANCD2 (gene of the DNA\_REPAIR pathway) is selected as:
  - 1<sup>st</sup> (out of 8 features) for POLQ with a regression coefficient of 0.4926;
  - 1<sup>st</sup> (out of 9 features) for OGG1 with a regression coefficient of 0.4220;
  - 2<sup>nd</sup> (out of 9 features) for MLH1 with a regression coefficient of 0.2916;
  - 2<sup>nd</sup> (out of 11 features) for RAD51 with a regression coefficient of 0.3216;
  
- SUZ12 (regulatory gene of the DNA\_REPAIR pathway) is selected as:
  - 1<sup>st</sup> (out of 5 features) for CDK12 with a regression coefficient of 0.3550;
  - 2<sup>nd</sup> (out of 8 features) for BRCA1 with a regression coefficient of 0.1691;
  - 3<sup>rd</sup> (out of 9 features) for PALB2 with a regression coefficient of 0.1983;
  - 6<sup>th</sup> (out of 9 features) for OGG1 with a regression coefficient of 0.1212;
  
- ZHX1 (regulatory gene of the DNA\_REPAIR pathway) is selected as:
  - 2<sup>nd</sup> (out of 1 features) for XPA with a regression coefficient of 0.2350;
  - 3<sup>rd</sup> (out of 7 features) for ERCC5 with a regression coefficient of 0.1853;
  - 7<sup>th</sup> (out of 11 features) for FANCF with a regression coefficient of 0.1212;
  - 11<sup>th</sup> (out of 13 features) for ERCC1 with a regression coefficient of -0.1759.

Model M5: the expression of each target gene is regulated by genes in the DNA\_REPAIR pathway, genes encoding for transcription factors binding to the target gene promoters or to the promoters of other genes in this pathway, genes in the STEM\_CELLS pathway and their regulatory genes.

The Adj. R<sup>2</sup> score is higher than 0.6 for 8 genes of the pathway (POLB, FANCC, POLQ, TP53BP1, FANCD2, ERCC2, POLE and ERCC1) and it reaches a maximum value of around 0.68 for gene POLB. *Table 7.2* shows an excerpt of the whole set of relevant features in model M5 for some of the best genes in the DNA\_REPAIR pathway. The only features discarded a priori are the ones corresponding to missing values in TCGA, i.e., the methylation of POLQ and the expression of the regulatory gene EMSY.

For 4 genes of the pathway the gene methylation is selected as one of the relevant features involved in the regulation of their expression:

- BRCA1, with a regression coefficient of -0.3475;

Table 7.2: Model M5 best DNA\_REPAIR genes and their features.

| GENE           | Significant Feature | Adj.R2  | Regression Coefficient | Feature Description                                 |
|----------------|---------------------|---|------------------------|---|
| <b>POLB</b>    | THAP1               | <b>0.68</b>   | 0.5420                 | Candidate regulatory gene of the DNA_REPAIR pathway |
|                | SRSF7               |   | 0.2957                 | Candidate regulatory gene of the STEM_CELLS pathway |
|                | IKBKB               |   | 0.2249                 | Gene of the STEM_CELLS pathway                      |
|                | ZNF207              |   | 0.1338                 | Candidate regulatory gene of the STEM_CELLS pathway |
|                | XRCC3               |   | 0.1098                 | Candidate regulatory gene of the STEM_CELLS pathway |
|                | ITGB1               |   | 0.0837                 | Gene of the STEM_CELLS pathway                      |
|                | MCM3                |   | -0.0770                | Candidate regulatory gene of the STEM_CELLS pathway |
|                | PTCH1               |   | -0.1124                | Gene of the STEM_CELLS pathway                      |
|                | SRSF1               |   | -0.1350                | Candidate regulatory gene of the STEM_CELLS pathway |
|                | ELF1                |   | -0.1629                | Candidate regulatory gene of the DNA_REPAIR pathway |
| ZBTB1          | -0.1823             | Candidate regulatory gene of the STEM_CELLS pathway |                        |   |
| <b>FANCC</b>   | XPA                 | <b>0.67</b>   | 0.3552                 | Gene of the DNA_REPAIR pathway                      |
|                | POLE                |   | 0.2505                 | Gene of the DNA_REPAIR pathway                      |
|                | MCM7                |   | 0.1869                 | Candidate regulatory gene of the STEM_CELLS pathway |
|                | BRCA1               |   | 0.1826                 | Gene of the DNA_REPAIR pathway                      |
|                | ZBTB5               |   | 0.1681                 | Candidate regulatory gene of the DNA_REPAIR pathway |
|                | EPCAM               |   | 0.1157                 | Gene of the STEM_CELLS pathway                      |
|                | ITGB1               |   | 0.1146                 | Gene of the STEM_CELLS pathway                      |
|                | THRA                |   | 0.1119                 | Candidate regulatory gene of the STEM_CELLS pathway |
|                | XRCC3               |   | 0.0985                 | Candidate regulatory gene of the STEM_CELLS pathway |
|                | SOX2                |   | 0.0624                 | Gene of the STEM_CELLS pathway                      |
|                | SIRT1               |   | -0.1191                | Gene of the STEM_CELLS pathway                      |
| <b>TP53BP1</b> | ZSCAN29             | <b>0.65</b>   | 0.4646                 | Candidate regulatory gene of the model gene TP53BP1 |
|                | ZBTB40              |   | 0.3737                 | Candidate regulatory gene of the model gene TP53BP1 |
|                | MAML1               |   | 0.1917                 | Gene of the STEM_CELLS pathway                      |
| <b>ERCC1</b>   | ERCC2               | <b>0.60</b>   | 0.5420                 | Gene of the DNA_REPAIR pathway                      |
|                | HNRNPUL1            |   | 0.2743                 | Candidate regulatory gene of the STEM_CELLS pathway |
|                | PTTG1               |   | 0.1799                 | Candidate regulatory gene of the DNA_REPAIR pathway |
|                | NCOA3               |   | -0.0821                | Candidate regulatory gene of the STEM_CELLS pathway |
|                | ZHX1                |   | -0.1278                | Candidate regulatory gene of the model gene ERCC1   |
|                | E2F4                |   | -0.1561                | Candidate regulatory gene of the DNA_REPAIR pathway |
|                | POLR2A              |   | -0.1605                | Candidate regulatory gene of the model gene ERCC1   |
|                | RCOR1               |   | -0.1712                | Candidate regulatory gene of the model gene ERCC1   |
|                | SKIL                |   | -0.1874                | Candidate regulatory gene of the model gene ERCC1   |
|                | PTBP1               |   | -0.2344                | Candidate regulatory gene of the model gene ERCC1   |

- FANCF, with a regression coefficient of -0.1659;
- ERCC4, with a regression coefficient of -0.2608;
- ERCC5, with a regression coefficient of -0.1774.

Differently from the previous model, gene ERCC1 loses its methylation from the set of significant features, because more relevant regulators are found among the set of candidate regulatory genes of the pathway: in particular its methylation, along with the contribution of its regulatory genes GATAD2B and TBP and the regulatory genes of the DNA\_REPAIR pathway, ZC3H8 and ZMIZ1, is improved by selected regulatory genes of the STEM\_CELLS pathway, HNRNPUL1 and NCOA3.

The small set of genes in *Table 7.2* highlights the strong interrelationship between the DNA\_REPAIR and the STEM\_CELLS pathway and the relevant impact that the genes in the latter pathway have in regulating the activity of genes involved in the DNA damages recovery mechanisms.

As also reported in the networks at the end of the paragraph, these evaluation can be generalized on the whole set of genes in the DNA\_REPAIR pathway.

There is one single most frequent regulator, selected as a relevant regulatory feature for 5 DNA-repair genes:

- XRCC3 (regulatory gene of the STEM\_CELLS pathway) is selected as:
  - 2<sup>nd</sup> (out of 11 features) for POLQ with a regression coefficient of 0.1916;
  - 4<sup>th</sup> (out of 9 features) for POLE with a regression coefficient of 0.1625;
  - 5<sup>th</sup> (out of 11 features) for POLE with a regression coefficient of 0.1098;
  - 9<sup>th</sup> (out of 11 features) for FANCC with a regression coefficient of 0.0985;
  - 9<sup>th</sup> (out of 12 features) for PARP1 with a regression coefficient of -0.0767;

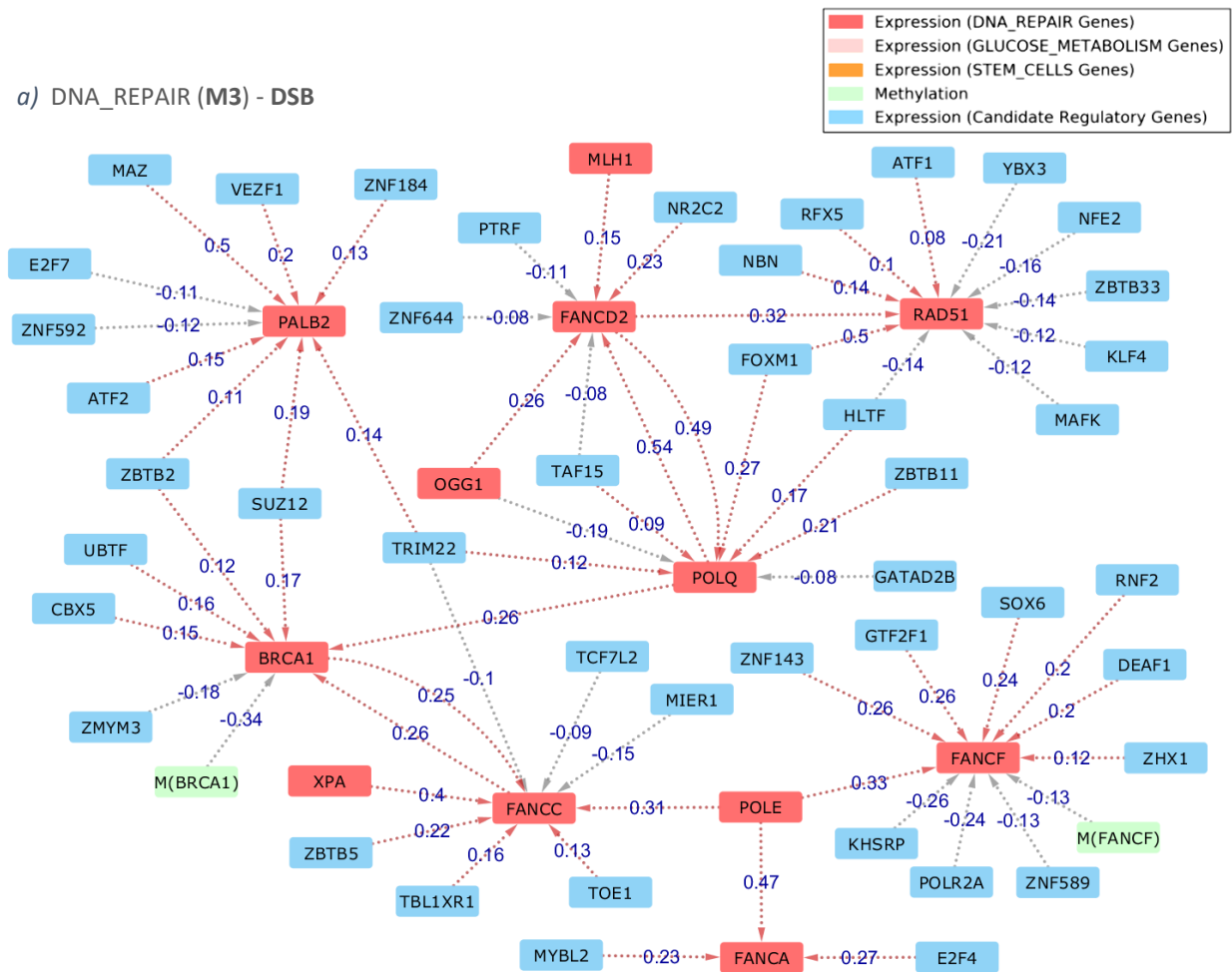
Finally, all the results are graphically represented through expression networks in Cytoscape, as described in *Chapter 3*. An example is reported in *Figure 7.1*, showing model M3 (a) and model M5 (b) networks for the genes of the *DNA Double Strand Breaks* (DSB) subclass.

In general, all the drawn networks follow the same color-legend: in red genes in the DNA\_REPAIR pathway, in orange genes in the STEM\_CELLS pathway, in pink genes in the GLUCOSE\_METABOLISM pathway, in green gene methylations and in light blue regulatory genes, regardless of the pathway they belong to. The relationships are represented by directed edges starting in the regulatory element and incoming in the target gene: red edges correspond to positive regulations, while grey edges correspond negative regulations. This code of colors allows an easy reading of the results and gives us the possibility of assessing the effect of both methylation and transcription factors, as well as the interrelationships between pathways of interest (i.e., how each target gene is regulated and how it in turn participates, if so, in regulating other target genes).

Networks pictured in *Figure 7.1* describe the regulation systems of a subset of 8 genes from the DNA\_REPAIR pathway (BRCA1, FANCA, FANCC, FANCD2, FANCF, PALB2, POLQ and RAD51), showing how the regulation of their expression changes from considering only target and regulatory genes of the DNA\_REPAIR pathway, to including also activity related to genes of the STEM\_CELLS pathway.

As a whole, we could not find common transcription factors regulating the expression of genes involved in the same DNA\_REPAIR pathway. This is partly unexpected, but there can be several reasons for this finding. A possible explanation could be that all these DNA repair pathways are multistep processes involving many different proteins having roles in other cellular processes. It may be that under specific conditions (i.e., DNA damage or other stress cellular conditions) different processes are activated. Our analysis relies on the expression of genes in basal conditions, not taking into consideration what occurs in activated/stress conditions or after therapy.

a) DNA\_REPAIR (M3) - DSB



b) DNA\_REPAIR (M5) - DSB

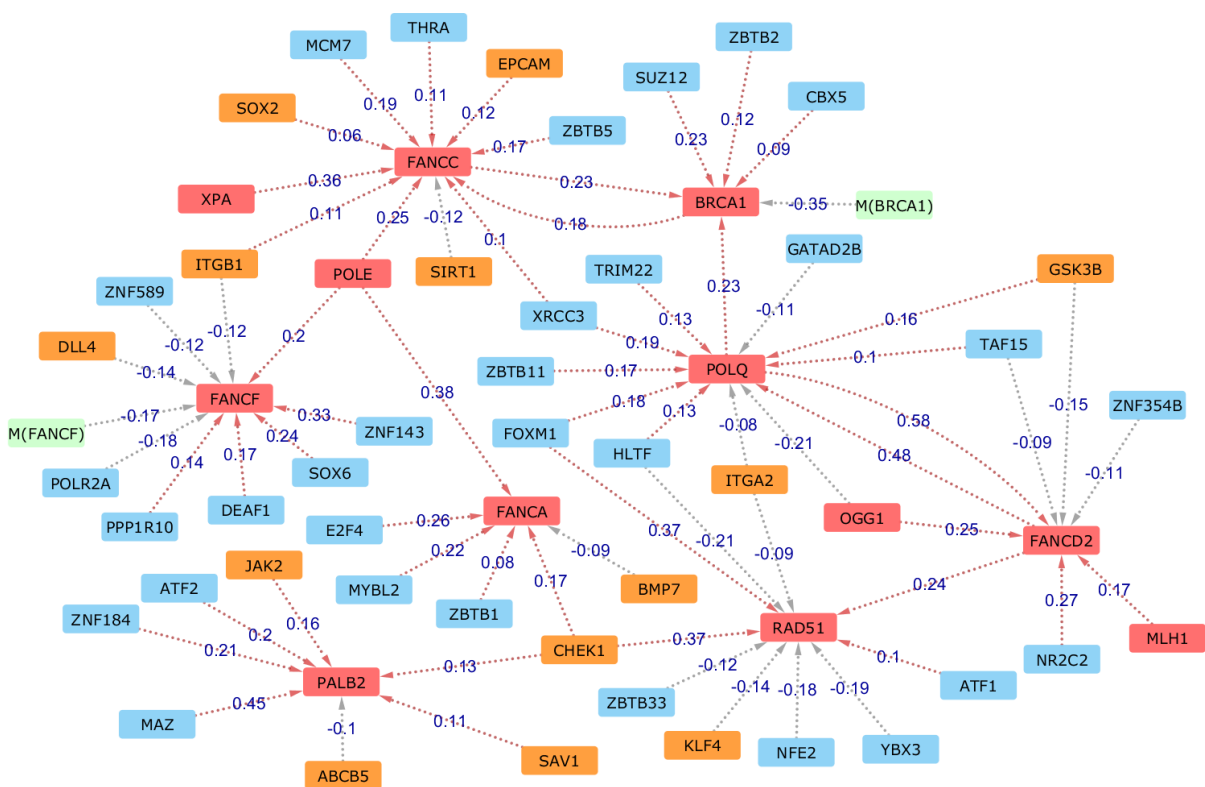


Figure 7.1: Expression networks from linear regression models M3 (a) and M5 (b) of the DSB subclass.

## 7.1.2 STEM\_CELL pathway

The genes involved in stem cells overall showing the best linear fit in the regression models are 11: AXL, CHEK1, DNMT1, ENG, ITGA4, JAK2, LATS1, MAML1, NOTCH2, PECAM1 and PTPRC (*Table 7.3*).

Even for this pathway, in general, models accuracy increases while progressively adding new features. However, some genes (ENG, NOTCH2, PECAM1) better behave in either M2 model or M3 model, showing their regulation systems mainly depend on the activity of genes in the same pathway and on their regulatory genes.

Model M2: the expression of each target gene is regulated by genes in the STEM\_CELL pathway and genes encoding for transcription factors binding to the target gene promoters.

The Adj.  $R^2$  score is higher than 0.6 for 5 genes of the pathway (PTPRC, ITGA4, PECAM1, ENG and LATS1) and it reaches a maximum value of around 0.77 for gene PTPRC. The only features discarded a priori are the ones corresponding to missing values in TCGA, i.e., the methylation of CD24, DLL4, HDAC1, JAK2, LIN28B, NANOG, NOTCH2 genes.

For 10 genes of the pathway the gene methylation is selected as one of the relevant features regulating their expression:

- ATM, with a regression coefficient of -0.2246;
- CD34, with a regression coefficient of -0.1010;
- CHEK1, with a regression coefficient of -0.1121;
- ENG, with a regression coefficient of -0.0896;
- EPCAM, with a regression coefficient of -0.3359;
- CXCL8, with a regression coefficient of -0.1422;
- MAML1, with a regression coefficient of -0.1017;
- PLAT, with a regression coefficient of -0.2656;
- POU5F1, with a regression coefficient of -0.2565;
- SAV1, with a regression coefficient of -0.2558.

There is one single most frequent regulator, selected as a relevant regulatory feature for 9 stem cells target genes:

- PTPRC (gene of the STEM\_CELL pathway) is selected as:
  - 1<sup>st</sup> (out of 4 features) for CD38 with a regression coefficient of 0.4178;
  - 1<sup>st</sup> (out of 6 features) for CD44 with a regression coefficient of 0.3582;
  - 1<sup>st</sup> (out of 5 features) for ITGA4 with a regression coefficient of 0.5299;
  - 1<sup>st</sup> (out of 10 features) for JAK2 with a regression coefficient of 0.3359;
  - 1<sup>st</sup> (out of 4 features) for MS4A1 with a regression coefficient of 0.4193;
  - 1<sup>st</sup> (out of 5 features) for PECAM1 with a regression coefficient of 0.4741;
  - 3<sup>rd</sup> (out of 5 features) for AXL with a regression coefficient of 0.1786;
  - 6<sup>th</sup> (out of 6 features) for DLL4 with a regression coefficient of -0.2629;
  - 7<sup>th</sup> (out of 8 features) for SMO with a regression coefficient of -0.2176.



Table 7.3: STEM\_CELLs genes with either M3 or M5 model score higher than the 0.6 threshold.

|               | M2   |        |                       |             | M3   |        |                       |             | M5   |        |                       |             |
|---------------|------|--------|-----------------------|-------------|------|--------|-----------------------|-------------|------|--------|-----------------------|-------------|
|               | R2   | Adj.R2 | Most Relevant Feature | Coefficient | R2   | Adj.R2 | Most Relevant Feature | Coefficient | R2   | Adj.R2 | Most Relevant Feature | Coefficient |
| <b>PTPRC</b>  | 0.78 | 0.77   | ITGA4                 | 0.4530      | 0.82 | 0.82   | IKZF1                 | 0.5958      | 0.83 | 0.82   | IKZF1                 | 0.5857      |
| <b>ITGA4</b>  | 0.70 | 0.69   | PTPRC                 | 0.5299      | 0.76 | 0.75   | PTPRC                 | 0.4427      | 0.77 | 0.76   | PTPRC                 | 0.4511      |
| <b>DNMT1</b>  | 0.43 | 0.42   | SIN3B                 | 0.3627      | 0.69 | 0.69   | SMARCA4               | 0.6022      | 0.71 | 0.70   | SMARCA4               | 0.5505      |
| <b>LATS1</b>  | 0.64 | 0.63   | ARID1B                | 0.5622      | 0.72 | 0.71   | ARID1B                | 0.4710      | 0.71 | 0.70   | ARID1B                | 0.5248      |
| <b>MAML1</b>  | 0.55 | 0.53   | HCFC1                 | 0.3149      | 0.68 | 0.67   | ZNF354B               | 0.4484      | 0.70 | 0.69   | ZNF354B               | 0.4257      |
| <b>JAK2</b>   | 0.54 | 0.52   | PTPRC                 | 0.3359      | 0.69 | 0.68   | TRIM22                | 0.3378      | 0.69 | 0.68   | TRIM22                | 0.3314      |
| <b>CHEK1</b>  | 0.56 | 0.55   | NFRKB                 | 0.4920      | 0.64 | 0.63   | NFRKB                 | 0.4643      | 0.67 | 0.66   | NFRKB                 | 0.4343      |
| <b>PECAM1</b> | 0.64 | 0.64   | PTPRC                 | 0.4741      | 0.65 | 0.65   | PTPRC                 | 0.3183      | 0.64 | 0.64   | PTPRC                 | 0.3595      |
| <b>NOTCH2</b> | -    | -      | -                     | -           | 0.68 | 0.68   | CSDE1                 | 0.7630      | 0.64 | 0.64   | CSDE1                 | 0.7441      |
| <b>AXL</b>    | 0.49 | 0.48   | ITGA4                 | 0.504       | 0.64 | 0.63   | TITGA4                | 0.358       | 0.64 | 0.63   | ITGA4                 | 0.3422      |
| <b>ENG</b>    | 0.64 | 0.64   | ZEB2                  | 0.4696      | 0.63 | 0.62   | ZEB2                  | 0.4547      | 0.62 | 0.62   | ZEB2                  | 0.4504      |

**Model M3:** the expression of each target gene is regulated by genes in the STEM\_CELLs pathway and genes encoding for transcription factors binding to the target gene promoters or to the promoters of other genes in this pathway.

The Adj.  $R^2$  score is higher than 0.6 for 11 genes of the pathway (PTPRC, ITGA4, LATS1, DNMT1, JAK2, NOTCH2, MAML1, PECAM1, CHEK1, AXL and ENG) and it reaches a maximum value of around 0.82 for gene PTPRC. The only features discarded a priori are the ones corresponding to missing values in TCGA, i.e., the methylation of CD24, DLL4, HDAC1, JAK2, LIN28B, NANOG, NOTCH2 and the expression of the regulatory gene EMSY.

For 10 genes of the pathway the gene methylation is selected as one of the relevant features regulating their expression:

- ATM, with a regression coefficient of -0.1591;
- CD34, with a regression coefficient of -0.1733;
- CHEK1, with a regression coefficient of -0.0960;
- DACH1, with a regression coefficient of -0.1215;
- EPCAM, with a regression coefficient of -0.2276;
- CXCL8, with a regression coefficient of -0.1628;
- MAML1, with a regression coefficient of -0.0723;
- PLAT, with a regression coefficient of -0.2482;
- POU5F1, with a regression coefficient of -0.1984;
- SAV1, with a regression coefficient of -0.1725.

Differently from the previous model, methylation becomes relevant for gene DACH1, while gene ENG loses its methylation from the set of relevant features, because more relevant regulators are found among the set of candidate regulatory genes of the pathway: in particular its methylation, along with the contribution of its regulatory genes L3MBTL2 and ATF1 and the gene SMO of the current STEM\_CELLs pathway, is improved by selected regulatory genes of the STEM\_CELLs pathway, TEAD2 and ZNF143.



There is one single most frequent regulator, selected as a relevant regulatory feature for 10 stem cells target genes:

- PAX8 (regulatory gene of the STEM\_CELLS pathway) is selected as:
  - 1<sup>st</sup> (out of 1 feature) for NOS2 with a regression coefficient of -0.1792;
  - 2<sup>nd</sup> (out of 8 features) for BMP7 with a regression coefficient of 0.2424;
  - 2<sup>nd</sup> (out of 3 features) for NANOG with a regression coefficient of -0.1257;
  - 6<sup>th</sup> (out of 7 features) for LIN28B with a regression coefficient of -0.1251;
  - 6<sup>th</sup> (out of 6 features) for MYCN with a regression coefficient of -0.1533;
  - 7<sup>th</sup> (out of 7 features) for KLF4 with a regression coefficient of -0.1542;
  - 8<sup>th</sup> (out of 8 features) for CD44 with a regression coefficient of -0.1997;
  - 10<sup>th</sup> (out of 12 features) for MYC with a regression coefficient of -0.1491;
  - 11<sup>th</sup> (out of 13 features) for ETFA with a regression coefficient of -0.1424;
  - 11<sup>th</sup> (out of 11 features) for ITGA4 with a regression coefficient of -0.1075;

Model M5: the expression of each target gene is regulated by genes in the STEM\_CELLS pathway, genes encoding for transcription factors binding to the target gene promoters or to the promoters of other genes in this pathway, genes in the DNA\_REPAIR pathway and their regulatory genes.

The Adj. R<sup>2</sup> score is higher than 0.6 for the same 11 genes of the pathway as in M3 (PTPRC, ITGA4, DNMT1, LATS1, MAML1, JAK2, CHEK1, PECAM1, NOTCH2, AXL and ENG) and it reaches a maximum value a little higher than 0.82 for gene PTPRC. *Table 7.4* shows an excerpt of the whole set of relevant features in model M5 for some of the best genes in the STEM\_CELLS pathway. The only features discarded a priori are the ones corresponding to missing values in TCGA, i.e., the methylation of CD24, DLL4, HDAC1, JAK2, LIN28B, NANOG, NOTCH2 and the expression of the regulatory gene EMSY.

For 9 genes of the pathway the gene methylation is selected as one of the relevant features regulating their expression, i.e., the same as in M3, with the exception of gene DACH1:

- ATM, with a regression coefficient of -0.1591;
- CD34, with a regression coefficient of -0.1733;
- CHEK1, with a regression coefficient of -0.0960;
- EPCAM, with a regression coefficient of -0.2276;
- CXCL8, with a regression coefficient of -0.1628;
- MAML1, with a regression coefficient of -0.0723;
- PLAT, with a regression coefficient of -0.2482;
- POU5F1, with a regression coefficient of -0.1984;
- SAV1, with a regression coefficient of -0.1725.

The same most frequent regulator is present, selected as a relevant regulatory feature for the same 10 stem cells target genes as in model M3:

- PAX8 (regulatory gene of the STEM\_CELLS pathway) is selected as:
  - 2<sup>nd</sup> (out of 2 features) for NOS2 with a regression coefficient of -0.1858;

Table 7.4: Model M5 best STEM\_CELLs genes and their features.

| GENE         | Significant Feature | Adj.R2      | Regression Coefficient | Feature Description                                 |
|--------------|---------------------|-------------|------------------------|---|
| <b>PTPRC</b> | IKZF1               | <b>0.82</b> | 0.5857                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | ITGA4               |             | 0.2392                 | Gene of the STEM_CELLs pathway                      |
|              | CD38                |             | 0.1394                 | Gene of the STEM_CELLs pathway                      |
|              | CD44                |             | 0.1217                 | Gene of the STEM_CELLs pathway                      |
|              | AXL                 |             | 0.0729                 | Gene of the STEM_CELLs pathway                      |
|              | ERCC2               |             | -0.1169                | Gene of the DNA_REPAIR pathway                      |
|              | ELF4                |             | -0.1252                | Candidate regulatory gene of the STEM_CELLs pathway |
| <b>ITGA4</b> | PTPRC               | <b>0.76</b> | 0.4511                 | Gene of the STEM_CELLs pathway                      |
|              | AXL                 |             | 0.2539                 | Gene of the STEM_CELLs pathway                      |
|              | MEF2A               |             | 0.2216                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | ERCC2               |             | 0.1325                 | Gene of the DNA_REPAIR pathway                      |
|              | MERTK               |             | 0.1047                 | Gene of the STEM_CELLs pathway                      |
|              | ZZZ3                |             | 0.0966                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | MLH1                |             | 0.0961                 | Gene of the DNA_REPAIR pathway                      |
|              | SMAD1               |             | 0.0924                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | SKIL                |             | -0.0715                | Candidate regulatory gene of the STEM_CELLs pathway |
|              | XPA                 |             | -0.0735                | Gene of the DNA_REPAIR pathway                      |
|              | NR2F2               |             | -0.0804                | Candidate regulatory gene of the STEM_CELLs pathway |
|              | PAX8                |             | -0.0921                | Candidate regulatory gene of the STEM_CELLs pathway |
|              | SOX6                |             | -0.1051                | Candidate regulatory gene of the STEM_CELLs pathway |
| <b>DNMT1</b> | SMARCA4             | <b>0.70</b> | 0.5505                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | POLE                |             | 0.1738                 | Gene of the DNA_REPAIR pathway                      |
|              | CHEK1               |             | 0.1565                 | Gene of the STEM_CELLs pathway                      |
|              | CC2D1A              |             | 0.1552                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | MTA2                |             | 0.1208                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | MCM3                |             | 0.1049                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | DEAF1               |             | -0.1112                | Candidate regulatory gene of the STEM_CELLs pathway |
|              | TEAD2               |             | -0.1282                | Candidate regulatory gene of the STEM_CELLs pathway |
| <b>CHEK1</b> | NFRKB               | <b>0.66</b> | 0.4343                 | Candidate regulatory gene of the model gene CHEK1   |
|              | RAD51               |             | 0.2605                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | FANCA               |             | 0.2058                 | Gene of the DNA_REPAIR pathway                      |
|              | CBX5                |             | 0.1290                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | SRSF1               |             | 0.1290                 | Candidate regulatory gene of the STEM_CELLs pathway |
|              | DACH1               |             | 0.1073                 | Gene of the STEM_CELLs pathway                      |
|              | YBX3                |             | 0.0818                 | Candidate regulatory gene of the DNA_REPAIR pathway |
|              | METHYLATION (CHEK1) |             | -0.1144                | Methylation of the model gene CHEK1                 |
|              | MAFK                |             | -0.1611                | Candidate regulatory gene of the STEM_CELLs pathway |
|              | ZNF384              |             | -0.1680                | Candidate regulatory gene of the STEM_CELLs pathway |
|              | RB1                 |             | -0.1680                | Candidate regulatory gene of the model gene CHEK1   |
|              | ZKSCAN1             |             | -0.1872                | Candidate regulatory gene of the STEM_CELLs pathway |

- 2<sup>nd</sup> (out of 8 fetatures) for BMP7 with a regression coefficient of 0.2264;
- 2<sup>nd</sup> (out of 3 features) for NANOG with a regression coefficient of -0.1257;
- 6<sup>th</sup> (out of 9 features) for LIN28B with a regression coefficient of -0.1277;
- 6<sup>th</sup> (out of 7 features) for MYCN with a regression coefficient of -0.1625;
- 8<sup>th</sup> (out of 8 features) for CD44 with a regression coefficient of -0.1997;
- 9<sup>th</sup> (out of 9 features) for KLF4 with a regression coefficient of -0.1344;
- 10<sup>th</sup> (out of 12 features) for ETFA with a regression coefficient of -0.1213;
- 11<sup>th</sup> (out of 13 features) for MYC with a regression coefficient of -0.1132;
- 12<sup>th</sup> (out of 12 features) for ITGA4 with a regression coefficient of -0.0921;

Unlike what happens swapping the pathways, in this case it is interesting to notice the limited impact of the DNA\_REPAIR pathway on these genes, whose regulation systems mainly depends on genes or regulatory genes of the STEM\_CELLS pathway itself.

Finally, all the results are graphically represented through expression networks in Cytoscape. An example for this pathway is reported in *Figure 7.2*, showing model M3 (a) and model M5 (b) networks for the genes of the *Cancer Therapeutic Targets* subclass: it describes the regulation systems of a subset of 17 genes from the STEM\_CELLS pathway (ABCG2, ATM, AXL, CHEK1, DDR1, DKK1, EPCAM, FZD7, GSK3B, ID1, IKKKB, JAK2, KLF17, NFKB1, PTCH1, SMO and STAT3), showing how the regulation of their expression changes from considering only target and regulatory genes of the STEM\_CELLS pathway, to including also activity related to genes of the DNA\_REPAIR pathway, whose impact is here limited.

### 7.1.3 GLUCOSE\_METABOLISM pathway

The genes involved in glucose metabolism overall showing the best linear fit in the regression models are 10: ACLY, ACO2, ALDOA, DLAT, HK3, MDH1, PHKA1, PRPS1, SDHD and TPI1 (*Table 7.5*).

Even for this third pathway, models accuracy increases while progressively adding new features, except for MDH1 which better behave in model M3.

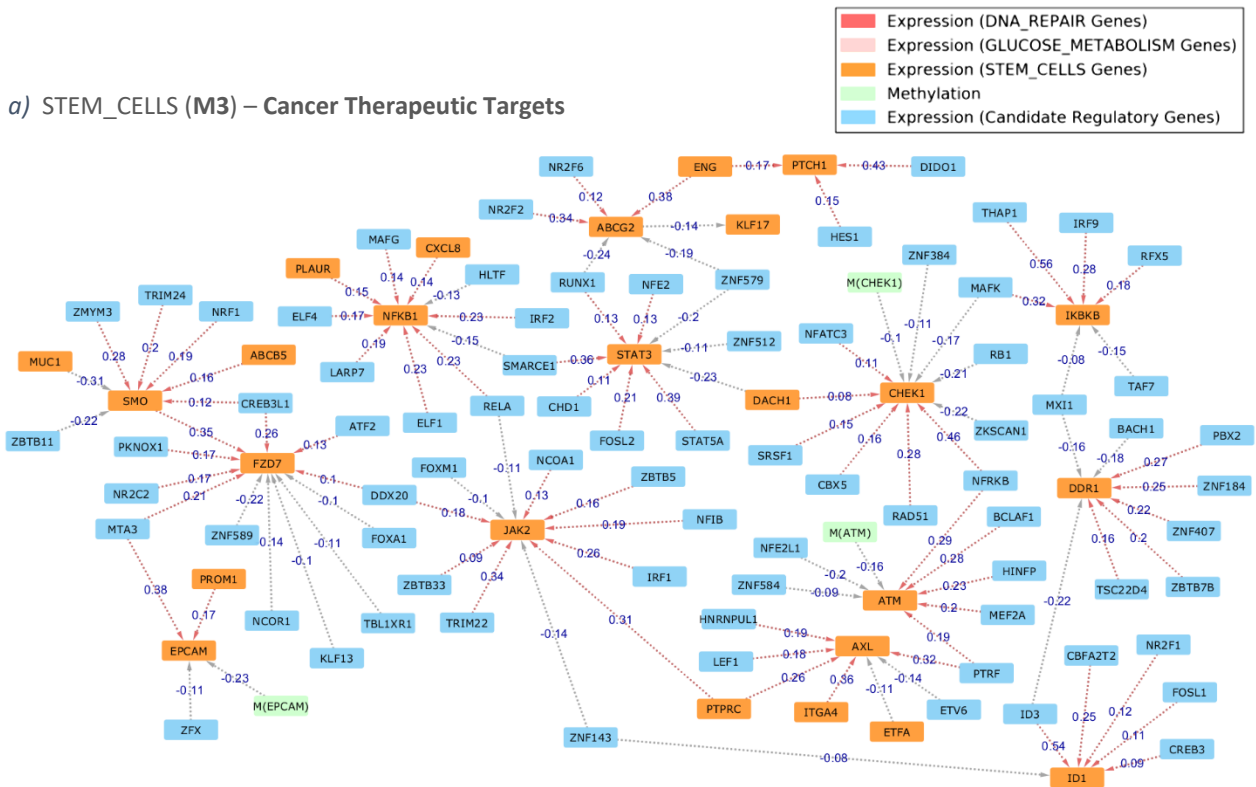
Model M2: the expression of each target gene is regulated by genes in the GLUCOSE\_METABOLISM pathway and genes encoding for transcription factors binding to the target gene promoters.

The Adj. R<sup>2</sup> score is higher than 0.6 only for 3 genes of the pathway (SDHD, DLAT and ACO2) and it reaches a maximum value of around 0.78 for gene SDHD. The only features discarded a priori are the ones corresponding to missing values in TCGA, i.e., the methylation of HK2, PDPR, SUCLG2 genes.

For 14 genes of the pathway the gene methylation is selected as one of the relevant features involved in the regulation of their expression:

- AGL, with a regression coefficient of -0.2198;
- ALDOC, with a regression coefficient of -0.3634;
- DLD, with a regression coefficient of -0.2418;
- IDH3A, with a regression coefficient of -0.1198;
- IDH3B, with a regression coefficient of -0.2608;
- MDH2, with a regression coefficient of -0.1573;
- PCK1, with a regression coefficient of -0.3248;
- PCK2, with a regression coefficient of -0.1278;
- PDHA1, with a regression coefficient of -0.1490;
- PDK4, with a regression coefficient of -0.1237;
- PGM3, with a regression coefficient of -0.2025;
- PYGM, with a regression coefficient of -0.1122;
- RPE, with a regression coefficient of -0.1445;
- SDHA, with a regression coefficient of -0.2296.

a) STEM\_CELLS (M3) – Cancer Therapeutic Targets



b) STEM\_CELLS (M5) - Cancer Therapeutic Targets

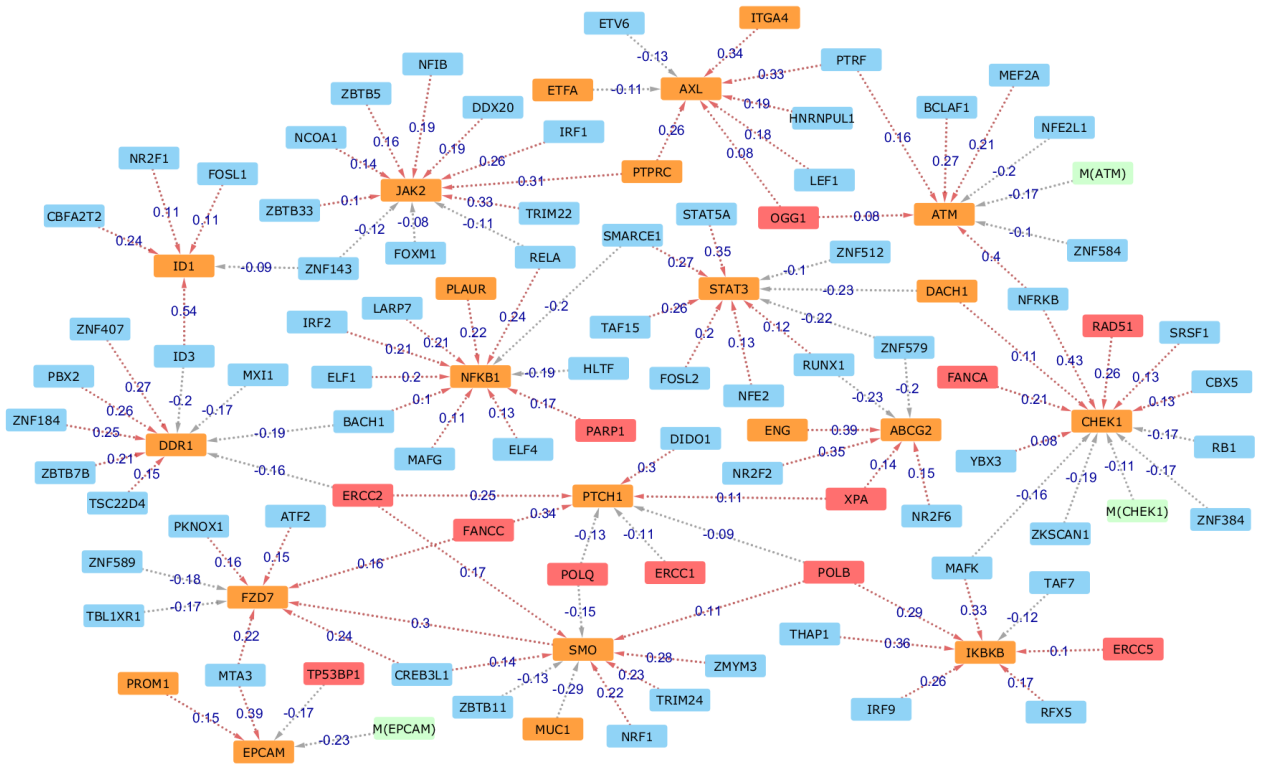


Figure 7.2: Expression networks from linear regression models M3 (a) and M5 (b) of the Cancer Therapeutic Targets subclass.

Table 7.5: GLUCOSE\_METABOLISM genes with either M3 or M5 model score higher than the 0.6 threshold.

|              | M2   |        |                       |             | M3   |        |                       |             | M5   |        |                       |             |
|--------------|------|--------|-----------------------|-------------|------|--------|-----------------------|-------------|------|--------|-----------------------|-------------|
|              | R2   | Adj.R2 | Most Relevant Feature | Coefficient | R2   | Adj.R2 | Most Relevant Feature | Coefficient | R2   | Adj.R2 | Most Relevant Feature | Coefficient |
| <b>SDHD</b>  | 0.78 | 0.78   | DLAT                  | 0.7363      | 0.79 | 0.78   | DLAT                  | 0.7314      | 0.79 | 0.79   | DLAT                  | 0.7345      |
| <b>TPI1</b>  | 0.61 | 0.60   | ENO2                  | 0.4773      | 0.72 | 0.71   | PHB2                  | 0.3495      | 0.76 | 0.76   | ENO2                  | 0.2825      |
| <b>DLAT</b>  | 0.69 | 0.68   | SDHD                  | 0.7490      | 0.73 | 0.72   | SDHD                  | 0.8270      | 0.76 | 0.75   | SDHD                  | 0.7804      |
| <b>HK3</b>   | 0.55 | 0.54   | FBP1                  | 0.6465      | 0.59 | 0.58   | FBP1                  | 0.6356      | 0.71 | 0.71   | FBP1                  | 0.5655      |
| <b>ACO2</b>  | 0.69 | 0.68   | L3MBTL2               | 0.5472      | 0.65 | 0.64   | L3MBTL2               | 0.5814      | 0.70 | 0.69   | L3MBTL2               | 0.5854      |
| <b>ACLY</b>  | 0.51 | 0.49   | SUZ12                 | 0.3391      | 0.56 | 0.55   | SUZ12                 | 0.3655      | 0.68 | 0.67   | STAT3                 | 0.2612      |
| <b>PRPS1</b> | 0.59 | 0.57   | PGK1                  | 0.3770      | 0.64 | 0.63   | PDK3                  | 0.2626      | 0.64 | 0.63   | PDK3                  | 0.2945      |
| <b>PHKA1</b> | 0.41 | 0.40   | PRPS1                 | 0.2500      | 0.52 | 0.50   | TAF1                  | 0.3303      | 0.65 | 0.63   | TAF1                  | 0.3758      |
| <b>ALDOA</b> | 0.50 | 0.49   | PHKG2                 | 0.3397      | 0.53 | 0.52   | PHKG2                 | 0.3136      | 0.64 | 0.62   | PHKG2                 | 0.3420      |
| <b>MDH1</b>  | 0.57 | 0.55   | UGP2                  | 0.4278      | 0.66 | 0.65   | UGP2                  | 0.3906      | 0.57 | 0.56   | SRSF7                 | 0.3855      |

There are 2 most frequent regulators, selected as relevant regulatory features for 10 glucose metabolism target genes:

- SUCLG1 (gene of the GLUCOSE\_METABOLISM pathway) is selected as:
  - 1<sup>st</sup> (out of 10 features) for FH with a regression coefficient of 0.1944;
  - 1<sup>st</sup> (out of 8 features) for PDK2 with a regression coefficient of 0.2395;
  - 1<sup>st</sup> (out of 10 features) for SDHC with a regression coefficient of 0.3659;
  - 2<sup>nd</sup> (out of 5 features) for RPIA with a regression coefficient of 0.2701;
  - 2<sup>nd</sup> (out of 8 features) for SDHD with a regression coefficient of 0.1968;
  - 3<sup>rd</sup> (out of 10 features) for SDHB a regression coefficient of 0.1796;
  - 3<sup>rd</sup> (out of 5 features) for SUCLA2 with a regression coefficient of 0.1769;
  - 4<sup>th</sup> (out of 11 features) for MDH1 with a regression coefficient of 0.2055;
  - 5<sup>th</sup> (out of 13 features) for ACO2 with a regression coefficient of 0.1116;
  - 7<sup>th</sup> (out of 7 features) for ENO2 with a regression coefficient of -0.3067;
- PHKB (gene of the GLUCOSE\_METABOLISM pathway) is selected as:
  - 2<sup>nd</sup> (out of 3 features) for PDP2 with a regression coefficient of 0.1832;
  - 2<sup>nd</sup> (out of 9 features) for SUCLG2 with a regression coefficient of 0.2254;
  - 3<sup>rd</sup> (out of 10 features) for SDHC with a regression coefficient of 0.2183;
  - 4<sup>th</sup> (out of 6 features) for H6PD with a regression coefficient of 0.2405;
  - 4<sup>th</sup> (out of 4 features) for PDPR with a regression coefficient of 0.0921;
  - 4<sup>th</sup> (out of 5 features) for PKLR with a regression coefficient of -0.1542;
  - 5<sup>th</sup> (out of 9 features) for IDH3A with a regression coefficient of 0.0977;
  - 6<sup>th</sup> (out of 11 features) for MDH1 with a regression coefficient of -0.0955;
  - 6<sup>th</sup> (out of 9 features) for PHKG1 with a regression coefficient of -0.0912;
  - 10<sup>th</sup> (out of 10 features) for PDHB with a regression coefficient of -0.1710;

Model M3: the expression of each target gene is regulated by genes in the GLUCOSE\_METABOLISM pathway and genes encoding for transcription factors binding to the target gene promoters or to the promoters of other genes in this pathway.

The Adj.  $R^2$  score is higher than 0.6 for 6 genes of the pathway (SDHD, TPI1, DLAT, MDH1, ACO2, and PRPS1) and it reaches a maximum value of around 0.78 for gene SDHD. The only features discarded a priori are the ones corresponding to missing values in TCGA, i.e., the methylation of HK2, PDPR, SUCLG2 and the expression of the regulatory gene EMSY.

For 12 genes of the pathway the gene methylation is selected as one of the relevant features involved in the regulation of their expression:

- AGL, with a regression coefficient of -0.0917;
- ALDOC, with a regression coefficient of -0.3614;
- DLD, with a regression coefficient of -0.2463;
- IDH3B, with a regression coefficient of -0.2262;
- MDH2, with a regression coefficient of -0.1419;
- PCK1, with a regression coefficient of -0.3192;
- PDK3, with a regression coefficient of -0.1905;
- PDK4, with a regression coefficient of -0.1139;
- PGM3, with a regression coefficient of -0.1658;
- PYGM, with a regression coefficient of -0.1122;
- RPE, with a regression coefficient of -0.1595;
- SDHA, with a regression coefficient of -0.2319.

There is one single most frequent regulator, selected as a relevant regulatory feature for 8 glucose metabolism target genes:

- ILK (regulatory gene of the GLUCOSE\_METABOLISM pathway) is selected as:
  - 1<sup>st</sup> (out of 4 features) for ALDOC with a regression coefficient of 0.2099;
  - 2<sup>nd</sup> (out of 8 features) for TALDO1 with a regression coefficient of 0.2258;
  - 3<sup>rd</sup> (out of 7 features) for DLST with a regression coefficient of 0.1962;
  - 5<sup>th</sup> (out of 7 features) for ACO2 with a regression coefficient of 0.1340;
  - 5<sup>th</sup> (out of 11 features) for ENO1 with a regression coefficient of 0.1271;
  - 5<sup>th</sup> (out of 9 features) for SUCLG1 with a regression coefficient of 0.1253;
  - 6<sup>th</sup> (out of 7 features) for DLAT with a regression coefficient of 0.1025;
  - 6<sup>th</sup> (out of 7 features) for RPIA with a regression coefficient of -0.1845.

Model M5: the expression of each target gene is regulated by genes in the GLUCOSE\_METABOLISM pathway, genes encoding for transcription factors binding to the target gene promoters or to the promoters of other genes in this pathway, genes in the DNA\_REPAIR and STEM\_CELLS pathways and their regulatory genes.

Table 7.6: Model M5 best GLUCOSE\_METABOLISM genes and their features.

| GENE        | Significant Feature | Adj.R2      | Regression Coefficient | Feature Description   |
|-------------|---------------------|-------------|------------------------|---|
| <b>SDHD</b> | DLAT                | <b>0.79</b> | 0.7345                 | Gene of the GLUCOSE_METABOLISM pathway                      |
|             | SUCLG1              |             | 0.1897                 | Gene of the GLUCOSE_METABOLISM pathway                      |
|             | HCFC1               |             | -0.0735                | Candidate regulatory gene of the model gene SDHD            |
|             | MERTK               |             | -0.0849                | Gene of the STEM_CELLS pathway                              |
|             | MAZ                 |             | -0.1266                | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
|             | TP53BP1             |             | -0.1406                | Gene of the DNA_REPAIR pathway                              |
|             | ACLY                |             | -0.1521                | Gene of the GLUCOSE_METABOLISM pathway                      |
| <b>TPI1</b> | ENO2                | <b>0.76</b> | 0.2825                 | Gene of the GLUCOSE_METABOLISM pathway                      |
|             | PHB2                |             | 0.2656                 | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
|             | ZNF384              |             | 0.2360                 | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
|             | PGM1                |             | 0.2176                 | Gene of the GLUCOSE_METABOLISM pathway                      |
|             | YBX3                |             | 0.1689                 | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
|             | PTTG1               |             | 0.1673                 | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
|             | ETFA                |             | 0.1334                 | Gene of the STEM_CELLS pathway                              |
|             | CREB3               |             | 0.1176                 | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
|             | FOXP1               |             | -0.0982                | Gene of the STEM_CELLS pathway                              |
|             | NR2C2               |             | -0.1001                | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
|             | ATM                 |             | -0.1768                | Gene of the STEM_CELLS pathway                              |
| <b>DLAT</b> | SDHD                | <b>0.75</b> | 0.7804                 | Gene of the GLUCOSE_METABOLISM pathway                      |
|             | ATM                 |             | 0.2195                 | Gene of the STEM_CELLS pathway                              |
|             | MAZ                 |             | 0.1710                 | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
|             | ACO2                |             | 0.1541                 | Gene of the GLUCOSE_METABOLISM pathway                      |
|             | MAML1               |             | 0.1416                 | Gene of the STEM_CELLS pathway                              |
|             | MERTK               |             | 0.1310                 | Gene of the STEM_CELLS pathway                              |
|             | ZBTB1               |             | -0.0636                | Candidate regulatory gene of the STEM_CELLS pathway         |
| <b>HK3</b>  | FBP1                | <b>0.71</b> | 0.5655                 | Gene of the GLUCOSE_METABOLISM pathway                      |
|             | PTPRC               |             | 0.4751                 | Gene of the STEM_CELLS pathway                              |
|             | MERTK               |             | 0.0898                 | Gene of the STEM_CELLS pathway                              |
|             | PLAT                |             | -0.0713                | Gene of the STEM_CELLS pathway                              |
|             | ALDH1A1             |             | -0.0812                | Gene of the STEM_CELLS pathway                              |
|             | MS4A1               |             | -0.0924                | Gene of the STEM_CELLS pathway                              |
|             | ZHX1                |             | -0.1459                | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
| <b>ACO2</b> | L3MBTL2             | <b>0.69</b> | 0.5854                 | Candidate regulatory gene of the model gene ACO2            |
|             | ESRRA               |             | 0.2703                 | Candidate regulatory gene of the model gene ACO2            |
|             | IDH3A               |             | 0.2294                 | Gene of the GLUCOSE_METABOLISM pathway                      |
|             | ZHX2                |             | 0.1470                 | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
|             | HNRNPUL1            |             | 0.1261                 | Candidate regulatory gene of the STEM_CELLS pathway         |
|             | ILK                 |             | 0.1245                 | Candidate regulatory gene of the GLUCOSE_METABOLISM pathway |
|             | SUCLG1              |             | 0.1078                 | Gene of the GLUCOSE_METABOLISM pathway                      |
|             | NOTCH2              |             | 0.0843                 | Gene of the STEM_CELLS pathway                              |
|             | CHEK1               |             | -0.1221                | Gene of the STEM_CELLS pathway                              |
|             | LATS1               |             | -0.2098                | Gene of the STEM_CELLS pathway                              |

The Adj. R<sup>2</sup> score is higher than 0.6 for 9 genes of the pathway (SDHD, TPI1, DLAT, HK3, ACO2, ACLY, PRPS1, PHKA1 and ALDOA) and it reaches a maximum value of around 0.79 for gene SDHD. Table 7.6 shows an excerpt of the whole set of relevant features in model M5 for the best genes in the GLUCOSE\_METABOLISM pathway. The only features discarded a priori are the ones corresponding to missing values in TCGA, i.e., the methylation of HK2, PDPR, SUCLG2 and the expression of the regulatory gene EMSY.

For 12 genes of the pathway the gene methylation is selected as one of the relevant features involved in the regulation of their expression (the same genes as in M3, with the exception of gene PDK4 and the addition of gene TKT):

- AGL, with a regression coefficient of -0.0896;
- ALDOC, with a regression coefficient of -0.3334;



- DLD, with a regression coefficient of -0.2559;
- IDH3B, with a regression coefficient of -0.2132;
- MDH2, with a regression coefficient of -0.1513;
- PCK1, with a regression coefficient of -0.3547;
- PDK3, with a regression coefficient of -0.2408;
- PGM3, with a regression coefficient of -0.1134;
- PYGM, with a regression coefficient of -0.1141;
- RPE, with a regression coefficient of -0.1718;
- SDHA, with a regression coefficient of -0.2191;
- TKT, with a regression coefficient of -0.0819.

Starting from the small set of genes in *Table 7.6*, it is clear the limited effect of the DNA\_REPAIR pathway in the regulation of glucose metabolism involved genes and the higher interrelationship between GLUCOSE\_METABOLISM and STEM\_CELLS pathways, with the latter pathway having a key role in the regulation systems of the former one.

As reported in the networks at the end of the paragraph, these evaluation can be generalized on the whole set of genes in the GLUCOSE\_METABOLISM pathway.

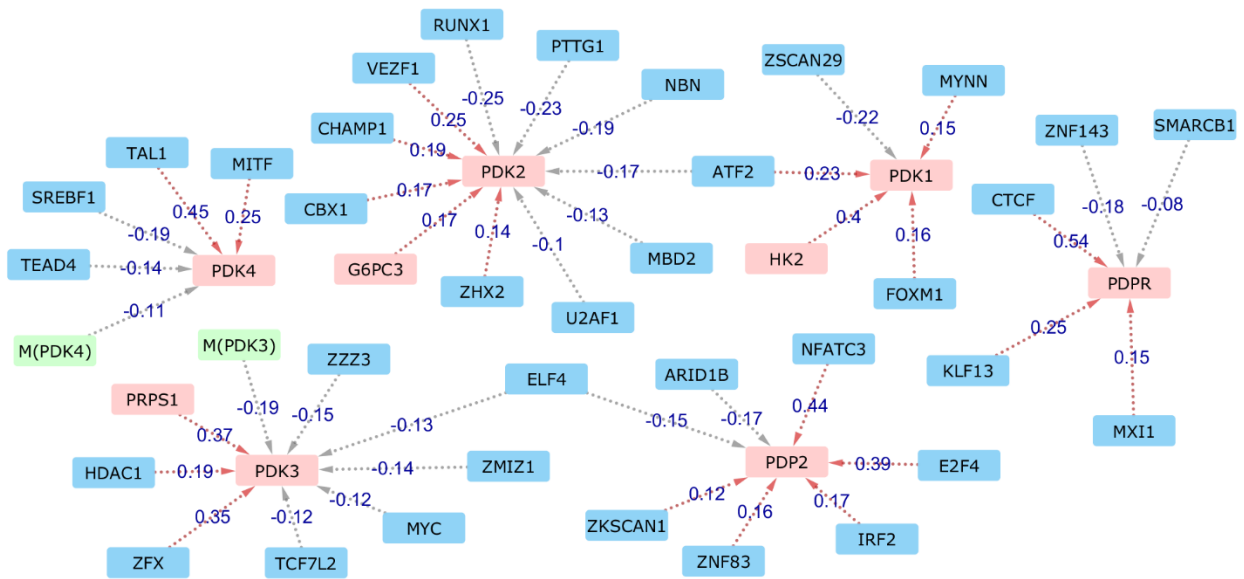
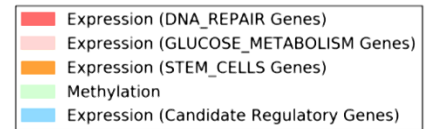
There is one single most frequent regulator, selected as a relevant regulatory feature for 10 glucose metabolism target genes:

- TSC22D4 (regulatory gene of the STEM\_CELLS pathway) is selected as:
  - 1<sup>st</sup> (out of 5 features) for PCK1 with a regression coefficient of 0.1532;
  - 4<sup>th</sup> (out of 9 features) for MDH2 with a regression coefficient of 0.1929;
  - 5<sup>th</sup> (out of 8 features) for PFKL with a regression coefficient of 0.2117;
  - 6<sup>th</sup> (out of 12 features) for BPGM with a regression coefficient of 0.1340;
  - 6<sup>th</sup> (out of 10 features) for H6PD with a regression coefficient of -0.1458;
  - 7<sup>th</sup> (out of 13 features) for CS with a regression coefficient of -0.0804;
  - 8<sup>th</sup> (out of 12 features) for SUCLG2 with a regression coefficient of 0.1094;
  - 9<sup>th</sup> (out of 9 features) for AGL with a regression coefficient of -0.1076;
  - 12<sup>th</sup> (out of 12 features) for ENO with a regression coefficient of -0.1792;
  - 13<sup>th</sup> (out of 13 features) for PGM3 with a regression coefficient of -0.2090;

Finally, all the results are graphically represented through expression networks in Cytoscape. An example for this pathway is reported in *Figure 7.3*, showing model M3 (a) and model M5 (b) networks for the genes of the *Regulation of Glucose Metabolism* subclass: it describes the regulation systems of a subset of 6 genes from the GLUCOSE\_METABOLISM pathway (PDK1, PDK2, PDK3, PDK4, PDP2 and PDPR), showing how the regulation of their expression changes from considering only target and regulatory genes of the GLUCOSE\_METABOLISM pathway, to including also activity related to genes of the DNA\_REPAIR and STEM\_CELLS pathways.



a) GLUCOSE\_METABOLISM (M3) – Regulation of Glucose Metabolism



b) GLUCOSE\_METABOLISM (M5) – Regulation of Glucose Metabolism

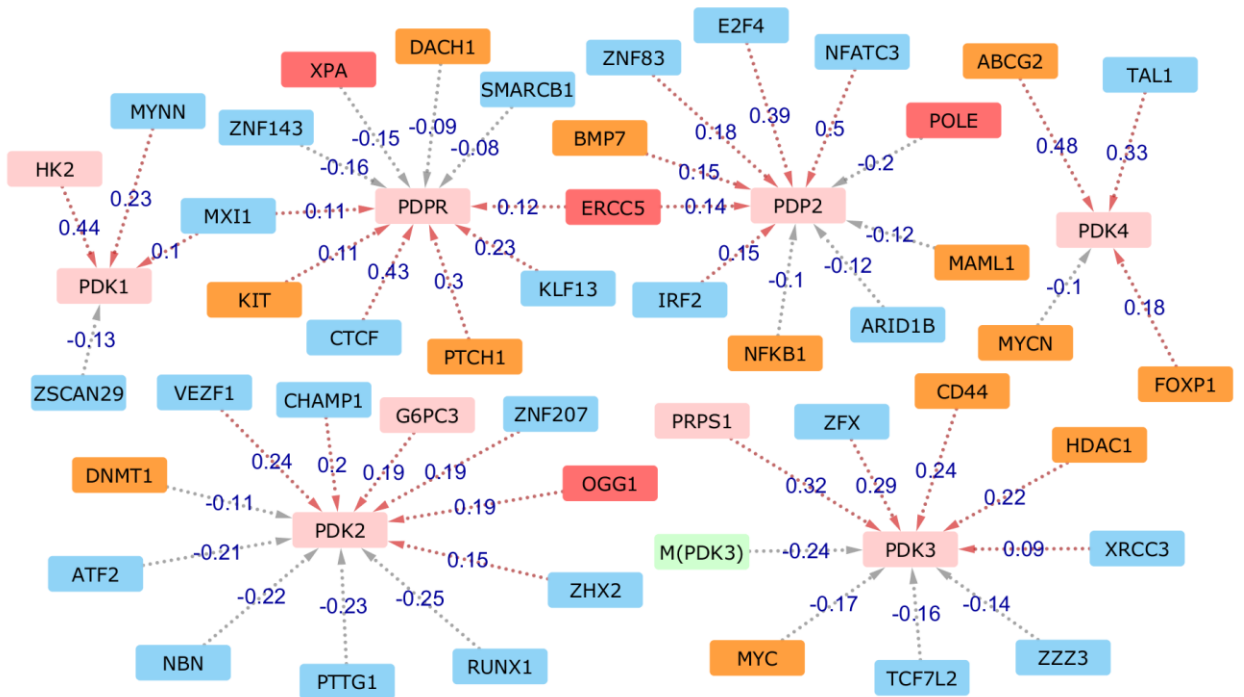


Figure 7.3: Expression networks from linear regression models M3 (a) and M5 (b) of the Regulation of Glucose Metabolisms subclass.

Table 7.7: Number of relevant features selected for each genetic pathway: target genes in the pathways (a) and regulatory genes (b).

a)

|                    | Features (Genes in the genetic pathways of interest) |                            |                                    |
|--------------------|--|----------------------------|------------------------------------|
|                    | DNA_REPAIR<br>Target Genes                           | STEM_CELLS<br>Target Genes | GLUCOSE_METABOLISM<br>Target Genes |
| DNA_REPAIR         | 13   | 29                         | -                                  |
| STEM_CELLS         | 16   | 37                         | -                                  |
| GLUCOSE_METABOLISM | 18   | 55                         | 61                                 |

b)

|                    | Features (Regulatory Genes encoding TFs binding to target genes promoters) |                                |  |
|--------------------|--|--------------------------------|--|
|                    | DNA_REPAIR<br>Regulatory Genes   | STEM_CELLS<br>Regulatory Genes | GLUCOSE_METABOLISM<br>Regulatory Genes |
| DNA_REPAIR         | 60   | 14                             | -                                      |
| STEM_CELLS         | 3  | 154                            | -                                      |
| GLUCOSE_METABOLISM | 0  | 9                              | 157                                    |

Since the main purpose of this thesis is inferring the regulatory-based relationships among the 177 ovarian cancer-related target genes and the set of their 249 distinct regulatory genes extracted, starting from the most comprehensive model M5 Table 7.7 summarizes the number of distinct features identified as relevant for genes in each pathway: this helps assessing the impact of each genetic pathway on the others.

We recall that our computational approach defined oriented target/features relationships: this means that it is possible that gene  $G1$  is a regulator of gene  $G2$ , but  $G2$  is not a relevant feature for  $G1$ . As a consequence pathway  $P1$  may have a relevant impact in the regulation of pathway  $P2$ , but the opposite is not necessarily true.

## 7.2 Validation

Validating the models we generated and their related results is a key point for our analysis. The outcomes of the regression must be valid from both a computational and a biological standpoint.

### 7.2.1 Computational validation

Multiple computational methods exist for computing expression correlations among sets of genes and for computationally inferring mutual functional relationships, but no gold standards are defined. The most effective and relevant among these methods is ARACNe.

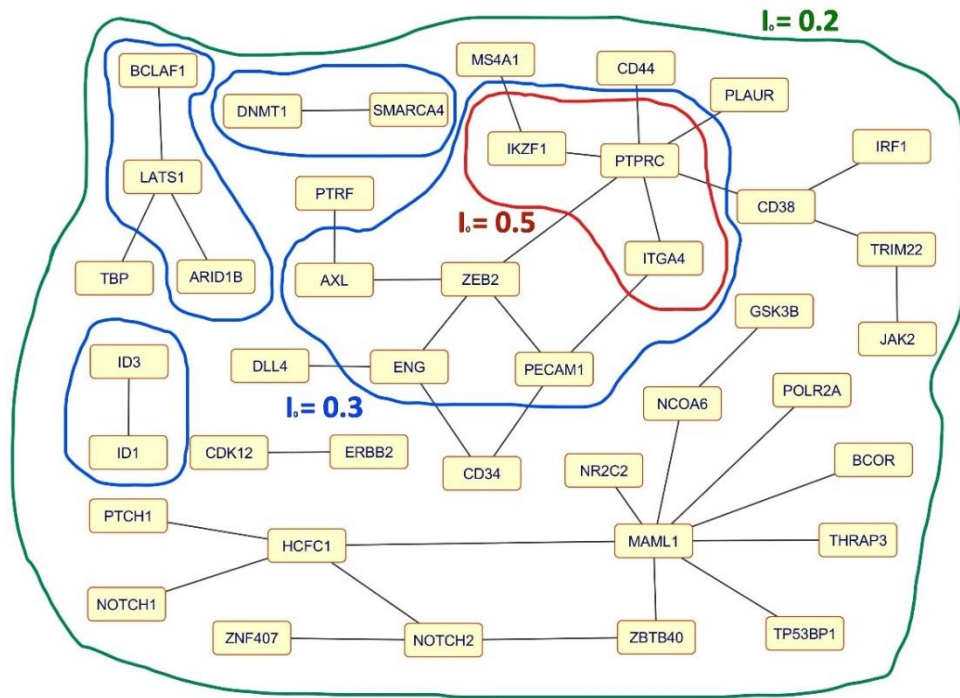


Figure 7.4: Example of different mutual information thresholds in ARACNe.

ARACNe is an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context: starting from DNA microarray data, it generates gene expression networks by considering triplets of genes and removing the weakest of the three relationships at each iteration, according to its assigned *Mutual Information* (MI) value and to an arbitrary threshold ( $I_0$ ) that is set a priori by the user (i.e., only relationships between genes quantified by a MI value higher than  $I_0$  are considered and, for each triplet, the edge with the lowest MI value is removed). The threshold allows to define the dimension of the generated network: the higher is  $I_0$ , the smaller are the network and the number of displayed relationships. An example of how the dimension of the output network changes in relation to different values of the threshold is reported in Figure 7.4: nodes represent the human genes of interest, while edges identify their bidirectional correlations retrieved by ARACNe.

ARACNe works under a specific set of initial approximations that, although they are proved to be reasonable, may lead the algorithm to fail. In general, this is a computational method and it does not guarantee reliable results, so we may fall into wrong or incomplete relationships.

We use ARACNe for validating the associations and correlations found during the data analysis, by comparing our ovarian cancer regression models (i.e., M3 and M5) with corresponding networks generated by the ARACNe algorithm.

Depending on how this comparison is made and due to the fact that the linear regression algorithm is a different computational procedure with respect to the ARACNe algorithm, differences are expected: specifically, according to our computational approach, the computed regression models usually detect less correlations than ARACNe, although some of them may not be found by ARACNe. In addition, ARACNe is made for building only gene expression networks without taking methylation into account, as allowed by our OV regression models, instead.

In the end, as a result of this comparison, for most target genes a set of common features between the two approaches is expected, i.e., a set of the main relationships revealed by the OV regression models which are also extracted by the ARACNe algorithm: this allows to verify that the initial plan and its implementation through a combined feature selection / linear regression approach leads to meaningful results that are worth investigating through a further biological experimental analysis.

We also adopt a second computational validation approach: since a validation on another set of ovarian cancer samples cannot be performed, because no other samples are available, results are validated according to another biological model, as similar as possible to the OV tumor. Specifically, the focus is on a biomolecularly equivalent tumor, i.e., the basal-like breast cancer.

Validation on breast cancer follows two different paths: on one hand, the OV tumor regression models are applied on this other tumor data, in order to evaluate if the same exact set of relationships may reasonably be valid also for this specific sub-type of breast cancer; on the other hand, considering each single gene independently, the objective is verifying if a similar set of relevant features may be found also in the basal breast cancer data, by re-running the same analysis procedure defined for the ovarian cancer (the full linear regression process is executed on this other tumor data and the corresponding set of BRCA models is generated, in order to assess if this approach is valid and the results are related to what obtained for ovarian cancer).

Basal-like breast cancer mainly compares with the ovarian cancer, because they are both considerably affected by DNA damages. In general, common relevant features are expected for the same genes in both models, though differences may be there, due to peculiar properties typical of each tumor. In particular, since damages to the DNA have a key role in both tumors, a better result is expected from comparing OV models and BRCA models of genes involved in the DNA\_REPAIR pathway.

### 7.2.1.1 Comparison with the ARACNe processing

Using ARACNe through the Cyni Toolbox in Cytoscape, we process information contained in both M3 and M5 models, generating two corresponding networks.

As already done for the OV linear regression procedure, this processing is performed for each pathway, i.e., two networks, respectively corresponding to M3 and M5, are generated using ARACNe and then used for the comparison: the former network takes the genes of the pathway and the complete set of their candidate regulators as input, along with their expression values in the different ovarian cancer TCGA samples, to allow a complete processing using the list of genes in the pathway as hubs of the network; the latter one adds expression information about genes of other pathways and their candidate regulators, to allow a complete processing by setting the list of genes in the considered pathway as nodes and the list of their candidate regulatory genes as transcription factors.

The whole process is executed by setting the lowest possible Mutual Information threshold, i.e.,  $I_0 = 0.001$ , in order to obtain the highest number of correlations. Generated networks are then compared with models M3 and M5 of the corresponding pathway. *Figure 7.5* summarizes the set operations implemented for validating our regression models through ARACNe.

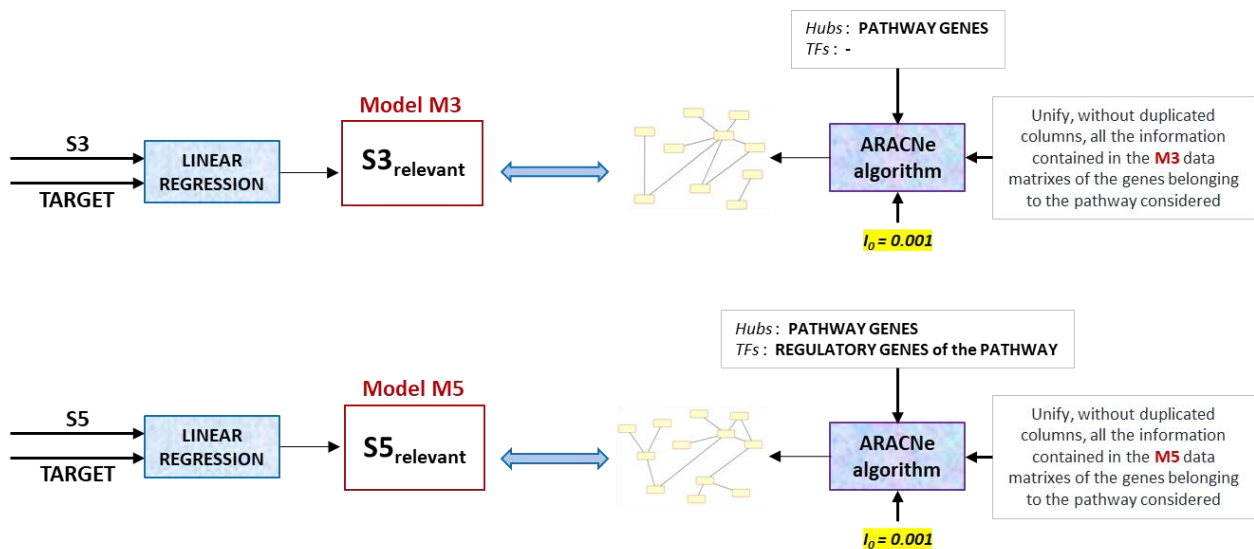


Figure 7.5: Validation of ovarian cancer regression models based on the ARACNe processing.

Table 7.8: Validation of ovarian cancer regression models based on the ARACNe processing: comparison results.

|                    | Total N° Genes in the pathway | M3  |   | M5  |   |
|--------------------|-------------------------------|---|---|---|---|
|                    |                               | % of genes with common features in OV Models and ARACNe | Mean % OV Models features found by ARACNe | % of genes with common features in OV Models and ARACNe | Mean % OV Models features found by ARACNe |
| DNA_REPAIR         | 20                            | 90.0%   | 37.4%                                     | 95.0%   | 35.4%                                     |
| GLUCOSE_METABOLISM | 84                            | 83.3%   | 27.4%                                     | 84.5%   | 27.8%                                     |
| STEM_CELLS         | 73                            | 72.6%   | 26.5%                                     | 73.9%   | 26.4%                                     |

The results of the comparison show that for a high percentage of target genes common correlations in the two approaches are found and an average of 30% of the whole set of relevant features identified by the OV regression models is verified also by ARACNe, as reported in *Table 7.8*.

The last step of the validation consists in evaluating the ranking of the common features. In particular, for each gene, relevant features found using both procedures are sorted in ascending order according to their regression coefficients in the OV models and their mutual information value in the ARACNe networks. The two rankings are compared: finding the same ranking is an additional validation of the regression results, indicating that these features and their relevance on the expression of the model gene are very likely correct.

60% of the genes having common features according to the two approaches, also have the same feature ranking, while the other 40% are characterized by very slightly different rankings, where some pairs of feature are swapped in their order.

*Table 7.9* shows an example of the detailed M5 comparison for some genes, performed by matching relevant features found in the OV models and in the ARACNe network.

Table 7.9: Validation of ovarian cancer regression models based on the ARACNe processing: detailed results for a sample set of genes.

| N° ARACNe Features  | N° OV Models Features | Common Features | Mutual Information | Regression Coefficient | ARACNe Ranking | OV Models Ranking | % OV Models features found by ARACNE |
|---------------------|-----------------------|-----------------|--------------------|------------------------|----------------|-------------------|--------------------------------------|
| <b>Gene FANCA</b>   |                       |                 |                    |                        |                |                   |                                      |
| 19                  | 6                     | POLE            | 0.2337             | 0.3805                 | 1              | 1                 | 66.7%                                |
|                     |                       | E2F4            | 0.1112             | 0.2595                 | 3              | 2                 |                                      |
|                     |                       | MYBL2           | 0.1796             | 0.218                  | 2              | 3                 |                                      |
|                     |                       | ZBTB1           | 0.0425             | 0.0841                 | 4              | 4                 |                                      |
| <b>Gene TP53BP1</b> |                       |                 |                    |                        |                |                   |                                      |
| 94                  | 3                     | ZSCAN29         | 0.5494             | 0.4646                 | 1              | 1                 | 100%                                 |
|                     |                       | ZBTB40          | 0.3848             | 0.3737                 | 2              | 2                 |                                      |
|                     |                       | MAML1           | 0.2499             | 0.1917                 | 3              | 3                 |                                      |
| <b>Gene CHEK1</b>   |                       |                 |                    |                        |                |                   |                                      |
| 46                  | 12                    | NFRKB           | 0.1022             | 0.4343                 | 3              | 1                 | 50%                                  |
|                     |                       | RAD51           | 0.15               | 0.2605                 | 1              | 2                 |                                      |
|                     |                       | FANCA           | 0.1048             | 0.2058                 | 2              | 3                 |                                      |
|                     |                       | CBX5            | 0.0937             | 0.129                  | 4              | 4                 |                                      |
|                     |                       | SRSF1           | 0.0825             | 0.129                  | 5              | 5                 |                                      |
|                     |                       | RB1             | 0.0132             | -0.168                 | 6              | 6                 |                                      |
| <b>Gene PTPRC</b>   |                       |                 |                    |                        |                |                   |                                      |
| 13                  | 7                     | IKZF1           | 0.7776             | 0.5857                 | 1              | 1                 | 57.1%                                |
|                     |                       | ITGA4           | 0.5026             | 0.2392                 | 2              | 2                 |                                      |
|                     |                       | CD38            | 0.254              | 0.1394                 | 4              | 3                 |                                      |
|                     |                       | CD44            | 0.2987             | 0.1217                 | 3              | 4                 |                                      |
| <b>Gene DLAT</b>    |                       |                 |                    |                        |                |                   |                                      |
| 13                  | 7                     | SDHD            | 0.3366             | 0.7804                 | 1              | 1                 | 28.6%                                |
|                     |                       | ATM             | 0.12               | 0.2195                 | 2              | 2                 |                                      |

### 7.2.1.2 Models application on basal-like breast cancer data

The first validation approach using breast cancer data consists in a simple computational procedure able to verify if the same OV models are exploitable on a different set of data that may share properties with the initial set of observations.

Despite the smaller dimensions of this set (122 samples) with respect to the available ovarian cancer samples, some key correlations among the genes of interest are expected to be valid also for basal-like breast cancer. Thus, for each target gene, the set of its relevant features, along with their associated regression coefficients, is selected from the OV model and used on breast cancer methylation and expression data (*Chapter 5*, paragraph 5.4.2) to compute the estimated value of the target gene expression in each breast cancer TCGA aliquot, according to the definition of linear regression:

$$EXPR_G = C_1 V_{f1} + C_2 V_{f2} + \dots + C_n V_{fn}$$

where:

$V_{fi}$  : expression or methylation values in breast cancer data sample corresponding to relevant features extracted in the considered OV regression model for gene  $G$

$C_i$  : regression coefficients assigned to the  $n$  relevant features extracted in the considered OV regression model for gene  $G$

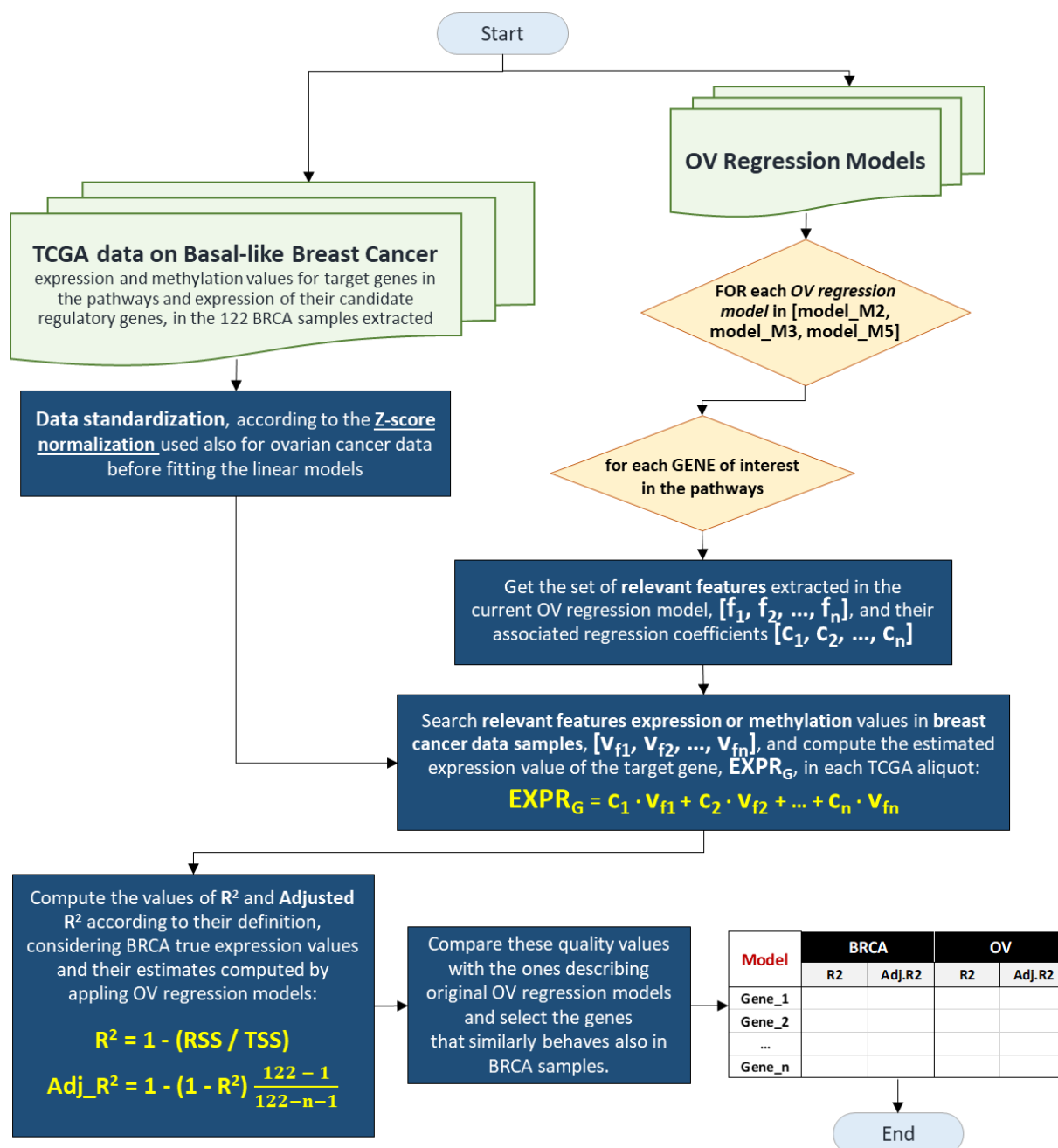


Figure 7.6: Ovarian cancer regression models application on basal-like breast cancer data: complete workflow.

Finally, the values of  $R^2$  and Adjusted  $R^2$  are manually computed according to their formulas and compared with the ones from the original OV models. Figure 7.6 shows all the steps of the applied validation procedure, as implemented in a parametric Python script.

From comparing Adjusted  $R^2$  values, we realize that OV models can be applied to BRCA data, obtaining reasonable results for some target genes and inferring a small set of key genes which appear to be similarly regulated in the two tumors (some even show a better performance in BRCA rather than OV), as reported in Table 7.10.



Table 7.10: Ovarian cancer regression models application on basal-like breast cancer data: excerpt of the results comparison for model M5.

| Model<br>M5 | BRCA |        | OV   |        | PATHWAY            |
|-------------|------|--------|------|--------|--------------------|
|             | R2   | Adj.R2 | R2   | Adj.R2 |                    |
| PTPRC       | 0.90 | 0.89   | 0.83 | 0.82   | STEM_CELLS         |
| IKBKB       | 0.58 | 0.55   | 0.50 | 0.49   | STEM_CELLS         |
| TP53BP1     | 0.71 | 0.69   | 0.66 | 0.65   | DNA_REPAIR         |
| GSK3A       | 0.54 | 0.51   | 0.52 | 0.51   | GLUCOSE_METABOLISM |
| FBP1        | 0.48 | 0.47   | 0.50 | 0.50   | GLUCOSE_METABOLISM |
| DNMT1       | 0.69 | 0.67   | 0.71 | 0.70   | STEM_CELLS         |
| RPIA        | 0.31 | 0.26   | 0.35 | 0.33   | GLUCOSE_METABOLISM |
| GSK3B       | 0.46 | 0.41   | 0.50 | 0.49   | GLUCOSE_METABOLISM |
| NFKB1       | 0.54 | 0.49   | 0.59 | 0.58   | STEM_CELLS         |
| PECAM1      | 0.55 | 0.53   | 0.64 | 0.64   | STEM_CELLS         |
| BRCA1       | 0.51 | 0.48   | 0.60 | 0.59   | DNA_REPAIR         |
| JAK2        | 0.61 | 0.56   | 0.69 | 0.68   | STEM_CELLS         |
| LATS1       | 0.62 | 0.58   | 0.71 | 0.70   | STEM_CELLS         |
| MLH1        | 0.28 | 0.22   | 0.36 | 0.35   | DNA_REPAIR         |
| RPE         | 0.42 | 0.35   | 0.50 | 0.48   | GLUCOSE_METABOLISM |
| PALB2       | 0.39 | 0.34   | 0.49 | 0.47   | DNA_REPAIR         |
| ITGA4       | 0.67 | 0.63   | 0.77 | 0.76   | STEM_CELLS         |
| TPI1        | 0.65 | 0.61   | 0.76 | 0.76   | GLUCOSE_METABOLISM |
| GSK3B       | 0.48 | 0.42   | 0.60 | 0.59   | STEM_CELLS         |
| SDHD        | 0.61 | 0.59   | 0.79 | 0.79   | GLUCOSE_METABOLISM |
| IDH3A       | 0.39 | 0.34   | 0.56 | 0.55   | GLUCOSE_METABOLISM |
| HK3         | 0.51 | 0.47   | 0.71 | 0.71   | GLUCOSE_METABOLISM |
| DLAT        | 0.55 | 0.51   | 0.76 | 0.75   | GLUCOSE_METABOLISM |
| POLB        | 0.49 | 0.44   | 0.69 | 0.68   | DNA_REPAIR         |
| PGK1        | 0.39 | 0.32   | 0.60 | 0.58   | GLUCOSE_METABOLISM |

Given the proved genetic similarity between these two types of cancer, but also their own peculiarities, the overall results of this computational step contribute to the positive validation of our OV regression models.

### 7.2.1.3 Linear regression of individual genes on breast cancer samples

The last computational validation is still performed on basal-like breast cancer data, but it involves a complete re-computation of the models. The same procedure described in this document is repeated considering the 122 TCGA aliquots that identify patients with basal-like breast cancers and the regression models M2, M3 and M5 are re-computed according to breast cancer data values.

In general, lower quality results are expected, due to the limited number of available samples, which is about a third of those used for the ovarian cancer regression. BRCA models and OV models are finally compared by analyzing for each gene of interest the set of common relevant features and their regression coefficients in the two models, both for M3 and M5.



Table 7.11: Ovarian - Breast cancer regression models comparison results.

|                    | M3                            |   |  |  | M5  |  |  |
|--------------------|-------------------------------|---|--|--|---|--|--|
|                    | Total N° Genes in the pathway | % of genes with common features in OV and BRCA Models | Mean % OV features also found in BRCA Models | Mean % BRCA features also found in OV Models | % of genes with common features in OV and BRCA Models | Mean % OV features also found in BRCA Models | Mean % BRCA features also found in OV Models |
| DNA_REPAIR         | 20                            | 60.0%   | 10.0%  | 38.0%  | 55.0%   | 8.7%   | 33.8%  |
| STEM_CELLS         | 84                            | 17.8%   | 2.7%   | 16.0%  | 23.3%   | 4.7%   | 19.8%  |
| GLUCOSE_METABOLISM | 73                            | 29.8%   | 4.6%   | 23.4%  | 34.5%   | 4.1%   | 20.3%  |

Table 7.12: Ovarian - Breast cancer regression models comparison: detailed results for a sample set of genes.

| N° OV Models Features | N° BRCA Models Features | Common Features     | OV Coefficient | BRCA Coefficient |
|-----------------------|-------------------------|---------------------|----------------|------------------|
| <b>Gene BRCA1</b>     |                         |                     |                |                  |
| 6                     | 3                       | SUZ12               | 0.2317         | 0.5717           |
|                       |                         | METHYLATION (BRCA1) | -0.3475        | -0.3685          |
| <b>Gene OGG1</b>      |                         |                     |                |                  |
| 11                    | 4                       | FANCD2              | 0.4351         | 0.4574           |
|                       |                         | MEF2A               | -0.1581        | 0.1688           |
| <b>Gene ERCC2</b>     |                         |                     |                |                  |
| 13                    | 4                       | ERCC1               | 0.5332         | 0.4886           |
| <b>Gene DNMT1</b>     |                         |                     |                |                  |
| 8                     | 4                       | SMARCA4             | 0.5505         | 0.5592           |
|                       |                         | POLE                | 0.1738         | 0.2994           |
| <b>Gene DLAT</b>      |                         |                     |                |                  |
| 13                    | 7                       | SDHD                | 0.3366         | 0.7804           |
|                       |                         | ATM                 | 0.12           | 0.2195           |
| <b>Gene PTPRC</b>     |                         |                     |                |                  |
| 7                     | 1                       | IKZF1               | 0.5857         | 0.9495           |
| <b>Gene NFKB1</b>     |                         |                     |                |                  |
| 11                    | 3                       | RELA                | 0.2379         | 0.2773           |
|                       |                         | IRF2                | 0.2096         | 0.4674           |

What turns out is extremely interesting and confirms the expectations: most relevant features in the OV models are found and most common outcomes involve genes in the DNA\_REPAIR pathway. Biologists consider this to be a positive result, mainly because of the DNA-damages-related similarities of these two types of tumor.

The regression coefficients assigned to common features in the two models are interesting too: in most cases, they almost have the same value, indicating the impact of that feature on the expression of the target gene is similar in both tissues, confirming tumors similarity and the opportunity to extend ovarian cancer results also to basal-like breast cancer.

Instead, different regression coefficients for the same feature in the two models are potentially associated to tumor-related peculiarities. Comparison details are reported in *Table 7.11* and *Table 7.12*.

## 7.2.2 Biological validation

The computational work of this project must be supported by a biological validation of the ovarian cancer models, through experimental analyses in laboratory, expected to be carried out as a natural follow-up. Nonetheless, some specific experiments conducted by the *DNA Repair Unit* of the *Molecular Pharmacology* laboratory at “Mario Negri” Institute have already confirmed some regression results.

In addition, biological and genomic literature helps validating the data analysis, by proving a significant set of relationships revealed by the regression models.

### 7.2.2.1 PCR data

The DNA Repair Unit of the Molecular Pharmacology laboratory at “Mario Negri” Institute have measured the expression of all genes in the DNA\_REPAIR pathway and their correlations, starting from ovarian *Patient Derived Xenografts* (PDX) and using the *Polymerase Chain Reaction* (PCR) technique [60]: portions of tumor tissues taken from different patients are implanted and left growing into immunodeficient mice. Measurements taken from these PDXs using the PCR technique show an important set of correlation between genes that are also present in our OV regression models.

*Table 7.13* shows which correlations of our models are validated by the PDXs experiment: some significant correlations in the PDXs (in orange) are present also in the M2 regression model of the corresponding target gene in the DNA\_REPAIR pathway (in green), which is the model mainly taking genes of the same pathway into consideration, as potential relevant features.

We do not expect a complete matching between the two approaches, since M2 model also comprises candidate regulatory genes for each target gene, while PDXs only investigates genes within the DNA\_REPAIR pathway.

This is a very consistent validation, because even if this PCR experiment is a totally different process than the regression approach adopted in this thesis, it still reflects most correlations highlighted for DNA\_REPAIR genes.

### 7.2.2.2 Literature-confirmed results

For reasons of clarity and brevity, we limit the discussion of relevant literature to the pathway of DNA repair genes.

As described in paragraph 7.2.2.1, our analysis on the expression of DNA repair genes in ovarian tumor samples reveals some associations that have been recently reported by the DNA Repair group at “Mario Negri”, that have evaluated the gene expression profile of the genes involved in the DNA repair pathways within a recently established biobank of Patients Derived Ovarian Xenografts.

The prominent role of BRCA1 methylation on the regulation of the corresponding gene that we have found in our analysis is well supported by several studies. In fact, reduced BRCA1 expression as a result of promoter methylation has been reported in 5%–30% epithelial ovarian cancers [61, 62, 63, 64].

A similar role of methylation is found in our study for ERCC1 gene. To the best of our knowledge, distinct ERCC1 DNA methylation profiles in ovarian tumors and subsequent silencing of the gene have

Table 7.13: Correlations among the expression of the genes studied in the ovarian PDXs and their comparison with ovarian M2 model.

|             |         | BER   |      |       |       | DSB     |       |       |       |       |       |       |       | NER    |       |       |       | CDK12 |       |       |       |
|-------------|---------|-------|------|-------|-------|---------|-------|-------|-------|-------|-------|-------|-------|--------|-------|-------|-------|-------|-------|-------|-------|
|             |         | MLH1  | OGG1 | PARP1 | POLB  | TP53BP1 | BRCA1 | RAD51 | PALB2 | POLQ  | POLH  | FANCA | FANCC | FANCD2 | FANCF | XPA   | ERCC2 |       | ERCC4 | ERCC5 | ERCC1 |
|             | MLH1    | 1.00  | 0.35 | 0.05  | -0.14 | -0.12   | -0.06 | 0.01  | -0.01 | 0.22  | 0.11  | 0.15  | 0.11  | 0.26   | 0.31  | 0.03  | 0.07  | 0.08  | -0.06 | -0.04 | 0.12  |
| B           | OGG1    | 0.35  | 1.00 | 0.41  | 0.18  | 0.16    | 0.00  | 0.41  | 0.41  | 0.52  | 0.49  | 0.45  | 0.65  | 0.69   | 0.53  | 0.24  | 0.18  | 0.47  | 0.24  | 0.12  | 0.25  |
|             | PARP1   | 0.05  | 0.41 | 1.00  | 0.15  | 0.32    | -0.06 | 0.25  | 0.44  | 0.27  | 0.44  | 0.33  | 0.18  | 0.25   | 0.05  | 0.01  | 0.02  | 0.29  | 0.23  | -0.04 | 0.29  |
| R           | POLB    | -0.14 | 0.18 | 0.15  | 1.00  | 0.30    | -0.10 | 0.04  | 0.23  | -0.02 | 0.02  | -0.11 | 0.24  | 0.21   | 0.02  | -0.15 | 0.56  | 0.13  | 0.05  | 0.18  | 0.43  |
| D<br>S<br>B | TP53BP1 | -0.12 | 0.16 | 0.32  | 0.30  | 1.00    | 0.50  | 0.24  | 0.38  | 0.14  | 0.18  | 0.10  | 0.08  | 0.09   | 0.16  | -0.04 | 0.24  | 0.24  | 0.03  | 0.02  | 0.62  |
|             | BRCA1   | -0.06 | 0.00 | -0.06 | -0.10 | 0.50    | 1.00  | -0.10 | -0.03 | 0.03  | -0.06 | 0.01  | -0.11 | -0.24  | 0.32  | 0.12  | -0.11 | -0.07 | -0.12 | -0.30 | 0.27  |
|             | RAD51   | 0.01  | 0.41 | 0.25  | 0.04  | 0.24    | -0.10 | 1.00  | 0.65  | 0.55  | 0.22  | 0.25  | 0.44  | 0.72   | 0.03  | 0.31  | 0.01  | 0.72  | 0.03  | -0.11 | 0.33  |
|             | PALB2   | -0.01 | 0.41 | 0.44  | 0.23  | 0.38    | -0.03 | 0.65  | 1.00  | 0.41  | 0.28  | 0.13  | 0.16  | 0.48   | -0.08 | -0.06 | 0.09  | 0.79  | 0.21  | -0.05 | 0.45  |
|             | POLQ    | 0.22  | 0.52 | 0.27  | -0.02 | 0.14    | 0.03  | 0.55  | 0.41  | 1.00  | 0.13  | 0.42  | 0.37  | 0.71   | 0.45  | 0.34  | 0.38  | 0.46  | 0.28  | 0.24  | 0.17  |
|             | POLH    | 0.11  | 0.49 | 0.44  | 0.02  | 0.18    | -0.06 | 0.22  | 0.28  | 0.13  | 1.00  | 0.28  | 0.13  | 0.16   | 0.12  | -0.04 | -0.22 | 0.24  | -0.07 | -0.25 | -0.07 |
|             | FANCA   | 0.15  | 0.45 | 0.33  | -0.11 | 0.10    | 0.01  | 0.25  | 0.13  | 0.42  | 0.28  | 1.00  | 0.54  | 0.29   | 0.37  | 0.34  | -0.17 | 0.07  | -0.05 | -0.14 | -0.01 |
|             | FANCC   | 0.11  | 0.65 | 0.18  | 0.24  | 0.08    | -0.11 | 0.44  | 0.16  | 0.37  | 0.13  | 0.54  | 1.00  | 0.59   | 0.44  | 0.45  | 0.23  | 0.29  | 0.16  | 0.24  | 0.24  |
|             | FANCD2  | 0.26  | 0.69 | 0.25  | 0.21  | 0.09    | -0.24 | 0.72  | 0.48  | 0.71  | 0.16  | 0.29  | 0.59  | 1.00   | 0.26  | 0.28  | 0.36  | 0.65  | 0.04  | 0.15  | 0.26  |
|             | FANCF   | 0.31  | 0.53 | 0.05  | 0.02  | 0.16    | 0.32  | 0.03  | -0.08 | 0.45  | 0.12  | 0.37  | 0.44  | 0.26   | 1.00  | 0.21  | 0.29  | -0.04 | 0.06  | 0.25  | 0.19  |
| N<br>E<br>R | XPA     | 0.03  | 0.24 | 0.01  | -0.15 | -0.04   | 0.12  | 0.31  | -0.06 | 0.34  | -0.04 | 0.34  | 0.45  | 0.28   | 0.21  | 1.00  | -0.01 | 0.24  | 0.26  | -0.10 | -0.05 |
|             | ERCC2   | 0.07  | 0.18 | 0.02  | 0.56  | 0.24    | -0.11 | 0.01  | 0.09  | 0.38  | -0.22 | -0.17 | 0.23  | 0.36   | 0.29  | -0.01 | 1.00  | 0.07  | 0.33  | 0.71  | 0.37  |
|             | ERCC4   | 0.08  | 0.47 | 0.29  | 0.13  | 0.24    | -0.07 | 0.72  | 0.79  | 0.46  | 0.24  | 0.07  | 0.29  | 0.65   | -0.04 | 0.24  | 0.07  | 1.00  | 0.10  | -0.11 | 0.42  |
|             | ERCC5   | -0.06 | 0.24 | 0.23  | 0.03  | 0.03    | -0.12 | 0.03  | 0.21  | 0.28  | -0.07 | -0.05 | 0.16  | 0.04   | 0.06  | 0.26  | 0.33  | 0.10  | 1.00  | 0.61  | 0.03  |
|             | ERCC1   | -0.04 | 0.12 | -0.04 | 0.18  | 0.02    | -0.30 | -0.11 | -0.05 | 0.24  | -0.25 | -0.14 | 0.24  | 0.15   | 0.25  | -0.10 | 0.71  | -0.11 | 0.61  | 1.00  | 0.08  |
|             | CDK12   | 0.12  | 0.25 | 0.29  | 0.43  | 0.62    | 0.27  | 0.33  | 0.45  | 0.17  | -0.07 | -0.01 | 0.24  | 0.26   | 0.19  | -0.05 | 0.37  | 0.42  | 0.03  | 0.08  | 1.00  |

not been described yet. This observation, confirmed also in basal breast cancer samples, could be very interesting and could open the way to improve treatment strategies.

Another important relationship highlighted by our models is the one between ERCC1 and ERCC2: their gene expression is highly positively correlated and such association is maintained also after the introduction of different traits (both in M3 and M5). Moreover, the same holds true when basal breast cancer samples are used for the analysis. Both genes belong to the *Nucleotide Excision Repair* (NER) pathway, involved in the repair of UV-induced DNA damage [65]. In addition, the pathway has a key role in the repair of DNA adducts induced by cisplatin, the current golden standard treatment for ovarian cancer.

A similar correlation was observed when analyzing ERCC1 and ERCC2 gene expression in a number of triple negative breast cancer patients (but not in Luminal A breast cancers) [66], as reported in the graphs of Figure 7.7. This is interesting considering the fact that platinum-sensitive triple-negative breast cancers (TNBC) and serous ovarian cancers have been reported to carry extensive genomic rearrangements and allelic imbalance, suggesting that these cancers may share similar defects in the DNA repair [67]. This observation deserves further validation in different experimental models and at the protein level.

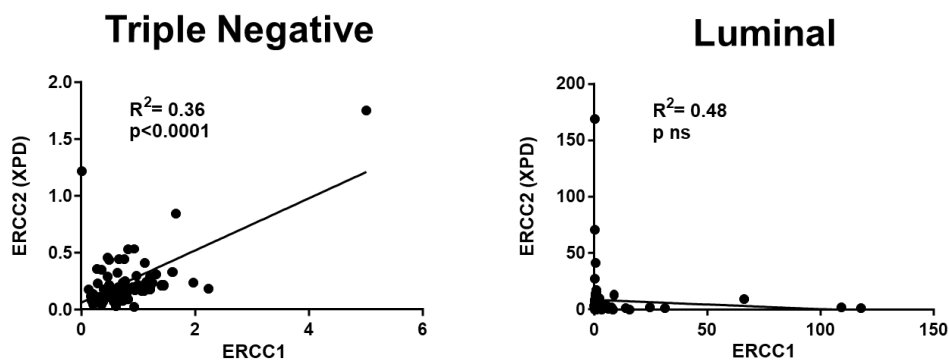


Figure 7.7: ERCC1 – ERCC2 correlation in Triple Negative breast cancer and in the breast cancer Luminal subtype.



## 8. Conclusions

*« Progress looks like a bunch of failures. And you can have feeling about that because it is sad, but you can't fall apart.*

*And then one day we will succeed and we will save a person's life and we will walk on the moon.*

*Figuratively, anyway. »*

Meredith

With this work we developed a computational and statistical analysis of the regulation systems of 177 target genes involved in specific biological functions which are strictly related to ovarian cancer development progression.

Ovarian cancer prognosis and therapy are nowadays still very poor and tumor-related knowledge limited. The key role in the process of acquisition of tumor-related properties is mainly played by a misregulation of gene expression that leads to significant changes in the activity of ovarian cancer relevant target genes. So, it is fundamental to understand how ovarian cancer gene expression is regulated, in order to improve knowledge and hopefully cancer therapies.

We focused only on a restricted set of regulatory elements: there are several other factors that could affect gene expression that we did not account for in the present study. Our results are limited by design to the best-predicting sets of features, leaving out potential regulators with important biological functions, but with a lower predicting power.

In the first part of the project, the sets of heterogeneous data needed for the analysis were extracted from the main biological and genomic data sources (i.e., ENCODE and TCGA) using the GMQL engine, focusing on a set of 372 ovarian cancer patients. Data on transcription factors and on the expression of their encoding genes, along with expression and methylation values associated with target genes, were then arranged, for each target gene, in multiple data matrices with a fixed number of rows (i.e., the patients data samples) and a variable number of columns (i.e., the set of potential features affecting the model gene expression), gradually increasing according to pre-defined rules.

In the second part of the project, we used these sets of features as inputs for building three predictive models for each target gene: a preliminary step of feature selection was followed by the application of the linear regression algorithm for inferring most relevant features, either up-regulating or down-regulating the expression of the model gene.

The analysis of the 531 built models allowed to explain the relations between each gene and its related biological processes and the interconnections between the different genomic pathways, as well as to evaluate the relevant regulatory elements at the single gene level: this allowed to identify already known regulators or genes correlations (e.g., hypermethylation of BRCA1) and to unveil a set of still unknown and potentially extremely interesting biological relationships (e.g., hypermethylation of ERCC1, the correlation between ERCC1 and ERCC2 or the correlation between CHEK1 and DNMT1), as the basis for an experimental follow-up.



## 9. Future developments

*« In the practice of medicine, change is inevitable. New surgical techniques are created.*

*Procedure are updated. Levels of expertise increase. Innovation is everything.*

*Nothing remains the same for long. We either adapt to change or we get left behind. »*

Meredith

This thesis is only the computational part of a wider project that has main biological future developments.

The most significant and newsworthy development involves laboratory work: conducting a deep experimental analysis to apply the present results on ovarian cancer to biological experiments, with the final goal of predicting the potential oncogenic role of target genes relevant regulators.

Our project allows to directly observe various correlations between the expression of multiple genes; however, since the implemented method is mainly computational, the existence of a specific correlation suggests that genes involved in this relationship may be functionally related from a biological standpoint, although it does not prove it.

The idea is trying to isolate the most interesting cases, mainly the ones showing a good fit in the model, and to verify whether the fact that one gene is correlated to another one means that it regulates this other gene; more precisely, biologists analyze the behavior of this latter gene when the expression of the former one is reduced or completely suppressed (i.e., the gene is turned off).

In the “Mario Negri” Institute’s laboratory all the tools for suppressing gene CHEK1 are already available: this is a great starting point, since this is one of the genes for which ovarian cancer regression models showed significant relationships according to the biologists.

A second in-depth study can be conducted on gene PTPRC of the STEM\_CELLS pathway, which showed a very interesting behavior while comparing the linear regression model results with the ARACNe processing, proving to have a correlation with a large set of genes in the whole human genome. Biologists believe this gene could be involved in the activation of cells of the immune system. Since TCGA provides mixed tissue samples, where both the tumor and its entire micro-environment are observed, this gene could be related to the immune system activation, highlighting the presence of *T cells* (i.e, lymphocytes that kill tumors) that are trying to kill the tumor. This is of great interest, because new antitumor therapies are immune therapies that block the immune system of cells T to enhance their activity, recognizing the tumor as an unwanted guest.

Finally, a third possible development consists in focusing on the metabolomics (i.e., the study of chemical processes involving metabolites, the small intermediate end products of metabolism), in order to correlate gene expression data of GLUCOSE\_METABOLISM pathway genes with the actual presence of metabolites.





# Bibliography and Webliography

- [1] Brown TA. Genomes. Department of Biomolecular Sciences, UMIST, Manchester, UK. 2<sup>nd</sup> Edition. Oxford: Wiley-Liss; 2002.
- [2] Masseroli M. Bioinformatics and Computational Biology, Lectures. [Online]  
Available from: <http://www.bioinformatics.deib.polimi.it/masseroli/BCB/>
- [3] Masseroli M. Genomic Computing, Lectures and Practices. [Online]  
Available from: [http://www.bioinformatics.deib.polimi.it/courses/PhD/genomic\\_computing/](http://www.bioinformatics.deib.polimi.it/courses/PhD/genomic_computing/)
- [4] Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cells. 4<sup>th</sup> Edition. New York: Garland Science; 2002. PartII: Basic Genetic Mechanisms. Chapter 4: DNA and Chromosomes; p.191-234.
- [5] Medical News Today. What Is DNA and How Does It Work? (only images). [Online]  
Available from: <https://www.medicalnewstoday.com/>
- [6] Clan Henderson Society. Introduction to DNA (only images). [Online]  
Available from: <http://www.clanhendersonsociety.org/intoduction-dna/>
- [7] Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis E. The Next-Generation Sequencing Revolution and its Impact on Genomics. Cell. 2013;155(1):27-38.
- [8] Guo J. Transcription: the epicenter of gene expression. Journal of Zhejiang University Science B. 2014; 15(5):409-411.
- [9] Venngage. DNA Transcription (only images). [Online]  
Available from: <https://infograph.venngage.com/p/223385/dna-transcription>
- [10] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;10(1):57-63.
- [11] Wikipedia. RNA-sequencing (only images). [Online]  
Available from: <https://en.wikipedia.org/wiki/RNA-Seq>
- [12] Latchman DS. Transcription factors: an overview. International Journal of Experimental Pathology. 1993;74:417-422.

- [13] Mundade R, Gulcin Ozer H, Wei H, Prabhu L, Lu T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic mark and beyond. *Cell Cycle*. 2014;13(18):2847-2852.
- [14] Stomp on Step 1. Epigenetics, Prader-Willi Syndrome, Angelman Syndrome (only images). [Online] Available from: <http://www.stomponstep1.com/epigenetics-prader-willi-syndrome-angelman-syndrome/>
- [15] Jin B, Li Y, Robertson KD. DNA Methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer*. 2001;2(6):607-617.
- [16] Chen Q, Zhu X, Li Y, Meng Z. Epigenetic regulation and cancer (review). *Oncology Reports*. 2013; 31(2):523-532.
- [17] Ricci F, Fratelli M, Guffanti F, Porcu L, Spriano F, Dell'Anna T, Fruscio R, Damia G. Patient-derived ovarian cancer xenografts re-growing after a cisplatin treatment are less responsive to a second drug re-challenge: a new experimental setting to study response to therapy. *Oncotarget*. 2017;8(5):7441-7451.
- [18] Pastò A, Pagotto A, Pilotto G, De Paoli A, De Salvo G, Baldoni A, Nicoletto M, Ricci F, Damia G, Bellio C, Indraccolo S, Amadori A. Resistance to glucose starvation as metabolic trait of platinum-resistant human epithelial ovarian cancer cells. *Oncotarget*. 2017;8(4):6433-6445.
- [19] Carrassa L, Damia G. DNA damage response inhibitors: Mechanisms and potential applications in cancer therapy. *Cancer Treat Reviews*. 2017;60:139-151.
- [20] Chilà R, Guffanti F, Damia G. Role and therapeutic potential of CDK12 in human cancers. *Cancer Treat Reviews*. 2016;50:83-88.
- [21] Cancer Genome Atlas Network. Study Reveals Genomic Similarities between Breast Cancer and Ovarian Cancer. *Nature*. 2012;490(7418):61-70.
- [22] The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*. 2008; 36(Database issue):D190-D195.
- [23] UniProt. [Online] Available from: <http://www.uniprot.org/>
- [24] The ENCODE Project Consortium. A user's guide to the Encyclopedia of DNA elements (ENCODE). *PLoS Biology*. 2011;9(4):e1001046.
- [25] ENCODE: Encyclopedia of DNA Elements. [Online] Available from: <https://www.encodeproject.org/>
- [26] HGNC database of human gene names | HUGO Gene Nomenclature Committee. [Online] Available from: <https://www.genenames.org/>

- [27] Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*. 2015;19(1A):A68-A77.
- [28] The Cancer Genome Atlas. [Online] Available from: <https://cancergenome.nih.gov/>
- [29] TCGA Wiki Home - National Cancer Institute - Confluence Wiki. [Online] Available from: <https://wiki.nci.nih.gov/display/TCGA/TCGA+Wiki+Home>
- [30] National Cancer Institute. GDC Docs. [Online] Available from: <https://docs.gdc.cancer.gov/>
- [31] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M et al. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Research*. 2012;22(9):1760-74.
- [32] Frankish A, Uszczyńska B, Ritchie G, Gonzalez J, Pervouchine D, Petryszak R et al. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics*. 2015;16(Suppl 8):S2.
- [33] McCarthy D, Humburg P, Kanapin A, Rivas M, Gaulton K, Cazier J et al. Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*. 2014;6(3):26.
- [34] GENCODE. Release 22 (GRCh38.p2). [Online] Available from: <https://www.encodegenes.org/releases/22.html>
- [35] OpenGDC. The Genomic Data Common Extraction Tool. OpenGDC File Format Definition. [Online]. Available from: [http://bioinf.iasi.cnr.it/opengdc/data/OpenGDC\\_format\\_definition.pdf](http://bioinf.iasi.cnr.it/opengdc/data/OpenGDC_format_definition.pdf)
- [36] GenoMetric Query Language [Online]. Available from: [http://www.bioinformatics.deib.polimi.it/genomic\\_computing/GMQL/](http://www.bioinformatics.deib.polimi.it/genomic_computing/GMQL/)
- [37] Masseroli M, Pinoli P, Venco F, Kaitoua A, Jalili V, Palluzzi F et al. GenoMetric Query Language: a novel approach to large-scale genomic data management. *Bioinformatics*. 2015;31(12):1881-1888.
- [38] Ceri S, Kaitoua A, Masseroli M, Pinoli P, Venco F. Data Management for Heterogeneous Genomic Datasets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2016;14(6):1251-1264.
- [39] Masseroli M, Kaitoua A, Pinoli P, Ceri S. Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. *Methods*. 2016;111:3-11.
- [40] Kaitoua A, Pinoli P, Bertoni M, Ceri S. Framework for Supporting Genomic Operations. *IEEE Transactions on Computers*. 2016;66(3):443-457.

- [41] Bertoni M, Ceri S, Kaitoua A, Pinoli P. Evaluating Cloud Frameworks on Genomic Applications. 2015 IEEE International Conference on Big Data. Santa Clara, CA, USA. 2015:193-202.
- [42] GMQL System. Documentation. GMQL Introduction to the language. [Online] Available from: [http://www.bioinformatics.deib.polimi.it/genomic\\_computing/GMQLsystem/doc/GMQL\\_introduction\\_to\\_the\\_language.pdf](http://www.bioinformatics.deib.polimi.it/genomic_computing/GMQLsystem/doc/GMQL_introduction_to_the_language.pdf)
- [43] GMQL. [Online] Available from: <http://gmql.eu/gmql-rest/>
- [44] Lutz M. Learning Python. 5<sup>th</sup> Edition. USA: O'Reilly Media; 2013.
- [45] DEIB-GECO/PyGMQL. Github. [Online] Available from: <https://github.com/DEIB-GECO/PyGMQL>
- [46] Welcome to PyGMQL's documentation! — PyGMQL 0.0.9 documentation. [Online] Available from: <http://pygmql.readthedocs.io/en/latest/>
- [47] Python. Pandas Library. [Online] Available from: <https://pandas.pydata.org/>
- [48] Python. Statsmodels Library. [Online] Available from: <http://www.statsmodels.org/stable/index.html>
- [49] Python. Scikit-learn Library: machine learning in Python. [Online] Available from: <http://scikit-learn.org/stable/index.html>
- [50] Python. Scikit-learn Library. Preprocessing data. [Online] Available from: <http://scikit-learn.org/stable/modules/preprocessing.html>
- [51] Raschka S. MLxtend: providing machine learning and data science utilities and extensions to python's scientific computing stack. The Journal of Open Source Software. 2018;3(24):638
- [52] Python. mlxtend. Feature Selection [Online]. Available from: [https://rasbt.github.io/mlxtend/api\\_subpackages/mlxtend.feature\\_selection/](https://rasbt.github.io/mlxtend/api_subpackages/mlxtend.feature_selection/)
- [53] Python. Networkx Library. [Online] Available from: <https://networkx.github.io/>
- [54] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research. 2003;13(11):2498-504.
- [55] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNe: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7(1):S7.

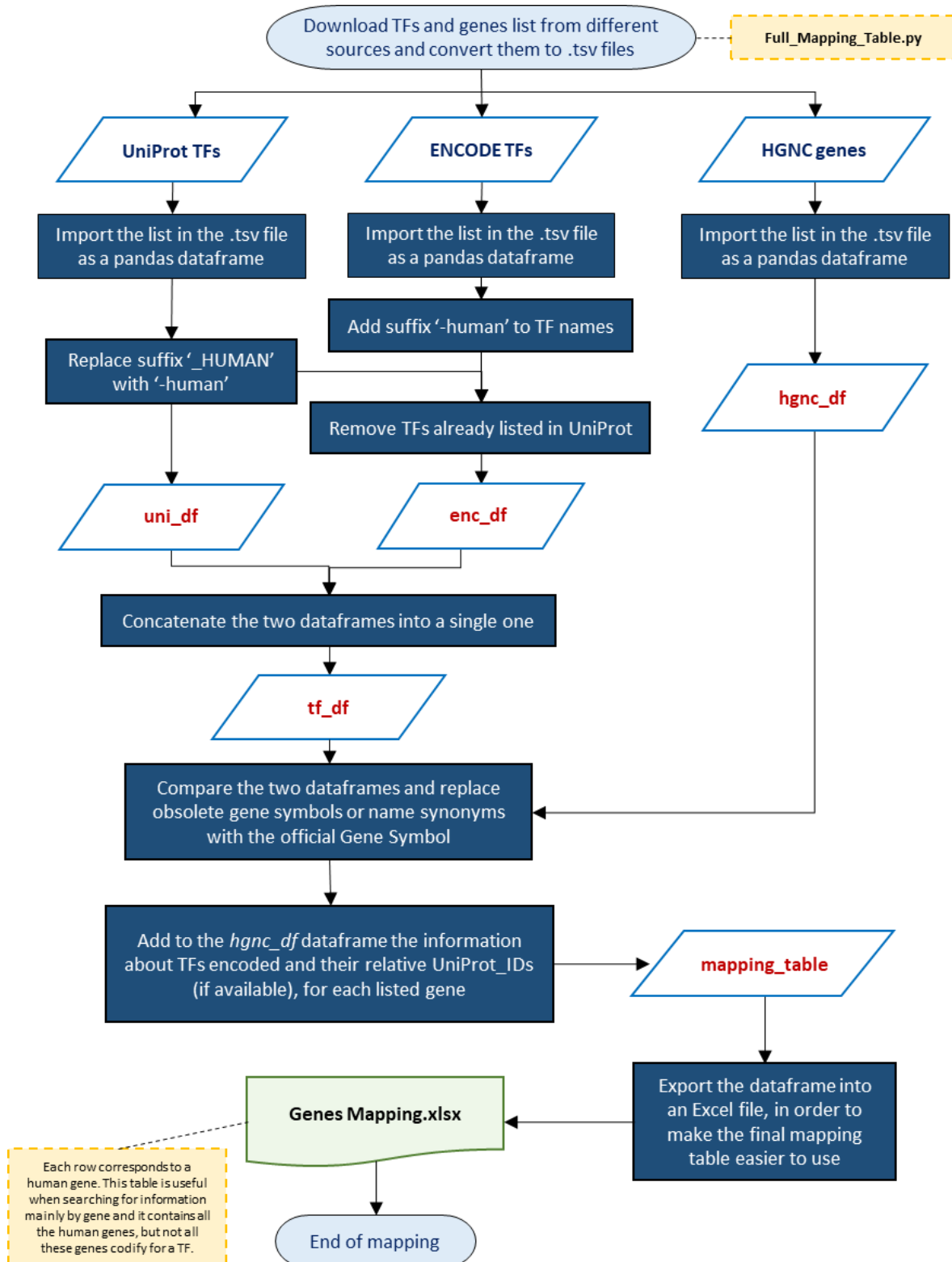
- [56] Lachmann A, Giorgi FM, Lopez G, Califano A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*. 2016;32(14):2233-5.
- [57] Guitart-Pla O, Kustagi M, Rugheimer F, Califano A, Schwikowski B. The Cyni framework for network inference in Cytoscape. *Bioinformatics*. 2015;31(9):1499-1501.
- [58] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning (with application in R). New York: Springer; 2014.
- [59] Matteucci M. Machine Learning, Teacher Slides. [Online] Available from: [http://chrome.ws.dei.polimi.it/index.php/Machine\\_Learning#Teaching\\_Material\\_.28the\\_textbook.29](http://chrome.ws.dei.polimi.it/index.php/Machine_Learning#Teaching_Material_.28the_textbook.29)
- [60] Guffanti F, Fratelli M, Ganzinelli M, Bolis M, Ricci F, Bizzarro F, Chilà R, Sina FP, Fruscio R, Lupia M, Cavallaro U, Cappelletti MR, Ganerali D, Giavazzi R, Damia G. Platinum sensitivity and DNA repair in a recently established panel of patient-derived ovarian carcinoma xenografts *Oncotarget*. 2018;9(37):24707-24717.
- [61] The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474:609-615.
- [62] Baldwin ROL, Nemeth E, Tran H, Shvartsman H, Cass I, Narod S, Karlan BY. BRCA1 promoter region hypermethylation in ovarian carcinoma. *Cancer Research*. 2000;60(19):5329-33.
- [63] Lederman J, Harter P, Gourley C, Friedlander M et al. Olaparib maintenance therapy in platinum-sensitive relapsed ovarian cancer. *The New England Journal of Medicine*. 2012;366:1382-1392.
- [64] Rikagos G, Razis E. BRCAness: finding the Achilles heel in ovarian cancer. *Oncologist*. 2012;17(7):956-62.
- [65] Zhu Q, Wani AA. Nucleotide Excision Repair: finely tuned molecular orchestra of early pre-incision events. *Photochemistry and Photobiology*. 2017;93(1):166-177.
- [66] Riberio E et al. Triple negative breast cancers have a reduced expression of DNA repair genes. *PLoS One*. 2013;8(6):e66243.
- [67] Bikbak NJ et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discovery*. 2012;2(4):366-375.
- [68] Citations from "Grey's Anatomy". Shondaland, ABC Studios, 2005-2018.



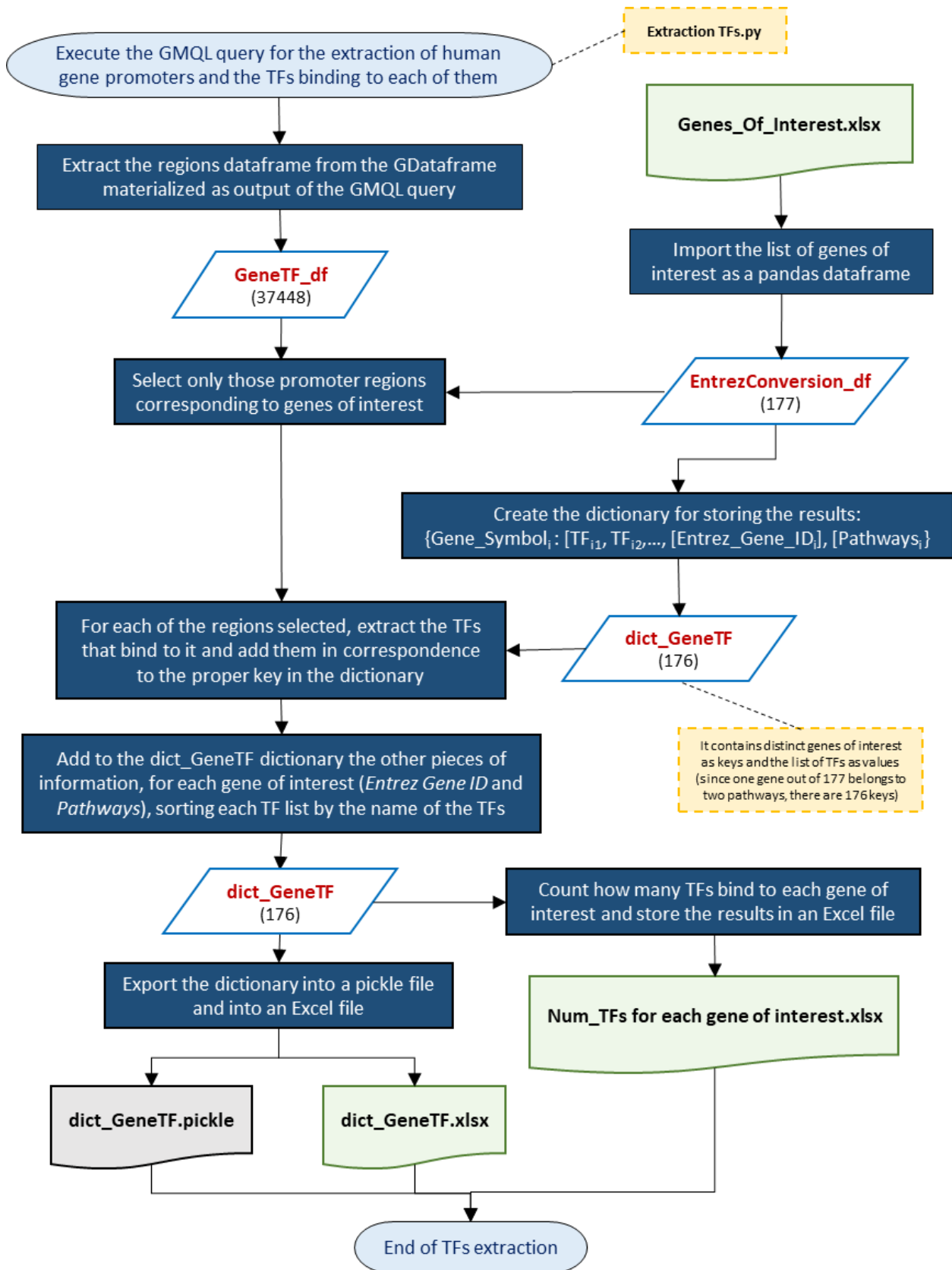
# Appendix A

## Python scripts flowcharts

### A.1 Genes – transcription factors mapping

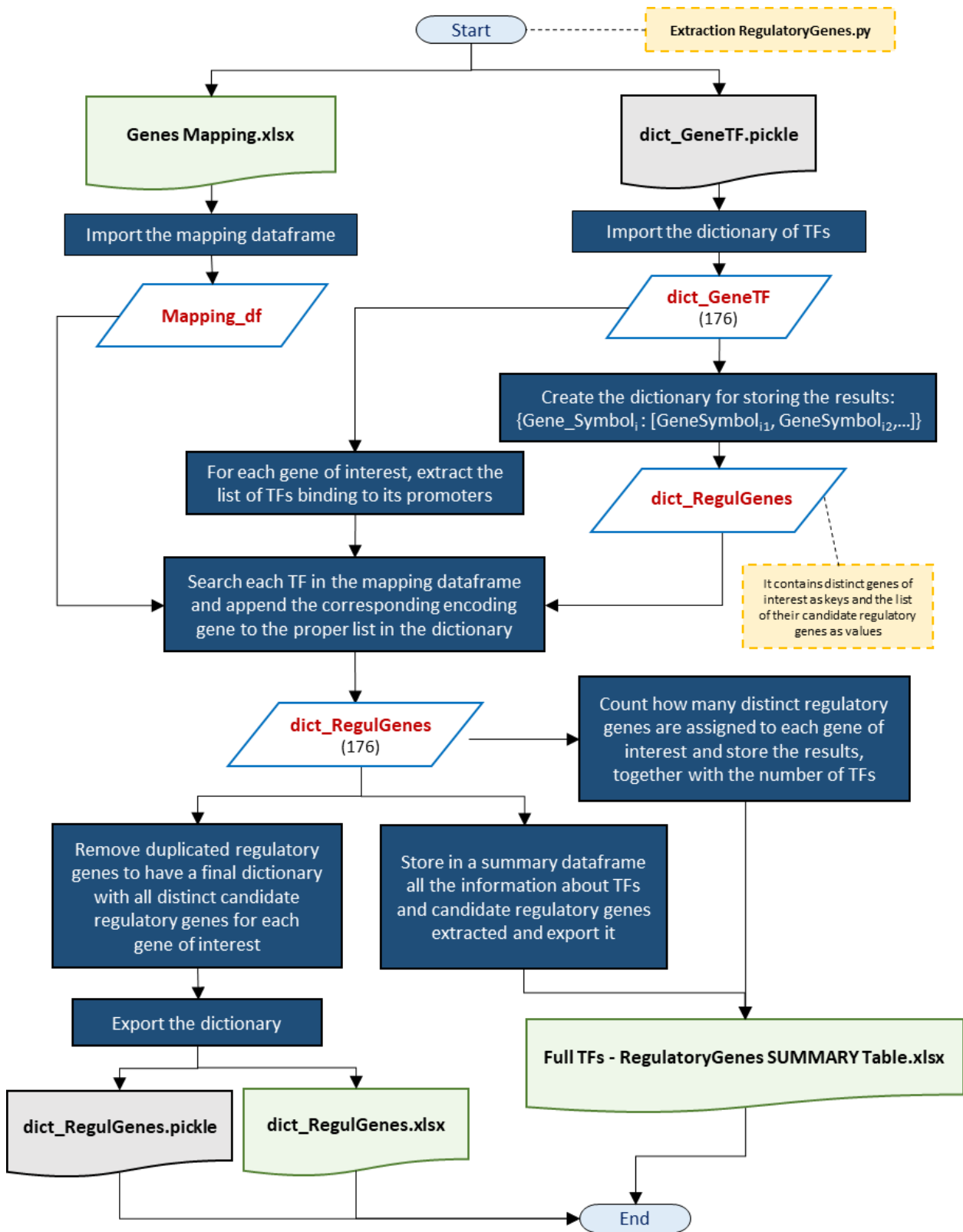


## A.2 Selection of transcription factors

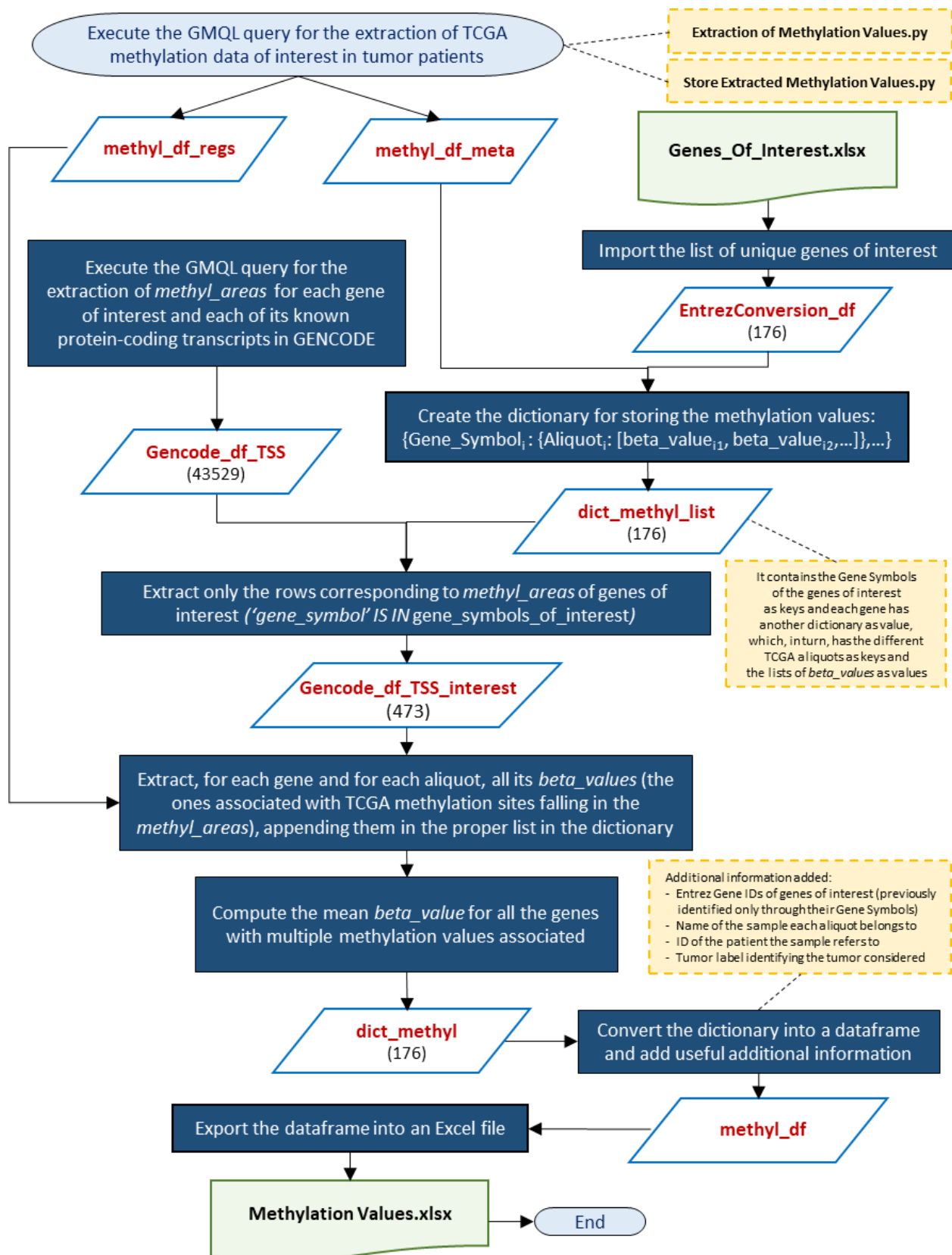




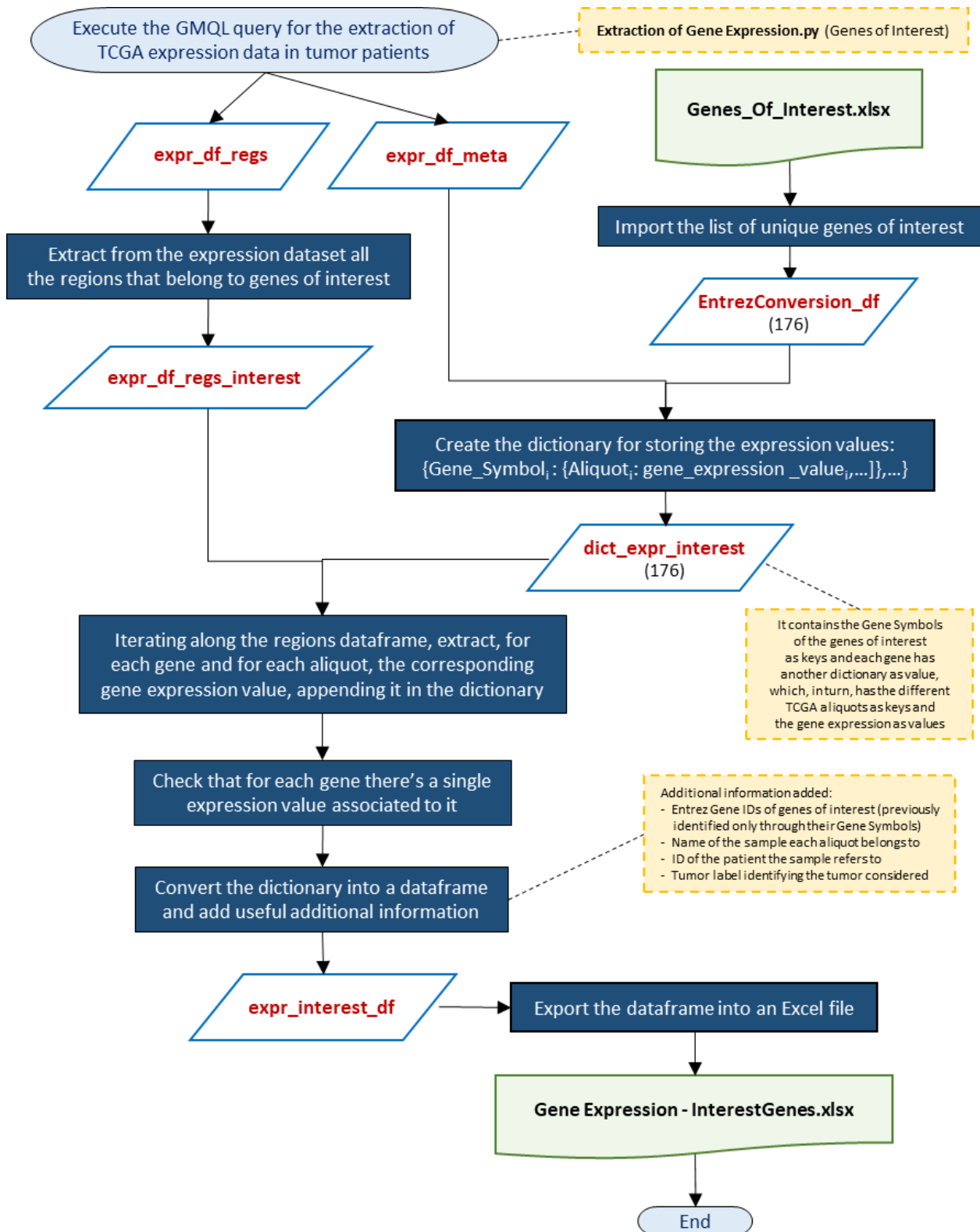
### A.3 Identification of candidate regulatory genes



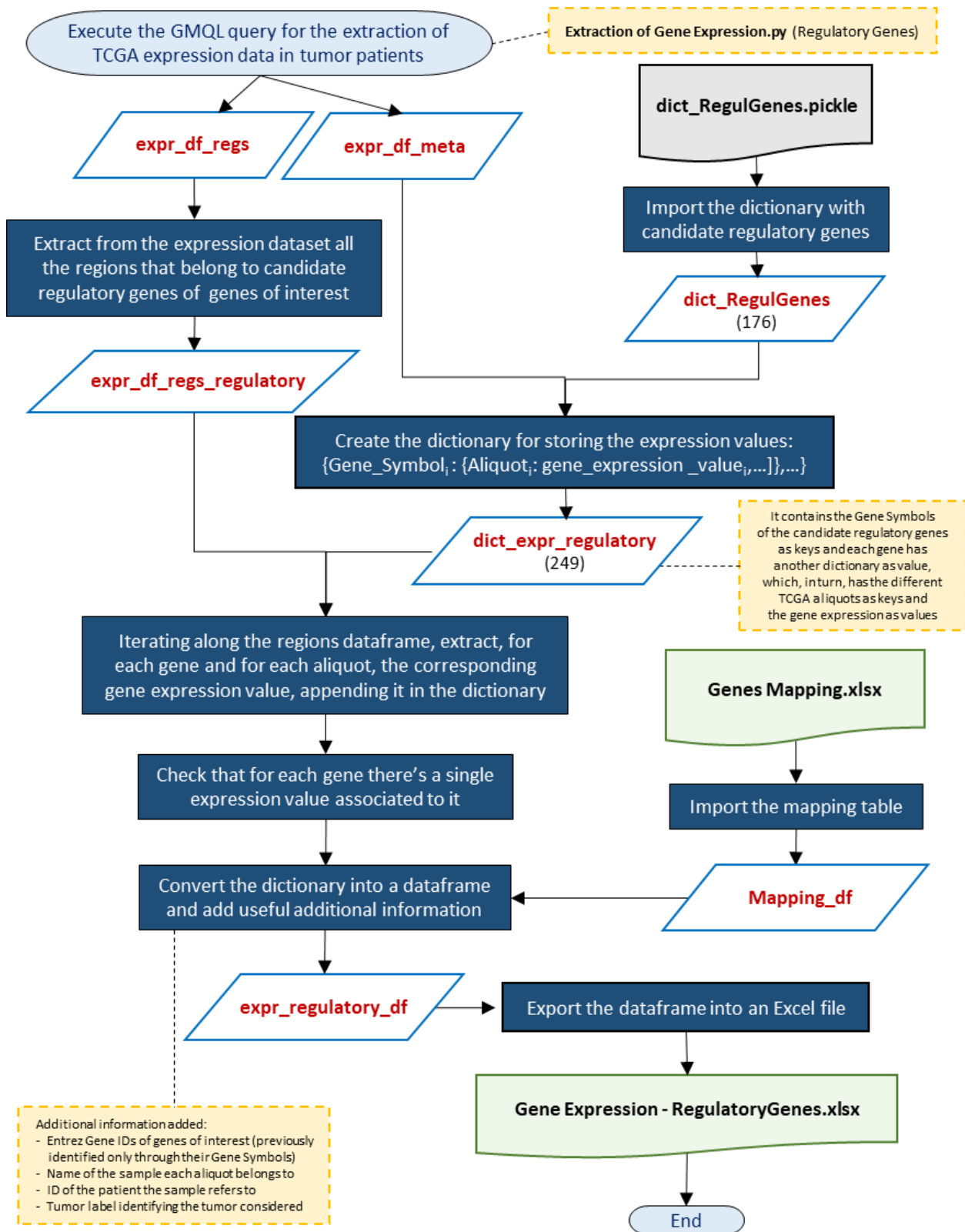
## A.4 Extraction of methylation values



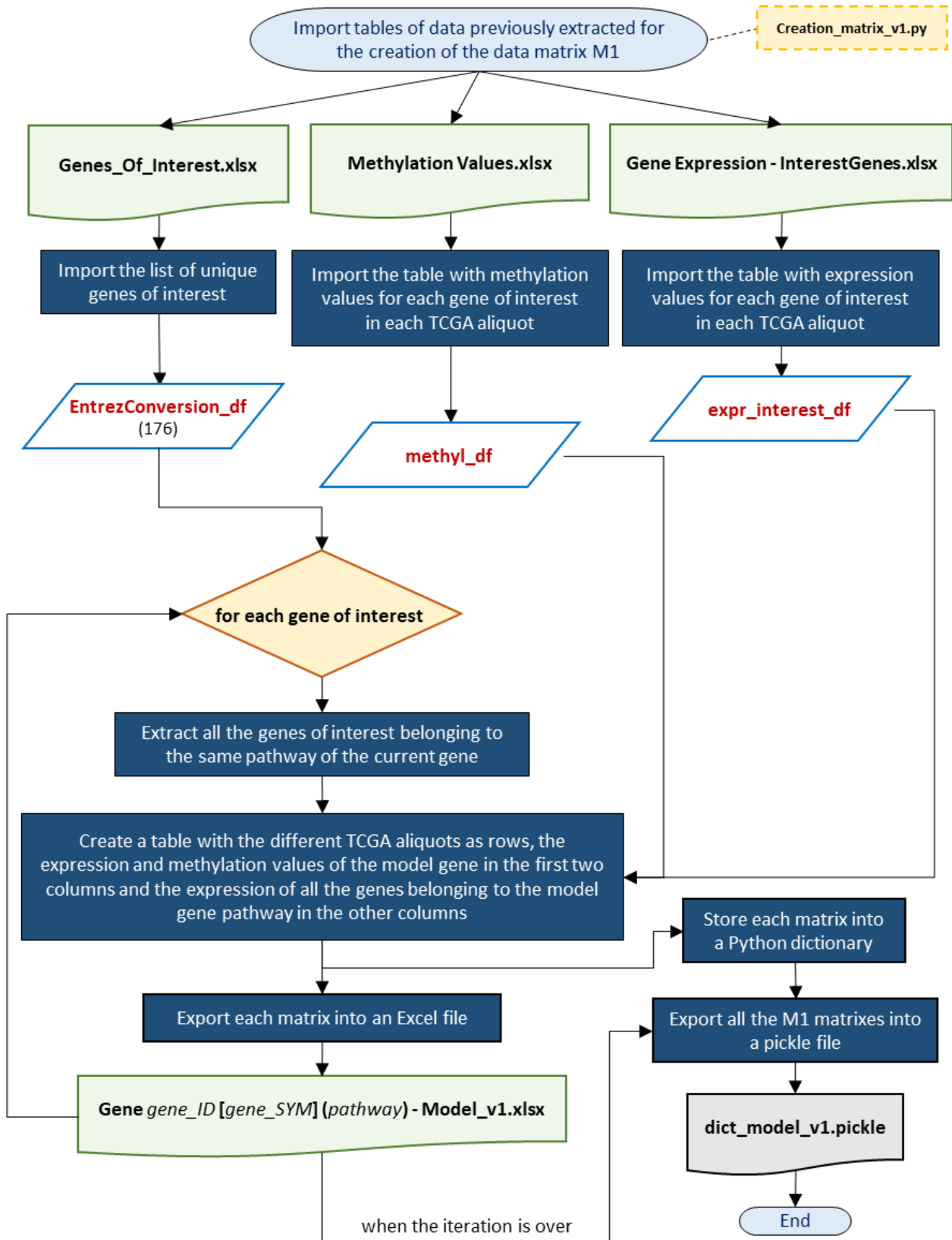
## A.5 Extraction of gene expression values (genes of interest)



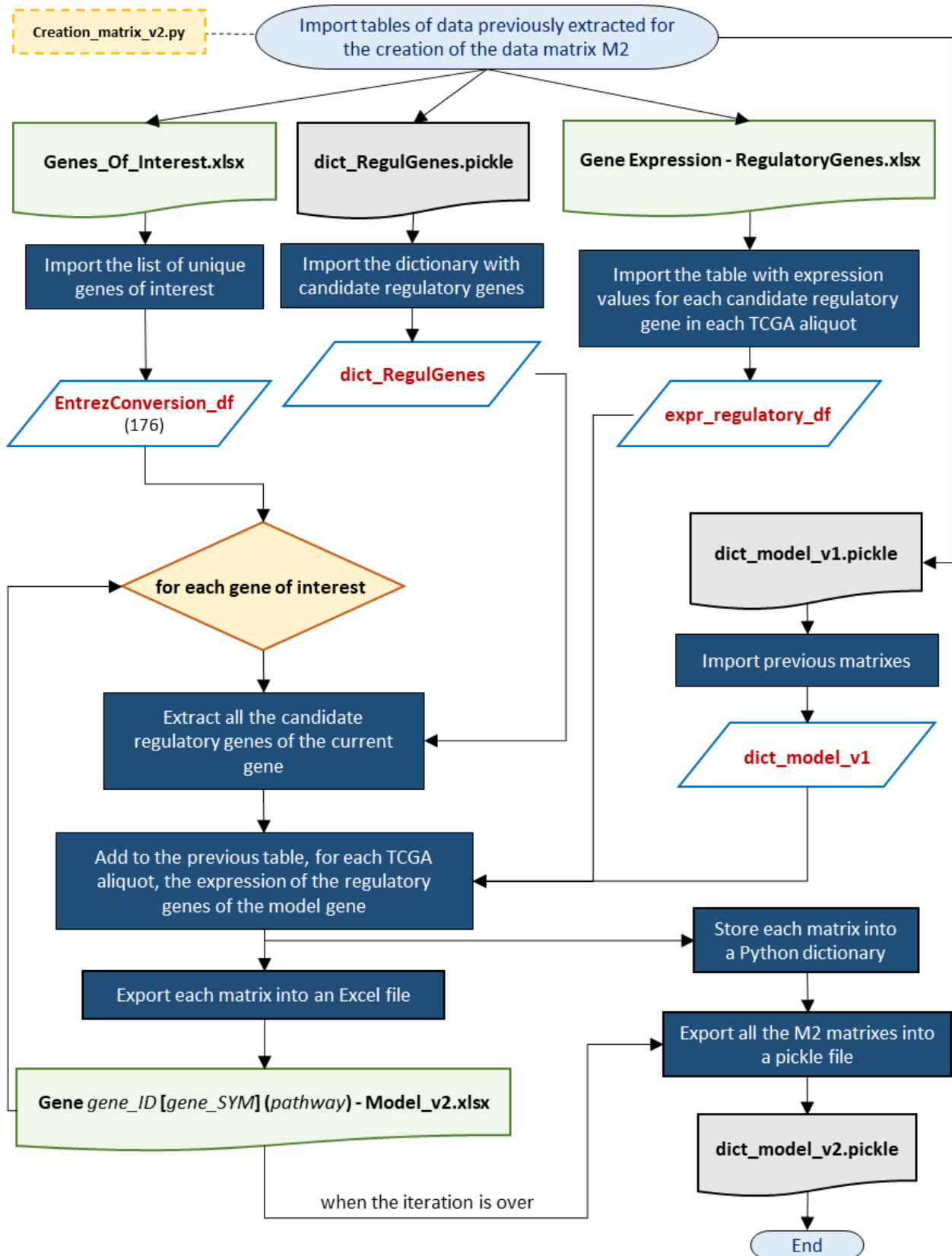
## A.6 Extraction of gene expression values (candidate regulatory genes)



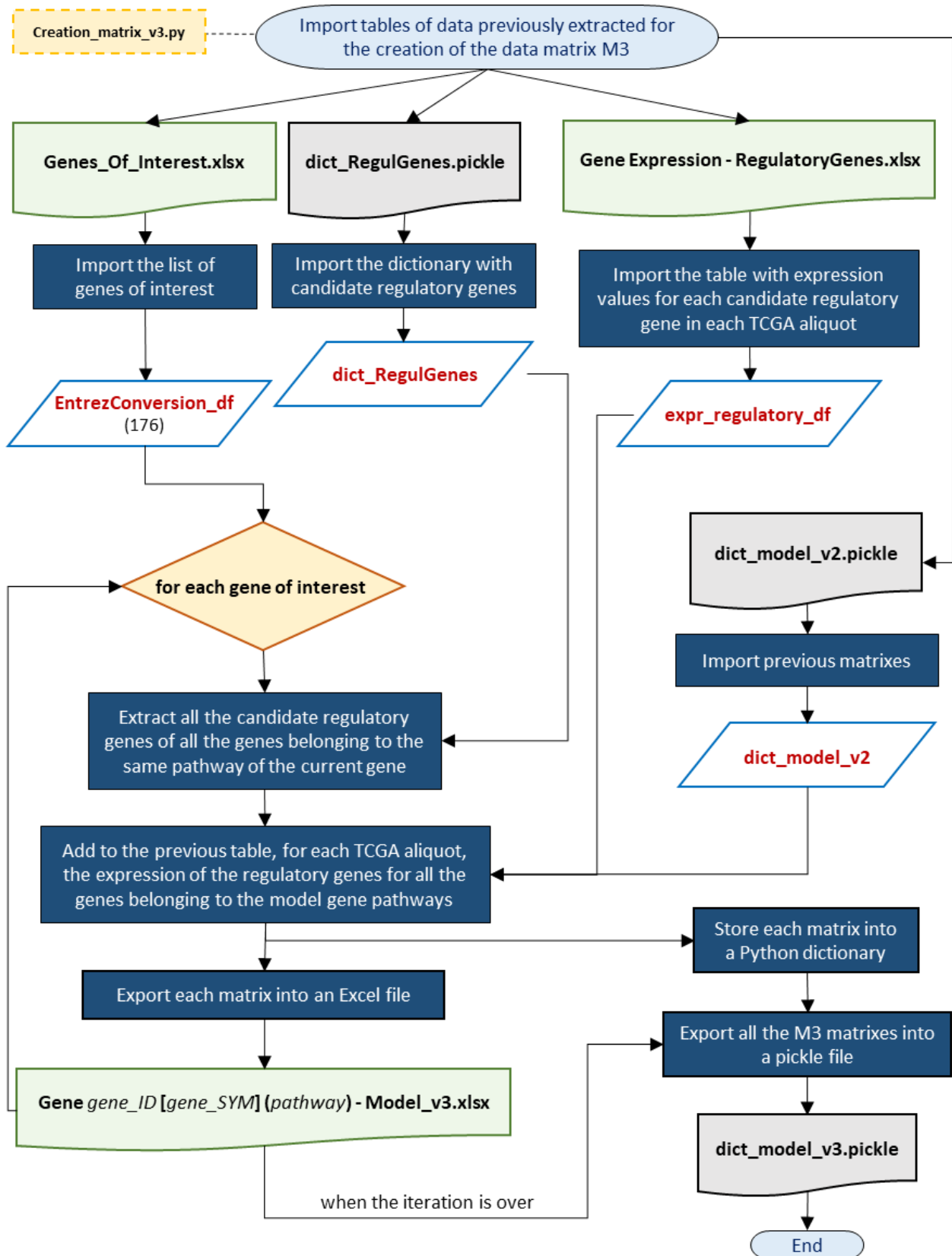
## A.7 Data matrix construction: M1



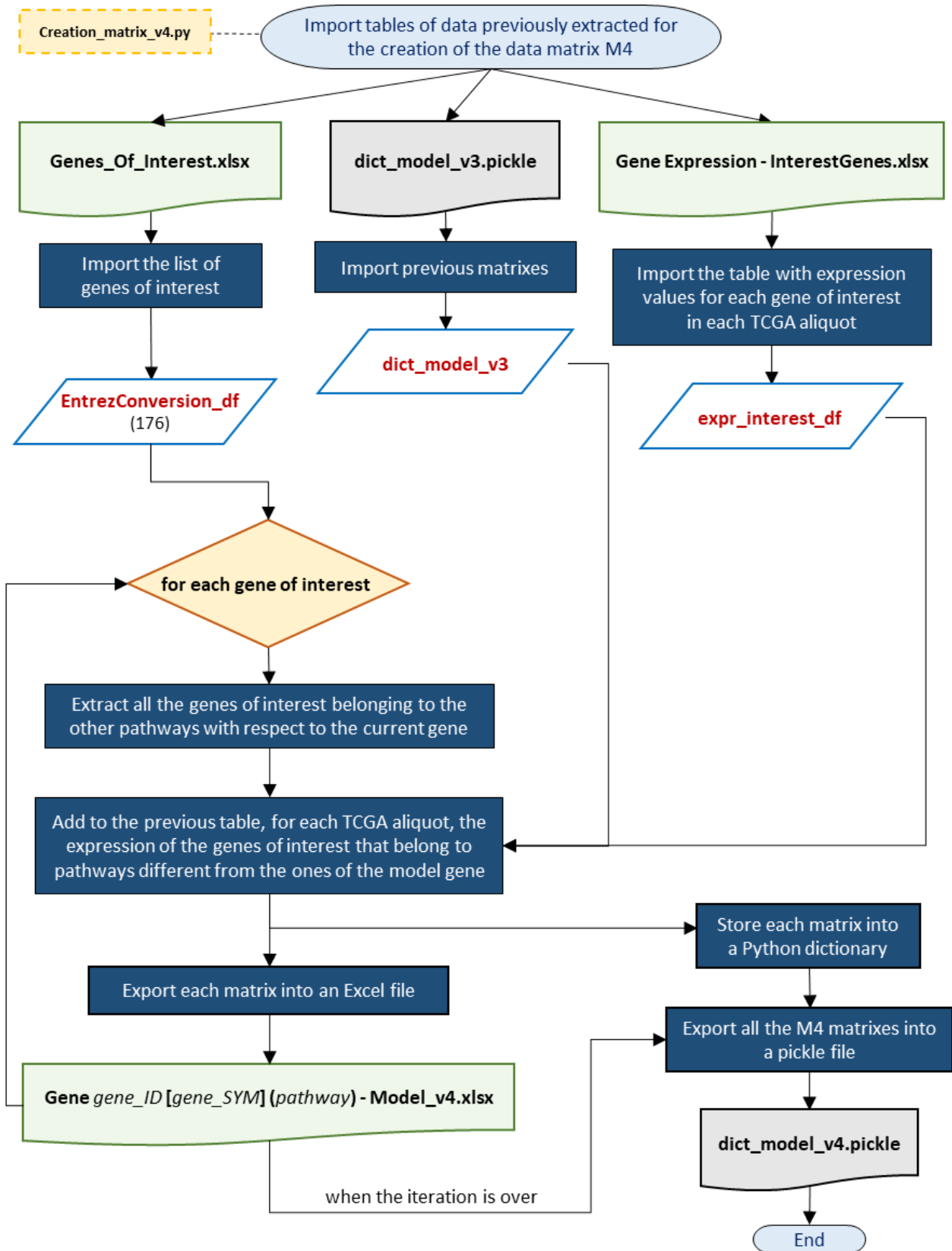
## A.8 Data matrix construction: M2



## A.9 Data matrix construction: M3

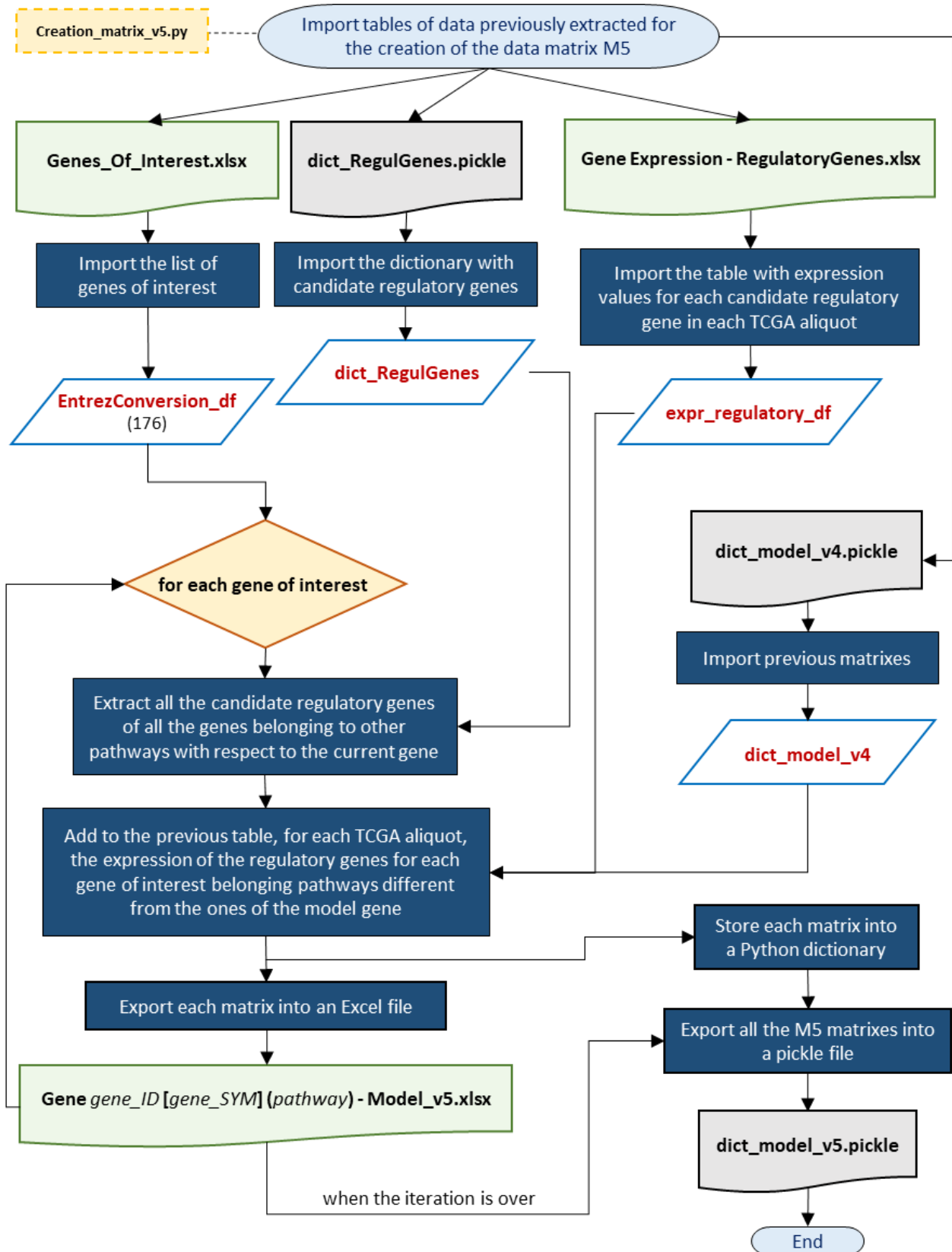


## A.10 Data matrix construction: M4

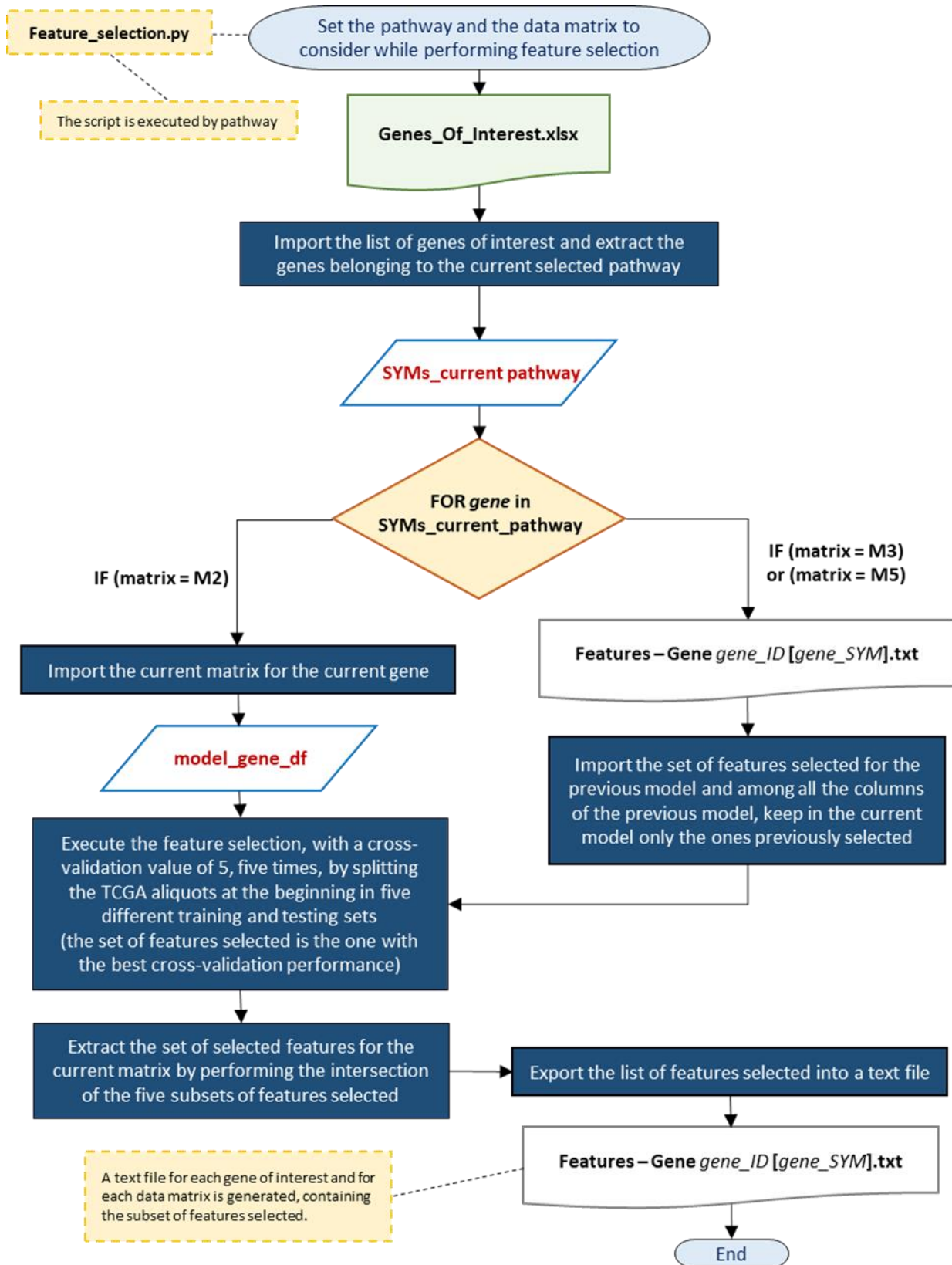




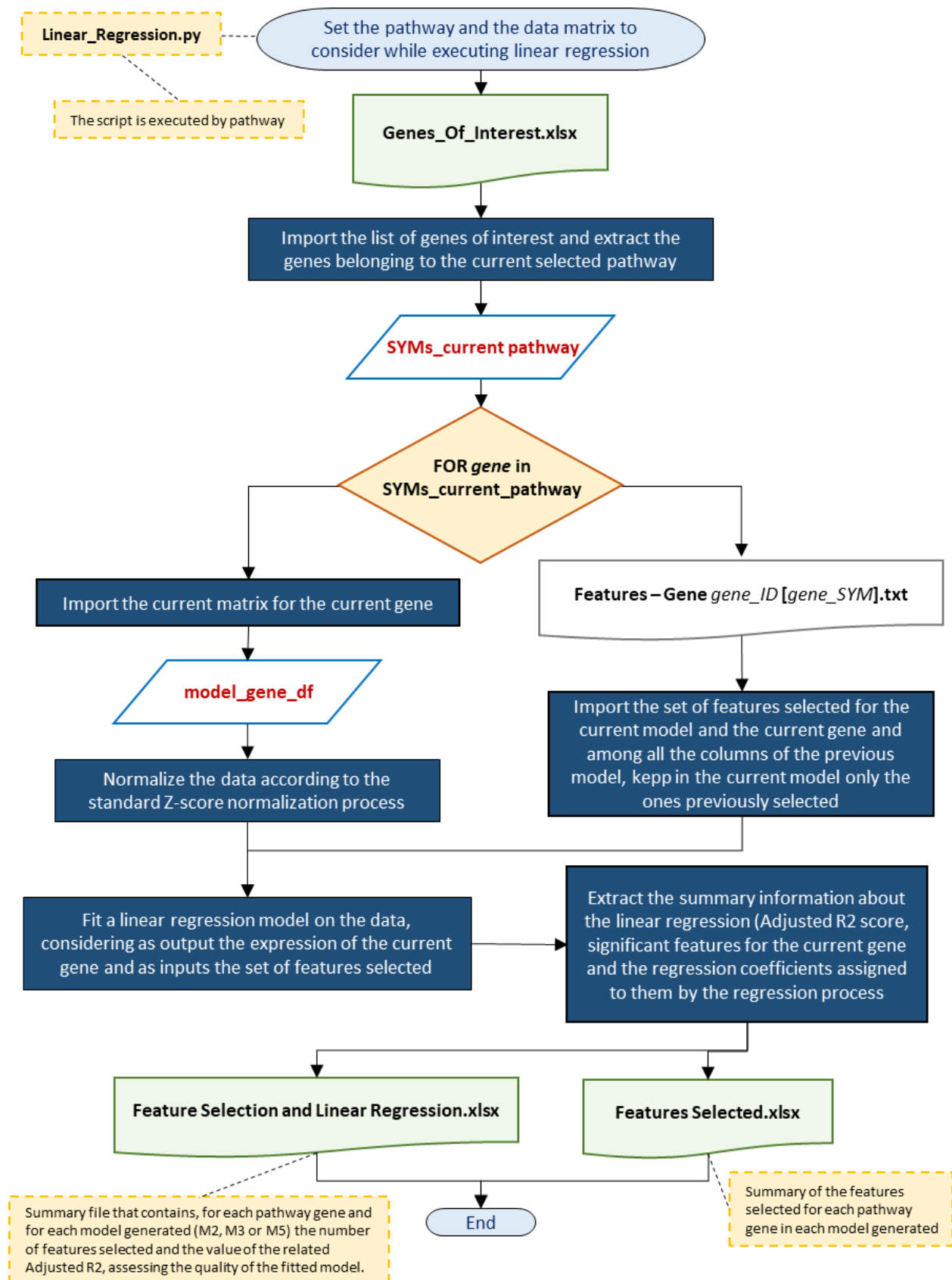
## A.11 Data matrix construction: M5



## A.12 Feature/gene selection



### A.13 Linear regression of individual genes on ovarian tumor samples

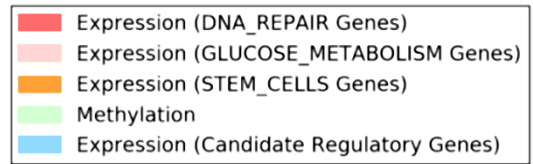




# Appendix B

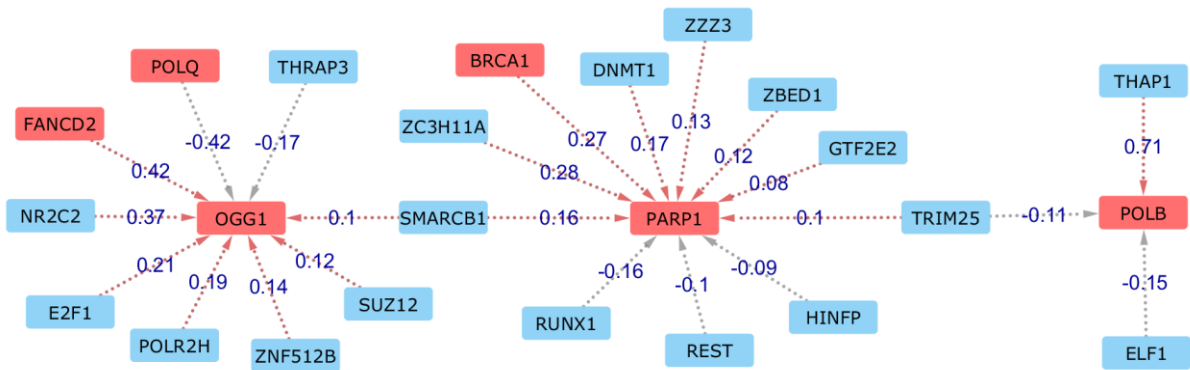
## Genetic expression networks from linear regression models

Legend

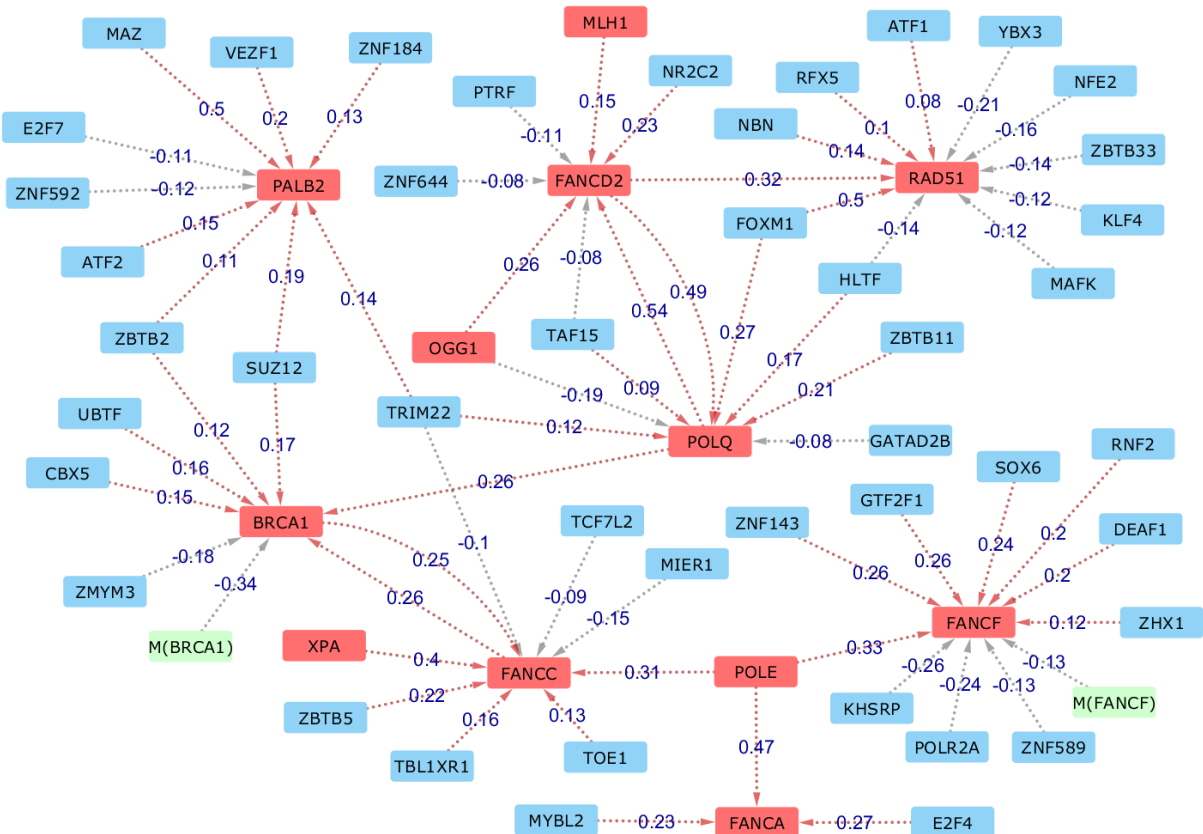


### B.1 DNA\_REPAIR (M3)

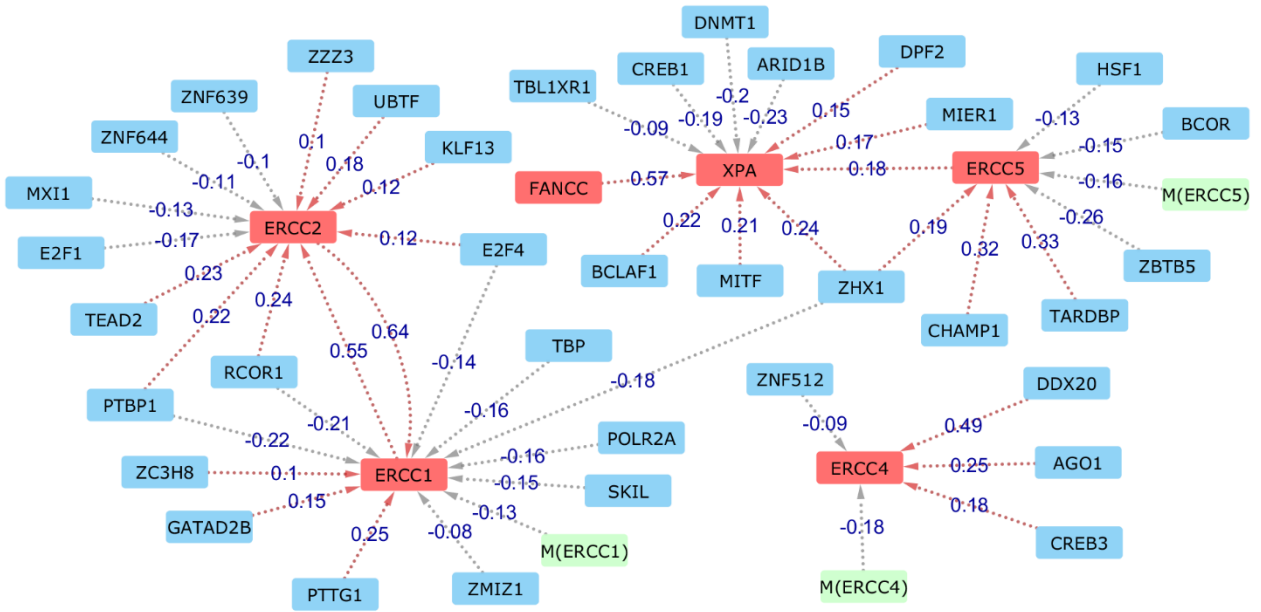
BER



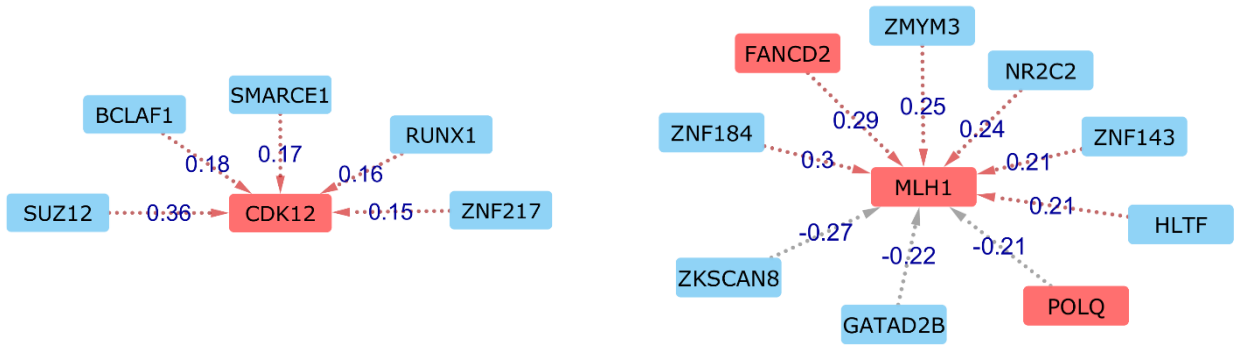
DSB



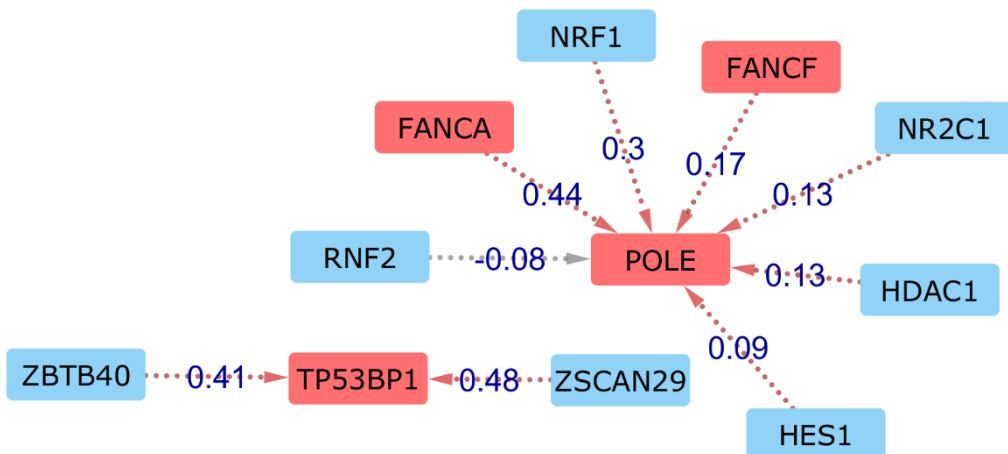
NER



Gene CDK12 Group & Gene MLH1 Group

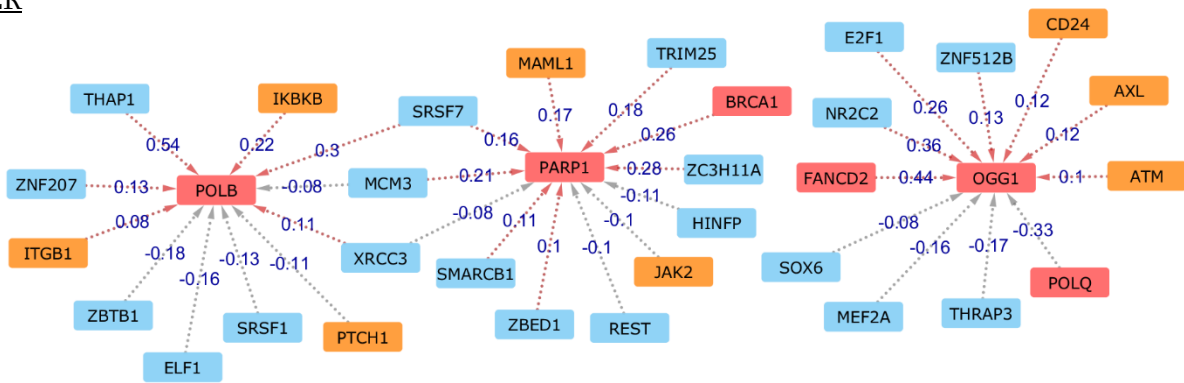


Unclassified

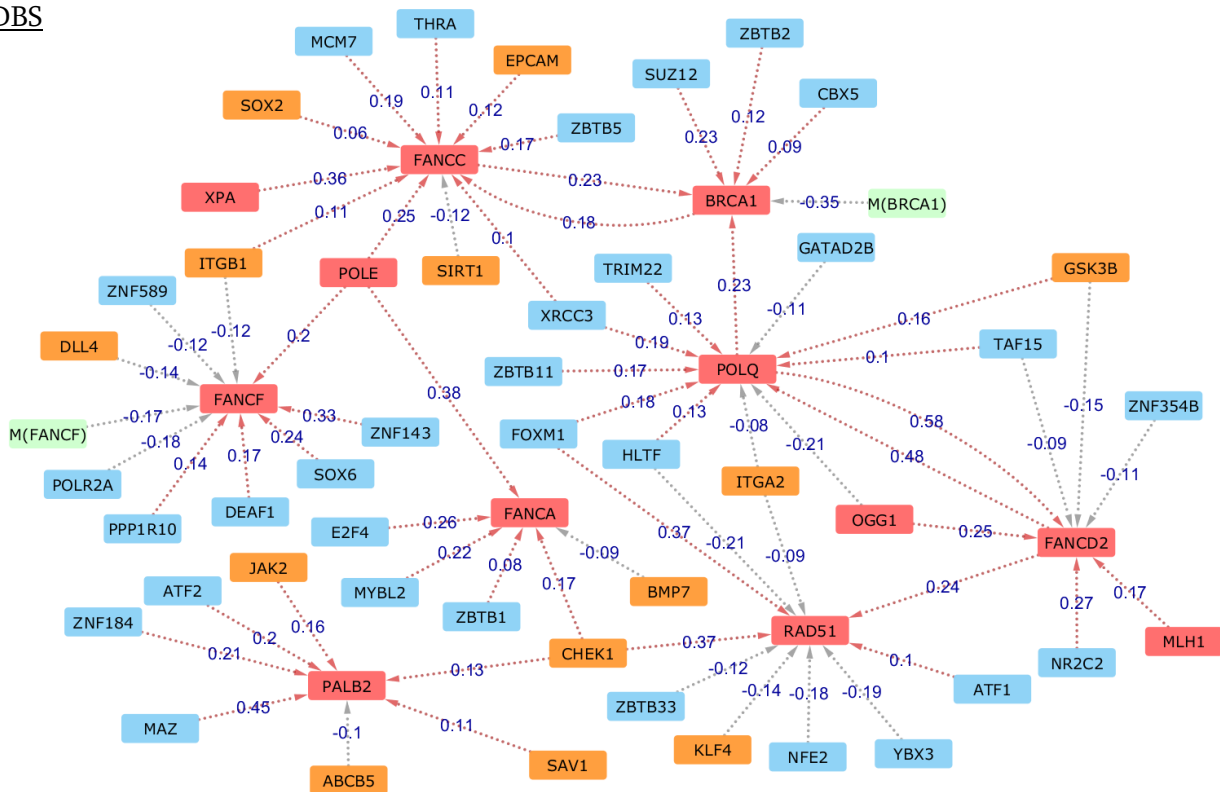


## B.2 DNA\_REPAIR (M5)

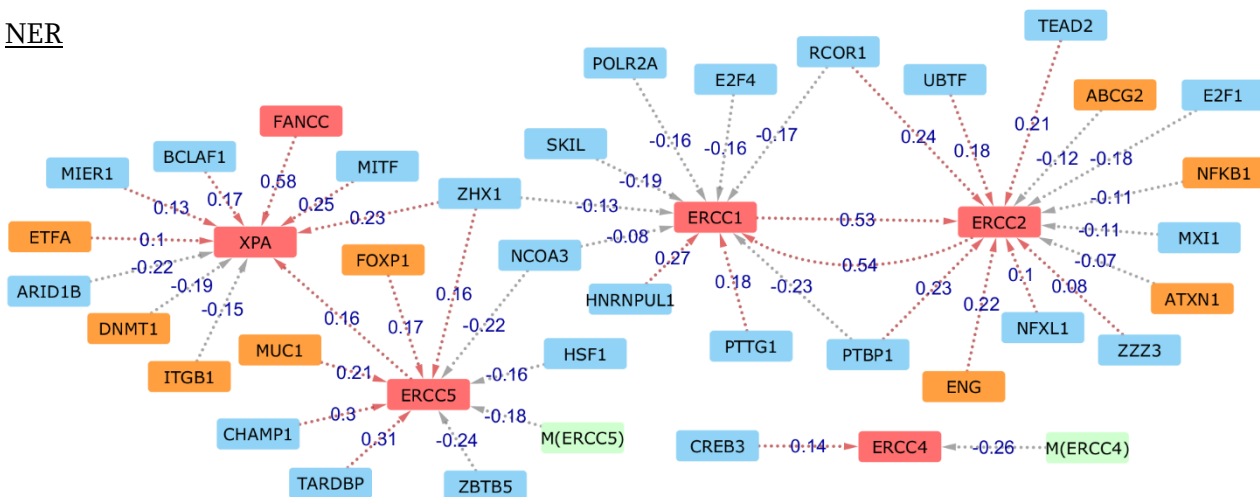
### BER



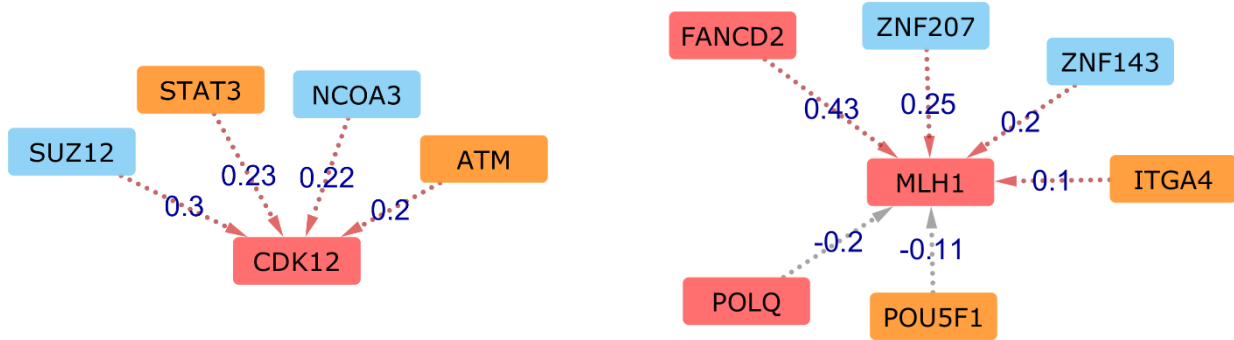
### DBS



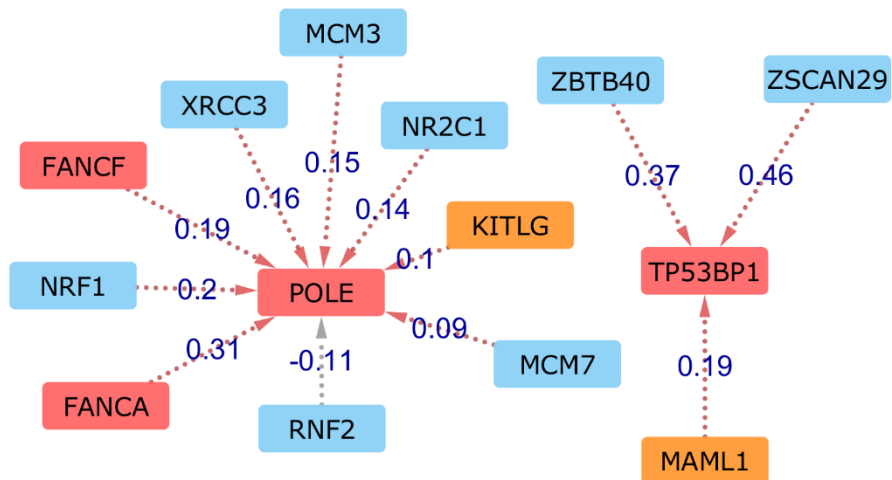
### NER



Gene CDK12 Group & Gene MLH1 Group

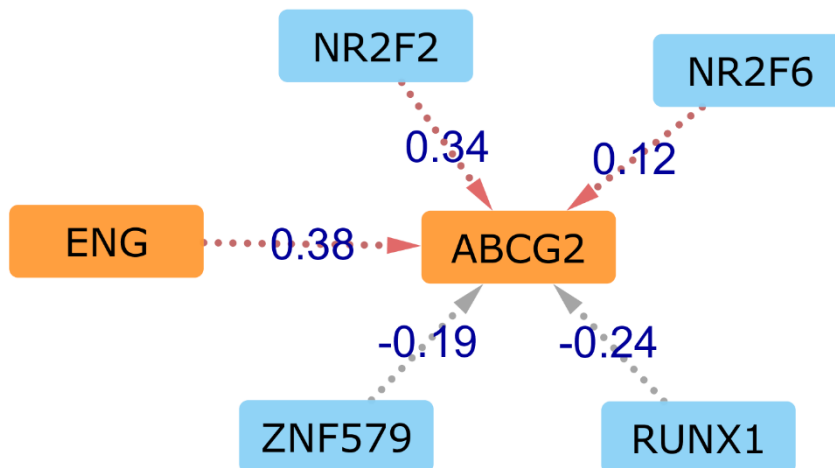


Unclassified



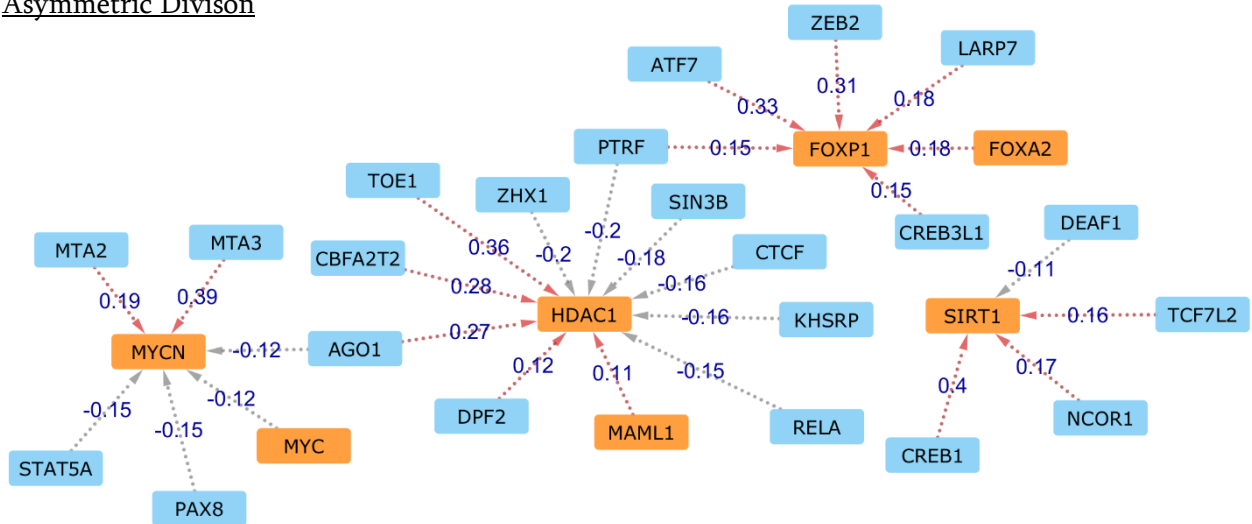
**B.3 STEM\_CELLS (M3)**

AKT & PI3 Kinase – mTOR Signaling

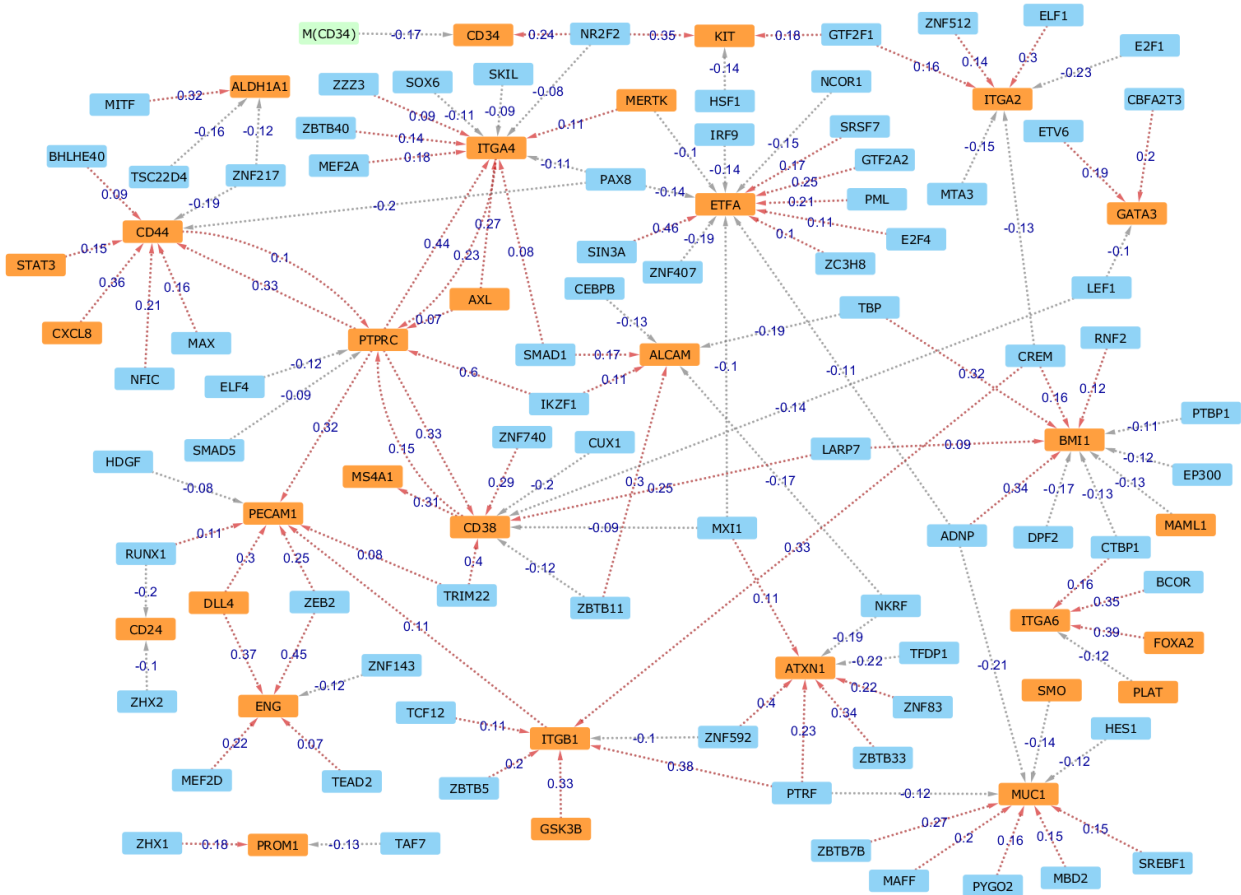




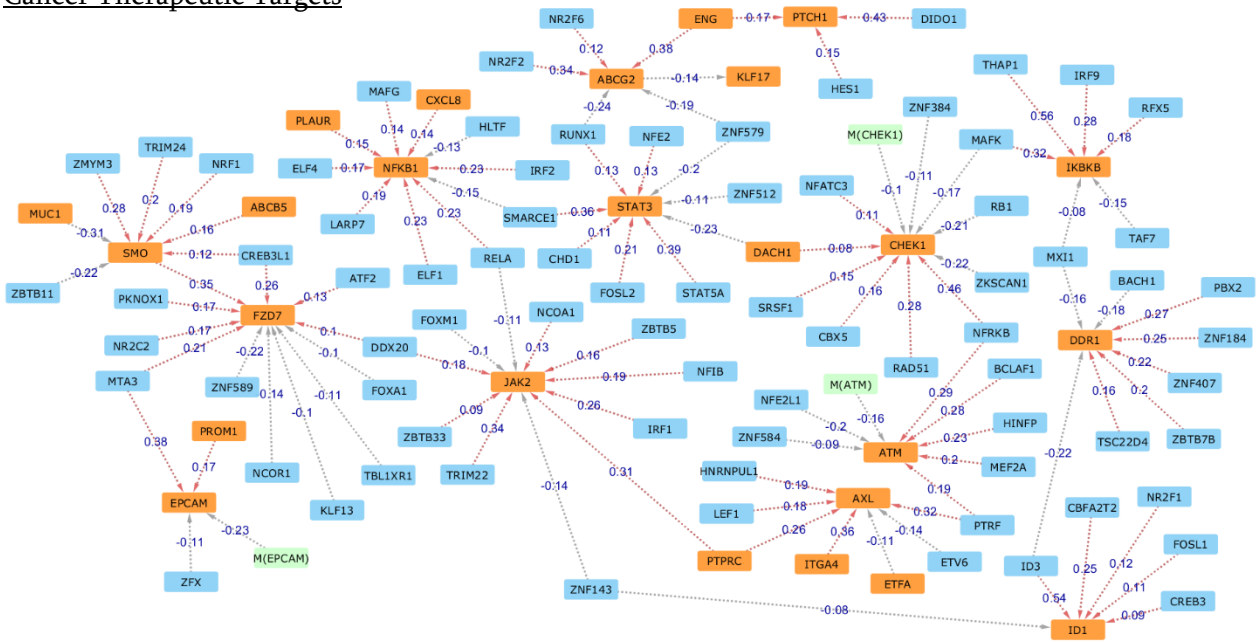
## Asymmetric Divison



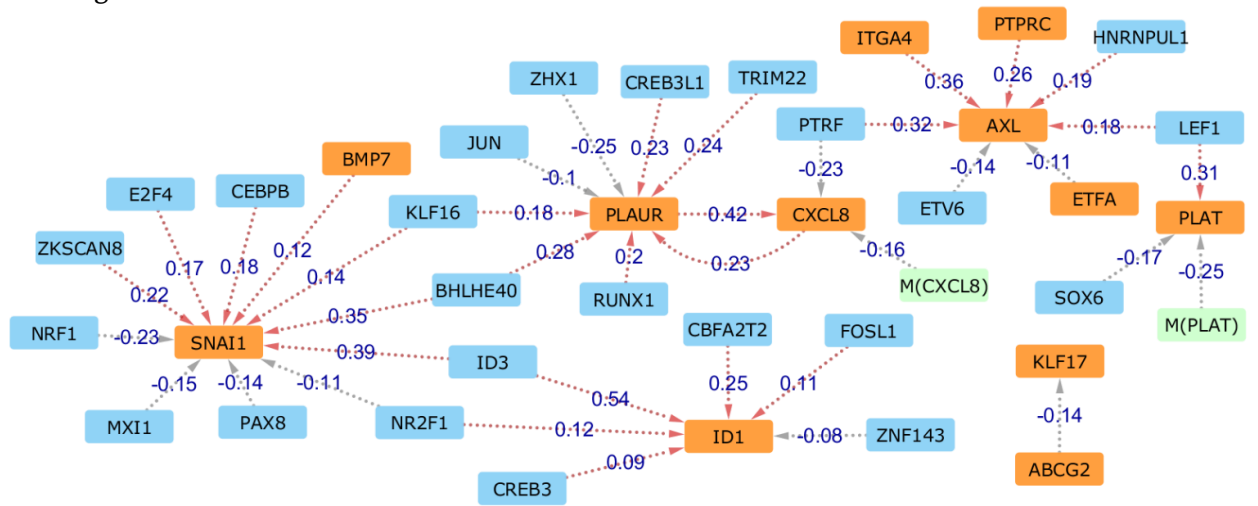
## Cancer Stem Cells Markers



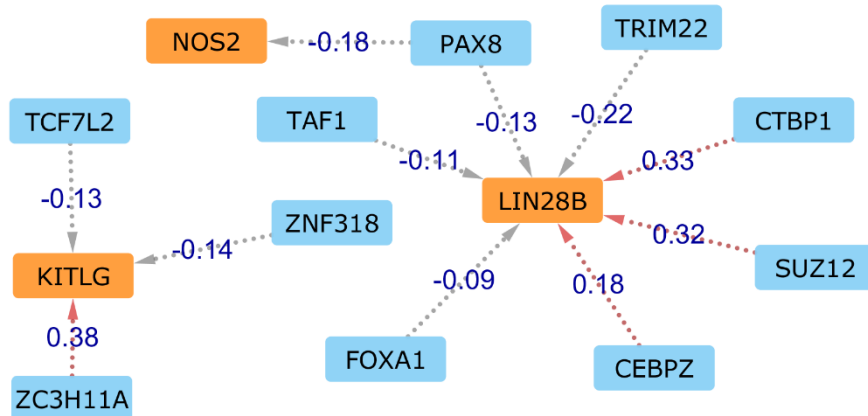
## Cancer Therapeutic Targets



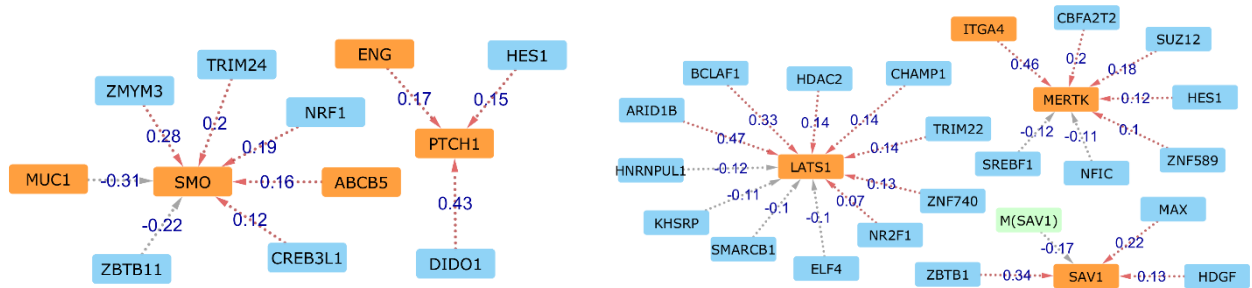
## Cell Migration & Metastasis



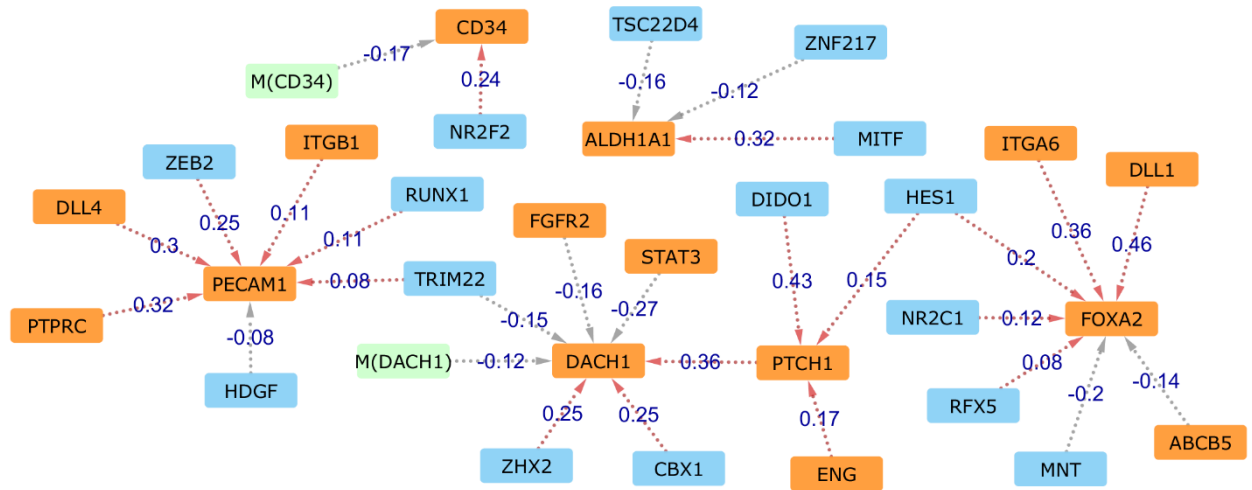
## Cell Proliferation



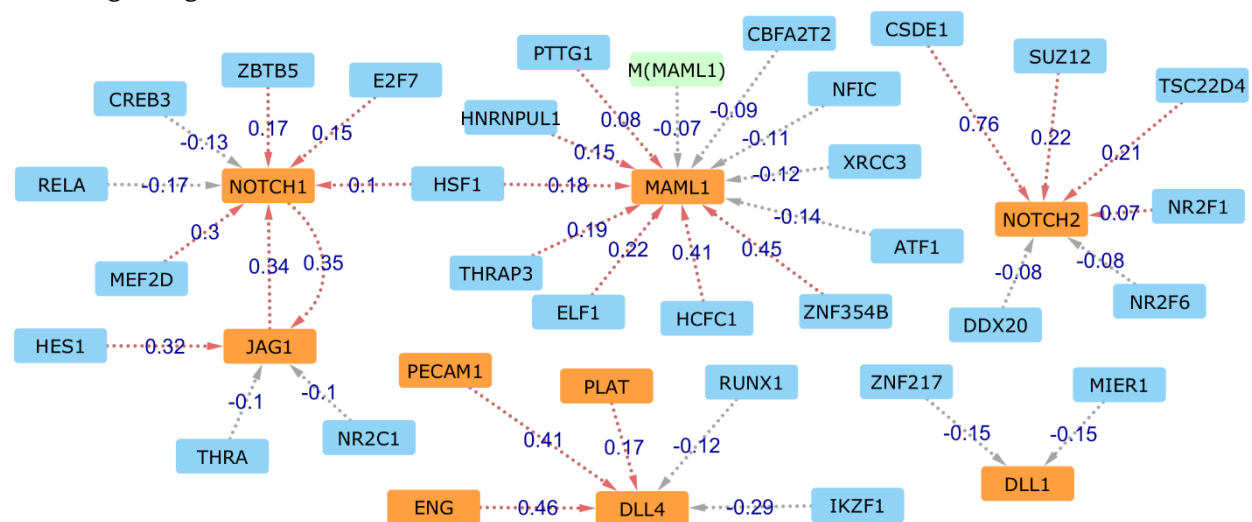
## Hedgehog Signaling & Hippo Signaling



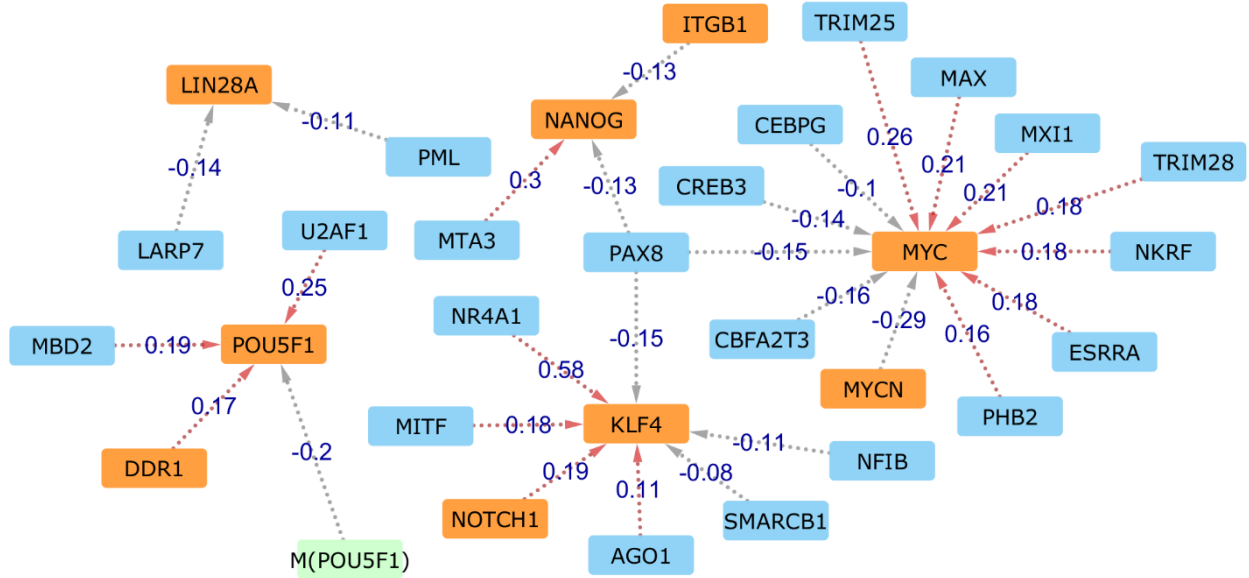
## Loss of Stemness



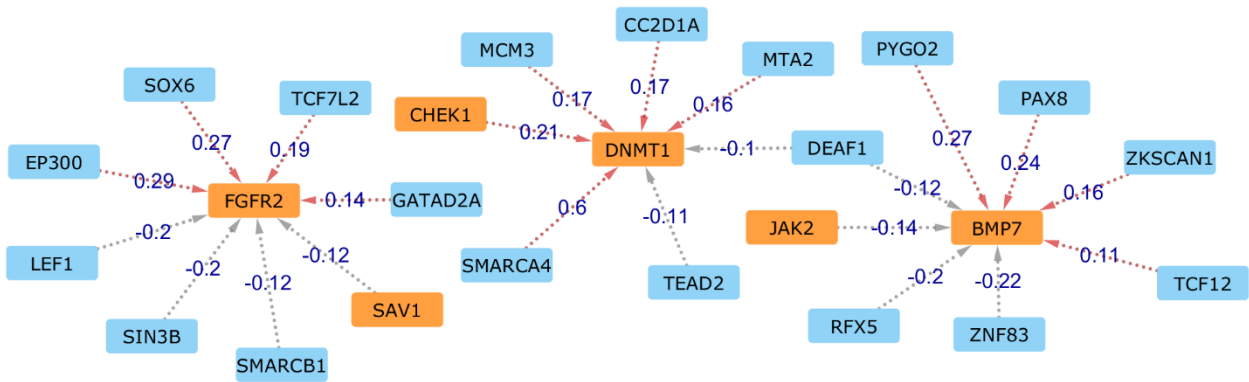
## Notch Signaling



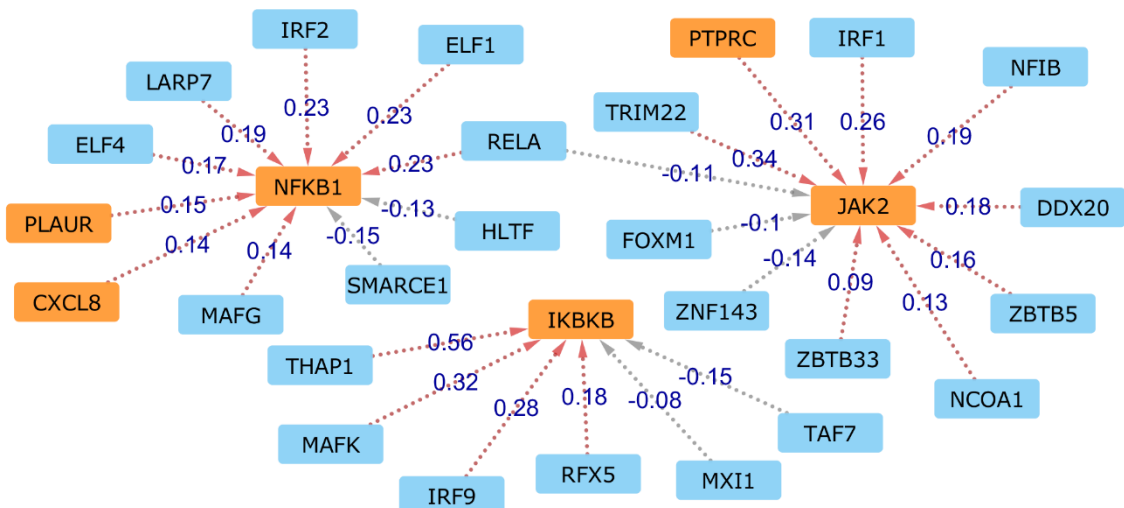
## Pluripotency



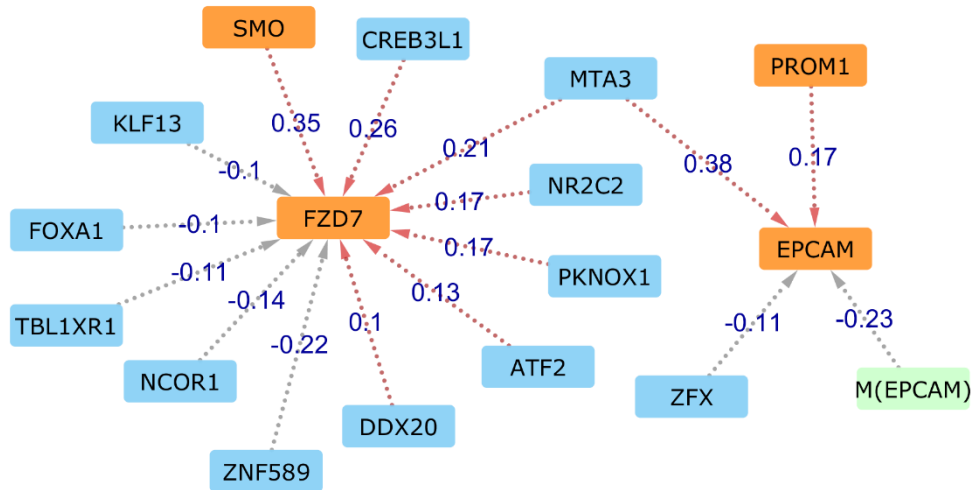
## Self-Renewal



## STAT-NFkB Signaling

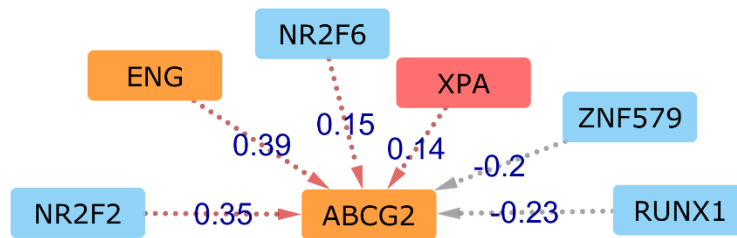


## WNT Signaling

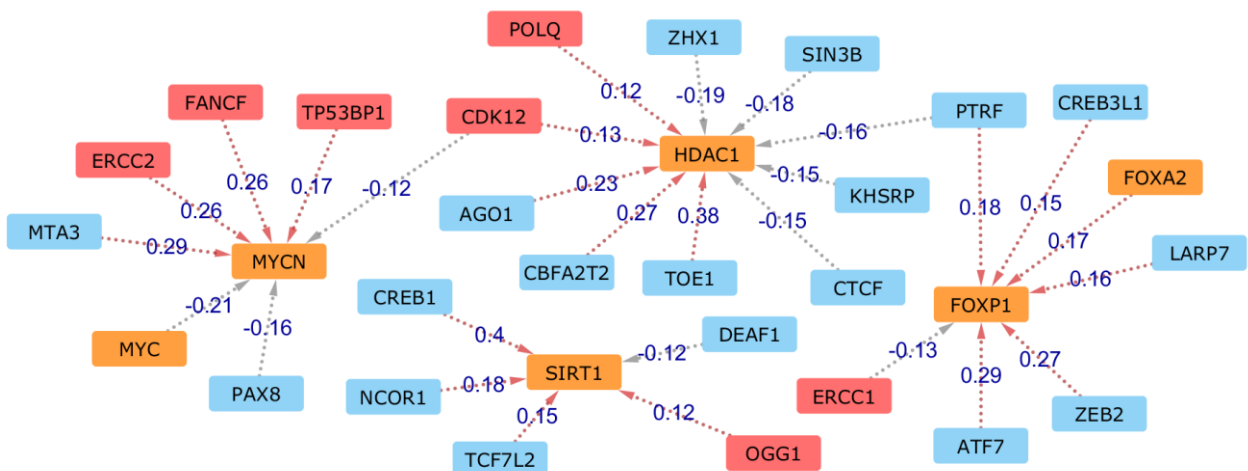


## B.4 STEM\_CELLS (M5)

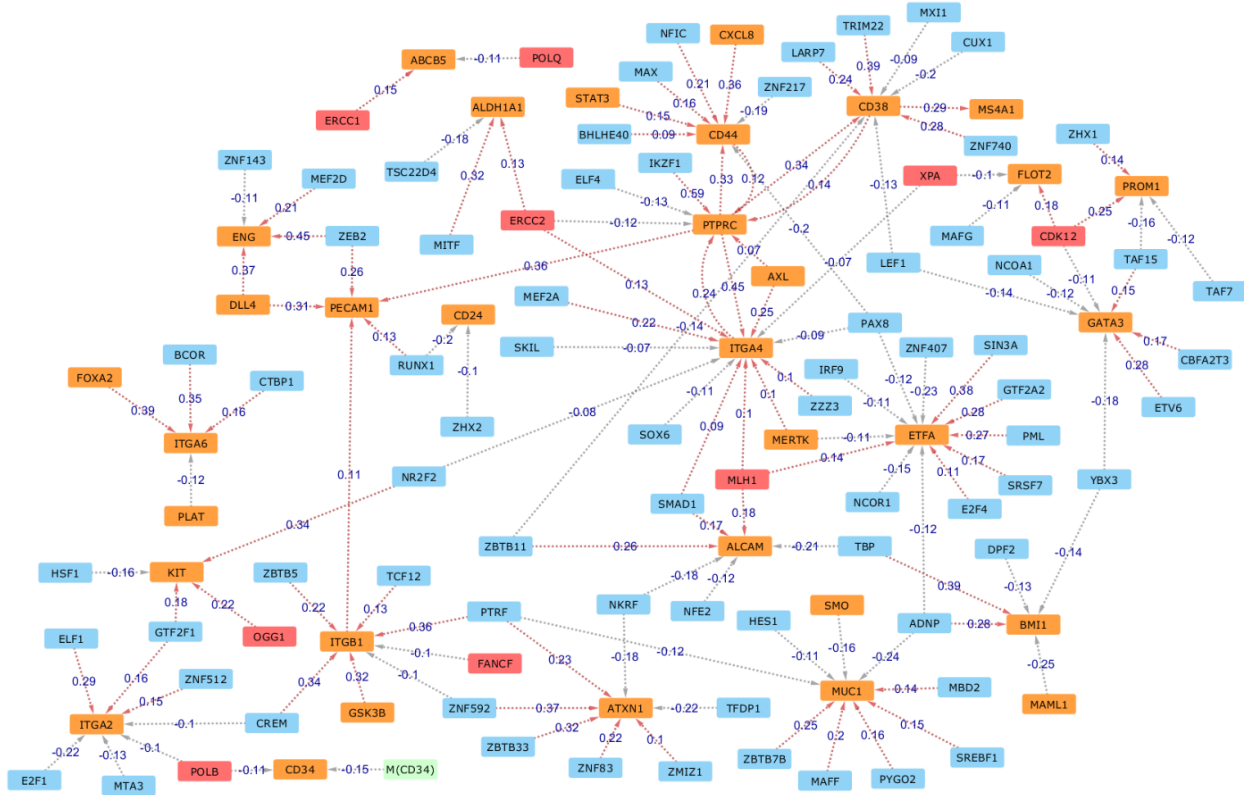
### AKT & PI3 Kinase – mTOR Signaling



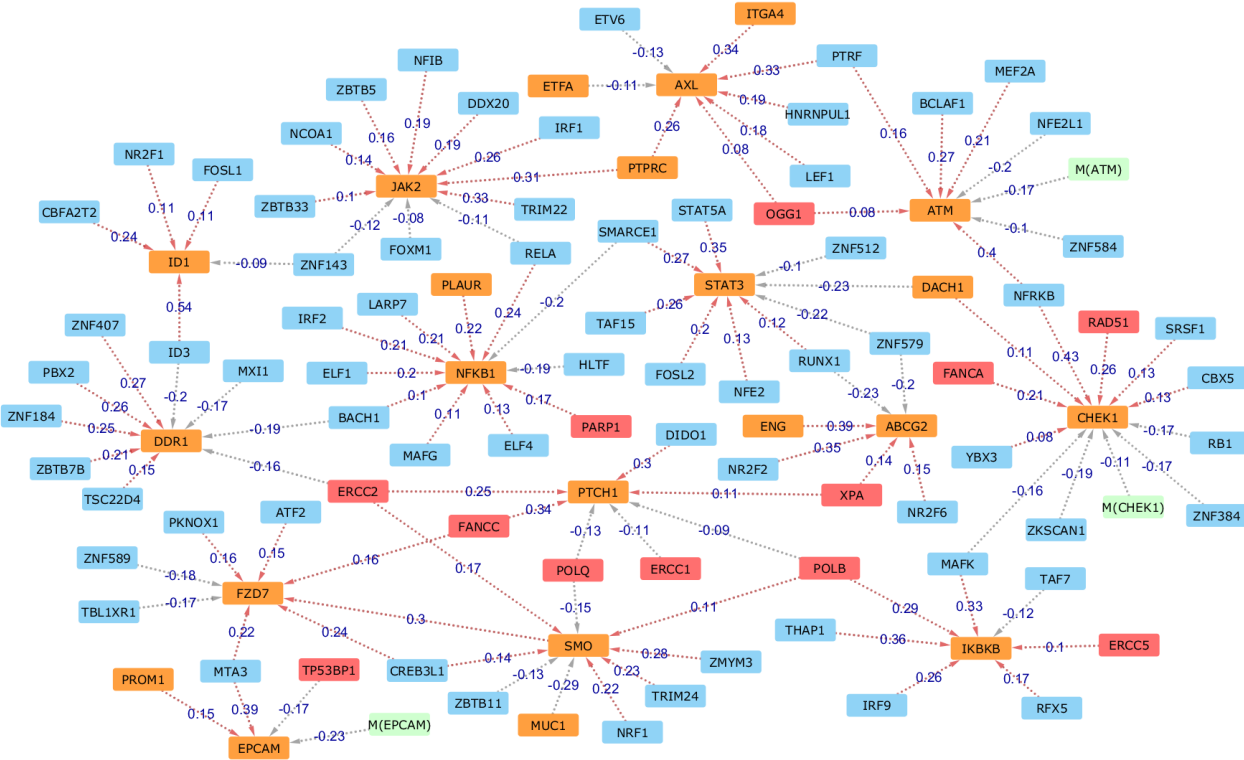
## Asymmetric Divison



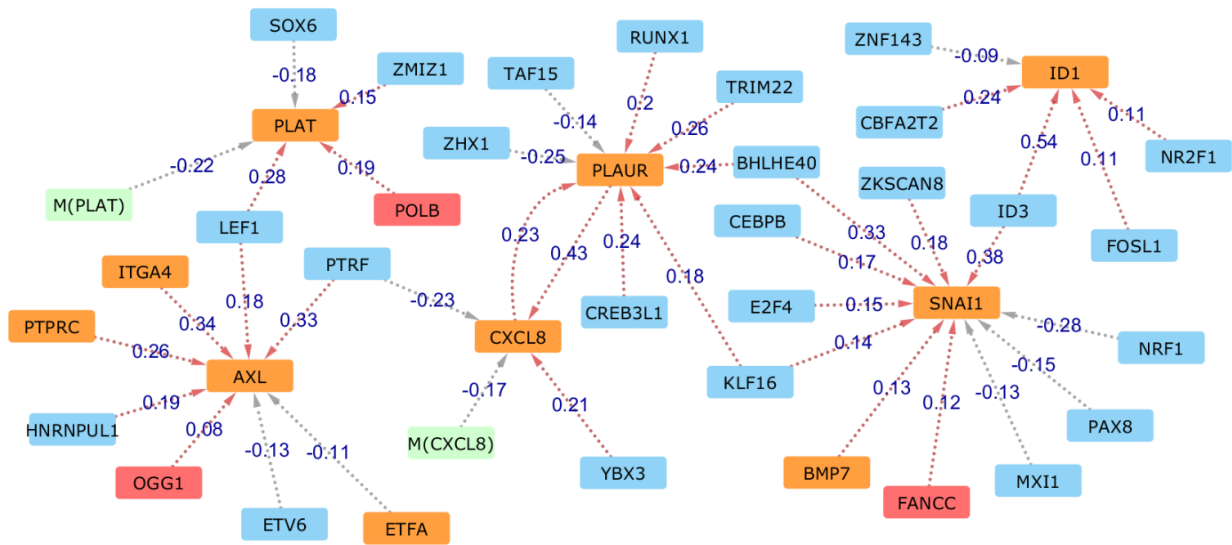
## Cancer Stem Cells Markers



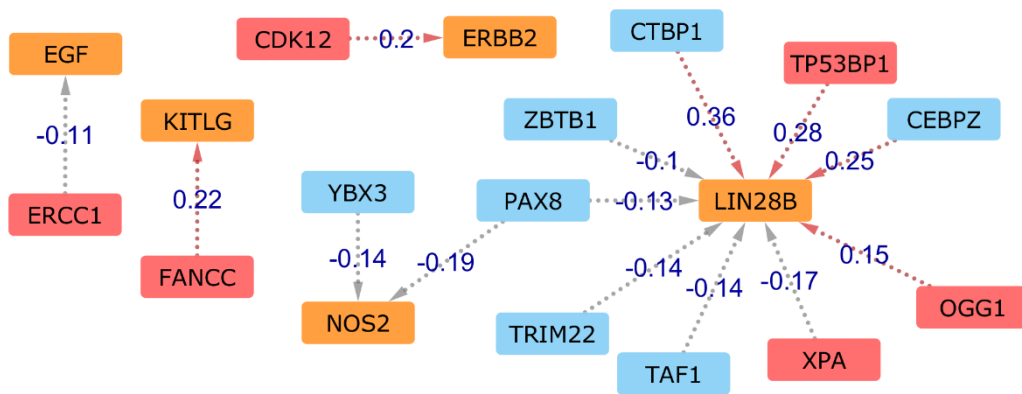
## Cancer Therapeutic Targets



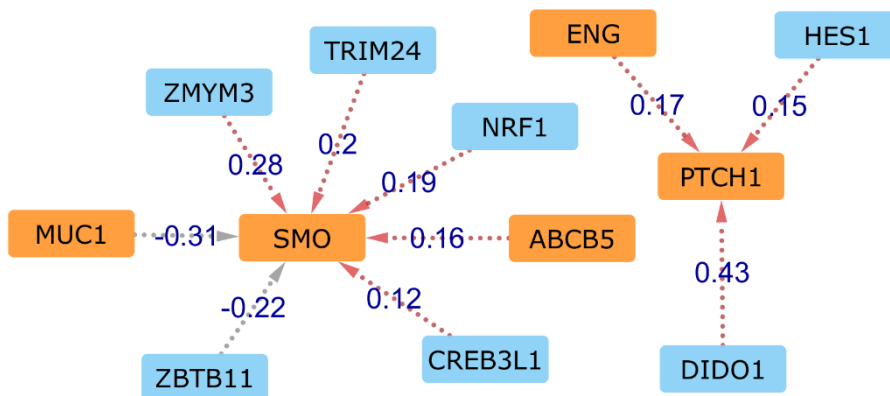
## Cell Migration & Metastasis



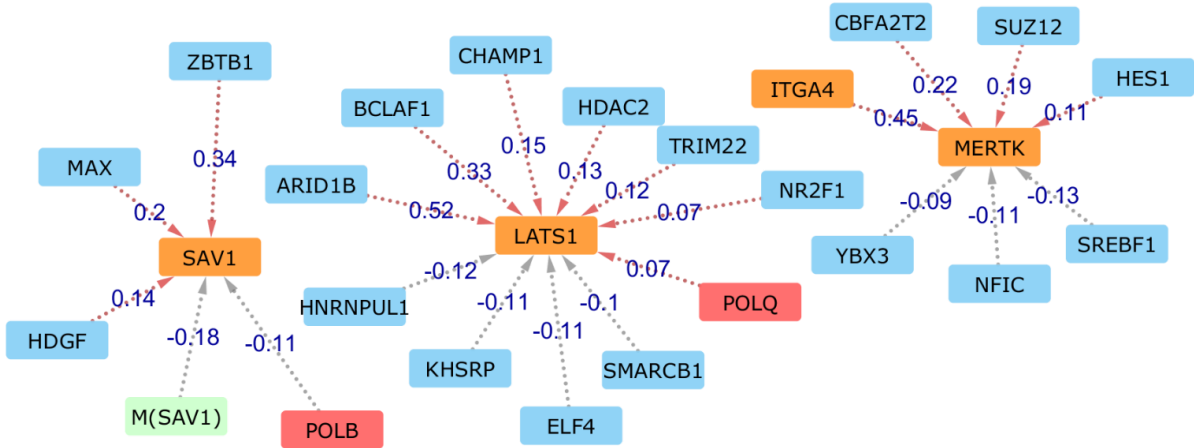
## Cell Proliferation



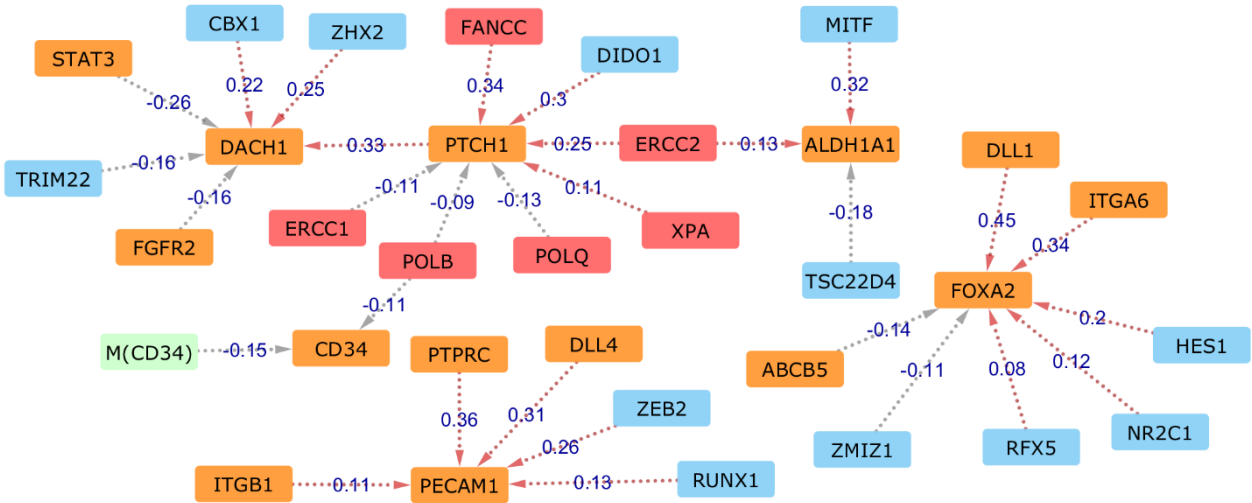
## Hedgehog Signaling



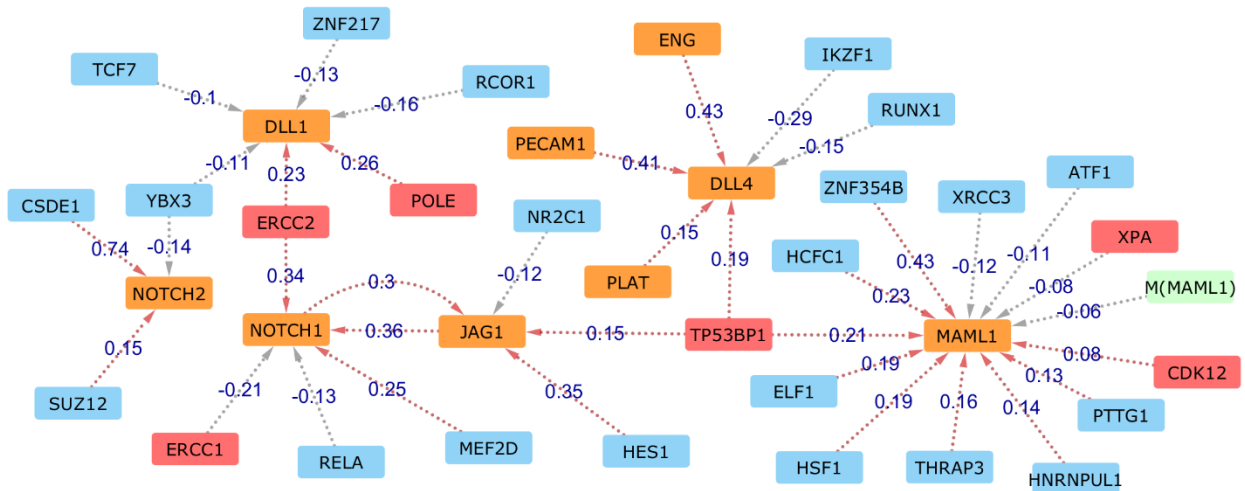
## Hippo Signaling



## Loss of Stemness

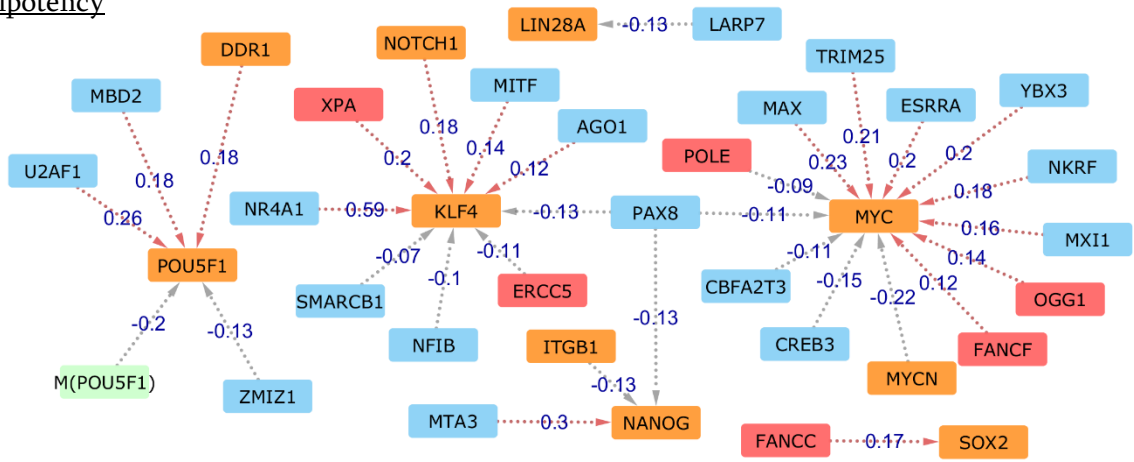


## Notch Signaling

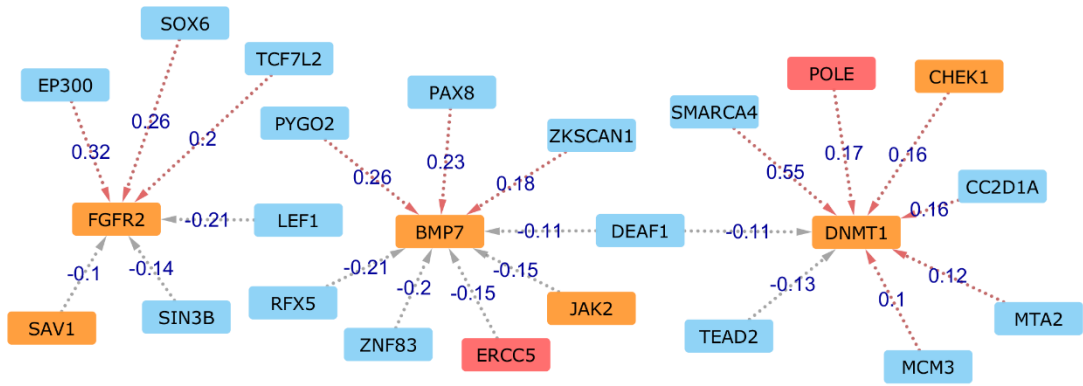




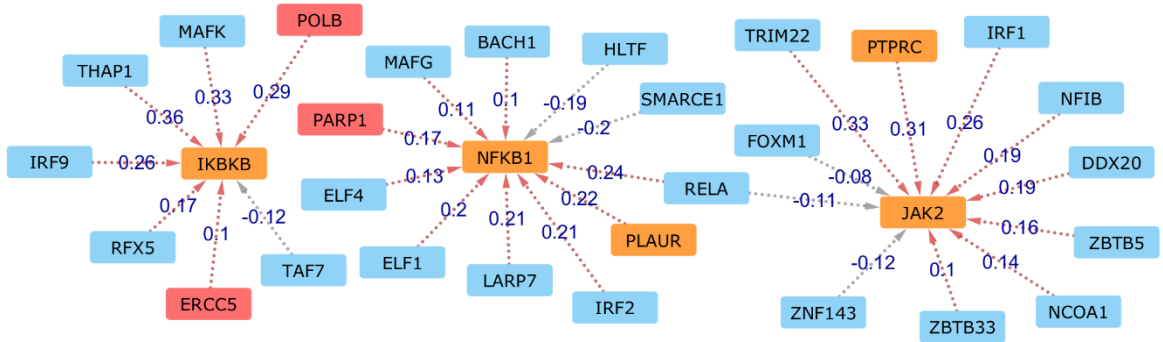
Pluripotency



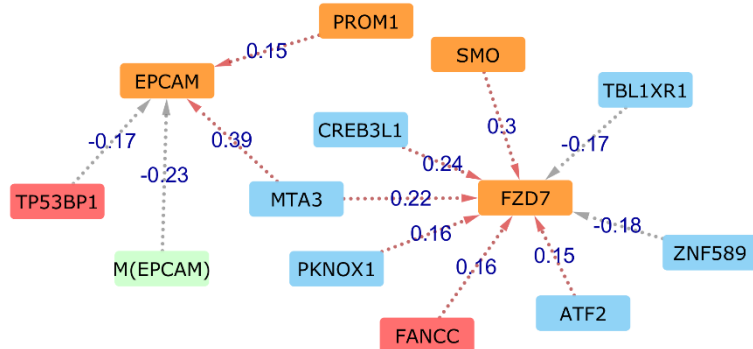
Self-Renewal



STAT-NFkB Signaling

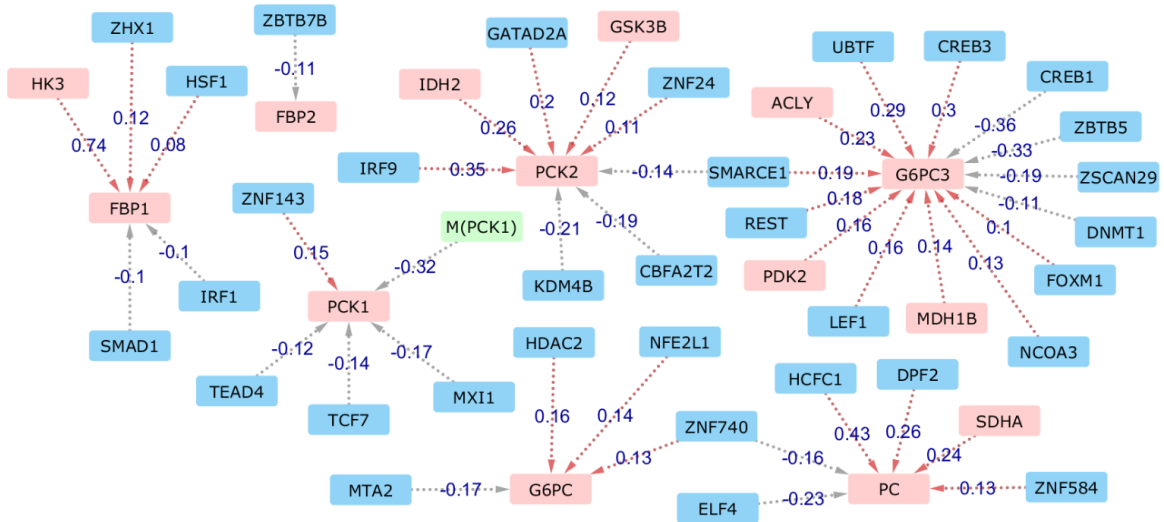


WNT Signaling

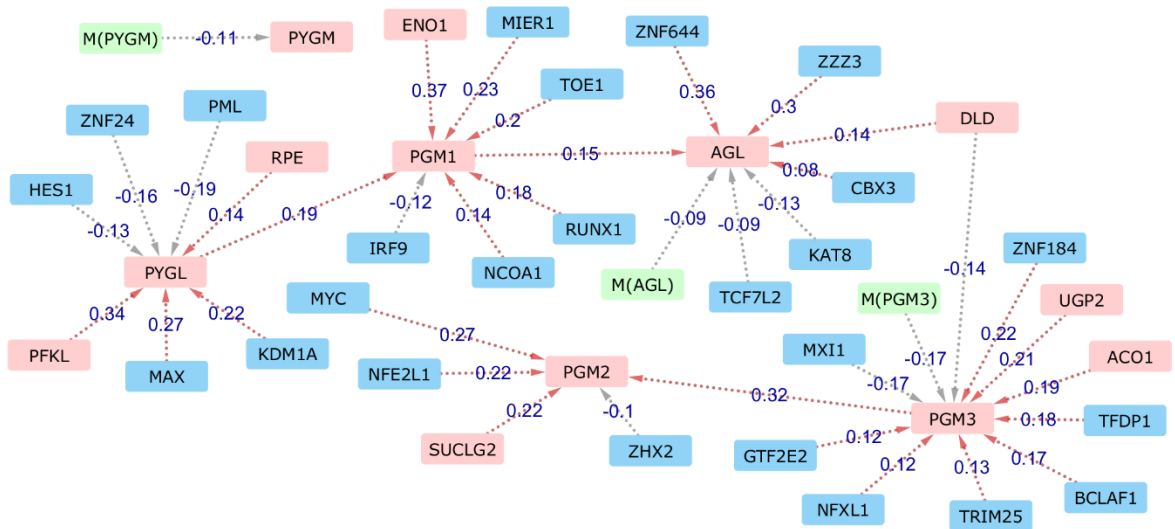


## B.5 GLUCOSE\_METABOLISM (M3)

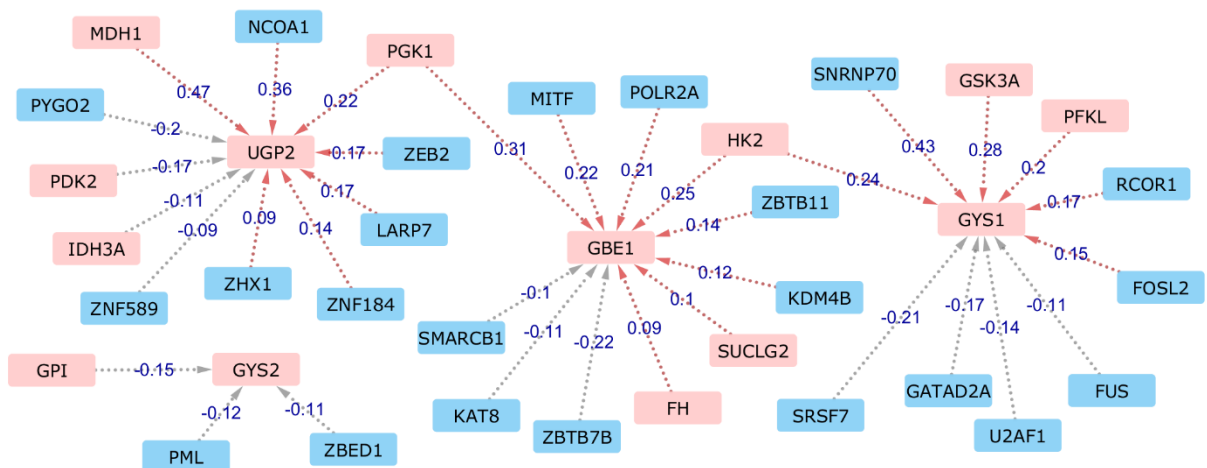
### Gluconeogenesis



### Glycogen Degradation

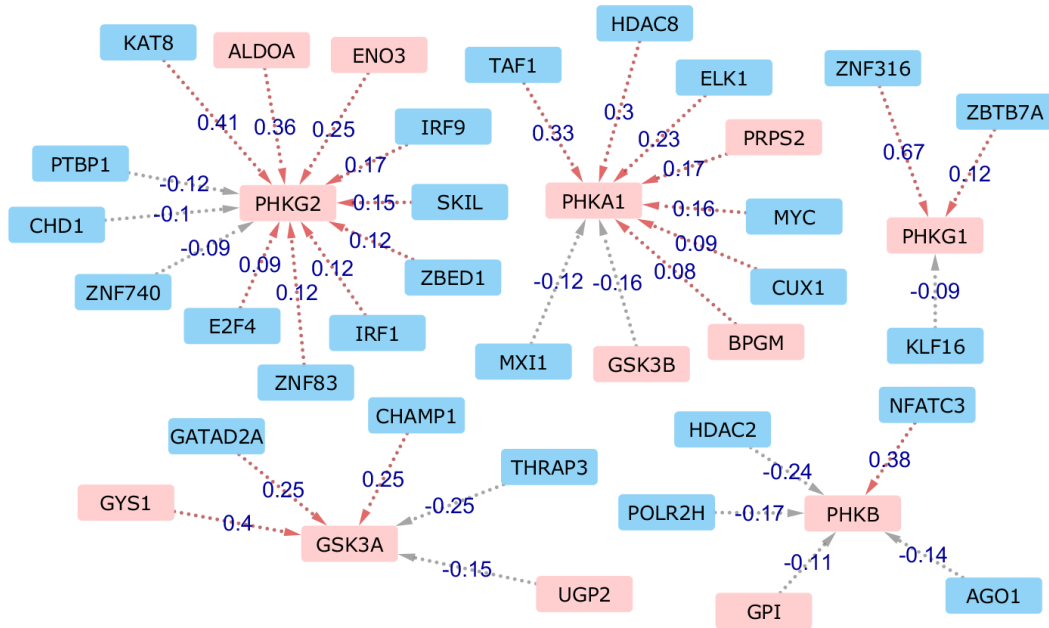


### Glycogen Synthesis



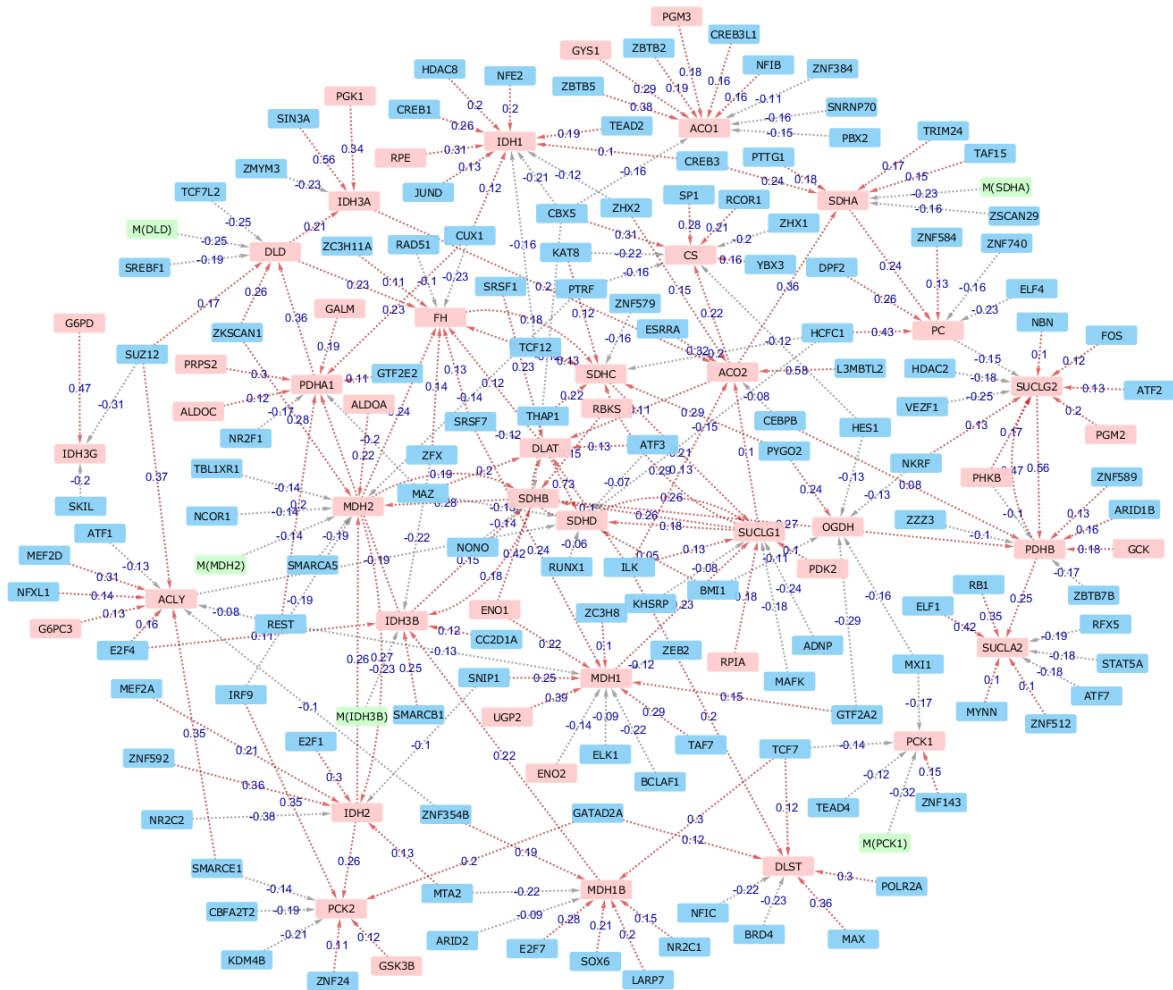


## Regulation of Glycogen Metabolism

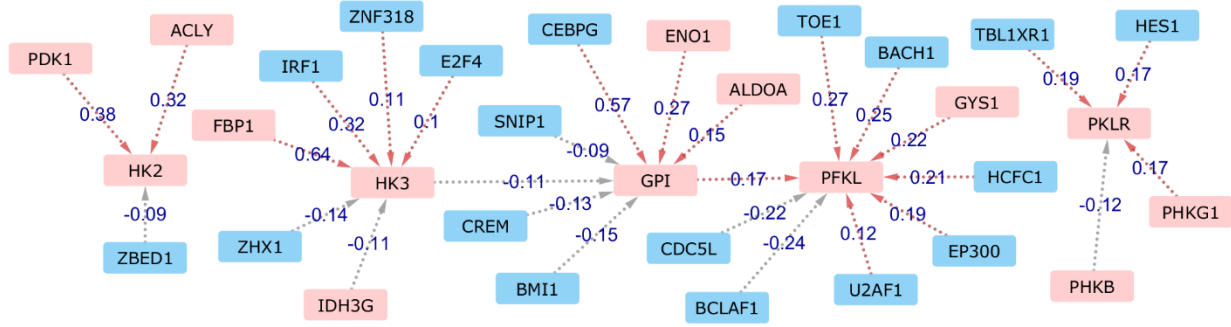


## Tricarboxylic Acid Cycle

This network is almost unreadable on paper, due to the high number of genes of the GLUCOSE\_METABOLISM pathway belonging to the TAC sub-class. The best way to visualize it is using the original network file and opening it with Cytoscape, in order to be able to retrieve all the correlations between the genes set as nodes.

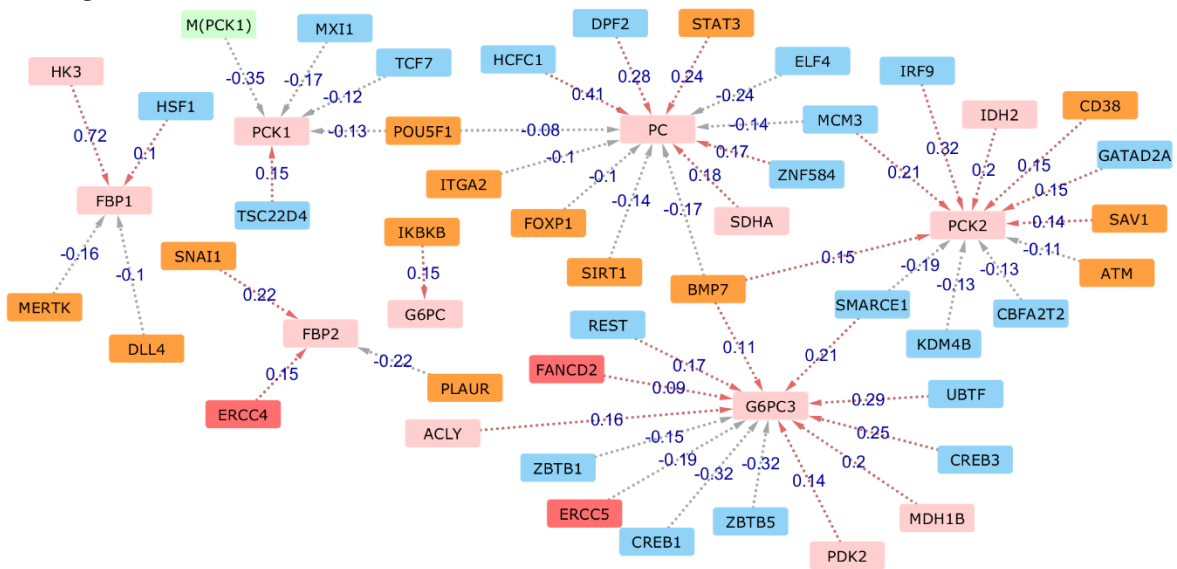


## Unclassified

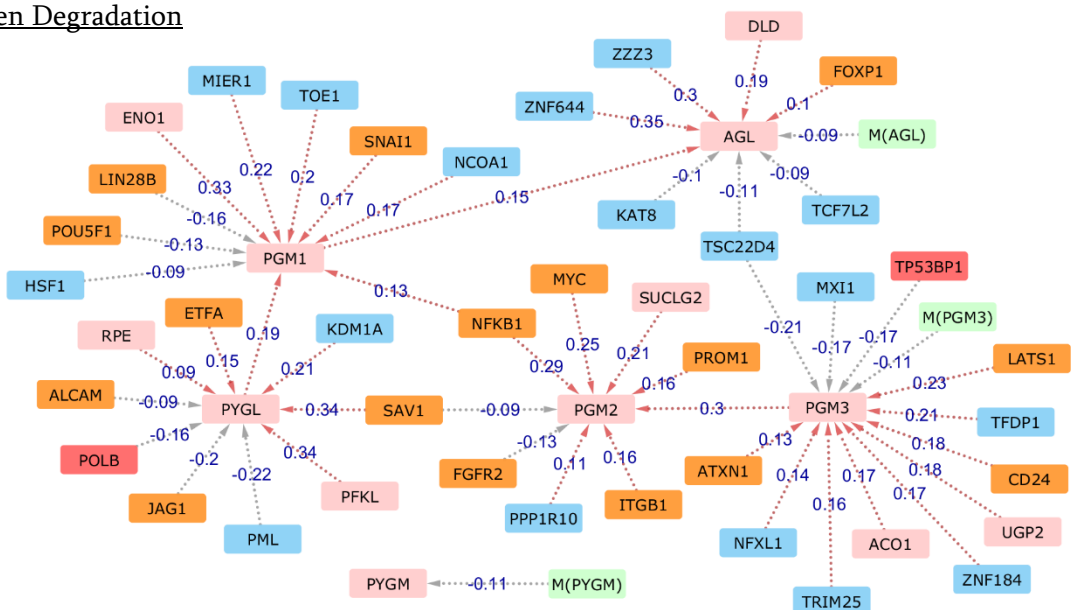


## B.6 GLUCOSE\_METABOLISM (M5)

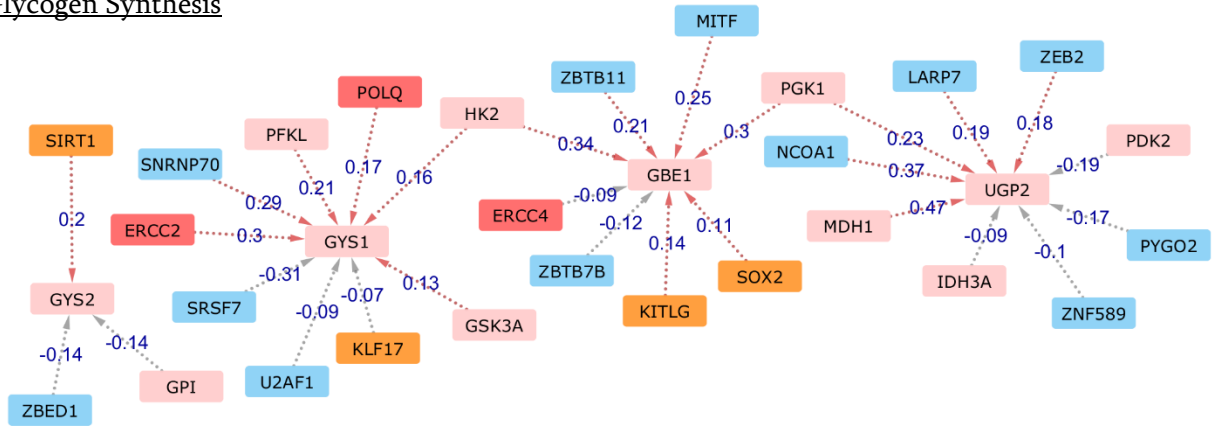
### Gluconeogenesis



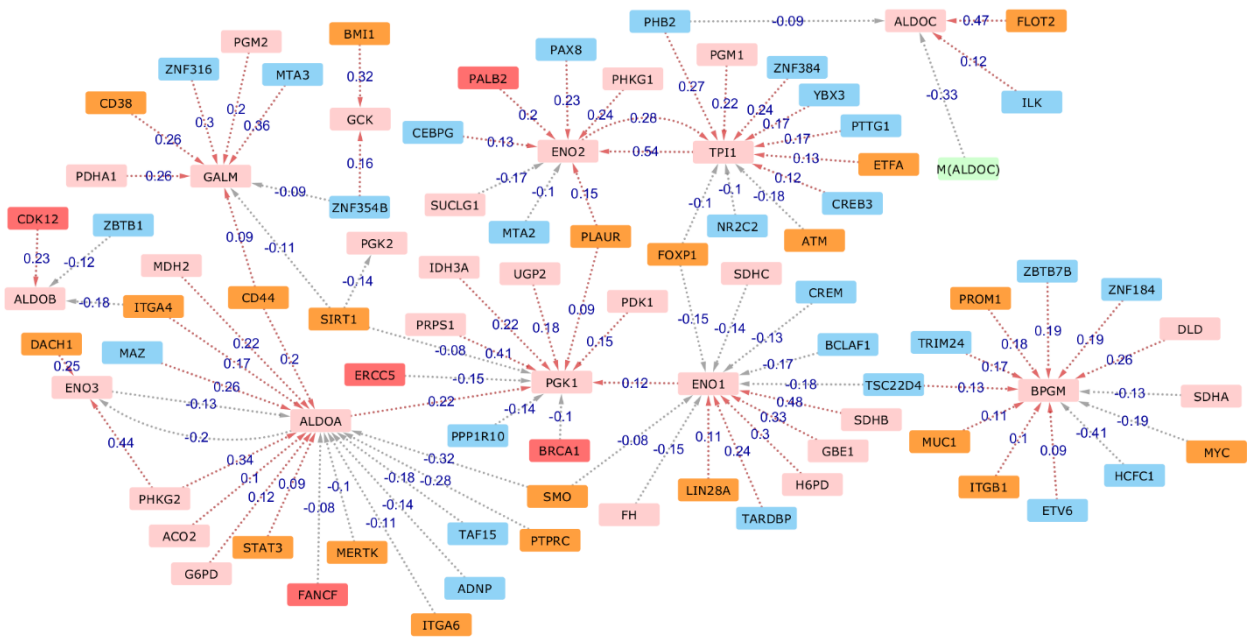
### Glycogen Degradation



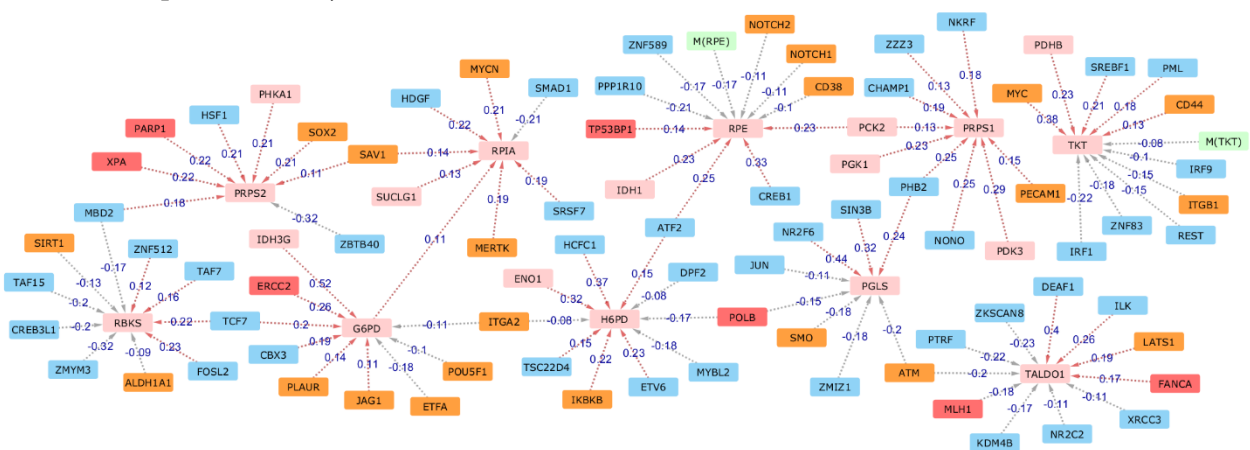
## Glycogen Synthesis



## Glycolysis

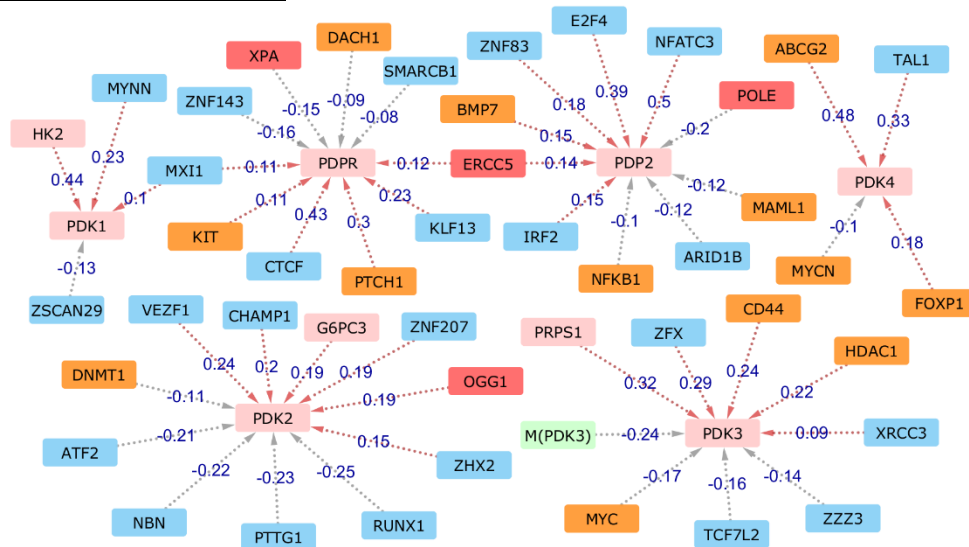


## Pentose Phosphate Pathway

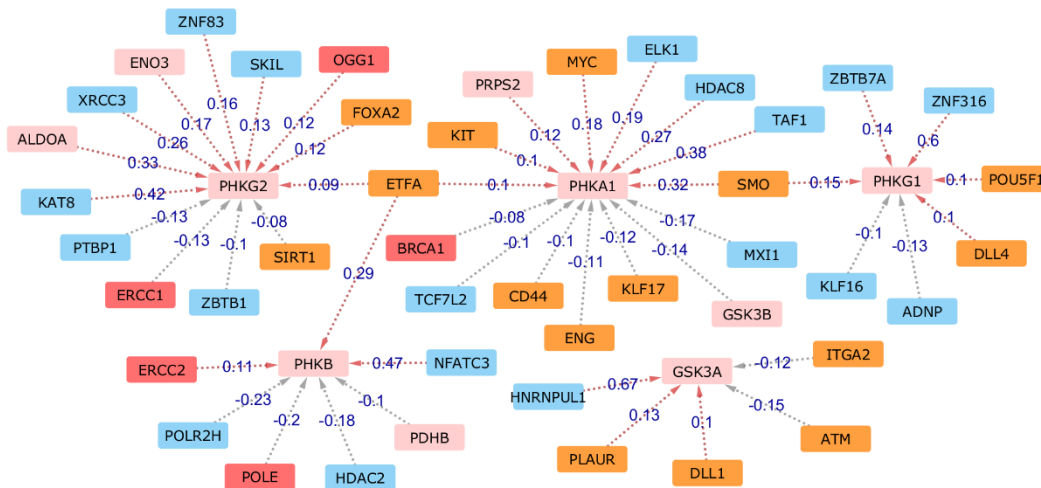




## Regulation of Glucose Metabolism



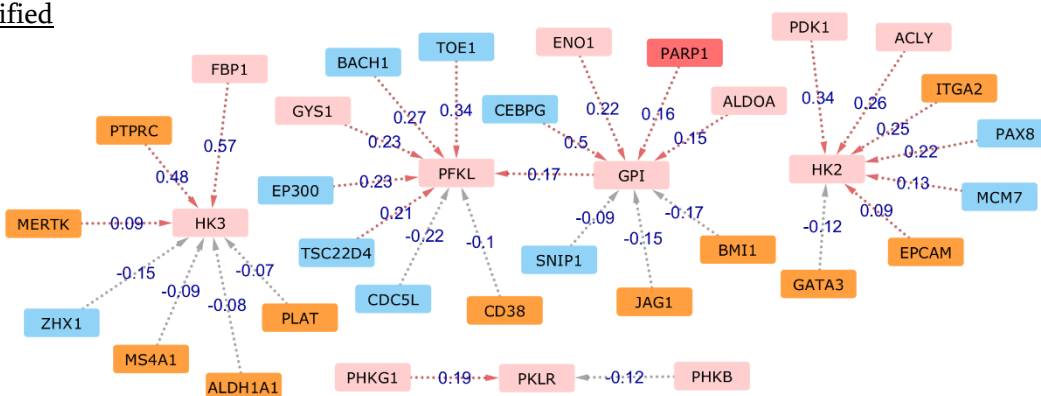
## Regulation of Glycogen Metabolism



## Tricarboxylic Acid Cycle

This network is completely unreadable on paper: what said for model M3 holds also for model M5 and its resulting network, which comprises more genes and a higher number of correlations. For this reason, TAC sub-class network for model M5 is not reported.

## Unclassified







*<< Whenever we think we know the future, even for a second, it changes.  
Sometimes the future changes quickly and completely.  
And we're left only with the choice of what to do next.  
We can choose to be afraid of it, to stand there, trembling,  
not moving, assuming the worst that can happen.  
Or we step forward, into the unknown, and assume it will be brilliant. >>*

Cristina