UNIVERSITY OF SÃO PAULO
RIBEIRÃO PRETO MEDICAL SCHOOL (FMRP)

DOCTORAL THESIS

Bioinformatic tool to integrate and understand aberrant epigenomic and genomic changes associated with cancer

Methods, development and analysis

By

TIAGO CHEDRAOUI SILVA

November 2017

Bioinformatic tool to integrate and understand aberrant epigenomic and genomic changes associated with cancer: Methods, development and analysis

Tiago Chedraoui Silva

Advisor: **Prof. Dr. Houtan Noushmehr**

Doctoral dissertation submitted to the *Ribeirão Preto Medical School* (FMRP/USP), in partial fulfillment of the requirements for the degree of the Doctorate Specialization in Normal and Neoplastic Cell Differentiation. *EXAMINATION BOARD PRESENTATION COPY*.

USP – Ribeirão Preto
November 2017

Ficha catalográfica

Silva, Tiago Chedraoui
S586b Ferramenta de bioinformática para integrar e compreender as mudanças epigenômicas e genômicas aberrantes associadas com câncer: métodos, desenvolvimento e análise / Tiago Chedraoui Silva; orientador Houtan Noushmehr.- Ribeirão Preto - SP, 2017.
105 p.

Tese Doutorado (Doutorado - Especialização em Diferenciação celular normal e neoplásicas) - Faculdade de medicina de Ribeirão Preto (FMRP/USP), USP, 2017.

1. Câncer. 2. metilação do DNA. 3. redes reguladoras de genes. 4. melhoradores. 5. interações de cromatina. 6. sitios de ligação do fator de transcrição. 7. epi-genética. 8. ferramentas computacionais. I. Noushmehr, Houtan, Orientador.

Tiago Chedraoui Silva

Bioinformatic tool to integrate and understand aberrant epigenomic and genomic changes associated with cancer: Methods, development and analysis

Doctoral dissertation submitted to the *Ribeirão Preto Medical School* (FMRP/USP), in partial fulfillment of the requirements for the degree of the Doctorate Specialization in Normal and Neoplastic Cell Differentiation. *EXAMINATION BOARD PRESENTATION COPY.*

Approved in Ribeirão Preto – SP, _____/_____/_____:

Examining Committee

Prof.Dr.: _____ Institution: _____

Verdict: _____ Signature: _____

Prof.Dr.: _____ Institution: _____

Verdict: _____ Signature: _____

Prof.Dr.: _____ Institution: _____

Verdict: _____ Signature: _____

Prof.Dr.: _____ Institution: _____

Verdict: _____ Signature: _____

To my family who always loved me and gave me all the support to achieve my dreams.

ACKNOWLEDGEMENTS

First of all, I would like to thank all my friends, from my undergraduate degree at Unicamp, from my double diploma at Telecom ParisTech, from my year of doctorate sandwich at Cedars-Sinai, from the handball teams. I ask my sincere excuses if I have failed them during these last years, but I will eternally thank for all the support, for all the laughter, for all the hugs, for all understanding in those my depressive days.

I would like to thank my parents for all their support in these moments of indecision, for being more understanding with each passing day, for always supporting me to pursue my dreams and for all the love they have always given me. Thanks even for the moments of scolding, because I know they were all to make me a better person.

I also would like to thank my sister for teaching me to never doubt yourself, for all conversations, for all trips together, to be a great friend and a wonderful sister.

Notably, I would like to thank my great friend, Dr. Hona, for supporting my family these last years. I still remember the first day I met her. She was in a wheelchair, her health was not good, despite all this she had that incredible smile on her face. With whom I have had the most profound conversations, she has always given me her extremely knowledgeable advice. I do not consider only a great friend, but an enormous inspiration as a person. I thank her for reminding me that the noblest thing a man can do in his life is to help others; that happiness lies not in the great moments of life, but in those little quotidian moments that we forget to enjoy. In these darkest moments of humanity, people like her brought me hope of a better world. Her recent passing has brought us sad times, but she will forever remain a part of our lives. Thank you so much for everything.

I would like to thank Dr. Bernardo Mantovani for being a great inspiration and for teaching me the importance of the Public University in society. I'm also grateful to my scientific initiation mentor, Dra. Islene Calciolari, for all the support when everything went wrong, and for being an excellent teacher and friend.

Finally, I also would like to thank FAPESP for the financial support, which was fundamental to carry out all the research activities. And to my advisers Dr. Houtan Noushmehr and Dr. Benjamin Berman for all guidance and support during this project.

"Our greatest weakness lies in giving up.
The most certain way to succeed is always
to try just one more time. "
(Thomas A. Edison)

Resumo

SILVA, T. C.. **Ferramenta de bioinformática para integrar e compreender as mudanças epigenômicas e genômicas aberrantes associadas com câncer: métodos, desenvolvimento e análise**. 2017. 105 f. Tese Doutorado (Doutorado – Especialização em Diferenciação celular normal e neoplásicas) – Faculdade de medicina de Ribeirão Preto, Ribeirão Preto – SP.

O câncer configura uma das maiores causas de mortalidade no mundo, caracterizando-se como uma doença complexa orquestrada por alterações genômicas e epigenômicas capazes de alterar a expressão gênica e a identidade celular. Nova evidência obtida por meio de um estudo genômico em larga escala e cujos dados encontram-se disponíveis no banco público do TCGA sugere que um em cada dez pacientes portadores de câncer pode ser classificado com maior eficácia tendo como base a taxonomia molecular quando comparada à histologia. Dessa maneira, nós hipotetizamos que o estabelecimento de mapas genômicos exibindo a localização de sítios de ligação de fatores de transcrição combinada à identificação de regiões diferencialmente metiladas e perfis alterados de expressão gênica possa nos auxiliar a caracterizar e explorar, ao nível molecular, fenótipos associados ao câncer.

Avanços tecnológicos e bancos de dados públicos a exemplo do The Cancer Genome Atlas (TCGA), The Encyclopedia of DNA Elements (ENCODE) e o NIH Roadmap Epigenomics Mapping Consortium (Roadmap) têm proporcionado um recurso inestimável para interrogar o (epi)genoma de linhagens de células tumorais em cultura, bem como de tecidos normais e tumorais em alta resolução. Todavia, a informação biológica encontra-se armazenada em diferentes formatos e não há ferramentas computacionais para integrar esses dados, evidenciando um cenário atual que requer, com urgência, o desenvolvimento de ferramentas de bioinformática e softwares capazes de direcionar a solução deste obstáculo. Nesse contexto, o objetivo principal deste estudo consiste em implementar o desenvolvimento de ferramentas de bioinformática, na linguagem de programação R que, ao final do estudo, será submetido à comunidade científica do projeto Bioconductor sob a licença de código aberto GNU GPL versão 3. Além disso, ajudaremos nossos colaboradores com o aperfeiçoamento do ELMER, um pacote R/Bioconductor que identifica elementos reguladores usando dados de expressão gênica, de metilação do DNA e análise de motivo.

Nossa expectativa é que essas ferramentas possam automatizar com acurácia a pesquisa, o download e a análise dos dados (epi)genômicos que se encontram atualmente disponíveis nas bases de dados públicas dos consórcios internacionais TCGA, ENCODE e Roadmap, além de integrá-los facilmente aos dados genômicos e epigenômicos gerados por pesquisadores por meio de experimentos em larga escala. Além disso, realizaremos também o processamento e a análise manual dos dados que serão automatizados pelas ferramentas, visando validar sua capacidade em descobrir assinaturas epigenômicas que possam redefinir subtipos de câncer. Por

fim, as usaremos para investigar as diferenças moleculares entre dois subgrupos de gliomas recentemente descobertos por nosso laboratório.

Palavras-chave: Câncer, metilação do DNA, redes reguladoras de genes, melhoradores, interações de cromatina, sitios de ligação do fator de transcrição, epi-genética, ferramentas computacionais.

Abstract

SILVA, T. C.. **Bioinformatic tool to integrate and understand aberrant epigenomic and genomic changes associated with cancer: Methods, development and analysis**. 2017. 105 f. Doctoral dissertation (Doctorate Candidate – Specialization in Normal and Neoplastic Cell Differentiation) – *Ribeirão Preto Medical School* (FMRP/USP), Ribeirão Preto – SP.

Cancer, which is one of the major causes of mortality worldwide, is a complex disease orchestrated by aberrant genomic and epigenomic changes that can modify gene regulatory circuits and cellular identity. Emerging evidence obtained through high-throughput genomic data deposited within the public TCGA international consortium suggests that one in ten cancer patients would be more accurately classified by molecular taxonomy versus histology. Therefore, we have hypothesized that the establishment of genome-wide maps of the de novo DNA binding motifs localization coupled with differentially methylated regions and gene expression changes might help to characterize and exploit cancer phenotypes at the molecular level.

Technological advances and public databases like The Cancer Genome Atlas (TCGA), The Encyclopedia of DNA Elements (ENCODE), and The NIH Roadmap Epigenomics Mapping Consortium (roadmap) have provided unprecedented opportunities to interrogate the epigenome of cultured cancer cell lines as well as normal and tumor tissues with high resolution. Markedly however, biological information is stored in different formats and there is no current tool to integrate the data, highlighting an urgent need to develop bioinformatic tools and/or computational softwares to overcome this challenge. In this context, the main purpose of this study is the development of bioinformatics tools in R programming language that will be submitted to the larger open-source Bioconductor community project under the GNU GPL3 (General Public License version 3). Also, we will help our collaborators improve of the R/Bioconductor ELMER package that identifies regulatory enhancers using gene expression, DNA methylation data and motif analysis.

Our expectation is that these tools can effectively automate search, retrieve, and analyze the vast (epi)genomic data currently available from TCGA, ENCODE, and Roadmap, and integrate genomics and epigenomics features with researchers own high-throughput data. Furthermore, we will also navigate through these data manually in order to validate the capacity of these tools in discovering epigenomic signatures able to redefine subtypes of cancer. Finally, we will use them to investigate the molecular differences between two subgroups of gliomas, one of the most aggressive primary brain cancer, recently discovered by our laboratory.

Key-words: Cancer, DNA methylation, gene regulatory networks, enhancers, chromatin interactions, transcription factor binding sites, epi-genetics, computational tools.

LIST OF FIGURES

Figure 1 – Granges object . . . 17
Figure 2 – Sumarized Experiment object . . . 18
Figure 3 – Example of survival curve. . . . 27
Figure 4 – Overview of TCGAbiolinks functions. . . . 33
Figure 5 – TCGAbiolinks download summary . . . 37
Figure 6 – TCGAbiolinksGUI: The volcano plot menu . . . 39
Figure 7 – TCGAbiolinksGUI: Gene expression download: . . . 39
Figure 8 – TCGAbiolinksGUI: Visualizing mutation summary . . . 41
Figure 9 – TCGAbiolinksGUI: Enrichement analysis of genes . . . 41
Figure 10 – TCGAbiolinksGUI: Visualizing DMR results as heatmap . . . 42
Figure 11 – TCGAbiolinksGUI: Visualizing mutation as an oncoplot . . . 42
Figure 12 – TCGAbiolinksGUI: Survival analysis . . . 43
Figure 13 – ELMER workflow . . . 47
Figure 14 – Supervised vs unsupervised mode . . . 50
Figure 15 – Case study - LGG downstream analysis with gene expression . . . 54
Figure 16 – Case study - Integrative data analysis of Colon Adenocarcinoma . . . 56
Figure 17 – Schematic plot gene-probe pairs . . . 64
Figure 18 – Scatter plot for significant probe (cg04723436) gene (GATA3) pair. . . . 65
Figure 19 – Heatmap of anticorrelated pairs of DNA methylation probes and gene . . . 66
Figure 20 – Enrichment of paired probes and chromatin states of encode cells. . . . 68
Figure 21 – Motif enrichment plot . . . 70
Figure 22 – TF ranking plot: ANDR motif . . . 73
Figure 23 – MCF7 ChIA-PET validation . . . 73
Figure 24 – G-CIMP analysis: Volcano plot . . . 77
Figure 25 – G-CIMP analysis: Heatmap of paired probes and distal genes . . . 78
Figure 26 – G-CIMP analysis: Odds Ratio plot . . . 79
Figure 27 – HOCOMOCO: HOX-related factors family . . . 81

LIST OF SOURCE CODES

Source code 1 – "Selection of probes within biofeatures" 57
Source code 2 – "Create MultiAssayExperiment" 58
Source code 3 – "Verifying MultiAssayExperiment" 58
Source code 4 – "Identify significantly different DNA methylation probes in tumor and normal samples" 59
Source code 5 – "Identify putative target genes for differentially methylated distal probes" 62
Source code 6 – "Schematic plot to visualize gene-probe pairs" 63
Source code 7 – "Scatterplot to visualize correlation between gene expression and DNA methylation levels at probe" 63
Source code 8 – "Heatmap to visualize gene-probe pairs" 67
Source code 9 – "Motif enrichment analysis on the selected probes" 69
Source code 10 – "Identifying regulatory Transcript Factors" 69

LIST OF TABLES

Table 1 – Translational effect of variant allele . . . 10
Table 2 – Types of DNA Mutations and Their Impact . . . 12
Table 3 – Example TCGA barcode . . . 16
Table 4 – Histone and epigenomic marks . . . 21
Table 5 – Hypothesis tests . . . 23
Table 6 – Type I and II Errors. $\alpha = P(\text{Type I Error})$, $\beta = P(\text{Type II Error})$ . . . 23
Table 7 – Summary of parametric and nonparametric procedures. . . . 26
Table 8 – Survival time example . . . 27
Table 9 – Kaplan-Meier method example . . . 27
Table 10 – Comparing TCGAbiolinks to competing software . . . 36
Table 11 – Main differences between ELMER old version (v.1) and the new version (v.2) 45
Table 12 – Identification of distal probes with significant differential DNA methylation (i.e. DMCs) . . . 60
Table 13 – Identification of putative target gene(s) for differentially methylated distal probes 62
Table 14 – Identification of master regulator Transcription Factors (TF) for each enriched motif . . . 71
Table 15 – Candidate regulatory TFs for each molecular subtype found in a pairwise comparison. . . . 75
Table 16 – G-CIMP analysis: Sample summary . . . 76
Table 17 – G-CIMP analysis: ELMER arguments values . . . 76
Table 18 – G-CIMP analysis: Summary results . . . 76
Table 19 – G-CIMP analysis: TF ranking plot . . . 80

LIST OF ABBREVIATIONS AND ACRONYMS

5mc 5-methylcytosine
ANOVA .. Analysis of variance
AUC ..... Area Under the Curve
AWGs .... Analysis working groups
BAM ..... Binary Alignment/Map
CBS ..... Circular Binary Segmentation
CCG ..... Center for Cancer Genomics
ChIP-seq . ChIP-sequencing
CNP ..... Copy-number polymorphism
CNV ..... copy number variation
COAD ... Colon adenocarcinoma
DEA ..... differential expression analysis
DEGs .... differentially expressed genes
DMCs .... differentially methylated CpGs
DMR ..... Differentially methylated regions
DNA ..... deoxyribonucleic acid
Dnmts .... DNA methyltransferases
ELMER .. Enhancer Linking by Methylation/Expression Relationship
ENCODE . The Encyclopedia of DNA Elements
EPIC ..... MethylationEPIC BeadChip
FC ..... fold change
FDR ..... false discovery rate
FIMO .... Find Individual Motif Occurrences
FOXA1 ... Forkhead box protein A1
FPKM .... Fragments Per Kilobase of transcript per Million mapped reads
FWER ... Family-wise error rate
GATA3 ... GATA-binding protein 3
GBMs .... Glioblastomas
GDC ..... Genomic Data Commons
GHRH ... growth hormone-releasing hormone
GNU GPL3 GNU General Public License version 3

GRanges . GenomicRanges
GRCh36 .. Genome Reference Consortium Human Build 36
GRCh37 .. Genome Reference Consortium Human Build 37
GRCh38 .. Genome Reference Consortium Human Build 38
GSEA .... gene set enrichment analysis
GUI ..... Graphical User Interface
H3K36 ... histone H3 lysine 9
H3K4me3 H3K4 trimethylation
H3K9 .... histone H3 lysine 9
HM450 ... HumanMethylation450 BeadChip
HOCOMOCO HOmo sapiens COmprehensive MOdel COllection
HOMER .. Hypergeometric Optimization of Motif EnRichment
HOXD13 . Homeobox D13
HOXD3 .. Homeobox D3
KW ..... Kruskal-Wallis test
LGG ..... lower-grade glioma
MAE ..... MultiAssayExperiment
MAF ..... Mutation Annotation Format
MWU .... Mann-Whitney U-test
NGS ..... next-generation sequencing
NHGRI ... National Human Genome Research Institute
NIH ..... National Institute of Health
OR ..... Odds Ratio
P63 ..... tumor protein p63
PI3K ..... Phosphatidylinositol 3-kinase
PPIN ..... Protein-protein interaction networks
PPIs ..... Protein-protein interactions
PTEN .... Phosphatase and tensin homolog
PWM .... position weight matrix
qCML .... quantile-adjusted conditional maximum likelihood
RAS ..... Rat Sarcoma
RB ..... retinoblastoma-associated
ROADMAP The NIH Roadmap Epigenomics Mapping Consortium
ROC ..... Receiver Operator Characteristics
RPM ..... Reads per Millions
SAM ..... S-adenyl methionine
SE ..... SummarizedExperiment

SNP ..... single nucleotide polymorphism
SOX2 .... Sex determining region Y-box 2
TARGET . Therapeutically Applicable Research to Generate Effective Treatments
TCGA .... The Cancer Genome Atlas
TF ..... Transcription Factor
TFBSs ... Transcription factor binding sites
TSSs ..... transcription start sites
WGBS ... whole-genome bisulfite sequencing

LIST OF SYMBOLS

H_0 — Null hypothesis

H_1 — Alternative hypothesis

α — Significance level

β — rate of the type II error

$S(t)$ — Kaplan–Meier estimator

CONTENTS

Page

Acknowledgements ix

Resumo xi

Abstract xiii

List of Figures xvii

List of Charts xvii

List of Algorithms xvii

List of Source codes xix

List of Tables xxi

List of Abbreviation and acronyms xxv

List of Symbols xxvii

1 Introduction 1
1.1 Objectives 1
1.1.1 Specific aims 1
1.2 Motivations 2
1.3 Organization of the Remainder of the Document 3

2 Review of essential concepts 5
2.1 Cancer 5
2.2 Genetics and epigenetics alterations 7
2.2.1 DNA methylation 7
2.2.2 Histone modifications 8
2.2.3 Genetics alterations 9
2.3 Databases and data structure 13
2.3.1 Databases 13
2.3.2 Data and data structure 14
2.3.2.1 Data 15

2.3.2.2 Data structures . . . 16
2.4 (Epi)Genomic data analysis . . . 17
2.4.1 Integrative analysis . . . 17
2.4.1.1 Functional enrichment analysis of cancer signatures . . . 19
2.4.1.2 Protein interaction networks and cancer signatures . . . 19
2.4.1.3 Transcriptional targets and cancer signatures . . . 19
2.4.1.4 Human Epigenomes . . . 20
2.4.2 Statistical analysis . . . 22
2.4.2.1 Hypothesis Testing . . . 22
2.4.2.2 Making a decision: P-value approach . . . 23
2.4.2.3 Correcting for multiple testing . . . 24
2.4.2.4 Nonparametric and parametric tests . . . 25
2.4.3 Survival analysis . . . 26
2.4.3.1 Kaplan–Meier method: Estimating the survival curve . . . 26
2.4.3.2 Comparing survival curves of two groups using the log-rank test 27
2.4.4 Machine Learning . . . 28
2.4.4.1 Supervised learning . . . 28
2.4.4.2 Unsupervised learning . . . 29

3 Development of methods and softwares for cancer data analysis . . . 31
3.1 TCGAbiolinks: An R/Bioconductor package to download and analyze data from GDC . . . 31
3.1.1 The TCGAbiolinks package . . . 32
3.1.2 Comparisons . . . 35
3.1.3 Software availability . . . 37
3.1.4 Public reception . . . 37
3.2 TCGAbiolinksGUI: A graphical user interface to analyze GDC cancer molecular and clinical data . . . 38
3.2.1 Infrastructure . . . 38
3.2.2 Graphical user interface design . . . 38
3.2.3 Documentation . . . 40
3.2.4 Docker container . . . 40
3.2.5 Comparison of alternative software . . . 43
3.3 Enhancer Linking by Methylation/Expression Relationships (ELMER) 44
3.3.1 Implementation . . . 46
3.3.1.1 Organization of data as a MultiAssayExperiment object . . . 46
3.3.1.2 Selecting distal probes . . . 48
3.3.1.3 Identification of differentially methylated CpGs (DMCs) . . . 48
3.3.1.4 Identification of putative target gene(s) . . . 49

3.3.1.5 Characterization of chromatin state context of enriched probes using FunciVar . . . 50
3.3.1.6 Motif enrichment analysis . . . 50
3.3.1.7 Identification of master regulator TFs . . . 51

4 Cancer data analysis . . . **53**
4.1 Use cases using TCGAbiolinks . . . 53
4.1.1 Lower-grade glioma downstream analysis with gene expression 53
4.1.2 Downstream analysis integration of gene expression and methylation data . . . 54
4.2 Use cases using ELMER . . . 55
4.2.1 Breast Invasive Carcinoma (unsupervised approach) . . . 55
4.2.2 BRCA molecular subtypes analysis (supervised approach) . . . 72
4.3 Glioma analysis . . . 74

5 Conclusion . . . **83**
5.1 Conclusions . . . 83
5.2 Conclusions and future studies . . . 83
5.2.1 Conclusions and future works of TCGAbiolinks . . . 83
5.2.2 Conclusions and future works of TCGAbiolinksGUI . . . 84
5.2.3 Conclusions and future works of ELMER . . . 84
5.3 Publications, presentations and softwares of the Doctorate Period . . . 84
5.3.1 First-authored papers . . . 85
5.3.2 Co-authored papers . . . 86
5.3.3 First-authored softwares . . . 87
5.3.4 Co-authored softwares . . . 87
5.3.5 Workshops and workflows . . . 87
5.3.6 Conferences & presentations . . . 88

Bibliography . . . **89**

APPENDIX A Dispensa comitê de ética . . . **105**

CHAPTER

1

INTRODUCTION

1.1 Objectives

The main goal of this project is to develop tools for searching, retrieving and analyzing pan-cancer genomic data from several databases, such as the NCI's Genomic Data Commons (GDC), which contains data from the The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET), The Encyclopedia of DNA Elements (ENCODE), and The NIH Roadmap Epigenomics Mapping Consortium (ROADMAP). For a better transparency, all tools will be open source and for their scalability and interoperability they will be published in the Bioconductor project, an environment that provides a broad range of powerful statistical and graphical methods for the analysis of genomic data. Furthermore, We also aim to investigate the intergenic epigenomic changes associated with distinct biological and clinical subgroups of gliomas first discovered by our laboratory (CECCARELLI *et al.*, 2016b). Specifically, using the tools developed, we will integrate DNA methylation, gene expression, mutation and copy number data as well as important epigenomic marks defined by histone modifications in normal samples in order to identify candidate regulatory elements associated with glioma progression.

1.1.1 Specific aims

1. Download and process transcription factor (TF) ChIP-seq data for each cancer cell and tissue type through the ENCODE dataset;

2. Download and process DNA methylation data for both cancer and non-tumor control cases through the TCGA consortium via HM450K platform;

3. Identify statistically Differentially methylated regions (DMR) at the single CpG resolution;

4. Determine statistically enriched proximal transcription factor binding sites (TFBSs) to altered DNA-methylated regions at the level of individual DNA/protein site interaction;

5. Within known DMRs, classify and identify statistically known and novel DNA binding motifs;
6. Download and process RNA-seq data from both cancer and non-tumor control cases;
7. Use standard data structure to organize the data and the metadata;
8. Correlate the DNA methylation status of Transcription factor binding sites (TFBSs) with target gene RNA-seq expression in order to determine regulatory networks that might alter the pan-cancer genome;
9. Use learning machine algorithms for classifying an independent set of gliomas based on newly identified regulatory networks as related to pan-cancer deregulation;
10. Develop tools to automate the previous steps;
11. Use those tools to investigate the intergenic epigenomic changes associated with distinct biological and clinical subgroups of gliomas g-cimp-low and g-cimp-high discovered by our laboratory and collaborators;
12. Compare the automated results with ones found manually in order to validate the package capacity in providing searching, retrieving and downstream biological analysis to discover pan-cancer epigenomic signatures able to redefine subgroups of gliomas;
13. Submit those set of tools to be freely available in the open-source Bioconductor environment (available at http://www.bioconductor.org).

1.2 Motivations

Unravelling the genomic, epigenomic, and proteomic features using high-throughput methodologies is a central question for understanding regulatory gene networks in cancer. In this line of evidence, thousands of tumor and normal samples have been massively sequenced and a large amount of data are publicly deposited by the three main international consortia: TCGA, ENCODE, and Roadmap. However, a major challenge is the fact that the biological information necessary for a complete gene regulatory analysis is spread over different databases that store data in different formats. Moreover, to date there is a lack of computational tools and methods that can integrate and interpret such information. Consequently, the process of analysis is performed manually by the end user, who must access all databases, select and process the data necessary to the project, and integrate that data using multiple downstream analysis tools to extract and interpret the relevant biological information. To overcome these limitations, here we propose to implement tools for searching genomic and epigenomic data acquired from several biological databases, and to provide key scientific analysis steps and methodology, thereby allowing other researchers to apply strategies for in-depth bioinformatics analysis. These tools

will be submitted to Bioconductor, and then, our expectation is that researchers may integrate all relevant data from the most important international consortia in the genomics field with their own experiments. Providing our package through Bioconductor, enables us to access a broader scientific community of advanced informatics users and developers worldwide, who can test the package with their own microarray and next-generation sequencing data, submit bug reports, criticize the methodology, provide new contributions, and ensure the quality of our package in terms of code and documentation. In addition, storing the outputs inside an open-source software like R, allows one to utilize the many available statistical and analytical packages commonly used by researchers (Editorial Nature Genetics, 2014). Then, we expect that these tools will provide insights and novel discoveries into unanticipated regulatory circuits in complex disease and normal developmental biology, which will be verified through the molecular analysis between the newly identified groups G-CIMP-low and G-CIMP-high.

1.3 Organization of the Remainder of the Document

This paper is organized into chapters as follows: In Chapter 2, fundamental concepts required to the development of this project are introduced. We review the concepts of cancer and their epigenetic and genetic alterations, followed by the description of the biological data generated through experiments conducted to identify these changes and the data structure used to store it, finally we reviewed some of the data analysis methods used in the genetics field. Chapters 3, 4, and 5 highlights the results of this work. Specifically, in Chapter 3, we detail the methods and computational tools developed, while in Chapter 4 we show their application in the real world. Enclosing this thesis, in Chapter 5, we draw the main conclusions of this work, as well as the scientific contributions derived from this project, possible future works and the main papers that have been published during the Doctorate period.

CHAPTER 2

REVIEW OF ESSENTIAL CONCEPTS

This chapter explains some of the fundamental concepts used in the development in this thesis. In section 2.1 cancer introduced. Its characteristics when compared to normal cells, the molecular mechanisms responsible for its onset and those that facilitate its maintenance and proliferation. In section 2.2 the genetic and epigenetic changes in cancer cells are detailed. In section 2.3 the clinical, genetic and epigenetic data publicly available used for cancer data analysis is described. Their computational representation and data structures available to manipulate large collections of data are detailed. Finally, in section 2.4 several techniques of data analysis used to extract biological knowledge from this data is presented.

2.1 Cancer

Tumors are complex tissues composed of multiple distinct cell types that participate in heterotypic interactions with one another. For the onset of tumorigenesis, cancer cells create a "tumor microenvironment" recruiting normal cells which are active participants in tumorigenesis rather than passive bystanders. If compared to normal cells, cancer cells acquired capabilities which includes sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming of energy metabolism and evading immune destruction (HANAHAN; WEINBERG, 2011). Some examples of these acquired capabilities are described below.

Sustaining Proliferative Signaling While normal tissues carefully control the production and release of growth-promoting signals the cancer cells deregulate these signals in ways that promote proliferation. Cancer cells can acquire the capability to sustain proliferative signaling in a number of alternative ways: They may produce growth factor ligands themselves, they may send signals to stimulate normal cells within the supporting tumor-associated stroma, to supply them with various growth factors (CHENG *et al.*, 2008), or they may elevate the levels of receptor proteins displayed on the cancer cell surface, rendering such

cells hyperresponsive to otherwise-limiting amounts of growth factor ligand. Cancer cell also disrupts Negative-Feedback Mechanisms that Attenuate Proliferative Signaling, for example, loss-of-function mutations or promoter methylation of the Phosphatase and tensin homolog (PTEN), a tumor suppressor, amplify Phosphatidylinositol 3-kinase (PI3K) signaling and promote tumorigenesis in a variety of experimental models of cancer (JIANG; LIU, 2009). A second example is the oncogene Rat Sarcoma (RAS), some mutations affecting it compromise Ras GTPase activity, a protein that normally regulates the inactivation of Ras. In addition to the oncogenic mutations in Ras, another mechanism affecting this gene showed silence of the regulatory Ras GTPase proteins through CpG methylation (JIN *et al.*, 2007).

Evading growth suppressors Cancer cells must also circumvent actions of tumor suppressor genes that negatively regulate cell proliferation. One example of tumor suppressor proteins is the retinoblastoma-associated (RB), that integrates signals from diverse extracellular and intracellular sources and, in response, decides whether or not a cell should proceed through its growth-and-division cycle (BURKHART; SAGE, 2008). Cancer cells with defects in RB pathway function are thus missing the services of a critical gatekeeper of cell-cycle progression whose absence permits persistent cell proliferation. Another example is the TP53 proteins that receives inputs from stress and abnormality sensors that function within the cell's intracellular operating systems. This protein can activate apoptosis if it received a signal of irreparable damage to the cellular subsystems, or it can interrupt the cell-cycle progression if it receives signals that the degree of damage to the genome is excessive, or if the levels of nucleotide pools, growth-promoting signals, glucose, or oxygenation are suboptimal (HANAHAN; WEINBERG, 2011). Moreover, disruption of such self-attenuating signaling may contribute to the development of adaptive resistance toward drugs targeting mitogenic signaling. To evade those growth suppressors proteins tumor cells developed mechanisms of contact inhibition and its evasion

Resisting Cell Death The loss of TP53 tumor suppressor function, which eliminates this critical damage sensor from the apoptosis-inducing circuitry, is a tumor cell strategies to limit or circumvent apoptosis (HANAHAN; WEINBERG, 2011).

Those capabilities, which enable tumor growth and metastatic dissemination, were acquired mainly due to a genetic diversity generated by genome instability (HANAHAN; WEINBERG, 2011). A multistep tumor progression can be portrayed as a succession of clonal expansions, which can be triggered by non mutational changes (epigenetic mechanisms such as DNA methylation and histone modifications (BERDASCO; ESTELLER, 2010)) or mutational changes affecting the regulation of gene expression.

To increase the rates of mutation, cancer cells often breakdown components of the genomic maintenance machinery that normally monitor genomic integrity and force genetically

damaged cells into either senescence or apoptosis (JACKSON; BARTEK, 2009), more specifically those involved in detecting DNA damage and activating the repair machinery, in directly repairing damaged DNA, and in inactivating or intercepting mutagenic molecules before they have damaged the DNA (NEGRINI; GORGOULIS; HALAZONETIS, 2010). Another major source of tumor-associated genomic instability is the loss of telomeric DNA in many tumors that generates karyotypic instability and associated amplification and deletion of chromosomal segments (ARTANDI; DEPINHO, 2009). Recurrence of these amplifications and deletions aberrations at particular sites in the genome indicates that they are likely to harbor genes whose alteration favors neoplastic progression (KORKOLA; GRAY, 2010).

2.2 Genetics and epigenetics alterations

While genetics is the study of heritable changes in gene activity or function due to the direct alteration of the DNA sequence such as point mutations, deletions, insertions, and translocation, epigenetics is the study of those not associated with any change of the DNA sequence (MOORE; LE; FAN, 2013). Even though all cells in an organism contain the same genetic information, this epigenetic mechanism controls which gene are expressed enabling the existence of a diversified gene expression profiles in a variety of cells and tissues in multicellular organisms.

2.2.1 DNA methylation

DNA methylation is an epigenetic mechanism involving the transfer of a methyl group to the C5 position of the cytosine to form 5-methylcytosine, which regulates gene expression by either blocking the binding of Transcription factors or recruiting proteins involved in gene repression (MOORE; LE; FAN, 2013). The pattern of DNA methylation in the genome can change by DNA methylation and demethylation process, these processes are involved in cell differentiation as unique DNA methylation pattern is able to regulate tissue-specific gene transcription (MOORE; LE; FAN, 2013).

The DNA methylation process is mediated by the family of DNA methyltransferases (Dnmts) proteins which catalyze the transfer of a methyl group from S-adenyl methionine (SAM) to the fifth carbon of cytosine residue to form 5-methylcytosine (5mc). The main proteins responsible for this transfer are the Dnmt3a and Dnmt3b, which are involved in establishing a new methylation pattern to unmodified DNA, and Dnmt1, which is responsible to the methylation of a daughter strand during the DNA replication process.

The majority of DNA methylation occurs on cytosines that precede a guanine nucleotide (CpG sites), however Xie *et al.* (2012) it has already been reported a significant percentage of methylated non-CpG sites (XIE *et al.*, 2012).

The role of DNA methylation varies in different genomic regions. Within intergenic regions, the DNA methylation represses the expression of potentially harmful genetic elements, while within promoter regions containing CpG islands, stretches of DNA roughly 1000 base pairs long that have a higher CpG density than the rest of the genome and often are not methylated (BIRD *et al.*, 1985), results in stable silencing of gene expression (MOHN *et al.*, 2008). Also, within the gene body, region of the gene past the first exon, the methylation of the first exon leads to gene silencing (BRENET *et al.*, 2011).

2.2.2 Histone modifications

This process of transcription regulation involving DNA methylation, works in association with histone modifications and noncoding RNAs.

Normally, the DNA is wrapped around the histone proteins forming small, packaged sections called nucleosomes. This association is capable of inhibiting gene expression since a more compact region hinders the accessibility of transcription factors.

Those histones might have chemical modifications (methylation, acetylation, ubiquitination, and phosphorylation) added to three specific amino acids on their N-terminal tail which influences how DNA strands are packaged and their transcriptional activity. Those modifications that loosen DNA association with histones generally provide a permissive environment for transcription, whereas the ones that tights them repress gene expression (MOORE; LE; FAN, 2013).

These modifications involved in adding and/or stripping histone markers in order to impose a repressive state on a gene region, are lead by histone-modifying enzymes and are in general in cooperation with the Dnmts proteins which are involved in the mechanism of DNA methylation. For example, Dnmt1 and Dnmt3 are known to bind to the histone methyltransferase SUV39H1 that restricts gene expression by methylation on histone H3 lysine 9 (H3K9) (FUKS *et al.*, 2003). Other examples are Dnmt1 and Dnmt3b than can both bind to histone deacetylases that remove acetylation from histones to make DNA pack more tightly and restrict access for transcription (FUKS *et al.*, 2000; GEIMAN *et al.*, 2004). Also, histone H3 lysine 9 (H3K36) trimethylation, a repressive histone mark, stimulates the methyltransferase activity, the H3K4 trimethylation (H3K4me3) prevents it (OOI *et al.*, 2007; ZHANG *et al.*, 2010).

The miRNAs have emerged as another important epigenetic mechanism that influences gene expression. It can repress gene expression by inhibiting translation or inducing RNA degradation (BEREZIKOV, 2011) and it not only can regulate histone modifications and Dnmt expression thus regulating DNA methylation (BENETTI *et al.*, 2008; SINKKONEN *et al.*, 2008), but also the DNA methylation can regulate the expression of miRNAs (HAN *et al.*, 2007; LUJAMBIO *et al.*, 2008).

Overall, miRNAs, DNA methylation, and histone modifications work closely together to

regulate gene expression (MOORE; LE; FAN, 2013).

2.2.3 Genetics alterations

Although the haploid human genome consists of 3 billion nucleotides, changes in even a single base pair might result in dramatic physiological malfunctions. A mutation, defined as any alteration in the DNA sequence, can be a germ-line mutation, which occurs in a germ-line cell (one that will give rise to gametes) and can be passed to a descendant of the organism, and somatic mutations, which occurs in a somatic cell (one that develops into the body tissues) and are never transmit to their descendants (CLANCY, 2008).

A single base pair alteration that is present in at least 1% of the population is called single nucleotide polymorphism (SNP) and is generally used to refer to a normal variation (does not directly cause disease) (CLANCY, 2008).

An example of a point mutations that is able to alter proteins is the insertion or the deletion of a single base, called frameshift mutations. This change of nucleotides alters the codons (a group of three nucleotides) read by the ribosomes, which affects the protein resulted during the translation process. Table 1 show a classification of translational effect of variant allele (INSTITUTE, 2017; GDC, 2017).

Table 1 – Translational effect of variant allele. (INSTITUTE, 2017)

Variant_Classification	**Group**
Frame_Shift_Del	Deletion that moves the coding sequence out of frame
Frame_Shift_Ins	Insertion that moves the coding sequence out of frame
Missense_Mutation	The point mutation alters the protein structure by one amino acid
Nonsense_Mutation	A premature stop codon is created by the variant
Silent	Variant is in coding region of the chosen transcript, but protein structure is identical
Splice_Site	The variant is within two bases of a splice site.
In_Frame_Del	Deletion that keeps the sequence in frame
In_Frame_Ins	Insertion that keeps the sequence in frame
Translation_Start_Site	Point mutation, insertion or deletion that overlaps the start codon
Nonstop_Mutation	variant removes stop codon.
3'UTR	The variant is on the 3'UTR for the chosen transcript
3'Flank	The variant is downstream of the chosen transcript (within 3kb)
5'UTR	The variant is on the 5'UTR for the chosen transcript
5'Flank	The variant is upstream of the chosen transcript (within 3kb)
IGR	Intergenic region. Does not overlap any transcript.
Intron	The variant lies between exons within the bounds of the chosen transcript.
De_novo_Start_InFrame	New start codon is created in frame relative to the coded protein.
De_novo_Start_OutOfFrame	New start codon is created out of frame relative to the coded protein.
RNA	The variant lies on one of the RNA transcripts.
lincRNA	The variant lies on one of the lincRNAs.

Not only changes in DNA can occur on a single nucleotide called point mutations but they can also occur at the level of the chromosome, in which large segments of chromosomes are altered (deleted, duplicated, inverted, translocated to different chromosomes, rearranged) resulting in levels of gene expression, the complete absence of genes, or the alteration of gene sequence. Feuk, Carson and Scherer (2006) define structural variants as genomic alterations that involve segments of DNA that are larger than $1kb$ which can be microscopic ($> 3Mb$) or submicroscopic ($1kb$ to $3Mb$) variants, and also define smaller ($< 1kb$) variations or polymorphisms that involve the copy-number change of a segment of DNA as insertions or deletions (indels). A type of structural variant is the copy number variation (CNV) which refers to a segment of DNA that is $1kb$ or larger and is present at a variable copy number in comparison with a reference genome, while Copy-number polymorphism (CNP) are CNVs that occurs in more than 1% of the population. Also, it is common to refer to changes in copy number that have arisen in somatic tissue as copy number alterations/aberrations (CNAs or SCNAs) and changes in copy number in germline cells as copy number variations (CNVs). Table 2 summarizes the types of mutations.

Table 2 – Types of DNA Mutations and Their Impact (COMISC, 2017; CLANCY, 2008; FEUK; CARSON; SCHERER, 2006)

Class of Mutation	**Type of Mutation**	**Description**
Point mutation	Substitution	One base is incorrectly added during replication and replaces the pair in the corresponding position on the complementary strand
	Insertion	One or more extra nucleotides are inserted into replicating DNA, often resulting in a frameshift
	Deletion	One or more nucleotides is "skipped" during replication or otherwise excised, often resulting in a frameshift
Chromosomal mutation	Inversion	A segment of DNA that is reversed in orientation with respect to the rest of the chromosome.
	Deletion	A region of a chromosome is lost, resulting in the absence of all the genes in that area
	Duplication or low-copy repeat	A segment of DNA $< 1kb$ in size that occurs in two or more copies per haploid genome, with the different copies sharing $> 90\%$ sequence identity.
	Translocation	A change in position of a chromosomal segment within a genome that involves no change to the total DNA content. Translocations can be intra- or inter- chromosomal.
Copy number variation	Gain/Amplification	A single-copy gain or a multi-copy amplification of the entire gene
	Loss/Deletion	A loss for part or all of a coding region within the gene footprint in a single allele/copy (hemizygous deletion) or deletion for part or all of a coding region within the gene footprint in both alleles/copies (homozygous deletion). (COMISC, 2017)

2.3 Databases and data structure

Recent technological developments allowed the deposition of large amounts of genomic and epigenomic data, such as gene expression, DNA methylation, and genomic localization of transcription factors, into freely available public international consortia like The Cancer Genome Atlas (TCGA), The NCI's Genomic Data Commons (GDC), The Encyclopedia of DNA Elements (ENCODE), and The NIH Roadmap Epigenomics Mapping Consortium (Roadmap) (HAWKINS; HON; REN, 2010). Subsection 2.3.1 presents an overview of those consortia and subsection 2.3.2 presents some of the computational data formats and data structures used to represent the biological data.

2.3.1 Databases

The Cancer Genome Atlas (TCGA): The TCGA consortium, which was a National Institute of Health (NIH) initiative, made publicly available molecular and clinical information for more than 30 types of human cancers including exome (variant analysis), single nucleotide polymorphism (SNP), DNA methylation, transcriptome (mRNA), microRNA (miRNA) and proteome. Sample types available at TCGA are primary solid tumors, recurrent solid tumors, blood-derived normal and tumor, metastatic, and solid tissue normal (WEINSTEIN *et al.*, 2013). This project was ended in mid-2016 and its data was moved to the NCI Genomic Data Commons (GDC).

The Genomic Data Commons (GDC): The NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine. It supports numerous cancer genome programs at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) (NCI, 2017d). Those data is harmonized against GRCh38 (hg38) using GDC Bioinformatics Pipelines which provides methods to the standardization of biospecimen and clinical data, the re-alignment of DNA and RNA sequence data against a common reference genome build GRCh38, and the generation of derived data. Also, this repository has a Legacy Archive to provide access to a unmodified copy of data that was previously stored in CGHub (WILKS *et al.*, 2014) and in the TCGA Data Portal hosted by the TCGA Data Coordinating Center (DCC), in which uses as references GRCh37 (hg19) and GRCh36 (hg18).

The Encyclopedia of DNA Elements (ENCODE): Found in 2003 by the National Human Genome Research Institute (NHGRI), the project aims to build a comprehensive list of functional elements that have an active role in the genome, including regulatory elements that govern gene expression. Biosamples include immortalized cell lines, tissues, primary cells and stem cells (CONSORTIUM *et al.*, 2011).

The NIH Roadmap Epigenomics Mapping Consortium: This was launched with the goal of producing a public resource of human epigenomic data in order to analyze biology and disease-oriented research. Roadmap maps DNA methylation, histone modifications, chromatin accessibility, and small RNA transcripts in stem cells and primary ex vivo tissues (FINGERMAN *et al.*, 2011; BERNSTEIN *et al.*, 2010).

Briefly, these consortia provide large-scale epigenomic data onto a variety of microarrays and next-generation sequencing (NGS) platforms. Each consortium encompasses specific types of biological information on specific type of tissue or cell and, when analyzed together, it provides an invaluable opportunity for research laboratories to better understand the developmental progression of normal cells to cancer state at the molecular level and importantly, correlate these phenotypes with tissue of origins.

2.3.2 Data and data structure

If previously the main bottleneck to scientific progress in cellular biology was data collection, now this bottleneck shifted to analysis of data (MCPHERSON, 2009). The large-scale genomic data mining, which is the process of using many diverse datasets to address a specific biological question, involves three main tasks: establishing methodology for efficiently querying large data collections; assembling data from appropriate repositories, and integrating information from a variety of experimental data types (HUTTENHOWER; HOFMANN, 2010). In data science, raw data refers to data that have not been changed since acquisition and after its editing, cleaning or modification results in processed data. In genomics the raw data, sequencing read data, from next-generation sequencing machines is stored in a FASTQ, a common file format data that combines both the sequence and an associated per base quality score (COCK *et al.*, 2009). Those stored reads are then aligned to a reference genome sequence with software like Bowtie 2 (LANGMEAD; SALZBERG, 2012), generating SAM files, the accepted standard for storing short read alignment data, which are subsequently compressed to binary format (Binary Alignment/Map (BAM)) via SAMtools (LI *et al.*, 2009). It in common to consider this set of SAM, BAM, and FASTQ files as raw data. Some databases classify this raw data as level 1, protected level or raw level. It is important to highlight that due to privacy concerns, the access to this raw data is controlled because it could be used in identity tracing attacks (ERLICH; NARAYANAN, 2014; AYDAY *et al.*, 2014) as generally includes individually identifiable data such as low-level genomic sequencing data, germline variants, SNP6 genotype data, and certain clinical data elements (NCI, 2017e). This raw data are processed through bioinformatics pipelines generating processed data files or matrices (usually in the form of tab-delimited text files). Some databases classify this processed data as level 2 or high or as open. For example, the previously stored data in CGHub, TCGA Data Portal and Broad Institute's GDAC Firehose, were provided as different levels or tiers that were defined in terms of a specific combination of both processing level (raw, normalized, integrated) and access level (controlled or open access).

Level 1 indicated raw and controlled data, level 2 indicated processed and controlled data, level 3 indicated Segmented or Interpreted Data and open access and level 4 indicated region of interest and open access data.

In the next subsections, we will focus on the processed data and their data structures used in this project.

2.3.2.1 Data

Array-based DNA methylation data: DNA methylation data files contain information on raw and normalized signal intensities, detection confidence and calculated beta values for methylated and unmethylated probes. The processed data from DNA methylation array-based platforms such as GoldenGate, Infinium Human Methylation27, and the Infinium HD 450K methylation array, are presented in the form of beta-values that uses a scale ranging from 0.0 (probes completely unmethylated) up to 1.0 (probes completely methylated). In each sample file, each row represents a 1 bp region (probe) and its correspondent beta-value. Also, probes overlapping with SNPs are normally masked as they are based largely on ad hoc assumptions and subjective criteria (ZHOU; LAIRD; SHEN, 2016).

Somatic Variant data: The DNA-Seq Somatic Variant Analysis algorithms, such as MuSE (FAN *et al.*, 2016), Mutect2 (CIBULSKIS *et al.*, 2013), SomaticSniper (LARSON *et al.*, 2011) and Varscan2 (KOBOLDT *et al.*, 2012), identify and characterizes somatic mutations by comparing reference alignments from tumor and normal samples from the same case. The results of the variant calls are stored in a Mutation Annotation Format (MAF), which are filtered to remove any potentially erroneous or germline variant calls. Each row of these represents a mutation in the genome of a sample (NCI, 2017a).

RNA-Seq Gene Expression data : The processed data contains read counts measured on a gene level, which might be normalized using some methods such as the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) (NCI, 2017b) and FPKM Upper Quartile (FPKM-UQ) (NCI, 2017c).

miRNA-Seq data: The processed data contains expression levels measured and normalized post-alignment. Normalization is performed using the Reads per Millions (RPM) method. Expression levels of known miRNAs and observed miRNA isoforms are generated for each sample.

Copy Number Variation data: The processed data is a table that associates contiguous chromosomal segments with genomic coordinates, mean array intensity (log2 ratio segment means), and the number of probes that binds to each segment. This information might be masked to remove segments with probes known to contain germline mutations. From the Circular Binary Segmentation (CBS) methods output described previously, algorithms

Table 3 – Example TCGA barcode

TCGA	02	0001	01	C	01	D	0182	01
Project	Tissue source site	Participant	sample	vial	portion	analyte	place	center

such as GISTIC2 (MERMEL *et al.*, 2011) are used to identify significantly altered regions of amplification or deletion across sets of patients.

Histone ChIP-Seq data: The processed data contains either the fold change over control and signal p-value stored in a bigWig format or as peaks in a bed and bigBed format (ENCODE, 2017). The ENCODE consortium highlight that these peak files are not meant to be interpreted as definitive binding events, but are rather intended to be used as input for subsequent statistical comparison of replicates (ENCODE, 2017). This data is used as input for software to infer chromatin-state such as ChromHMM (ERNST; KELLIS, 2012) and StatePaintR (COETZEE *et al.*, 2017).

To identify each of these samples, each database uses a unique identifier, which is able to help users to find a certain sample in the database or to find associated information. The TCGA database, for example, created an ID in two forms, a human-readable version called barcode and a non-human readable called UUID. A TCGA barcode is composed of 28 alphanumeric elements, representing the sample information. How those elements are grouped and which information they have is shown in table 3, a code table for each field value is found in the GDC website (NCI, 2017f). The barcode identifies if a sample is a primary solid tumor sample or a normal tissue sample, but if there was a mistake in the metadata and the sample which was normal was set to a tumor sample, the barcode is then invalid. Due to this issue, the TCGA UUID, composed of 32 randomly generated digits, was created. The TCGA barcode was the primary identifier of biospecimen data since the pilot project began. However, since for any one sample, the barcode can change as the meta-data associated with it changes, the TCGA project transitioned to using UUIDs as the primary identifier.

2.3.2.2 Data structures

Although there exists a wealth of possibilities (KANNAN *et al.*, 2015) in accessing cancer associated data, Bioconductor (https://www.bioconductor.org/) represent the most comprehensive set of open source, updated and integrated professional tools for the statistical analysis of large-scale genomic data. It uses the R statistical programming language and has been developing several data structures to handle genetics, epigenetics and clinical data. Some of the main data structures from the R/Bioconductor used in this project are described below.

DataFrame: A data frame is used for storing data tables. It is a list of vectors of equal length.

GenomicRanges (GRanges): This data structure was created to representation and manipulate genomic intervals and variables defined along a genome GRanges is a vector of genomic locations, which are composed of the sequence names (chr1, chr2), the sequence range (start and end) and the strand information, and associated annotations (metadata) (LAWRENCE *et al.*, 2013). An example of GRanges object is shown in Figure 1.

```
GRanges object with 10 ranges and 2 metadata columns:
    seqnames     ranges strand |     score                GC
       <Rle>  <IRanges>  <Rle> | <integer>         <numeric>
  a     chr1 [101, 111]      - |         1                 1
  b     chr2 [102, 112]      + |         2 0.888888888888889
  c     chr2 [103, 113]      + |         3 0.777777777777778
  .      ...         ...    ... .       ...               ...
  h     chr3 [108, 118]      + |         8 0.222222222222222
  i     chr3 [109, 119]      - |         9 0.111111111111111
  j     chr3 [110, 120]      - |        10                 0
  -------
  seqinfo: 3 sequences from an unspecified genome; no seqlengths
```

Figure 1 – Example of Granges object.

SummarizedExperiment (SE): This data structure, exemplified in Figure 2, is a matrix-like container where rows represent ranges of interest (as a GRanges object) and columns represent samples (with sample data summarized as a DataFrame). It can contains one or more assays, each represented by a matrix-like object of numeric or other mode.

MultiAssayExperiment (MAE): This data structure is an integrative environment where multiple assays are managed and preprocessed for genomic data analysis. It is composed of at least two metadata matrices, one with the phenotype metadata (i.e. clinical data) and one table mapping each data column from different assays to a entry in the phenotype metadata. The data handle by this data structure can be several SummarizedExperiment objects (i.e DNA methylation, gene expression, copy number, histone modification signals).

2.4 (Epi)Genomic data analysis

2.4.1 Integrative analysis

Often the onset and progression of cancerous diseases are linked to the aberrant function of proteins and alterations in gene expression, which has led research in genetics in search of the molecular alterations responsible for such aberrant behavior. Among the types of alterations genetic and epigenetic that can impact gene function are, gene copy number (CN), DNA methylation, single nucleotide variations (SNV), and indels (small insertions and deletions). These

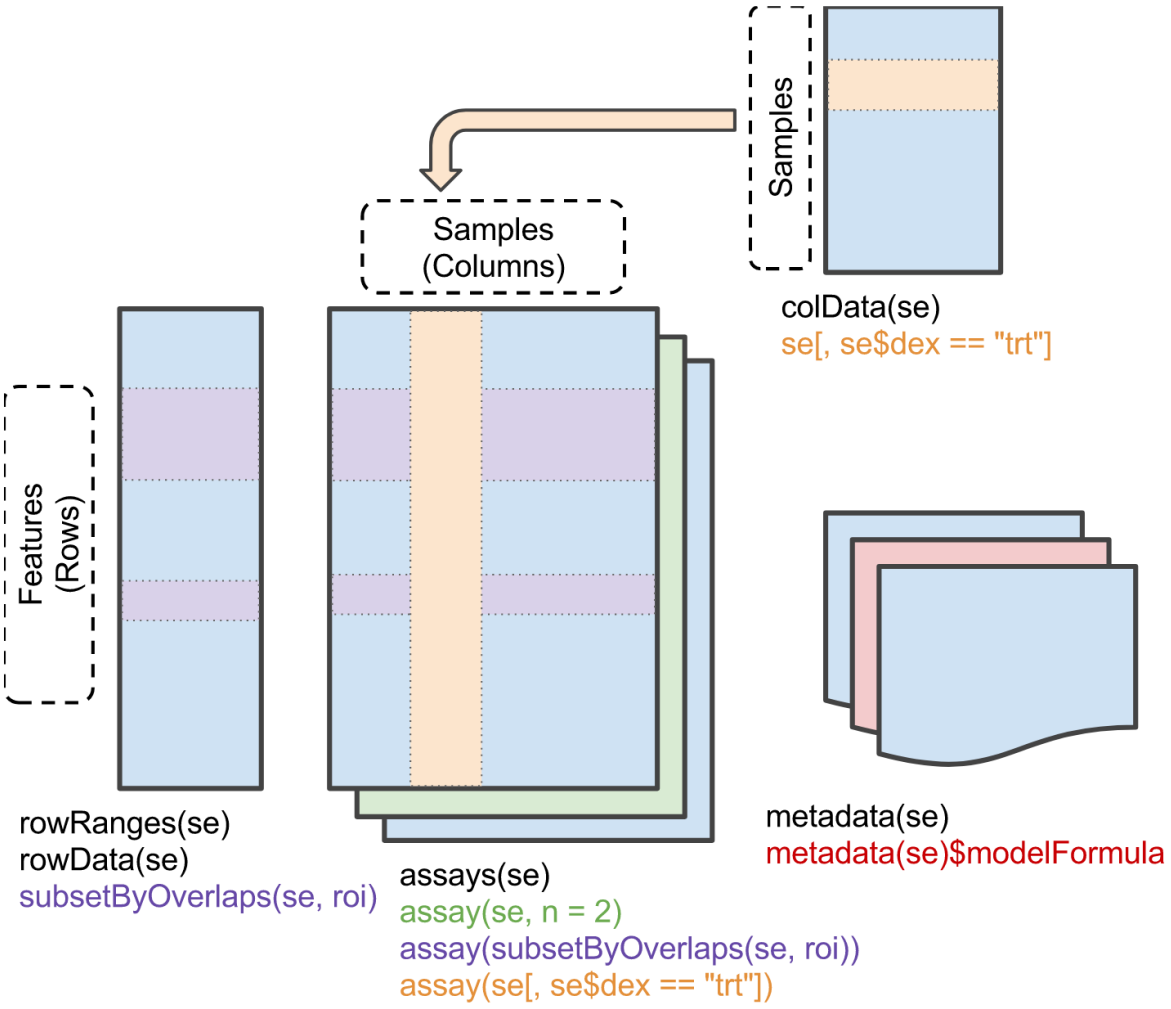


Figure 2 – Example of Sumarized Experiment object. Figure reproduced from SummarizedExperiment manual (MORGAN *et al.*, 2017).

variations can have a direct effect by modifying the function of the gene product, for example, a indel in a coding region, or an indirect effect such as the modification of regulatory regions which can interfere with gene expression by inhibiting transcription (THINGHOLM *et al.*, 2016).

Furthermore, recent technological developments have enabled the creation of genome-wide data for multiple types of variations. Although individual analyzes of these variations helped increasing knowledge of the genome and of complex disorders, integrative analyses that evaluate cancer transcriptome data in the context of other data sources are often capable of extracting deeper biological insight from the data (RHODES; CHINNAIYAN, 2005).

In this section we will highlight several integrative approaches, including meta-analysis for extracting robust profiles from independent data sets, enrichment analysis for identifying coordinately regulated functional gene modules, protein interaction networks for detecting interaction complexes deregulated in cancer, transcriptional networks for inferring regulatory mechanisms in cancer and analyses of model system profiles with human tumor profiles for inferring activity of oncogenic pathways.

2.4.1.1 Functional enrichment analysis of cancer signatures

A gene set enrichment analysis (GSEA) or functional enrichment analysis is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins and may have an association with disease phenotypes. This type of analysis integrates differentially expressed genes identified by a differential expression analysis (DEA) with external functional information which is necessary for interpreting and summarizing large cancer signatures. Most approaches use external annotation databases such as Gene Ontology which is a database of controlled vocabulary gene annotations describing the biological processes, molecular functions and cellular localizations of genes as a resource for enrichment analysis in cancer signatures.

Let L be a list of genes ranked by their differential expression between the classes, the goal of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) this list, which might correlate with the phenotypic class distinction (SUBRAMANIAN *et al.*, 2005).

2.4.1.2 Protein interaction networks and cancer signatures

A major objective of systems biology is to organize molecular interactions as networks (VINAYAGAM *et al.*, 2014). Protein-protein interactions (PPIs) are essential to almost every process in a cell and are not only crucial for understanding cell physiology in normal and disease states, since the disruption of protein-protein interactions may result in disruption of the cell component or process to which they contribute compromising the cell viability or even leading to cell death, but also for the drug development, since drugs can affect PPIs (EMBL-EBI, 2017; ALZATE, 2009). These interactions are represented as Protein-protein interaction networks (PPIN), a mathematical representation of the physical contacts between proteins in the cell. The totality of PPIs that happen in a cell, an organism or a specific biological context is called interactome.

A detailed human interactome network that captures the entire cellular network is invaluable in interpreting cancer signatures, allowing one to infer activated subnetworks and specific proteins that are most important to a subnetwork. For example, for a given set of overexpressed proteins in the same subnetwork, one specific protein that interacts with the entire subnetwork might be the key to control the expression the entire subnetwork.

A lot of this information is available through molecular interaction databases such as IntAct (http://www.ebi.ac.uk/intact) (ORCHARD *et al.*, 2013).

2.4.1.3 Transcriptional targets and cancer signatures

Similar to the Protein-protein interaction networks, global transcriptional networks, which defines directional pathways (i.e. which gene actives are activated by a given gene), have the potential to improve the interpretation of cancer signatures. For example, if the targets of all

transcription factors were known, then one could easily infer which transcription factors must be activated in a tumor to yield the observed cancer signature. With that information, it is possible to reduce a complex cancer signature to a small number of activated transcriptional factors which will be potential therapeutic targets.

The identification of transcription factor-binding sites (TFBSs) is done using high throughput experimental methods such as ChIP-Chip and ChIP-sequencing (ChIP-seq) which identifies a region of 100–1000 base pairs (b.p.) in which the TFBS (typically 9-15 b.p.) resides (JAYARAM; USVYAT; MARTIN, 2016). In-silico sequence-based methods were developed to predicted TFBSs. These methods scan a DNA sequence of interest with a position weight matrix (PWM),a $4xn$ matrix of scores for each of the 4 bases across each position in the binding motif, for a transcription factor of interest and perform a pattern-matching. PWM models can be obtained from a number of resources including the the open access database JASPAR (PORTALES-CASAMAR *et al.*, 2009), HOCOMOCO (KULAKOVSKIY *et al.*, 2013) and HOMER (HEINZ *et al.*, 2010). Among the existing software that scans a sequence database for individual matches to each of the motifs are Find Individual Motif Occurrences (FIMO) (GRANT; BAILEY; NOBLE, 2011), HOMER (HEINZ *et al.*, 2010) and PATSER (TURATSINZE *et al.*, 2008).

2.4.1.4 Human Epigenomes

The human body contains more than 200 different cell types each one has an identical copy of the genome but expresses a distinct set of genes, due to their epigenome which in each cell regulates gene expression in a number of ways - by organizing the nuclear architecture of the chromosomes, restricting or facilitating transcription factor access to DNA, and preserving a memory of past transcriptional activities (RIVERA; REN, 2013). The integrative analysis of epigenomic maps, which references to the collection of DNA methylation state and covalent modification of histone proteins along the genome (BONASIO; TU; REINBERG, 2010), has been shown to be important in the study of the gene regulatory programs (GIFFORD *et al.*, 2013; HAWKINS *et al.*, 2010; RADA-IGLESIAS *et al.*, 2012). Table 4 shows a summary of the epigenomic marks and their associate role.

Table 4 – Core set of five histone modification marks and other epigenomic marks

Histone marks	**Role**
Histone H3 lysine 4 trimethylation (H3K4me3)	Promoter regions (HEINTZMAN *et al.*, 2007; BERNSTEIN *et al.*, 2005)
Histone H3 lysine 4 monomethylation (H3K4me1)	Enhancer regions (HEINTZMAN *et al.*, 2007)
Histone H3 lysine 36 trimethylation (H3K36me3)	Transcribed regions
Histone H3 lysine 27 trimethylation (H3K27me3)	Polycomb repression (BONASIO; TU; REINBERG, 2010)
Histone H3 lysine 9 trimethylation (H3K9me3)	Heterochromatin regions (PETERS *et al.*, 2003)
Histone H3 acetylated at lysine 27 (H3K27ac)	Increase activation of genomic elements (HEINTZMAN *et al.*, 2009; RADA-IGLESIAS *et al.*, 2011; CREYGHTON *et al.*, 2010)
Histone H3 lysine 9 acetylation (H3K9ac)	Transcriptional activation (NISHIDA *et al.*, 2006)
DNase hypersensitivity	Regions of accessible chromatin (THURMAN *et al.*, 2012)
DNA methylation	Repressed regulatory regions (CEDAR; BERGMAN, 2009; MOORE; LE; FAN, 2013)

Using these set of epigenetic marks, it is possible to discover epigenomic states which would aid in understanding "non-coding" genomic elements (COETZEE *et al.*, 2017). Several tools can use those marks to infer chromatin-state such as ChromHMM (ERNST; KELLIS, 2012) and StatePaintR (COETZEE *et al.*, 2017). A 15-state model created from ROADMAP data using ChromHMM is presented as follows: active TSS-proximal promoter states (TssA, TssAFlnk), a transcribed state at the 5' and 3' end of genes showing both promoter and enhancer signatures (TxFlnk), actively-transcribed states (Tx, TxWk), enhancer states (Enh, EnhG), and a state associated with zinc finger protein genes (ZNF/Rpts). The inactive states consist of constitutive heterochromatin (Het), bivalent regulatory states (TssBiv, BivFlnk, EnhBiv), repressed Polycomb states (ReprPC, ReprPCWk), and a quiescent state (Quies) (KUNDAJE *et al.*, 2015). Those precomputed chromatin-state segmentation from public databases such as Blueprint, ENCODE and ROADMAP using ChromHMM (http://compbio.mit.edu/ChromHMM/) are available at http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state and using StatePaintR are available at www.statehub.org.

These data are extremely useful to characterize each different region of the genome. For example, using those information Kundaje *et al.* (2015) were able to identify enhancer regions containing strong H3K27ac signal which showed a higher DNA accessibility, lower methylation, and higher TF binding than enhancers regions lacking H3K27ac. Those regions with the strong H3K27ac signal are more likely to have a regulatory role.

2.4.2 Statistical analysis

2.4.2.1 Hypothesis Testing

Human genetic studies aim to identify if a phenotype is related to the genotypes at various loci, that is, if genetic variations have an influence on the risk of disease or other health-related phenotypes. Statistical analysis is a crucial to present the findings in an interpretable and objective manner (SHAM; PURCELL, 2014).

The most popular hypothesis testing approach used to test if genotypes and phenotypes are related is the frequentist significance testing approach. This is a classical approach that involves setting up two competing hypothesis: a null hypothesis (H_0) and an alternative hypothesis (H_1).

Computing the statistical significance can be done using a one-tailed or a two-tailed test. A two-sided test is appropriate to evaluate both directions of the test, for example, is the estimated value smaller or higher than the reference, which actually tests if the estimated value is different from the reference. A one-sided test is appropriate to evaluate only one direction of the test, for example, is the estimated p-value smaller than the reference. An example in genetic studies for a two-sided test would be H_0 hypothesis that genotypes have no effect on the phenotypes while the H_1 hypothesis is that there is an effect. Table 5 shows other examples.

Table 5 – Example of three hypothesis tests about the population mean μ. In genetics it could be the mean level of expression of a gene.

Type	Null	Alternative
Right-tailed	$H_0 : \mu = 0$	$H_1 : \mu > 0$
Left-tailed	$H_0 : \mu = 0$	$H_1 : \mu < 0$
Two-tailed	$H_0 : \mu = 0$	$H_1 : \mu \neq 0$

Table 6 – Type I and II Errors. $\alpha = P(\text{Type I Error})$, $\beta = P(\text{Type II Error})$

Decision	H_0 is True	H_0 is False
Do Not Reject H_0	Correct Decision	Incorrect Decision (1 - β)
Reject H_0	Incorrect Decision (1 - α)	Correct Decision

2.4.2.2 Making a decision: P-value approach

The decision to reject or accept H_0 is made based on the calculation of a test statistic (T) from the observed data. As the value of T depends on particular individuals in the population, repeating the study using different random samples from the population would provide of many different values for T. These set of T can be summarized as a probability distribution.

Even though, the decision made to reject or accept H_0 just state that we had enough evidence to behave one way or the other. The rejection of the null hypothesis does not prove that the alternative hypothesis is true as the acceptance the null hypothesis does not prove that the null hypothesis is true. It might happen that null hypothesis was rejected when it was true, or it was not rejected when it was false. The first error in statistics is called a Type I error ("false positive"), while the second is called a Type II error ("false negative"). Table 6 shows the relations between truth/falseness of the null hypothesis and outcomes of the test.

The rate of the type I error or significance level α is the probability of having a false positive. Normally, the significance level is set to 5%, implying that it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis. With the same logic, the rate of the type II error is denoted by β .

To make a decision whether to reject or accept the null hypothesis, the concept of probability value was introduced. A p-value (p-value $\in [0,1]$) is the probability under a specified statistical model, constructed under a set of assumptions (normally “null hypothesis") that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value (WASSERSTEIN; LAZAR, 2016). That means, the smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis and greater the p-value more compatible is the data with the null hypothesis. In summary, if P-value is small (e.g.P-value $\leq \alpha$) then the null hypothesis is rejected, otherwise, it is not rejected.

It is important to highlight that a p-value does not measure the size of an effect or the importance of a result. It might happen that a very small effect produces smaller p-values if the sample size is big or measurement precision is high. On the other hand, a large effect might produce higher p-values if the sample size is small or measurements are imprecise.

2.4.2.3 Correcting for multiple testing

When performing a set of statistical inferences simultaneously more likely erroneous inferences are to occur. For example, if 100 tests are carried out, then 5% of them (that is 5 tests) are expected to have $P-value < 0.05$ by chance when H_0 is, in fact, true for all the tests. Compared to a single test (equations 2.1a and 2.1b), the probability of having a type 1 error multiple test is given by the equations 2.1c and 2.1d (ŠIDÁK, 1967).

$$P(\text{Making an error}) = \alpha \tag{2.1a}$$

$$P(\text{Not making an error}) = 1-\alpha \tag{2.1b}$$

$$P(\text{Not making an error in m tests}) = (1-\alpha)^m \tag{2.1c}$$

$$P(\text{Making at least 1 error in m tests}) = 1-(1-\alpha)^m \tag{2.1d}$$

To handle this multiple statistical testing problems, some techniques to re-calculating probabilities obtained from a statistical test which was repeated multiple times have been developed to prevent the inflation of false positive rates.

Among the different approaches to control type I errors we have Family-wise error rate (FWER) which controls the probability of at least one type I error, and False discovery rate (FDR) which controls the expected proportion of Type I errors among the rejected hypotheses. Compared to FDR, controlling FWER is extremely conservative approach as the power to detect H_1 gets very small.

Among the different adjustment methods to control FWER includes the Bonferroni correction in which the p-values are multiplied by the number of comparisons ($M * P_i < \alpha$) and the Holm correction, $P-adjusted_i = P_i * (M+1-i)$, where $i \in \{1,2,\ldots,n\}$ and smaller the p-value is smaller will the index i be (AICKIN; GENSLER, 1996). The Benjamini-Hochberg (BH) method to control FDR procedure will identify the largest k, such that $P_k \leq \frac{k}{m}\alpha$, all null hypotheses H_i for $i \in \{1,\ldots,k\}$ are rejected.

These methods make the assumption that the tests are independent tests, which often is not valid for genomics data. For dependent tests, permutation methods are often used to calculate p-values. This approach recalculates a p-value comparing the P-value calculated from the real data test with random ones, which are performed by randomly shuffling the case–control (or phenotype) labels. All M tests are recalculated on the reshuffled data set, with the smallest P value of these M tests being recorded. The procedure is repeated for many times to construct an

empirical frequency distribution of the smallest P values. This empirical adjusted P value (P_*) is given by:

$$P_* = \frac{r+1}{n+1}$$

where n is the number of permutation carried out, and r is the number of permuted p-values smaller than P-value calculated from the real data.

For example, considering P-value = 0.1 and the permuted p-values

$$P_{permu} = \{0.001, 0.01, 0.02, 0.03, 0.05, 0.2, 0.5, 0.6, 1\}$$

the first 5 permuted p-values are smaller than the original p-value, which would give us $r = 5$, resulting in:

$$P_* = \frac{r+1}{n+1} = \frac{5+1}{9+1} = 0.6$$

It is important to highlight that a high number of permutations is required in order to produce reliable permuted p-value adjusted. (DAVISON; HINKLEY, 1997; NORTH; CURTIS; SHAM, 2002; NORTH; CURTIS; SHAM, 2003; SHAM; PURCELL, 2014).

2.4.2.4 Nonparametric and parametric tests

Statistical procedures can be classified into two groups: Parametric and nonparametric. Parametric statistical procedures rely on assumptions about the shape of the distribution (i.e., assume a normal distribution) in the underlying population and about the form or parameters (i.e., means and standard deviations) of the assumed distribution. While nonparametric statistical procedures don't rely or rely on only a few assumptions about the shape or parameters of the population distribution from which the sample was drawn. Some of these procedures are summarized in Table 7.

The most used parametric test for comparing the means of two independent groups is the t-test, which assumes that the data are normally distributed, that samples from different groups are independent and that the variances between the groups are equal (KITCHEN, 2009). The most commonly used nonparametric test for indepedent samples is the Mann-Whitney U-test (MWU), which assumes that observations from the different groups are random samples (i.e. independent and identically distributed) from their respective populations, are mutually independent and are ordinal or continuous measurements. If matched or dependent samples are compared the nonparametric Wilcoxon signed-rank test is used (WHITLEY; BALL, 2002). When there are more than two groups being compared, the nonparametric test used is the Kruskal-Wallis test (KW), a generalization of the MWU and the parametric test used is the Analysis of variance (ANOVA) (PARAB; BHALERAO, 2010).

It is also important to highlight that for larger sample sizes (greater than 20 or 30) P values can be calculated using a Normal distribution, on other words parametric tests could be used instead of the nonparametric ones (VICKERS, 2005).

Table 7 – Summary of parametric and nonparametric procedures.

Analysis Type	Example	Parametric	Nonparametric
Compare means between two distinct/independent groups	Is the mean TP53 gene expression for control group different from the mean for treatment group?	Two-sample t-test	Wilcoxon ranksum test
Compare two quantitative measurements taken from the same individual	Was there a change in gene expression after the treatment?	Paired t-test	Wilcoxon signedrank test
Compare means between three or more distinct/independent groups	For a given three groups (e.g., placebo, drug #1, drug #2), is the TP53 gene expression different among the three groups?	Analysis of variance (ANOVA)	Kruskal-Wallis test
Estimate the degree of association between two quantitative variables	Is age related to the TP53 gene expression?	Pearson coefficient of correlation	Spearman's rank correlation

2.4.3 Survival analysis

In the medical sciences, in several circumstances, it is necessary to evaluate if a treatment had a beneficial effect on the survival of patients. For this, it is measured the fraction of patients living for a certain amount of time after treatment.

It is called "Survival times" data that measure follow-up time from a defined starting time to the occurrence of a given event (e.g. from the diagnosis of a disease to death) (BEWICK; CHEEK; BALL, 2004). These survival times are "censored" when there is a follow-up time but the event has not yet occurred or is not known to have occurred. That might happen if a patient drops out of the study before its end, or if you are studying a treatment and it is not over yet.

2.4.3.1 Kaplan–Meier method: Estimating the survival curve

It is defined survival function $S(t)$ is defined as the probability of surviving at least to time t, while a graph of $S(t)$ against t is called the survival curve. To estimate this curve there exists the Kaplan-Meier method:

$$S(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

where d_i are the number of events and n_i the total individuals at risk at time i. Table 8 shows an example of survival data and Table 9 shows an example of applying Kaplan-Meier method. Figure 3 shows a more real data example. It is important to highlight that for a censored time the proportion surviving will be 1, that means an individual is considered to be at risk of dying in the next event of the censoring but not in subsequent events (BLAND; ALTMAN, 2004).

Table 8 – Survival time and status for a group of patients.

Patient ID	Survival times (in days)	Status
1	1	Dead
2	1	Dead
3	2	Alive (censored)
4	3	Dead

Table 9 – Kaplan-Meier method example for table 8

Interval	n_i: patient at risk at time t_i^-	d_i = deaths at time t_i	c_i = censored at time t_i	$S(t)$
$[0,1)$	4	0	0	1
$[1,3)$	4 - 0 = 4	2	1	$1-\frac{2}{4}=0.5$
[3, End of study]	4 - 2 - 1 = 1	1	0	$0.5*(1-\frac{1}{1})=0$

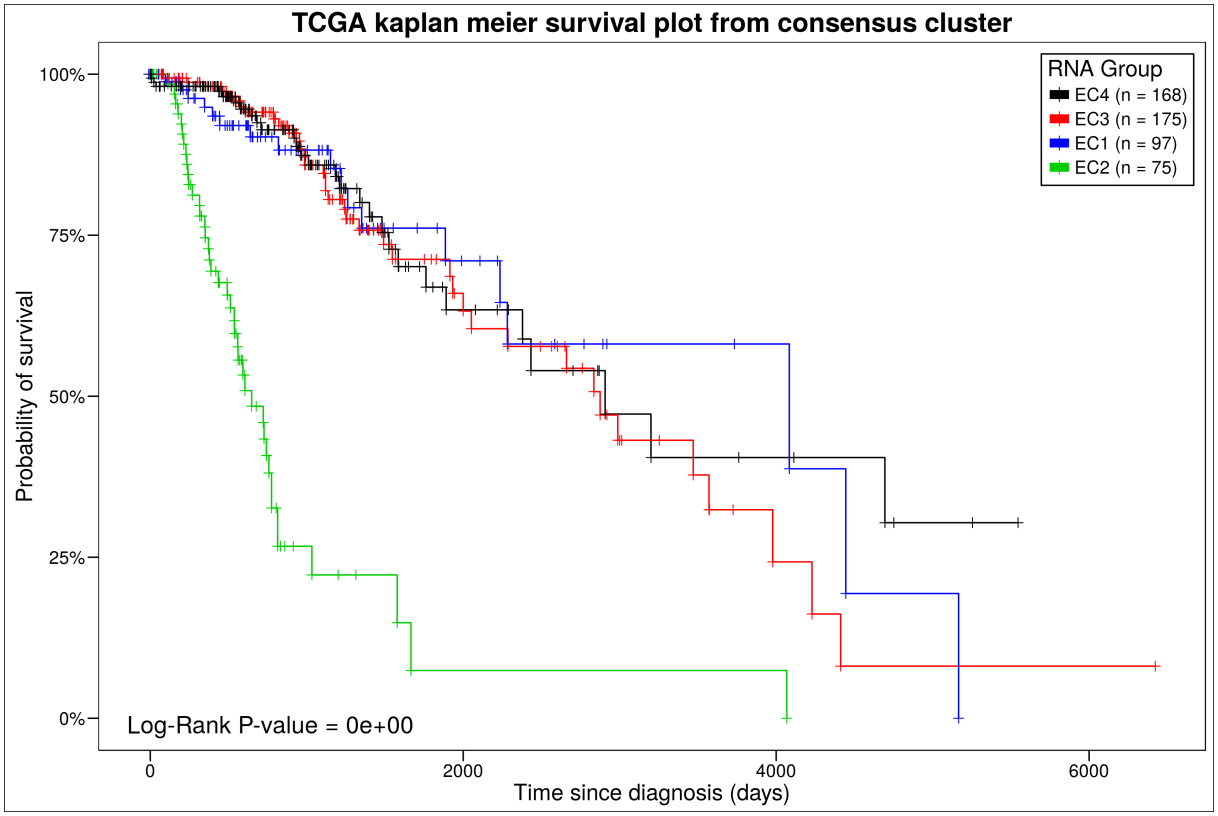


Figure 3 – Example of survival curve for TCGA samples clustered by RNA expresssion levels. Group 2 has the worst survival. Censored data is marked (+) in the plot.

2.4.3.2 Comparing survival curves of two groups using the log-rank test

To compare the survival distributions of two or more groups, the hypothesis test log-rank test is used to test the null hypothesis that there is no difference between the populations in the probability of an event at any time point. The approximated statistics used for comparison purposes for k groups is

$$T = \sum_{i \in \{1,\ldots,k\}} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed numbers of death in group i while E_i is its expected numbers of deaths. If the null hypothesis is true, T is distributed approximately as a χ^2_{k-1} (MATTHEWS; FAREWELL *et al.*, 1996). If T calculated is 9.44, and $k = 2$, evaluating the quantile function (also known as "inverse CDF" or "ICDF") of the chi-squared distribution the significance level of these data is equal to $P_r(\chi^2_1 \geq 9.44) = 0.002$ (YAU, 2012). For a given cut-off, normally 0.05, the results with a p-value smaller are considered significant.

2.4.4 Machine Learning

Machine learning is a field of computer science focused on the development and application of algorithms that improve with experience (MITCHELL *et al.*, 1997). In genetics and genomics, it has been applied for the interpretation of large genomic datasets and annotation of a wide variety of genomic sequence elements. For example, for the detection of transcription start sites (TSSs) locations, which have proven hard to detect in silico due to the complexity and the fairly diffuse structures of Eukaryotic promoters, Down and Hubbard (2002) developed a machine-learning method is able to build useful models of promoters for more than 50% of the human transcription start sites (DOWN; HUBBARD, 2002).

The machine learning techniques can be classified into two main categories: supervised and unsupervised learning (MITCHELL *et al.*, 1997). The supervised learning, which aims to infer data labels by learning from already labeled data, has three stages: design, model, and test. The first stage refers to the selection of a learning algorithm used to learn from data (e.g. choose between support vector machines or random forest algorithms) and its training data. The second stage is the creation of a model from labeled data using the algorithm selected previously. The last stage uses this generated model to predict the labels of unlabeled data. The unsupervised learning methods, on the other hand, cluster the data without using labels, which requires an additional step in which semantics must be manually assigned to each cluster. As this discovery is not tied to previously defined classes, these methods have as benefit the ability to identify potentially novel types of genomic elements.

2.4.4.1 Supervised learning

Supervised learning is a type of machine learning algorithm that identifies patterns in data to make predictions. Specifically, the algorithm takes a known set of input data and known responses to the data (labels) and trains a model to generate predictions for the response to new data. The supervised learning algorithms can be classified into two categories: classification and regression.

In classification, the goal is to assign a label to an observation. That is, responses are categorical variables. For example, given dataset known to be an enhancer or not enhancer, one could want to predict the locations of enhancers in the genome.

In regression, the goal is to predict a continuous measurement for an observation. That is, the responses variables are real numbers. In genomics, these methods are used in genetic epidemiology to detect association of genetic variants with a trait or disease of interest (DASGUPTA *et al.*, 2011).

2.4.4.2 Unsupervised learning

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. These algorithms seek to group a set of objects into clusters (groups) such that those in the same group are more similar to each other than to those in another cluster. The major unsupervised learning techniques used in bioinformatics are described below.

Hierarchical clustering: This clustering method builds a hierarchy of clusters, through either a agglomerative ("bottom-up") or a divisive ("top-down") procedures. In the agglomerative procedures, each n observation starts in its own cluster and until only one cluster remains the groups with the smallest dissimilarity are merged. On the other hand, in the divisive procedures, all observations start in one cluster and until all observations are in their own cluster the group is split into two groups with the biggest dissimilarity. To decide which clusters to merge or divide a metric to measure of dissimilarity between sets of observations is required. Among the existing agglomerative techniques are the single linkage, also known as the nearest-neighbour technique, which defines the smallest dissimilarity between two points in each group as the dissimilarity between two groups (FLOREK *et al.*, 1951), the complete linkage which defines the largest dissimilarity between two points in each group as the dissimilarity between two groups (DEFAYS, 1977), the average linkage which defines the average dissimilarity overall points as the dissimilarity between two groups (SOKAL, 1958) and the Ward's method which at each merge step minimizes the increase in the total within-cluster error sum of squares, which means that the groups leading to a minimum increase in total within-cluster variance are merged (JR, 1963; EVERITT *et al.*, 2011). The results of hierarchical clustering are usually represented in a dendrogram a branching diagram in which the objects are represented in one of the axes while the similarity between clusters are represented in the other axis by the length of the connection which joins them (MANNING; SCHÜTZE *et al.*, 1999).

Optimization clustering techniques: These are a set of non-hierarchical clustering techniques that cluster individuals into a specified number of groups, by either minimizing or maximizing some numerical criterion (EVERITT *et al.*, 2011). One of these techniques is the k-means algorithms, which iteratively updates a partition by simultaneously relocating each object to the group to whose mean it was closest and then recalculating the group means until the groups no longer change. More explicitly, let $S = \{S_1, S_2, \ldots, S_k\}$ be the

set of k clusters, this algorithm tries to minimize the squared distances of the elements from the cluster center:

$$N(n,k) = min \sum_{g=1}^{k} \sum_{j \in S_g} \|x_j - \mu_g\|^2,$$

where x_j is a data element that belongs to a group g, and μ_g is the center of points of the g group. A counterpoint of these techniques also requirement to 'estimate' the number of clusters in the data, for which a variety of methods have been suggested and most are relatively informal and involve, essentially, plotting the value of the clustering criterion against the number of groups in which large changes of levels in the plot are usually taken as suggestive of a particular number of groups. Some formal methods were also proposed such as the Silhouette index by Rousseeuw (1987). And the second counterpoint of this technique is the choice of initial starting values which might lead to a local optimal solution, some solution to this problem such as a bootstrap-like approach to 'refine' initial seeds are suggested (STEINLEY, 2003).

Principal component analysis (PCA): This method transforms the variables in a multivariate data set into new variables which are uncorrelated with each other and a linear combination of the original variables. This technique provides a means of projecting the data into a lower dimensional space, which is very useful mainly due to the problems of high dimensional data which is called "curse of dimensionality". Generally, Verleysen and François (2005) defines the curse of dimensionality as the expression of all phenomena that appear with high-dimensional data, and that have most often unfortunate consequences on the behavior and performances of learning algorithms. In summary, the number of learning data should grow exponentially with the dimension, that means if 10 data is reasonable to learn a 1-dimensional model, for a 2-dimensional model it will need 100 data). For example, in genetics, microarray experiments have information for thousands of genes, if each one is considered a variable we would have a thousand dimensions, and we would need an enormous amount of observations to obtain a reliable result. If a dimensionality reduction is not performed, with a increase of data features (dimensions), it is very likely that, by chance, one feature perfectly separates the training examples into positive and negative classes, which would lead to good performance on the training data but poor generalization to data that were not used in training (LIBBRECHT; NOBLE, 2015).

CHAPTER 3

DEVELOPMENT OF METHODS AND SOFTWARES FOR CANCER DATA ANALYSIS

This chapter introduces three R/Bioconductor packages developed during the doctoral project. These tools are complementary and together they are able to perform a deep integrative analysis of genomic and epigenomic data from different databases, from the data acquisition stage to the integrative analysis.

First, to deal with the problem of data acquisition and analysis we created the R/Bioconductor TCGAbiolinks package which searches, downloads and organizes data and metadata from the National Cancer Institute's (NCI) Genomic Data Commons (GDC) into the latest R/Bioconductor data structures. Second, to deal with the problem of integrative analysis, we worked closely with Dr. Benjamin P. Berman's group to create a new version of the R/Bioconductor ELMER package, which infers regulatory element landscapes and transcription factor networks from cancer methylomes (YAO *et al.*, 2015). This version not only was greatly improved in terms of stability, performance, and extensibility, but it also added a number of new features. Finally, to improve the usability of the tools developed, we used the R web application framework shiny (CHANG *et al.*, 2016) to create the R/Bioconductor TCGAbiolinksGUI package, which provides a graphical interface for our packages and others of the Bioconductor project.

3.1 TCGAbiolinks: An R/Bioconductor package to download and analyze data from GDC

The aim of TCGAbiolinks is four-fold: (i) to facilitate data retrieval via GDC's API; (ii) to prepare the data using the appropriate preprocessing strategies; (iii) to provide a means to conduct different standard analyses and advanced integrative analyses and (iv) to allow the user to easily reproduce earlier research results. We introduce public methods used in several marker papers to integrate DNA methylation and gene expression data. In addition, our tool extracts published molecular subtype information for each TCGA sample within a tumor type (generally

embedded in supplementary tables, PDFs or external websites). The tool was developed in the R language specifically for integration within the Bioconductor project, thus we have provided most of the data objects as the Bioconductor specified 'SummarizedExperiment' class (HUBER *et al.*, 2015), thereby allowing easy integration with other data types and statistical methods that are common in the Bioconductor repository.

3.1.1 The TCGAbiolinks package

TCGAbiolinks is an R package, which is licensed under the General Public License (GPLv3), and is freely available through the Bioconductor repository (GENTLEMAN *et al.*, 2004). By conforming to the strict guidelines for package submission to Bioconductor, we were able to utilize and incorporate existing R/Bioconductor packages and statistics to assist in identifying differentially altered genomic regions defined by mutation, copy number, expression or DNA methylation; to reproduce previous TCGA marker studies; and to integrate data types both within TCGA and across other data types outside of TCGA. TCGAbiolinks consists of functions that can be grouped into three main levels: Data, Analysis and Visualization. More specifically, the package provides multiple methods for the analysis of individual experimental platforms (e.g. differential expression analysis or identifying differentially methylated regions or copy number alterations) and methods for visualization (e.g. survival plots, volcano plots and starburst plots) to facilitate the development of complete analysis pipelines. In addition, TCGAbiolinks offers in-depth integrative analysis of multiple platforms, such as copy number and expression or expression and DNA methylation, as demonstrated and applied in our recent TCGA study of 1122 gliomas (CECCARELLI *et al.*, 2016b). These functions can be used independently or in combination to provide the user with fully comprehensible analysis pipelines applied to TCGA data. A schematic overview of the package is presented in Figure 4. The next subsections describe each of the three main levels (Data, Analysis and Visualization), highlighting the importance and utility of each associated function and sub-function.

Data

TCGA data is accessible via the NCI Genomic Data Commons (GDC) data portal, GDC Legacy Archive and the Broad Institute's GDAC Firehose. The GDC Data Portal provides access to the subset of TCGA data that has been harmonized against Genome Reference Consortium Human Build 38 (GRCh38) (hg38) using GDC Bioinformatics Pipelines which provides methods to the standardization of biospecimen and clinical data, the re-alignment of DNA and RNA sequence data against a common reference genome build GRCh38, and the generation of derived data. Whereas the GDC Legacy Archive provides access to an unmodified copy of data that was previously stored in CGHub (WILKS *et al.*, 2014) and in the TCGA Data Portal hosted by the TCGA Data Coordinating Center (DCC), in which uses as references Genome Reference Consortium Human Build 37 (GRCh37) (hg19) and Genome Reference Consortium Human

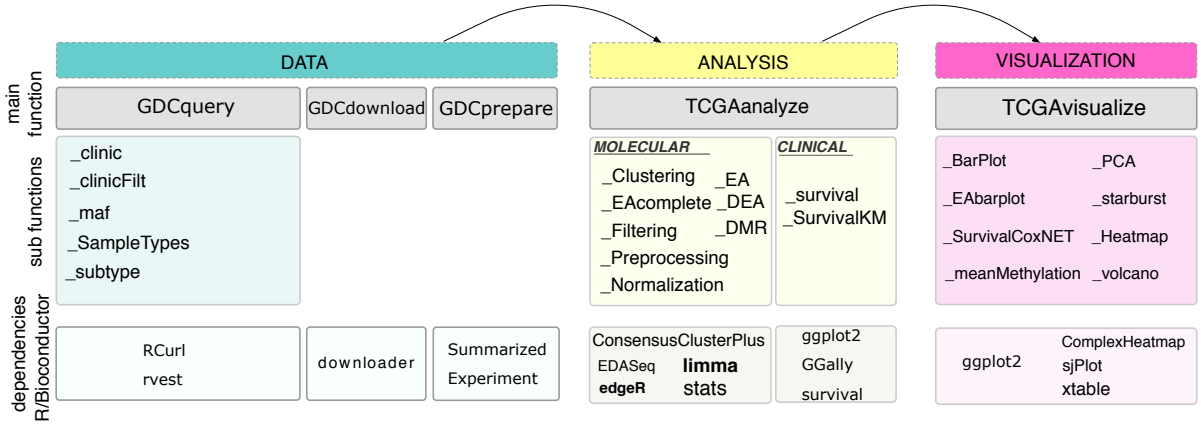


Figure 4 – Overview of TCGAbiolinks functions. TCGAbiolinks is organized in three categories. In the first category (Data), functions to query the GDC database, to download the data and to prepare it are made available. The second category (Analysis) contains functions that allow the user to carry out different types of analyses; these include clustering (*TCGAanalyze_Clustering*), differential expression analysis (*TCGAanalyze_DEA*) and enrichment analysis (*TCGAanalyze_EA*). Finally, the obtained results can be visualized using the functions in the third category (Visualization): these include principal component analysis (*TCGAvisualize_PCA*), starburst plots (*TCGAvisualize_starburst*) and survival curves (*TCGAvisualize_SurvivalCoxNET*). The different dependencies to other R/Bioconductor packages are specified in the last row of the figure.

Build 36 (GRCh36) (hg18).

The previously stored data in CGHub, TCGA Data Portal and Broad Institute's GDAC Firehose, were provided as different levels or tiers that were defined in terms of a specific combination of both processing level (raw, normalized, integrated) and access level (controlled or open access). Level 1 indicated raw and controlled data, level 2 indicated processed and controlled data, level 3 indicated Segmented or Interpreted Data and open access and level 4 indicated region of interest and open access data. While the TCGA data portal provided level 1 to 3 data, Firehose only provides level 3 and 4. An explanation of the different levels can be found at TCGA Wiki (https://wiki.nci.nih.gov/display/TCGA/Data+level). However, the GDC data portal no longer uses this based classification model in levels. Instead a new data model was created, its documentation can be found in GDC documentation at https://gdc.nci.nih.gov/developers/gdc-data-model/gdc-data-model-components. In this new model, data can be open or controlled access. While the GDC open access data does not require authentication or authorization to access it and generally includes high level genomic data that is not individually identifiable, as well as most clinical and all biospecimen data elements, the GDC controlled access data requires dbGaP authorization and eRA Commons authentication and generally includes individually identifiable data such as low level genomic sequencing data, germline variants, SNP6 genotype data, and certain clinical data elements. The process to obtain access to controlled data is found in GDC web site at https://gdc.nci.nih.gov/access-data/obtaining-access-controlled-data.

TCGAbiolinks divides the GDC data retrieval into three main functions: *GDCquery*, *GDCdownload* and *GDCprepare*. *GDCquery* allows the user to query data from the the The NCI's Genomic Data Commons (GDC) data portal or GDC Legacy Archive by accessing the

database API. Up to the moment, the GDC data portal provides data from two programs The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) on more than 38 diseases type cancer types and 5 different molecular data types (Transcriptome Profiling, Raw Sequencing Data, Copy Number Variation, DNA Methylation) as well as 3 different types of clinical reports (clinical tables, pathology reports and histology image slides). The pathology reports and histology image slides are not prepared but are downloaded from the GDC Legacy Archive to a directory if requested by the user. *GDCdownload* receives the output of *GDCquery*, a file manifest with all the metadata necessary to download it. In order to organize the files downloaded, this function will save the data with the following pattern "Root directory/project/source/data_category/data_type/file_id/file_name" (i.e Example: GDCdata/TCGA-GBM/harmonized/DNA_Methylation/Methylation_Beta_Value/079fcaff-3ae6-4150-b2e6-2b7330ffbcd9/jhu-usc.edu_GBM.HumanMethylation450.10.lvl-3.TCGA-19-A6J5-01A-21D-A33U-05.gdc_hg38.txt) If a file was previously downloaded it will not be re-downloaded. *GDCprepare* is a function that reads open processed data and prepares them for downstream analysis. Specifically, the objects are organized in a *SummarizedExperiment* object to allow easy integration with other Bioconductor packages, such as GRanges (LAWRENCE *et al.*, 2013), IRanges (LAWRENCE *et al.*, 2013), limma (RITCHIE *et al.*, 2015) and edgeR (ROBINSON; MCCARTHY; SMYTH, 2010). The samples are always referred to by their given TCGA/TARGET barcode. If the user prefers the data not to be prepared in a *SummarizedExperiment*, there is an option to set the argument *SummarizedExperiment* to FALSE; the data are then prepared as a standard data frame object (rows and columns).

Analysis

The analysis functions and subfunctions are designed to analyze TCGA data through both common and novel methods. The main function, called *TCGAanalyze*, comprises two distinct types of analysis: molecular analysis and clinical analysis. Once the data are prepared into data matrices (genes/loci in rows and samples in columns) or a SummarizedExperiment, the downstream analysis can be divided into (i) supervised analysis: differential expression analysis, enrichment analysis and master regulator analysis or (ii) unsupervised analysis: inference of gene regulatory network, clustering, classification, Receiver Operator Characteristics (ROC) (SONEGO; KOCSOR; PONGOR, 2008), Area Under the Curve (AUC), feature selection and survival analysis. *TCGAanalyze_Normalization* allows users to normalize mRNA transcripts and miRNA using the EDASeq package (RISSO *et al.*, 2011). This function uses within-lane normalization procedures to adjust for GC-content effects (or other gene-level effects) on read counts: LOESS robust local regression and global-scaling, full-quantile and between-lane normalization procedures to adjust for distributional differences between lanes (e.g. sequencing depth). *TCGAanalyze_DEA* allows the user to identify differential expression or regions between two populations or conditions. In particular, we used the edgeR package from Bioconductor, which uses the quantile-adjusted conditional maximum likelihood (qCML) method for experiments

with a single factor to detect differentially expressed genes (DEGs) (ROBINSON; MCCARTHY; SMYTH, 2010). Compared to several other estimators, qCML is the most reliable in terms of bias on a wide range of conditions; specifically, qCML performs best in situations involving many small samples with a common dispersion (ROBINSON; SMYTH, 2007). The P-values generated from the analysis are sorted in ascending order and corrected using the Benjamini & Hochberg procedure for multiple testing correction (BENJAMINI; HOCHBERG, 1995a). After running *TCGAanalyze_DEA*, it is possible to filter the output by fold change and/or significance and to use the *TCGAanalyze_LevelTab* function to create a table of DEGs, including fold change (FC), false discovery rate (FDR), gene expression levels of samples under conditions of interest and delta values (the difference in gene expression multiplied by logFC). *TCGAanalyse_DMR* allows the user to identify differentially methylated regions (DMRs) between two groups with a DNA methylation difference above a certain threshold. To calculate P-values, this subfunction uses the Wilcoxon ranksum statistical non-parametric test and adjusts the values using the FDR method. *TCGAanalyze_Clustering* allows the user to perform a hierarchical cluster analysis through two methods: ward.D2 and ConsensusClusterPlus (WILKERSON; HAYES, 2010).

Visualization

The visualization section allows the user to visualize the results generated by the analysis sections using heatmap, cluster, plots with incremental layers (ggplot2), pathway enrichment analysis and PCA. Furthermore, we provide methods to generate a starburst plot, which integrates gene expression and DNA methylation data (NOUSHMEHR *et al.*, 2010).

3.1.2 Comparisons

Recently, several tools to retrieve TCGA data sets have been made available, as summarized in Table 10. These tools include TCGA-Assembler (ZHU; QIU; JI, 2014), CGDS-R (GAO *et al.*, 2013), canEnvolve (SAMUR *et al.*, 2013), Firehose (DENG *et al.*, 2017), RCTCGAtoolbox (SAMUR, 2014), and cBioPortal (CERAMI *et al.*, 2012) . These tools can be divided into three representative categories. The first category comprises tools mainly used to download cancer genomics data, such as TCGA-Assembler and CGDSR. The second category includes tools that focus mainly on data analysis and integration, such as canEnvolve. The third category comprises tools to download and analyze data, such as RTCGAToolbox, Firehose and cBioPortal.

RTCGAToolbox is a tool that systematically accesses the Broad GDAC Firehose (https://gdac.broadinstitute.org/) preprocessed data and performs basic analysis and visualization of an individual data type (expression, mutation or DNA methylation). Despite the existence of TCGA specific software packages, none of these tools perform the integrative analysis harnessing methodologies designed by TCGA Analysis working groups (AWGs), such as identifying epigenetically silenced genes (represented in a starburst plot (NOUSHMEHR *et al.*, 2010) or functional copy number identification (CECCARELLI *et al.*, 2016b). Although RTCGAToolbox

Table 10 – Each column represents a software tool compared with TCGAbiolinks, and each row represents a feature. The cells checked with *X* indicates features that exists in the tool. Available platform abbreviations are defined as: R (R script); C (R package deposited in CRAN); B (Bioconductor package); W (available only as a web portal);

		Packages						
Features	**Sub-features**	TCGAbiolinks	TCGAAssembler	canEnvolve	TCGA2stat	Firehose-FirebrowserR	RTCGAtoolbox	cBio Portal CGDS-R
Availability	Platform	B	R	W	C	CW	B	CW
Genome of reference	Access to data aligned against the GRCh38/hg38	X	X					
	Access to data aligned against the GRCh37/hg19	X	X	X	X	X	X	X
Query TCGA Cases	Individual TCGA samples (e.g. TCGA-01-0001)	X	X			X		
Download	All TCGA platforms	X						
Data type analysis	mRNA	X		X	X	X	X	X
	miRNA	X		X	X	X	X	X
	copy number	X		X	X	X	X	X
	DNA methylation	X			X	X	X	X
	Clinical	X		X	X	X	X	X
	Protein			X		X		X
	Mutatation	X		X	X	X	X	X
Integrative analysis	DNA methylation and gene expression	X				X		
Other	Extensible to other BioC packages	X						

can download and analyze Firehose-generated data, neither tool can provide the downloaded data as a "SummarizedExperiment" object, which is critical for allowing the full integration and use of other popular Bioconductor packages, an integral aspect of Bioconductor (GENTLEMAN *et al.*, 2004, 2004). Briefly, the SummarizedExperiment class is a matrix-like container in which rows represent ranges of interest (as a GRanges or GRangesList object) and columns represent samples (with sample data summarized as a DataFrame). A Summarized Experiment contains one or more assays, each represented by a matrix-like object of numeric or other mode. Finally, TCGAbiolinks is able to access data aligned against the Genome Reference Consortium Human Build 38 (hg38) via the NCI Genomic Data Commons (GDC) data portal, and data aligned against the Genome Reference Consortium Human Build 19 (hg19) via the GDC Legacy Archive. This feature is only available using TCGA-Assembler.

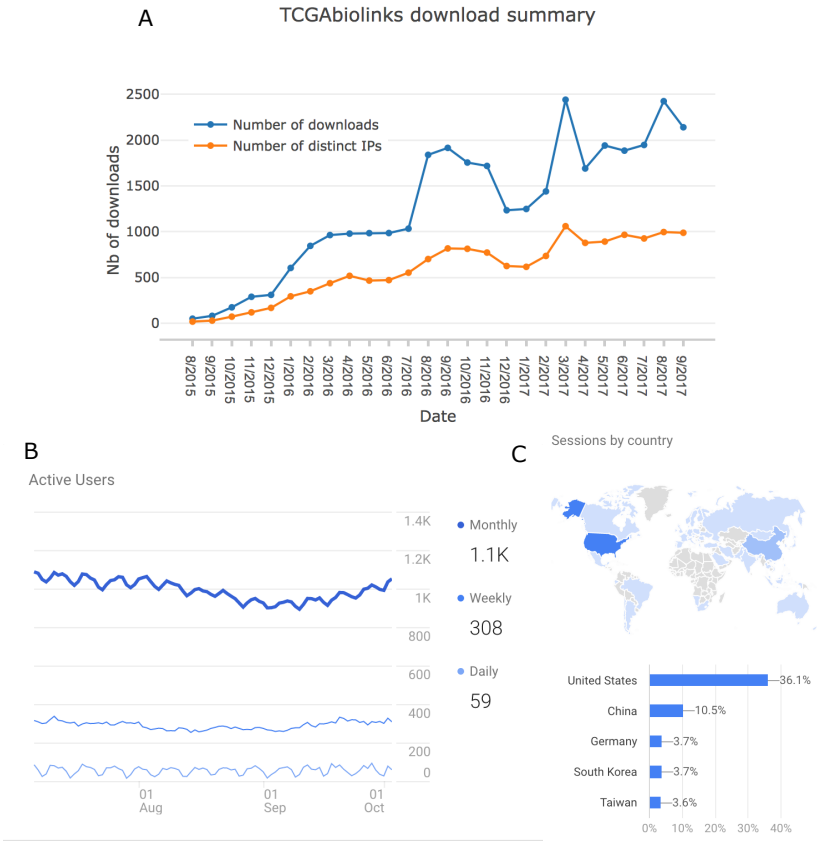


Figure 5 – TCGAbiolinks download summary. A) Number of downloads per month. B) Number of active users. C) Percentage of sessions connect to the documentation by country.

3.1.3 Software availability

TCGAbiolinks is available under the GNU General Public License version 3 (GNU GPL3). Its source code is available at https://github.com/BioinformaticsFMRP/TCGAbiolinks and a binary version for windows, macOSX and Linux is freely available through the Bioconductor repository at http://bioconductor.org/packages/TCGAbiolinks/. To execute this tool, it is required to have installed a R version $\geq$ 3.3.

3.1.4 Public reception

TCGAbiolinks had a good reception the research community being among the top 5% of the most downloaded tools of the Bioconductor project. In October 2017, the tool already had more than 17 thousand downloads, with an average of visits to the documentation pages of a thousand users per month. The Figure 5 shows a summary since the beginning of the project.

3.2 TCGAbiolinksGUI: A graphical user interface to analyze GDC cancer molecular and clinical data

Although TCGAbiolinks is a suitable R package for most data analysts with a strong knowledge and familiarity with R specifically those who can comfortably write R commands, we developed TCGAbiolinksGUI to enable user access to the methodologies offered in TCGAbiolinks and to give users the flexibility of point-and-click style analysis without the need to enter specific arguments. TCGAbiolinksGUI takes in all the important features of TCGAbiolinks and offers a graphics user interface (GUI) thereby eliminating any need to familiarize TCGAbiolinks' key functions and arguments.

3.2.1 Infrastructure

The TCGAbiolinksGUI user interface was created using Shiny, a Web Application Framework for R, and uses several packages to provide advanced features that can enhance Shiny apps, such as shinyjs to add JavaScript actions (ATTALI, 2017), shinydashboard to add dashboards (CHANG; Borges Ribeiro, 2017) and shinyFiles (PEDERSEN, 2016) to provide access to the server file system.

The following R/Bioconductor packages are used as back-ends for the data retrieval and analysis: TCGAbiolinks (COLAPRICO *et al.*, 2016) which allows to search, download and prepare data from the NCI's Genomic Data Commons (GDC) data portal into an R object and perform several downstream analysis; ELMER (Enhancer Linking by Methylation/Expression Relationship) (YAO *et al.*, 2015; SILVA *et al.*, 2017a) which identifies DNA methylation changes in distal regulatory regions and correlate these signatures with the expression of nearby genes to identify transcriptional targets associated with cancer; ComplexHeatmap (GU; EILS; SCHLESNER, 2016) to visualize data as oncoprint and heatmaps, pathview (LUO; BROUWER, 2013) which offers pathway based data integration and visualization; and maftools (MAYAKONDA; KOEFFLER, 2016) to analyze, visualize and summarize Mutation Annotation Format (MAF) files.

3.2.2 Graphical user interface design

The user interface has been divided into three main Graphical User Interface (GUI) menus. The first menu defines the acquisition of GDC data. The second defines the analysis steps which subdivides according to the molecular data types. And the third is dedicated to harnessing integrative analyses. We present below a brief description of each menu and their features that can be accessed through a side panel (see figure 6):

- **GDC Data:** Provides a guided approach to search for published molecular subtype information, clinical and molecular data. In addition, it downloads and processes the molecular

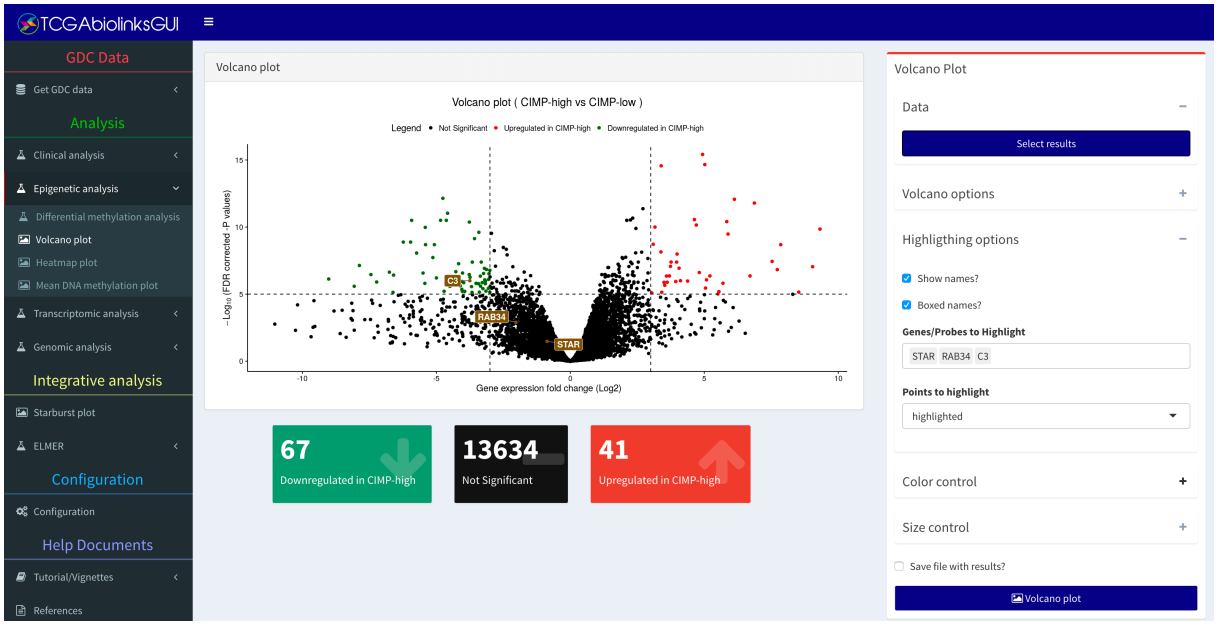


Figure 6 – The volcano plot menu of TCGAbiolinksGUI: The panel on the left shows the menus divided by different analyses, the panel on the right shows the controls available for the menu selected. In the center is a volcano plot window from the analysis menu. It is possible to control the colors, to change cut-offs and to export results into a CSV document.

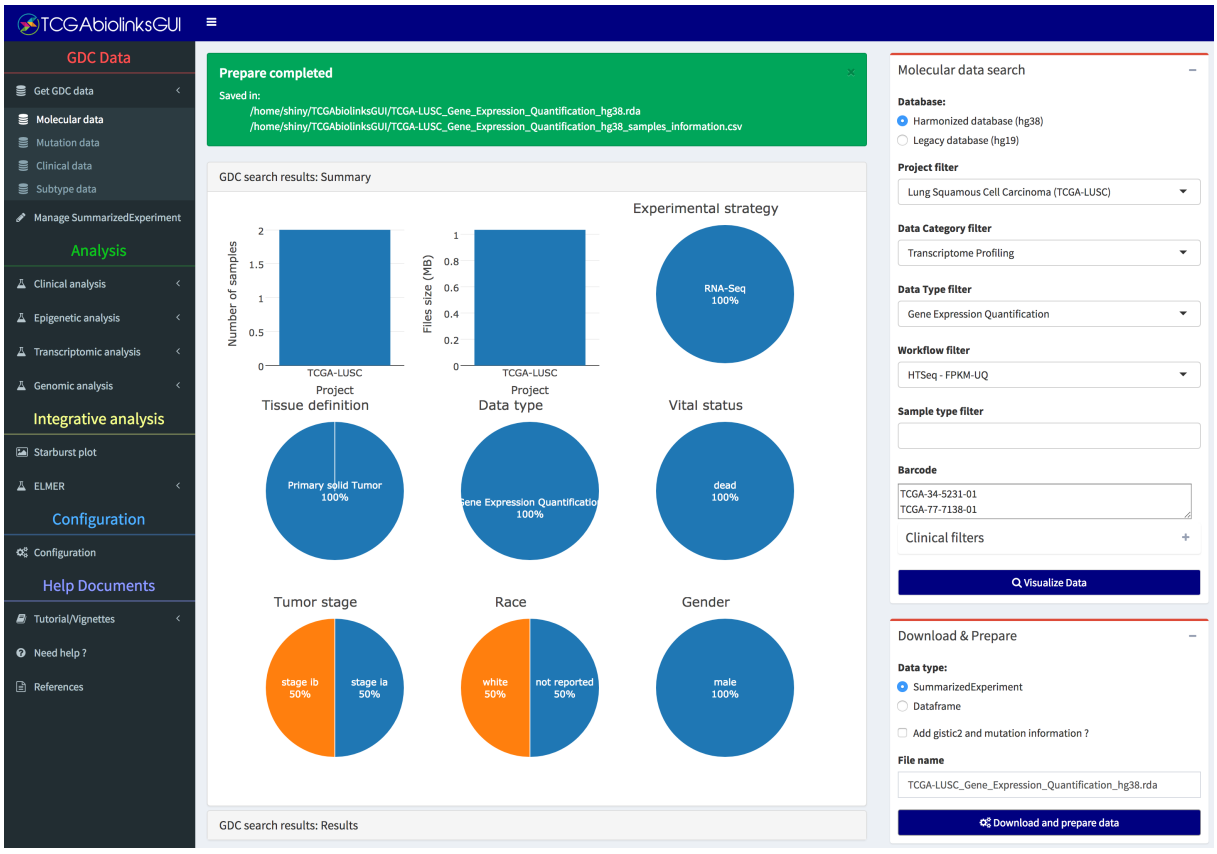


Figure 7 – Molecular data download: Gene expression Search, download and prepare into an R object of gene expression data for two TCGA-LUSC samples ("TCGA-34-5231-01","TCGA-77-7138-01"). The Summarized Experiment object created is saved as an RData file (TCGA-LUSC-Gene_Expression_Quantification_hg38.rda).

data into an R object that can be used for further analysis (Figure 7)

- **Clinical analysis:** Performs survival analysis to quantify and test survival differences between two or more groups of patients and draws survival curves with the 'number at risk' table, the cumulative number of events table and the cumulative number of censored subjects table using the R/CRAN package survminer (KASSAMBARA; KOSINSKI, 2017) (Figure 12).

- **Epigenetic analysis:** Performs a Differentially methylated regions (DMR) analysis, visualizes the results through both volcano and heatmap plots, and visualizes the mean DNA methylation level by groups (Figure 10).

- **Transcriptomic analysis:** Performs a Differential Expression Analysis (DEA), and visualizes the results through both volcano and heatmap plots. For the genes found as upregulated or downregulated an enrichment analysis can be performed and pathway data can be integrated (LUO; BROUWER, 2013) (Figure 9).

- **Genomic analysis:** Visualize and summarize the mutations from MAF (Mutation Annotation Format) files through summary plots and oncoplots using the R/Bioconductor maftools package (GU; EILS; SCHLESNER, 2016; MAYAKONDA; KOEFFLER, 2016) (Figures 11 and 8).

- **Integrative analysis:** Integrate the DMR and DEA results through a starburst plot. Also, using the DNA methylation data and the gene expression data the R/Bioconductor ELMER package can be used to discover functionally relevant genomic regions associated with cancer (YAO *et al.*, 2015; SILVA *et al.*, 2017a).

3.2.3 Documentation

We provide a guided tutorial for users via a vignette document which details each step and menu function available at http://bit.do/TCGAbiolinksDocs, via online documents available at http://bit.ly/TCGAbiolinks_PDFTutorials, and via YouTube video instructions showing step by step how each menu works available at http://bit.ly/TCGAbiolinksGUI_videoTutorials, which assist end-users in taking full advantage of TCGAbiolinksGUI. A demonstration version of the tool is available at http://tcgabiolinks.fmrp.usp.br:3838/. Users are encouraged to report and file bug reports or feature requests via our GitHub repository BioinformaticsFMRP/TCGAbiolinksGUI/issues.

3.2.4 Docker container

To further simplify the usability and accessibility of our tool, we provide a docker image compatible with most popular operating system available at

Figure 8 – TCGAbiolinksGUI: Visualizing mutation summary. This maftools plot available thought TCGAbiolinksGUI shows a summary of the MAF file. Highlighting the most mutated genes, SNV class and variant classification distributions.

Figure 9 – TCGAbiolinksGUI: Enrichement analysis of genes. TCGAbiolinks uses Gene Ontology which defines concepts/classes used to describe gene function, to perform an enrichement analysis. The plots shows molecular function, biological process, celullar components, and pathway gene Ontology classes. Each barplot represents a class with the numnber of genes in the class. The width of the barplot represents the significance, and the red line represents the percentage genes inputed that is found class, for example of all genes in the biological process regulation of endothelial cell migration our 2 genes inputed represents 20% of all genes in this process.

Figure 10 – TCGAbiolinksGUI: Visualizing DMR results as heatmap. Plot shows the DMR results: hypermethylated probes in Solid tissue normal samples compared to the Primary Solid tumor samples. Each column is a sample, while each row is a probe. Blue colors represents probes with low levels of DNA methylation and red the ones with high level.

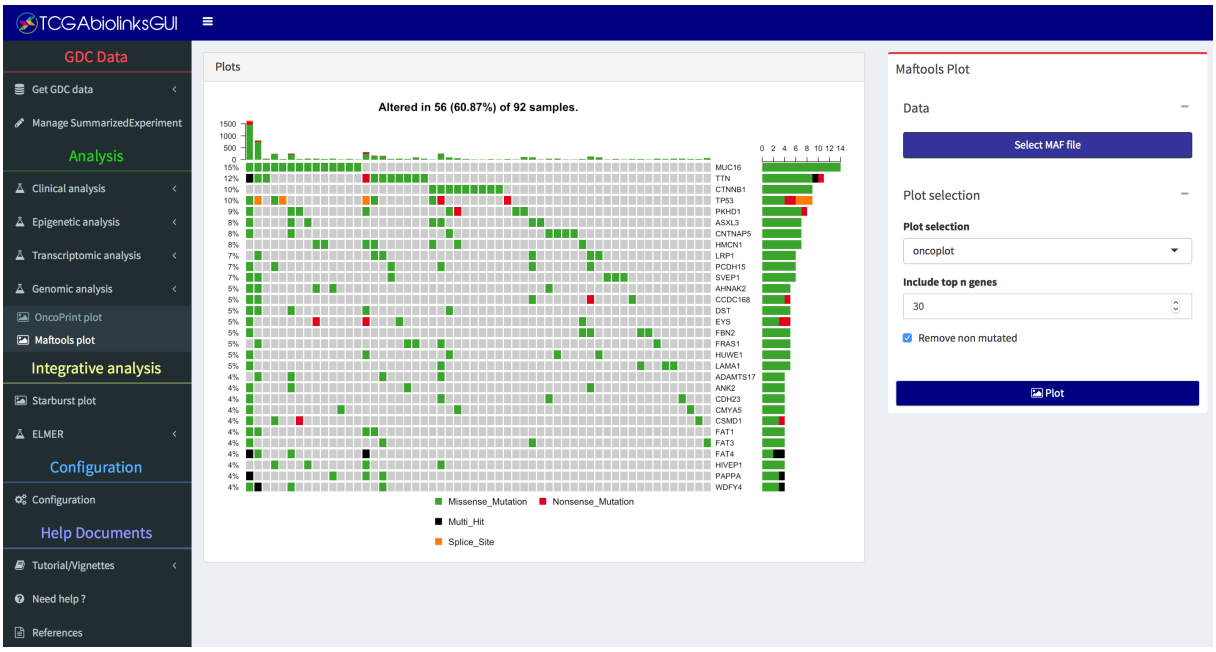


Figure 11 – TCGAbiolinksGUI:Visualizing mutation as an oncoplot. Each column represents a sample and each row a different gene. The top barplot has the frequency of mutations for each patient, while the right barplot has the frequency of mutations for each gene. The plot by default is ordered by the most mutated genes.

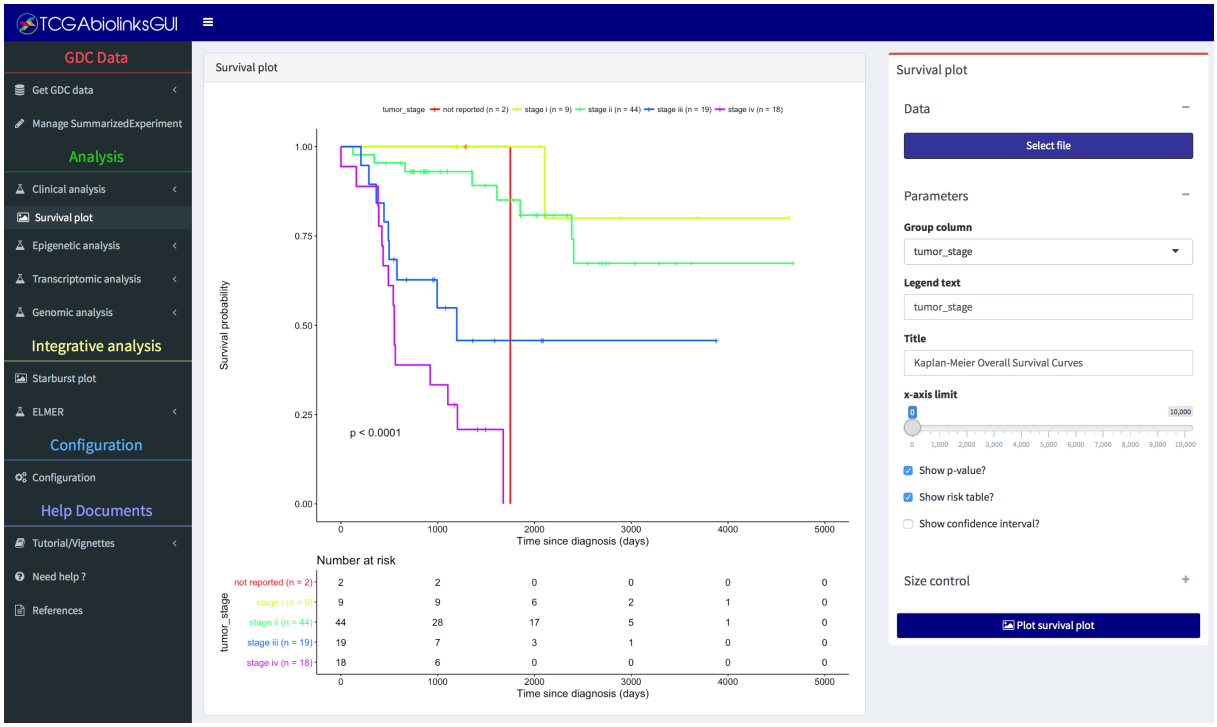


Figure 12 – TCGAbiolinksGUI: Survival analysis. The plot shows the Kaplan-Meier Overall Survival Curves for ACC (Adrenocortical carcinoma) stratified by tumor stage. As expected, hihger levels which are more agressive have a lower survival.

https://hub.docker.com/r/tiagochst/tcgabiolinksgui/. This file allows users to run TCGAbiolinksGUI without the need to install associated dependencies or configure system files, common steps required to run R installations and load R/Bioconductor packages.

3.2.5 Comparison of alternative software

Web tools used for cancer data analysis might be classified into two broad groups. The first group only provides an interface to existing software analysis tools. The Galaxy project (https://galaxyproject.org/), which is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research, is an example of such a tool that belongs to this group. The other group is composed of exploratory tools mainly focused on the visualization of processed data and pre-computed results. The cBioPortal project (GAO *et al.*, 2013; CERAMI *et al.*, 2012), by providing several visualizations for mining the TCGA data, is an example of a tool that falls within this classification.

If one were to classify TCGAbiolinksGUI, it would belong to the first group. Compared to the Galaxy project, TCGAbiolinksGUI offers an open platform which improves the accessibility of R/Bioconductor packages, allowing users an advantage to integrate their features with existing Bioconductor packages without the need to go beyond the R/Shiny frameworks as a common feature from the Galaxy project, which requires the interface elements to be structured through XML files (TURAGA *et al.*, 2016). In addition, going beyond the R/Bioconductor environment requires more software dependencies which make the process to install Galaxy to

use R/Bioconductor packages laborious. On the other hand, compared to cBioPortal, TCGAbiolinksGUI allows users to perform deep integrative analysis by comparing different subtypes of data (i.e. performing an integrative analysis to compare breast cancer samples with a mutation on FOXA1 gene compared to wild-type samples using DNA methylation, gene expression, and motif enrichment analysis on genomic regions of interest). Although cBioPortal offers these features, it would require users to process each step independently and download outside of cBioPortal in order to perform such integrative analysis.

3.3 Enhancer Linking by Methylation/Expression Relationships (ELMER)

Motivated by the discovery of transcriptional enhancers in tissue DNA methylation data (BERMAN *et al.*, 2012), and subsequent approaches to linking these enhancers to transcriptional targets using a chromQTL approach (ARAN; HELLMAN, 2013) (reviewed in Yao *et al.*), Yao *et al.* developed the the R/Bioconductor *ELMER* (Enhancer Linking by Methylation/Expression Relationships) package, a tool which infers regulatory element landscapes and transcription factor networks from cancer methylomes.

This tool combined DNA methylation and gene expression data from human tissues to infer multi-level cis-regulatory networks through several steps which included the identification of distal enhancer probes with significantly altered DNA methylation levels in primary tumor tissues compared to normal tissues, followed by the identification of putative target genes, and a comprehensive gene regulatory network analysis which combined transcription factor motifs at the altered enhancers with TF expression to identify the underlying master regulators. This approach identified several known and unknown master regulators in TCGA data, such as GATA-binding protein 3 (GATA3) and Forkhead box protein A1 (FOXA1) in breast cancer, and tumor protein p63 (P63) and Sex determining region Y-box 2 (SOX2) in squamous cell lung carcinoma (YAO *et al.*, 2015; SILVA *et al.*, 2016).

Based on user feedback and a full review of the source code, we identified and implemented a number of software improvements, which are summarized in table 11: (i) The original package contained no standard data structure to handle multiple assays (DNA methylation, gene expression, and clinical data), which would be required for an integrative genomic data analysis. Recently, the Bioconductor team provided such a data structure through the MultiAssayExperiment package. (ii) All auxiliary databases (human TF list, classification of TF in families, gene annotation, DNA methylation annotation and motif occurrences within probe sites) used in the package were created and maintained manually, thereby making the upgrade process laborious; thus, we automated this process. (iii) The package was developed to analyze primary tumor tissue samples compared to normal tissues samples, thus not allowing arbitrary subgroups to be compared (for instance mutants vs. non-mutants, treated vs. untreated, etc.) (iv)

Table 11 – Main differences between ELMER old version (v.1) and the new version (v.2)

Features	ELMER Version 1	ELMER Version 2
Primary data structure	mee object (custom data structure)	MAE object (Bioconductor data structure)
Auxiliary data	Manually created	Programmatically created
Number of human TFs	1,982	2,014 (UniProt database)
Number of TF motifs	91	771 (HOCOMOCO v11 database)
TF classification	78 families	82 families and 331 subfamilies (TFClass database, HOCOMOCO)
Analysis performed	Normal vs tumor samples	Group 1 vs group 2
Statistical grouping	Unsupervised only	Unsupervised or supervised using labeled groups
TCGA data source	The Cancer Genome Atlas (TCGA) (not available)	The NCI's Genomic Data Commons (GDC)
Genome of reference	GRCh37 (hg19)	GRCh37 (hg19)/GRCh38 (hg38)
DNA methylation platforms	HM450	EPIC and HM450
Graphical User Interface (GUI)	None	TCGAbiolinksGUI
Automatic report	None	HTML summarizing results
Annotations	None	StateHub

Our original approach used known epigenomic markers for enhancers to constrain the genomic regions searched for differential methylation. However, this selection could limit our algorithm to identifying regulatory networks for tissue types that exist in the epigenomic databases; we found this constraint problematic, and thus now search *all* distal regulatory regions without any such filter. (v) The function used to download data from The Cancer Genome Atlas (TCGA) data portal (TOMCZAK *et al.*, 2015) broke when the TCGA site was shutdown and its data transferred to The NCI's Genomic Data Commons (GDC) (GROSSMAN *et al.*, 2016); we now have a more general data provider interface that supports GDC as the default provider. (vi) The package only supported data aligned to Genome Reference Consortium GRCh37 (hg19), and we now provide support for Genome Reference Consortium GRCh38 (hg38). (vii) There was no support to the recent HumanMethylationEPIC (EPIC) array. In addition to the specific improvements listed above, we substantially re-wrote most of the code to be more efficient and maintainable, also most of the output plots generated were improved.

In this section, we present a new version of the R *ELMER* package, which addresses all the issues described above.

3.3.1 Implementation

Here we describe each of following analysis steps shown in figure 13.

- Organize data as a *MultiAssayExperiment* object
- Identify distal probes with significantly different DNA methylation level when comparing two sample groups.
- Identify putative target genes for differentially methylated distal probes, using methylation vs. expression correlation
- Identify enriched motifs for each probe belonging to a significant probe-gene pair
- Identify master regulatory Transcription Factors (TF) whose expression associate with DNA methylation changes at multiple regulatory regions.

3.3.1.1 Organization of data as a MultiAssayExperiment object

To facilitate the analysis of experiments and studies with multiple samples the Bioconductor team created the *SummarizedExperiment* class (HUBER *et al.*, 2015), a data structure able to store data and metadata for a single experiment but not for data spanning several experiments for the same sample. To overcome this problem, recently, the MultiAssay SIG (Special Interest Group) created the MultiAssayExperiment class (SIG, 2017) a data structure to manage and preprocess multiple assays for integrated genomic analysis. This data structure is now an input for all main functions of *ELMER* and can be generated by the *createMAE* function.

To perform *ELMER* analyses, users need to populate a *MultiAssayExperiment* with a DNA methylation matrix or *SummarizedExperiment* object from HumanMethylation450 BeadChip (HM450) or MethylationEPIC BeadChip (EPIC) platform; a gene expression matrix or SummarizedExperiment object for the same samples; a matrix mapping DNA methylation samples to gene expression samples; and a matrix with sample metadata (i.e. clinical data, molecular subtype, etc.). If TCGA data are used, the last two matrices will be automatically generated. If using non-TCGA data, the matrix with sample metadata should be provided with at least a column with a patient identifier and another one identifying its group which will be used for analysis, if samples in the methylation and expression matrices are not ordered and with same names, a matrix mapping for each patient identifier their DNA methylation samples and their gene expression samples should be provided to the *createMAE* function. Based on the genome of reference selected, metadata for the DNA methylation probes, such as genomic coordinates, will be added from Zhou, Laird and Shen (2016); and metadata for gene expression and annotation is added from ENSEMBL database (YATES *et al.*, 2015) using biomaRt (DURINCK *et al.*, 2009).

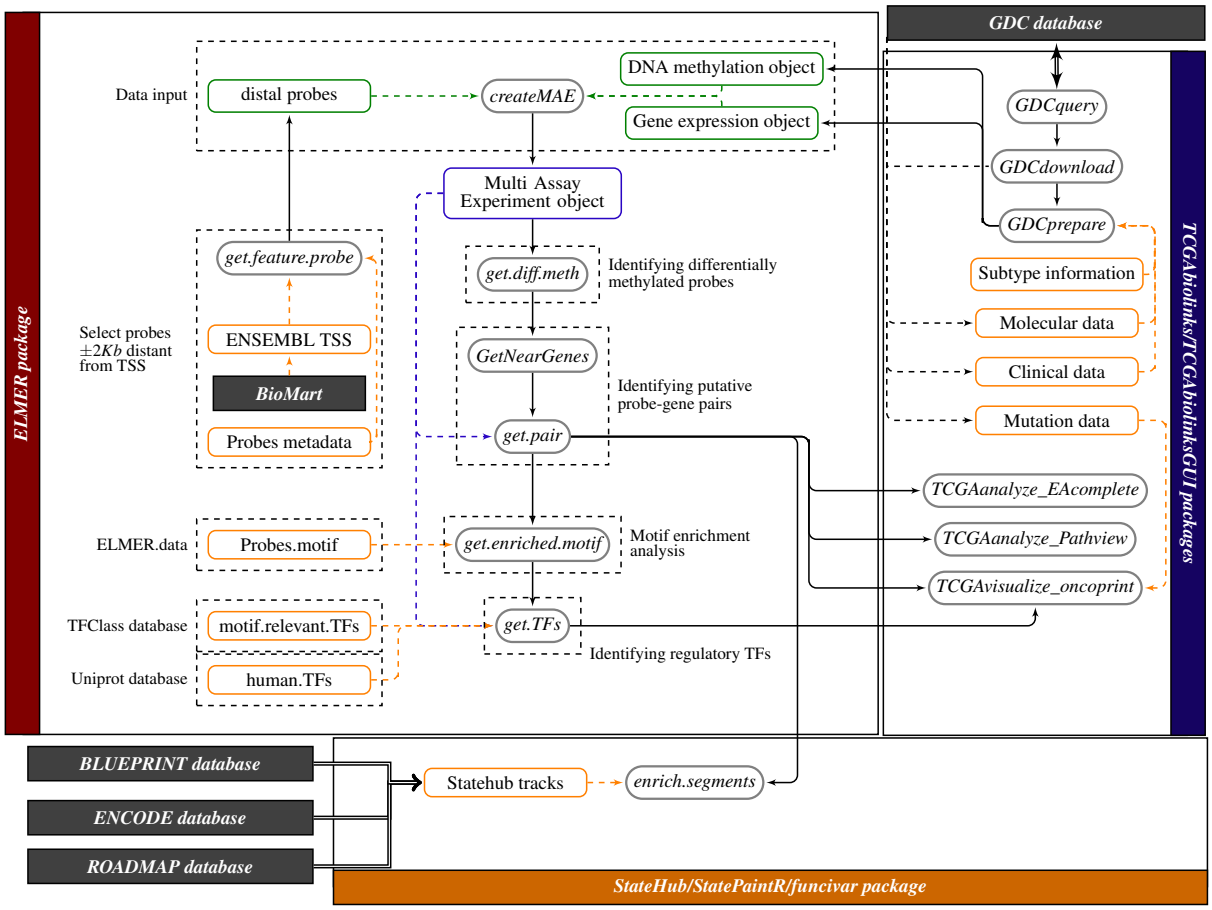


Figure 13 – ELMER workflow: ELMER receives as input a DNA methylation object, a gene expression object (a matrix or a SummarizedExperiment object) and a Genomic Ranges (GRanges) object with distal probes to be used as filter which can be retrieved using the *get.feature.probe* function. The function *createMAE* will create a Multi Assay Experiment object keeping only samples that have both DNA methylation and gene expression data. Genes will be mapped to genomic position and annotated using ENSEMBL database (AKEN *et al.*, 2016), while for probes it will add annotation from ZHOU; LAIRD; SHEN (http://zwdzwd.github.io/InfiniumAnnotation) . This MAE object will be used as input to the next analysis functions. First, it identifies differentially methylated probes followed by the identification of their nearest genes (10 upstream and 10 downstream) through the *get.diff.meth* and *GetNearGenes* functions respectively. For each probe, it will verify if any of the nearby genes were affected by its change in the DNA methylation level and a list of gene and probes pairs will be outputted from *get.pair* function. For the probes in those pairs, it will search for enriched regulatory Transcription Factors motifs with the *get.enriched.motif* function. Finally, the enriched motifs will be correlated with the level of the transcription factor through the *get.TFs* function. In the figure green Boxes represents user input data, blue boxes represent output object, orange boxes represent auxiliary pre-computed data and gray boxes are functions.

3.3.1.2 Selecting distal probes

Probes from HumanMethylationEPIC (EPIC) array and Infinium HumanMethylation450 (HM450) array are removed from the analysis if they have either internal SNPs close to the $3'$ end of the probe; non-unique mapping to the bisulfite-converted genome; or off-target hybridization due to partial overlap with non-unique elements (ZHOU; LAIRD; SHEN, 2017). This probe metadata information is included in *ELMER.data* package, populated from the source file at http://zwdzwd.github.io/InfiniumAnnotation (ZHOU; LAIRD; SHEN, 2017). To limit ELMER to the analysis of distal elements, probes located in regions of $\pm 2kb$ around transcription start sites (TSSs) were removed.

3.3.1.3 Identification of differentially methylated CpGs (DMCs)

For each distal probe, samples of each group (group 1 and group 2) are ranked by their DNA methylation beta values, those samples in the lower quintile (20% samples with the lowest methylation levels) of each group are used to identify if the probe is hypomethylated in group 1 compared to group 2, using an unpaired one-tailed t-test. The 20% is a parameter to the *diff.meth* function called *minSubgroupFrac*. For the (ungrouped) cancer case, this is set to 20% as in Yao *et al.* (2015), because we typically wanted to be able to detect a specific molecular subtype among the tumor samples; these subtypes often make up only a minority of samples, and 20% was chosen as a lower bound for the purposes of statistical power (high enough sample numbers to yield t-test p-values that could overcome multiple hypothesis corrections, yet low enough to be able to capture changes in individual molecular subtypes occurring in 20% or more of the cases.) This number can be set arbitrarily as an input to the *diff.meth* function and should be tuned based on sample sizes in individual studies. In the *Supervised* mode, where the comparison groups are implicit in the sample set and labeled, the *minSubgroupFrac* parameter is set to 100%. An example would be a cell culture experiment with 5 replicates of the untreated cell line, and another 5 replicates that include an experimental treatment.

To identify hypomethylated differentially methylated CpGs (DMCs), a one-tailed t-test is used to rule out the null hypothesis: $\mu_{group1} \geq \mu_{group2}$, where μ_{group1} is the mean methylation within the lowest group 1 quintile (or another percentile as specified by the *minSubgroupFrac* parameter) and μ_{group2} is the mean within the lowest group 2 quintile. Raw p-values are adjusted for multiple hypothesis testing using the Benjamini-Hochberg method (BENJAMINI; HOCHBERG, 1995b), and probes are selected when they had adjusted p-value less than 0.01 (which can be configured using the *pvalue* parameter). For additional stringency, probes are only selected if the methylation difference: $\Delta = \mu_{group1} - \mu_{group2}$ was greater than 0.3. The same method is used to identify hypermethylated DMCs, except we use the *upper* quintile, and the opposite tail in the t-test is chosen.

3.3.1.4 Identification of putative target gene(s)

For each differentially methylated distal probe (DMC), the closest 10 upstream genes and the closest 10 downstream genes are tested for inverse correlation between methylation of the probe and expression of the gene (the number 10 can be changed using the *numFlankingGenes* parameter). To select these genes, the probe-gene distance is defined as the distance from the probe to the transcription start site specified by the ENSEMBL gene level annotations (YATES *et al.*, 2015) accessed via the R/Bioconductor package biomaRt (DURINCK *et al.*, 2009; DURINCK *et al.*, 2005). By choosing a constant number of genes to test for each probe, our goal is to avoid systematic false positives for probes in gene rich regions. This is especially important given the highly non-uniform gene density of mammalian genomes. Thus, exactly 20 statistical tests were performed for each probe, as follows.

For each probe-gene pair, the samples (all samples from both groups) are divided into two groups: the M group, which consisted of the upper methylation quintile (the 20% of samples with the highest methylation at the enhancer probe), and the U group, which consists of the lowest methylation quintile (the 20% of samples with the lowest methylation.) The 20% ile cutoff is a configurable parameter *minSubgroupFrac* in the *get.pair* function. As with its usage in the *diff.meth* function, the default value of 20% is a balance, allowing for the identification of changes in a molecular subtype making up a minority (i.e. 20%) of cases, while also yielding enough statistical power to make strong predictions. For larger sample sizes or other experimental designs, this could be set even lower.

For each candidate probe-gene pair, the Mann-Whitney U test is used to test the null hypothesis that overall gene expression in group M is greater than or equal than that in group U. This non-parametric test was used in order to minimize the effects of expression outliers, which can occur across a very wide dynamic range. For each probe-gene pair tested, the raw p-value P_r is corrected for multiple hypothesis using a permutation approach as follows. The gene in the pair is held constant, and x random methylation probes are chosen to perform the same one-tailed U test, generating a set of x permutation p-values P_p. We chose the x random probes only from among those that were "distal" (farther than $2kb$ from an annotated transcription start site), in order to draw these null-model probes from the same set as the probe being tested (SHAM; PURCELL, 2014). An empirical p-value P_e value was calculated using the following formula (which introduces a pseudo-count of 1):

$$P_e = \frac{num(P_p \leq P_r) + 1}{x + 1} \tag{3.1}$$

Notice that in the *Supervised* mode, no additional filtering is necessary to ensure that the M and U group segregate by sample group labels. The two sample groups are segregated by definition, since these probes were selected for their differential methylation, with the same directionality, between the two groups (Figure 14).

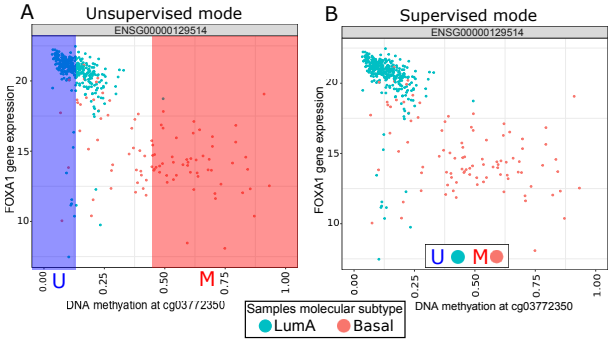


Figure 14 – Supervised mode maximizes statistical power. Difference of groups *U* and *M* definition in *supervised* and *unsupervised* mode. A: *unsupervised* mode; when minSubgroupFrac argument is set to 40%, the methylated group is defined as the highest quintile and the unmethylated group as the lowest quintile; B: *supervised* mode; methylated and unmethylated group are defined as one of the known molecular subtypes. For example, the unmethylated group is represented by all the LumA samples while the methylated group is represented by all the Basal samples. The t-test p-value achieved for the Unsupervised mode is $8.3E-25$, while the Supervised mode is : $1.43E-43$.

3.3.1.5 Characterization of chromatin state context of enriched probes using FunciVar

Unlike version 1 of *ELMER*, we now consider *all* distal probes in the identification of regulatory elements. DNA methylation is known to affect several different classes of distal chromatin state element, including active enhancers, poised enhancers, and insulators. In order to provide a functional interpretation of the regulatory elements identified by *ELMER*, we perform a chromatin state enrichment analysis of the probes within significant probe-gene pairs, using the *statePaintR* tools from the www.statehub.org (COETZEE *et al.*, 2017), along with our new FunciVar package (FUNCIVAR, 2017). Enrichment of the putative pairs within chromatin states is calculated against a background model that uses the distal probe set that the putative pairs are drawn from.

3.3.1.6 Motif enrichment analysis

In order to identify enriched motifs and potential upstream regulatory TFs, all probes with occurring in significant probe-gene pairs are combined for motif enrichment analysis. Hypergeometric Optimization of Motif EnRichment (HOMER) (HEINZ *et al.*, 2010) is used to find motif occurrences in a $\pm 250bp$ region around each probe, using HOCOMOCO (HOmo sapiens COmprehensive MOdel COllection) v11 (KULAKOVSKIY *et al.*, 2016). Transcription factor (TF) binding models are available at http://hocomoco.autosome.ru/downloads (using the HOMER specific format with threshold score levels corresponding to p-value $\leq 10^{-4}$).

For each probe set tested (i.e. the set of all probes occurring in significant probe-gene pairs), we quantify enrichments using Fisher's exact test (where a is the number of probes within

the selected probe set that contains one or more motif occurrences; b is the number of probes within the selected probe set that do not contain a motif occurrence; c and d are the same counts within the entire array probe set drawn from the same set of distal-only probes using the same definition as the primary analysis) and multiple testing correction with the Benjamini-Hochberg procedure (FISHER, 1922).

A probe set was considered significantly enriched for a particular motif if the 95% confidence interval of the Odds Ratio was greater than 1.1 (specified by option *lower.OR*, 1.1 is default), the motif occurred at least 10 times (specified by option *min.incidence*, 10 is default) in the probe set and $FDR < 0.05$.

3.3.1.7 Identification of master regulator TFs

When a group of enhancers is coordinately altered in a specific sample subset, this is often the result of an altered upstream *master regulator* transcription factor in the gene regulatory network. *ELMER* tries to identify such transcription factors corresponding to each of the TF binding motifs enriched from the previous analysis step. For each enriched motif, *ELMER* takes the average DNA methylation of all distal probes (in significant probe-gene pairs) that contain that motif occurrence (within a $\pm 250bp$ region) and compares this average DNA methylation to the expression of each gene annotated as a human TF.

A statistical test is performed for each motif-TF pair, as follows. All samples are divided into two groups: the M group, which consists of the 20% of samples with the highest average methylation at all motif-adjacent probes, and the U group, which consisted of the 20% of samples with the lowest methylation. This step is performed by the *get.TFs* function, which takes *minSubgroupFrac* as an input parameter, again with a default of 20%. For each candidate motif-TF pair, the Mann-Whitney U test is used to test the null hypothesis that overall gene expression in group M is greater or equal than that in group U. This non-parametric test was used in order to minimize the effects of expression outliers, which can occur across a very wide dynamic range. For each motif tested, this results in a raw p-value (P_r) for each of the human TFs. All TFs are ranked by their $-log_{10}(Pr)$ values, and those falling within the top 5% of this ranking were considered candidate upstream regulators. The best upstream TFs which are known to recognize to specific binding motif are automatically extracted as putative regulatory TFs, and rank ordered plots are created to visually inspect these relationships, as shown in the example below. Because the same motif can be recognized by many transcription factors of the same binding domain family, we define these relationships at both the family and subfamily classification level using the classifications from TFClass database (WINGENDER; SCHOEPS; DÖNITZ, 2013). Use of this database is a major change from version 1 of ELMER, which used custom curations for DNA binding domain families. Use of the TFClass database is preferable because it is well curated and regularly updated to reflect new findings.

Data availability

The TCGA data was downloaded from the NCI Genomic Data Commons (GDC) data portal (GROSSMAN *et al.*, 2016) using TCGAbiolinks R/Bioconductor package (COLAPRICO *et al.*, 2015; SILVA *et al.*, 2016). Gene annotations were retrieved from ENSEMBL (YATES *et al.*, 2015) database via biomaRt R/Bioconductor package (DURINCK *et al.*, 2005; DURINCK *et al.*, 2009). DNA methylation microarrays metadata were retrieved from http://zwdzwd.github.io/InfiniumAnnotation (ZHOU; LAIRD; SHEN, 2017). Transcription factor (TF) binding models can be downloaded at HOCOMOCO database (http://hocomoco.autosome.ru/) (KULAKOVSKIY *et al.*, 2016). The list of human TF can be accessed at http://www.uniprot.org/ (APWEILER *et al.*, 2004). The classification of human transcription factors (TFs) can be viewed at http://tfclass.bioinf.med.uni-goettingen.de/tfclass (WINGENDER; SCHOEPS; DÖNITZ, 2013).

Software availability

ELMER source code is available at https://github.com/tiagochst/ELMER and the auxiliary data files are available https://github.com/tiagochst/ELMER.data. *ELMER* is available under the GNU General Public License version 3 (GNU GPL3).

CHAPTER 4

CANCER DATA ANALYSIS

This chapter presents results of the analysis of cancer data using the software and algorithms presented in the previous chapter. In section 4.1 we present some use cases using mainly the TCGAbiolinks package. In section 4.2 we present some data analysis using the ELMER package. In section 4.3, using the previous tools described we perfom a glioma analysis focused specially in the two molecular subtypes G-CIMP-low and G-CIMP-high discovered by our laboratory and collaborators.

4.1 Use cases using TCGAbiolinks

In this section, we introduce and describe the utility and application of TCGAbiolinks through some use cases. In subsection 4.1.1 we show a lower-grade glioma downstream analysis with gene expression. In subsection 4.1.2 we show a downstream analysis integration of gene expression and methylation data of colon adenocarcinoma data, describing how to generate a starburst plot (NOUSHMEHR *et al.*, 2010), which was introduced to illustrate the results of integrating DNA methylation and gene expression data.

4.1.1 Lower-grade glioma downstream analysis with gene expression

For this case study, we used the recently available lower-grade glioma (LGG) data to investigate gene expression differences between the reported molecular subtypes (IDHmutant, IDHwildtype and IDHmutant codels) (PLATFORMS, 2015). In particular, we used TCGAbiolinks to download 293 samples profiled using messenger RNA expression (IlluminaHiSeq RNASeqV2) with available molecular subtypes. The data was normalized using the *TCGAanalyze_Normalization* function and we applied three filters to remove features/mRNAs with low signals across samples, obtaining 4578, 4284 and 1187 mRNAs, respectively. A clustering analysis was then applied using the ConsensusClusterPlus package (WILKERSON; HAYES, 2010) which identified four distinct groups of samples (EC1-EC4) (Figure 15A). The survival curves for each cluster were generated using *TCGAanalyze_survival* and are shown in Figure 15B. As expected,

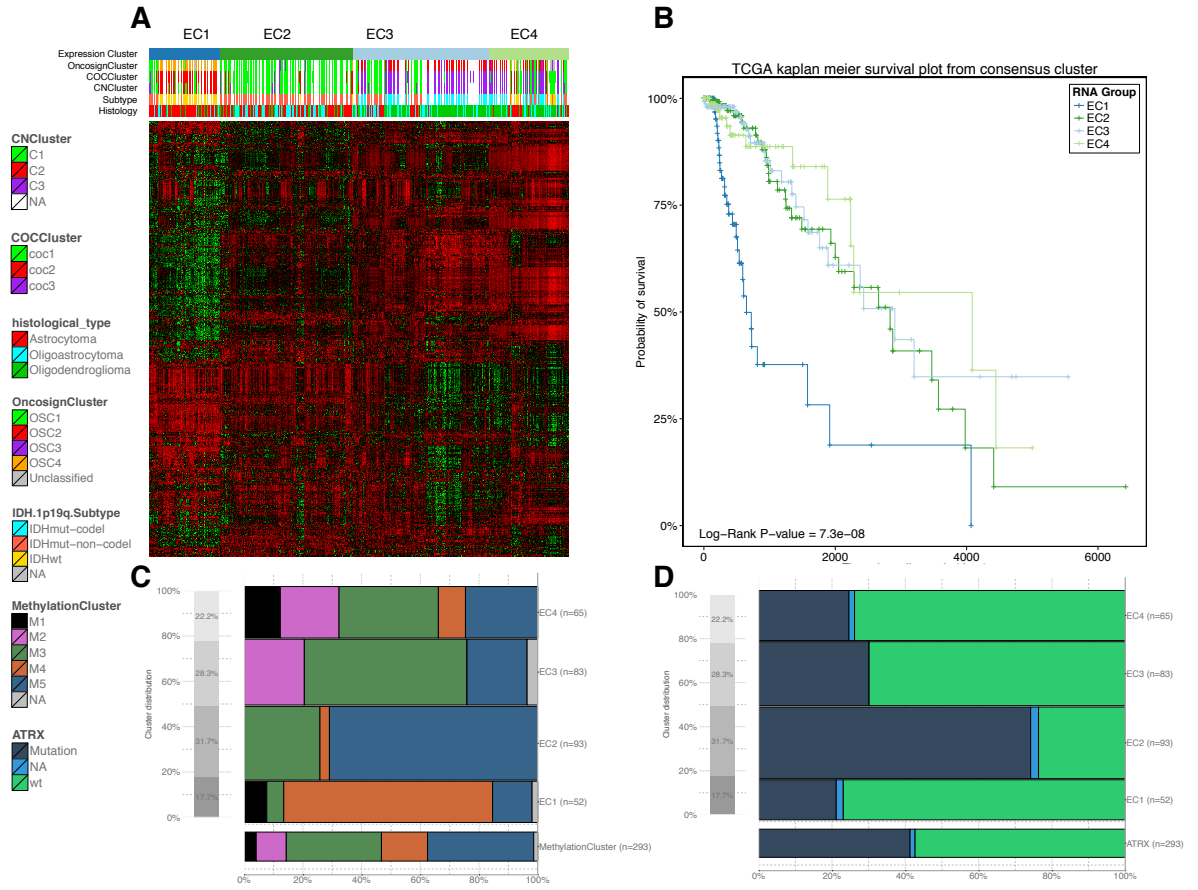


Figure 15 – Case study - Integrative (or Downstream) analysis of gene expression and clinical data from LGG disease with unsupervised clustering and crossing expression clusters with clinical and molecular information. **(A)** Heatmap of 1187 more variables genes clustered with tree $k = 4$ in EC1, EC2, EC3, EC4. **(B)** Kaplan Meier survivals plot for EC clusters. **(C and D)** Distribution of the DNA Methylation clusters and ATRX mutation within the EC clusters.

each cluster effectively separated IDHwildtype tumors (EC1) from IDHmutant-non-codel (EC2) and IDHmutant-codel tumors (EC3 and EC4) (Figure 15). Additional biological subtypes (DNA methylation subtypes) were reproduced as expected (Figure 15D) (PLATFORMS, 2015).

4.1.2 Downstream analysis integration of gene expression and methylation data

The DNA methylation of specific promoter CpG islands has the potential to influence gene expression. In this case study, we used TCGAbiolinks to examine the biological relationship between DNA methylation and gene expression in Colon adenocarcinoma (COAD). Using *GDCquery*, *GDCdownload* and *GDCprepare*, we obtained DNA methylation data (Infinium HumanMethylation450 and Infinium HumanMethylation27 platforms) and gene expression data (IlluminaGA RNASeqV2 platform) for the same TCGA COAD samples (NETWORK *et al.*, 2012). A supervised analysis was performed on the molecular subtypes CIMP-Low [CIMP.L] and CIMP-High [CIMP.H]. The gene expression analysis started by the identification of outliers,

followed by the normalization methods. Using *TCGAanalyze_DEA*, 34 DEGs ($log_2FC \geq 3.0$ and FDR $\leq 10^{-4}$) were identified. The result of this analysis is represented in a volcano plot (Figure 16A) created using *TCGAVisualize_volcano*. For the DNA methylation analysis, using *TCGAanlayze_DMR* we identified 73 CpG-methylated probes ($\Delta\overline{\beta} \geq 0.25$ and $FDR \leq 10^{-5}$; (Figure 16B). The DNA methylation and gene expression results were integrated as in the previous TCGA marker paper (NOUSHMEHR *et al.*, 2010; NETWORK *et al.*, 2012), by generating a starburst plot (Figure 16C) in which the x-axis is the log_{10} of the correct P-value for DNA methylation and the y-axis is the log_{10} of the correct P-value for the expression data. The starburst plot highlights nine distinct quadrants. To incorporate the DNA methylation difference cut-off into the graph, we highlighted genes that might have the potential for silencing due to epigenetic alterations. We highlighted five genes, EYA1, SIX2, ACSL6, OGDHL and SLC30A2, that showed a $\Delta\overline{\beta} \geq 0.25$ and a $log_2FC \geq 3.0$ between CIMP.L and CIMP.H.

4.2 Use cases using ELMER

4.2.1 Breast Invasive Carcinoma (unsupervised approach)

In this subsection, we describe how to perform *ELMER* analysis on TCGA BRCA (Breast Invasive Carcinoma) data retrieved from the GDC server. We first describe how the data can be downloaded and organized to the default *ELMER* input, followed by the following analysis steps:

- Identification of distal probes with significant differential DNA methylation (i.e. DMCs) in tumor vs. normal samples
- Identification of putative target gene(s) for differentially methylated distal probes
- Characterization of chromatin state context of significant probe regions using FunciVar
- Identification of enriched motifs within set of probes in significant probe-gene pairs
- Identification of master regulator Transcription Factors (TF) for each enriched motif

In addition to these standard steps, we also compared the putative probe-gene pairs to those derived from deep-sequenced ChIA-PET data from MCF7 cells (as shown in Yao *et al.* (2015)).

Downloading TCGA data

The function *getTCGA* uses the TCGAbiolinks package (COLAPRICO *et al.*, 2015) to download TCGA data for all samples for a given disease (such as BLCA, LGG, GBM). Its main arguments are the *genome* that if set to "hg19" will download data from GDC legacy archive, and if set to "hg38" it will download data from the main GDC harmonized data portal.

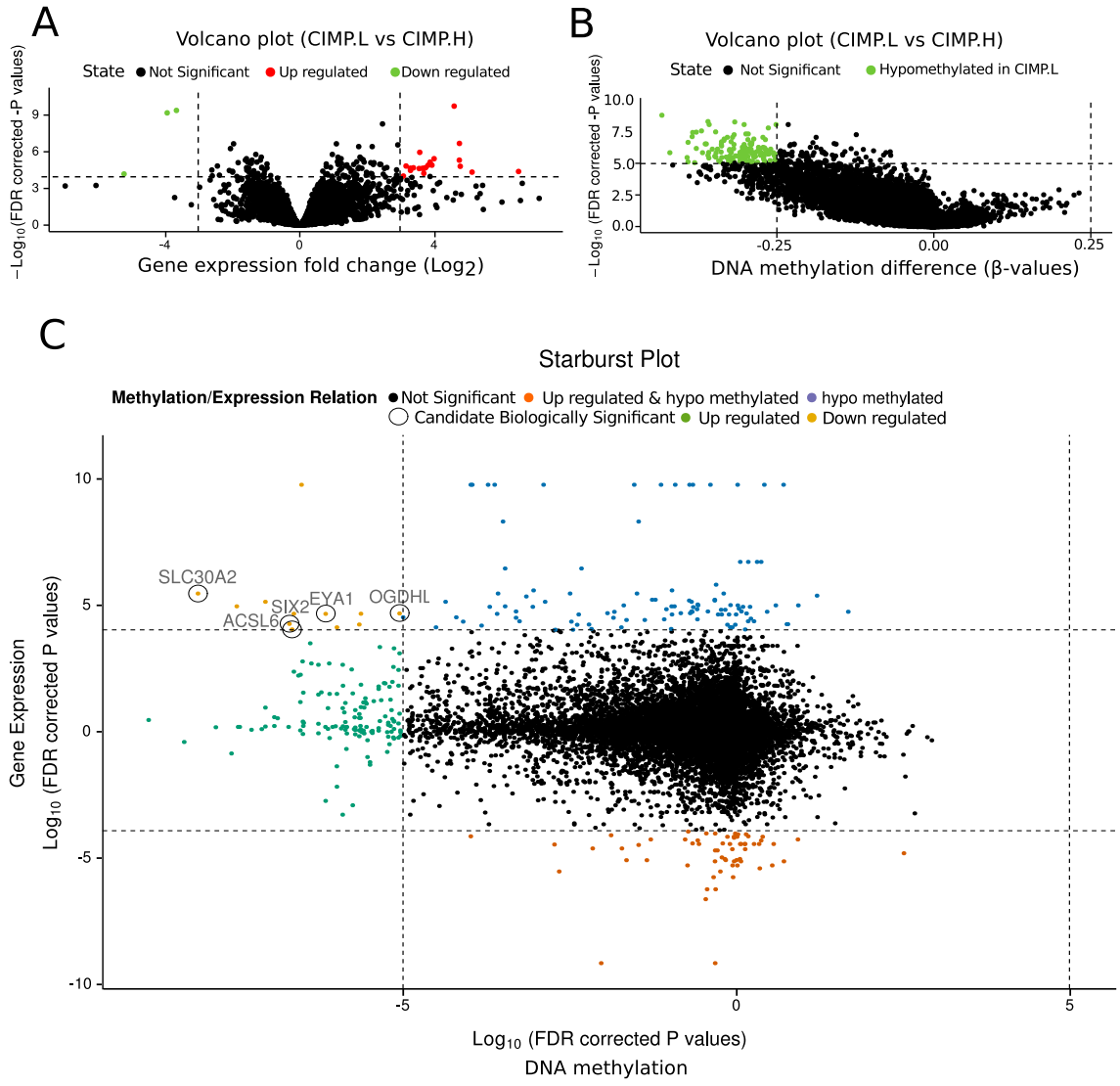


Figure 16 – Case study - Integrative analysis of gene expression and DNA methylation data from COAD disease, comparing groups CIMP.L and CIMP.H. **(A)** Expression volcano plot: fold change of expression data versus significance. **(B)** DNA methylation volcano plot: difference of DNA methylation versus significance. **(C)** Starburst plot: DNA methylation significance versus gene expression significance.

If the *getTCGA* function called before was successful it will create the following objects and folders:

```
--- DATA/BRCA/
  |---------- BRCA_meth_hg38.rda (object with DNA methylation)
  |---------- BRCA_RNA_hg38.rda (object with gene expression)
  |---------- BRCA_clinic.rda   (object with indexed clinical information)
  |---------- Raw/ (folder: contains All raw data from GDC)
```

Selecting distal probes

The function *get.feature.probe*, shown in Listing 1, is used to select HM450K/EPIC probes located away from any TSS (at least $2Kb$ away). Its main arguments are the genome of reference ("hg38"/"hg19") and DNA methylation platform ("450K"/"EPIC"). The *feature* argument is used to limit the region of probes; as we want all distal probes, we set it to NULL.

Source code 1 – "Selection of probes within biofeatures"

```
# get distal probes that are 2kb away from TSS
distal.probes <- get.feature.probe(feature = NULL,
                                   genome = "hg38",
                                   met.platform = "450K")
# 168644 probes
```

Organizing data into a MultiAssayExperiment object

The function *createMAE* is used to organize the gene expression and DNA methylation data into a MultiAssayExperiment (MAE) object. Listing 2 shows how to use it with the data created in the previous steps. Its main arguments are described below:

- *exp* : An R object or a path to a file containing a gene expression matrix or SummarizedExperiment with gene counts.
- *met* : An R object or a path to a file containing a DNA methylation matrix or SummarizedExperiment with beta values.
- *met.platform*: DNA methylation platform. "EPIC" for Infinium MethylationEPIC or "450K" for Infinium HumanMethylation450.
- *genome*: The genome of reference ("hg19" or "hg38") used to select the correct metadata. Genes genomic ranges will be annotated using ENSEMBL database and DNA methylation probes using metadata available at http://zwdzwd.github.io/InfiniumAnnotation.
- *linearize.exp*: this step will take the $log_2(gene\ expression + 1)$ in order to linearize the relationship between gene expression and DNA methylation.

- *filter.probes*: genomic ranges (i.e. distal regions) within which probes from DNA methylation data should be kept.

- *met.na.cut*: maximum percentage of empty values (NA) a probe might have to be considered in the analysis. The default is 20% (i.e if 50% of samples has empty values for a given probe, it will be removed).

- *colData*: A matrix with samples metadata (i.e. clinical data , molecular subtype information). If argument TCGA is set to *TRUE* this matrix will be created automatically. In this case, the *colData* argument is optional.

- *sampleMap*: A matrix mapping DNA methylation data and gene expression data to samples. *ELMER* uses only samples with both data. Otherwise, it will be removed. If argument TCGA is set to *TRUE* this matrix will be created automatically. In this case *sampleMap* argument is optional.

Source code 2 – "Create MultiAssayExperiment"

```
1 mae <- createMAE(exp = "DATA/BRCA/BRCA_RNA_hg38.rda",
2                   met = "DATA/BRCA/BRCA_meth_hg38.rda",
3                   met.platform = "450K",
4                   genome = "hg38",
5                   linearize.exp = TRUE,
6                   filter.probes = distal.probes,
7                   met.na.cut = 0.2,
8                   save = TRUE,
9                  TCGA = TRUE)
```

Listing 3 shows information about the object created. There are 866 samples with both gene expression and DNA methylation data, and among those 5 are metastatic samples, 778 are Primary Solid Tumor and 83 are Solid Tissue Normal.

Source code 3 – "Verifying MultiAssayExperiment"

```
1
2 > mae
3 A MultiAssayExperiment object of 2 listed
4  experiments with user-defined names and respective classes.
5  Containing an ExperimentList class object of length 2:
6  [1] DNA methylation: RangedSummarizedExperiment with 135331 rows and 866 columns
7  [2] Gene expression: RangedSummarizedExperiment with 57035 rows and 866 columns
8 Features:
9  experiments() - obtain the ExperimentList instance
10  colData() - the primary/phenotype DataFrame
11  sampleMap() - the sample availability DataFrame
12  `$`, `[`, `[[` - extract colData columns, subset, or experiment
13  *Format() - convert ExperimentList into a long or wide DataFrame
14  assays() - convert ExperimentList to a list of rectangular matrices
15 > table(mae$definition)
16          Metastatic Primary solid Tumor Solid Tissue Normal
17                   5                 778                  83
```

Identification of distal probes with significant differential DNA methylation (i.e. DMCs) in tumor vs. normal samples

The function *get.diff.meth* is used to identify regions differentially methylation between two groups. Listing 4 shows how to use it to select hypomethylated probes in "Primary solid tumor" samples when compared to "solid tissue normal" samples ($FDR \leq 0.01, \Delta\overline{\beta} \geq 0.3$), using those samples in the lower quintile ($minSubgroupFrac = 0.2$) of DNA methylation levels for each probe. Its main arguments are described below:

- *data* A multiAssayExperiment with DNA methylation and Gene Expression data.
- *group.col* A column defining the groups of the sample. You can view the available columns using: colnames(MultiAssayExperiment::colData(data)).
- *group1* A group from group.col. *ELMER* will run group1 vs group2. That means, if the direction is hyper, get probes hypermethylated in group 1 compared to group 2.
- *group2* A group from group.col. *ELMER* will run group1 vs group2. That means, if the direction is hyper, get probes hypermethylated in group 1 compared to group 2.
- *diff.dir* Differential methylation direction. It can be "hypo" which is only selecting hypomethylated probes in group 1 when compared to group 2; "hyper" which is only selecting hypermethylated probes;
- *minSubgroupFrac* A number ranging from 0 to 1, specifying the fraction of extreme samples from group 1 and group 2 that are used to identify the differential DNA methylation. The default is 0.2 because we typically want to be able to detect a specific (possibly unknown) molecular subtype among tumor; these subtypes often make up only a minority of samples, and 20% was chosen as a lower bound for the purposes of statistical power. If you are using pre-defined group labels, such as treated replicates vs. untreated replicated, use a value of 1.0 (*Supervised* mode)
- *pvalue* A number specifying the significant P value (adjusted P value by Benjamini-Hochberg procedure) cutoff for selecting significant hypo/hyper-methylated probes. The default is 0.01.
- *sig.dif* A number specifying the smallest DNA methylation difference as a cutoff for selecting significant hypo/hyper-methylated probes. The default is 0.3.

Source code 4 – "Identify significantly different DNA methylation probes in tumor and normal samples"

```
1 diff.probes <- get.diff.meth(data = mae,
2                               group.col = "definition",
3                               group1 = "Primary solid Tumor",
4                               group2 = "Solid Tissue Normal",
```

```
5                                   diff.dir = "hypo", # Get probes hypometh. in group 1
6                                   cores = 1,
7                                   minSubgroupFrac = 0.2, # % group samples  used.
8                                   pvalue = 0.01,
9                                   sig.dif = 0.3,
10                                  dir.out = "Results_hypo/",
11                                  save = TRUE)
```

If the *save* argument is set to TRUE, in the *dir.out* folder two files will be created: *getMethdiff.hypo.probes.csv* containing all probes from the DNA methylation data with the difference means of the groups and the significance values, *getMethdiff.hypo.probes.significant.csv* will contain only probes that respect the thresholds. Table 12 shows the first rows of *getMethdiff.hypo.probes.significant.csv* file.

probe	pvalue	Primary.solid.Tumor_Minus_Solid.Tissue.Normal	adjust.p
cg00001809	1.97e-35	-0.32	1.26e-34
cg00008695	1.62e-67	-0.44	3.72e-66
cg00009553	6.84e-31	-0.525	3.61e-30

Table 12 – Identification of distal probes with significant differential DNA methylation (i.e. DMCs): First three rows of getMethdiff.hypo.probes.significant.csv file.

Identification of putative target gene(s) for differentially methylated distal probes

The function *get.pair* is used to link enhancer probes with methylation changes to target genes with expression changes and report the putative target gene for selected probes. Listing 5 shows how to select the 20 nearest genes (10 downstream and 10 upstream) and evaluate if each pair is anti-correlated (probes with higher methylation levels have lower gene expression levels). Its main arguments are described below:

- *nearGenes:* Output of *GetNearGenes* function.
- *mode:* Algorithm mode: "unsupervised" or "supervised". If unsupervised is set the *U* (unmethylated) and *M* (methylated) groups will be selected among all samples of both groups based on methylation of each probe. Otherwise *U* group and *M* group will set as all the samples of group1 or group2 as described below: If diff.dir is "hypo, *U* will be the group 1 and *M* the group2. If diff.dir is "hyper" *M* group will be the group1 and *U* the group2.
- *minSubgroupFrac:* A number ranging from 0 to 1, specifying the fraction of extreme samples that define group U (unmethylated) and group M (methylated), which are used to link probes to genes. The default is 0.4 (the lowest quintile of samples is the U group and the highest quintile samples is the M group) because we typically want to be able to detect a specific (possibly unknown) molecular subtype among tumor; these subtypes often make up only a minority of samples, and 20% was chosen as a lower bound for the purposes of statistical power. This argument is Only used if mode is "supervised", otherwise if you are using pre-defined group labels ("supervised" mode), such as treated replicates vs. untreated replicated, it will use all samples.
- *permu.size:* Number of permutation. The default is 10000. *Note*: This parameter can strongly impact run time.
- *raw.pvalue:* Raw p-value cutoff for defining significant pairs. The default is 0.001.
- *Pe:* Empirical p-value cutoff for defining significant pairs. The default is 0.001.
- *filter.probes:* Should probes be filtered by selecting only those which have at least a certain number of samples below and above a certain cut-off ? If true, arguments *filter.probes* and *filter.percentage* will be used.
- *filter.portion:* A number specifying the cut point to define binary methylation level for probe loci. The default is 0.3. When the beta value is above 0.3, the probe is methylated and vice versa. For one probe, the percentage of methylated and unmethylated samples should be above filter.percentage value. Only used if *filter.probes* is TRUE.

- *filter.percentage:* Minimum percentage of samples to be considered in methylated and unmethylated for the filter.portion option. Default 5%. Only used if *filter.probes* is TRUE.

Source code 5 – "Identify putative target genes for differentially methylated distal probes"

```
1
2 # For each differently methylated probes we will get the
3 # 20 nearby genes (10 downstream and 10 upstream)
4 nearGenes <- GetNearGenes(data = mae,
5                           probes =  diff.probes$probe,
6                           numFlankingGenes = 20,
7                           cores = 1)
8
9 Hypo.pair <- get.pair(data = mae,
10                       nearGenes = nearGenes,
11                       group.col = "definition",
12                       group1 = "Primary solid Tumor",
13                       group2 = "Solid Tissue Normal",
14                       permu.dir = "Results_hypo/permu",
15                       permu.size = 10000,
16                       mode = "unsupervised",
17                       minSubgroupFrac = 0.4,
18                       raw.pvalue = 0.001,
19                       Pe = 0.001,
20                       filter.probes = TRUE,
21                       filter.percentage = 0.05,
22                       filter.portion = 0.3,
23                       dir.out = "Results_hypo",
24                       cores = 1,
25                       label = "hypo")
```

The output of this function is shown in table 13. Probe and GeneID columns show the significant pair and the column P_e shows the adjusted p-value.

Probe	GeneID	Symbol	Distance	Sides	Raw.p	Pe
cg14058239	ENSG00000141424	SLC39A6	0	R1	5.17e-56	9.99e-5
cg14986386	ENSG00000183323	CCDC125	0	L2	3.26e-55	9.99e-5
cg04723436	ENSG00000107485	GATA3	110455	R3	6.00e-55	4.9.99e-4

Table 13 – Identification of putative target gene(s) for differentially methylated distal probes: First three rows of getPair.hypo.pairs.significant.csv file.

To visualize the relationship between the probe-gene pairs inferred, there are two auxiliary functions in ELMER. The function *schematic.plot*, shown in Listing 6, which will plot genes and probes in a specified genomic region, highlighting the significant pairs identified by plotting a genomic interactions track and highlight the genes in the pair in red (Figure 17). Also, using the function *scatter.plot* (Listing 7) it is possible to visualize the correlation between gene expression and DNA methylation levels at probe (Figure 18). Finally, an overall summary of the DNA methylation levels and gene expression levels can be visualized using the auxiliary function *heatmapPairs*, as shown in Listing 8. This function creates a heatmap for all samples as shown in Figure 19.

Source code 6 – "Schematic plot to visualize gene-probe pairs"

```
# by probe and with detail about DNA methylation
schematic.plot(data = mae,
               group.col = "definition",
               group1 = "Primary solid Tumor",
               group2 = "Solid Tissue Normal",
               pair = Hypo.pair,
               statehub.tracks = "hg38/ENCODE/mcf-7.16mark.segmentation.bed",
               byProbe = "cg04723436")
```

Source code 7 – "Scatterplot to visualize correlation between gene expression and DNA methylation levels at probe"

```
scatter.plot(data = mae,
             byPair = list(probe = "cg04723436",
                           gene = "ENSG00000107485"),
             save = T,
             category = "definition",
             lm = TRUE)
```

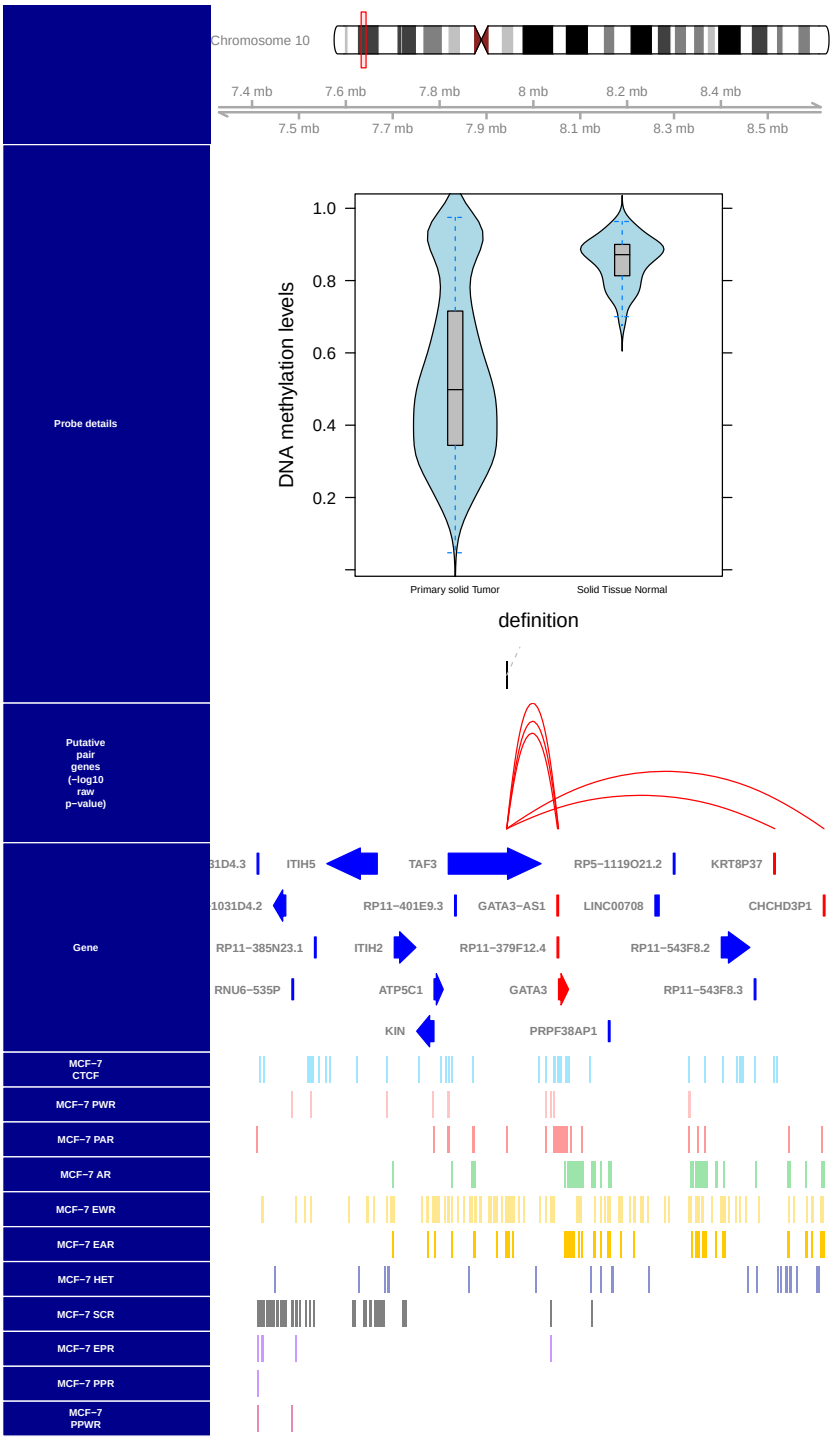


Figure 17 – Plot probe-gene pairs with annotation track for MCF-7 cell line from StateHub.org. Significant probes and gene pairs are highlighted in red.

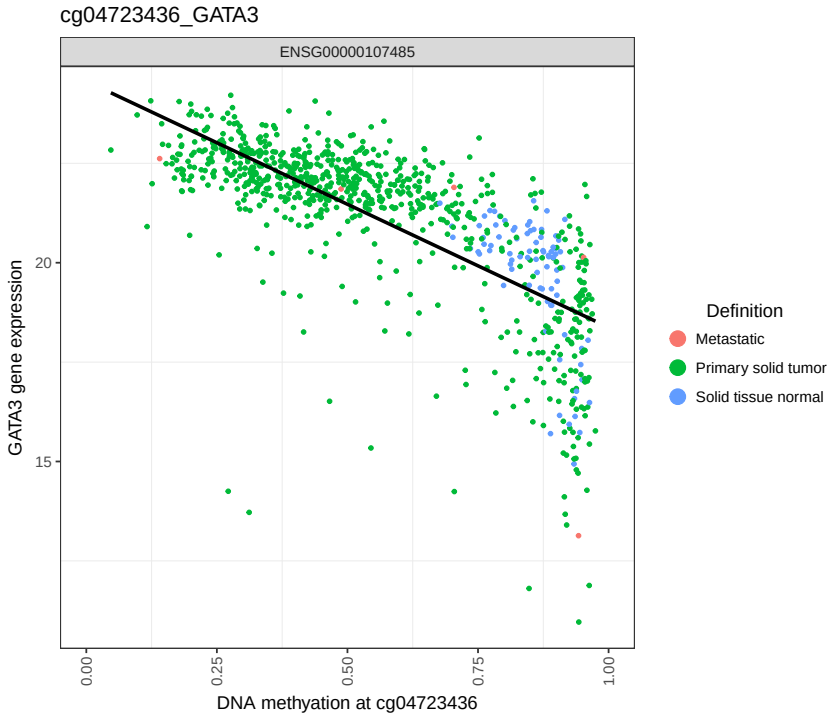


Figure 18 – Scatter plot for significant probe (cg04723436) gene (GATA3) pair.

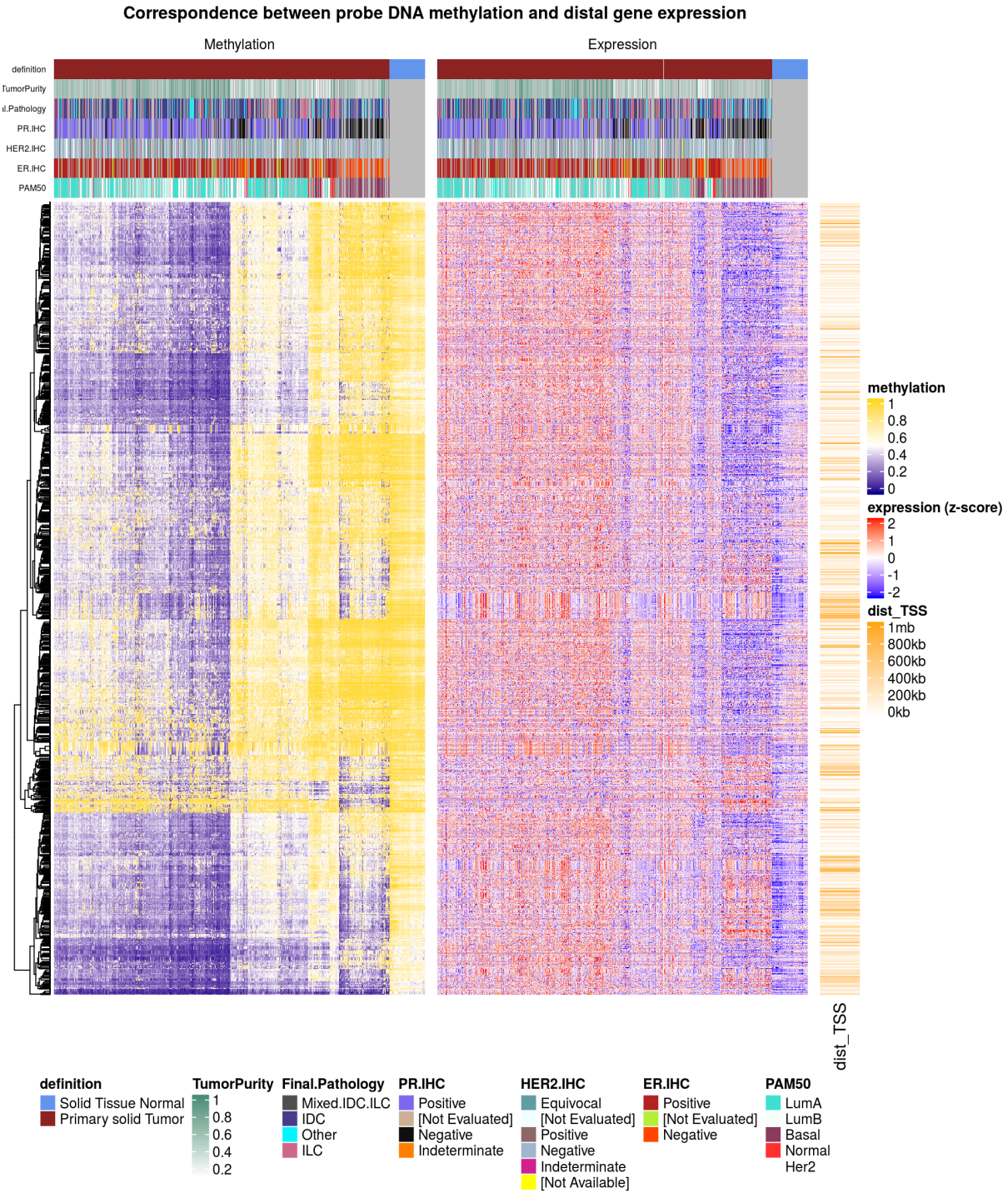


Figure 19 – Heatmap of paired probes and distal genes. The first heatmap (left one) shows DNA methylation β levels ranging from 0 (non-methylated) up to 1 (methylated probes). The second heatmap (middle one) shows z-scores for gene expression levels (standard deviations from each gene means). The last heatmap (right heatmap) shows the distance between the probe and gene anti-correlated.

Source code 8 – "Heatmap to visualize gene-probe pairs"

```
# Get molecular subytpes/purity from cell paper
# https://doi.org/10.1016/j.cell.2015.09.033
file <- "http://ars.els-cdn.com/content/image/1-s2.0-S0092867415011952-mmc2.xlsx"
downloader::download(file, basename(file))
subtypes <- readxl::read_excel(basename(file), skip = 2)

subtypes$sample <- substr(subtypes$Methylation,1,16)
meta.data <- merge(colData(mae),subtypes,by = "sample",all.x = T)
meta.data <- meta.data[match(colData(mae)$sample,meta.data$sample),]
meta.data <- S4Vectors::DataFrame(meta.data)
rownames(meta.data) <- meta.data$sample
stopifnot(all(meta.data$patient == colData(mae)$patient))
colData(mae) <- meta.data

heatmapPairs(data = mae,
             group.col = "definition",
             group1 = "Primary solid Tumor",
             group2 = "Solid Tissue Normal",
             annotation.col = c("TumorPurity",
                                "Final.Pathology",
                                "PR.IHC",
                                "HER2.IHC",
                                "ER.IHC",
                                "PAM50"),
             pairs = Hypo.pair,
             filename = "heatmap.pdf")
```

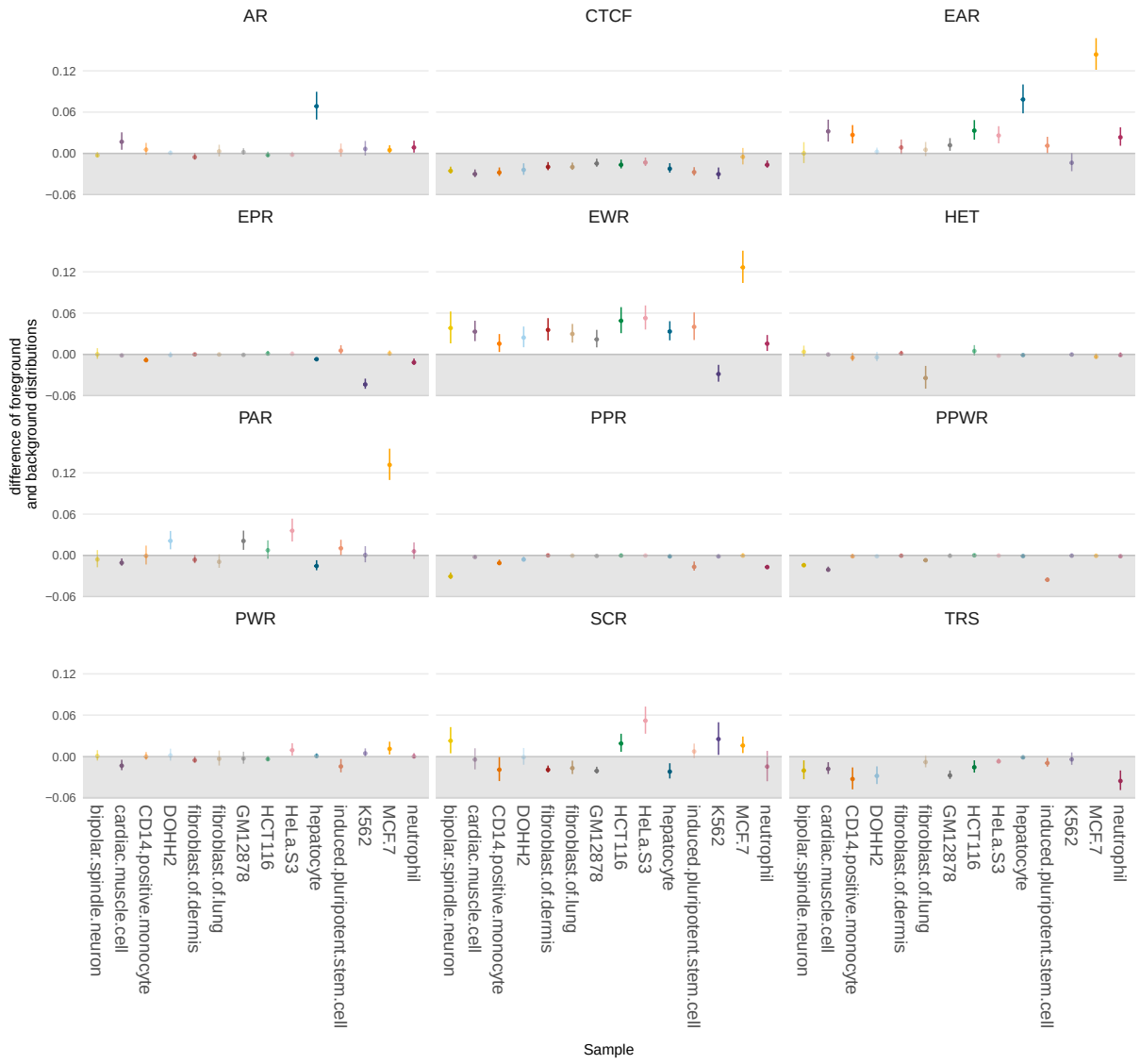


Figure 20 – Enrichment of paired probes and chromatin states of encode cells. The plot shows enrichment for enhancer active region, weak enhancer and active promoter region for MCF-7 cell. Acronyms - AR: Active region, EAR: active enhancer, EWR: Weak Enhancer, EPR: poised enhancer, PAR: active promoter, PWR: Weak Promoter, PPR: poised promoter, PPWR: Weak Poised Promoter, CTCF: architectural complex, TRS: transcribed, HET: heterochromatin, SCR: Polycomb Repressed Silenced

Characterization of chromatin state context of significant probe regions using FunciVar

To understand and compare our set of probes identified in the probe-gene pairs inferred we used chromatin state of IHEC cell types from http://statehub.org/, to calculate the relative enrichment of different states. This procedure uses code from the statepaintR (COETZEE *et al.*, 2017) and FunciVar (FUNCIVAR, 2017) packages. Figure 20 shows the enrichment for 14 encode cells lines. The plot shows enrichment for enhancer active region (EAR), weak enhancer (EWR) and active promoter region (PAR) in MCF-7 cell (human breast adenocarcinoma cell line) while for other cell lines this enrichment is not visible.

Identification of enriched motifs within set of probes in significant probe-gene pairs

The function *get.enriched.motif* is used to identify enriched motif in a set of probes. The main arguments are described below:

- *lower.OR* The motif with lower boundary of 95% confidence interval for Odds Ratio $\geq$ *lower.OR* are the significantly enriched motifs.
- *min.incidence* Minimum number of probes having the motif signature (default: 10) required for a motif to be enriched.

Source code 9 – "Motif enrichment analysis on the selected probes"

```
enriched.motif <- get.enriched.motif(data = mae,
                                     min.motif.quality = "DS",
                                     probes = unique(Hypo.pair$Probe),
                                     dir.out = "Results_hypo",
                                     label = "hypo",
                                     min.incidence = 10,
                                     lower.OR = 1.1)
```

Identification of master regulator Transcription Factors (TF) for each enriched motif

The function *get.TFs* is used to identify regulatory TF whose expression associates with TF binding motif DNA methylation which.

Source code 10 – "Identifying regulatory Transcript Factors"

```
TF <- get.TFs(data = mae,
              group.col = "definition",
              group1 = "Primary solid Tumor",
              group2 = "Solid Tissue Normal",
              minSubgroupFrac = 0.4, # Set to 1 if supervised mode
              enriched.motif = enriched.motif,
              dir.out = "Results_hypo",
              cores = 1,
              label = "hypo")
```

The result of this function is shown in table 14 and in figure 22.

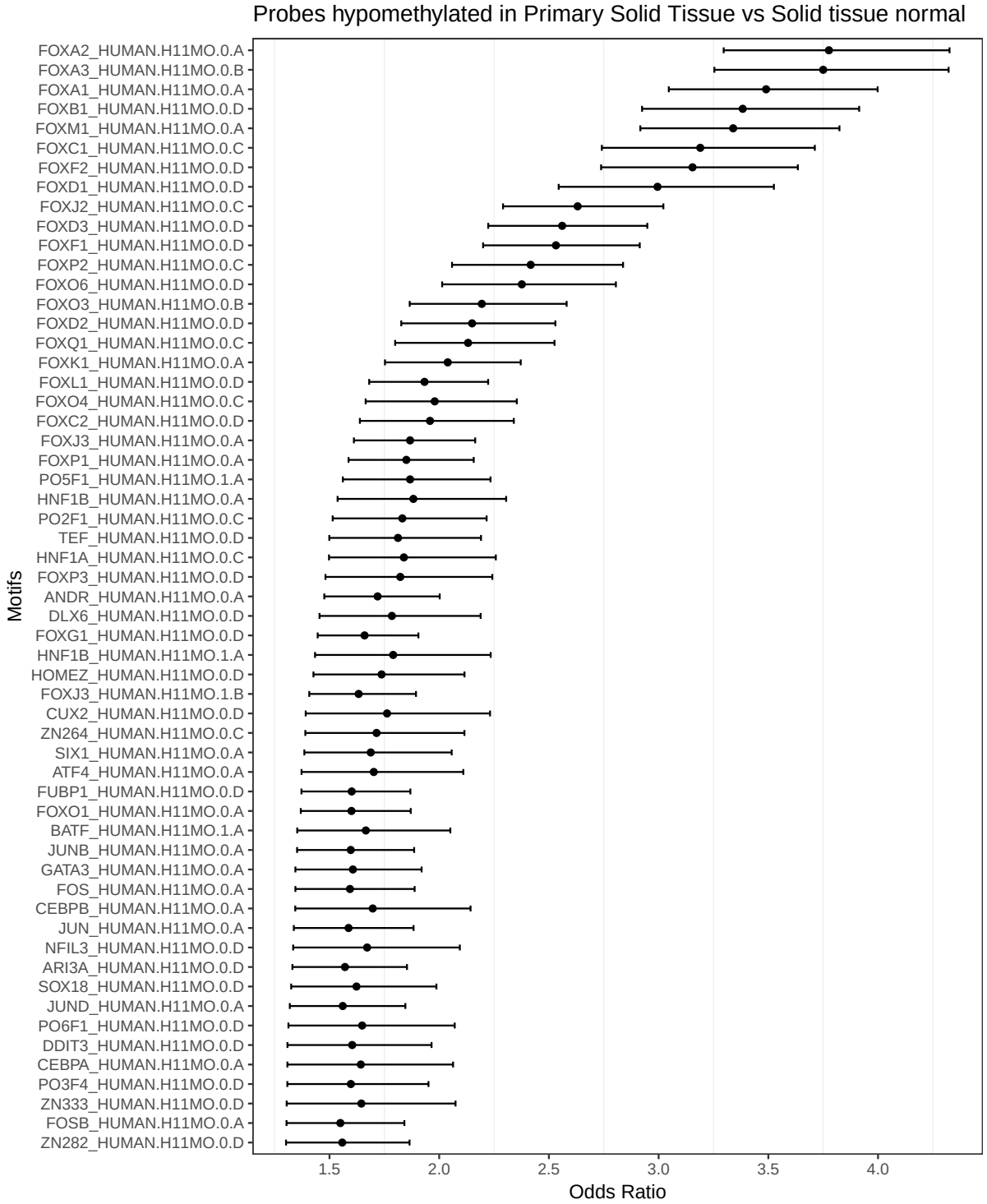


Figure 21 – Motif enrichment plot shows the enrichment levels ($OR \geq 2.0$) for the selected motifs. This plot represents a subset of the enriched motifs for $lower.or = 1.1$, only selected for representational purposes.

motif	top.potential.TF.family	top.potential.TF.subfamily	potential.TF.family	potential.TF.subfamily	top_5percent_TFs
AIRE.C	NA	NA	NA	NA	FOXA1;GATA3;ESR1;SPDEF;RARA;...
ANDR.A	ESR1	AR	ESR1;AR	AR	FOXA1;GATA3;ESR1;RARA;SPDEF;...
ANDR2.A	ESR1	AR	ESR1;AR	AR	FOXA1;GATA3;ESR1;RARA;SPDEF;...
BSH.D	EMX1	NA	EMX1;LBX2	NA	FOXA1;GATA3;ESR1;SPDEF;RARA;...
CDC5L.D	MYB	MYB	MYB	MYB	FOXA1;GATA3;ESR1;SPDEF;RARA;...
DLX2.D	EMX1	NA	EMX1;LBX2	NA	FOXA1;GATA3;ESR1;SPDEF;CXXC5;...
EMX1.D	EMX1	EMX1	EMX1;LBX2	EMX1	FOXA1;GATA3;ESR1;SPDEF;CXXC5;...
EMX2.D	EMX1	EMX1	EMX1;LBX2	EMX1	FOXA1;GATA3;ESR1;SPDEF;CXXC5;...
ERR3.B	ESR1	ESR1	ESR1;AR	ESR1	FOXA1;GATA3;ESR1;SPDEF;CXXC5;...
EVI1.B	ZNF467	NA	ZNF467;PATZ1	NA	FOXA1;GATA3;ESR1;SPDEF;RARA;...
EVX1.D	NA	NA	NA	NA	FOXA1;GATA3;ESR1;CXXC5;SPDEF;...
FOSB.A	NA	NA	NA	NA	FOXA1;GATA3;ESR1;RARA;CXXC5;...
FOSL1.A	NA	NA	NA	NA	FOXA1;GATA3;ESR1;CXXC5;RARA;...
FOSL2.A	NA	NA	NA	NA	FOXA1;GATA3;ESR1;CXXC5;RARA;...
FOS.A	NA	NA	NA	NA	FOXA1;GATA3;ESR1;CXXC5;RARA;...
FOXA1.A	FOXA1	FOXA1	FOXA1	FOXA1	FOXA1;GATA3;ESR1;SPDEF;RARA;...
FOXA2.A	FOXA1	FOXA1	FOXA1	FOXA1	FOXA1;GATA3;ESR1;SPDEF;RARA;...
FOXA3.B	FOXA1	FOXA1	FOXA1;FOXD2	FOXA1	FOXA1;GATA3;ESR1;SPDEF;RARA;...

Table 14 – First twenty rows of the *getTF.hypo.significant.TFs.with.motif.summary.csv* file created by *get.Tfs* function (suffix "_HUMAN.H11MO" was removed from motifs names). First column shows the enriched motif, "top_5percent_TFs" shows the top 5% TFs ranked (the same as all TFs to the left of the dashed line in figure 22), "potential.TFs.family" are the TF from the "top_5percent" that belongs to the same family as the TF of the motif, "top.potential.TFs.family" is the highest ranked TF belonging to the same family as the TF of the motif (same as the first TF from "potential.TFs.family" column). The columns "potential.TFs.subfamily" and "top.potential.TFs.subfamily" are the same as "potential.TFs.family" and "top.potential.TFs.family" but considering the subfamily classification instead. For example, the motif ANDR has two TFs in the top 5% that belongs to the same TF family (Steroid hormone receptors): ESR1 and AR, but if considering subfamilies only AR in considered.

Comparing inferred results with MCF-7 chIA-PET

As shown in Yao *et al.* (2015), we compared the putative pairs inferred to the chromatin loops derived from deep-sequenced ChIA-PET data from MCF7 cells (LI *et al.*, 2012). First, we identify the number of *ELMER* pairs overlapping the ChIA-PET loops, then we repeat using randomly generated pairs with properties similar to the *ELMER* pairs. For each true ELMER probe in a probe-gene pair, we randomly select a different probe from the complete set of distal probes. We then choose the nth nearest gene to the random probe, where n is the same as the adjacency of the true ELMER probe (i.e. if the true probe is linked to the second gene upstream, the random probe will also be linked to its second gene upstream). Thus, the random linkage set has both the same number of probes and the same number of linked genes as the true set. One hundred such random datasets were generated to arrive at a 95% CI ($\pm 1.96 * SD$). The result is shown in Figure 23. Of the 2124 putative pairs identified in breast cancer tumors, 316 (approximately 14.9%) were also identified as loops in the MCF7 ChIA-PET data. This was a three-fold enrichment over randomized probe-gene pairs (see Additional file for the code).

4.2.2 BRCA molecular subtypes analysis (supervised approach)

Several studies identified distinct molecular Breast cancer classes and divided them as the following subclasses: luminal-like (Luminal A and Luminal B), which are Estrogen receptor-positive (ER-positive), and the basal-like, ErbB2-positive and normal-like subclasses, which are the ER-negative groups (PEROU *et al.*, 2000; YERSAL; BARUTCA, 2014; SØRLIE *et al.*, 2001). To perform *ELMER* analysis comparing known molecular subtypes (Her2, Luminal A, Luminal B and Basal-like) a TCGA BRCA dataset classification was retrieved from Ciriello *et al.* (2015).

The main arguments changed were the percentage of samples used to identify the differentially methylated probes in function *get.diff.meth*, which was set to 100% (use all samples from each group), and the mode in function *get.pair* and in function *get.TFs* which was set to "supervised". In this mode instead of defining the *U* (unmethylated) group as the samples with lowest quintile of DNA methylation levels and the *M* (methylated) group as the highest quintile, the *U* and *M* group were defined as all samples of one known molecular subtype. For example, if the first step identified probes hypomethylated in Luminal A group when compared to Basal-like group, the next steps will use the Luminal A samples as *U* group and the Basal-like samples as the *M* group.

The unsupervised analysis identified several Luminal type Master Regulators (MRs) such as FOXA1, GATA3, and ESR1. In order to identify MRs for the other subtypes, we created a table (Table 15) of candidate MRs identified by each pairwise ELMER run.

Interestingly, several new MRs are identified for the Basal-like group, and these were mostly consistent in comparisons against Luminal and HER2+ subtypes. One group of MRs

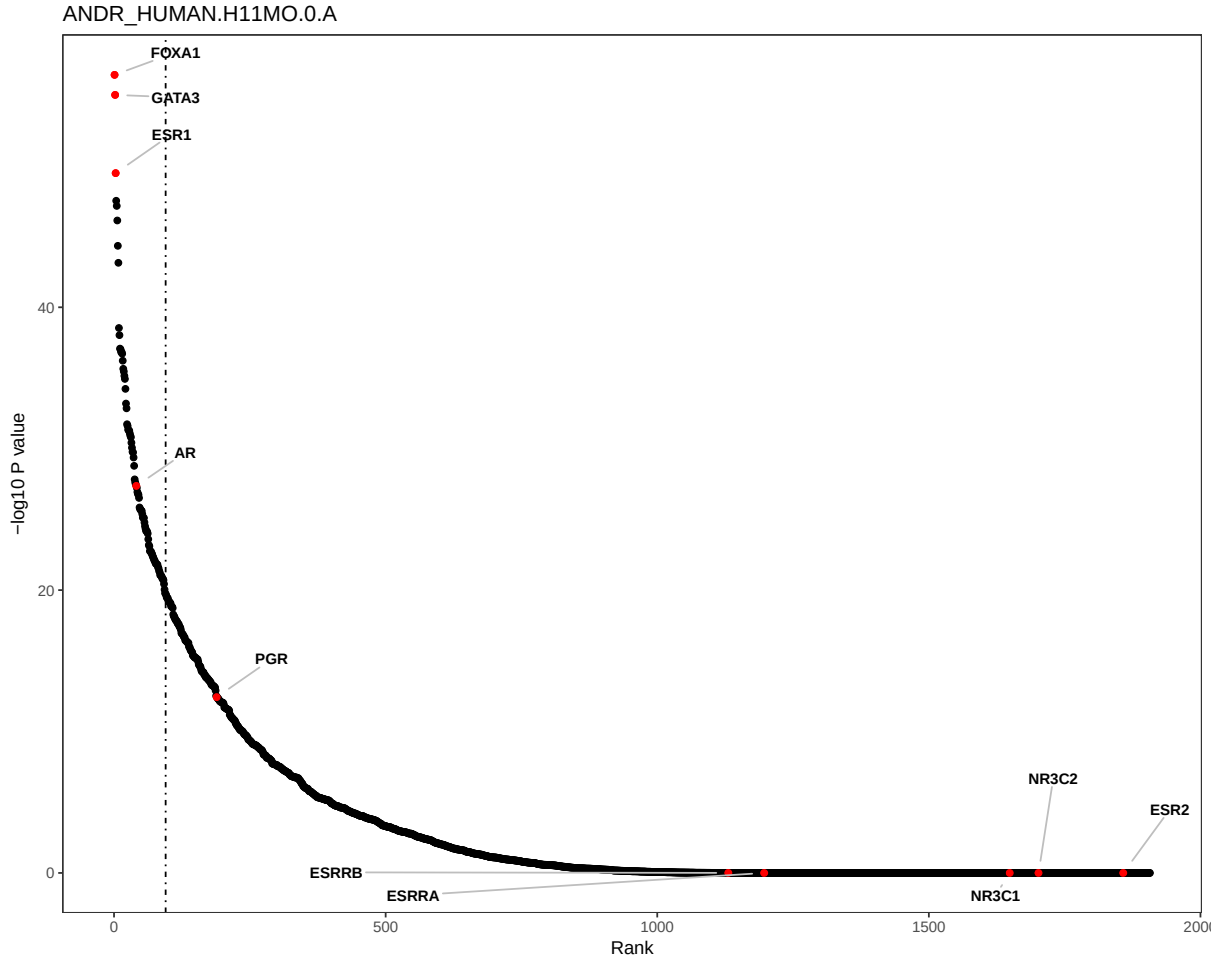


Figure 22 – TF ranking plot shows statistic $-log_{10}(P-value)$ assessing the anti-correlation level of TFs expression level with average DNA methylation level at sites with a given motif. By default, the top 3 associated TFs and the TF family members (dots in red) that are associated with that specific motif are labeled in the plot. But there is also an option to highlight only TF sub-family members (TCClass database classification)

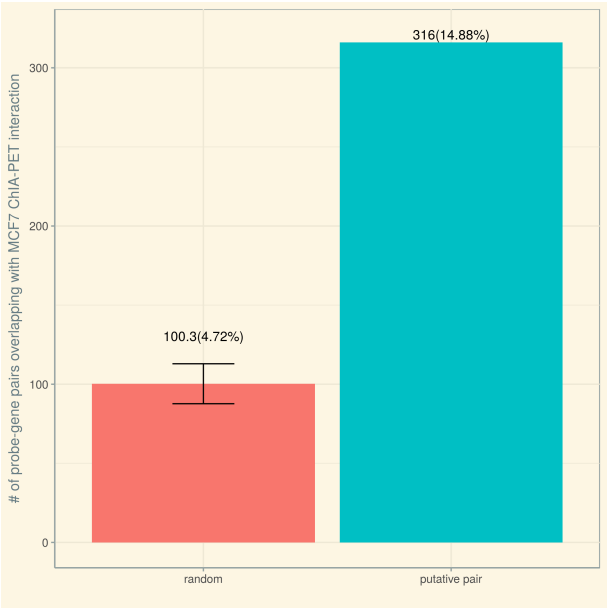


Figure 23 – The graph shows the comparison of the number of probe-gene pairs identified within MCF7 ChIA-PET data using the putative pairs from BRCA vs. random pairs

identified are the *SOX10* and *SOX9* TF signatures. For these signatures, the regulatory TF candidate identified are the *SOX9* (Sry-related HMG box-9) TF and *SOX11* (Sry-related HMG box-11) TF; this correlation between basal-like and SOX11 was recently described by (SHEPHERD *et al.*, 2016) and *SOX9* was described by (GONG *et al.*, 2015). Most interestingly, we found KLF5 to be a consistently predicted MR for the Basal-like breast subtype. KLF5 is a master pluripotency factor of embryonic stem cells, and has been associated with a number of different cancers. In breast cancer, it's overexpression has been linked to aggressive, ER-negative and basal-like breast cancers (BEN-PORATH *et al.*, 2008).

4.3 Glioma analysis

The Enhancer Linking by Methylation/Expression Relationship (ELMER) tool was used to analyze the molecular differences between the newly identified G-CIMP-low subtype of glioma that has some regions of the genome with a lower DNA methylation level and was associated with significantly worse survival compared to the G-CIMP-high, recently described by Dr. Noushmehr and his lab (CECCARELLI *et al.*, 2016b).

For this analysis, TCGA data from the NCI's Genomic Data Commons (GDC) was downloaded using our R/Bioconductor TCGAbiolinks package described in the previous chapter. Table 16 summarizes the number of samples in each group that have both deoxyribonucleic acid (DNA) methylation data for the Illumina HumanMethylation450 platform (HM450) and gene expression data (RNA-Seq) and Table 17 summarizes the main values for the ELMER arguments. This analysis was performed with data aligned against the genome of reference hg38 and using ELMER "supervised" mode, which ensures that in all the steps the comparisons are made between each group using all their samples.

The DMCs (differentially methylated CpGs) analysis, which uses a one-tailed t-test to find probes with higher mean level in G-CIMP-high group compared to the G-CIMP-low one. The result of this analysis is summarized in Figure 24. The volcano plot shows the difference of DNA methylation ($\Delta\overline{\beta}$) versus significance $-log_{10}$(FDR corrected P-values). Using as default cuttoffs $\Delta\overline{\beta} \geq 0.3$ and $-log_{10}$(FDR corrected P-values) ≤ 0.01, 7540 probes were identified to be differentially methylated.

Those differentially methylated were then paired with their 10 upstream and 10 downstream genes and an anti-correlation test which searches for genes highly expressed when the DNA methylation levels decreases was performed. Using as cutt-offs $-log_{10}$(raw P-values) $\leq$ 0.001 and $-log_{10}$(Permuted corrected P-values) ≤ 0.001 for a 10000 permutation correction approach, 886 pairs of gene and probes were identified to be anti-correlated. The result of this analysis is summarized in Figure 25. The heatmap is composed of other three heatmaps, one

Table 15 – Candidate regulatory TFs for each molecular subtype found in a pairwise comparison.

TF	**LUMA (vs basal)**	**LUMB (vs basal)**	**LUMA (vs normal)**	**LumB (vs normal)**	**Basal (vs LumA)**	**Basal (vs LumB)**	**Basal (vs HER2)**	**HER2 (vs Basal)**
AR	x	x	x					
BATF3						x		
BCL11A					x	x	x	
CBFB					x			
CEBPB					x	x	x	
CEBPG					x			
E2F3						x	x	
ELF5					x			
EMX1	x	x	x					
ESR1	x	x	x	x				
ELF5						x		
ETS2						x		
ETV6					x	x	x	
FOSL1						x		
FOXA1	x	x	x	x				x
FOXM1				x				
FOXD2		x						
FOXP1	x	x						x
GATA2	x							x
GATA3	x	x	x	x				x
GLI1	x	x						
HOMEZ		x						
HOXB1	x	x						
HOXB2	x	x						x
HOXB3								x
HOXB6								x
HOXC10								x
HOXC11								x
KLF5						x	x	
LMX1B	x	x	x					
MAZ			x					
MNX1								x
MSX2	x	x						
MYB	x			x				
MYBL1				x				
MYBL2				x	x			
NFATC4	x							
NFIB					x			
NFIL3					x	x		
NR2E3	x	x						
OVOL2			x					
PATZ1		x	x					
PBX1		x		x				
POU2F1			x	x				
PGR	x	x						
RARA	x	x	x					
RELB						x		
RORC	x							
RUNX3					x	x		
SOX8						x	x	
SOX9						x	x	
SOX11					x	x		
SPIB						x		
VEZF1		x						
ZBTB4	x							
ZNF281			x	x				
ZNF423	x							
ZNF467	x	x	x					
ZIC1					x	x	x	

Table 16 – G-CIMP-high vs G-CIMP-low analysis: number of samples with both DNA methylation (HM450) and gene expression (RNA-seq) data.

Group	Number of samples
G-CIMP-high	233
G-CIMP-low	11

Table 17 – G-CIMP-high vs G-CIMP-low analysis: ELMER arguments values.

Step	Argument	Value
createMAE	genome	hg38
Pair probe-gene correlation/master TF	Mode	Supervised
All	minSubgroupFrac	100%
All	group1	GGimp-high
All	group2	GGimp-low
All	direction	hypermethylated probes
DMCs	$\min \Delta\overline{\beta}$	0.3
DMCs	p-value adj cut-off	0.01
Pair probe-gene correlation	# permutations	10000
Pair probe-gene correlation	raw p-value cut-off	0.001
Pair probe-gene correlation	empirical p-values cut-off	0.001
Motif enrichment	minimum # probes	10
Motif enrichment	lower OR	1.1

heatmap for the DNA methylation levels, one heatmap for the gene expression levels and one with the distance between the gene and the probe.

Table 18 – G-CIMP-high vs G-CIMP-low analysis: number of distal probes, significant differentially methylated probes, significant anti-correlated paired gene-probes and enriched motifs

Result	Value
Distal probes	135374
Significant differentially methylated probes	7540
Significant anti-correlated paired gene-probes	886
Enriched motifs	84

The enriched motifs analysis results are summarized in Figure 26, which shows the Odds Ratio (OR) (x axis) for the enriched motifs, and in Table 19, which shows the candidate regulatory TFs whose expression anti-correlated with the DNA methylation level on the probes of each enriched motifs. From the most anti-correlated ones, ELMER uses the TFClass classification (WINGENDER *et al.*, 2015) to identifies which TFs are known to bind in those motifs. This classification has two levels, family and subfamily, which groups motifs with a similar signature.

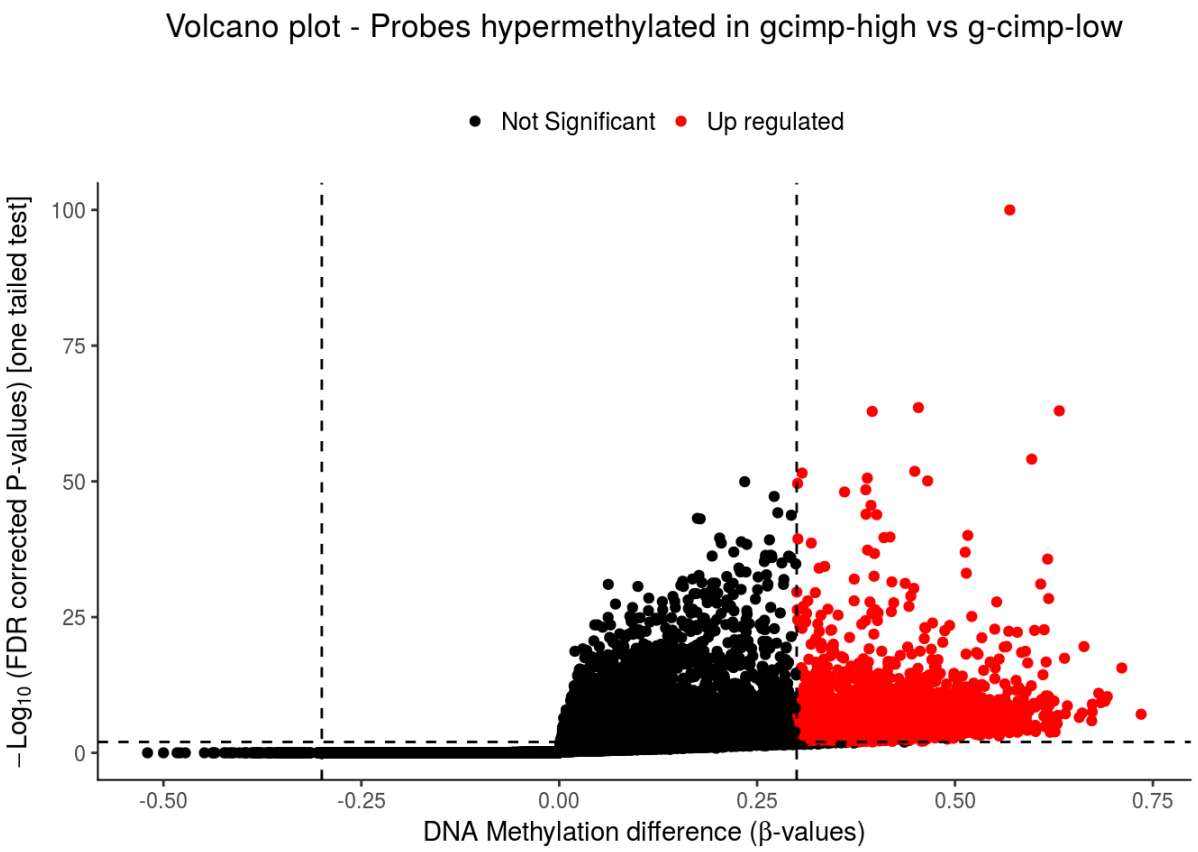


Figure 24 – G-CIMP analysis: DNA methylation Volcano plot. Difference of DNA methylation ($\min \Delta\overline{\beta}$) versus significance $-log_{10}$(FDR corrected P-values).

Depending on the motif, its family classification might have a very similar signature, otherwise, the subfamily classification is the most indicated.

For example, in Table 19 the motif HXD3 has as potential TF candidate the HOXD13 if we consider the family Transcription Factor (TF) classification and the HOXD3 TF candidate considering the subfamily TF classification. Figure 27 shows the motif signature for the HOX-related factors family from HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) database. The transcription factors Homeobox D13 (HOXD13) and Homeobox D3 (HOXD3) are in the same family (HOX-related factors) but in different subfamilies.

Overall, the results suggests that those regions that lost DNA methylation are bound by the regulatory TF Hmx3, which has been related to activation and maintenance of Gsh1 expression and subsequent downstream generation of growth hormone-releasing hormone (GHRH) expressing neurons (MORALES-DELGADO *et al.*, 2014) and Pozsgai *et al.* (2010) conducted test with GHRH antagonists in glioblastoma cells showing efficacy of those drugs for experimental treatment of glioblastoma, HOXD13, which have been reported to be substantially up regulated in Glioblastomas (GBMs) and may contribute to malignancy (LEE *et al.*, 2015), PAX3, which is essential for gliomagenesis (XIA *et al.*, 2013) also FOXM1 was shown by Wang *et al.* (2015) to promote the development and progression of GBM and to be a novel therapeutic target

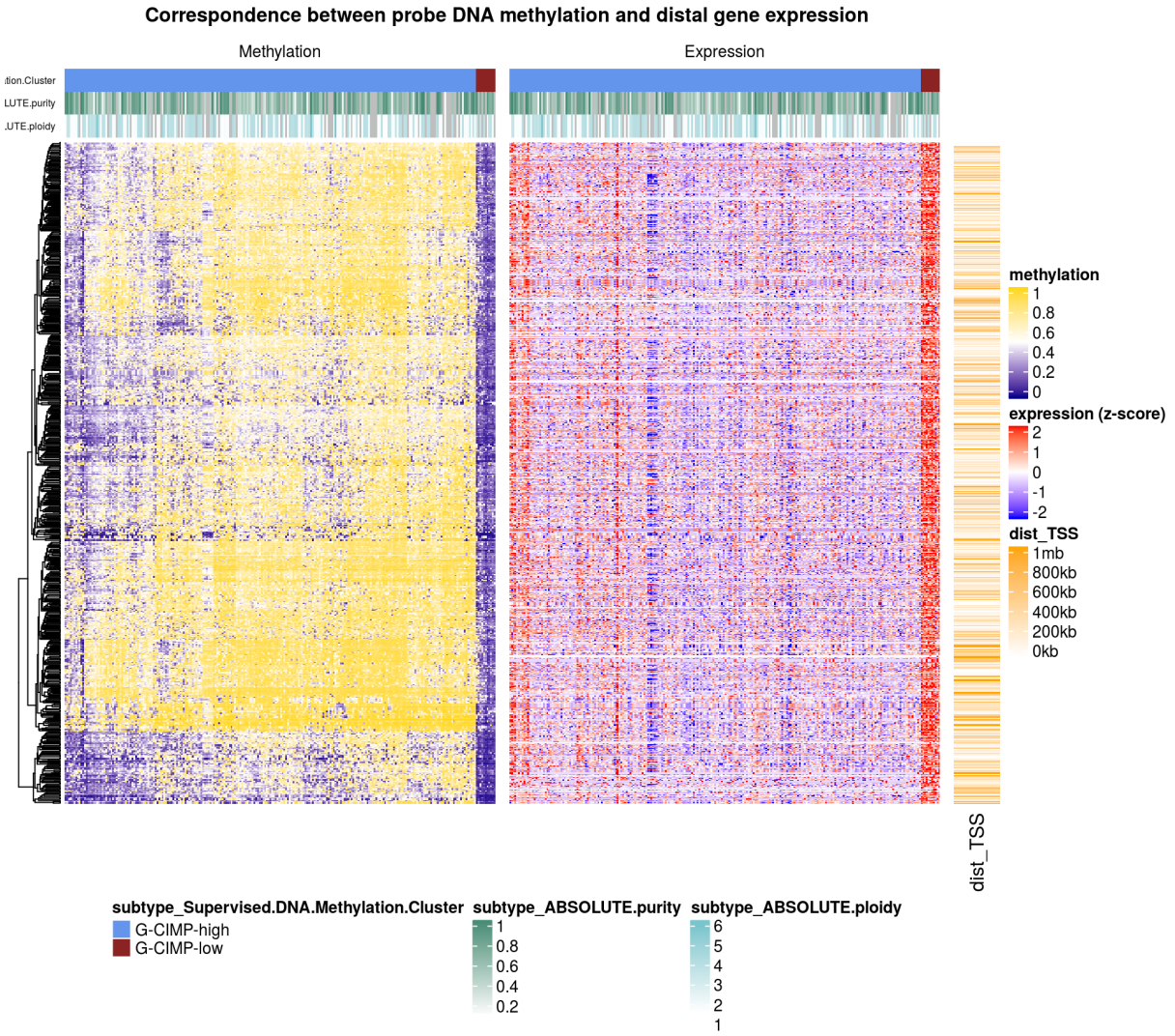


Figure 25 – G-CIMP analysis: Heatmap of paired probes and distal genes. The first heatmap (left one) shows DNA methylation β levels ranging from 0 (non-methylated) up to 1 (methylated probes). The second heatmap (middle one) shows z-scores for gene expression levels (standard deviations from each gene means). The last heatmap (right heatmap) shows the distance between the probe and gene anti-correlated. The top annotation in the heatmap was retried from Ceccarelli *et al.* (2016a)

against GBM. To validate these findings, biological experiments are needed which by either knocking down these TFs or by regulating the DNA methylation levels of those binding regions will be able to verify if the downstream genes are being regulated.

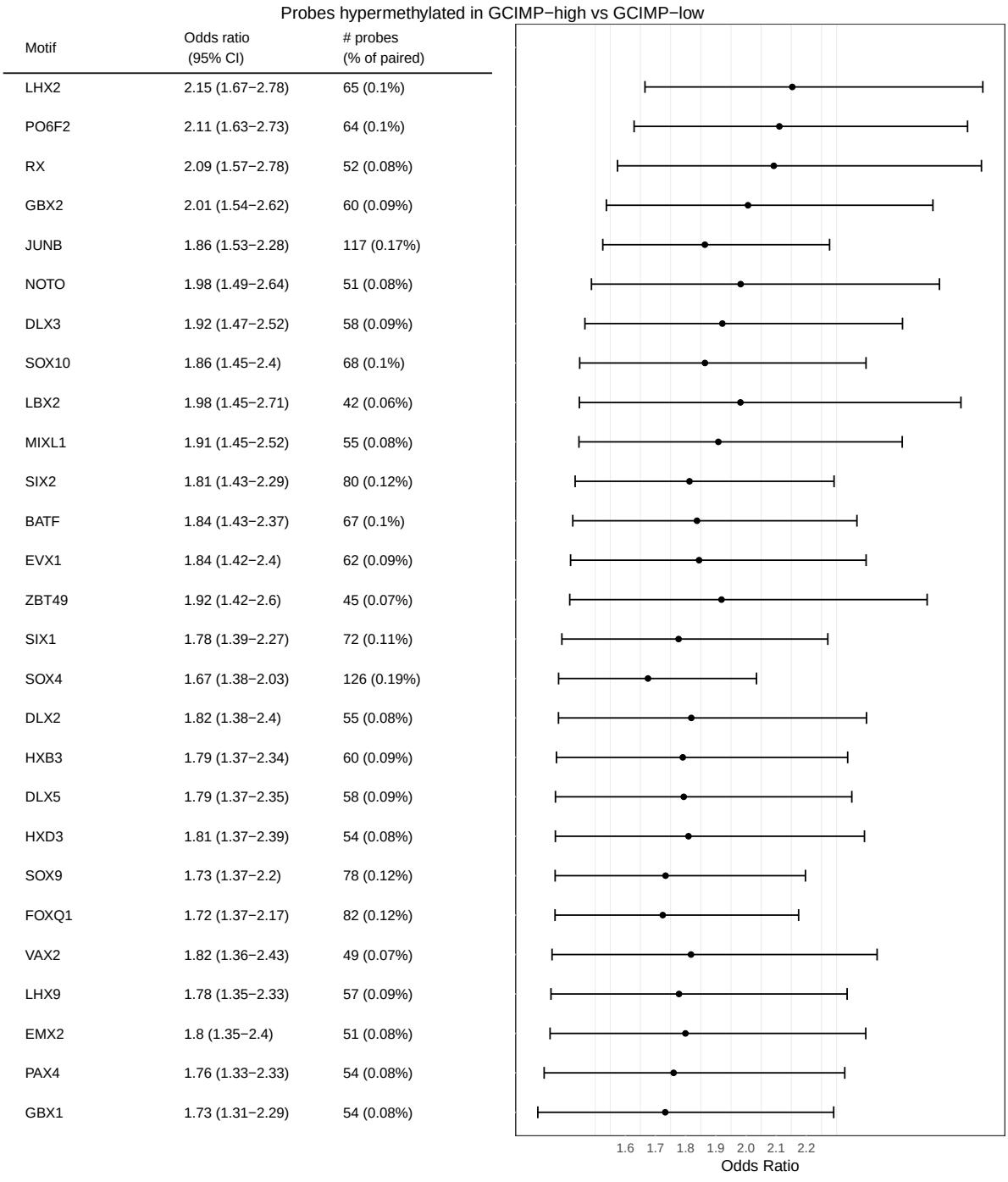


Figure 26 – Motif enrichment analysis: Odds Ratio (x axis) for the selected motifs with lower OR above 1.3. The range shows the 95% confidence interval for each Odds Ratio.

Table 19 – TF ranking analysis: statistic For each enriched motif the anti-correlation level of all human TFs expression level with average DNA methylation level at sites with a given motif was access and ranked by the $-log_{10}(P_{value})$, the most relevant one that belongs to the same family as the motif is shown in column *top.potential.TF.family* while the most relevant within the same sub-family classification is shown in column *top.potential.TF.subfamily*

motif	top.potential.TF.family	top.potential.TF.subfamily
LHX2_HUMAN.H11MO.0.A	NA	NA
PO6F2_HUMAN.H11MO.0.D	NA	NA
RX_HUMAN.H11MO.0.D	NA	NA
GBX2_HUMAN.H11MO.0.D	HOXD13	NA
JUNB_HUMAN.H11MO.0.A	NA	NA
NOTO_HUMAN.H11MO.0.D	HMX3	NA
DLX3_HUMAN.H11MO.0.C	HMX3	NA
SOX10_HUMAN.H11MO.0.B	SOX11	NA
LBX2_HUMAN.H11MO.0.D	HMX3	NA
MIXL1_HUMAN.H11MO.0.D	NA	NA
SIX2_HUMAN.H11MO.0.A	NA	NA
BATF_HUMAN.H11MO.1.A	NA	NA
EVX1_HUMAN.H11MO.0.D	HOXD13	NA
ZBT49_HUMAN.H11MO.0.D	ZSCAN16	ZSCAN16
SIX1_HUMAN.H11MO.0.A	NA	NA
SOX4_HUMAN.H11MO.0.B	SOX11	SOX11
DLX2_HUMAN.H11MO.0.D	HMX3	NA
HXB3_HUMAN.H11MO.0.D	HOXD13	HOXD3
DLX5_HUMAN.H11MO.0.D	HMX3	NA
HXD3_HUMAN.H11MO.0.D	HOXD13	HOXD3
SOX9_HUMAN.H11MO.0.B	SOX11	NA
FOXQ1_HUMAN.H11MO.0.C	FOXM1	NA
VAX2_HUMAN.H11MO.0.D	HMX3	NA
LHX9_HUMAN.H11MO.0.D	NA	NA
EMX2_HUMAN.H11MO.0.D	HMX3	NA
PAX4_HUMAN.H11MO.0.D	PAX3	NA
GBX1_HUMAN.H11MO.0.D	HOXD13	NA

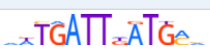
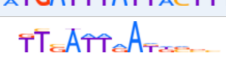
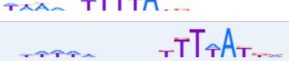
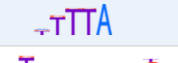
Model	LOGO	Transcription factor	Quality	TF family	TF subfamily
			D	HOX-related	
HXD3_HUMAN.H11MO.0.D		HOXD3 (GeneCards)	D	HOX-related factors{3.1.1}	HOX3{3.1.1.3}
HXD4_HUMAN.H11MO.0.D		HOXD4 (GeneCards)	D	HOX-related factors{3.1.1}	HOX4{3.1.1.4}
HXA5_HUMAN.H11MO.0.D		HOXA5 (GeneCards)	D	HOX-related factors{3.1.1}	HOX5{3.1.1.5}
HXA7_HUMAN.H11MO.0.D		HOXA7 (GeneCards)	D	HOX-related factors{3.1.1}	HOX6-7{3.1.1.6}
HXB6_HUMAN.H11MO.0.D		HOXB6 (GeneCards)	D	HOX-related factors{3.1.1}	HOX6-7{3.1.1.6}
HXC6_HUMAN.H11MO.0.D		HOXC6 (GeneCards)	D	HOX-related factors{3.1.1}	HOX6-7{3.1.1.6}
HXC8_HUMAN.H11MO.0.D		HOXC8 (GeneCards)	D	HOX-related factors{3.1.1}	HOX8{3.1.1.7}
HXD8_HUMAN.H11MO.0.D		HOXD8 (GeneCards)	D	HOX-related factors{3.1.1}	HOX8{3.1.1.7}
HXA11_HUMAN.H11MO.0.D		HOXA11 (GeneCards)	D	HOX-related factors{3.1.1}	HOX9-13{3.1.1.8}
HXC10_HUMAN.H11MO.0.D		HOXC10 (GeneCards)	D	HOX-related factors{3.1.1}	HOX9-13{3.1.1.8}
HXC11_HUMAN.H11MO.0.D		HOXC11 (GeneCards)	D	HOX-related factors{3.1.1}	HOX9-13{3.1.1.8}
HXC12_HUMAN.H11MO.0.D		HOXC12 (GeneCards)	D	HOX-related factors{3.1.1}	HOX9-13{3.1.1.8}
HXC13_HUMAN.H11MO.0.D		HOXC13 (GeneCards)	D	HOX-related factors{3.1.1}	HOX9-13{3.1.1.8}
HXD10_HUMAN.H11MO.0.D		HOXD10 (GeneCards)	D	HOX-related factors{3.1.1}	HOX9-13{3.1.1.8}
HXD11_HUMAN.H11MO.0.D		HOXD11 (GeneCards)	D	HOX-related factors{3.1.1}	HOX9-13{3.1.1.8}
HXD12_HUMAN.H11MO.0.D		HOXD12 (GeneCards)	D	HOX-related factors{3.1.1}	HOX9-13{3.1.1.8}
HXD13_HUMAN.H11MO.0.D		HOXD13 (GeneCards)	D	HOX-related factors{3.1.1}	HOX9-13{3.1.1.8}

Figure 27 – HOCOMOCO V11: HOX-related factors family. Transcription factors HOXD13 and HOXD3 are in the same family (HOX-related factors) but in different subfamilies.

CHAPTER

5

CONCLUSION

5.1 Conclusions

The main goal of this project has been to develop tools and workflows to perform integrative analysis as well as their application in the analysis of public cancer data. The motivations came from the fact that integrating data from different databases, stored in different formats and with different biological meanings is a complex process from both a computational point of view by dealing with big datasets which requires an optimized use of computational resources as well as from a biological point of view for dealing with a biological process still under study, often unknown, or even without a scientific consensus.

5.2 Conclusions and future studies

In this section, we list the conclusions and future studies for the results reported in the previous chapters.

5.2.1 Conclusions and future works of TCGAbiolinks

This work has presented an effort towards the search, download and prepare of data from the NCI's Genomic Data Commons (GDC) data portal for downstream analysis. It offered several single-dataset exploratory analysis such as PCA, clustering methods, box plots, differential gene expression analysis, differential methylated CpGs (DMCs) analysis, survival analysis and some integrative analysis such as the integration of differential methylated CpGs results with differential gene expression analysis to identify regions different methylated with the nearest gene level expression changed, and gene set enrichment analysis. As future works, it is required the improvement of integrative analysis of the starburst plot, this method integrates the results from two different analysis DEA (differential expression analysis) and DMR (differential methylated regions). There is, however, a problem that the results are analyzed considering the set and not the same sample, that means the expression and methylation data of a sample are not compared

to each other. It is inferred if a data set has a differentiated average of methylation and a mean of expression of the nearest differentiated gene, and may even be different populations. This flaw is corrected in ELMER algorithms.

5.2.2 Conclusions and future works of TCGAbiolinksGUI

This work created a Graphical User Interface (GUI) to our command-line tools, with the purpose to help users without programming knowledge to perform a deeper downstream analysis. Among possible improvements in future works are the export of figures in vector formats (PDF, SVG) that do not lose quality if altered (enlarged or reduced), the facilitation of the incorporation of external user data, which although can be incorporated if formatted in the currently defined standard, this can still be improved, the creation of a modularization of the tool, which would load only the packages chosen by the user in order to decrease the number of libraries needed to run the interface, which would be a challenge and may not be possible due to the limitations of the development tools used R/Shiny.

5.2.3 Conclusions and future works of ELMER

This work has presented an effort towards the integrative analysis performed using RNA-seq, DNA methylation, and histone marks to identify a candidate regulatory network. It mainly identifies distal regions with a difference in DNA methylation and correlates them with the gene expression levels of upstream and downstream genes. A motif enrichment analysis is performed in the regulatory regions from the anti-correlated pairs (loss of DNA methylation and gain o gene expression) to identify potential regulatory TF candidates. As future works we suggest to expand the algorithm to the promoter regions, to expand the DNA methylation analysis to accept whole-genome bisulfite sequencing (WGBS) data and to use mutation information instead of DNA methylation to identify regulatory regions mutated that might have affected the regulation of a upstream/downstream genes.

5.3 Publications, presentations and softwares of the Doctorate Period

The work produced during the Doctorate period in form of scientific articles, softwares and conference presentations is shown in the next subsections. The published scientific articles were divided into two groups, one with the first authorship ones, and the other with the co-authorship ones. Those are listed in the following subsection: "First-authored papers", "Co-authored papers","First-authored softwares" and "Co-authored softwares". The subsection "workshops and workflows" contains all the material created to help users to use the developed

tools. Finally, the subsection "Conferences & presentations" list all oral and poster presentations in international and national conferences.

5.3.1 First-authored papers

- COLAPRICO, A.; SILVA, T. C.; OLSEN, C.; GAROFANO, L.; CAVA, C.; GAROLINI, D.; SABEDOT, T. S.; MALTA, T. M.; PAGNOTTA, S. M.; CASTIGLIONI, I.; CECCARELLI, M.; BONTEMPI, G.; NOUSHMEHR, H. TCGAbiolinks: an r/bioconductor package for integrative analysis of tcga data. **Nucleic Acids Research**, v. 44, n. 8, p. e71, 2016. Available at: http://nar.oxfordjournals.org/content/44/8/e71.abstract.

  - Main contributions to the paper and tool: responsible for structuring of the package according to the standards of the Bioconductor project, creation of all data functions (query, download and prepare), some function for analysis and visualization (survival analysis, DNA methylation analysis and plots and integration of DNA methylation and gene expression and visualization in a starburst plot function), creation of unitary tests, creation of the documentation and content of all functions listed above, and the creation of half of the use case presented in the paper and package.

- SILVA, T.; COLAPRICO, A.; OLSEN, C.; D'ANGELO, F.; BONTEMPI, G.; CECCARELLI, M.; NOUSHMEHR, H. TCGA workflow: Analyze cancer genomics and epigenomics data using bioconductor packages [version 2; referees: 1 approved, 1 approved with reservations]. **F1000Research**, v. 5, n. 1542, 2016.

  - Main contributions to the paper and tool: Responsible for the development and testing the sections “Experimental data”, “DNA methylation analysis”, “Motif analysis” and “Integrative analysis”, creating and maintaining the workflow version available in the Bioconductor website.

- SILVA, T. C.; COLAPRICO, A.; OLSEN, C.; BONTEMPI, G.; CECCARELLI, M.; BERMAN, B. P.; NOUSHMEHR, H. Tcgabiolinksgui: A graphical user interface to analyze gdc cancer molecular and clinical data. **bioRxiv**, Cold Spring Harbor Labs Journals, 2017. Available at: http://www.biorxiv.org/content/early/2017/08/17/147496.

  - Main contributions to the paper and tool: Responsible for the creation of the graphical user interface structure and all the menus except the "Transcriptomic analysis - Network of inference and differential expression analysis" menus, the creation of the documentation, docker image, and tutorials (PDF and youtube videos).

- SILVA, T. C.; COETZEE, S. G.; YAO, L.; HAZELETT, D. J.; NOUSHMEHR, H.; BERMAN, B. P. Enhancer linking by methylation/expression relationships with the r package elmer version 2. **bioRxiv**, Cold Spring Harbor Labs Journals, p. 148726, 2017.

  - Main contributions to the paper and tool: responsible for re-writing the code to provide greater stability, performance, and ease of use, changing the main data structure to a standard Bioconductor data structures, integrating ELMER and TCGAbiolinks for data import from GDC, creating graphical user interface, creation of an interactive HTML reports, expansion of the algorithm to consider supervised cases, creation of the case of study in the article, restructuring of all the documentation by changing from a PDF format to an HTML and adding unit tests to the tool.

5.3.2 Co-authored papers

- LIN, D.-C.; DINH, H. Q.; XIE, J.-J.; MAYAKONDA, A.; SILVA, T. C.; JIANG, Y.-Y.; DING, L.-W.; HE, J.-Z.; XU, X.-E.; HAO, J.-J.; WANG, M.-R.; LI, C.; XU, L.-Y.; LI, E.-M.; BERMAN, B. P.; KOEFFLER, H. P. Identification of distinct mutational patterns and new driver genes in oesophageal squamous cell carcinomas and adenocarcinomas. **Gut**, BMJ Publishing Group, 2017. ISSN 0017-5749. Available at: http://gut.bmj.com/content/early/2017/09/02/gutjnl-2017-314607.

  - Main contributions to the paper: performed integrative analysis of esophageal cancer using the ELMER package.

- CECCARELLI, M.; BARTHEL, F.; MALTA, T.; SABEDOT, T.; SALAMA, S.; MURRAY, B.; AL., O. M. et. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. **Cell**, v. 164, n. 3, p. 550 – 563, 2016. ISSN 0092-8674. Available at: http://www.sciencedirect.com/science/article/pii/S009286741501692X.

  - Main contributions to the paper: All the DNA methylation and clinical data used in the paper was redownloaded and re-analyzed to validate the findings. Also, the analysis methods used in this article were made available in the TCGAbiolinks package.

- MALTA, T. M.; SOUZA, C. F.; SABEDOT, T. S.; SILVA, T. C.; MOSELLA, M. Q.; KALKANIS, S. N.; SNYDER, J.; CASTRO, A. V. B.; NOUSHMEHR, H. Glioma cpg island methylator phenotype (g-cimp): Biological and clinical implications. **bioRxiv**, Cold Spring Harbor Labs Journals, p. 169680, 2017.

  - Main contributions to the paper: helped to write the text about bioinformatic tools and analysis.

- CAVA, C.; COLAPRICO, A.; BERTOLI, G.; GRAUDENZI, A.; SILVA, T. C.; OLSEN, C.; NOUSHMEHR, H.; BONTEMPI, G.; MAURI, G.; CASTIGLIONI, I. Spidermir: An r/bioconductor package for integrative analysis with mirna data. **International journal of molecular sciences**, Multidisciplinary Digital Publishing Institute, v. 18, n. 2, p. 274, 2017.

  - Main contributions to the paper: responsible for the integration of TCGAbiolinks tool to the SpidermiR, help with package and documentation structure.

- GOOD, E. E.; MCCORMACK, E. P.; SILVA, T. C.; PARONETT, E. M.; NOUSHMER, H.; LEE, N. H.; MAYNARD, T. M.; LAMANTIA, A. S.; SHERMAN, J. H. Foxj1, a potential biomarker for glioma. Manuscript in prep.

  - Main contributions to the paper: performed differential expression analysis for high expressed FOXJ1 samples vs low expressed, integrated results with copy number alteration and mutation data using TCGAbiolinks.

5.3.3 First-authored softwares

TCGAbiolinks (version 2.0) An R/Bioconductor package for integrative analysis of TCGA data. Published in Bioconductor http://bioconductor.org/packages/TCGAbiolinks/. Source code available in GitHub https://github.com/BioinformaticsFMRP/TCGAbiolinks.

TCGAbiolinksGUI A Graphical User Interface to analyze cancer genomics and epigenomics data. Published in GitHub https://github.com/BioinformaticsFMRP/TCGAbiolinksGUI.

ELMER (version 2.0) Enhancer Linking by Methylation/Expression Relationship (ELMER) is package to identify tumor-specific changes in DNA methylation within distal enhancers, and link these enhancers to downstream target genes. Published in Bioconductor http://bioconductor.org/packages/ELMER/. Source code available in GitHub https://github.com/tiagochst/ELMER.

5.3.4 Co-authored softwares

SpidermiR: An R/Bioconductor package for integrative network analysis with miRNA data Published in Bioconductor http://bioconductor.org/packages/SpidermiR/. Source code available in GitHub https://github.com/claudiacava/SpidermiR.

5.3.5 Workshops and workflows

Workshop Integrative analysis workshop with TCGAbiolinks and ELMER. Dana Farber Cancer Institute, Boston, MA. Link to workshop: https://bioinformaticsfmrp.github.io/Bioc2017.TCGAbiolinks.ELMER/index.html. Source code available in GitHub https://github.com/BioinformaticsFMRP/Bioc2017.TCGAbiolinks.ELMER.

Workflow TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. Available at https://www.bioconductor.org/help/workflows/TCGAWorkflow/. Source code available in GitHub https://github.com/BioinformaticsFMRP/TCGAWorkflow.

5.3.6 Conferences & presentations

SNOLA 2016 - Update on Neuro-Oncology - Oral presentation - 04/19/2016 "EPIGENOMIC AND TRANSCRIPTOMIC ANALYSIS OF ADULT GLIOMA REVEALS CANDIDATE DRIVER TRANSCRIPTION FACTORS INVOLVED IN GLIOMA PROGRESSION." Windsor Barra Hotel, Rio de Janeiro - Brazil

Chromatin and Epigenetics - Poster presentation - 5/5/2017 TCGAbiolinksGUI: A Graphical User Interface to analyze cancer genomics and epigenomics data. EMBL, Heidelberg, Germany

Omics Seminar - Oral presentation - 06/06/2017 Enhancer Linking by Methylation/Expression Relationship: a case study using Breast Cancer. Cedars-Sinai Medical Center, Los Angeles, California.

Bioc2017 - Oral presentation - 07/28/2017 Workshop: Integrative analysis workshop with TCGAbiolinks and ELMER. Dana Farber Cancer Institute, Boston, MA. Link to workshop: https://bioinformaticsfmrp.github.io/Bioc2017.TCGAbiolinks.ELMER/index.html

From Single to Multiomics - Poster presentation - 11/13/2017 Enhancer Linking by Methylation/Expression Relationships with the R package (ELMER) version 2. EMBL, Heidelberg, Germany

BIBLIOGRAPHY

AICKIN, M.; GENSLER, H. Adjusting for multiple testing when reporting research results: the bonferroni vs holm methods. **American journal of public health**, American Public Health Association, v. 86, n. 5, p. 726–728, 1996. Cited on page 24.

AKEN, B. L.; AYLING, S.; BARRELL, D.; CLARKE, L.; CURWEN, V.; FAIRLEY, S.; BANET, J. F.; BILLIS, K.; GIRÓN, C. G.; HOURLIER, T.; HOWE, K.; KÄHÄRI, A.; KOKOCINSKI, F.; MARTIN, F. J.; MURPHY, D. N.; NAG, R.; RUFFIER, M.; SCHUSTER, M.; TANG, Y. A.; VOGEL, J.-H.; WHITE, S.; ZADISSA, A.; FLICEK, P.; SEARLE, S. M. J. The ensembl gene annotation system. **Database**, v. 2016, p. baw093, 2016. Available at: +http://dx.doi.org/10.1093/database/baw093. Cited on page 47.

ALZATE, O. **Neuroproteomics**. [S.l.]: CRC Press, 2009. Cited on page 19.

APWEILER, R.; BAIROCH, A.; WU, C. H.; BARKER, W. C.; BOECKMANN, B.; FERRO, S.; GASTEIGER, E.; HUANG, H.; LOPEZ, R.; MAGRANE, M. *et al.* Uniprot: the universal protein knowledgebase. **Nucleic acids research**, Oxford Univ Press, v. 32, n. suppl 1, p. D115–D119, 2004. Cited on page 52.

ARAN, D.; HELLMAN, A. Dna methylation of transcriptional enhancers and cancer predisposition. **Cell**, Elsevier, v. 154, n. 1, p. 11–13, 2013. Cited on page 44.

ARTANDI, S. E.; DEPINHO, R. A. Telomeres and telomerase in cancer. **Carcinogenesis**, Oxford University Press, v. 31, n. 1, p. 9–18, 2009. Cited on page 7.

ATTALI, D. **shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds**. [S.l.], 2017. R package version 0.9.1. Available at: https://CRAN.R-project.org/package=shinyjs. Cited on page 38.

AYDAY, E.; RAISARO, J. L.; HENGARTNER, U.; MOLYNEAUX, A.; HUBAUX, J.-P. Privacy-preserving processing of raw genomic data. In: **Data Privacy Management and Autonomous Spontaneous Security**. [S.l.]: Springer, 2014. p. 133–147. Cited on page 14.

BEN-PORATH, I.; THOMSON, M. W.; CAREY, V. J.; GE, R.; BELL, G. W.; REGEV, A.; WEINBERG, R. A. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. **Nature genetics**, Nature Publishing Group, v. 40, n. 5, p. 499–507, 2008. Cited on page 74.

BENETTI, R.; GONZALO, S.; JACO, I.; MUÑOZ, P.; GONZALEZ, S.; SCHOEFTNER, S.; MURCHISON, E.; ANDL, T.; CHEN, T.; KLATT, P. *et al.* A mammalian microrna cluster controls dna methylation and telomere recombination via rbl2-dependent regulation of dna methyltransferases. **Nature structural & molecular biology**, Nature Publishing Group, v. 15, n. 3, p. 268–279, 2008. Cited on page 8.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society. Series B (Methodological)**, Blackwell Publishing for the Royal Statistical Society, v. 57, n. 1, p. 289–300, 1995. ISSN 00359246. Available at: http://dx.doi.org/10.2307/2346101. Cited on page 35.

____. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the royal statistical society. Series B (Methodological)**, JSTOR, p. 289–300, 1995. Cited on page 48.

BERDASCO, M.; ESTELLER, M. Aberrant epigenetic landscape in cancer: how cellular identity goes awry. **Developmental cell**, Elsevier, v. 19, n. 5, p. 698–711, 2010. Cited on page 6.

BEREZIKOV, E. Evolution of microrna diversity and regulation in animals. **Nature Reviews Genetics**, v. 12, n. 12, 2011. Cited on page 8.

BERMAN, B. P.; WEISENBERGER, D. J.; AMAN, J. F.; HINOUE, T.; RAMJAN, Z.; LIU, Y.; NOUSHMEHR, H.; LANGE, C. P.; DIJK, C. M. van; TOLLENAAR, R. A. *et al.* Regions of focal dna hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. **Nature genetics**, Nature Publishing Group, v. 44, n. 1, p. 40–46, 2012. Cited on page 44.

BERNSTEIN, B. E.; KAMAL, M.; LINDBLAD-TOH, K.; BEKIRANOV, S.; BAILEY, D. K.; HUEBERT, D. J.; MCMAHON, S.; KARLSSON, E. K.; KULBOKAS, E. J.; GINGERAS, T. R. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. **Cell**, Elsevier, v. 120, n. 2, p. 169–181, 2005. Cited on page 21.

BERNSTEIN, B. E.; STAMATOYANNOPOULOS, J. A.; COSTELLO, J. F.; REN, B.; MILOSAVLJEVIC, A.; MEISSNER, A.; KELLIS, M.; MARRA, M. A.; BEAUDET, A. L.; ECKER, J. R.; FARNHAM, P. J.; HIRST, M.; LANDER, E. S.; MIKKELSEN, T. S.; THOMSON, J. A. The NIH Roadmap Epigenomics Mapping Consortium. **Nat. Biotechnol.**, v. 28, n. 10, p. 1045–1048, Oct 2010. Cited on page 14.

BEWICK, V.; CHEEK, L.; BALL, J. Statistics review 12: survival analysis. **Critical care**, BioMed Central, v. 8, n. 5, p. 389, 2004. Cited on page 26.

BIRD, A.; TAGGART, M.; FROMMER, M.; MILLER, O. J.; MACLEOD, D. A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. **Cell**, Elsevier, v. 40, n. 1, p. 91–99, 1985. Cited on page 8.

BLAND, J. M.; ALTMAN, D. G. The logrank test. **Bmj**, British Medical Journal Publishing Group, v. 328, n. 7447, p. 1073, 2004. Cited on page 26.

BONASIO, R.; TU, S.; REINBERG, D. Molecular signals of epigenetic states. **science**, American Association for the Advancement of Science, v. 330, n. 6004, p. 612–616, 2010. Cited 2 times on pages 20 and 21.

BRENET, F.; MOH, M.; FUNK, P.; FEIERSTEIN, E.; VIALE, A. J.; SOCCI, N. D.; SCANDURA, J. M. Dna methylation of the first exon is tightly linked to transcriptional silencing. **PloS one**, Public Library of Science, v. 6, n. 1, p. e14524, 2011. Cited on page 8.

BURKHART, D. L.; SAGE, J. Cellular mechanisms of tumour suppression by the retinoblastoma gene. **Nature Reviews Cancer**, v. 8, n. 9, 2008. Cited on page 6.

CAVA, C.; COLAPRICO, A.; BERTOLI, G.; GRAUDENZI, A.; SILVA, T. C.; OLSEN, C.; NOUSHMEHR, H.; BONTEMPI, G.; MAURI, G.; CASTIGLIONI, I. Spidermir: An r/bioconductor package for integrative analysis with mirna data. **International journal of molecular sciences**, Multidisciplinary Digital Publishing Institute, v. 18, n. 2, p. 274, 2017. Cited on page 86.

CECCARELLI, M.; BARTHEL, F.; MALTA, T.; SABEDOT, T.; SALAMA, S.; MURRAY, B.; AL., O. M. et. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. **Cell**, v. 164, n. 3, p. 550 – 563, 2016. ISSN 0092-8674. Available at: http://www.sciencedirect.com/science/article/pii/S009286741501692X. Cited 2 times on pages 78 and 86.

CECCARELLI, M.; BARTHEL, F. P.; MALTA, T. M.; SABEDOT, T. S.; SALAMA, S. R.; MURRAY, B. A.; MOROZOVA, O.; NEWTON, Y.; RADENBAUGH, A.; PAGNOTTA, S. M. *et al.* Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. **Cell**, Elsevier, v. 164, n. 3, p. 550–563, 2016. Cited 4 times on pages 1, 32, 35, and 74.

CEDAR, H.; BERGMAN, Y. Linking dna methylation and histone modification: patterns and paradigms. **Nature Reviews Genetics**, v. 10, n. 5, p. 295, 2009. Cited on page 21.

CERAMI, E.; GAO, J.; DOGRUSOZ, U.; GROSS, B. E.; SUMER, S. O.; AKSOY, B. A.; JACOBSEN, A.; BYRNE, C. J.; HEUER, M. L.; LARSSON, E. *et al.* **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data**. [S.l.]: AACR, 2012. Cited 2 times on pages 35 and 43.

CHANG, W.; Borges Ribeiro, B. **shinydashboard: Create Dashboards with 'Shiny'**. [S.l.], 2017. R package version 0.6.1. Available at: https://CRAN.R-project.org/package=shinydashboard. Cited on page 38.

CHANG, W.; CHENG, J.; ALLAIRE, J.; XIE, Y.; MCPHERSON, J. **shiny: Web Application Framework for R**. [S.l.], 2016. R package version 0.14. Available at: https://CRAN.R-project.org/package=shiny. Cited on page 31.

CHENG, N.; CHYTIL, A.; SHYR, Y.; JOLY, A.; MOSES, H. L. Transforming growth factor-β signaling–deficient fibroblasts enhance hepatocyte growth factor signaling in mammary carcinoma cells to promote scattering and invasion. **Molecular Cancer Research**, AACR, v. 6, n. 10, p. 1521–1533, 2008. Cited on page 5.

CIBULSKIS, K.; LAWRENCE, M. S.; CARTER, S. L.; SIVACHENKO, A.; JAFFE, D.; SOUGNEZ, C.; GABRIEL, S.; MEYERSON, M.; LANDER, E. S.; GETZ, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. **Nature biotechnology**, Nature Research, v. 31, n. 3, p. 213–219, 2013. Cited on page 15.

CIRIELLO, G.; GATZA, M. L.; BECK, A. H.; WILKERSON, M. D.; RHIE, S. K.; PASTORE, A.; ZHANG, H.; MCLELLAN, M.; YAU, C.; KANDOTH, C. *et al.* Comprehensive molecular portraits of invasive lobular breast cancer. **Cell**, Elsevier, v. 163, n. 2, p. 506–519, 2015. Cited on page 72.

CLANCY, S. Genetic mutation. **Nature Education**, v. 1, n. 1, p. 187, 2008. Cited 2 times on pages 9 and 12.

COCK, P. J.; FIELDS, C. J.; GOTO, N.; HEUER, M. L.; RICE, P. M. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. **Nucleic acids research**, Oxford University Press, v. 38, n. 6, p. 1767–1771, 2009. Cited on page 14.

COETZEE, S. G.; RAMJAN, Z.; DINH, H. Q.; BERMAN, B. P.; HAZELETT, D. J. Statehubstatepaintr: rapid and reproducible chromatin state evaluation for custom genome annotation.

bioRxiv, Cold Spring Harbor Labs Journals, p. 127720, 2017. Cited 4 times on pages 16, 22, 50, and 68.

COLAPRICO, A.; SILVA, T. C.; OLSEN, C.; GAROFANO, L.; CAVA, C.; GAROLINI, D.; SABEDOT, T. S.; MALTA, T. M.; PAGNOTTA, S. M.; CASTIGLIONI, I. *et al.* TCGAbiolinks: an r/bioconductor package for integrative analysis of tcga data. **Nucleic acids research**, Oxford Univ Press, p. gkv1507, 2015. Cited 2 times on pages 52 and 55.

COLAPRICO, A.; SILVA, T. C.; OLSEN, C.; GAROFANO, L.; CAVA, C.; GAROLINI, D.; SABEDOT, T. S.; MALTA, T. M.; PAGNOTTA, S. M.; CASTIGLIONI, I.; CECCARELLI, M.; BONTEMPI, G.; NOUSHMEHR, H. TCGAbiolinks: an r/bioconductor package for integrative analysis of tcga data. **Nucleic Acids Research**, v. 44, n. 8, p. e71, 2016. Available at: http://nar.oxfordjournals.org/content/44/8/e71.abstract. Cited 2 times on pages 38 and 85.

COMISC. **CONAN - Search**. 2017. http://cancer.sanger.ac.uk/cosmic/help/conan. Accessed: 2017-09-30. Cited on page 12.

CONSORTIUM, E. P. *et al.* A user's guide to the encyclopedia of dna elements (encode). **PLoS Biol**, v. 9, n. 4, p. e1001046, 2011. Cited on page 13.

CREYGHTON, M. P.; CHENG, A. W.; WELSTEAD, G. G.; KOOISTRA, T.; CAREY, B. W.; STEINE, E. J.; HANNA, J.; LODATO, M. A.; FRAMPTON, G. M.; SHARP, P. A. *et al.* Histone h3k27ac separates active from poised enhancers and predicts developmental state. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 107, n. 50, p. 21931–21936, 2010. Cited on page 21.

DASGUPTA, A.; SUN, Y. V.; KÖNIG, I. R.; BAILEY-WILSON, J. E.; MALLEY, J. D. Brief review of regression-based and machine learning methods in genetic epidemiology: the genetic analysis workshop 17 experience. **Genetic epidemiology**, Wiley Online Library, v. 35, n. S1, 2011. Cited on page 29.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application**. [S.l.]: Cambridge university press, 1997. Cited on page 25.

DEFAYS, D. An efficient algorithm for a complete link method. **The Computer Journal**, Oxford University Press, v. 20, n. 4, p. 364–366, 1977. Cited on page 29.

DENG, M.; BRÄGELMANN, J.; KRYUKOV, I.; SARAIVA-AGOSTINHO, N.; PERNER, S. Firebrowser: an r client to the broad institute's firehose pipeline. **Database**, Oxford University Press, v. 2017, n. 1, p. baw160, 2017. Cited on page 35.

DOWN, T. A.; HUBBARD, T. J. Computational detection and location of transcription start sites in mammalian genomic dna. **Genome research**, Cold Spring Harbor Lab, v. 12, n. 3, p. 458–461, 2002. Cited on page 28.

DURINCK, S.; MOREAU, Y.; KASPRZYK, A.; DAVIS, S.; MOOR, B. D.; BRAZMA, A.; HUBER, W. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. **Bioinformatics**, Oxford Univ Press, v. 21, n. 16, p. 3439–3440, 2005. Cited 2 times on pages 49 and 52.

DURINCK, S.; SPELLMAN, P. T.; BIRNEY, E.; HUBER, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. **Nature protocols**, Nature Publishing Group, v. 4, n. 8, p. 1184–1191, 2009. Cited 3 times on pages 46, 49, and 52.

Editorial Nature Genetics. Credit for code. **Nat Genet**, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., v. 46, n. 1, p. 1–1, Jan 2014. ISSN 1061-4036. Editorial. Available at: http://dx.doi.org/10.1038/ng.2869. Cited on page 3.

EMBL-EBI. **Protein-protein interaction networks**. 2017. https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/protein-protein-interaction-networks. Accessed: 2017-10-17. Cited on page 19.

ENCODE. **Transcription Factor ChIP-seq Data Standards and Processing Pipeline**. 2017. https://www.encodeproject.org/chip-seq/transcription_factor/. Accessed: 2017-10-17. Cited on page 16.

ERLICH, Y.; NARAYANAN, A. Routes for breaching and protecting genetic privacy. **Nature Reviews Genetics**, Nature Research, v. 15, n. 6, p. 409–421, 2014. Cited on page 14.

ERNST, J.; KELLIS, M. Chromhmm: automating chromatin-state discovery and characterization. **Nature methods**, Nature Research, v. 9, n. 3, p. 215–216, 2012. Cited 2 times on pages 16 and 22.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. Hierarchical clustering. **Cluster Analysis, 5th Edition**, Wiley Online Library, p. 71–110, 2011. Cited on page 29.

FAN, Y.; XI, L.; HUGHES, D. S.; ZHANG, J.; ZHANG, J.; FUTREAL, P. A.; WHEELER, D. A.; WANG, W. Muse: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. **Genome biology**, BioMed Central, v. 17, n. 1, p. 178, 2016. Cited on page 15.

FEUK, L.; CARSON, A. R.; SCHERER, S. W. Structural variation in the human genome. **Nature Reviews Genetics**, v. 7, n. 2, 2006. Cited 2 times on pages 11 and 12.

FINGERMAN, I. M.; MCDANIEL, L.; ZHANG, X.; RATZAT, W.; HASSAN, T.; JIANG, Z.; COHEN, R. F.; SCHULER, G. D. NCBI Epigenomics: a new public resource for exploring epigenomic data sets. **Nucleic Acids Res.**, v. 39, n. Database issue, p. D908–912, Jan 2011. Cited on page 14.

FISHER, R. A. On the interpretation of χ 2 from contingency tables, and the calculation of p. **Journal of the Royal Statistical Society**, JSTOR, v. 85, n. 1, p. 87–94, 1922. Cited on page 51.

FLOREK, K.; ŁUKASZEWICZ, J.; PERKAL, J.; STEINHAUS, H.; ZUBRZYCKI, S. Sur la liaison et la division des points d'un ensemble fini. In: **Colloquium Mathematicae**. [S.l.: s.n.], 1951. v. 2, n. 3-4, p. 282–285. Cited on page 29.

FUKS, F.; BURGERS, W. A.; BREHM, A.; HUGHES-DAVIES, L.; KOUZARIDES, T. Dna methyltransferase dnmt1 associates with histone deacetylase activity. **Nature genetics**, Nature Publishing Group, v. 24, n. 1, p. 88–92, 2000. Cited on page 8.

FUKS, F.; HURD, P. J.; DEPLUS, R.; KOUZARIDES, T. The dna methyltransferases associate with hp1 and the suv39h1 histone methyltransferase. **Nucleic acids research**, Oxford University Press, v. 31, n. 9, p. 2305–2312, 2003. Cited on page 8.

FUNCIVAR. 2017. https://github.com/Simon-Coetzee/funcivar. Accessed: 2017-06-09. Cited 2 times on pages 50 and 68.

GAO, J.; AKSOY, B. A.; DOGRUSOZ, U.; DRESDNER, G.; GROSS, B.; SUMER, S. O.; SUN, Y.; JACOBSEN, A.; SINHA, R.; LARSSON, E. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. **Science signaling**, Science Signaling, v. 6, n. 269, p. pl1–pl1, 2013. Cited 2 times on pages 35 and 43.

GDC. **GDC MAF format**. 2017. https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/. Accessed: 2017-09-30. Cited on page 9.

GEIMAN, T. M.; SANKPAL, U. T.; ROBERTSON, A. K.; ZHAO, Y.; ZHAO, Y.; ROBERTSON, K. D. Dnmt3b interacts with hsnf2h chromatin remodeling enzyme, hdacs 1 and 2, and components of the histone methylation system. **Biochemical and biophysical research communications**, Elsevier, v. 318, n. 2, p. 544–555, 2004. Cited on page 8.

GENTLEMAN, R. C.; CAREY, V. J.; BATES, D. M.; BOLSTAD, B.; DETTLING, M.; DUDOIT, S.; ELLIS, B.; GAUTIER, L.; GE, Y.; GENTRY, J. *et al.* Bioconductor: open software development for computational biology and bioinformatics. **Genome biology**, BioMed Central Ltd, v. 5, n. 10, p. R80, 2004. Cited 2 times on pages 32 and 36.

GIFFORD, C. A.; ZILLER, M. J.; GU, H.; TRAPNELL, C.; DONAGHEY, J.; TSANKOV, A.; SHALEK, A. K.; KELLEY, D. R.; SHISHKIN, A. A.; ISSNER, R. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. **Cell**, Elsevier, v. 153, n. 5, p. 1149–1163, 2013. Cited on page 20.

GONG, C. *et al.* Foxa1 repression is associated with loss of brca1 and increased promoter methylation and chromatin silencing in breast cancer. **Oncogene**, Nature Publishing Group, v. 34, n. 39, p. 5012–5024, 2015. Cited on page 74.

GOOD, E. E.; MCCORMACK, E. P.; SILVA, T. C.; PARONETT, E. M.; NOUSHMER, H.; LEE, N. H.; MAYNARD, T. M.; LAMANTIA, A. S.; SHERMAN, J. H. Foxj1, a potential biomarker for glioma. Manuscript in prep. Cited on page 87.

GRANT, C. E.; BAILEY, T. L.; NOBLE, W. S. Fimo: scanning for occurrences of a given motif. **Bioinformatics**, Oxford University Press, v. 27, n. 7, p. 1017–1018, 2011. Cited on page 20.

GROSSMAN, R. L.; HEATH, A. P.; FERRETTI, V.; VARMUS, H. E.; LOWY, D. R.; KIBBE, W. A.; STAUDT, L. M. Toward a shared vision for cancer genomic data. **New England Journal of Medicine**, Mass Medical Soc, v. 375, n. 12, p. 1109–1112, 2016. Cited 2 times on pages 45 and 52.

GU, Z.; EILS, R.; SCHLESNER, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. **Bioinformatics**, 2016. Available at: http://bioinformatics.oxfordjournals.org/content/early/2016/05/20/bioinformatics.btw313.abstract. Cited 2 times on pages 38 and 40.

HAN, L.; WITMER, P. D. W.; CASEY, E.; VALLE, D.; SUKUMAR, S. Dna methylation regulates microrna expression. **Cancer biology & therapy**, Taylor & Francis, v. 6, n. 8, p. 1290–1294, 2007. Cited on page 8.

HANAHAN, D.; WEINBERG, R. A. Hallmarks of cancer: the next generation. **cell**, Elsevier, v. 144, n. 5, p. 646–674, 2011. Cited 2 times on pages 5 and 6.

HAWKINS, R. D.; HON, G. C.; LEE, L. K.; NGO, Q.; LISTER, R.; PELIZZOLA, M.; EDSALL, L. E.; KUAN, S.; LUU, Y.; KLUGMAN, S. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. **Cell stem cell**, Elsevier, v. 6, n. 5, p. 479–491, 2010. Cited on page 20.

HAWKINS, R. D.; HON, G. C.; REN, B. Next-generation genomics: an integrative approach. **Nat. Rev. Genet.**, v. 11, n. 7, p. 476–486, Jul 2010. Cited on page 13.

HEINTZMAN, N. D.; HON, G. C.; HAWKINS, R. D.; KHERADPOUR, P.; STARK, A.; HARP, L. F.; YE, Z.; LEE, L. K.; STUART, R. K.; CHING, C. W. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. **Nature**, Nature Publishing Group, v. 459, n. 7243, p. 108–112, 2009. Cited on page 21.

HEINTZMAN, N. D.; STUART, R. K.; HON, G.; FU, Y.; CHING, C. W.; HAWKINS, R. D.; BARRERA, L. O.; CALCAR, S. V.; QU, C.; CHING, K. A. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. **Nature genetics**, Nature Publishing Group, v. 39, n. 3, p. 311–318, 2007. Cited on page 21.

HEINZ, S.; BENNER, C.; SPANN, N.; BERTOLINO, E.; LIN, Y. C.; LASLO, P.; CHENG, J. X.; MURRE, C.; SINGH, H.; GLASS, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. **Molecular cell**, Elsevier, v. 38, n. 4, p. 576–589, 2010. Cited 2 times on pages 20 and 50.

HUBER, W.; CAREY, V. J.; GENTLEMAN, R.; ANDERS, S.; CARLSON, M.; CARVALHO, B. S.; BRAVO, H. C.; DAVIS, S.; GATTO, L.; GIRKE, T. *et al.* Orchestrating high-throughput genomic analysis with bioconductor. **Nature methods**, Nature Publishing Group, v. 12, n. 2, p. 115–121, 2015. Cited 2 times on pages 32 and 46.

HUTTENHOWER, C.; HOFMANN, O. A quick guide to large-scale genomic data mining. **PLoS computational biology**, Public Library of Science, v. 6, n. 5, p. e1000779, 2010. Cited on page 14.

INSTITUTE, B. **GTAK**. 2017. https://gatkforums.broadinstitute.org/gatk/discussion/8815/oncotator-variant-classification-and-secondary-variant-classification. Accessed: 2017-09-30. Cited 2 times on pages 9 and 10.

JACKSON, S. P.; BARTEK, J. The dna-damage response in human biology and disease. **Nature**, Nature Publishing Group, v. 461, n. 7267, p. 1071–1078, 2009. Cited on page 7.

JAYARAM, N.; USVYAT, D.; MARTIN, A. C. Evaluating tools for transcription factor binding site prediction. **BMC bioinformatics**, BioMed Central, p. 1, 2016. Cited on page 20.

JIANG, B.-H.; LIU, L.-Z. Pi3k/pten signaling in angiogenesis and tumorigenesis. **Advances in cancer research**, Elsevier, v. 102, p. 19–65, 2009. Cited on page 6.

JIN, H.; WANG, X.; YING, J.; WONG, A. H.; CUI, Y.; SRIVASTAVA, G.; SHEN, Z.-Y.; LI, E.-M.; ZHANG, Q.; JIN, J. *et al.* Epigenetic silencing of a ca2+-regulated ras gtpase-activating protein rasal defines a new mechanism of ras activation in human cancers. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 104, n. 30, p. 12353–12358, 2007. Cited on page 6.

JR, J. H. W. Hierarchical grouping to optimize an objective function. **Journal of the American statistical association**, Taylor & Francis, v. 58, n. 301, p. 236–244, 1963. Cited on page 29.

KANNAN, L.; RAMOS, M.; RE, A.; EL-HACHEM, N.; SAFIKHANI, Z.; GENDOO, D. M.; DAVIS, S.; GOMEZ-CABRERO, D.; CASTELO, R.; HANSEN, K. D. *et al.* Public data and open source tools for multi-assay genomic investigation of disease. **Briefings in bioinformatics**, Oxford Univ Press, p. bbv080, 2015. Cited on page 16.

KASSAMBARA, A.; KOSINSKI, M. **survminer: Drawing Survival Curves using 'ggplot2'**. [S.l.], 2017. R package version 0.4.0. Available at: https://CRAN.R-project.org/package=survminer. Cited on page 40.

KITCHEN, C. M. Nonparametric versus parametric tests of location in biomedical research. **American journal of ophthalmology**, NIH Public Access, v. 147, n. 4, p. 571, 2009. Cited on page 25.

KOBOLDT, D. C.; ZHANG, Q.; LARSON, D. E.; SHEN, D.; MCLELLAN, M. D.; LIN, L.; MILLER, C. A.; MARDIS, E. R.; DING, L.; WILSON, R. K. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. **Genome research**, Cold Spring Harbor Lab, v. 22, n. 3, p. 568–576, 2012. Cited on page 15.

KORKOLA, J.; GRAY, J. W. Breast cancer genomes—form and function. **Current opinion in genetics & development**, Elsevier, v. 20, n. 1, p. 4–14, 2010. Cited on page 7.

KULAKOVSKIY, I. V.; MEDVEDEVA, Y. A.; SCHAEFER, U.; KASIANOV, A. S.; VORONTSOV, I. E.; BAJIC, V. B.; MAKEEV, V. J. Hocomoco: a comprehensive collection of human transcription factor binding sites models. **Nucleic acids research**, Oxford Univ Press, v. 41, n. D1, p. D195–D202, 2013. Cited on page 20.

KULAKOVSKIY, I. V.; VORONTSOV, I. E.; YEVSHIN, I. S.; SOBOLEVA, A. V.; KASIANOV, A. S.; ASHOOR, H.; BA-ALAWI, W.; BAJIC, V. B.; MEDVEDEVA, Y. A.; KOLPAKOV, F. A. *et al.* Hocomoco: expansion and enhancement of the collection of transcription factor binding sites models. **Nucleic acids research**, Oxford Univ Press, v. 44, n. D1, p. D116–D125, 2016. Cited 2 times on pages 50 and 52.

KUNDAJE, A.; MEULEMAN, W.; ERNST, J.; BILENKY, M.; YEN, A.; HERAVI-MOUSSAVI, A.; KHERADPOUR, P.; ZHANG, Z.; WANG, J.; ZILLER, M. J. *et al.* Integrative analysis of 111 reference human epigenomes. **Nature**, Nature Publishing Group, v. 518, n. 7539, p. 317–330, 2015. Cited on page 22.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with bowtie 2. **Nature methods**, Nature Research, v. 9, n. 4, p. 357–359, 2012. Cited on page 14.

LARSON, D. E.; HARRIS, C. C.; CHEN, K.; KOBOLDT, D. C.; ABBOTT, T. E.; DOOLING, D. J.; LEY, T. J.; MARDIS, E. R.; WILSON, R. K.; DING, L. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. **Bioinformatics**, Oxford University Press, v. 28, n. 3, p. 311–317, 2011. Cited on page 15.

LAWRENCE, M.; HUBER, W.; PAGES, H.; ABOYOUN, P.; CARLSON, M.; GENTLEMAN, R.; MORGAN, M. T.; CAREY, V. J. Software for computing and annotating genomic ranges. **PLoS computational biology**, Public Library of Science, v. 9, n. 8, p. e1003118, 2013. Cited 2 times on pages 17 and 34.

LEE, E.-J.; RATH, P.; LIU, J.; RYU, D.; PEI, L.; NOONEPALLE, S. K.; SHULL, A. Y.; FENG, Q.; LITOFSKY, N. S.; MILLER, D. C. *et al.* Identification of global dna methylation signatures

in glioblastoma-derived cancer stem cells. **Journal of Genetics and Genomics**, Elsevier, v. 42, n. 7, p. 355–371, 2015. Cited on page 77.

LI, G.; RUAN, X.; AUERBACH, R. K.; SANDHU, K. S.; ZHENG, M.; WANG, P.; POH, H. M.; GOH, Y.; LIM, J.; ZHANG, J. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. **Cell**, Elsevier, v. 148, n. 1, p. 84–98, 2012. Cited on page 72.

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; MARTH, G.; ABECASIS, G.; DURBIN, R. The sequence alignment/map format and samtools. **Bioinformatics**, Oxford University Press, v. 25, n. 16, p. 2078–2079, 2009. Cited on page 14.

LIBBRECHT, M. W.; NOBLE, W. S. Machine learning in genetics and genomics. **Nature Reviews. Genetics**, NIH Public Access, v. 16, n. 6, p. 321, 2015. Cited on page 30.

LIN, D.-C.; DINH, H. Q.; XIE, J.-J.; MAYAKONDA, A.; SILVA, T. C.; JIANG, Y.-Y.; DING, L.-W.; HE, J.-Z.; XU, X.-E.; HAO, J.-J.; WANG, M.-R.; LI, C.; XU, L.-Y.; LI, E.-M.; BERMAN, B. P.; KOEFFLER, H. P. Identification of distinct mutational patterns and new driver genes in oesophageal squamous cell carcinomas and adenocarcinomas. **Gut**, BMJ Publishing Group, 2017. ISSN 0017-5749. Available at: http://gut.bmj.com/content/early/2017/09/02/gutjnl-2017-314607. Cited on page 86.

LUJAMBIO, A.; CALIN, G. A.; VILLANUEVA, A.; ROPERO, S.; SÁNCHEZ-CÉSPEDES, M.; BLANCO, D.; MONTUENGA, L. M.; ROSSI, S.; NICOLOSO, M. S.; FALLER, W. J. *et al.* A microrna dna methylation signature for human cancer metastasis. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 105, n. 36, p. 13556–13561, 2008. Cited on page 8.

LUO, W.; BROUWER, C. Pathview: an r/bioconductor package for pathway-based data integration and visualization. **Bioinformatics**, Oxford Univ Press, v. 29, n. 14, p. 1830–1831, 2013. Cited 2 times on pages 38 and 40.

MALTA, T. M.; SOUZA, C. F.; SABEDOT, T. S.; SILVA, T. C.; MOSELLA, M. Q.; KALKANIS, S. N.; SNYDER, J.; CASTRO, A. V. B.; NOUSHMEHR, H. Glioma cpg island methylator phenotype (g-cimp): Biological and clinical implications. **bioRxiv**, Cold Spring Harbor Labs Journals, p. 169680, 2017. Cited on page 86.

MANNING, C. D.; SCHÜTZE, H. *et al.* **Foundations of statistical natural language processing**. [S.l.]: MIT Press, 1999. Cited on page 29.

MATTHEWS, D. E.; FAREWELL, V. T. *et al.* **Using and understanding medical statistics**. [S.l.]: Karger Basel, Switzerland;, 1996. Cited on page 28.

MAYAKONDA, A.; KOEFFLER, P. H. Maftools: Efficient analysis, visualization and summarization of maf files from large-scale cohort based cancer studies. **BioRxiv**, 2016. Cited 2 times on pages 38 and 40.

MCPHERSON, J. D. Next-generation gap. **Nature methods**, Nature Publishing Group, v. 6, p. S2–S5, 2009. Cited on page 14.

MERMEL, C. H.; SCHUMACHER, S. E.; HILL, B.; MEYERSON, M. L.; BEROUKHIM, R.; GETZ, G. Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. **Genome biology**, BioMed Central, v. 12, n. 4, p. 1, 2011. Cited on page 16.

MITCHELL, T. M. *et al.* **Machine learning. WCB**. [S.l.]: McGraw-Hill Boston, MA:, 1997. Cited on page 28.

MOHN, F.; WEBER, M.; REBHAN, M.; ROLOFF, T. C.; RICHTER, J.; STADLER, M. B.; BIBEL, M.; SCHÜBELER, D. Lineage-specific polycomb targets and de novo dna methylation define restriction and potential of neuronal progenitors. **Molecular cell**, Elsevier, v. 30, n. 6, p. 755–766, 2008. Cited on page 8.

MOORE, L. D.; LE, T.; FAN, G. Dna methylation and its basic function. **Neuropsychopharmacology**, Nature Publishing Group, v. 38, n. 1, p. 23, 2013. Cited 4 times on pages 7, 8, 9, and 21.

MORALES-DELGADO, N.; CASTRO-ROBLES, B.; FERRÁN, J. L.; TORRE, M. Martinez-de-la; PUELLES, L.; DÍAZ, C. Regionalized differentiation of crh, trh, and ghrh peptidergic neurons in the mouse hypothalamus. **Brain Structure and Function**, Springer, v. 219, n. 3, p. 1083–1111, 2014. Cited on page 77.

MORGAN, M.; OBENCHAIN, V.; HESTER, J.; PAGÈS, H. **SummarizedExperiment: SummarizedExperiment container**. [S.l.], 2017. R package version 1.7.5. Cited on page 18.

NCI. **GDC MAF Format v.1.0.0**. 2017. https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/. Accessed: 2017-10-17. Cited on page 15.

____. **HTSeq-FPKM**. 2017. https://docs.gdc.cancer.gov/Encyclopedia/pages/HTSeq-FPKM/. Accessed: 2017-10-17. Cited on page 15.

____. **HTSeq-FPKM-UQ**. 2017. https://docs.gdc.cancer.gov/Encyclopedia/pages/HTSeq-FPKM-UQ/. Accessed: 2017-10-17. Cited on page 15.

____. **The Next Generation Cancer Knowledge Network**. 2017. https://gdc.cancer.gov/. Accessed: 2017-10-17. Cited on page 13.

____. ____. 2017. https://gdc.cancer.gov/access-data/data-access-processes-and-tools. Accessed: 2017-10-17. Cited on page 14.

____. **TCGA Code Tables**. 2017. https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables. Accessed: 2017-10-17. Cited on page 16.

NEGRINI, S.; GORGOULIS, V. G.; HALAZONETIS, T. D. Genomic instability—an evolving hallmark of cancer. **Nature reviews Molecular cell biology**, Nature Publishing Group, v. 11, n. 3, p. 220–228, 2010. Cited on page 7.

NETWORK, C. G. A. *et al.* Comprehensive molecular portraits of human breast tumors. **Nature**, NIH Public Access, v. 490, n. 7418, p. 61, 2012. Cited 2 times on pages 54 and 55.

NISHIDA, H.; SUZUKI, T.; KONDO, S.; MIURA, H.; FUJIMURA, Y.-i.; HAYASHIZAKI, Y. Histone h3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell. **Chromosome Research**, Springer, v. 14, n. 2, p. 203–211, 2006. Cited on page 21.

NORTH, B.; CURTIS, D.; SHAM, P. A note on the calculation of empirical p values from monte carlo procedures. **American journal of human genetics**, Elsevier, v. 72, n. 2, p. 498, 2003. Cited on page 25.

NORTH, B. V.; CURTIS, D.; SHAM, P. C. A note on the calculation of empirical p values from monte carlo procedures. **American journal of human genetics**, Elsevier, v. 71, n. 2, p. 439, 2002. Cited on page 25.

NOUSHMEHR, H.; WEISENBERGER, D. J.; DIEFES, K.; PHILLIPS, H. S.; PUJARA, K.; BERMAN, B. P.; PAN, F.; PELLOSKI, C. E.; SULMAN, E. P.; BHAT, K. P. *et al.* Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. **Cancer cell**, Elsevier, v. 17, n. 5, p. 510–522, 2010. Cited 3 times on pages 35, 53, and 55.

OOI, S. K.; QIU, C.; BERNSTEIN, E.; LI, K.; JIA, D.; YANG, Z.; ERDJUMENT-BROMAGE, H.; TEMPST, P.; LIN, S.-P.; ALLIS, C. D. *et al.* Dnmt3l connects unmethylated lysine 4 of histone h3 to de novo methylation of dna. **Nature**, NIH Public Access, v. 448, n. 7154, p. 714, 2007. Cited on page 8.

ORCHARD, S.; AMMARI, M.; ARANDA, B.; BREUZA, L.; BRIGANTI, L.; BROACKES-CARTER, F.; CAMPBELL, N. H.; CHAVALI, G.; CHEN, C.; DEL-TORO, N. *et al.* The mintact project—intact as a common curation platform for 11 molecular interaction databases. **Nucleic acids research**, Oxford University Press, v. 42, n. D1, p. D358–D363, 2013. Cited on page 19.

PARAB, S.; BHALERAO, S. Choosing statistical test. **International journal of Ayurveda research**, Medknow Publications, v. 1, n. 3, p. 187, 2010. Cited on page 25.

PEDERSEN, T. L. **shinyFiles: A Server-Side File System Viewer for Shiny**. [S.l.], 2016. R package version 0.6.2. Available at: https://CRAN.R-project.org/package=shinyFiles. Cited on page 38.

PEROU, C. M.; SORLIE, T.; EISEN, M. B.; RIJN, M. V. D. *et al.* Molecular portraits of human breast tumours. **nature**, Nature Publishing Group, v. 406, n. 6797, p. 747, 2000. Cited on page 72.

PETERS, A. H.; KUBICEK, S.; MECHTLER, K.; O'SULLIVAN, R. J.; DERIJCK, A. A.; PEREZ-BURGOS, L.; KOHLMAIER, A.; OPRAVIL, S.; TACHIBANA, M.; SHINKAI, Y. *et al.* Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. **Molecular cell**, Elsevier, v. 12, n. 6, p. 1577–1589, 2003. Cited on page 21.

PLATFORMS, A. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. Mass Medical Soc, 2015. Cited 2 times on pages 53 and 54.

PORTALES-CASAMAR, E.; THONGJUEA, S.; KWON, A. T.; ARENILLAS, D.; ZHAO, X.; VALEN, E.; YUSUF, D.; LENHARD, B.; WASSERMAN, W. W.; SANDELIN, A. Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. **Nucleic acids research**, Oxford University Press, v. 38, n. suppl_1, p. D105–D110, 2009. Cited on page 20.

POZSGAI, E.; SCHALLY, A. V.; ZARANDI, M.; VARGA, J. L.; VIDAURRE, I.; BELLYEI, S. The effect of ghrh antagonists on human glioblastomas and their mechanism of action. **International journal of cancer**, Wiley Online Library, v. 127, n. 10, p. 2313–2322, 2010. Cited on page 77.

RADA-IGLESIAS, A.; BAJPAI, R.; PRESCOTT, S.; BRUGMANN, S. A.; SWIGUT, T.; WYSOCKA, J. Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. **Cell stem cell**, Elsevier, v. 11, n. 5, p. 633–648, 2012. Cited on page 20.

RADA-IGLESIAS, A.; BAJPAI, R.; SWIGUT, T.; BRUGMANN, S. A.; FLYNN, R. A.; WYSOCKA, J. A unique chromatin signature uncovers early developmental enhancers in humans. **Nature**, Nature Publishing Group, v. 470, n. 7333, p. 279–283, 2011. Cited on page 21.

RHODES, D. R.; CHINNAIYAN, A. M. Integrative analysis of the cancer transcriptome. **Nature genetics**, Nature Publishing Group, v. 37, p. S31–S37, 2005. Cited on page 18.

RISSO, D.; SCHWARTZ, K.; SHERLOCK, G.; DUDOIT, S. Gc-content normalization for rna-seq data. **BMC bioinformatics**, BioMed Central Ltd, v. 12, n. 1, p. 480, 2011. Cited on page 34.

RITCHIE, M. E.; PHIPSON, B.; WU, D.; HU, Y.; LAW, C. W.; SHI, W.; SMYTH, G. K. limma powers differential expression analyses for rna-sequencing and microarray studies. **Nucleic acids research**, Oxford Univ Press, p. gkv007, 2015. Cited on page 34.

RIVERA, C. M.; REN, B. Mapping human epigenomes. **Cell**, Elsevier, v. 155, n. 1, p. 39–55, 2013. Cited on page 20.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edger: a bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, Oxford Univ Press, v. 26, n. 1, p. 139–140, 2010. Cited 2 times on pages 34 and 35.

ROBINSON, M. D.; SMYTH, G. K. Small-sample estimation of negative binomial dispersion, with applications to sage data. **Biostatistics**, Oxford University Press, v. 9, n. 2, p. 321–332, 2007. Cited on page 35.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987. Cited on page 30.

SAMUR, M. K. Rtcgatoolbox: A new tool for exporting tcga firehose data. 2014. Cited on page 35.

SAMUR, M. K.; YAN, Z.; WANG, X.; CAO, Q.; MUNSHI, N. C.; LI, C.; SHAH, P. K. canevolve: a web portal for integrative oncogenomics. **PLoS One**, Public Library of Science, v. 8, n. 2, p. e56228, 2013. Cited on page 35.

SHAM, P. C.; PURCELL, S. M. Statistical power and significance testing in large-scale genetic studies. **Nature reviews. Genetics**, Nature Publishing Group, v. 15, n. 5, p. 335, 2014. Cited 3 times on pages 22, 25, and 49.

SHEPHERD, J. H.; URAY, I. P.; MAZUMDAR, A.; TSIMELZON, A.; SAVAGE, M.; HILSENBECK, S. G.; BROWN, P. H. The sox11 transcription factor is a critical regulator of basal-like breast cancer growth, invasion, and basal-like gene expression. **Oncotarget**, Impact Journals, LLC, v. 7, n. 11, p. 13106, 2016. Cited on page 74.

ŠIDÁK, Z. Rectangular confidence regions for the means of multivariate normal distributions. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 62, n. 318, p. 626–633, 1967. Cited on page 24.

SIG, M. **MultiAssayExperiment: Software for the integration of multi-omics experiments in Bioconductor**. [S.l.], 2017. R package version 1.2.0. Available at: https://github.com/waldronlab/MultiAssayExperiment/wiki/MultiAssayExperiment-API. Cited on page 46.

SILVA, T.; COLAPRICO, A.; OLSEN, C.; D'ANGELO, F.; BONTEMPI, G.; CECCARELLI, M.; NOUSHMEHR, H. TCGA workflow: Analyze cancer genomics and epigenomics data using bioconductor packages [version 2; referees: 1 approved, 1 approved with reservations]. **F1000Research**, v. 5, n. 1542, 2016. Cited 2 times on pages 52 and 85.

SILVA, T. C.; COETZEE, S. G.; YAO, L.; HAZELETT, D. J.; NOUSHMEHR, H.; BERMAN, B. P. Enhancer linking by methylation/expression relationships with the r package elmer version 2. **bioRxiv**, Cold Spring Harbor Labs Journals, 2017. Available at: http://www.biorxiv.org/content/early/2017/06/11/148726. Cited 2 times on pages 38 and 40.

SILVA, T. C.; COETZEE, S. G.; YAO, L.; HAZELETT, D. J.; NOUSHMEHR, H.; BERMAN, B. P. Enhancer linking by methylation/expression relationships with the r package elmer version 2. **bioRxiv**, Cold Spring Harbor Labs Journals, p. 148726, 2017. Cited on page 85.

SILVA, T. C.; COLAPRICO, A.; OLSEN, C.; D'ANGELO, F.; BONTEMPI, G.; CECCARELLI, M.; NOUSHMEHR, H. Tcga workflow: Analyze cancer genomics and epigenomics data using bioconductor packages. **F1000Research**, v. 5, 2016. Cited on page 44.

SILVA, T. C.; COLAPRICO, A.; OLSEN, C.; BONTEMPI, G.; CECCARELLI, M.; BERMAN, B. P.; NOUSHMEHR, H. Tcgabiolinksgui: A graphical user interface to analyze gdc cancer molecular and clinical data. **bioRxiv**, Cold Spring Harbor Labs Journals, 2017. Available at: http://www.biorxiv.org/content/early/2017/08/17/147496. Cited on page 85.

SINKKONEN, L.; HUGENSCHMIDT, T.; BERNINGER, P.; GAIDATZIS, D.; MOHN, F.; ARTUS-REVEL, C. G.; ZAVOLAN, M.; SVOBODA, P.; FILIPOWICZ, W. Microrna control de novo dna methylation through regulation of transcriptional repressors in mouse embryonic stem cells. **Nature structural & molecular biology**, Nature Publishing Group, v. 15, n. 3, p. 259–267, 2008. Cited on page 8.

SOKAL, R. R. A statistical method for evaluating systematic relationship. **University of Kansas science bulletin**, v. 28, p. 1409–1438, 1958. Cited on page 29.

SONEGO, P.; KOCSOR, A.; PONGOR, S. Roc analysis: applications to the classification of biological sequences and 3d structures. **Briefings in bioinformatics**, Oxford University Press, v. 9, n. 3, p. 198–209, 2008. Cited on page 34.

SØRLIE, T.; PEROU, C. M.; TIBSHIRANI, R.; AAS, T.; GEISLER, S.; JOHNSEN, H.; HASTIE, T.; EISEN, M. B.; RIJN, M. V. D.; JEFFREY, S. S. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 98, n. 19, p. 10869–10874, 2001. Cited on page 72.

STEINLEY, D. Local optima in k-means clustering: what you don't know may hurt you. **Psychological methods**, American Psychological Association, v. 8, n. 3, p. 294, 2003. Cited on page 30.

SUBRAMANIAN, A.; TAMAYO, P.; MOOTHA, V. K.; MUKHERJEE, S.; EBERT, B. L.; GILLETTE, M. A.; PAULOVICH, A.; POMEROY, S. L.; GOLUB, T. R.; LANDER, E. S. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 102, n. 43, p. 15545–15550, 2005. Cited on page 19.

THINGHOLM, L. B.; ANDERSEN, L.; MAKALIC, E.; SOUTHEY, M. C.; THOMASSEN, M.; HANSEN, L. L. Strategies for integrated analysis of genetic, epigenetic, and gene expression variation in cancer: Addressing the challenges. **Frontiers in genetics**, Frontiers Media SA, v. 7, 2016. Cited on page 18.

THURMAN, R. E.; RYNES, E.; HUMBERT, R.; VIERSTRA, J.; MAURANO, M. T.; HAUGEN, E.; SHEFFIELD, N. C.; STERGACHIS, A. B.; WANG, H.; VERNOT, B. *et al.* The accessible chromatin landscape of the human genome. **Nature**, NIH Public Access, v. 489, n. 7414, p. 75, 2012. Cited on page 21.

TOMCZAK, K.; CZERWINSKA, P.; WIZNEROWICZ, M. *et al.* The cancer genome atlas (tcga): an immeasurable source of knowledge. **Contemp Oncol (Pozn)**, v. 19, n. 1A, p. A68–A77, 2015. Cited on page 45.

TURAGA, N.; FREEBERG, M.; BAKER, D.; CHILTON, J.; NULL, n.; NEKRUTENKO, A.; TAYLOR, J. A guide and best practices for r/bioconductor tool integration in galaxy [version 1; referees: 1 approved, 1 approved with reservations]. **F1000Research**, v. 5, n. 2757, 2016. Cited on page 43.

TURATSINZE, J.-V.; THOMAS-CHOLLIER, M.; DEFRANCE, M.; HELDEN, J. V. Using rsat to scan genome sequences for transcription factor binding sites and cis-regulatory modules. **Nature protocols**, Nature Publishing Group, v. 3, n. 10, p. 1578, 2008. Cited on page 20.

VERLEYSEN, M.; FRANÇOIS, D. The curse of dimensionality in data mining and time series prediction. In: SPRINGER. **IWANN**. [S.l.], 2005. v. 5, p. 758–770. Cited on page 30.

VICKERS, A. J. Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. **BMC medical research methodology**, BioMed Central, v. 5, n. 1, p. 35, 2005. Cited on page 25.

VINAYAGAM, A.; ZIRIN, J.; ROESEL, C.; HU, Y.; YILMAZEL, B.; SAMSONOVA, A. A.; NEUMÜLLER, R. A.; MOHR, S. E.; PERRIMON, N. Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. **Nature methods**, Nature Research, v. 11, n. 1, p. 94–99, 2014. Cited on page 19.

WANG, Z.; ZHANG, S.; SIU, T. L.; HUANG, S. Glioblastoma multiforme formation and emt: role of foxm1 transcription factor. **Current pharmaceutical design**, Bentham Science Publishers, v. 21, n. 10, p. 1268–1271, 2015. Cited on page 77.

WASSERSTEIN, R. L.; LAZAR, N. A. **The ASA's statement on p-values: context, process, and purpose**. [S.l.]: Taylor & Francis, 2016. Cited on page 23.

WEINSTEIN, J. N.; COLLISSON, E. A.; MILLS, G. B.; SHAW, K. R. M.; OZENBERGER, B. A.; ELLROTT, K.; SHMULEVICH, I.; SANDER, C.; STUART, J. M.; NETWORK, C. G. A. R. *et al.* The cancer genome atlas pan-cancer analysis project. **Nature genetics**, Nature Publishing Group, v. 45, n. 10, p. 1113–1120, 2013. Cited on page 13.

WHITLEY, E.; BALL, J. Statistics review 6: Nonparametric methods. **Critical care**, BioMed Central, v. 6, n. 6, p. 509, 2002. Cited on page 25.

WILKERSON, M. D.; HAYES, D. N. Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. **Bioinformatics**, Oxford Univ Press, v. 26, n. 12, p. 1572–1573, 2010. Cited 2 times on pages 35 and 53.

WILKS, C.; CLINE, M. S.; WEILER, E.; DIEHKANS, M.; CRAFT, B.; MARTIN, C.; MURPHY, D.; PIERCE, H.; BLACK, J.; NELSON, D. *et al.* The cancer genomics hub (cghub): overcoming cancer through the power of torrential data. **Database**, Oxford University Press, v. 2014, p. bau093, 2014. Cited 2 times on pages 13 and 32.

WINGENDER, E.; SCHOEPS, T.; DÖNITZ, J. Tfclass: an expandable hierarchical classification of human transcription factors. **Nucleic acids research**, Oxford Univ Press, v. 41, n. D1, p. D165–D170, 2013. Cited 2 times on pages 51 and 52.

WINGENDER, E.; SCHOEPS, T.; HAUBROCK, M.; DÖNITZ, J. Tfclass: a classification of human transcription factors and their rodent orthologs. **Nucleic Acids Research**, v. 43, n. D1, p. D97–D102, 2015. Available at: +http://dx.doi.org/10.1093/nar/gku1064. Cited on page 76.

XIA, L.; HUANG, Q.; NIE, D.; SHI, J.; GONG, M.; WU, B.; GONG, P.; ZHAO, L.; ZUO, H.; JU, S. *et al.* Pax3 is overexpressed in human glioblastomas and critically regulates the tumorigenicity of glioma cells. **Brain research**, Elsevier, v. 1521, p. 68–78, 2013. Cited on page 77.

XIE, W.; BARR, C. L.; KIM, A.; YUE, F.; LEE, A. Y.; EUBANKS, J.; DEMPSTER, E. L.; REN, B. Base-resolution analyses of sequence and parent-of-origin dependent dna methylation in the mouse genome. **Cell**, Elsevier, v. 148, n. 4, p. 816–831, 2012. Cited on page 7.

YAO, L.; SHEN, H.; LAIRD, P.; FARNHAM, P.; BERMAN, B. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. **Genome biology**, v. 16, n. 1, p. 105–105, 2015. Cited 7 times on pages 31, 38, 40, 44, 48, 55, and 72.

YATES, A.; AKANNI, W.; AMODE, M. R.; BARRELL, D.; BILLIS, K.; CARVALHO-SILVA, D.; CUMMINS, C.; CLAPHAM, P.; FITZGERALD, S.; GIL, L. *et al.* Ensembl 2016. **Nucleic acids research**, Oxford Univ Press, p. gkv1157, 2015. Cited 3 times on pages 46, 49, and 52.

YAU, C. R tutorial with bayesian statistics using openbugs. **URL http://www. r-tutor. com/-conten 'dr-tutorial-ebook. Indice de instrucciones**, 2012. Cited on page 28.

YERSAL, O.; BARUTCA, S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. **World journal of clinical oncology**, Baishideng Publishing Group Inc, v. 5, n. 3, p. 412, 2014. Cited on page 72.

ZHANG, Y.; JURKOWSKA, R.; SOEROES, S.; RAJAVELU, A.; DHAYALAN, A.; BOCK, I.; RATHERT, P.; BRANDT, O.; REINHARDT, R.; FISCHLE, W. *et al.* Chromatin methylation activity of dnmt3a and dnmt3a/3l is guided by interaction of the add domain with the histone h3 tail. **Nucleic acids research**, Oxford University Press, v. 38, n. 13, p. 4246–4253, 2010. Cited on page 8.

ZHOU, W.; LAIRD, P. W.; SHEN, H. Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes. **Nucleic Acids Research**, Oxford Univ Press, p. gkw967, 2016. Cited 2 times on pages 15 and 46.

_____. ____. **Nucleic Acids Research**, v. 45, n. 4, p. e22, 2017. Available at: +http://dx.doi.org/10.1093/nar/gkw967. Cited 3 times on pages 47, 48, and 52.

ZHU, Y.; QIU, P.; JI, Y. Tcga-assembler: open-source software for retrieving and processing tcga data. **Nature methods**, Nature Publishing Group, v. 11, n. 6, p. 599–600, 2014. Cited on page 35.

APPENDIX

A

DISPENSA COMITÊ DE ÉTICA

Faculdade de Medicina de Ribeirão Preto
Universidade de São Paulo

Ilma. Senhora Profa. Dra. Marcia Guimarães Villanova,
(Coordenadora do Comitê de Ética em Pesquisa do HCFMRP-USP)

Venho, por meio desta, solicitar dispensa de apreciação ética do projeto de pesquisa de Doutorado intitulado *"Bioinformatic tool to integrate and understand aberrant epigenomic and genomic changes associated with cancer:Methods, development and analysis"* ao Comitê de Ética em Pesquisa do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (HCFMRP-USP). A solicitação de dispensa justifica-se uma vez que esta pesquisa consiste no desenvolvimento de uma ferramenta de bioinformática para a análise dados genômicos e epigenômicos e utiliza somente dados genômicos de amostras humanas que estão publicamente disponíveis à comunidade científica no banco de dados internacional *The Cancer Genome Atlas* ou TCGA (disponível em https://tcga-data.nci.nih.gov/tcga/), com livre acesso a todos os usuários para baixar e analisar os dados sem necessidade de cadastro ou identificação prévios. Este projeto de Doutorado está sendo desenvolvido no Departamento de Genética da FMRP-USP pelo aluno de Pós-Graduação Tiago Chedraoui Silva e sob orientação do Professor da Faculdade de Medicina de Ribeirão Preto Dr. Houtan Noushmehr. As amostras estão de acordo com a ética e a legislação norte-americana vigente, tendo em vista que o consórcio TCGA, responsável pelo banco de dados, mantém em sigilo todas as informações que possam levar à identificação dos pacientes.

Ribeirão Preto, 11 de maio de 2016.

Tiago Chedraoui Silva

Prof. Dr. Houtan Noushmehr

COMITÊ DE ÉTICA EM PESQUISA - HCRP
15:25 h.
Recebi
11/05/16

Mas há necessidade de submissão ao Comitê de Ética em Pesquisa 23/05/16

DRª MARCIA GUIMARÃES VILLANOVA
Coordenadora do Comitê de Ética em Pesquisa
do HCFMRP-USP e da FMRP-USP