UNIVERSIDAD DE GRANADA

INTEGRATION OF HETEROGENEOUS GENE EXPRESSION SOURCES IN HUMAN CANCER PATHOLOGIES, EMPLOYING HIGH PERFORMANCE COMPUTING AND MACHINE LEARNING TECHNIQUES

Doctoral Thesis submitted by

DANIEL CASTILLO SECILLA

To obtain the International Ph.D. degree as part of the

PROGRAMA DE DOCTORADO EN TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN

Supervisors

IGNACIO ROJAS RUIZ
LUIS JAVIER HERRERA MALDONADO

January 20, 2020

Editor: Universidad de Granada. Tesis Doctorales
Autor: Daniel Castillo Secilla
ISBN: 978-84-1306-459-8
URI: http://hdl.handle.net/10481/60017

En honor a mis abuelos

In honor of my grandparents

«Your time is limited, so don't waste it living someone else's life. Don't be trapped by dogma, which is living with the results of other people's thinking. Don't let the noise of others opinions drown out your own inner voice. And most important, have the courage to follow your heart and intuition. They somehow already know what you truly want to become. Everything else is secondary. Stay Hungry. Stay Foolish.»

— Steve Jobs

ACKNOWLEDGEMENTS

Si quieres ir rápido, camina sólo. Si quieres ir lejos, ve acompañado. Esta frase me ha identificado durante el trascurso de mi camino desde que tengo uso de razón. A decir verdad, yo no habría llegado hastá aquí sin la ayuda inconmensurable de mi familia, así como la voluntad de creer en los avances médicos de un hombre. El Doctor José Ayala Montoro, a quién durante toda mi vida le agradeceré haberme dado la mínima posibilidad de poder caminar por mí mismo, cuando otros no lo hicieron. Gracias.

No obstante, nada de esto habría sido posible sin unos padres que, mas hallá de rendirse, decidieron buscar segundas y terceras y cuartas opiniones, hasta que dieron con una sola que arrojaba algo de esperanza. Mamá, Papá, gracias por luchar para traerme hasta donde estoy hoy, yo solo he puesto mis ideas, el resto del mérito es todo vuestro. Nunca podré agradeceros lo suficiente lo que habéis luchado, y aún luchais, por mí.

Tampoco puedo olvidar el amor de mis abuelos y el apoyo que me dieron, así como toda su sabiduría. Este logro en mi vida, uno de los más altos que una persona puede obtener, os lo dedico a vosotros. Abuelo Ricardo, Abuela Rosa, Abuela Juanita, os tengo cada día presentes y guiais cada paso que he dado y daré. Abuelo Boni, como siempre me dices, ***Mirada al frente, paso corto y mala leche***, una frase que aunque graciosa, esconde una actitud de vida necesaria para llegar a donde uno se propone. Gracias a los cuatro por vuestras lecciones de vida.

Aún recuerdo cuando decidí estudiar Ingeniería Informática después de ver lo que mi hermano Jose era capaz de hacer y el mundo de posibilidades que esta disciplina albergaba. El verano antés de llegar a la carrera, me enseñaste los pasos para programar una calculadora muy básica en C que me sirvió para no ir a ciegas mi primer año. A partir de ahí te convertiste en un modelo a seguir, fuiste mi director de TFG, hicé el master en Granada guiado por tí y, gracias a ello, me convertiré finalmente en Doctor, al igual que tú. Sin duda alguna, he seguido tus pasos y consejos y gracias a ello ya somos dos doctores en la familia. Gracias por tanto "Tate".

A mi hermano Jesús tengo que agradecerle siempre su gran corazón, sus charlas de ánimo y su predisposición a ayudar cuando haga falta a lo que haga falta. Ojalá hubiera más personas como tú. Me has demostrado después de 10 años de lucha, que los objetivos se alcanzan tarde o temprano y que todo depende de lo mucho que desees llegar a la meta. Mas que gracias tengo que decirte Enhorabuena Agente Castillo.

Después 4 años y pico conviviendo contigo, y algunos más de compañeros de carrera y de frikismos, prácticamente eres como otro hermano para mí Juan Carlos Gómez López. Han sido mucho los momentos malos y buenos que hemos tenido estos años, pero siempre nos hemos terminado riendo, y mucho, de las cientos de tonterías que se nos ocurren a diario. Ahora te cedo el relevo, te toca a tí afrontar este camino largo y bonito que supone empezar una tesis doctoral, ánimo que llegarás al final sin lugar a dudas. Aunque dentro de relativamente poco tengamos que tomar cáminos separados seguramente, seguiremos muy unidos aunque sea a través de nuestras partidas interminables a la Xbox. Dale caña KosmoBox.

Tampoco me puedo olvidar de tí, José Manuel Martín Carrión. Desde los 4 años que nos conocemos ya ha llovido bastante y, aun así, aquí seguimos con una hermandad que ya es de por vida y que muchas personas querrían. Me alegro mucho que hallas conseguido al fin tu ansiado título de Ingeniero Electrónico, se que las cosas te van a ir genial a partir de ahora y espero estar siempre ahí para verlo con mis propios ojos. Ah! Y por supuesto, gracias por la imagen de portada tan bonita que me has hecho para esta tesis. A seguir comiendote el mundo!

Durante estos bonitos años he tenido la oportunidad de tratar con maravillosos compañeros, con los cuales espero no perder el contacto nunca. Gracias también a todos vosotros. Al único e inimitable Dr. Antonio Fernández Ares, por las charlas videojueguiles en el gimnasio, y por todos los consejos que me has dado a lo largo de estos años relacionados con mi tesis. Al Dr. Juan José Escobar Pérez, porque detrás de esa apariencia dura, se esconde una de las mejores personas que he conocido y con mejor corazón. He disfrutado y me he reído mucho contigo en los congresos y cursos a los que hemos asistido juntos. Por último, al Dr. Juan Manuel Gálvez Gómez, sin el cual esta tesis habría sido mucho más difícil de sacar adelante. Hemos sido y somos compañeros de batalla desde que empezamos los dos prácticamente a la vez con esto de la Bioinfomática. Estoy seguro de que a ambos nos depara un gran futuro en este campo, algún día conseguiremos ese Nature! Seguimos trabajando compi.

Casí para terminar ya he de agradecer a mis dos directores la oportunidad que me dieron de encontrar un campo tan bonito y con tanta proyección como la Bioinformática. Pensar en usar la tecnología para intentar avanzar en la lucha contra el cáncer era algo apasionante para mí y, a partir de ahora puedo llevarlo a la realidad. Por ello, Gracias Dr. Ignacio Rojas Ruiz, porque de tí he aprendido cualidades que me han hecho crecer como persona y como profesional y por ello espero que podamos seguir trabajando codo con codo en el futuro. Eres un gran Director y lo has demostrado durante todos estos años. Y gracias también a tí, Dr. Luis Javier Herrera Maldonado, por la cantidad de horas y revisiones que has echado conmigo sin desesperarte, incluso por videollamadas en verano en tus vacaciones. Hemos sacado trabajos muy interesantes juntos y he aprendido infinidad de cosas de tí que me han hecho mejorar como persona y como investigador. Releer, darle vueltas a la escritura, dejarlo bonito y si aun así nada, a borrar y escribirlo de cero. Me ha costado entender ese proceso, pero viendo los resultados que da no dudes que lo seguiré usando y mejorando el resto de mi carrera. Gracias a ambos, no podría haber contado con mejores directores.

Y por último, a la persona más importante, aquella que me ha alegrado la vida desde que está a mi lado y con la que he compartido el camino hacia este logro. Inmaculada Rosa Estepa te lo digo siempre que puedo y lo seguiré haciendo, te quiero y me has hecho entender que significa compartir un proyecto de vida y un futuro en común con una persona. Me has apoyado de todas las formas posibles y me has levantado el ánimo en los momentos de bajón que he tenido en el camino. Juntos iremos consiguiendo grandes cosas y esto solo es un paso más para ello. Ahora te toca a tí darlo todo para lograr tu meta y yo estaré ahí para ser tu apoyo en los malos momentos y para disfrutar juntos de los buenos. De eso se trata esto del amor, apoyarnos incondicionalmente sin importar lo que venga. Te amo cariño.

ACRONYMS

A

ACC	Adenocarcinoma
AJCC	American Joint Committee on Cancer
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
ANOVA	ANalysis Of VAriance

B

BAM	Binary Alignment Map
BP	Base pair
BPr	Biological Process

C

CC	Cellular Component
cDNA	complementary DNA
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
CNV	Copy Number Variation
COV	Coverage
CPUs	Central Processing Units
CSV	Comma-Separated Values
CUDA	Compute Unified Device Architecture
CV	Cross-Validation

D

DA	Daltons
DA-FS	Disease Association Feature Selection
DCIS	Ductal Carcinoma In Situ
DDBB	Databases
DEGs	Differentially Expressed Genes
DEPs	Differentially Expressed Proteins
DNA	Deoxyribonucleic acid
dNTPs	deoxyribose Nucleoside Triphosphates
dsDNA	double-stranded DNA

F

FDR	False Discovery Rate

FPGAs	Field-Programmable Gate Arrays
FS	Feature Selection

G

GB	Gigabyte
Gbps	Gigabit per second
GDC	Genomic Data Commons
Ghz	Gigahertz
GOs	Gene Ontologies
GPUs	Graphic Processing Units
GRCh	Genome Reference Consortium human
GTF	Gene Transfer Format
GWAS	Genome-wide association study

H

HPC	High Performance Computing
HTS	High Throughput Sequencing
HUGO	HUman Genome Organization

I

INDELS	Insertion and Deletions

K

k-NN	k-Nearest Neighbour
KEGG	Kyoto Encyclopedia of Genes and Genomes

L

LCLC	Large Cell Lung Carcinoma
LFC	Log-Fold Change

M

MAP	MinION Access Program
MF	Molecular Function
MI	Mutual Information
ML	Machine Learning
mRMR	minimum Redundancy Maximum Relevance

N

NAS	Network Attached Storage
NB	Naive Bayes

NCBI/GEO National Center for Biotechnology Information / Gene Expression Omnibus
NFS Network File System
NGS Next Generation Sequencing
NSCLC Non Small Cell Lung Carcinoma

O

ONT Oxford Nanopore Technology
OS Operative System

P

PCR Polymerase Chain Reaction
PPI Protein-Protein Interactions
PTP PicoTiter Plate

R

RAM Random Access Memory
RF Random Forest
RMA Robust Multi-Array Average
RNA Ribonucleic acid
RUV Remove Unwanted Variation

S

SAS Serial Attached SCSI
SBS Sequencing By Synthesis
SCC Squamous Cell Cancer
SCLC Small Cell Lung Cancer
SCSI Small Computer System Interface
SMRT Single Molecule Real Time
SNPs Single Nucleotide Polymorphisms
SRA Sequence Read Archive
SSD Solid-State Drive
ssDNA single-stranded DNA
SVA Surrogate Variable Analysis
SVM Support Vector Machine

T

TB Terabyte
TCGA The Cancer Genome Atlas
TNM Tumor, Node, Metastasis
TSV Tab-Separated Values

U

USB Universal Serial Bus

W

WGAS Whole Genome Association Study
WGS Whole Genome Sequencing
WHO World Health Organisation

Z

ZMWs Zeromode Waveguides

ABSTRACT

Battle against cancer has arisen as one of the main challenges for humanity. This is largely due to the annual increase of the amount of people who suffer from any type of cancer. The growth in the life expectancy, the unhealthy lifestyle or the pollution, are possible factors for this increase. For that, scientists and researchers are focused on the study and comprehension of the development of this genetic disease. The treatment and analysis of biological data coming from different omics sources are helping to address the study of cancer from different perspectives, with the purpose of achieving an early diagnosis and increase the life expectancy and the survival rate. Moreover, the consolidation and the cost reduction of Next Generation Sequencing technologies and platforms has lead to a notable increase in the precision, quality and quantity of the omics studies and the available data. In addition to this, thanks to the use of machine learning techniques applied to the study and evaluation of omics data, the search of groups of relevant biomarkers or possible gene signatures is tackled in forms impossible until now, due to the dimensionality of the problem.

On this basis, the main objective of this thesis is the search of relevant biomarkers at gene expression level, by using the integration of heterogeneous transcriptomic sources for different cancer pathologies. To carry out this search, heterogeneous public data from different databases have been gathered in order to find relevant biomarkers. Furthermore, through the use of advanced feature selection and machine learning techniques, relevant biomarkers are evaluated with the aim of discovering their potential to discern the state of a patient who suffer from cancer. All of this accompanied by a biological enrichment of the relevant genes for each case of study, making use of the literature. As culmination of this thesis, the design and implementation of a novel and public tool named as KnowSeq has been carried out. KnowSeq was designed with the purpose of yielding to the expert in bioinformatic and computational biology scope, an automatic tool to perform complete gene expression analyses in an easy and flexible way. The tool also counts with an advanced machine learning evaluation process, as well as an automatic biological enrichment for the final expressed genes.

RESUMEN

La batalla contra el cáncer se ha establecido como uno de los principales retos de la humanidad. Esto es debido al aumento año tras año del número de personas que padecen algún tipo concreto de cáncer. El aumento de la esperanza de vida, los malos hábitos de vida o la contaminación son factores ha tener en cuenta en este crecimiento. Por ello, la comunidad científica e investigadora tiene en uno de sus puntos de mira el estudio y comprensión del desarrollo de esta enfermedad multifactorial. El tratamiento y análisis de datos biológicos provenientes de las diferentes ómicas existentes ayuda a abordar el estudio del cáncer desde diferentes perspectivas, para así tratar de buscar nuevas formas de diagnóstico precoz y aumentar la esperanza de vida y supervivencia de los pacientes. Además, con la implantación y abaratamiento de las tecnologías y plataformas Next Generation Sequencing, la precisión, calidad y cantidad de los estudios se ha incrementado notablemente, permitiendo paulatinamente el avance de la sociedad hacía la medicina personalizada o de precisión. A todo esto se le añade el uso de técnicas de aprendizaje automático aplicadas al estudio y evaluación de datos ómicos, el cual ha permitido llevar a cabo la búsqueda de grupos de biomarcadores o posibles huellas génicas que antaño eran inviables por la dimensionalidad del problema.

Bajo estas premisas, el objetivo principal de esta tesis es la búsqueda de biomarcadores a nivel de expresión de gen, mediante la integración de fuentes heterogéneas de datos transcriptómicos para diferentes patologías de cáncer. Para llevar a cabo dicha búsqueda, se han recolectado datos públicos y heterogéneos de diferentes Bases de Datos para realizar su integración y análisis de expresión diferencial en busca de biomarcadores relevantes. Además, mediante el uso de técnicas avanzadas de selección de características y aprendizaje automático, dichos biomarcadores son evaluados con el fin de saber su potencial a la hora de discernir el estado de un paciente. Todo ello, acompañado de un estudio biológico a nivel de literatura del conjunto final de genes destacados en cada caso. Como colofón de esta tesis, se ha llevado a cabo el diseño e implementación de una herramienta actualmente pública en el lenguaje R llamada KnowSeq. Dicha herramienta se diseñó con el fin de brindar a los expertos en el ámbito de la bioinformática una manera de automatizar, bajo un solo paquete software, todos los procesos implicados en los análisis de expresión de gen.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

BACKGROUND & STATE OF THE ART

1 INTRODUCTION

CONTENTS

Nowadays, cancer is one of the most deadly worldwide diseases. For this reason, the number of biological data available to carry out analyses is growing exponentially. As a result, the number of studies and researches that try to find relevant biomarkers is gradually increasing too. With this enormous quantity of biological data, it is clear the necessity of using Machine Learning techniques to analyse and process them, and extract information and knowledge from them. In particular, the application of different Machine Learning methods can assist in finding hidden relations and relevant biomarkers never seen before.

However, the computational cost to process great amounts of genomic data can be considerable. To deal with it, it is highly recommendable to use optimization techniques and specific parallelization hardware such as computer clusters, GPUs and FPGAs.

GPUs: Graphic Processing Units
FPGAs: Field-Programmable Gate Arrays

It is important to highlight that this trend continues, and that the amount of biological data will keep on growing in the near future. In view of these circumstances, the future of the bioinformatics and computational biology scope is moving forward with the application of Machine Learning and high performance computing approaches.

1.1 CONTEXT AND MOTIVATION

For several decades, cancer has been one of the most studied diseases due to its high mortality rate. Although thanks to the research and medical advances, nowadays the disease has the best survival rate ever, there does not exist a real cure for cancer yet. For that, cancer still holds the second position in the most deadly diseases worldwide ranking with an estimated number of death in 2018 of 9.6 million people, only behind cardiovascular diseases [1]. The only possibility to significantly increase the cancer survival rate is achieving early diagnosis. However, the diagnosis is usually done when the patients present symptoms and, in many cases, it is too late. On the other hand, there are cancer types such as pancreatic cancer that usually is only diagnosed when the cancer is in a very advanced stage, and practically without any chance of curing it. For these reasons, it is primordial to research on new methods to find biomarkers that allow achieving those diagnostics in an early stage or, even before the emergence of the disease.

WGS: Whole Genome Sequencing

The amount of WGS generated data is massively growing due to several sequencing programs existing around the world, such as the 100K genomes in UK [2], the European 1+ Million Genomes Initiative [3] and All Of Us in America [4], among others. Because of that, the researchers have access to the largest number of samples ever. Furthermore, thanks to the use of Machine Learning techniques, unknown behaviours of the genome are now able to be detected and studied, allowing us to trace a biological profile for each type of cancer.

Nevertheless, due to this WGS massive data availability state, the amount of heterogeneous data sources, understood as data coming from different technologies (Microarray and RNA-Seq) or different platforms or manufacturers (Affymetrix, Illumina, etc.), is higher than never before. The availability of massive heterogeneous data sources presents certain challenges. A main one is the massive integration of heterogeneous data sources coming from different omics, or from different technologies inside the same omic, which can help to increase the data availability for a study, and ease the detection of hidden behaviours. In the case of the integration of different omics, a mutation on a set of genes or a genome region can be followed at gene expression and proteomic levels, so the direct effect of those mutations in the rest of biological processes can be observed. On the other hand, the integration of technologies or platforms coming from the same omics, allows achieving a huge number of samples and compensating possible unbalanced classes in order to ensure the robustness and the statistical significance of the studies. ML techniques learn more and better with a

ML: Machine Learning

higher number of samples to train the models. For that, if the classifiers count with more samples, they will obtain more generalised models, thus achieving thus best classification results when the models have to deal with unseen samples.

The increase in the number of WGS data and the advances in the integration of omics information and ML assessment, are opening the door towards the precision or personalised medicine instead of the traditional try and error medicine. The main strength of the precision medicine is the possibility to treat in a personalised way each patient depending on the mutations or gene expression of his genes. This is extremely useful in order to apply concrete drugs to each patient instead of standardised treatments. However, although there are many advances in the field, a proper infrastructure to storage and manage the huge volume of data generated to perform this implementation is still required [5]. Furthermore, it is still necessary to dig deeply in the research to understand all the underlying relations and processes involved in the development of cancer diseases. This is why the study using the integration of different types of omics data, in order to seek hidden knowledge about cancer, is now more important than ever before. These studies are changing the way we understand medicine and, in particular genetic diseases.

1.2 MAIN OBJECTIVES

Taking into account the main motivations presented in Section 1.1, and the current state of the art, which is later presented in Chapters 2 and 3 (which introduce the Biological and ML backgrounds of the works presented in this thesis), the main objective of this thesis is to make relevant contributions in the integration of heterogeneous transcriptomic data and subsequent differential expressed biomarkers analysis and assessment, through the application of feature selection and predictive ML models. The specific main objectives are detailed below.

1.2.1 *Integration and analysis of heterogeneous data*

The first specific objective is to design and propose and automatic pipeline to carry out the integration of data coming from heterogeneous transcriptomic technologies and platforms. Once the integration is

successfully performed, the pipeline follows carrying out a quality analysis over the integrated dataset. Finally, the appropriate relevant biomarkers extraction is performed. With that in mind, an in depth explanation of each of the steps above is given hereunder:

1. Before integrating all the samples, it is highly recommended to carry out a strong quality analysis and batch effect removal process for each series or dataset involved in the integration. The influence of outliers can introduce some noise and unwanted deviation in the results, thus they must be removed. On the other hands, in order to correct possible intrinsic deviations of the samples, it is desirable to treat the batch effect to ensure the best harmonization among the series and samples.

2. There exist many platforms that belong to different sequencing technologies. It is very interesting to take advantages of all of them with the purpose of avoiding the lack of samples that normally suffer many genetic diseases in public or controlled DDBB. Moreover, in many cases, an unbalanced problem appears due to the differences at number of available samples among the addressed classes or states. In such cases, it is better to complement the lack of samples with another datasets instead of applying imputation methods to create new samples.

DDBB: Databases

3. At last, if the quality analysis and integration have been performed, the integrated dataset can be employed to extract the DEGs. If all the process is followed in a rigorous and robustness way, the extracted biomarkers would be able to discern the studied cancer or sub-types of cancer.

DEGs: Differentially Expressed Genes

1.2.2 *Data assessment using Machine Learning approaches*

Once the candidates DEGs have been obtained, the second specific objective proposed in this thesis is to study and research the use of different ML algorithms for optimising and assessing biomarker sub-sets for different types of cancer. This was carried out in the the following sub-objectives related to the DEGs assessment and selection, making use of both feature selection and classification ML algorithms:

1. With the aim of achieving a reduced gene signature for the tackled disease, a feature selection step will be performed before the predictive model application. A feature selector has the capability to decide which are the best reduced sub-set of DEGs that will

achieved similar of equal classification results than the complete set of DEGs candidates. For that, the algorithms create a ranking that reorder those biomarkers.

2. To evaluate the feature selection ranking of biomarkers, different proposed classification algorithms will perform the applied, carrying out a CV process. This process will provide assessment in the search of a final sub-set of DEGs that reach optimal classification results.

CV: Cross-Validation

3. With the two previous considerations in mind, the final objective is to perform a test step of the different FS and classification algorithms evaluated. Several samples unseen before in the process will be used to classify them by using the predictive models trained with the reduced sub-set of DEGs. This final phase will provide the expected performance of the selected classification model and the selected sub-set of DEGs candidates, in their capability to discern people who suffer from the addressed disease from healthy people.

FS: Feature Selection

1.2.3 *Novel bioinformatic tool implementation*

Following the main thread of this thesis, the next objective is to encapsulate the complete proposed automatic pipeline implemented in the development of the previous objectives of this thesis, in one tool under the same programmatic language. This objective expects to bring a novel tool to help to the experts in the field to acquire robust knowledge and conclusions for the data and diseases to study. The idea of creating an automatic tool emerges due to the nonexistence of tools that embrace all those functionalities, specially including a complete Machine Learning step, under the same environment.

1.2.4 *Use of High Performance Computing Tools*

The final specific objective of this thesis is to use HPC to distribute and optimise the processes within the pipeline requiring a high computational cost. There exist a number of platforms to carry this out such as computer clusters, GPUs or FPGAs. First, the use of computer clusters to distribute the raw NGS data pre-processing is analysed and performed. Then, the use of GPUs is preliminary approached for

HPC: High Performance Computing

NGS: Next Generation Sequencing

the optimisation of the Machine Learning algorithms parallelization, however not finished at the end of this Ph. D. Finally, although there exists the possibility to take advantages of the potential of FPGAs for the processes optimization, this platforms will not be considered for it in the thesis.

1.3 THESIS STRUCTURE

This section provides a brief description of each of the chapters that make up this thesis with the objective of providing a global view of its structure. The document is divided into three main parts according to the content of the chapters they contain. First two parts comprise different chapters. The final one comprises appendices and bibliography. Therefore, the structure of the thesis is as follows:

PART I. BACKGROUND & STATE OF THE ART

- Chapter 1. Introduction: this chapter has introduced this thesis, presented its motivation and its objectives. An overview of the structure of the document is also provided in this section Section 1.3.

- Chapter 2. Biological Background, a sequencing review: this chapter presents a biological state of the art supporting the biological motivation of this thesis. Concretely, an in depth background about the sequencing technologies history will be provided, together with information about the he different omics available and about the origin of genetic diseases.

- Chapter 3. Predictive models applied to bioinformatic: this chapter, following the line of Chapter 2, presents the second state of the art of this thesis, summarising the most renowned Machine Learning techniques. Furthermore, the justification of the application of this type of techniques to bioinformatics and computational biology will be explained.

PART II. CASE STUDIES & DISCUSSION

- Chapter 4. Methodology: this chapter presents the experimental guidelines followed for the researches included in Chapters 5, 6 and Appendix A. The pipeline addressed for the evaluation of the data, such as the origin of the samples or series used, are also described. The chapter also presents the different Machine Learn-

ing algorithms and their particularities, and the characteristics of the platform and devices on which the codes will be executed.

- Chapter 5. Breast cancer profiling by integrating heterogeneous transcriptomic platforms: this chapter is focused on the integration of different heterogeneous Microarray and RNA-Seq platforms of breast cancer series from public DDBB. Moreover, once the integration was done, a search of relevant biomarker for automatic diagnosis via Machine Learning was implemented. Finally, a thorough study of DEGs candidates was performed to look to their biological and medical relationships with breast cancer [6].

- Chapter 6. Leukemia sub-types diagnosis by applying Machine Learning techniques: this chapter tackles a diagnosis problem with the main types of leukemia involved. For this study, heterogeneous Microarray and RNA-Seq datasets were integrated in order to acquire a significant number of samples available. Furthermore, a set of DEGs with the potential to discern among the different types of leukemia was searched. For that, a new parameter for multiclass genes selection was introduced and then, those DEGs were assessed with Machine Learning techniques. To conclude, clinical researchers provide a study of the final biomarkers to gather information about their relations with the different types of leukemia [7].

- Chapter 7. KnowSeq, Beyond the traditional gene expression pipeline: As a colophon of this doctoral thesis, once the first studies presented in Chapters 5 and 6 were done, the idea of combining all the functionalities implemented for them into one tool emerged. KnowSeq was born as an R package that combines the traditional gene expression steps with a Machine Learning set of steps under the same tool and language [8, 9]. This chapter shows in detail the KnowSeq characteristics, functionalities and possibilities, addressing a real Breast cancer RNA-Seq dataset study case. KnowSeq is already public in the most renowned bioinformatic repository, Bioconductor [10].

- Chapter 8. Conclusions: this chapter briefly presents the conclusions drawn from the results obtained and the contributions of this doctoral thesis. Furthermore, future work is also exposed.

PART III. APPENDICES & BIBLIOGRAPHY

- Appendix A. Impact of feature selection for biomarker detection in multiclass lung cancer: this appendix presents an extension

of an international conference publication, which addresses the assessment of different feature selection methods for multiclass lung cancer data, using microarray data and the KnowSeq R/Bioc software package presented in chapter 7. Moreover, we present a new Biologically-based Feature selection algorithm named as DA-FS that includes in its operation, information about the literature-extracted relation of the DEGs with the studied disease.

DA-FS: Disease Association Feature Selection

- Appendix B. KnowSeq User Documentation: this appendix contains the user documentation to learn all about KnowSeq. For that, all the functions included in KnowSeq are explained, even with example code.

- Appendix C. Publications: this appendix lists the different publications obtained during the course of the thesis. Publications in international journals with impact factor are included, as well as publications in international conferences.

- Appendix D. Grants and Special Acknowledgements: this appendix lists the different grants that support the development of this thesis and associated publications. Furthermore, the special acknowledgements are also included here.

- Bibliography: lists the scientific publications and web links that support the content of this thesis.

2

BIOLOGICAL BACKGROUND: A SEQUENCING REVIEW

CONTENTS

The technological and medical advances taken place over the last decades are allowing significant improvements in cancer detection and treatment. Thanks to that, the survival rates of the different types of cancer are higher nowadays. Furthermore, with the arrival of NGS technologies, the amount of accurate biological data coming from a variety of omics sources has greatly increased. These are publicly available in most of the cases, and have eventually replaced their predecessor: the microarrays. During this chapter, the state of the art regarding the different omics and the advances of the sequencing technologies along

the history will be summarised, setting up the conceptual biological framework that surrounds this thesis. For a more detailed information about the concepts and explanations described along this Chapter, the reader can refer to the original scientist sources [11–14], together with all the references along the text that support this review.

2.1 CANCER OVER THE LAST YEARS

Cancer is derived from the Greek word "karkinos", that is synonym of malignant tumour and malignant neoplasm among others, and it makes reference to a set of multifactorial diseases with genetic predisposition that share the same nature. Cancer was recognised and typified by Hypocrates (460-370 BC) and Galen (129-210 AD) [15]. All of the existing types of cancer are produced by several factors that leads to an uncontrollable cell division that invades nearby tissues and organs. The cells life cycle contains a growth step and division step to create new cells when the body needs them. When these cells grow old or are damaged, they die and are replaced by healthy cells. When a person suffers from cancer, this process goes uncontrollably and the old and damaged cells are kept while new cells are created, forming benign or malign tumours. These cancerous cells have the capability to ignore the biological signals that indicate that they have to stop their division and start their programmed cell death, also known by the term apoptosis [16]. These signals are the methods employed to remove unnecessary or useless cells, avoiding possible disorders caused by them.

As was mentioned before, cancer is a multifactorial disease caused by a combination of genetic and environmental factors working together in ways that are not yet fully understood. These alterations can be hereditary or can be produced by DNA damages caused by several factors. In most cases, the cells can correct those DNA damages but, otherwise, they can lead to the development of a genetic disease or a multifactorial disease such as cancer.

DNA: Deoxyribonucleic acid

Depending on the origin of the tissue of a tumour, it can be defined with different names (carcinoma, sarcoma, melanoma, lymphoma, among others related terms). Furthermore, even within a concrete type of cancer, there are many sub-groups that identify the aggressiveness of this cancer and its type of treatment.

Staging helps to know where a cancer is located, or where it has spread, and whether it is affecting other parts of the body. For many

types of cancer, clinicians commonly use the TNM system of the AJCC to describe a cancer's stage. TNM system is based on a set of four questions that are answered by using diagnostic tests, imaging scans, and surgery to remove or get a sample of the tumor [17].

TNM: Tumor, Node, Metastasis
AJCC: American Joint Committee on Cancer

- How large is the primary tumor? Where is it located? (Tumor, T)
- Has the tumor spread to the lymph nodes? If so, where and how many? (Node, N)
- Has the cancer spread to other parts of the body? If so, where and how much? (Metastasis, M)
- Are there any biomarkers or tumor markers linked to the cancer that may make it more or less likely to spread?

By combining the results raised by TNM system, clinicians can determine the stage of a cancer. Normally, all the cancers have four stages (I, II, III and IV) but, some of them have also a stage 0. Depending on the stage, the possibility to cure the disease and the survival rate varies widely. A brief description for each stage is given herein:

- Stage 0: This stage makes reference to cancer "in situ" or in the place where it started, because the cancer has not spread to nearby tissues. Stage 0 is often highly curable, by removing in many cases the whole tumour through surgery.
- Stage I: This stage represents a small cancer or tumour that has not grown deeply into nearby tissues. Furthermore, it has not also spread to the lymph nodes or other parts of the body. It is usually known as early-stage cancer.
- Stage II and III: Both stages indicate larger tumours that have grown more deeply into nearby tissue. They may have also reach lymph nodes but not other parts of the body.
- Stage IV: This stage means that the cancer has spread to other organs or parts of the body. It is also known as advanced or metastatic cancer.

Once a cancer is in stage IV it is considered that this cancer has metastasis. In that case, the survival rate is usually very poor, as cancerous cells are spread to other parts of the body, usually adjacent to the starting tumour place.

Nowadays, the improvement in the standard of living of the whole population is leading to a population with a longer life expectancy. However, the elderly persons DNA can not repair the DNA damages suffered due to mutations and external factors. Figure 2.1 represents the number of annual diagnosis cases of cancer between 1990 and 2017 [18]. It is interesting to see how from 1990 to 2017 the number of people of 50 years old and above that suffer from some type of cancer have considerably increased. However, the number of diagnoses cases of cancer has grown in a greater or lesser extent regardless the age. This seems to be due to external factors (Tobacco, Alcohol, Pollution...) that are affecting directly to the health of our society.

This trend is repeated in Figure 2.2, but in this case representing the number of annual worldwide death because of cancer between 1990 and 2017.

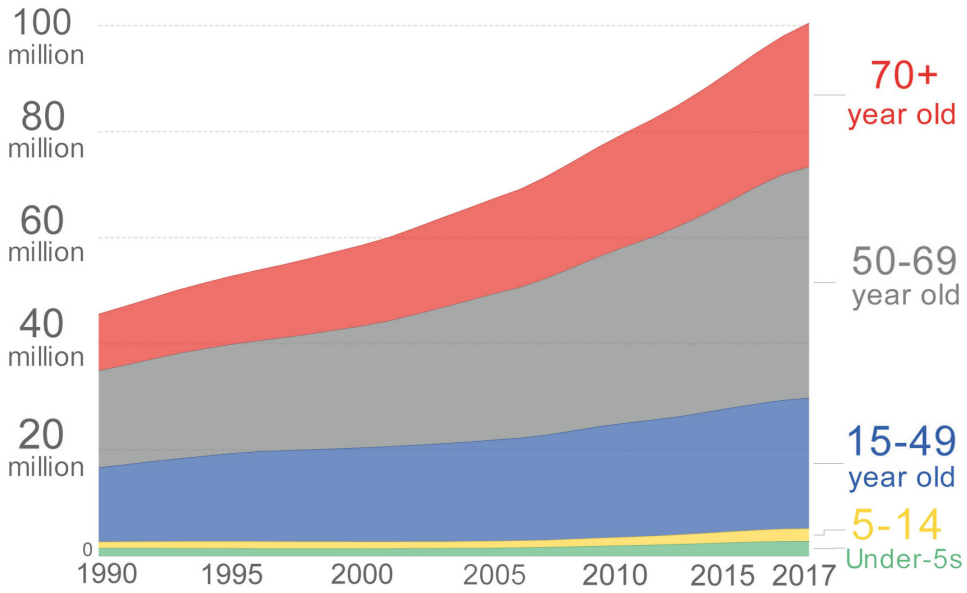

Figure 2.1: Total number of people worldwide with cancer differentiated by age. This is measured across all cancer types. Source: https://ourworldindata.org/cancer

2.2 HISTORY OF SEQUENCING

Genome keeps all the secrets about evolution, genetic diseases and the life itself. For that, from the end of the nineteenth century, with the original idea of the concept of gen (then called factors) proposed by Gregor Mendel, the mankind has been trying to discover and understand the complex world of genetics. During the twentieth century, advances in understanding genes and inheritance continued. Finally, through a set

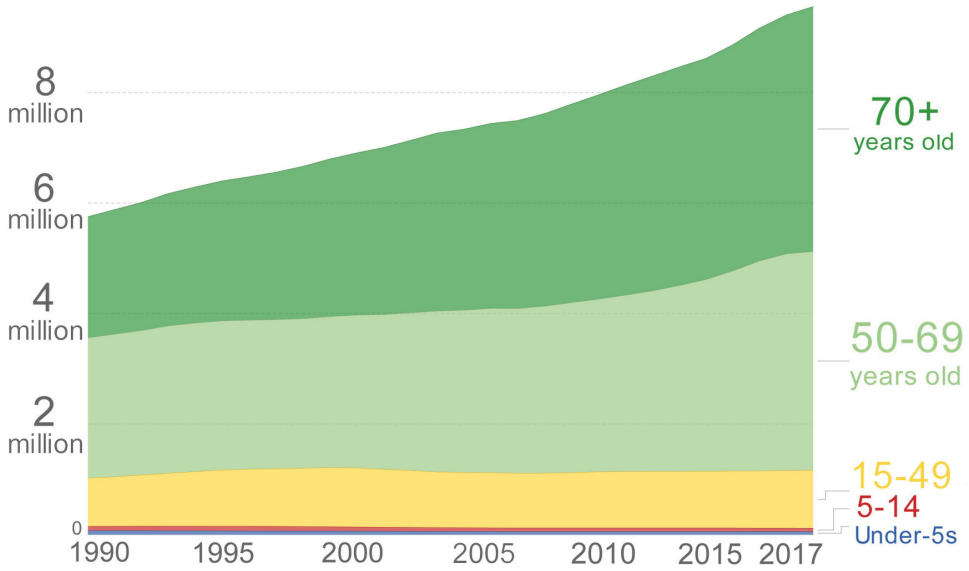

Figure 2.2: Total annual cancer deaths worldwide differentiated by age across both sexes. This is measured across all cancer types. Source: https://ourworldindata.org/cancer

of experiments in the 1940s to 1950s, it was demonstrated that DNA is the molecular repository of genetic information [19, 20]. Then, James D. Watson and Francis Crick published a model of the double-stranded DNA molecule whose paired nucleotide bases indicated a compelling hypothesis for the mechanism of genetic replication. Although Walter Fiers and his team were the first to determine the sequence of a gene [21], Frederick Sanger proposed the first improved method to sequence genes efficiently [22]. However, the whole human genome was not completely sequenced until 2003 with the Human Genome Project by using an automated version of the Sanger sequencing method [23].

Sequencing simply means determining the exact order of the bases in a strand of DNA. Because bases exist as pairs, and the identity of one of the bases in the pair determines the other member of the pair, researchers do not have to report both bases of the pair.

This section makes a journey through the different generations of genome sequencing [11]. These methods and technologies have contributed to improve the sequencing along the short history of this field. Furthermore, this section will take into account the near future of sequencing. The sequencing cost of a human genome has been drastically reduced from the start of the twentieth century to nowadays thanks to the technological improvements underneath the sequencing generations. As Figure 2.3 shows, with the arrival of NGS in 2007-2008, the evolution of the sequencing cost broke the Moore's Law trend, suffering

an exponential decrease of the sequencing cost per genome. Concretely, in early 2000s the cost per human genome was around 100.000.000 $, but at present time, sequencing a human genome cost only around 1000 $ depending on the technology and the conditions. Due to this enormous drop of the sequencing cost, it expects that the amount of available sequencing data will exponentially increase, allowing the development of more complex studies, requiring special pipelines and computer systems and architectures to analyse them.

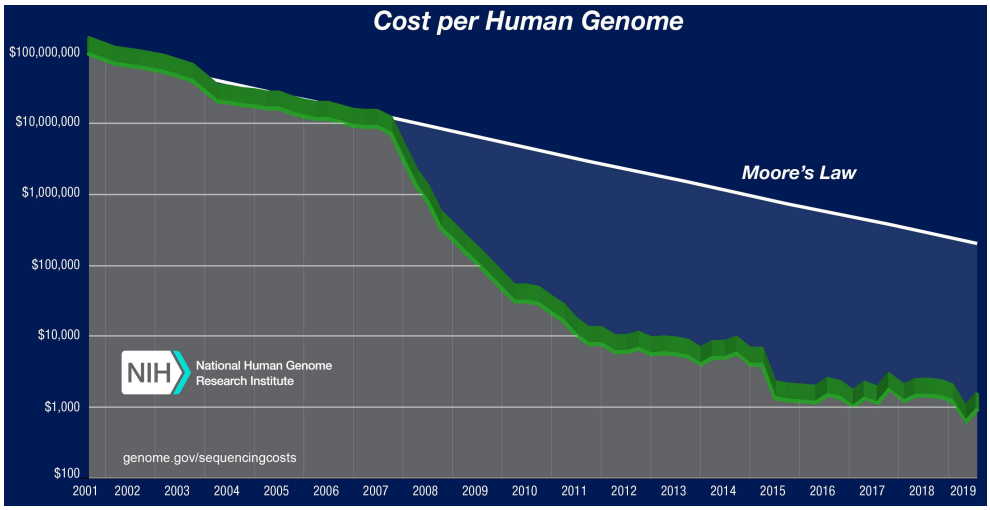

Figure 2.3: Sequencing cost evolution per human genome in last two decades. Source: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

2.2.1 *First Generation: Classic Sequencing*

This section will expose the classic sequencing techniques used until the arrival of NGS. Thanks to these techniques many advances in the sequencing of different species and diseases were made, achieving medical and biological milestones.

2.2.1.1 *Sanger Sequencing*

Frederick Sanger has been one of the few personalities in the world that has won twice a Nobel Prize. He won his second Nobel Prize thanks to a novel DNA sequencing method, that received his name, developed in 1975 [24]. Two years later, he used this method to obtain the first totally sequenced DNA of a living being in history, the bacteriophage Φ-X174 [22].

Sanger method is based on the DNA polymerisation and the use of dideoxynucleotides as reaction finishers. The first step in the sanger method is to heat the dsDNA to separate the two strands. Once the ssDNA is obtained, a primer or a small sequence of nucleic acids is introduced to initiate the sequencing process when this primer binds with the DNA polymerase.

dsDNA: DOUBLE-STRANDED DNA
ssDNA: SINGLE-STRANDED DNA

The primer is complementary to the start of the strand to sequence. The DNA polymerase keep going replicating until it find a nucleotide of stop (dideoxynucleotide). This process is repeated by using nucleotides of stop for each of the four nucleotides that conform the DNA: Adenine (A), Cytosine (C), Thymine (T) and Guanine (G).

Once the different sequenced DNA fragments are acquired, they are introduced inside of an electrophoresis tube. Inside the tube, the fragments are jointed to a fluorescent mark and subjected to an argon laser that allows to parallel measure the different fragments.

Figure 2.4 briefly shows the Sanger Sequencing. Firstly, The dsDNA fragment is denatured into two ssDNA fragments. Then, a ssDNA fragment is amplified into millions of copies. After that, a primer corresponding to one end of the fragment is added. It is then when the fragments are joined to four polymerase solutions. Subsequently, the chain grows until a termination nucleotide is randomly added and the resulting dsDNA fragments are denatured to acquire a set of ssDNA. Finally, the fragments are separated by electrophoresis and the sequence is read.

2.2.1.2 *Maxam-Gilbert Sequencing*

This sequencing method was developt by Maxam and Gilbert in the year 1976 [25]. Maxam-Gilbert method is very effective but limited, due to the necessity of chemical sequencing. This means that it is required the use of chemical processes to interrupt the DNA chains. Then, the resulting fragments run through a gel to resolve the sequence order.

The first step to acquire the sequence is to denature the dsDNA to obtain the ssDNA by applying heat. When the two strands are separated, Gamma-32P are joined with 5′ end of the DNA fragment by a kinase reaction.

After that, the DNA molecule is marked at specific nucleotides and then, broken with the purpose of obtaining a break for each of the reac-

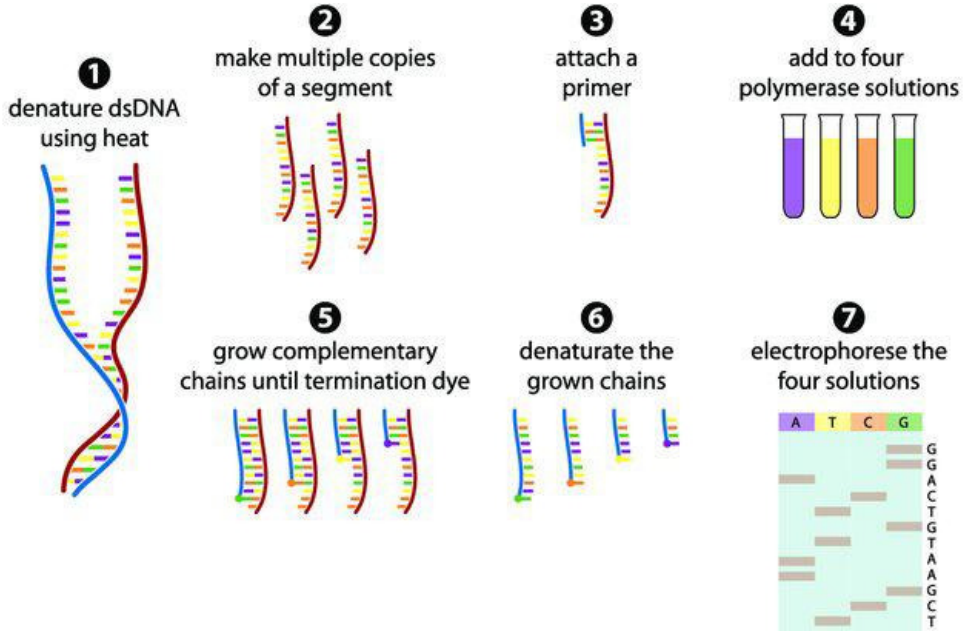

Figure 2.4: Sanger Sequencing method. Source: https://www.researchgate.net/publication/234248746_Simulation_of_polymer_translocation_through_small_channels_A_molecular_dynamics_study_and_a_new_Monte_Carlo_approach/figures

tions (G, A+G, T+C, C). The reactions are made through the following chemical agents:

- Dymethil Sulfate (G)
- Formic Acid (A+G)
- Hydrazine (T+C)
- Hydrazine plus salt (C)

Due to these reactions, a set of radioactively marked fragments are generated from the end until the place where the molecule was broke. Once the fragments are extracted, they are separated by size in four reaction tubes, using electrophoresis gel. To visualise the fragments of each reaction, the gels are placed under X-ray, which achieves a set of dark bands which represent the location of radiolabeled DNA molecules. Finally, the fragments are ordered by size and the DNA sequence can be deduced by inference. A graphical representation of the Maxam-Gilbert method can be seen at Figure 2.5.

2.2.2 *Second Generation: Next Generation Sequencing*

NGS makes reference to the second generation of sequencing technologies, also known as HTS. Nowadays, NGS has replaced the first generation sequencing methods and it has been set as the current standard in the market. These technologies are more powerful and cheaper than the classic technologies, thus they have revolutionised the way of studying the different omics.

HTS: High Throughput Sequencing

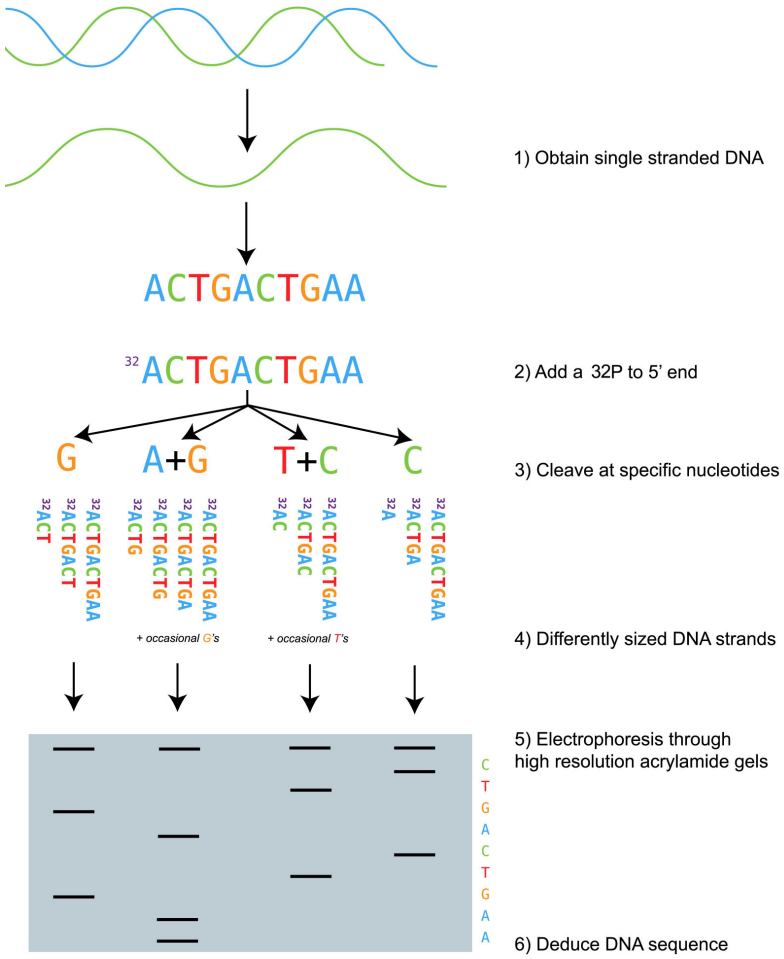

Figure 2.5: Maxam and Gilbert Sequencing method. Source: https://binf.snipcademy.com/lessons/dna-sequencing-techniques/maxam-gilbert

2.2.2.1 *Roche 454 sequencing*

This technology is considered the first available commercial NGS method [26]. It is required to perform a library preparation before the sequencing process. Firstly, the DNA is broken into small fragments of 300-800 BP, and different adapters are also added to both ends of each fragment.

BP: Base pair

The second step is to carry out the emulsion PCR. The most important feature of emulsion PCR is the creation of a large number of independent reaction space for DNA amplification. When the sample DNA amplification is finished, an aqueous solution that contains the emulsion PCR reaction components will be imbued into a mineral oil surface. With this, several small water droplets are formed and wrapped by mineral oil. Each of the small droplet builds an independent PCR reaction space. Finally, each tiny fragment will be amplified over 1 million times in order to achieve the minimum level needed by the sequencing.

PCR: Polymerase Chain Reaction

When the library creation and the amplification are correctly carried out, the last step to perform is the amplified fragments sequencing. Pyrosequencing method is usually used to achieve the sequences. For that, a small beads are inserted into the nanopores of a PTP and the sequencing reaction is started. DNA sequencing reaction is based on the ssDNAs which are amplified and fixed. In the reactions, each type of dNTPs produce a characteristic fluorescence colour, hence the DNA sequences are measured according to those fluorescence colours. At the end, the sequencing results will be processed by computer software. Figure 2.6 represents in detail the Roche 454 sequencing method.

PTP: PicoTiter Plate

dNTPs: deoxyribose Nucleoside Triphosphates

2.2.2.2 *Illumina SBS Sequencing*

The SBS, also known as Reversible Terminator Sequencing, is a widely used NGS. SBS was developed and established by Illumina as a commercial implementation in 2008 [27]. This sequencing technology is responsible for the 90% of the NGS worldwide data. The Illumina systems are able to carry out massive parallel sequencing that considerably reduce the sequencing time and cost. SBS was introduced by Illumina, with their HiSeq and MiSeq platforms. Concretely, HiSeq sequencer is the cheapest of the second generation sequencers with a cost of $0.02 per million bases.

SBS: Sequencing By Synthesis

Firstly, the DNA must be broken into more manageable fragments between 200 to 600 BP. Then, short sequences of DNA known as

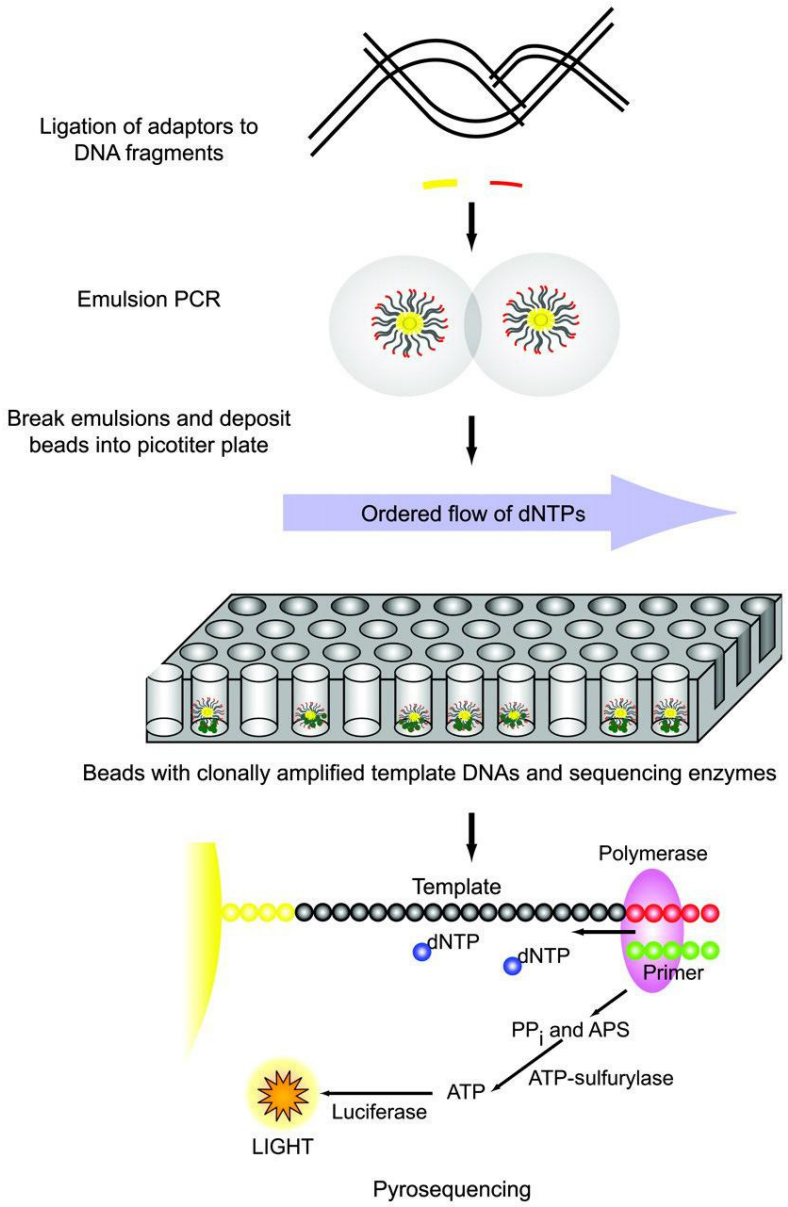

Figure 2.6: Roche 454 Sequencing method. Source: http://clinchem.aaccjnls.org/content/55/4/641/tab-figures-data

adaptors are joined to the fragments. The cDNA binds to primers on the surface of the flowcell and the DNA that does not adhere is washed away.

cDNA: COMPLEMENTARY DNA

It is necessary to amplify the DNA fragments, for that they are replicated to form small clusters of DNA with the same sequence. Thanks to this amplification, each cluster will emit a signal that is strong enough to be captured by a sensor. A set of bridges of dsDNA will be created between the primers on the flowcell surface.

The dsDNA is then denatured into ssDNA by applying heat, leaving million clusters with the same DNA sequences. Then, the primers and fluorescently labelled terminators that cut DNA synthesis are introduced to the flowcell.

The DNA polymerase is bound to the primer and the first fluorescently labelled terminator is added to the new DNA strand. Then, a set of lasers are passed over the flowcell to activate the fluorescent label on the nucleotide base. This fluorescence is detected by a sensor and registered on a computer. The terminator is then removed from the first base and the next terminator base can be joined alongside. Finally, the process goes on until all the clusters have been correctly sequenced. All this process is showed at Figure 2.7.

2.2.2.3 *ABI/SOLiD Sequencing*

SOLiD Sequencing is a NGS technology developed by Life Technologies. It is a very powerful technology with the capability of generating hundreds of million of small sequencing reads at the same time [28].

Starting with a sample to sequence, a library of DNA fragments is prepared, which will be used to also create clone populations. The joined fragments have an universal adaptor sequence in order to establish an identical and known initial sequence for each fragment.

After that, an amplification by emulsion PCR is carry out, just like for 454 Sequencing method. Thanks to the amplification process, each fragment could be amplified around 1 million times with the purpose of achieving the minimum level required by the sequencing method.

Finally, the sequencing process has several hybridisation rounds with 16 different nucleotides marked with 4 different colours. Employing a colour code each position is evaluated twice by two different primers and, due to that, the discrimination among sequencing errors and SNPs polymorphisms detection increase. A representation for this method is showed at Figure 2.8.

SNPs: Single Nucleotide Polymorphisms

2.2.2.4 *Ion Torrent sequencing*

Ion Torrent sequencing was released on February of 2010 by Ion Torrent system [29]. As happen with 454 sequencing, the first step for the sequencing is the library preparation. This process is very standardised

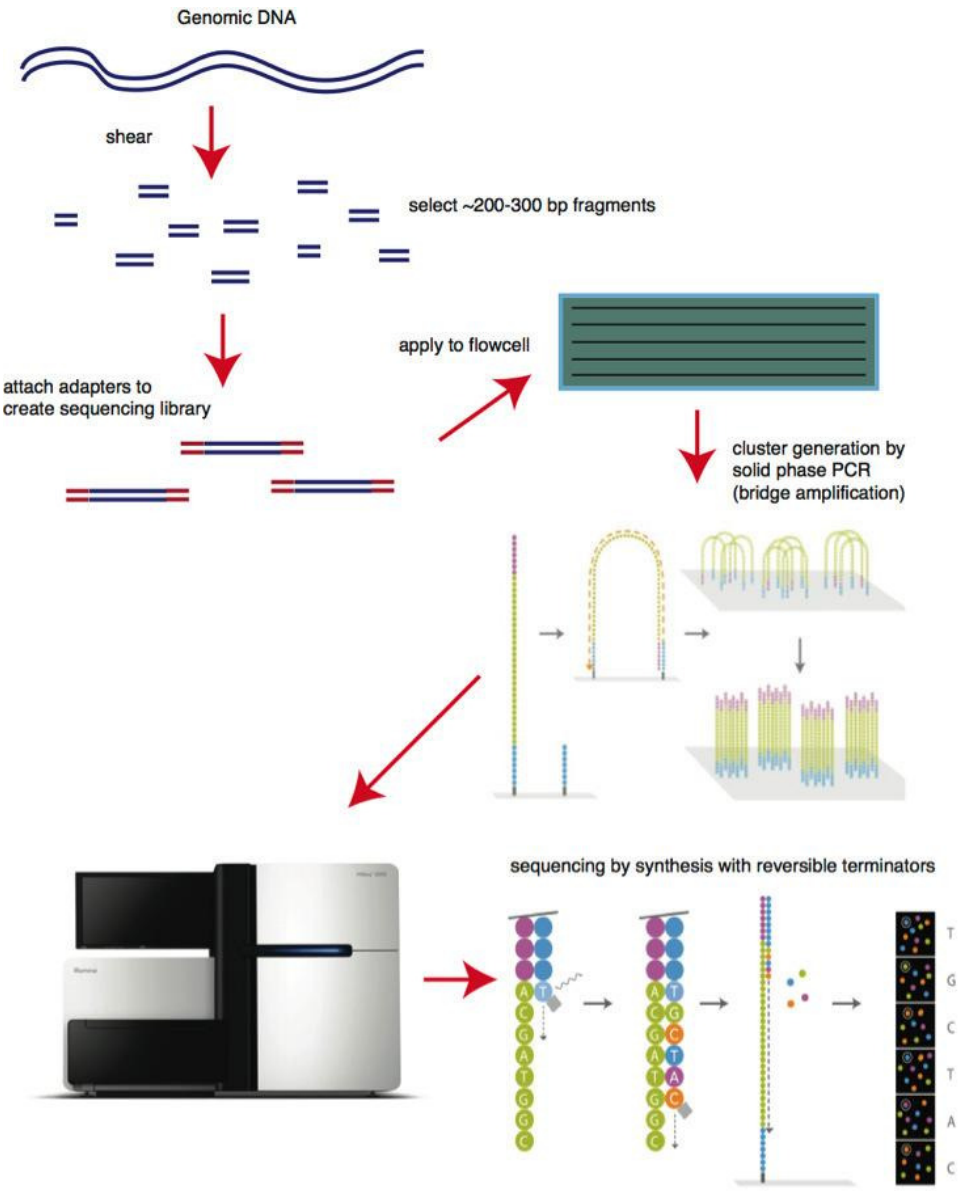

Figure 2.7: Illumina Reversible terminator Sequencing method. Source: https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/

and normally takes DNA, fragmenting it to a fragments between 200 and 400 BP. Then, those fragment are amplified by the emulsion PCR method explained above.

Ion torrent is base on the synthesis sequencing, that works through the polymerization of a DNA complementary chain with natural dNTPs. Furthermore, a semiconductor chip with the capability of detecting H+ ions is required. This method works as follow:

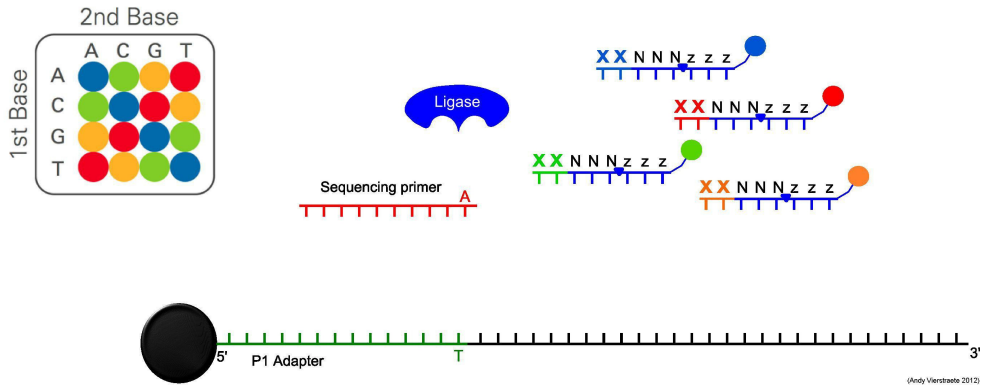

Figure 2.8: ABI SOLiD Sequencing method. Source: https://www.researchgate.net/publication/24043867_Next-Generation_Sequencing_From_Basic_Research_to_Diagnostics/figures?lo=1&utm_source=google&utm_medium=organic

When a nucleotide is joined to a DNA strand by a polymerase, an hydrogen cation is released as subproduct. This cation carry an ionised charge that can be detected by the ions sensor. Then, while the sequencer overwhelms with nucleotides the chip, any nucleotide incorporated to the template strand will be detected by the sensor due to the change in the voltage, and the system will report the corresponding base. If a nucleotide is added more than once consecutively, the detected signal will be more intense. The explanation of Ion Torrent method can also be seen at Figure 2.9.

2.2.3 *Third Generation: towards the future sequencing*

As was mentioned ahead, NGS has revolutionised the DNA analysis and it is the most widely used technology nowadays, consequently the largest amount of sequencing data comes from NGS at present. Nevertheless, NGS technologies need an amplification process through PCR. This amplification step usually is very expensive and takes a considerable amount of time. Furthermore, these technologies create relatively short reads that needs a posterior complex assembly step. With the aim of solving this particularities, scientists have developed the third generation sequencing. The technologies developed under the third generation are cheaper than their predecessors, and samples preparation is easier because they do not need PCR amplification. Moreover, third generation technologies have the capability to create long reads exceeding several kilobases for the resolution of the assembly problem.

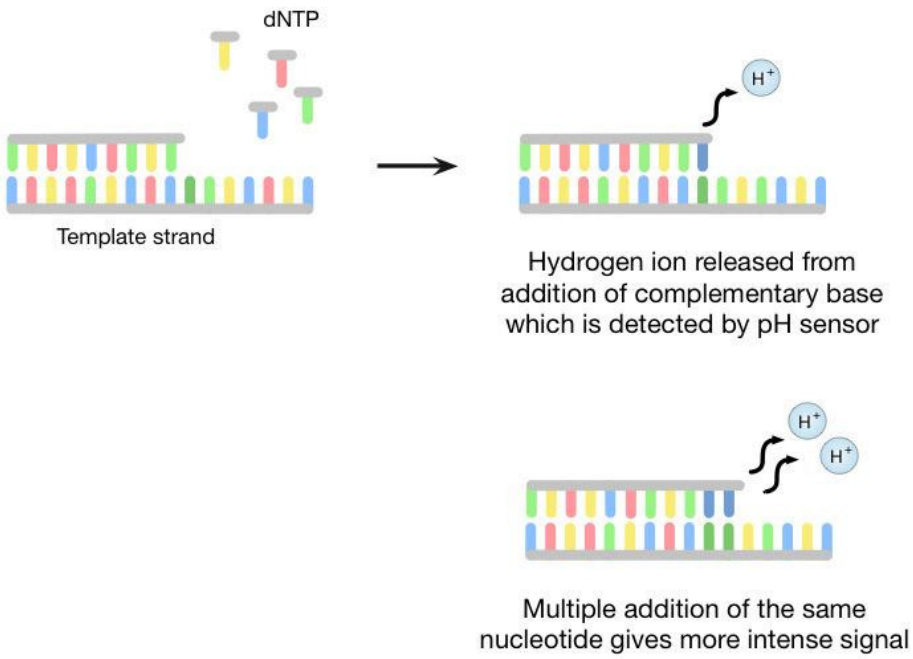

Figure 2.9: Ion torrent Sequencing method. Source: https://www.atdbio.com/content/58/Next-generation-sequencing

2.2.3.1 *Pacific Biosciences SMRT Sequencing*

Pacific Biosciences was the first company to developed the first SMRT sequencing method and, nowadays still being the most widespread third generation sequencing technology [30].

SMRT: Single Molecule Real Time

This method is based on the same fluorescent labelling than other technologies (Roche 454, Illumina SBS), but the signals of the nucleotides are detected in real time instead of by amplification. For that, Pacific Biosciences create a structure with many SMRT cells that contain microfabricated nanostructures or ZMWs. The ZMWs exploit the properties of light passing through openings with a diameter less than its wavelength, so light cannot be propagated. Each ZMWs has a DNA polymerase adhered to their bottom with the DNA fragment to sequencing.

ZMWs: Zeromode Waveguides

When the sequencing reaction starts, the DNA fragment is introduced by the DNA polymerase with fluorescent labeled nucleotides. Whenever a nucleotide is added, it releases a luminous signal that is recorded by sensors. The DNA sequence could be determined thanks to detection of these labeled nucleotides. The process is showed at Figure 2.10.

This SMRT method has some advantages in comparison with NGS technologies. The main advantage is the facility to prepare the sample, because it only takes 4 to 6 hours instead of days as happen with NGS. Furthermore, the long-read lengths currently are around 10 kbp but, individual very long reads can be until 60 kbp, which is longer than any of the NGS approaches. However, this systems have a high error rate of around 13%, which mainly are INDELS errors along the long reads.

INDELS: Insertion and Deletions

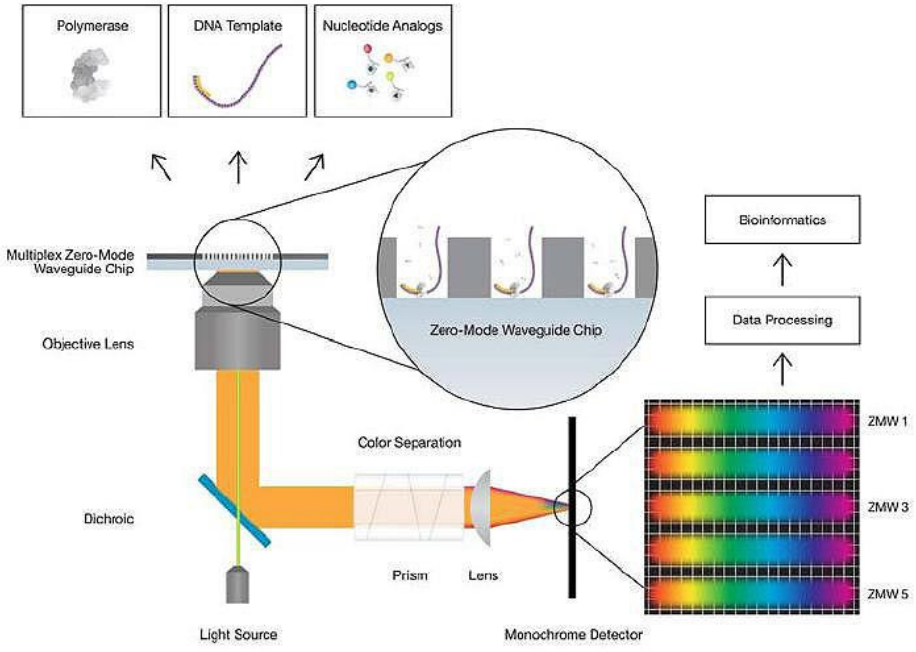

Figure 2.10: Pacific Biosciences SMRT Sequencing method. Source: https://www.researchgate.net/publication/281772504_EVALUATING_EMERGING_TECHNOLOGIES_APPLIED_IN_FORENSIC_ANALYSIS/figures?lo=1&utm_source=google&utm_medium=organic

2.2.3.2 *Oxford Nanopore DNA Sequencing*

ONT was developed in 2014 as a SMRT method to determine the DNA sequence. Firstly, the company released a portable device called MinION with the purpose of generating longer reads that will ensure better performances that NGS technologies. The device is connected to a laptop through the USB 3.0 port. In order to examine the performance of device, the company released a program MAP for testing it by a community of users.

ONT: Oxford Nanopore Technology

USB: Universal Serial Bus

MAP: MinION Access Program

To carry out the sequencing in this technology, the first strand of a DNA molecule is adhered by a hairpin to its complementary strand. Then, the DNA fragment is passed through a nanoscale hole, known as nanopore. This nanopore can be made by proteins or synthetic materials.

As Figure 2.11 shows, when the DNA fragment passes through the nanopore, it generates a variation of an ionic stream, this variation is recorded inside a graphic model and then analysed to distinguish the sequence. For this reason, the sequencing is made on real time over the strand, generating the template read. Finally, the hairpin structure is read followed by the inverse strand, creating the complement read.

The advantages for this methods are very similar to the advantages for Pacific Biosciences SMRT Sequencing. Firstly, the low cost and small size of this technology is a point of inflection in comparison to predecessor technologies. Furthermore, the MinION device provides portability and versatility, due to the sample is loaded into a normal laptop and data is displayed on the screen and generated in real time. Finally, the device can create very long reads which can improve the posterior assembly. Nevertheless, the device has a high error rate of 12%.

2.3 UNDERSTANDING THE MAIN OMICS

In the evolution of a cancer, there exists different processes to be analysed: from the genetic mutations or alterations to the uncontrolled massive cellular proliferation. These biological processes have been studied separately by different omics, depending on the biological type of the data involved and the information that each type provides. The word omics makes reference to a field of study in biology. The omics study point to the collective profiling and quantification of

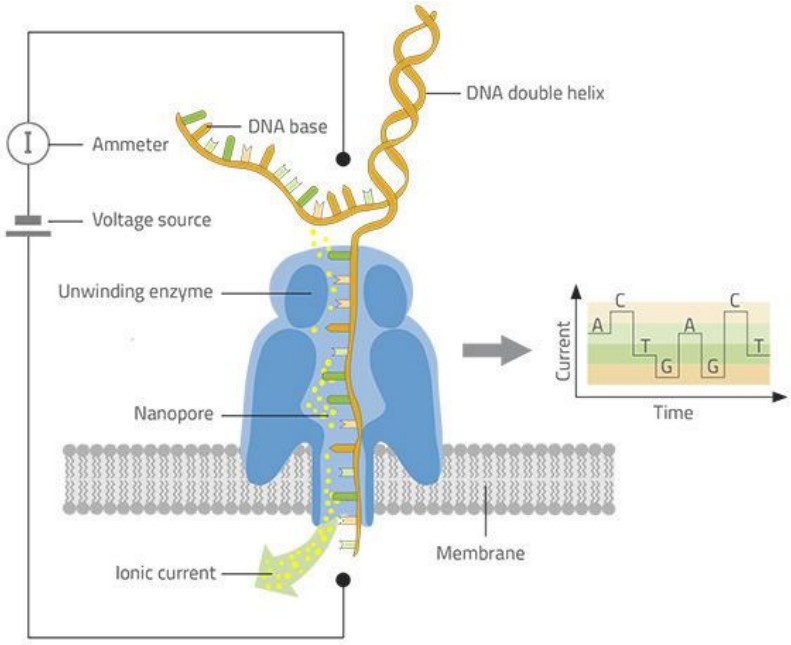

Figure 2.11: Nanopore Sequencing method. Source: https://www.scienceinschool.org/content/decoding-dna-pocket-sized-sequencer

clusters of biological molecules, that are translated into the structure and functions of the different organisms. Thanks to the biological characterisation technologies and the precision that they supply, there are many advances in the fight against rare diseases and cancer, by using those omics as well as their integration [31–36].

In the precision medicine and computational biology scope there are mainly four omics: Genomics, Transcriptomics, Proteomics and Metabolomics. Although they will be explained in depth in the next subsections, Figure 2.12 shows the relations among them. A change or mutation in the DNA could lead to changes at expression level of the genes measured in the RNA. Then, those expression variations could change also the proteins codified by the affected genes and hence, affect to the metabolites. Therefore, a mutation or variation in the genes could lead to a series of biological changes at different biological levels, which could end up arising a genetic disease.

RNA: Ribonucleic acid

2.3.1 *Genomics*

Genomics is focused on the structure, function, evolution, mapping, and editing of genomes. All the information about a person from birth

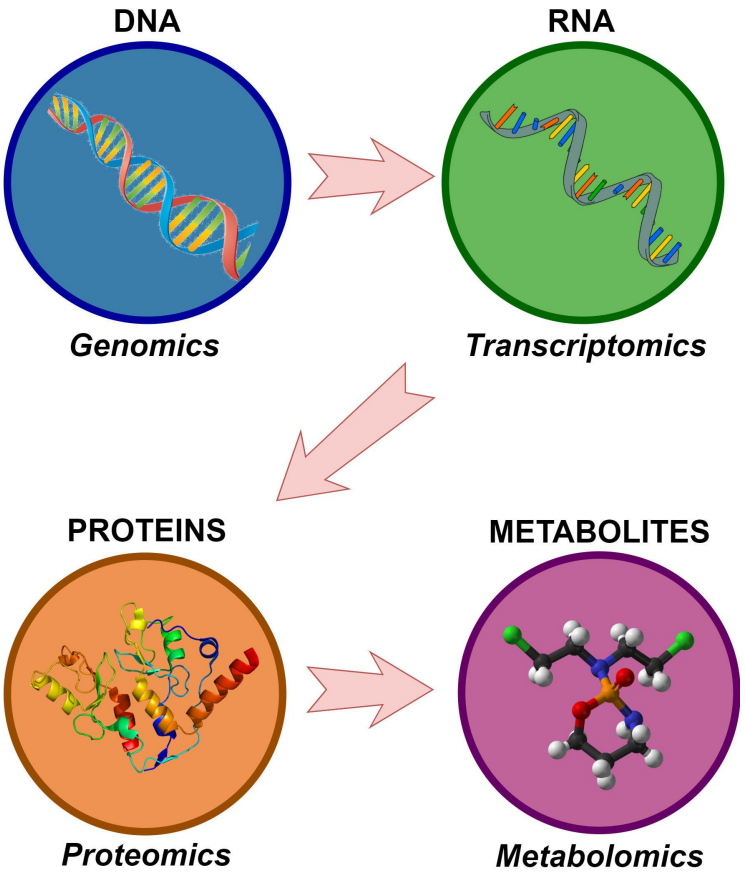

Figure 2.12: Relation among the four main omics: Genomics, Transcriptomics, Proteomics and Metabolomics. A change in one of them, could leads to a series of biological changes in the operation of its subordinates.

to death is codified in the genome. For that, in order to understand our life itself, the study and comprehension of the whole genome is required to understand our life itself. The eyes colour, the stature, the physical constitution, even the diseases that a person could suffer from are some of the codified things in the genome.

Genomics studies are usually known as GWAS or WGAS, and they try to find associations between genome variants and genetic diseases. Those variants or alterations in the genome might be beneficial to the organism, harmful or neutral. Moreover, if the variants affect to a 50 BP or less, they are considered as short variants. Taking this into account, SNPs and short INDELS are part of this group of variants. On the other hand, if the variants affect to more than 50 BP, they are considered as structural variants. Inside the structural variants can be found long INDELS, duplications, CNV, inversions and translocations [37]. All these variants are represented graphically at Figure 2.13.

GWAS: Genome-wide association study
WGAS: Whole Genome Association Study

CNV: Copy Number Variation

With that in mind, there are two type of genetic mutations that could lead to develop genetic diseases or cancer. The first type emerges from germinal mutations in the genetic constitution of the reproductive cells. These type of mutations might pass to the progeny of the individual and affect to population evolution over the years. There exists several studies that correlate germinal mutations across generations with inherited cancer and rare diseases [38–44].

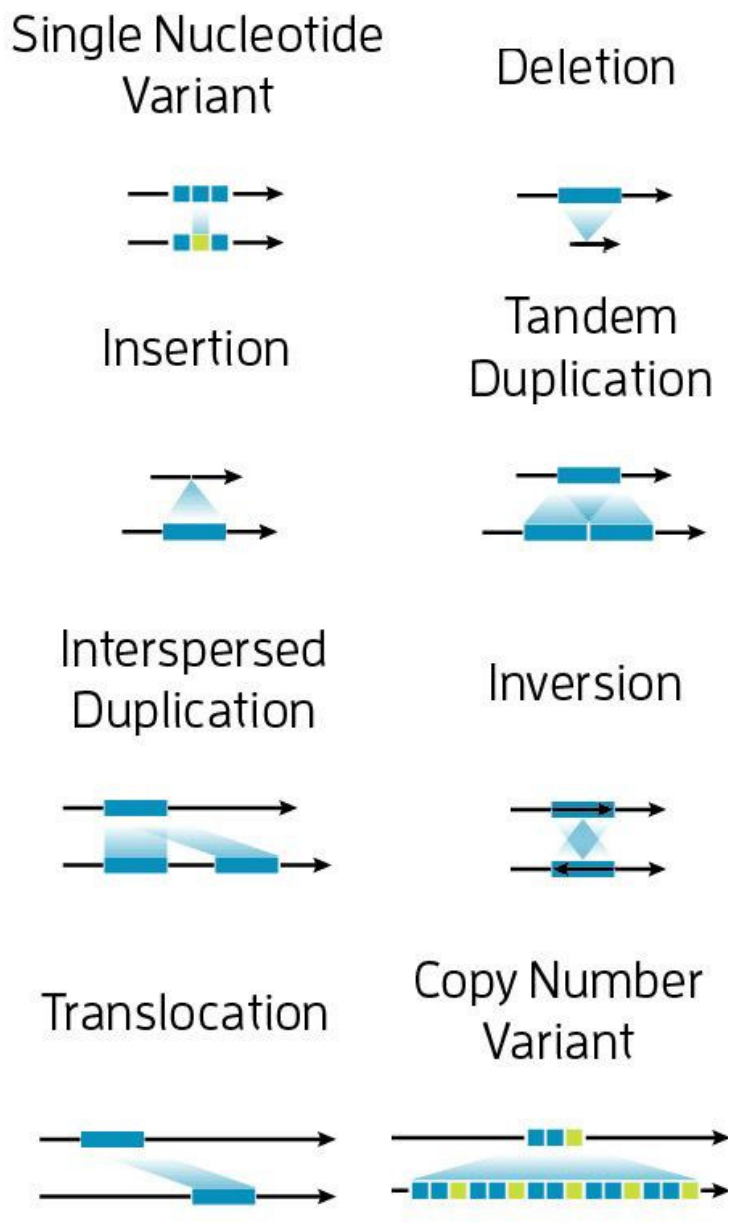

Figure 2.13: Different types of variants existing in the genome. Source: https://www.pacb.com/applications/whole-genome-sequencing/variant-detection/

The second type emerges from somatic mutations in the genome during the life of an individual. These mutations do not pass to next generations. Somatic mutations can accumulate in our cells and are mostly harmless, but may also have more serious effects such as cancer [45]. There a plenty of studies searching the role of somatic mutations for the different types of cancer [46–50].

2.3.2 *Transcriptomics*

The main goal of transcriptomics is the study and comprehension of the transcriptome, or the set of RNA present in a cell. The transcriptome shows what genes are expressed or inhibited at a specified time. With this information, it can be found those genes that are related with the development of the different types of cancer. For that, the transcriptomics studies are focused on the differences at expression level between both the cancerous cells and the healthy cells transcriptomes.

The production of the RNA is driven by the DNA of the cell, which supplies a pattern to create the messenger RNA. This creation of RNA is known as transcription. Then, through a process called translation, the messenger RNA is translated into the final codified proteins in the cells. Figure 2.14 shows graphically how the RNA is produced from the DNA replication, and its role in the proteins creation.

When a person suffers from a cancer, its genome has usually been affected by commonly several genetic mutations. This fact leads to the codification of erroneous messengers RNA when the transcription step is carried out. At the end, those wrong transcriptions could be translated in erroneous proteins, which directly affect to the biological processes of the cells, tissues or even the entire organism.

For this reason, one of the most important pillars in the battle against cancer is the research of the expression level of the genes, in order to determine the consequences of the variation of the expression in the protein codification. There is a large literature on transcriptomic changes with the objective of extracting biomarkers for different types of cancer [51–55], even sometimes predicting the survival rate [56–58].

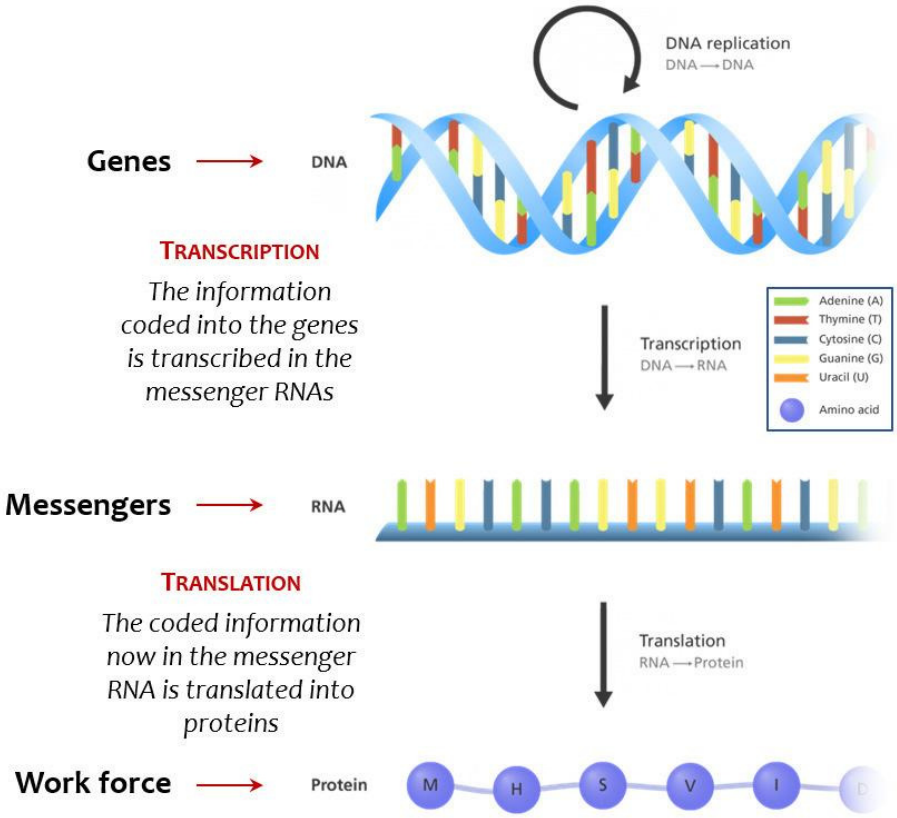

Figure 2.14: From genes to proteins through RNA. Source: https://www.lgmd2ifund.org/science-basics/from-gene-to-protein/attachment/reading-gene

2.3.3 *Proteomics*

Proteomics is focused on the large scale study of the proteomes. The proteome represents the set of codified proteins for an organism. However, the proteome changes along the time and differs from cell to cell. To a certain extent, the proteome reflects the underlying transcriptome because the messenger RNA is translated into proteins. Nevertheless, protein activity is also affected by other factors apart from the expression level of genes.

DEPs: Differentially Expressed Proteins

As can be seen at Figure 2.15, there exists different areas inside proteomics according to the experimental design. Many researches try to seek when and where the proteins are expressed (DEPs) together with how they affect to biological processes [59, 60]. There are also many interest in how those proteins affect metabolic pathways [61, 62]. Furthermore, there are studies focused on the proteins production, degradation and steady-state abundance [63–65].

However, there is a concrete field known as PPI which studies the proteins interaction and, in many cases, their relation with cancer. AA Ivanov et. al. exposes the potential of PPI as a anticancer strategy [66]. Moreover, Wang, S. et. al. target a concrete PPI as a new cancer therapeutics [67]. PPI is also applied to detect prognostic significance in breast cancer thanks to the interaction of proteins, as demonstrate Spears, M. et. al. [68]. Besides the mentioned studies, there are many more aiming to understand this complex field, and to improve the biological cancer mechanisms knowledge at proteins level to counteract them.

PPI: Protein-Protein Interactions

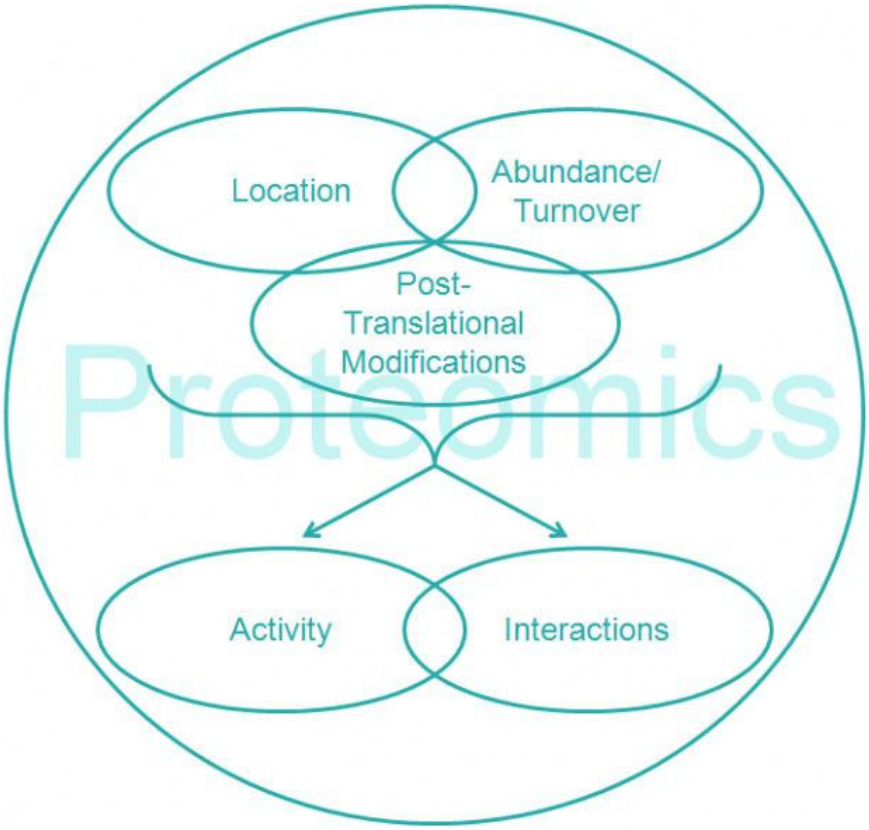

Figure 2.15: Different areas of study in proteomics. Source: https://www.ebi.ac.uk/training/online/course/proteomics-introduction-ebi-resources/what-proteomics

2.3.4 *Metabolomics*

Metabolomics makes reference to the in-depth study of the small molecules or metabolites inside a cell, tissue or organism and directly reflects the underlying biochemical activity and the state of cells or tissues. The metabolome is the total existing set of metabolites in a

biological sample under given genetic, nutritional or environmental conditions.

The metabolome is formed by small molecules or metabolites, defined as low molecular weight organic compounds, typically implicated in biological processes as substrates or products. Normally, metabolomics studies metabolites within a mass range between 50 and 1500 DA. Just as a curiosity, one DA is equivalent to $1.660\,54 \times 10^{-24}$ grams. Figure 2.16 shows some examples of small molecules or metabolites.

DA: Daltons

Due to the reaction of the metabolites, the metabolome is constantly changing. The small molecules are constantly being absorbed, synthesised, degraded and in interact with other molecules, both within and between biological systems, and with the environment. Figure 2.17 contains the possible reactions that affect the metabolites in a cell.

Metabolites analysis are an ideal tool for precision medicine due to its non-invasive nature and its close link to the phenotype. Biomarker discovery on cancer, and drug safety screens are two examples where metabolomics has already helped diagnosis and decision making [69–71].

Metabolites can be used as biomarkers to distinguish between two groups of samples (disease and control) or more. Taking this into account, a metabolite present in disease samples but not in healthy samples, would be selected as a metabolite biomarker. For example, samples coming from urine, saliva, bile, or seminal fluid can be used to discover biomarkers, due to the highly informative metabolites that each biological fluid contains. For that, this omic has the potential to identify hundreds of metabolites, giving the possibility to diagnose these diseases in an earlier stage. In this sense, there are a lot of studies seeking biomarkers for cancer early detection and diagnosis [72–76].

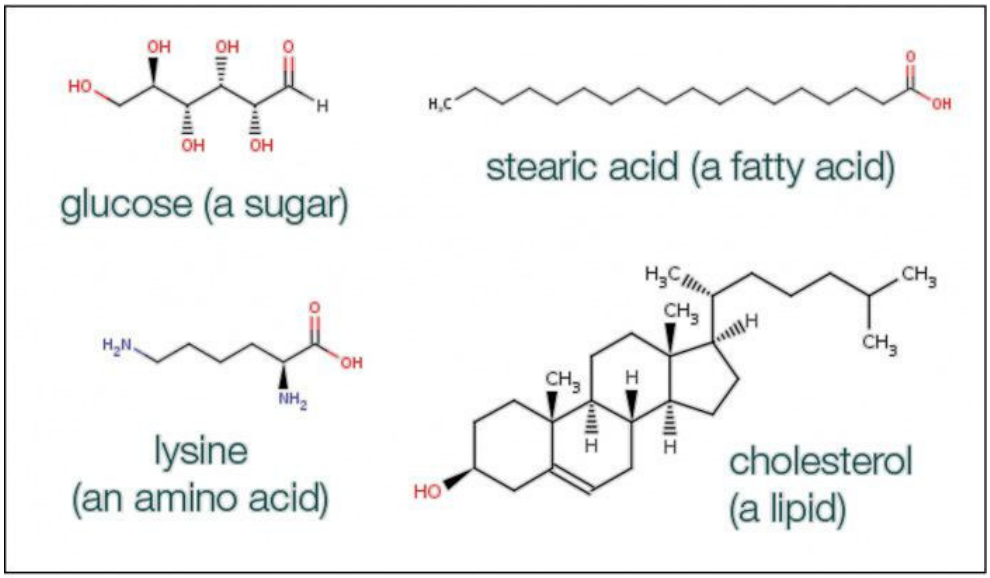

Figure 2.16: Small molecules or metabolites examples: Sugar, fatty acid, amino acid and lipid. Source: https://www.ebi.ac.uk/training/online/course/introduction-metabolomics/what-metabolomics/no-glossary-small-molecules-no-glossary

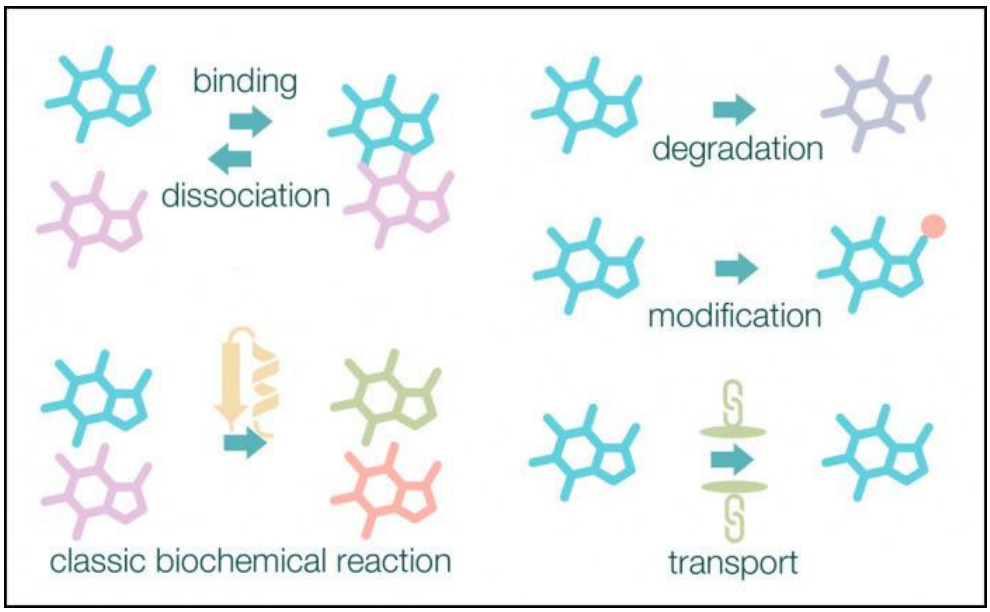

Figure 2.17: Metabolic reactions produced in a cell: Binding/Dissociation, degradation, modification, classic biochemical reaction and transport. Source: https://www.ebi.ac.uk/training/online/course/introduction-metabolomics/what-metabolomics/metabolome-and-metabolic-reactions

2.4 HETEROGENEOUS TRANSCRIPTOMICS SOURCES

Throughout this chapter, Sections 2.1, 2.2 and 2.3 have addressed several important definitions and the state of the art in relation to cancer and the different sequencing generations and omics. This overview is very important in order to understand the heterogeneous data gathering carried out in the experiments included in this doctoral thesis. The "heterogeneous" term refers to the fact that data used is a collection of datasets coming from different sources and different sequencing technologies. This last section of the chapter presents relevant information and studies about heterogeneous transcriptomics sources.

Gene expression is practically the most important source to find biomarkers in cancer. Because of that, this thesis is focused on the use of gene expression sources and their integration. Concretely, the integration of datasets coming from both Microarray and RNA-Seq technologies, which are explained at Subsections 2.4.1 and 2.4.2.

Biomedicine literature shows that, for each cancer, there are many genes playing a role in its development and also many combination of them. However, it is important to note that for each cancer there is not only a unique gene signature, or a single set of genes that have direct and unambiguous relation with it for clear diagnosis, prognosis or prediction of therapeutic response. Rather, the number of possible gene signatures variate is undefined, and literature show very different outcomes in biomarker set discovery for a given cancer, which might even depend on the statistics threshold and the number of samples used for a given the study.

The use of machine learning approaches for validating gene signatures is broader than ever before, due to the increase in the number of available samples and the current computational power [77]. Under this reality, many studies suggest possible gene signatures that are not clinically validated [78], although being statistically significant. Nevertheless, care must be taken when a gene signature is proposed, because applying machine learning without a properly biological interpretation could lead to a nonsense gene signatures [79].

There are several studies in the literature that propose different gene signatures for different types of cancer. Ru He and Shuguang Zuo provide a gene signature with 8 different genes for early-stage Non-small Cell Lung Cancer in their study [80]. On the other hand, Cardoso et. al. supply 70-gene signature for treatment decision in breast cancer [81]. In this sense, there are similar studies for almost all the existing can-

cer [82–84]. At sight of this, there is not only a unique gene signature that identify a concrete cancer because there are an enormous number of genes involved in cancerous processes.

2.4.1 *Microarray*

For several decades, Microarray technology has allowed studying alterations at gene expression level with the purpose of finding genes involved in pathologies of genetic source. This technology is highly widespread and well-known, and is based on the capability of the complementary molecules to hybridise among themselves. This allows determining the gene expression values of each studied gene in the analysed set of samples [85]. Through this process, the over-expressed or inhibited genes can be identified in tumor samples when comparing to normal samples.

The microarray operation is based on the DNA hybridisation process. Through this process, two DNA fragments are only hybridised if they are complementary between them. In order to be complementary, the Watson-Crick rule must be fulfilled. This rule establishes that in the DNA, the Adenine (A) is joined to the Thiamine (T) and the Cythosine (C) to the Guanine (G) as can be observed at Figure 2.18.

The process to measure the gene expression of a sample in a microarray is explained hereunder and represented in the Figure 2.19. The spots or oligonucleotides probes are adhered to a surface of 1 cm^2, creating a DNA array in which the probes are equidistant among them. Then, each DNA fragment is fluorescently labeled and incorporated into the array. Once this has been done, the genetic material in each probe that has not been hybridised is cleaned. Finally, the microarray is measured by a scanner when the fluorescent probes are subjected to a laser. Afterwards, the microarray image is analysed to know and quantify the proportion of hybridised samples.

The final result is stored at a .CEL file which can be loaded into a computer to analyse the gene expression values of an individual. In the majority of studies concerning to microarray, there are as CEL files as individuals, hence the volume of information to analyse is usually very high.

There are mainly two microarray technologies that have almost all the market, Illumina [86] and Affymetrix [87]. Several studies have emphasised the good correlation between the data from both technologies,

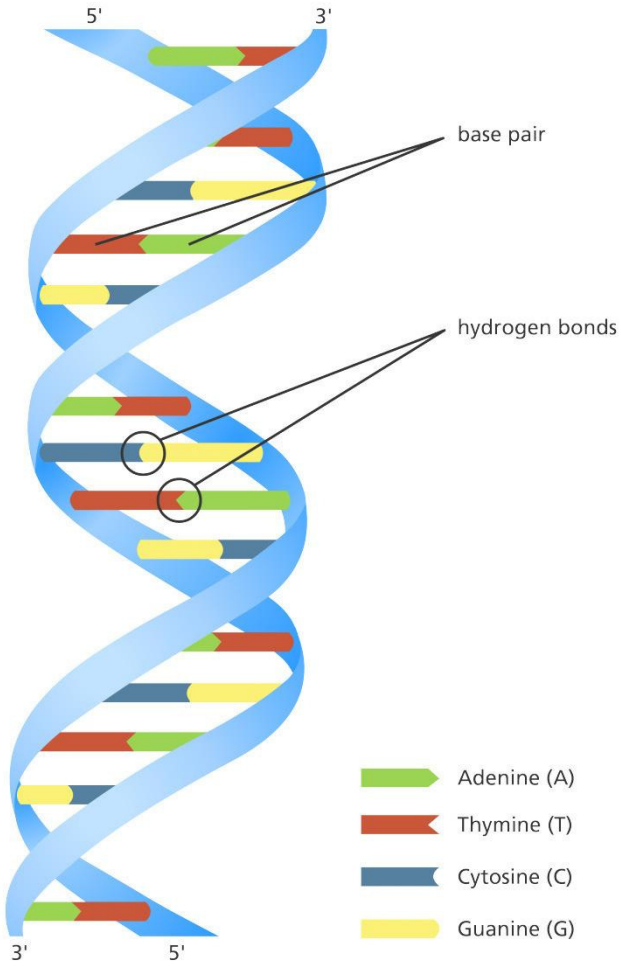

Figure 2.18: Hydrogen bonds and nucleotides of the DNA Double-Helix. Source: https://trickle.app/drip/15481-dna-is-the-blueprint-for-building-our-bodys-cells/

specially when differential expression analysis are addressed [88–90]. However, there are others important manufacturers or technologies as Agilent [91], Exiqon [92] or Taqman [93].

A feature of Illumina is that it provides the expression data not only at probe level but also at gene level. That can be achieved by implementing around 30 replies for each nucleotide in the array. This means that the same gene can be stored in different probes. Illumina can also provides expression data for each samples separately or grouped.

On the other hand, Affymetrix makes use of multiple probes as an internal controls to verify the correct functioning and not only for the hybridisation.

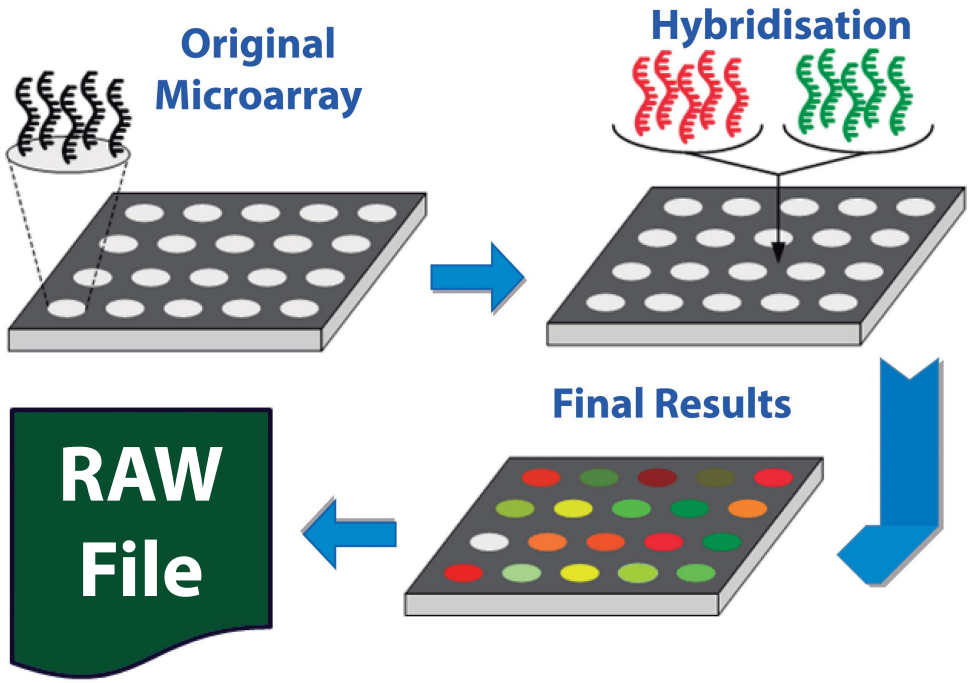

Figure 2.19: Microarray creation process. Through this process, a set of genes can be measure at expression level to carry out differential expression analysis between a chosen population.

Affymetrix counts with spotted cDNA, which are arrays-based GeneChips. The main difference with Illumina BeadArray is that each probe in the array is located in a specific and known position. The Illumina process is different due to there is a decoding step for the position of each probe depending on it molecular location. Figure 2.20 shows the two different Affymetrix and Illumina microarrays explained ahead.

Lastly, the Illumina hybridisation happens in parallel because several arrays are faced to the same substrate, while Affymetrix arrays are processed in different substrates.

2.4.2 *RNA-Seq*

Although Microarray had been the best gene quantification technology since the ninety decade, RNA-seq was consolidated as the most powerful and newest technology since the last decade [12]. As a natural evolutionary step in the gene quantification technologies, RNA-seq is gradually replacing the widespread use of Microarray. There exist many manufacturers that work with RNA-seq but, nowadays Illumina leads the RNA-seq sequencing technology market. Although its ap-

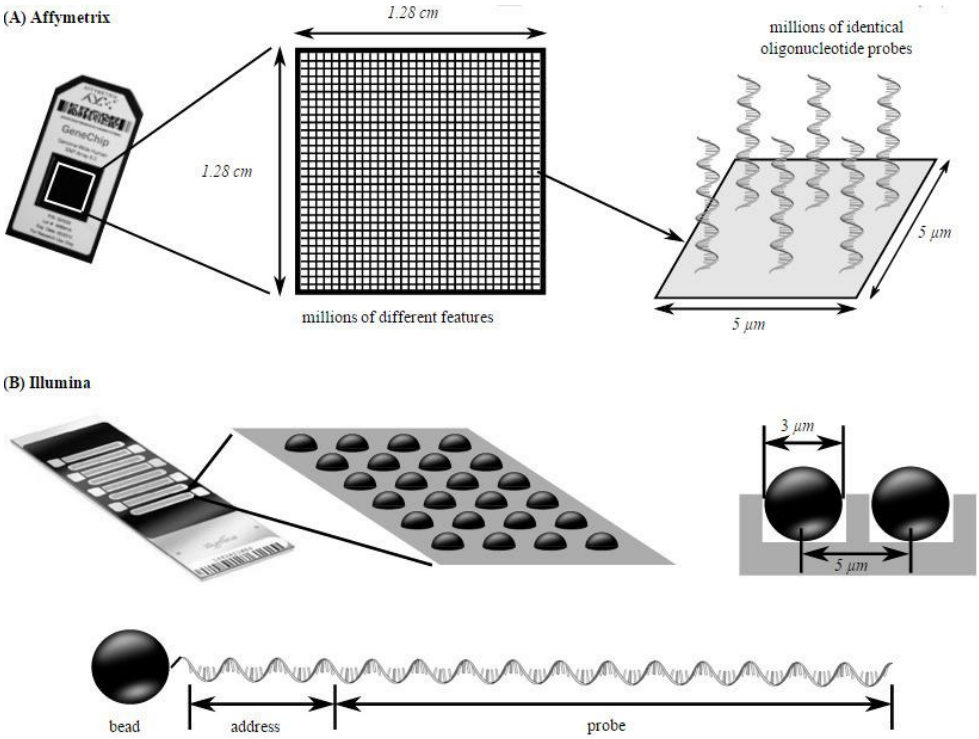

Figure 2.20: Affymetrix GeneChip and Illumina BeadArray representation

plication was originally intended for genomic transcription study, it also allows achieving a mapping between the levels of transcription and gene expression [94]. Thanks to this, it is possible to combine gene expression levels from both Microarray and RNA-seq. This is achieved through the quantification of the total number of reads that are mapped to each locus in the transcriptome assembly step. RNA-seq has many advantages in comparison with Microarray, which are explained herein.

RNA-seq offers an important number of advantages over Microarray that were clear described by Wang et al. [94], although the cost of RNA-seq experiments is still nowadays higher in some cases than Microarray technology's:

- RNA-Seq allows detecting the variation of a single nucleotide.
- RNA-Seq does not require genomic sequence knowledge.
- RNA-Seq provides quantitative expression levels.
- RNA-Seq provides isoform-level expression measurements.
- RNA-Seq offers a broader dynamic range than Microarray.

The classic RNA-Seq pipeline is showed at Figure 2.21. The standard process starts in the laboratory. Firstly, the RNA is extracted from a tissue or cell and then isolated. With the isolated RNA, the cDNA is synthesised and fragmented to create the sequencing library. Then, the library is sequenced with a depth between 10-30 million reads per samples by using a NGS platform (from Illumina in the most of cases). Normally, a SRA file is obtained as sequencer output, which contains all the sequenced fragments.

SRA: Sequence Read Archive

However, the process still continues once the reads are available, because they have to be aligned and/or assembled to a reference genome, in order to reconstruct the individual genome. With the aligned samples the counts can be estimated. Counts make reference to the number of reads that overlap a given feature like a gene. With the counts information, the equivalent gene expression can be found in order to achieve a way of discerning among states (i.e. Cancer and Control). Then, the samples follow a filtering and joint normalising process for acquiring a strong quality samples for the posterior statistical DEGs extraction model design.

There are three different types of generated reads, depending on the technology used for sequencing the samples as Figure 2.22 shows. Each type counts on some advantages and disadvantages. A brief explanation is given herein.

- Short-read cDNA sequencing: This is the most common read used for differential expression analysis. In this sense, Illumina short-read sequencers have generated more than 95% of the public RNA-seq worldwide data. This method has been consolidated as the standard method to detect and quantify gene expression, due to the high quality and performance for generating transcriptome data. This is the reason why this sequencing technology generates between 100-1000 time more reads per run than long-reads technologies and for that, it is the most suitable technology nowadays for differential expression analysis. However, this technology requires PCR amplification in all cases.

- Long-read cDNA sequencing: Currently, there are alternatives to Illumina short-reads sequencing. Thanks to the long-reads technologies from Pacific Biosciences and Oxford Nanopore, the necessity of short-reads assembly has been removed for the single-molecule sequencing of an individual. Furthermore, in comparison with short-reads technology, the ambiguity in the mapping of sequence reads is drastically reduced. In some protocols there is no needs of PCR amplification, accelerating the process. However,

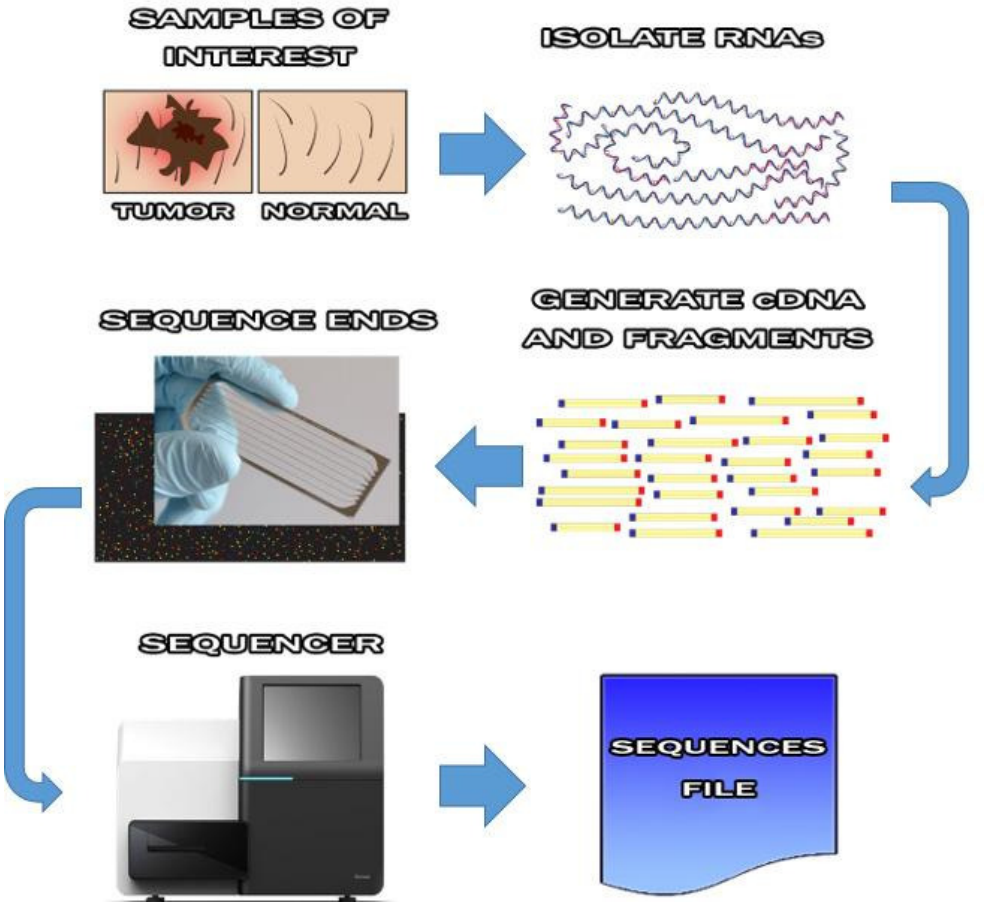

Figure 2.21: RNA Sequencing process. Through this process, the RNA is sequenced with the purpose of measuring the gene expression for transcriptomic analysis.

this technology is not suitable for differential expression analysis due to the low number of reads per run in comparison with short-reads technology. The main application of this technology is for isoform discovery, de novo transcriptome analysis and other complex transcriptome analysis.

- Long-read direct RNA sequencing: Oxford nanopore demonstrates that their technology can sequence RNA directly without PCR amplification or cDNA synthesis. Thanks to this technology, the generated biases are removed and the epigenetic information retained. As happen with Long-read cDNA, this technology is not suitable for differential expression analysis. However, Long-read direct RNA is commonly used in the same applications as Long-read cDNA and, also for detecting ribonucleotides modification. This is a very novel technology with a great future ahead.

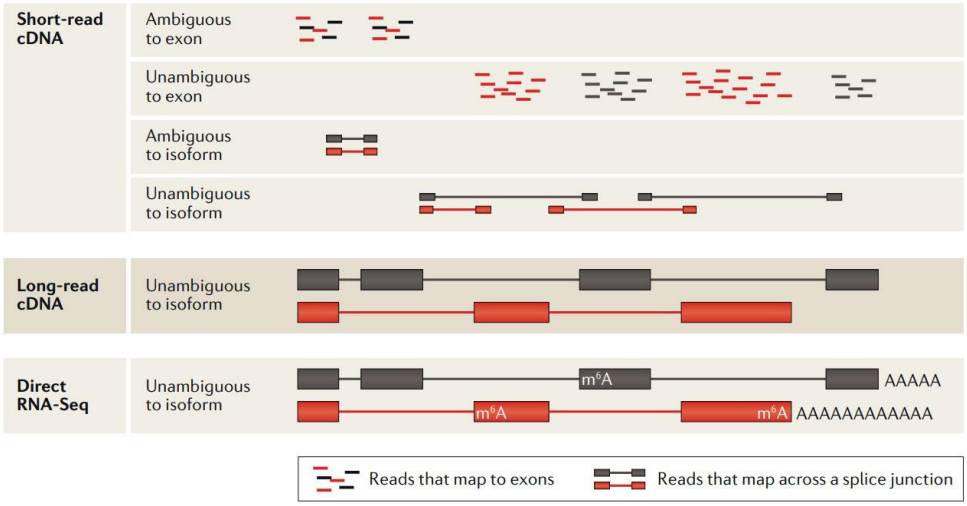

Figure 2.22: Different types of reads achieved depending on the considered RNA-Seq sequencing technology. Source: https://www.nature.com/articles/s41576-019-0150-2

For the development of this thesis, Illumina short-reads data were solely used because, as it was mentioned before, this type of reads are the best to carry out differential expression analysis. There are many researches that address complex RNA-Seq analysis for biomarkers detection in different cancer pathologies and rare diseases. In this sense, Kaczkowski, B. et al. found a set of candidate biomarkers with pan-cancer potential by analysing different tumour and healthy cell lines from different cancers [95]. On the other hand, Liang, J. et al. identified biomarkers for lung adenocarcinoma stages detection by using RNA-seq data [96]. However, RNA-seq has improved not only the cancer biomarkers detection but also the rare diseases diagnosis. For example, Kremer, L. S. et al. improve the genetic diagnosis of Mendelian disorders through the RNA-seq sequencing data [97]. Finally, scientists also take advantages of RNA-seq for fighting against brain diseases such as Alzheimer. Sutherland, G. T. et al. develop a study that shows the potential of this technology in comparison with Microarray for the transcriptomic analysis of this terrible disease [98].

2.4.3 *Microarray and RNA-Seq integration*

For some years now, integration of heterogeneous gene expression data is one of the main transcriptomic challenge to address. Normally, the available public series or datasets belong to a concrete transcriptomic platform or manufacturer, and contain a meaningless number of se-

quenced individuals to perform a robust study. In this sense, focusing solely on a concrete series or dataset leads to achieve results from a very tiny amount of population. In some cases, such as in rare diseases studies, they are made under these conditions due to the lack of available series or samples. Nevertheless, in the case of cancer studies, the number of available samples is usually higher and dispersed across different transcriptomic sources.

On this basis, integrating different datasets and series allows taking advantage of a larger number of samples and, consequently, improving the scale and the statistical significance of the results. However, this process is not trivial because each transcriptomic source has its own manner to quantify and calculate gene expression. For that reason, intrinsic biological information can be lost in the process. Because of that, several studies have been carried out to find what manufacturers and technologies have consistency among them, with the purpose of trying their joint integration. Concretely, M. Barnes et al. corroborate the correlation among Affymetrix and Illumina microarrays [88]. The classical Microarray analysis pipeline was focused on the extraction of different gene signatures for each series, and the obtaining of a final signature performing the intersection of all of them with a Venn Diagram. However, several methods appeared for carrying out a correct integration of microarray from Affymetrix and Illumina [99, 100], with the aim of exploiting the correlation between them. Afterwards, several tools were published to automatise this Microarray integration process such as VirtualArray [101], EMMA 2 [102] and FatiGO+ [103].

At present, Microarray has been gradually replaced with the arrival of RNA-seq, which is a clear evolution of its predecessor. Nevertheless, there are a considerable amount of public microarray without analysis yet. In this sense, those microarrays still are unexploited truthful sources of information. Under these considerations, the integration of RNA-seq together with Microarray is a way of exploiting the Microarray hidden potential. As Nookaew et al. explained, there is a high consistency between RNA-seq and Microarray, thus encouraging to continue using Microarray as a versatile tool for gene expression analysis [104].

With every passing year, the number of available new RNA-Seq samples increase while Microarray new samples is gradually disappearing. Due to that, although the integration of RNA-Seq and Microarray is a very promising process addressed in the develop of this thesis, in the near future the heterogeneous RNA-Seq integration data will be the most important and relevant. However, for multiclass experiments there still are a lack of RNA-Seq available samples that can be complemented with Microarray data. All these points of view and different integration

levels have been tackled along this doctoral thesis, producing different high impact publications [6–8, 105, 106]. In addition, the integration of different data sources can be supplied to the ML methods with the necessary quantity of samples for proper design of the predictive models, which usually are hard to gather only with only one technology. Although some advances were previously made in this line [107–109], in this thesis a novel pipeline to carry out the integration and the posterior ML assessment will be presented.

3

CLASSIFICATION AND FEATURE SELECTION MODELS APPLIED TO BIOINFORMATICS

CONTENTS

Machine learning revolutionised the way computers were able to perform advanced data analysis and infer patterns, behaviour and knowledge from data, in countless fields, not only those directly related with computer science. That is why nowadays there are uncountable applications applying machine learning techniques for predicting, classifying and decision making in all sciences and industry (such as automatic car driving, face recognition, etc.). Although many of those techniques exist since many years ago, the recent technological advances in the computational scope and the capacity of massive data generation, storage and processing have allow extracting the real potential of ML more thoroughly than ever.

With these considerations in mind, ML is the perfect way to tackle with the identification of optimal sets of biomarkers among all existing in the genome, in order to achieve proper early cancer diagnosis and treatments. In this thesis, a number of ML techniques are addressed and proposed, including well-known Supervised Classification Models and Feature Selection algorithms, with the purpose of assessing and identifying possible optimal DEGs candidates related to cancer diseases.

For a more detailed information about the concepts and explanations described along this Chapter, the reader can refer to the original scientist sources [**guyon2003introduction**, **kotsiantis2007supervised**, **bishop2006pattern**, **saeys2007review**, **larranaga2006machine**], together with all the references along the text that support this review.

3.1 SUPERVISED CLASSIFICATION MODELS

There are two types of classification or prediction models in ML: supervised and unsupervised learning algorithms. On one hand, the supervised algorithms know beforehand the labels/classes (output values) of both training and test input data. Based on this, the supervised model learns not only from the input data distribution itself but also from the labelled information. This type of learning models are ideal for problems in which the labels are known, and the model has to learn relations or differences among the existing classes. On the other hand, unsupervised learning methods do not know previously any information the information on the labels associated with the input data. In this sense, these algorithms try to learn patters or clusters existing in the input data. For that, the main goal for these models is to discover the underlying distribution in the data, with the purpose of learning complex and hidden relations. Output values are then associated with the input pattern or clusters according to experience or experts knowledge.

For the development of the experiments that support this thesis, supervised learning algorithms have been applied. When a differential expressed analysis is addressed, the label of each patient or individual used to perform the study is perfectly known. Thanks to this, it is easier to find relevant differences between the considered groups by taking into account these known labels. Then however, for the test phase of the development of the models, test data is totally left apart and operated as unseen data. This section reviews some of the most well-known supervised learning algorithms applied to bioinformatics data, and which will be applied in the experiments presented in this thesis.

3.1.1 *Naive Bayes*

NB: NAIVE BAYES

NB classifier is a classic supervised algorithm based on the Bayes Theorem [115]. This classifier assumes that the presence or absence

of a specific feature is independent of the presence or absence of any other feature for the studied variable. For example, each feature that define a person (Height, Two legs, Face, Hair, Sex...), contributes to the probability of being a person regardless of the presence or absence of the rest of features.

One of the main advantage for NB model is the low number of required training samples to estimate the means and variances for the classification. Due to the independence of the variable, it is not necessary to determine the covariance matrix. Under these considerations, this method is very powerful when the assumption of the independent features make sense. Moreover the computational cost for carrying out the training and test is very low in comparison with other more complex supervised algorithms.

As was mentioned above, NB is a conditional probabilistic model based on the Bayes Theorem. On one hand, the conditional probabilistic model establishes that a concrete event (A) will occur with a probability equal to the conditional probability of that event given a set of independent events (B_x) (Equation 3.1).

$$P(A|B_1, B_2, ..., B_n) \tag{3.1}$$

On the other hand, Bayes theorem determines the relation between the probabilities of two given events (Equation 3.2). By means of the Bayes Theorem is possible to obtain the probability of suffering from lung cancer depending on the fact of being smoker. On the other hand, and having information about patients who suffer from lung cancer, the probability of being smoker when suffering from cancer could be calculated.

$$P(B_i|A) = \frac{P(A|B_i) * P(B_i)}{P(A)} \tag{3.2}$$

NB classifier is designed taking into account the conditional probability along with the Bayes Theorem, as well as the naive concept. The naive concept means that all features or events (B_i) will have an independent influence over the event to study (A), regardless the presence or absence of any other events ($B_j \forall j \neq i$). Equation 3.3 shows the mathematical representation of NB probabilities identification.

$$P(A, B_1, B_2, ..., B_n) = P(A)P(B_1|A)P(B_2|A)...P(B_n|A) = P(A)\prod_{i=1}^{n} P(B_i|A) \tag{3.3}$$

In a problem to identify if an individual is a men or women given a set of characteristics (Weight, Height, foot size. i.e.), NB classifier calculates the possibilities for each class given each of the characteristics separately, assigning to the individual the sex which obtains the higher probability.

This classical classifier has been used in several research studies in a broad range of fields such as text classification [116], emotion recognition [117], network intrusion detection [118] or even for combining multi-species microRNA data [119], among many others.

3.1.2 *k-Nearest Neighbour*

k-NN: k-Nearest Neighbour

k-NN is one of the most widespread and old supervised classifiers due to its simplicity, powerful recognition capacities and low computational cost in comparison with other more complex methods [120, 121]. k-NN is based on classification by distance calculation, usually euclidean distance, from the sample to classify to the rest of samples. Once the vector of distances is determined and ascending reordered, the parameter k will indicate the number of nearest neighbours to the sample to classify to take into account. Thus, the classes of the samples in the first k positions of the vector of distances are consulted. k-NN will assign the majority class among the k-nearest neighbour to the samples to classify. A graphical representation of this method can be seen at Figure 3.1.

The euclidean distance can be calculated as Equation 3.4, where x is the sample to classify and $x^{'}$ represents one of the neighbours. Although euclidean distance is the most common distance for k-NN in the literature, others such as Manhattan, Chebyshev and Hamming distances are used too.

$$d(x, x^{'}) = \sqrt{(x_1 - x_1^{'})^2 + (x_2 - x_2^{'})^2 + ... + (x_n - x_n^{'})^2} = \sqrt{\sum_{i=1}^{n}(x_i - x_i^{'})^2} \tag{3.4}$$

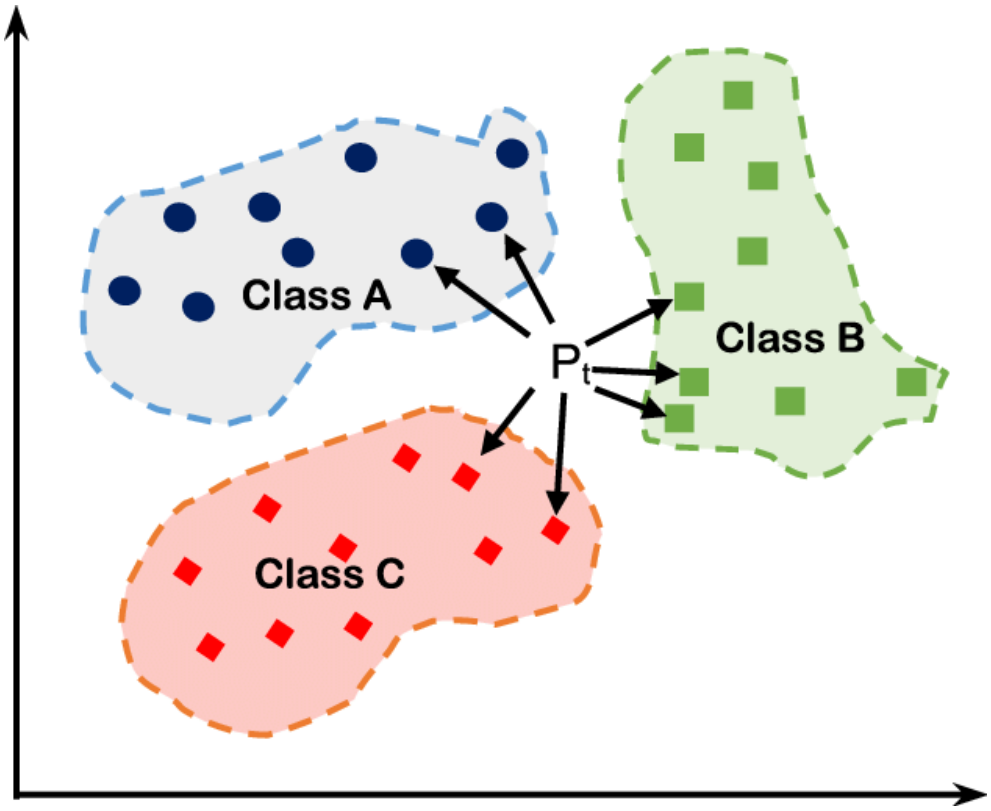

Figure 3.1: k-NN classifier graphical representation, where the number of nearest neighbour from the class B is higher than from the other two classes. Source: https://www.researchgate.net/figure/Example-on-KNN-classifier_fig1_331424423

k-NN is a supervised learning algorithm, which means that all samples are labeled and, this information is used to calculate the recognition rate of new unseen samples. Furthermore, k-NN is Non-Parametric, which means that no data distribution assumption is done, avoiding the threats of mismodeling the intrinsic data distribution. In addition, k-NN is an instance-base learning algorithm which means that the algorithm does not explicitly create a model for learning. Instead of creating a training model, k-NN chooses to retain the training instances which are utilised as knowledge for the posterior prediction step.

On a formal basis, according to a positive integer 'k', a new unseen sample 'x' and a closeness measure 'd', k-NN carries out the three steps listed herein. Moreover, the mathematical implementation for the calculation of the classes that correspond to each evaluated 'x' is given by the Equation 3.5.

1. The algorithm calculates through the entire training dataset the distance 'd' between the sample 'x' and each training sample. The

set A is formed by the 'k' points of the training dataset that are closest to 'x'.

2. Then, the algorithm calculates the conditional probability for each class for 'x', that corresponds to the fraction of points in 'A' with a concrete class label. As supplementary information, I(x) in the Equation 3.5 represents the function that returns 1 when the argument x is true and 0 otherwise.

3. At the end, the class that achieves the higher conditional probability is assigned to the sample 'x'.

$$P(y = j|x = i) = \frac{1}{k} \sum_{i \epsilon A} I(y^{(i)} = j) \tag{3.5}$$

k-NN is not only an algorithm for supervised learning but also used for feature extraction, dimension reduction, decision boundary, data reduction, data imputation, regression and outliers detection among others. Furthermore, k-NN has been applied in multitude of fields such as prediction of proteins location [122], authentication of smartphone users based on the way they walk [123], estimation of Mediterranean forest attributes [124], proactive detection of DDoS attacks [125],etc.

3.1.3 *Support Vector Machines*

SVM: Support Vector Machine

SVM is a complex classification and regression methodology for supervised machine learning published for the first time at 1995 by Cortes et.al [126]. It is included within the so-called kernel methods, a widespread and powerful set of algorithms with a higher computational cost in comparison with the methods explained above (NB and k-NN). SVM starts from the idea of trying to separate the points from different classes in a N-Dimensional space through the calculation of a hyperplane.

In practice, there are infinite hyperplanes that separate the classes in the space. Nevertheless, the objective is to find the hyperplane that obtains the best margins of separation between any point from the two classes and the hyperplane. This optimal hyperplane is obtained by convex optimization, given a set of hyperparameters, and the points defining it are also known as support vectors. A graphical representation of a hyperplane and the support vectors is given at Figure 3.4 in order to understand better SVM.

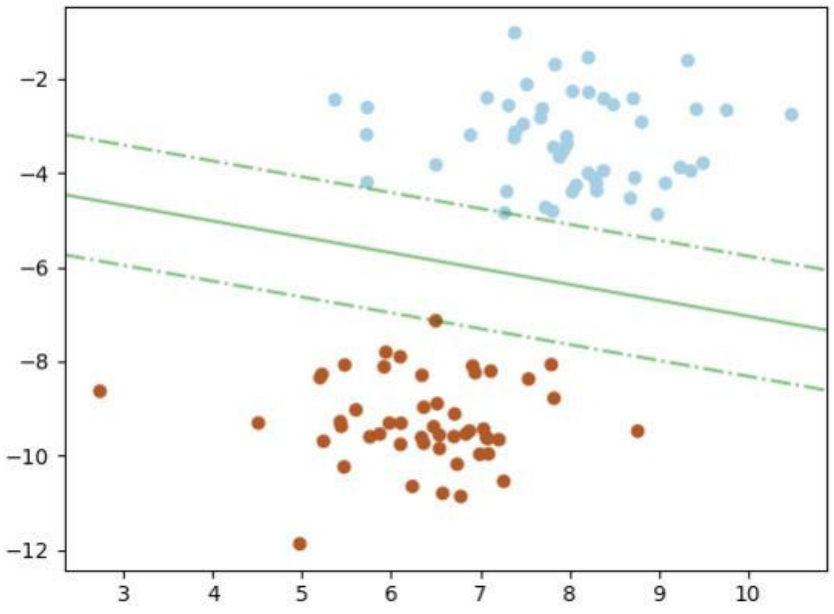

Figure 3.2: SVM classifier graphical representation, when there an hyperplane separating two classes with the most separated support vectors. Source: https://unipython.com/support-vector-machines-svm/

Hyperplanes are considered as decision boundaries with the purpose of performing the data points classification. Points falling on either side

of the hyperplane separated by support vectors can be assigned to one of the classes. Moreover, the number of features establish the dimension of the hyperplane. For example, for two input feature, the hyperplane is just a line. However, for three input feature, the hyperplane becomes a two-dimensional plane.

In SVM, the hyperplane is defined by a linear equation, for which, if the output of the linear function is positive, the point or sample to classify is identified with one class. Conversely, if the output is negative the sample is identified with another class. Support Vector get the values 1 or -1 in this equation, acting the range of values (-1,1) as the points falling within the margin of the hyperplane in the training data.

The maximisation of the margin between the data points is usually performed by means of the hinge loss function, which is represented at Equation 3.6. In it, the cost is 0 when both the predicted and the actual values have the same sign, otherwise, the loss value is calculated. Then a regularisation parameter(λ) is added to the cost function to find a balance between the margin maximisation and loss (Equation 3.7).

$$c(x, y, f(x)) = \begin{cases} 0 & if\ y * f(x) \geq 1 \\ 1 - y * f(x), & else \end{cases} \tag{3.6}$$

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+ \tag{3.7}$$

In order to find the absolute minimum, it is necessary to compute the gradients. For that, the partial derivatives with respect to the weights must be calculated (Equations 3.8 and 3.9). Through the use of those gradients, the weights can be progressively updated.

$$\frac{\partial}{\partial w_k} \lambda \|w\|^2 = 2\lambda w_k \tag{3.8}$$

$$\frac{\partial}{\partial w_k} (1 - y_i \langle x_i, w \rangle)_+ = \begin{cases} 0 & if\ y \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & else \end{cases} \tag{3.9}$$

On one hand, when the model predicts the classes without errors (no misclassification), the gradient is updated from the regularisation parameter (Equation 3.10).

$$w = w - \alpha * (2\lambda w) \tag{3.10}$$

On the other hand, when there is a misclassification, the gradient updating is carried out including the loss together with the regularisation parameter (Equation 3.11).

$$w = w + \alpha * (y_i x_i - 2\lambda w) \tag{3.11}$$

The most simple way to separate samples from different classes is through a line or hyperplane. However, in many cases the universe of features and samples to study is more complicated and harder to classify. This may be because several causes such as the presence of more than two classes to classify, a high number of features, which leads to the common fact that datasets that can not be linearly separated. To solve this problem, a Kernel function is usually implemented in SVM. A Kernel function allows operating in a higher-dimensional space, without computing the coordinates of the samples in that space. To achieve it, the inner products between the images of all pairs of samples in the feature space are computed. This operation is often computationally cheaper than the explicit computation of the coordinates.

In SVMs, the C hyperparameter establishes a trade off between the correct classification of training examples and the maximisation of the decision margin. Then, when a bi-class classification with a low number of features is addressed, it is common to use a linear kernel function. However, to deal with more complex problems, the gaussian kernel function is highly recommended. In this kernel, the σ hyperparameter manage the gaussian kernel width.

SVM has been used in multitude of applications, reaching very promising results in many cases. For that, this algorithm is one of the most standard methods for supervised learning in the literature. Among those applications there are a SVM kernel implementation for protein classification [127], SVM for recognising human actions [128], SVM for malware detection [129], even for the classification and visualisation of travel blog entries based on types of tourism [130], among many others.

3.1.4 *Random Forest*

RF is considered one of the top-used existing classifiers nowadays. This

RF: Random Forest

method combines individual decisions trees to make a RF, then each individual tree determines a class and, the final prediction by the forest will be the most voted class [131].

To understand better RF, it is primordial to know the basis of Decision Trees, as they are the foundations of RF. This method selects a set of features from the input data and tries to separate the data points taking into account the differences in the feature values existing within the data samples in each class. In the example showed at Figure 3.3, there are a set of numbers (five 1s and three 0s that are also the classes). Through a decision tree they will be separated using their features. The features selected to classify them are colour (red vs. blue) and if the number is underlined or not.

Although in real situations, input data will not be as easy as the example, the intrinsic logic of the Decision Tree is the same. Basically, at each node, the tree will query what feature will allow splitting the data in a form that the emerging groups are as different among themselves as possible.

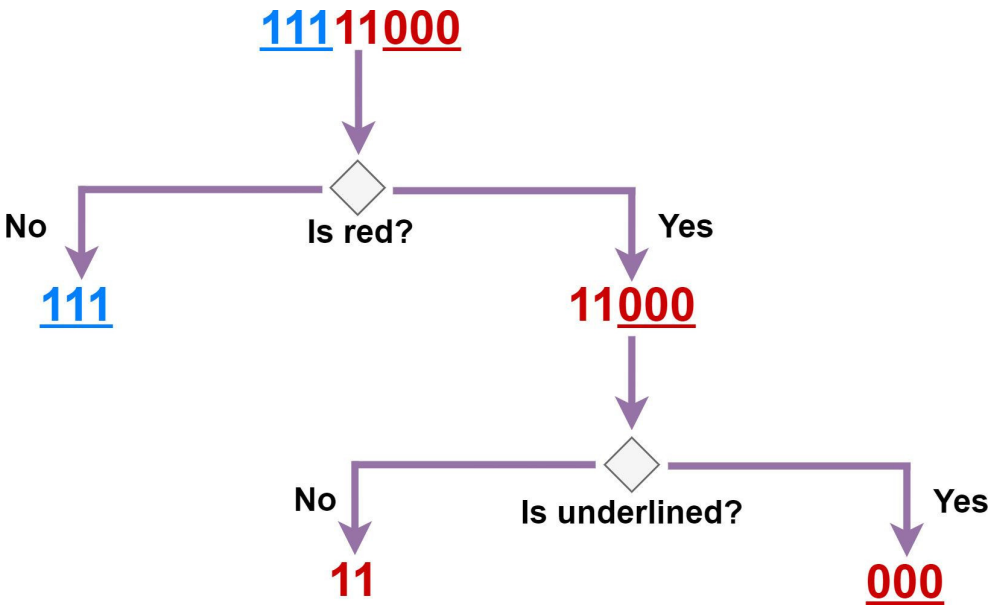

Figure 3.3: Decision Tree example in which a set of numbers is classified depending on their colour and underlining.

As it was mentioned above, RF is an ensemble conformed by many independent decision trees. The main advantage of RF relies on the large number of relatively uncorrelated models or trees that operate as a group, as they will outperform any of the individual models or decision trees. Working as individual entities, some trees may be wrong while other trees will be right. For that reason, when all the trees are

joined as a group, they are able to predict more accurate, using a voting scheme, protecting each other from their individual errors.

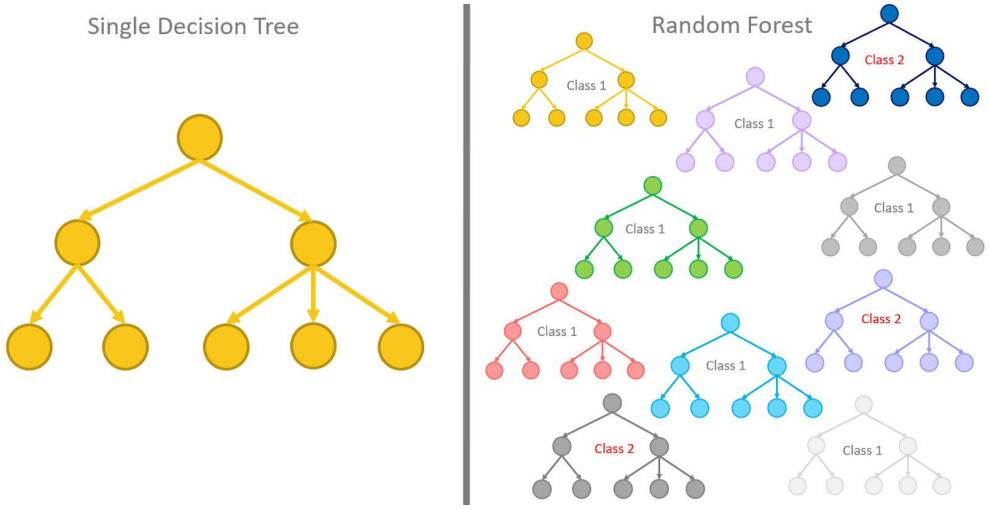

Figure 3.4: RF representation together with a single decision tree in order to see the comparison between them. Source: https://miro.medium.com/max/2000/0*YEwFetXQGPB8aDFV

Taking into account that RF is a predictor consisting of a set of M randomised and uncorrelated regression trees. Succinctly, at Equation 3.12 the calculation of the predicted value at the point x of a decision tree inside the RF is represented. In this equation, $m_n(x;\Theta_j)$ represents the predicted value at a certain point x by the j-th tree, where $\Theta_1,\ldots,\Theta_M$ are independent random variables, distributed as a generic random variable Θ, independent of the dataset D_n.

$$m_n(x;\Theta_j, D_n) \tag{3.12}$$

When all the trees that conform the RF perform their prediction at certain point x, a majority voting is required to determine the final prediction of the RF. For that, and for a binary problem, Equation 3.13 shows the evaluation along the M trees in the forest, in which at least the fifty percent of the trees have to vote the class 1 to assign this value to the final prediction, otherwise the value 0 will be assigned.

$$m_{M,n}(x;\Theta_1,...,\Theta_M, D_n) = \begin{cases} 1 & if\ \frac{1}{M}\sum_{j=1}^{M} m_n(x;\Theta_j, D_n) > \frac{1}{2} \\ 0 & otherwise \end{cases} \tag{3.13}$$

As happen with other ML techniques, RF also has a set of hyperparameters to tune in order to achieve a better fit of the models. The main

hyperparameter to tune is the number of trees created to training the model. For datasets and problems with high dimensionality, the number of trees to achieve a correct fit is usually higher than the number of trees in low-dimensional problems. However, there are some important hyperparameters to tune such as the maximum number of features considered for splitting a node, the maximum number of levels in each decision tree, the minimum number of data points allowed in a leaf node, or the minimum number of data points placed in a node before the node is split.

As it happens with the other supervised methods explained above, RF has been used for many applications along the science and other disciplines. In this sense, this approach was used for identifying SNPs predictive of phenotype [132], the prediction of protein–protein interactions [133], the imaging atmospheric of the Cherenkov telescope MAGIC [134] and for UAV remote sensing for urban vegetation mapping [135], amongst others.

3.2 FEATURE SELECTION

FS is a machine learning technique that consists in reducing the set of features employed in the design of the predictive model, in order to maintain only the truly relevant features for the addressed problem. Concretely, FS approaches are used for the following reasons:

- The simplification of the predictive models. Models with a more reduced set of features are easier to interpret and understand for the researchers or experts in the problem at hand.

- A reduced sub-set of feature also may dramatically shorten the training computation of the predictive models, especially in some specific methodologies whose complexity depends on the number of features used.

- The curse of dimensionality is a very common problem when datasets with more features than samples is handled. The dispersion within the input data space increases exponentially with the dimensionality of the input data, detracting from statistical significance to the results attained with any model used.

- Finally, FS usually is useful to enhance generalisation and minimising overfitting. Since predictive models with more irrelevant

or redundant features can generate models more prone to errors and values deviations in these features.

Along this section, three well-known feature selection algorithms will be briefly exposed here, which correspond to the feature selection methods applied in the experiments that support this doctoral thesis.

3.2.1 *Relief*

Relief is a well-known feature selection algorithm, whose main drawback is that it is highly sensitive to feature interactions [136]. Although the algorithm was originally created for 2-class classification problems, it was adapted too for multiclass classification. To perform the selection, Relief calculates a score for each feature, then the feature vector is reordered to create a ranking. Finally, the top scoring features are usually selected as the final sub-set of the FS.

For the score calculation, the algorithm detects the differences between the value in the nearest neighbour pairs of features. When a difference in the features values of a neighbouring pairs of features appears and they have the same class than the observed feature, the score for this feature decreases (hit). Conversely, when the difference in the features values of a neighbouring pairs of features appears with different class than the observed feature, the score for this feature increases (miss). A representation of Relief hit and miss given an observed instance is given at Figure 3.5.

Relief has the advantages of being not dependent on heuristics as well as having a low computational time. Nevertheless, the algorithm does not discern between redundant features, and a reduced numbers of training samples can lead to erroneous results.

The algorithm takes as input a dataset that contains n samples with m features, labeled through two known classes. Then, the algorithm starts with a weight vector of zeros (W) with a size equal to the number of features (m). The score calculation will be repeated k times in order to update the weights vector each iteration.

For each iteration, a features vector (X) from one random samples is taken along with the features vectors of the samples closest to X from each class, calculating the Euclidean distance or other well-known distance. As it was mentioned before, the closest same-class samples is known as near-hit, and the closest different-class samples is known as

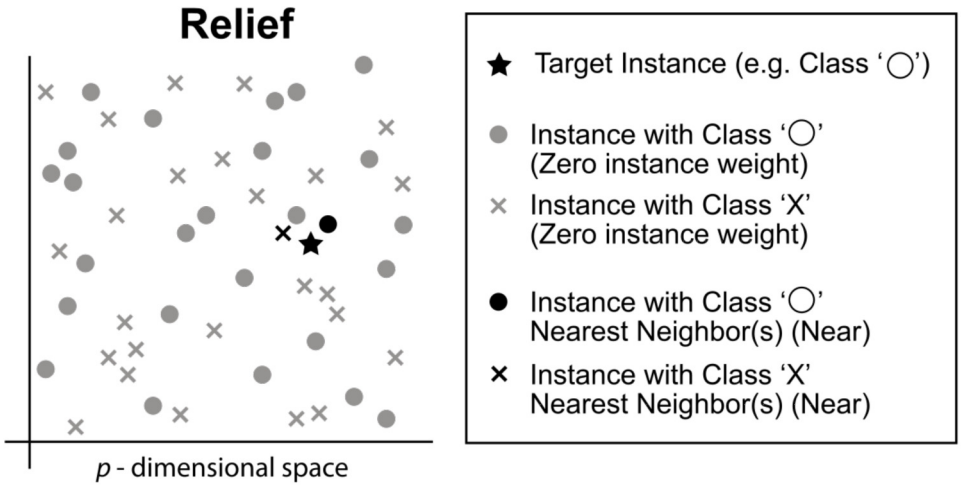

Figure 3.5: Relief near-hit and near-miss representation given an observed instance (Target Instance). Source: https://upload.wikimedia.org/wikipedia/commons/thumb/9/95/Relief_Wiki.svg/1200px-Relief_Wiki.svg.png

near-miss. Then, the Weight vector (W) is updated taking into account the near-hit and the near-miss, as it shown at Equation 3.14.

$$W_i = W_i - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \tag{3.14}$$

In this sense, the weight of a given feature decreases if it discerns from the same feature in closest samples of the same class more than closest samples of the other class, otherwise the weight of the observed feature increases.

Finally, once all the iterations are done, the Weight vector (W) is divided by the number of iterations k with the aim of calculating the relevance vector, selecting all those features that exceed a given threshold.

3.2.2 *minimum Redundancy Maximum Relevance*

mRMR: minimum Redundancy Maximum Relevance

mRMR is one of the most powerful and widespread feature selection algorithm among the scientific and ML community. The algorithm was originally developed by Peng et al. [137]. At its origin, mRMR was mainly designed to deal with the classification of DNA microarray data, with the purpose of reducing the extremely amount of genes or biomarkers candidates in comparison with the low number of samples usually available. However, the method has been used more extensively

in other fields and applications. Although mRMR has high complexity in computational terms and this complexity scales quadratically with the number of features and linearly with the number of samples, it also offers one of the best trade-offs between stability and accuracy as Brown et al demonstrates [138].

mRMR takes into account the importance or relevance of the features (genes) for a given classification task. mRMR creates a ranking of features based on their relevance to the classification, penalising also the redundancy among the features. The aim of the algorithm is to achieve the maximum relevance between a set of features X, and the class c but minimising the redundancy, taking into account the MI between the features. The MI measures the mutual dependence between the two variables or features. Equation 3.15 shows the calculation of the MI between a pair of features (A and B), which can be obtained if the marginal probabilities p(a) and p(b), and the joint probability p(a,b) are known.

MI: Mutual Information

$$I(A;B) = \sum_{b\epsilon B}\sum_{a\epsilon A} p(a,b)log(\frac{p(a,b)}{p(a)p(b)}) \tag{3.15}$$

Firstly, mRMR selects features such that they provide maximum relevance between the features set selected S and the class c. Equation 3.16 consists of the sum of the MI of all the features in S separately with the class c, divided by the number of elements of the feature vector S.

$$maxD(S,c); D = \frac{1}{|S|}\sum_{X_i\epsilon S} I(X_i;c) \tag{3.16}$$

Nevertheless. choosing features only according to the maximum relevance can bring a high redundancy, effect called also as colineality. Therefore, Equation 3.17 estimates the redundancy existing among the features in the subset S by summing the MI between those and dividing them by the cardinal of the features vector S.

$$minR(S); R = \frac{1}{|S|}\sum_{X_iX_j\epsilon S} I(X_i,X_j) \tag{3.17}$$

Finally, by combining the Maximum relevance and the Minimum redundancy equations, mRMR algorithm is designed. Equation 3.18 represents the junction of both criterion with is used in a greedy way

to iteratively select features forming a ranking, where S represents the set of currently selected features.

$$max_{X_i \notin S}[I(X_i; c)] - \frac{1}{|S|} \sum_{X_j \epsilon S} I(X_j; X_i)] \tag{3.18}$$

3.2.3 *Random Forest as Feature Selector*

Previously, RF was described as one of the most famous and powerful classifiers nowadays. However, RF is also used as feature selection method through a measure called feature importance, which is returned by the algorithm. RF was concretely used as feature selector for selecting biomarkers in Microarray data by Diaz et al. in 2006 [139], thus this method is used since several years ago. RF provides a good predictive performance in most of cases and easy interpretability. Focusing on the second advantage, the interpretability is provided as the method is straightforward to derive the importance of each feature on the different decision trees. Thanks to this, it is very simple to know how much each feature is affecting to the final decision.

To understand the feature importance, it is pertinent to remember the operation of RF, which consist in the creation of several decision trees, each of them designed over a random extraction of the samples and a random extraction of the features. For this reason, not every random tree uses all the features or all the samples, guarantying that the trees are uncorrelated and, hence less trended to over-fitting. Basically, each uncorrelated tree is a sequence of yes-no questions based on a single feature or a combination of them. At each node or question of a tree, the dataset is divided into 2 sub-sets, each of them contain samples that are more similar among themselves but different from the samples in the other sub-set. Consequently, the importance of each feature depending on how pure each of the sub-set is, understanding by pure the correct separation of the samples along the sub-set due to the feature influence.

The use of RF as feature selector by using tree derived feature importance is a very simple, fast and precise way of choosing features for the reduction of the dimensionality and noise for the final predictive model.

However, this method is based on impurity reduction and due to that, it is biased towards preferring features affecting more classes. Moreover, when the dataset contains correlated features, some of features can be

selected as predictor feature by the model, with no preference of one over the rest. Nevertheless, if one of them is used, the importance of the rest of features decrease significantly due to the impurity that these features can remove is already erased by the used feature. Consequently, these features will have a very low importance in comparison with the chosen feature.

3.3 MACHINE LEARNING FOR BIOMARKERS ASSESSMENT

The potential of ML has constantly increased in the last years due to the increment of both the data available and the computational performance of the current computing systems, including the almost unlimited potential that virtualisation and cloud computing provide to any research group or company at reasonable cost. For these reasons, although years ago ML practice seemed to be focused only on the computer sciences scope, nowadays they are widely used in practically every study field. That is the case with the biomarkers assessment in cancer, a very concrete application of ML inside the Bioinformatics and Computational Biology scope. In the fight against cancer, the search of possible new biomarkers or combinations of them is one of the most promising allies. Due to that, there are plenty of researches that propose and assess different candidate gene signatures for different cancer pathologies, rare diseases, Alzheimer, i.e.

Concretely, for gene expression analysis, the dimensionality of the experiments is usually very high due to the number of features involved. Taking into account that the human genome contains between 20.000-25.000 protein coding genes, it is moreover impossible to find and evaluate biomarkers manually or without applying any of the existing well-known statistical methods or ML approaches. In this sense, thanks to these techniques the traditional analysis of a very reduced set of biomarkers has turn into the possible analysis of the whole genome for discovering biomarkers whatever being their location inside the genome. For that, through the use of ML techniques, the possibilities of learning and detecting new biomarkers and hidden relations between them have increased remarkably. In addition, it is easier for biologists and clinicians to have a reduced and clear sub-set of candidate biomarkers with the aim of evaluating them at biological level to prove their real impact in the disease.

Although this thesis is focused only on transcriptomic biomarkers understood as DEGs, there are others biomarkers such as proteins or metabolites in which ML is applied too. For example, Swan et al. in 2013

perform a review of the use of ML for the classification and biomarker identification in post-genomics biology with proteomics data [140]. In this review, several well-known supervised learning classifiers and feature selection methods are applied in different proteomics issues, achieving very promising results.

Another relevant study were done by Abeel et al. in 2009, which address a robust biomarker identification for cancer diagnosis with ensemble feature selection methods by using SVM [141] approaches. For that, they used four microarrays, achieving a possible gene signature of a few tens of DEGs. The importance of a good FS implementation must be highlighted from this article, because they achieved an improvement of 15% in the classification accuracy thanks to the FS.

Biomarkers can be extracted even from images as Woo et al. shown in their article in 2015. They used human neuroimaging along with machine learning techniques to develop objective brain-based biomarkers of the neural functions and neuropathology that underlie chronic pain [142]. Furthermore, Azuaje et al. make use of ML techniques for cardiovascular biomarker discovery based on the combination of gene expression and functional network analyses [143].

Moreover, there are not only isolated studies but also tools and systems that try to automatise the biomarkers discovery. For example, Horng et al. proposed an expert system to classify Microarray gene expression data using gene selection by decision tree in 2009 [144]. Statnikov et al. created GEMS in 2005, which is a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data [145]. Even mRMR was originally designed by Peng et al. [137] for FS applied to biomarkers detection in DNA microarray data, as it was mentioned above.

As it has been explained along this section, the biomarkers detection is crucial for winning the battle against multifactorial and genetic diseases such as cancer. However, it is also very important not only to know those biomarkers but also how counteract their effect to make the disease diminish. In this sense, Vamathevan et al. have made a review of the applications of ML to drug discovery and development in 2019 [146].

To sum up, ML has a tremendous number of applications to help to understand the diseases like cancer and the way which they growth. In a near future, thanks to the medical, biological and computer sciences advances, the precision medicine will be a reality which will allow saving a lot of life from genetic and multifactorial diseases.

Part II

CASE STUDIES & CONCLUSIONS

4

METHODOLOGY & RESOURCES

CONTENTS

There are analysis in the literature that combine the differential gene expression analysis with ML techniques in different ways. However, those analysis are usually focused only on one type of cancer and in a bi-class approach, and make only use of a single dataset or series. In this sense, the development of this thesis is supported by the creation of an automatic pipeline to carry out complex differential gene expression analysis along with the posterior machine learning assessment, taking into account the integration of heterogeneous data sources. Furthermore, this automatic pipeline has been tested under different cancer pathologies and different number of classes (not only at bi-class level), achieving in all cases outstanding results. In this section, the different steps that conform the mentioned automatic pipeline will be in-depth explained. Moreover, as a colophon, the process to publish the automatic pipeline in the most important worldwide repository of Bioinformatics will be also detailed.

4.1 HARDWARE & SOFTWARE RESOURCES

The experiments that support this doctoral thesis have required a very high computational cost, specially for the RNA-Seq RAW data alignment. This RAW files usually need a huge storage capacity as each file contains a puzzle in which each piece is a fragment of the DNA of an individual. Because of that, it has become necessary to use HPC hardware. For the RAW files alignment two computer clusters have been used, reducing drastically the computational cost for aligning the samples and for counting the reads per genes for each individual. Then, for carrying out the biomarkers detection and analysis, R language was used along with Matlab for the ML assessment of the first version of the pipeline. Once the pipeline was finished, the ML step was implemented in R language too. All these steps were run in a Personal Computer prepared for complex and heavy analyses. The clusters and the personal computer characteristics are listed herein:

- BioATC: Computer cluster with a total of 19 computing nodes and, the following features per node:
  - 2 CPUs Intel Xeon E5520 (2.27 Ghz) (each processor with 4 cores, and 8 processing threads).
  - 16 GB of RAM.
  - Massive storage: SCSI Disk with 160 GB in the compute nodes. 2 TB of storage in the main node, available via NFS from the compute nodes.
  - Network: Gigabit Ethernet and Infiniband.
  - OS: CentOS 7.6.1810 64 bits.
  - Job Scheduling System: Slurm.
  - Monitoring: Ganglia.
- ATCBioSimul: Computer cluster with a total of 4 computing nodes and, the following features per node:
  - 2 CPUs Intel Xeon Silver 4110 8c (2.10 Ghz) (each processor with 8 cores, and 16 processing threads).
  - 32 GB of RAM.

CPUs: Central Processing Units
Ghz: Gigahertz
GB: Gigabyte
RAM: Random Access Memory
SCSI: Small Computer System Interface
TB: Terabyte
NFS: Network File System
OS: Operative System

  - 2 GPUs Asus Dual RTX-2080 A8G EVO 8 GB GDDR6 in the last compute node.
  - Massive storage: RAID with SAS disks (15 TB and 12 Gbps), and NAS (35 TB). Both available via NFS from the compute nodes.
  - Network: Gigabit Ethernet.
  - OS: Ubuntu 18.04.2 LTS 64 bits.
  - Job Scheduling System: Slurm.
  - Monitoring: Ganglia.

SAS: Serial Attached SCSI
Gbps: Gigabit per second
NAS: Network Attached Storage

- MSI GP62VR 7RF Leopard Pro: Personal Computer with the following features:
  - CPUs Intel Core i7-7700HQ (2.8 Ghz, 4 cores and 8 processing threads).
  - 32 GB of RAM.
  - GPUs Nvidia GeForce GTX 1060 3 GB GDDR5.
  - Massive storage: SSD with 256 GB and 2 hard disks with 1 TB per disk.
  - OS: Ubuntu 16.10 LTS 64 bits.

SSD: Solid-State Drive

4.2 ASSEMBLING AN INTELLIGENT DIFFERENTIAL EXPRESSION PIPELINE

Traditionally, the proposed methods for differential expression analysis in the literature do not include a ML learning step together with a biological enrichment process. Along this thesis, a whole pipeline not only for Microarray but also for RNA-Seq that includes RAW data pre-processing and quality treatment, heterogeneous integration, biomarkers detection, ML assessment and biological enrichment has been carefully designed. The main goal of this pipeline is to provide to the scientific community an automatic and integrated pipeline for performing all-in-one differential expression analysis under the same programmatic language (R) or environment. The pipeline has been

tested under different conditions and diseases and not only in one isolated experiment. This section contains an explanation for the different sub-steps that conform the pipeline.

4.2.1 *Heterogeneous Data Gathering*

The samples gathering is the first step to perform a differential gene expression analysis. This is a delicated process, because samples in bad condition can introduce noise and bias in the final results. In order to reduce possible deviations coming from the data, all the experiments of this thesis have been done by using Microarray and RNA-Seq RAW data with the aim of applying the same pre-processing strategies to all of them.

4.2.1.1 *Web-platform Databases*

To look for the samples, two of the most well-known public and controlled databases have been used. The first one is the public database NCBI/GEO, which has series from Microarray belonging to practically any type of cancer or multifactorial or genetic disease [147]. Moreover, NCBI/GEO stores series from RNA-Seq, although the amount of Microarray series is far larger than of RNA-Seq series. However, this database stores not only series coming from transcriptomic but also series from other different omics and technologies. The samples are not usually stored separately, instead, the samples are organised in series or platforms, which are groups of samples belonging to the same laboratory or experiment. It is very normal to find samples that have been previously treated with drugs, which can confound the biomarkers extraction and predictive models if some genes have been over-expressed or inhibited due to the effect of those drugs.

NCBI/GEO: National Center for Biotechnology Information / Gene Expression Omnibus

The other database is GDC Portal which has public and controlled samples depending on if the samples can identify an individual or not [148]. For example, BAM files are all of them under controlled access because a genome could identify an individual, however images and count files are public without any restriction. In this case, samples are separated by type of cancer, and are individually downloaded. The main advantage of GDC Portal is that samples are harmonised among them following the same alignment process, which ensure the outstanding quality of all samples. Furthermore, GDC Portal counts

GDC: Genomic Data Commons

BAM: Binary Alignment Map

with the clinical information that can be used to study the disease progression and lifestyle of each individual.

4.2.1.2 *Microarray RAW data processing*

Microarray RAW data are usually stored at .CEL files for Affymetrix microarrays and .TXT for Illumina microarrays. Affymetrix and Illumina also organise in a different manner the information inside the RAW files. Taking this into account, the easiest way to pre-process the RAW files from both manufacturers is through the use of the specific R packages for each of them. In the case of Affimetrix, the package affy can open .CEL files to extract the gene expression values from the information of an Affymetrix Microarray [149]. For that, the package counts with the RMA algorithm, which converts the intensities in the .CEL files into an expression measure in log base 2 scale [150].

RMA: Robust Multi-Array Average

Then, for Illumina microarrays, the .TXT files can be pre-processed with the lumi R package [151]. As happen with Affymetrix RAW files, the Illumina RAW files contain the intensities of the probes from the microarrays. This intensities can be translated into gene expression by using the function lumiExpresso from lumi package, which implements also a log base 2 scale. Keeping the same transformation scale between Affymetrix and Illumina Microarray is very important to achieve a correct integration between both technologies.

4.2.1.3 *RNA-Seq RAW alignment*

RNA-Seq RAW pre-preprocessing is a more complex process, involving more steps than for Microarray RAW pre-processing. BAM files are conformed by many fragments of DNA from an individual. Those fragments are unordered and need to be realigned by using a reference genome, which is a compendium of genomes from different volunteers that represents the standard human genome. This process takes a lot of time and is very heavy in terms of computation and, due to this reason computer clusters were used to carry out the alignment process. Furthermore, a GTF file is required depending on the aligner used for the alignment. It is a widely used format for storing gene annotation and structures. In the case of the experiments done for this thesis, the GTF file was necessary due to the Tophat2 [152] and Hisat2 [153] aligners, which were used for the RNA-Seq RAW alignment. Once the BAM files are created, it is required to count the number of times or reads per genes in the aligned genome. For that, Htseq-counts

GTF: Gene Transfer Format

allows counting those reads and create a .TXT files which contains the reads per gene [154]. Finally, through count files, the equivalent gene expression values can be calculated.

4.2.2 *Pre-processing*

Dataset pre-processing is the most sensitive step in any pipeline in which biological data is involved, due to the possibility of losing intrinsic information, or of introducing noise or bias inside the samples. To avoid as far as possible these problems, the pre-processig step has to be restrictive and robust, applying outliers detection and quality analyses strategies. Even so, when an integration is to be carried out and several heterogeneous data sources turn into one super-dataset.

4.2.2.1 *Outliers detection*

It is very common to find samples which differ from the numerically trend of the samples inside of a dataset or series, abruptly deviating the mean. These samples are considered as outliers and they have to be removed from the rest of samples. The first step to achieve a great concordance among the series to integrate is the outliers detection and removal with the aim of removing those samples that could introduce irregular values. This detection is usually performed by using statistical analyses which allows detecting those unusual values.

ArrayQualityMetrics R/Bioc package includes a set of statistical approaches with the objectives of catching possible outlier from different perspectives [155]. Mainly, the package computes four different test to find outliers. The distances between arrays to search for possible samples far from the others. The Kolmogorov-Smirnov K statistic which measures the probability that a concrete univariate series is drawn from the same parent population as a second series. The density of the standard deviation to measures any possible samples that introduce abnormal modifications into the standard deviation. At last, the MA plots are also represented through the calculation of Hoeffing's D-Statistic, which computes the differences among the density distributions of each array with a reference array that consists of the median across arrays. The outliers detected by those methods are removed to ensure the correct harmony among samples.

Although ArrayQualityMetrics was designed only for Microarray quality analysis, the package has been adapted to RNA-Seq in KnowSeq in order to compute the quality analysis regardless the transcriptomic technology.

4.2.2.2 *Data Suitability*

Having homogeneous data would be the ideal scenario when an integration is carried out. Nevertheless, each of the different manufacturers and technologies have their own particularities in the data, which must be removed or minimised to reach the integration. Firstly, all datasets have been pre-processed depending on the manufacturers they belong to, as it was previously mentioned. Then, a logarithmic transformation must be applied to those series or datasets which are not in log_2 scale. Furthermore, the bit depth of each series must be also equalised in order to keep the expression ranges of all series in the same dynamic range, avoiding the rise of erroneous DEGs. Finally, it is very common that each manufacturer has its own gene annotation, because of that genes have different identification between series from different technologies, making impossible their integration. It is required that annotations from different sources will be translated into standard existing annotations such as Entrez [156], Ensembl [157] or Gene Symbols (HUGO) [158]. After all these adequacy strategies, heterogeneous data should be ready to integrate them.

HUGO: HUman Genome Organization

4.2.2.3 *Batch Effect Treatment*

The presence of batch effects is one of the most problematic situation in omics data analysis, when there are different data sources involved. Batch effect makes reference to an intrinsic deviations of the samples given a set of factors (the laboratory, the lab technicians, the sequencers or even the environmental factors). Taking this into account, each series or datasets can suffer from a different batch effect among their samples. There are many debates on this topic because of it is very difficult to know if those deviations are due to batch effect or, otherwise biological deviations. Furthermore, even if batch effect is located, it is a challenge to truly know if data has been really corrected and, consequently, batch effect removed [159]. In this sense, there are different batch effect removal strategies to apply depending on if batch groups among samples are known or unknown.

When batch groups are known, the batch covariate can be calculated, in this case ComBat method is the most widespread strategy to remove batch effect. ComBat adjusts batch effect by applying parametric or non-parametric empirical Bayes frameworks following the methodology described by Johnson et al. in 2007 [160]. To carry out this correction, estimations taking into account several factors are calculated for each genes in each batch. The estimations are usually simple measures such as mean and variance. Those measures are estimated by means of extracting information through multiple genes with similar expression in each known batch.

However, when there are public datasets or series involved in a research, it is very difficult to know the batch groups because of a lack of information about the series creation and sequencing. Under these conditions, there are well-known methods to estimate the influence of batch effect in the samples. The chosen method in this thesis is the SVA algorithm to find surrogate variables [161]. The aim of this method is to remove all unwanted variations, protecting the contrasts due to the classes to compare. This is translated into the extraction of features that are truly different between groups, removing all common latent variations. Nevertheless, there are other methods such as RUV, also very useful for removing unwanted variations when there is no predefined factor of interest. It is also the case in which a normalisation is required without knowing which factors of interest will be studied. In this sense, RUV corrects the gene expression by estimating and removing the unwanted variation, without removing the unobserved variation of interest [162].

SVA: Surrogate Variable Analysis

RUV: Remove Unwanted Variation

4.2.2.4 *Heterogeneous Transcriptomic Integration*

Ensuring data quality for the integration requires an enormous effort to avoid losing biological information. Once these step are correctly performed, heterogeneous data sources are ready to be integrated. There are two important steps for an appropriate integration of heterogeneous data: data merging and normalisation.

The first step, data merging, requires the correct individual pre-processing of all available series in the study. Afterwards, heterogeneous integration was carried out using the *merge* function from the base R package.

The second step in the integration is the normalisation of the integrated dataset. To perform this normalisation, the *normalizeBetweenArrays* function was used. This function is based on quantile normalisation and

allows to remove any possible deviation or variation among different arrays or datasets. Concretely, the method selected for the normalisation in this function is the “Aquantile" that ensures that the A-values (average intensities) have the same empirical distribution across arrays, leaving the M-values (log-ratios) unchanged.

These tasks are essential in order to achieve a correct normalisation of the biological data and its subsequent processing. The integration is indeed a critical process in the study as if the merging or normalisation are not properly done then the extracted DEGs would be erroneous. This would introduce confusing values in the research, that would lead irretrievably to a misleading selection of DEGs and subsequent erroneous outcome of the machine learning process.

4.2.3 *Biomarkers Detection*

Once data is already pre-processed and/or integrated, the biomarkers detection can be carried out with the security of having a high quality dataset, which will ensure the correct DEGs candidates extraction. It has to be highlighted that through this process, the curse of dimensionality is partially avoided because this step acts as a FS, removing those features without enough information to discern between the addressed states (Tumour vs Normal i.e.). For this reason, threshold values involved in this step have to be wisely chosen to achieve a correct number of features with discerning potential.

In the literature, limma R package has been postulated as the most important tool for performing differential gene expression analysis [163, 164]. Limma was originally designed only for Microarray, however, over the years the package has been updated to support also RNA-Seq. Normally, DEGs are selected depending on their $Log_2FoldChange$ and their $P-value$ from limma output table. Fold change is a measure that describes how much a value changes between two different observation. For example, an initial value of 10 and a final value of 20 corresponds to a fold change of 2. Nevertheless, when a gene in a class is inhibited, its expression will be between 0 and 1 in comparison to the other class, and, if this gene is over-expressed in one class, its value will range from 1 to infinity. With the aim of equalising this situation, $Log_2FoldChange$ (LFC) is used instead of Fold Change, which allows representing inhibited genes with minus sign and over-expressed genes with plus sign. Equation 4.1 represents the $Log_2FoldChange$ formulation given two observation A and B. Otherwise, $P-value$ represents the threshold for considering a gene as a DEGs taking into account the FDR, which

LFC: Log-Fold Change

FDR: False Discovery Rate

represents the expected numbers of type I errors. These errors occur when the null hypothesis is incorrectly rejected (False positive).

$$LFC = log_2(B - A)/A \tag{4.1}$$

4.2.4 *Machine Learning Assessment*

Thanks to ML techniques, new biomarkers have been found and assessed over the last years for different multifactorial and genetic diseases. As it was mentioned at Section 3.3, it has become impossible to find in a manual way truth biomarkers among the massive quantity of transcriptomic generated data. In this sense, ML has allowed to extract and learn concrete and truthful biological information from a huge dimensionality space, providing a powerful tool for scientists and clinicians. Then, those biomarkers previously selected with ML can be corroborated in the laboratory in order to determine their real biological impact.

4.2.4.1 *Feature Selection*

Finding a reduced sub-set of DEGs with the capability of discerning among addressed states can be a challenging task. In this sense, FS techniques are usually implemented to obtain a reduced sub-set of DEGs that diminishes the model complexity while practically preserving the predictive model results, also avoiding possible overfitting. All the experiments in this thesis include a FS process that allows obtaining the final sub-set of DEGs presented for each cancer pathology addressed. Thanks to this process, a small set of DEGs candidates, even for multiclass experiments, were achieved and assessed obtaining in all cases outstanding results and cancer related DEGs.

4.2.4.2 *Predictive Model Implementation*

Once DEGs are obtained, they become potential biomarkers to discern among the states to study. For that, those DEGs have to be evaluated depending on their gene expression values. If those DEGs have truly differences between the expression values among the studied classes, they would have the desired discernment potential for these analysis. This is where predictive models are applied with the aim of carrying out

an evaluation of those DEGs, learning relevant information from them and using this information to predict unseen samples. Furthermore, the more available samples or observations, the larger generalisation capability the predictive models will have.

In order to accomplish this task, it is recommended to train the predictive models following a CV strategy. Concretely, for this thesis a k-fold CV was implemented, which means that the training dataset is splitted into k sub-training sets, leaving the rest of samples for validation. Thanks to this technique, all the training samples are used for both training and validation at least once. Finally, at the end of the validation process, there are as results as k iterations, which will be used to decide the final sub-set of DEGs for the test process.

The final sub-set of DEGs chosen in the previous validation process is used to training the final model along with the whole training dataset and, also for testing the model with unseen samples kept from the beginning only for this step.

Keeping a robust methodology ensures the correct DEGs assessment. Due to that, the addressed problems in this thesis have been carefully studied, warrantying the ML good practices in the pipeline.

4.2.5 *Biological Enrichment*

The aim of this type of analyses should be not only the search and assessment of candidates biomarkers, but also their biological enrichment in order to give to the results a truth biological sense. For this reason, all the experiments are accompanied by an exhaustive literature DEGs enrichment with references that related those DEGs with the specific cancer addressed in each study.

However, the pipeline has been expanded with new biological enrichment functionalities when it has been encapsulated under the KnowSeq R/Bioc package. Understanding the real impact of DEGs needs more than the appearing relation with cancer in the literature. For example, the Gene Ontology enrichment allows retrieving DEGs biological location and functionalities. Determining affected pathways by DEGs is also very useful in order to see how the gene expression variation affects to adjacent genes in the same pathway/s, which could lead to bifurcations and activation of different biological processes (Apoptosis, Angiogenesis, i.e.). Finally, finding a score that relates DEGs with a concrete cancer taking into account different factors (RNA expression,

affected pathways, literature evidences, i.e.) is possible and very useful to know and quantify this relation. All this improvements to the biological enrichment process allow achieving a real biological overview of those DEGs based on the existing information of them in different organisations and biological databases. As it was mentioned above, all these processes have been added to the automatic pipeline through our public KnowSeq R/Bioc package.

4.3 PUBLISHING A BIOCONDUCTOR PACKAGE

Once the pipeline was finished, the idea of encapsulating it in a tool publicly available for the scientist community emerged. For Bioinformatics, R language has been postulated as the most used language in the literature and, taking this into account the pipeline was designed and implemented under a R package (KnowSeq). Moreover, to maximise it diffusion Bioconductor repository was chosen to publish KnowSeq, as it stores the most well-known Bioinformatics R packages [165].

To achieve the Bionconductor quality requirements, a set of severe considerations were addressed for the package design. All the functions must be correctly documented and the examples have to run correctly, furthermore all the inefficient structures and functions were replaced following the R design guidelines. Finally, the package documentation was iteratively revised until its quality was certified by Bioconductor.

To submit KnowSeq, a public Github repository was done with the contribution request to the Bioconductor contributions Github repository. The issue to follow the KnowSeq Bioconductor revision is public and can be read by following this link: `https://github.com/Bioconductor/Contributions/issues/1121`. In this issue, Bioconductor staff and the main author (D. Castillo-Secilla) interact until KnowSeq passes all the tests to be accepted for being published in Bioconductor. The KnowSeq user manual is included in the Appendix B of this doctoral thesis as well as in Bioconductor repository (`https://bioconductor.org/packages/release/bioc/vignettes/KnowSeq/inst/doc/KnowSeq.pdf`).

5

BREAST CANCER PROFILING BY INTEGRATING HETEROGENEOUS TRANSCRIPTOMIC PLATFORMS

CONTENTS

The present chapter is a reorganised and extended version of the published manuscript "Integration of RNA-Seq data with heterogeneous Microarray data for breast cancer profiling" [6]. In this study, the first version of the integrated intelligent pipeline was presented, obtaining DEGs for differentiating between a breast cancer cell line and a normal cell line. All the samples used for the study coming from Microarray and RNA-Seq are publicly available at NCBI/GEO.

5.1 BACKGROUND

Nowadays, public repositories containing large Microarray gene expression datasets are available. However, the problem lies in the fact that microarray technology is less powerful and accurate than more recent Next Generation Sequencing technologies like RNA-Seq. In any case, information from Microarray is truthful and robust, thus it can be exploited through the integration of Microarray data with RNA-Seq data. Additionally, information extraction and acquisition of large number of samples in RNA-Seq still entails higher costs in terms of time and computational resources than Microarray. Along this Chapter, a new model to search a candidate gene expression breast cancer signature through the integration of heterogeneous data from different microarray and RNA-Seq series is proposed. Furthermore, a classification method is carried out in order to test the robustness of the DEGs when unseen data is presented for diagnosis.

WHO: World Health Organisation

Breast cancer is one of the five most dangerous cancers in the world, showing a high mortality rate according to WHO and being the cancer with the highest impact among the female population. Calculations have shown that about 1 in 8 women are diagnosed with breast cancer during their lifetime. There are some risk factors that promote breast cancer such as age, genetic disposition, previous diagnosis, previous non-cancerous (benign) breast lump, overweight, i.e. Concretely, in 2017, a total of 611.625 died due to breast cancer.

DCIS: Ductal Carcinoma In Situ

Normally, Breast cancer is divided into two subgroups: non-invasive breast cancer (carcinoma in situ) and invasive breast cancer. The first one is found in the ducts of the breast (DCIS) and it has not spread into the breast tissue surrounding the ducts. The other one occurs where the cancer cells have spread through the lining of the ducts into the surrounding breast tissue. Invasive breast cancer is the most common type of breast cancer.

Nowadays, many breast cancer diagnosis are performed when a patient presents several related symptoms, thus increasing the mortality risk. If the cancer has spread, treatment becomes more difficult, and generally the chances of surviving are significantly lower. However, cancers that are diagnosed at an early stage are more likely to be treated successfully. Therefore, it is primordial to find biomarkers that allow an early diagnosis of breast cancer.

There are many breast cancer transcriptomic data publicly available stored in Microarray. Those microarrays belong to experiments already

available and there are also even a high number of them that have not been analysed so far. Those series might have information that could reveal important facts and candidate biomarkers. In any case, there is no doubt that although RNA-Seq is the present of the transcriptomic profiling, it can also take advantage of the available data from Microarray technology. As Nookaew et al. explained, there is a high consistency between RNA-Seq and Microarray, thus encouraging to continue using microarray as a versatile tool for gene expression analysis [104].

The aim of this research is the search of possible breast cancer biomarkers, taking into account patient and control samples stored at NCBI/-GEO. On one hand, the training dataset is conformed by 108 microarray samples, 65 samples from Affymetrix and 43 from Illumina technologies and 24 RNA-Seq samples from Illumina Hi-Seq. On the other hand, a test dataset has been also designed due to the necessity of achieving samples that have never been seen in the DEGs extraction and validation processes. This test set is formed by 120 samples of microarray (108 of Illumina and 12 of Affymetrix) as well as 6 samples of RNA-Seq from Illumina Hi-Seq. Most of the previous studies in the selection of biomarkers perform this process through statistical tools over a given dataset and a given technology. However, this research takes an innovative step forward by combining different datasets and Microarray technologies with RNA-Seq data.

5.2 INTELLIGENT BREAST CANCER PIPELINE METHODOLOGY

5.2.1 *Breast cancer data gathering*

The gathered samples for this research coming from two different cell lines. A cell line can be defined as a standard cell culture which will proliferate indefinitely inside an appropriate medium and space. For control samples MCF10A cell line has been selected, which is a healthy non-tumorigenic epithelial cell line. Otherwise, MCF7 line cell has been chosen for cancer samples, as it is considered a breast cancer cell line. To achieved them, a widely searched through the NCBI-GEO platform have been done with the aim of finding series belong to the selected cell lines from Microarray and RNA-Seq.

Once the requirements for selecting the desired samples were established, an exhaustive search of Affymetrix and Illumina series was carried out for microarray data. On the other hand, RNA-Seq data was selected from Illumina HiSeq technology. Only public series containing

the above-mentioned cell lines were selected. Table 5.1 summarises the selected series for this study. As can be seen, the NCBI GEO database offers a larger availability of Microarray data in comparison with the number of RNA-Seq samples. Two different integrated sets have been designed, one for training predictive models, and the other for their test, both containing microarray as well as RNA-Seq samples. These series are publicly available at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=S.NAME` where S.NAME is the name of each series at NCBI GEO.

Table 5.1: Description of the training and test series considered with number of samples/outliers.

TRAINING SERIES					
Series	Platform	Technology	Quality Samples	Excluded Outliers	Samples Origin
GSE52712	Affymetrix	Microarray	19	1	Manchester (UK)
GSE40987	Affymetrix	Microarray	10	0	Boston (USA)
GSE52262	Affymetrix	Microarray	16	0	Houston (USA)
GSE12790	Affymetrix	Microarray	20	1	San Francisco (USA)
GSE46834	Illumina	Microarray	8	0	New York (USA)
GSE68651	Illumina	Microarray	35	1	Southampton (UK)
GSE74251	Illumina	RNA-Seq	12	0	Philadelphia (USA)
GSE74377	Illumina	RNA-Seq	12	0	Iowa (USA)
TOTAL	**Integrated**		**132**	**3**	
TEST SERIES					
Series	Platform	Technology	Quality Samples	Excluded Outliers	Samples Origin
GSE75292	Illumina	Microarray	6	1	Goyang (South Korea)
GSE29327	Affymetrix	Microarray	6	0	South San Francisco (USA)
GSE30931	Illumina	Microarray	12	0	Goettingen (Germany)
GSE48398	Illumina	Microarray	36	0	Texas (USA)
GSE35928	Affymetrix	Microarray	6	0	Piscataway (USA)
GSE57339	Illumina	Microarray	12	0	New Haven (USA)
GSE45715	Illumina	Microarray	42	0	Miami (USA)
TOTAL	**Integrated**		**126**	**1**	

5.2.2 *Microarray DEGs extraction*

As it was mentioned at Chapter 4, Microarray analysis includes several steps to ensure the correct data treatment. Figure 5.1 outlines the Microarray data analysis pipeline. As a reminder, RAW files from both Affymetrix and Illumina are subjected to a quality analysis together with an outliers detection and removal. This quality analysis also contains the intra-series normalisation along with the batch effect treatment. Before integrating all the Microarray series, it is required to retrieve the gene annotation for each technology in order to translate and unify the name of the genes across series. Following the integration strategy also defined at Chapter 4, Microarray data from the two different technology are combined, achieving a Microarray integrated dataset. Finally, DEGs extraction is carried out, acquiring a set of Microarray DEGs.

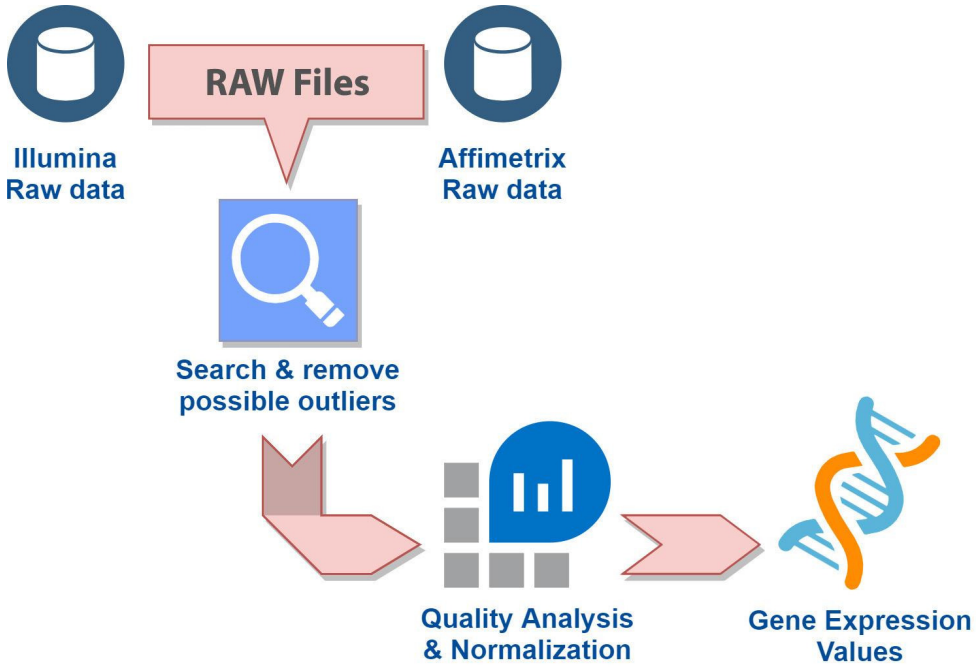

Figure 5.1: Microarray gene expression pipeline followed to extract and pre-process the microarray RAW data in this study.

5.2.3 *RNA-Seq DEGs extraction*

The proposed pipeline to treat RNA-Seq RAW data has been followed for the gene expression values retrieval. This pipeline has been explained at Chapter 4 but it is also summarised here. A graphical representation of the pipeline is shown at Figure 5.2. Starting from the SRA original files, FASTQ files are obtained by using SRA-Toolkit. Then, the reference genome and the GTF files are required to carry out the alignment process to obtain the BAM files by using TopHat2. Consequently, with the aligned files the counts files can be acquired and the gene expression values calculated. It is very important at this point the application of a quality analysis together with a batch effect removal strategies. At the end of the RNA-Seq pipeline, the dataset formed by all the RNA-Seq samples was used to retrieve a set of RNA-Seq DEGs.

5.2.4 *Intelligent Integrated Pipeline*

As a first approach of this new intelligent pipeline, an extension of the classical gene expression data analysis pipeline is proposed in this research. This pipeline counts with two main improvements. On one hand, this pipeline integrates data from both Microarray and RNA-Seq

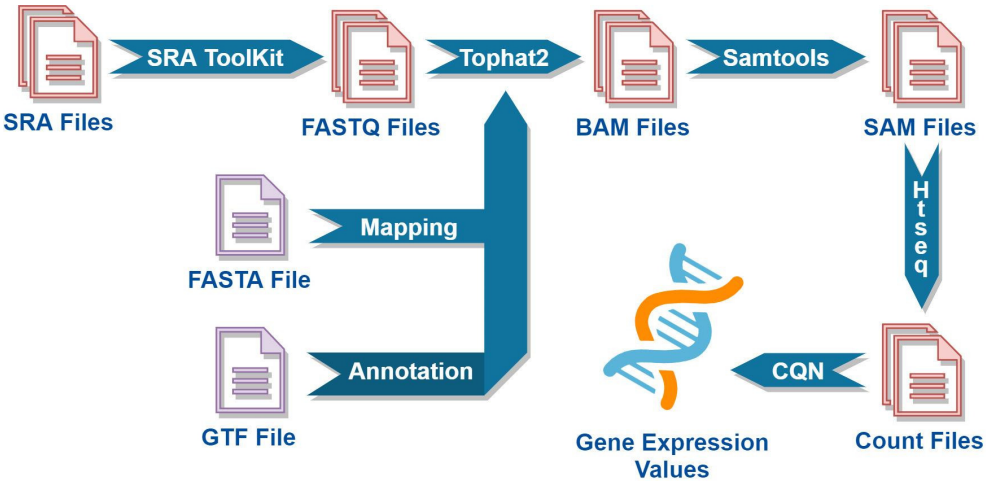

Figure 5.2: RNA-Seq gene expression pipeline implemented for extracting gene expression values from RNA-Seq RAW data.

technologies. On the other hand, a ML assessment through a feature selection and classification process by using separated training and test datasets was performed. Figure 5.3 represents the workflow of this research, including the new pipeline, for the extraction and assessment of DEGs related to Breast Cancer.

Firstly, the integration of the heterogeneous series from both Microarray and RNA-Seq technologies has been carried out following the integration strategy mentioned at Chapter 4. Thanks to this integration, the resultant dataset has more samples for the ML, giving to the predictive models more information to learn.

Although the integration of different gene expression sources is very useful to achieve a dataset with a high number of samples for ML, this process can introduce deviations for the DEGs selection. With the aim of minimising this possible deviations, a concrete strategy has been designed.

Concretely, DEGs extraction was performed at different levels using the limma R package, both at individual levels (Microarray and RNA-Seq separately) and at integrated level (joining Microarray and RNA-Seq data).

At the end of the DEGs extraction procedure, a total of three different sets of DEGs were obtained: the first one using only Microarray data, The second one using only RNA-Seq data and the last one with the integrated data from both technologies. Subsequently, the intersection of these sets was obtained. This intersection represents an invariant set of DEGs which are expressed independently of the used sequencing

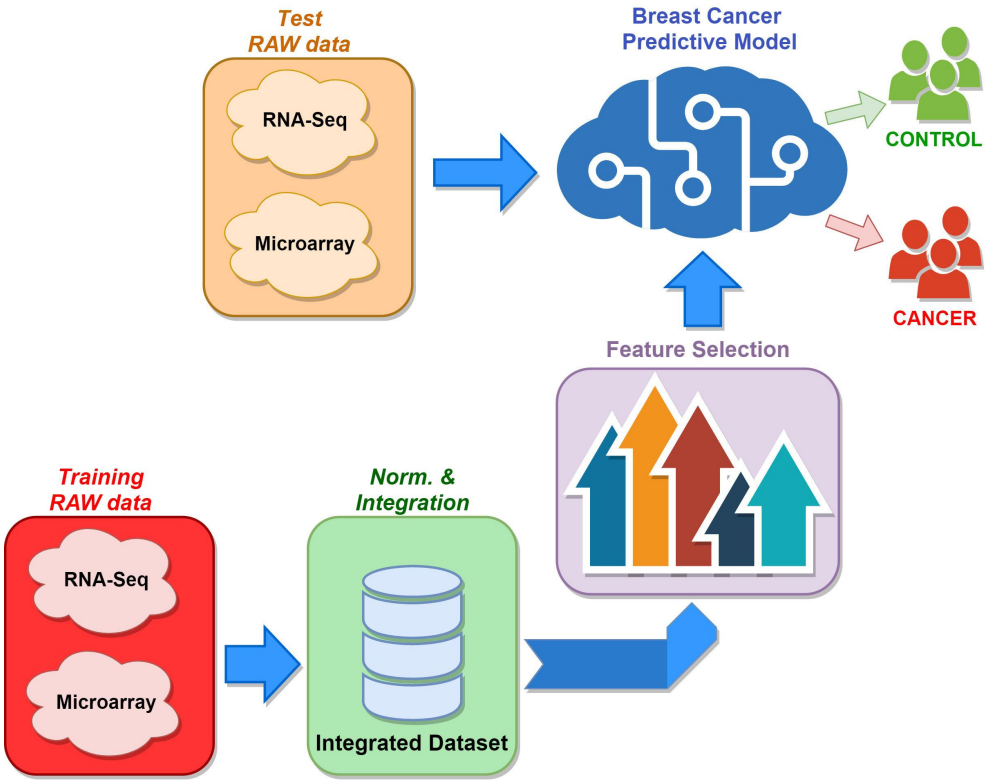

Figure 5.3: Integrated pipeline followed for this study

technology and the integration process. A Venn Diagram was used to perform this intersection.

5.2.5 *Predictive models*

Once a set of candidates DEGs which can be considered as biomarkers for breast cancer were retrieved, the assessment of those DEGs through two different classification technologies were made: SVM and RF. The main objective is the validation of the behaviour of the DEGs with the arrival of new unseen samples. The selected DEGs and the training dataset were used for building the predictive models that were later evaluated over the test dataset. Although both algorithms have been explained at Chapter 3, a briefly summary of both is given herein.

- SVM: This algorithm is based on the idea of separating the different categories in a problem through a hyperplane. The algorithm calculates the maximum-margin hyperplane that maximises the distance between different classes. For overlapped data, this type of models turn a reduced space into a higher space for performing the classification using a kernel function. Moreover, SVM allows

classification errors that are controlled by the γ hyperparameter in order to improve the generalisation capability of the model.

- RF: This method grows many single classification trees with the purpose of building a forest of uncorrelated classification trees. For the classification, the algorithm assigns the input vector to be classified to each tree of the forest. Once that each individual tree performs classification, the forest chooses the class having most votes over all the trees. After each tree is built, all of the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalised by dividing by the number of trees. Proximities are used with the aim of replacing missing data, locating outliers and producing illuminating low-dimensional views of the data.

10-fold cross-validation was used over the training dataset to obtain the optimal hyperparameters for both methodologies: σ (kernel width) and γ for SVM, and *number of trees* for RF.

5.2.6 *Feature selection*

Before the predictive models assessment, a FS process was made by applying mRMR algorithm over the candidate biomarkers with the objective of finding a reduced subset of genes that gives similar classification performances, reducing the number of genes. In this way, the reduction of the number of biomarkers allows creating a more simple and interpretable model, as well as more computationally efficient, while maintaining the robustness of the method. To sum up, mRMR sorts genes using mutual information as the criterion for computing the relevance and redundancy among biomarkers in this case. Taking this into account, mRMR will rank in first position the gene that contains the maximum relevance information but minimum redundancy information with respect to the rest of the genes, and so forth, it will proceed with the whole ranking.

5.3 RESULTS AND DISCUSSION

This section will focus on exposing and discussing the obtained results after the experimentation process followed for this breast cancer study.

The section is divided in two subsections: first subsection shows the results for obtaining the set of DEGs, while the second subsection will show classification results by making use of the former set of genes.

5.3.1 *Gene expression Analysis*

This subsection encompasses the achieved results for extracting the Breast cancer related DEGs. As it was previously stated, the heterogeneous transcriptomics series have been integrated. This integration have two main objective. The first one is to increase the number of available samples that will be used at the input of ML techniques with the aim of improving the robustness and stability of the results. The last one is to achieve an independence of the obtained DEGs from the technology, thanks to the intersection of the DEGs from different sources.

When working with heterogeneous transcriptomics data sources, normalisation is one of the most sensitive steps in the whole process, because a single mistake in this step could cause interpretation errors and it may lead to a false set of DEGs. Figure 5.4 shows the heterogeneity among samples coming from different series and technologies. Both training and test datasets have been subjected to a joint normalisation using normalizeBetweenArrays function from limma R package. Figure 5.5 shows the results once the joint normalisation was applied. As it can be seen, the dynamic range between samples has been equalised. From now on, only the training dataset will be used in the process for identifying DEGs.

The next phase in the pipeline is the identification of the DEGs both for each technology separately (Microarray & RNA-Seq) and for the integrated dataset. Several thresholds were imposed in order to determine the DEGs: the $Log_2FoldChange$ was set to be greater or equal than 2 and the $P - value$ was set to be less or equal than 0.001. These constraints ensure that the selected DEGs will be statistically significant, showing different behaviour between MCF7 and MCF10A samples. This restrictions were applied to all the datasets: Microarray, RNA-Seq and integrated datasets, acquiring three different sets of DEGs. At last, through the intersection of the three sets of DEGs, an total of 98 common DEGs were found. This genes comply with the restrictions and they are differentially expressed in all datasets as the intersection shows (Figure 5.6). Consequently, the obtained set of genes are differentially expressed independently of the gene expression technology.

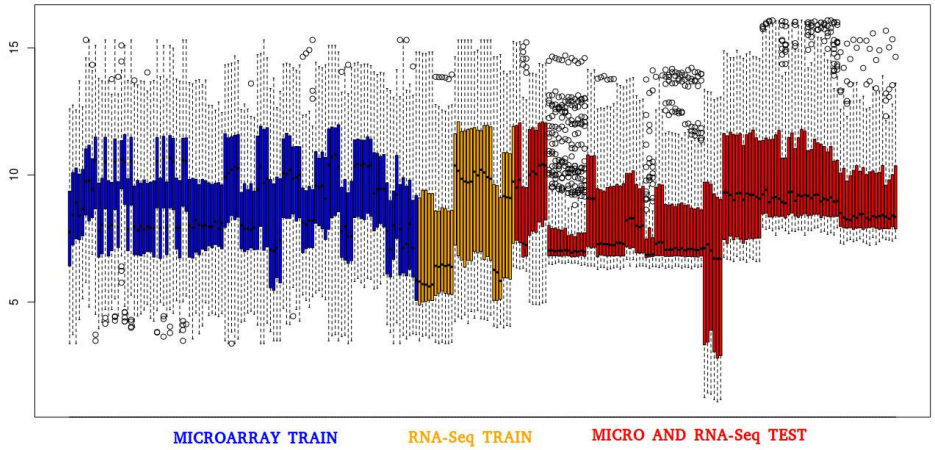

Figure 5.4: Expression levels of training and test datasets before normalisation

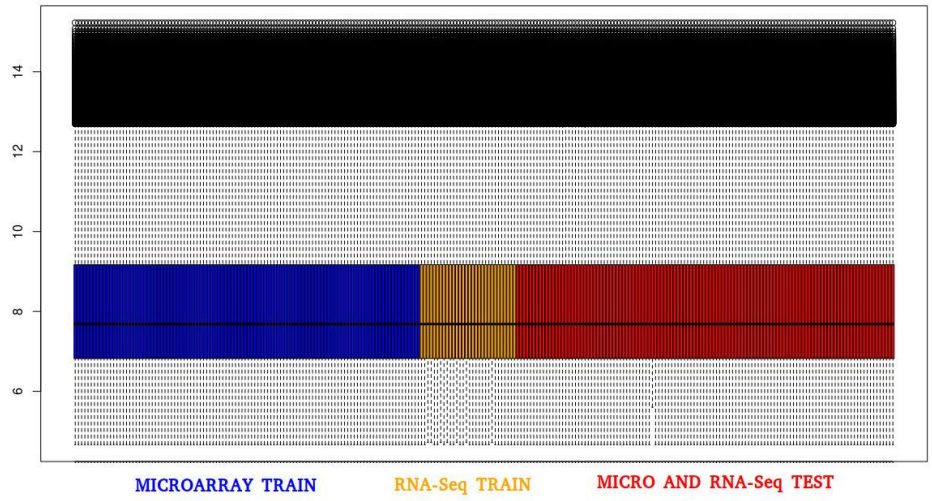

Figure 5.5: Expression levels of training and test datasets after normalisation

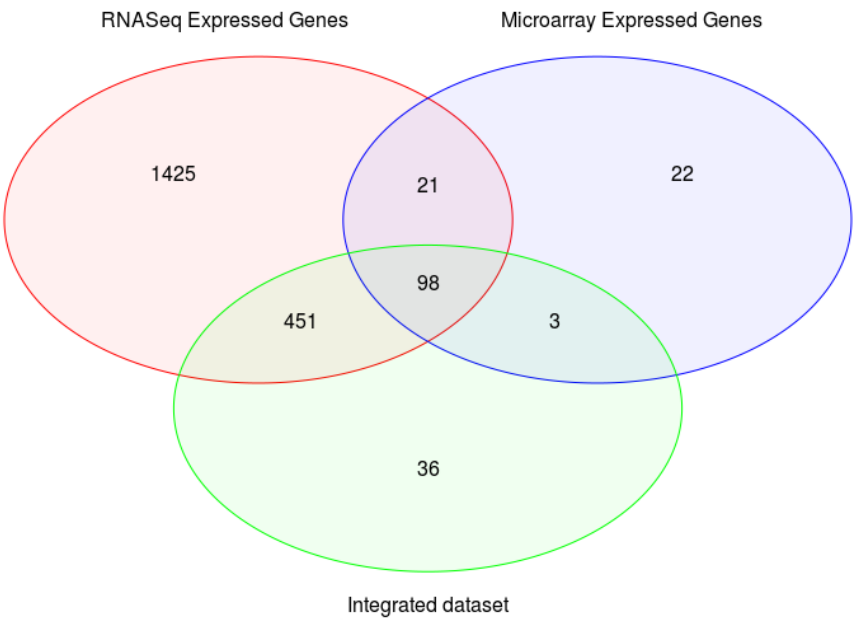

Figure 5.6: DEGs intersection among RNA-Seq, microarray and the integrated dataset.

It is very interesting to see the boxplots with the mean gene expression values for each samples in the training dataset. Figure 5.7 shows a clearly differentiation between the boxplots in the average value of MCF7 samples with regard to MCF10A samples (Cancer vs Control).

Table 5.2 shows the 98 common DEGs together with the five statistics values computed by the limma package. The log-fold change ($logFC$) represents the difference between breast cancer and control expressed values. If $\mid logFC \mid \geq 2$ it means that there exists significance differences between cancer and control values. The second value is the the moderated t-statistic, which has the same interpretation as the normal t-statistic but the standard errors have been reduced between the genes, effectively obtaining information from the set of genes to help with inference about each individual gene. The next value is the P-Value ($PVal$) which represents the probability of obtaining a result equal or higher than what it was observed when the null hypothesis is true. The adjusted P-Value indicates which proportion of comparisons within a family of comparisons (hypothesis tests) are significantly different. The B-statistic (B) is the log-odds that a given gene is differentially expressed.

Table 5.2: List of 98 common expressed genes obtained as the intersection of Microarray, RNA-Seq and integrated dataset.

Genes Names	$\mid logFC \mid \geq 2$	t	P.Value	adj.P.Val	B
KRT19	7.993	11.072	8.124E-21	2.449E-19	36.607
KRT6A	-7.800	-13.558	3.347E-27	2.503E-25	51.214
NNMT	-7.584	-11.544	4.951E-22	1.780E-20	39.384
VIM	-7.261	-15.117	3.917E-31	5.046E-29	60.213
AKR1B1	-6.943	-11.437	9.357E-22	3.265E-20	38.753
SFRP1	-6.866	-18.820	4.925E-40	1.904E-37	80.570
TGFBI	-6.701	-14.299	4.424E-29	4.174E-27	55.515
MT1E	-6.650	-15.281	1.537E-31	2.079E-29	61.142
C3	-6.569	-15.928	3.857E-33	6.589E-31	64.805
BMP7	6.406	13.058	6.330E-26	3.910E-24	48.292
KRT5	-6.229	-9.125	7.460E-16	1.062E-14	25.273
CXCL1	-6.145	-13.526	4.030E-27	2.986E-25	51.030
S100A2	-6.016	-9.582	5.249E-17	9.014E-16	27.902
KRT7	-5.991	-11.975	3.850E-23	1.643E-21	41.922
TNS4	-5.866	-25.125	1.651E-53	3.829E-50	111.284
EEF1A2	5.764	8.956	1.979E-15	2.656E-14	24.307
CLMP	-5.631	-11.238	3.037E-21	9.781E-20	37.583
IFI16	-5.543	-9.230	4.073E-16	6.036E-15	25.872
LAMC2	-5.426	-12.346	4.247E-24	2.015E-22	44.112
IGFBP4	5.412	13.779	9.173E-28	7.406E-26	52.501
FAM83A	-5.328	-14.042	1.974E-28	1.741E-26	54.028
SYTL2	5.283	11.883	6.617E-23	2.725E-21	41.384
SNAI2	-5.169	-9.731	2.204E-17	4.010E-16	28.762
DNER	-5.152	-11.859	7.620E-23	3.114E-21	41.244
PRKCDBP	-5.105	-10.241	1.105E-18	2.434E-17	31.730
ALOX15B	-5.088	-16.524	1.353E-34	2.896E-32	68.133
IGFBP5	5.085	8.165	1.755E-13	1.735E-12	19.871
BNC1	-5.072	-16.335	3.889E-34	7.697E-32	67.085
GFRA1	5.021	6.872	1.958E-10	1.223E-09	12.955
DSC3	-4.999	-17.145	4.296E-36	1.181E-33	71.561
PTGES	-4.990	-17.489	6.479E-37	1.947E-34	73.440
TFF1	4.925	4.857	3.168E-06	1.023E-05	3.497
RAB25	4.864	8.521	2.368E-14	2.683E-13	21.851
KRT14	-4.863	-6.445	1.768E-09	9.652E-09	10.794
EFEMP1	-4.855	-10.020	4.059E-18	8.275E-17	30.440
SLPI	-4.793	-10.194	1.455E-18	3.128E-17	31.457
SDPR	-4.728	-12.002	3.264E-23	1.401E-21	42.086
FBP1	4.707	6.789	3.017E-10	1.848E-09	12.530
EPCAM	4.662	8.150	1.906E-13	1.878E-12	19.790
GNA15	-4.570	-15.676	1.614E-32	2.495E-30	63.382
HTRA1	-4.527	-10.906	2.178E-20	6.152E-19	35.627
RAC2	-4.524	-11.727	1.669E-22	6.433E-21	40.465
CLCA2	-4.411	-9.272	3.189E-16	4.828E-15	26.115
GPX1	-4.384	-6.773	3.281E-10	1.994E-09	12.448
EMP3	-4.383	-9.299	2.728E-16	4.176E-15	26.269
SERPINB5	-4.371	-8.314	7.600E-14	8.016E-13	20.698
TSPYL5	4.317	6.297	3.735E-09	1.943E-08	10.062

	$\mid logFC \mid \geq 2$	t	P.Value	adj.P.Val	B
GSTP1	-4.242	-5.846	3.433E-08	1.523E-07	7.892
SLC2A10	4.216	11.411	1.088E-21	3.782E-20	38.602
LDHB	-4.182	-5.892	2.745E-08	1.238E-07	8.111
VSTM2L	-4.146	-11.277	2.409E-21	7.852E-20	37.813
BIRC3	-4.079	-13.064	6.110E-26	3.799E-24	48.327
ABLIM3	-4.000	-12.337	4.481E-24	2.113E-22	44.059
TFCP2L1	-3.874	-11.847	8.202E-23	3.344E-21	41.171
DSG3	-3.820	-8.387	5.035E-14	5.469E-13	21.105
SLC26A2	-3.798	-13.491	4.947E-27	3.632E-25	50.826
C3orf14	3.763	7.772	1.558E-12	1.358E-11	17.715
IL20RB	-3.667	-8.868	3.262E-15	4.229E-14	23.812
FXYD5	-3.623	-5.585	1.191E-07	4.882E-07	6.679
GSTM3	3.590	9.622	4.161E-17	7.268E-16	28.133
ADRB2	-3.572	-9.968	5.512E-18	1.099E-16	30.136
EMP1	-3.535	-7.622	3.543E-12	2.907E-11	16.905
IGFBP7	-3.530	-4.676	6.866E-06	2.104E-05	2.751
GJB5	-3.517	-12.456	2.225E-24	1.097E-22	44.755
HENMT1	3.514	7.953	5.732E-13	5.316E-12	18.702
ZBED2	-3.507	-6.452	1.705E-09	9.338E-09	10.830
MSLN	-3.504	-8.558	1.917E-14	2.217E-13	22.061
IL18	-3.415	-9.270	3.223E-16	4.864E-15	26.104
TRIM29	-3.395	-9.588	5.081E-17	8.735E-16	27.934
OSR2	3.346	8.380	5.238E-14	5.671E-13	21.066
LAMB1	-3.346	-6.972	1.162E-10	7.510E-10	13.468
UCP2	3.332	5.788	4.539E-08	1.979E-07	7.620
CPVL	-3.331	-7.870	9.043E-13	8.152E-12	18.253
KRT81	-3.320	-5.133	9.424E-07	3.334E-06	4.670
S100A8	-3.292	-5.698	6.982E-08	2.957E-07	7.200
TP53I3	-3.242	-11.149	5.160E-21	1.589E-19	37.057
FOXA1	3.226	5.576	1.241E-07	5.069E-07	6.640
SLC24A3	3.211	6.190	6.356E-09	3.184E-08	9.541
PNLIPRP3	-3.200	-7.998	4.470E-13	4.207E-12	18.948
INHBB	3.180	7.756	1.698E-12	1.468E-11	17.630
RAB38	-3.129	-9.539	6.781E-17	1.137E-15	27.649
ZBTB16	-3.112	-8.869	3.251E-15	4.217E-14	23.816
PLD5	-3.070	-11.039	9.925E-21	2.960E-19	36.408
DFNA5	-3.047	-7.565	4.835E-12	3.890E-11	16.599
FKBP5	-2.988	-10.435	3.528E-19	8.458E-18	32.863
CD109	-2.986	-7.196	3.541E-11	2.475E-10	14.637
CASP1	-2.955	-6.388	2.367E-09	1.267E-08	10.509
SULT1E1	-2.903	-7.749	1.763E-12	1.513E-11	17.594
FAM174B	2.779	5.557	1.353E-07	5.493E-07	6.555
PDZK1IP1	-2.752	-7.028	8.611E-11	5.667E-10	13.743
TNNI2	-2.750	-7.896	7.842E-13	7.133E-12	18.393
CAV1	-2.727	-5.028	1.503E-06	5.131E-06	4.217
IRX4	-2.714	-7.628	3.433E-12	2.825E-11	16.936
KRT80	2.706	5.268	5.131E-07	1.895E-06	5.259
FOXO1	-2.649	-8.921	2.408E-15	3.188E-14	24.113
SNCA	-2.635	-8.533	2.211E-14	2.526E-13	21.919
TBL1X	2.565	9.676	3.043E-17	5.434E-16	28.442

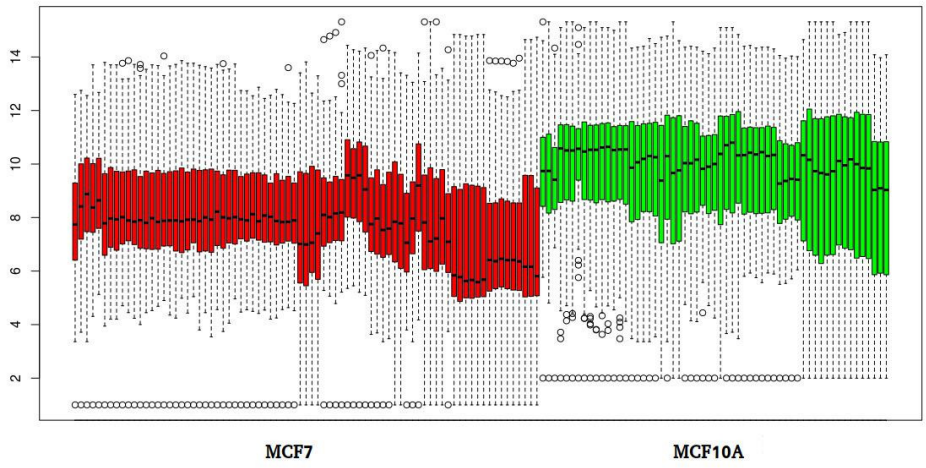

Figure 5.7: Gene expression values boxplot for the set of 98 expressed genes. Figure shows significant differences between expression values for MCF7 and MCF10A cell lines.

5.3.2 *Classification results*

This subsection assesses the performance of the selected DEGs through a feature selection and predictive models application. For that purpose, the classification algorithms SVM and RF have been implemented. The Training dataset has been used as the input data for the predictive models. The 98 common DEGs have been chosen as classification features ordered by a mutual information based ranking given by the mRMR FS algorithm. Moreover, for the posterior assessment of the models against new unseen samples, the test dataset has been used. This samples were correctly normalised as it was previously described in order to avoid possible errors with the dynamical midrange of the samples in the classification step.

One of the fundamental pillar of this novel pipeline is the ML techniques application to the set of common DEGs. Results for both algorithms (SVM and RF) in the validation stage with the training dataset, using the 98 DEGs reached an accuracy equal to 100%. Therefore, all samples belonging to the training dataset were correctly recognised and classified. When the model using 98 DEGs the test dataset was assessed, an accuracy of 97% was reached, confirming the robustness of the proposed pipeline approach.

Subsequently, the FS algorithm has been applied with the aim of reducing the cardinality of the 98 DEGs. In consequence, mRMR returned a gene ranking based on mutual information. Figure 5.8 shows the validation and test results using SVM algorithm. These results are above 96% using only the first gene of the ranking for classification. Furthermore, using the reduced set of the first six genes in the ranking it may be observed that the validation results for classification reached an accuracy of 98%. Furthermore, classification results when using the new 126 unseen samples from the test dataset are coherent reaching an accuracy of 96.5%. Therefore, the behaviour of the predictive model performs such as the validation results predict. Consequently, the main set of 98 common DEGs were reduced to the later six genes set that allow to discern if new samples are cancerous or not with a maximum of 3.5% of error. The results obtained by RF were slightly worst to the described SVM results, as it can be seen at Figure 5.9. With the purpose of exposing the precise test accuracy percentages for the reduced subset of six DEGs for each predictive model, Table 5.3 is also provided. This Table shows that SVM reaches better results than RF no matter the number of DEGs used.

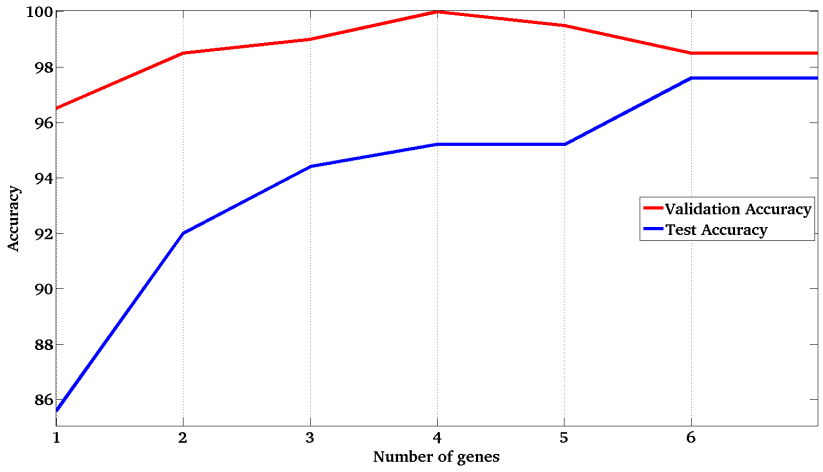

Figure 5.8: Validation and test classification results with SVM using the most relevant genes obtained by mRMR.

Table 5.3: Table with the test results from the predictive models after the feature selection step.

	Accuracy					
N Genes	1	2	3	4	5	6
SVM	85.6%	92.1%	94.8%	95.2%	95.3%	96.5%
RF	85.5%	85.5%	87.5%	91.2%	91%	88.4%

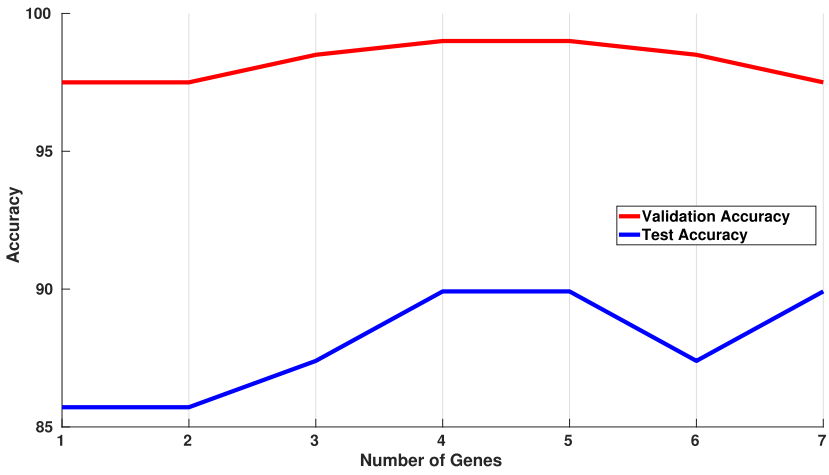

Figure 5.9: Validation and test classification results with RF using the most relevant genes obtained by mRMR.

Figure 5.10 shows a heatmap draw from the reduced six sub-set of DEGs. Two distinct groups are clearly identified: one matching MCF10A samples and the other matching MCF7 samples. Henceforth, this indicates that the expression profiles of these DEGs constitute a possible diagnosis criteria for breast cancer using MCF7 cell line.

Figure 5.11 contains two boxplots for each of the six DEGs, representing the average expression value for the cancerous samples (red) and control samples (green). As can be seen, average expression values between cancerous and control samples are clearly differentiated, thus reaffirming their potential as breast cancer biomarkers for MCF7 samples.

Finally, once the potential biomarker genes were identified as the reduced subset of six genes a literature review and biological study was done in order to reveal the relation between those genes and their involvement in breast cancer:

- Secreted frizzled-related protein 1 (SFRP1): Inhibition of SFRP1 increases the proliferation, migration and invasion of breast cancer cells. SFRP1 exerted this function by activating Wnt β-catenin signaling pathway in breast carcinogenesis [166, 167].

- Glutathione S-transferase mu 3 (GSTM3): GSTM3 is suggested as an important modifier that impacts on individual susceptibility to develop breast cancer among premenopausal women [168]. High expression of GSTM3 is related to protective genotypes against breast cancer.

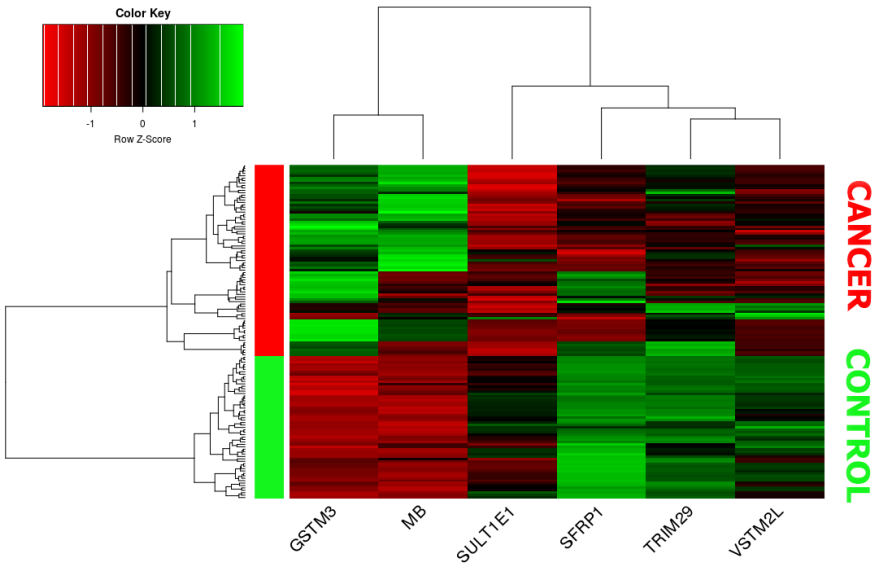

Figure 5.10: Hierarchical cluster over MCF10A and MCF7 samples using top 6 DEGs

- Gulfotransferase family 1E member 1 (SULT1E1): SULT1E1 is an enzyme that catalyzes the sulfation of active 17β-estradiol into inactive form. SULT1E1 is highly expressed in normal mammary epithelial cells and rarely expressed in breast cancer cells. However, its overexpression in breast carcinomas is considered to retarded tumor cell growth by arresting cell cycles and inducing apoptosis and may thus improve the prognosis of breast cancer [169, 170].

- Myoglobin (MB): MB plays a functional role in breast cancer progression by promoting the growth of fully oxygenated cells through the control of fatty acid homeostasis and lipogenesis [171, 172]. MB is dose-dependent downregulated by 17β-estradiol in breast cancer cells [173].

- Tripartite motif containing 29 (TRIM29): TRIM29 is considered a breast cancer tumor suppressor. Low TRIM29 expression in breast cancer is associated with more aggressive tumor features. Suppression of the oncogenic transcription factor TWIST1 expression is one mechanism suggested by which TRIM29 functions as a suppressor of breast cancer development [174].

- V-set and transmembrane domain containing 2 like (VSTML2): Although VSTM2L is detected in breast cancer tissues, to date

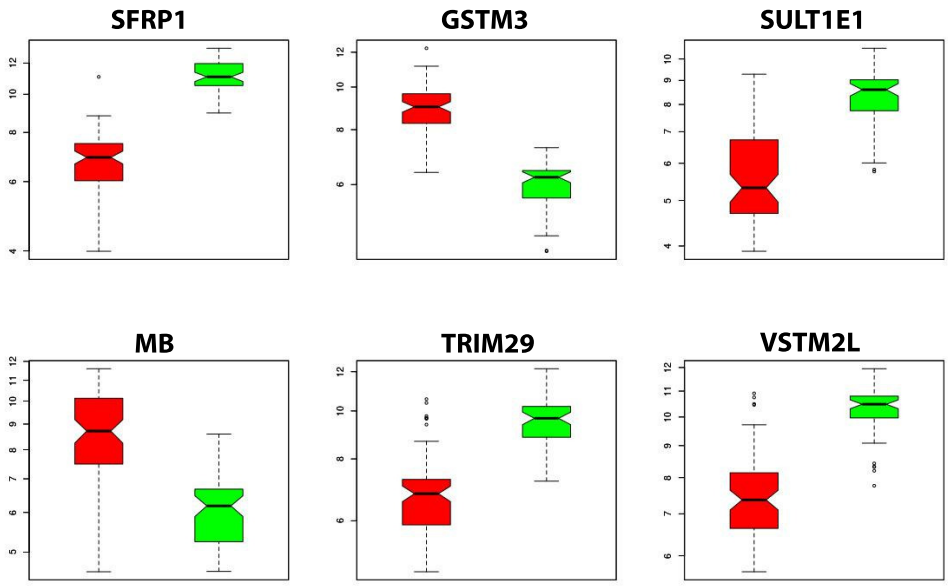

Figure 5.11: Average expression value boxplots of the six most relevant DEGs acquired in this research.

there are no relation between its expression and breast cancer development in the current literature.

The first five of these six genes have been formerly reported as genes related with breast cancer whilst the sixth gene is present in breast cancerous tissue but there is no evidence of a direct implication with breast cancer development. This means that the results following the proposed integrated pipeline are coherent as the reduced sub-set of six DEGs is formed by genes related with breast cancer. Furthermore, these DEGs can be used for classification and diagnosis purposes over new unseen samples, so that they can be designated as a new candidate biomarker signature when this type of data or cell lines from breast cancer are present.

5.4 CONCLUSIONS OF THE CHAPTER

Along this chapter, a first approach for integrating data from different heterogeneous transcriptomic sources has been designed and tested. To carry out this integration, an exhaustive search from the NCBI-GEO public repository has been done in order to collect breast cancer and control cell lines (MCF7 and MCF10A) samples from both technologies. The intersection of DEGs retrieved from RNA-Seq, Microarray and the

integrated dataset has allowed to identify a set of candidates biomarkers for the diagnosis of this disease using MCF7 cell line.

Afterwards, FS through mRMR was applied in order to select the most relevant biomarkers sub-set to achieve similar results, reducing the complexity. Then, both SVM and RF predictive models were built starting from the training dataset. The classifiers were validated with the test dataset achieving both outstanding results, overcoming though SVM to RF.

To conclude this chapter, results show that the DEGs can be designated as robust biomarkers for breast cancer diagnosis when specific cell lines samples are used. Furthermore, even with a small subset of six of those DEGs, a great validation accuracy was reached (98%). Also classification results over new unseen data show great accuracy (96.5%). Five of these top six genes have been formerly reported as genes that show biological relation with breast cancer, therefore reinforcing the designation of the expression profiles of these DEGs for breast cancer diagnosis by using MCF7 samples.

6

LEUKEMIA SUB-TYPES DIAGNOSIS BY APPLYING MACHINE LEARNING TECHNIQUES

CONTENTS

This research is a a reorganised and extended version of the published manuscript "Leukemia multiclass assessment and classification from Microarray and RNA-Seq technologies integration at gene expression level" [7], which presents a step forward in the pipeline exposed at Chapter 5 for Integration of heterogeneous data in Breast cancer [6].

6.1 BACKGROUND

As it was mentioned along this doctoral thesis, in more recent years, an important increase in the number of available public omics data has taken place due to the widespread use of NGS. Moreover, the continuous developments in ML and in the HPC areas, are allowing a faster and more efficient analysis and processing of this type of data. Nevertheless, biological information about a certain disease is normally widespread due to the use of different sequencing technologies and different manufacturers, in different experiments along the years around the world. Thus, nowadays it is of paramount importance to attain a correct integration of biologically-related data in order to achieve genuine benefits from them. In this research, an evolution of the pipeline presented at Chapter 5 has been addressed with the purpose of taking advantage of both RNA-Seq and Microarray. Furthermore, this integration has been done at multiclass level, due to the nature of the problem studied. At sight of this novel analysis, a new parameter has been introduced for extracting a set of candidate DEGs. This novel parameter will be called COV from now on and it will be used together with the LFC in order to find DEGs. This parameter aims to measure the "coverage" that a certain biomarker has over the different diseases analysed, i.e., the number of diseases it is able to discriminate.

COV: Coverage

In this Chapter, from the use of transcriptomic data coming from heterogeneous sources, different types of leukemia will be studied in order to find a leukemia gene signature for each of them. Leukemia, together with Lymphoma and Myeloma, is one of the three different existing blood cancer forms. People that suffer from leukemia produce an abnormal number of immature white blood cells, which collapse the bone marrow and inhibit the creation of the rest of vital blood cells for a balanced immune system and healthy blood. There are two main types of leukemia, being each divided into two subtypes:

- Acute Leukemia appears suddenly and progresses quickly so the treatment has to be urgent.
  - AML: is the most common leukemia in people around 70 years but it has impact in all ages. This malignancy is a heterogeneous group of neoplastic disorders, that are characterised by the proliferation and accumulation of immature hematopoietic cells in the bone marrow and blood. Different genetic factors have been identified that predispose to the development of AML. In this context, the germline predisposition and the existence of haematological disorders

AML: Acute Myeloid Leukemia

antecedent, have been associated with an increased risk of AML [175, 176].

- ALL: is the most common leukemia in children. About half the cases are in adults and half in children. ALL is also a very heterogeneous disease, characterised by impaired differentiation and proliferation of immature lymphoid cells in the bone marrow and peripheral blood. However, the prognosis of these patients has improved in the last years, especially in children, leading to cure rates approaching 80% to 90% due to the intensification of treatment, patient stratification based on clinical risk factors, and minimal residual disease (MRD) monitoring [177, 178].

ALL: Acute Lymphoblastic Leukemia

- Chronic Leukemia: symptoms appears more slowly, maybe in months or even years.

  - CML: it is very unusual and affects only 700 people per year. CML, defined as a clonal myeloproliferative disorder, was the first human malignancy associated to a consistent chromosomal abnormality and is characterized by the presence of the fusion oncogene BCR-ABL. Clinical symptoms associated to this disease include hypercellular bone marrow, anemia, platelet dysfunction, and an increase in the number of leukocytes, especially neutrophils and immature myeloid cells [179, 180].

CML: Chronic Myeloid Leukemia

  - CLL: it is more common in people over 60 years and is very rare in people under 40 years. CLL is a common B-cell tumor, characterised by the gradual accumulation of clonally expanded CD5+ B lymphocytes in peripheral lymphoid organs, secondary lymphoid organs, and bone marrow. It is also a genetic and biological complex disease, and the most commonly used factors to stratify CLL patients are the mutational status of the variable portion of the immunoglobulin gene, the deletion of the chromosome 17p and TP53 gene mutations [181, 182].

CLL: Chronic Lymphocytic Leukemia

This Chapter is twofold objective. The first one is the extraction of possible DEGs that allow discerning among the different forms of leukemia and people who does not suffers from the disease. The second one is to perform a classification stage for the DEGs assessment. Therefore, this study is about a multiclass classification problem in which a set of DEGs will be selected as long as they are useful to discern among the five classes: four different types of leukemia and healthy subjects. This

is a very novel research because most of the studies are dichotomous, addressing only two classes for their researches. To this end, the public repository NCBI/GEO has been used.

For the DEGs assessment, a set of smart leukemia classifiers to perform a differentiation among the different types of leukemia addressed when unlabeled samples are presented were build. To this end, mRMR FS algorithm was applied in order to select the most relevant DEGs to improve and perform the classification. Also, four different predictive models have been designed and their results compared. The classifiers are the following: SVM, RF, k-NN and NB.

6.2 INTEGRATED PIPELINE FOR MULTICLASS LEUKEMIA ANALYSIS

6.2.1 *Data gathering*

For this experiment, both Microarray and RNA-Seq samples have been collected in order to accomplish the integration of a wide range of heterogeneous data. As already mentioned, all series has been downloaded from the NCBI/GEO public database. A comprehensive search has been carried out with the purpose of gathering a notable number of samples belonging to the leukemia states addressed in this work. Furthermore, as all the samples must belong to the same tissue, only samples from cells of bone marrow have been used for both healthy and leukemia samples. With regard to Microarray samples, the two main platforms (Affymetrix and Illumina) have been taken into account. For RNA-Seq, samples have been solely collected from the most important sequencing platform known as Illumina HiSeq. Finally, a total amount of 11 series from Microarray and 2 series from RNA-Seq were selected for the research. These series are publicly available at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=S.NAME` where S.NAME is the name of each series at NCBI GEO shown at Table 6.1, which also includes information about the collected series.

Additionally, Table 6.2 shows the number of samples of each class for each gene quantification technology. It is worth noting that only Microarray technology presented samples for each of the studied classes. Indeed, there are not enough public samples of these states for RNA-Seq. This is an important motivation to keep using Microarray samples, taking advantage of them. In this sense, the integration that our pipeline performs is a significant step forward, opening the door to more complex studies.

Table 6.1: Relevant information about the series studied in this research. *Total Samples* column represents the total amount of samples that each series contains. *Accepted Samples* column denote the number of samples that belong to the different leukemia or healthy states and that will be analyzed in this study. The *Outliers* column quantifies the low quality samples that were removed from the *Accepted Samples*. Finally, the *Procedence* column reveals the genetic diversity in the origin of the series for thus study.

Series	Platform	Sequencing Technology	Total Samples	Accepted Samples	Outliers	Procedence
GSE6691	Affymetrix	Microarray	56	11	0	Salamanca (Spain)
GSE51082	Affymetrix	Microarray	139	55	1	Oregon (USA)
GSE12417	Affymetrix	Microarray	163	152	11	Munich (Germany)
GSE21029	Affymetrix	Microarray	62	19	1	Bethesda (USA)
GSE49067	Affymetrix	Microarray	12	12	0	Boston (USA)
GSE36474	Affymetrix	Microarray	7	3	0	Brussels (Belgium)
GSE33075	Affymetrix	Microarray	27	24	3	Salamanca (Spain)
GSE34860	Affymetrix	Microarray	78	78	0	Milan (Italy)
GSE61853	Illumina	Microarray	14	7	0	Daejeon (South Korea)
GSE11504	Affymetrix	Microarray	25	7	0	Oslo (Norway)
GSE13576	Affymetrix	Microarray	209	197	0	Padova (Italy)
GSE98310	Illumina	RNA-seq	22	22	0	Montreal (Canada)
GSE63646	Illumina	RNA-seq	71	71	0	Columbus (USA)
TOTAL	**-**	**-**	**885**	**658**	**16**	**-**

6.2.2 *Multiclass Workflow*

The Multiclass pipeline can be split into four steps or phases, as it is shown at Figure 6.1. For Microarray and RNA-Seq RAW data treatment, the strategies explained at Chapters 4 and 5 have been followed, although they are also represented at Figure 6.1.

Table 6.2: Number of categorized samples collected for each of the applied sequencing technologies. HBM stands for *Healthy Bone Marrow* and the rest represent the four types of leukemia. A lack of RNA-seq samples is clearly showed except for the AML state.

Type/State	HBM	AML	ALL	CML	CLL
Microarray	26	259	197	53	29
RNA-seq	0	93	0	0	0
Total	26	352	197	53	29

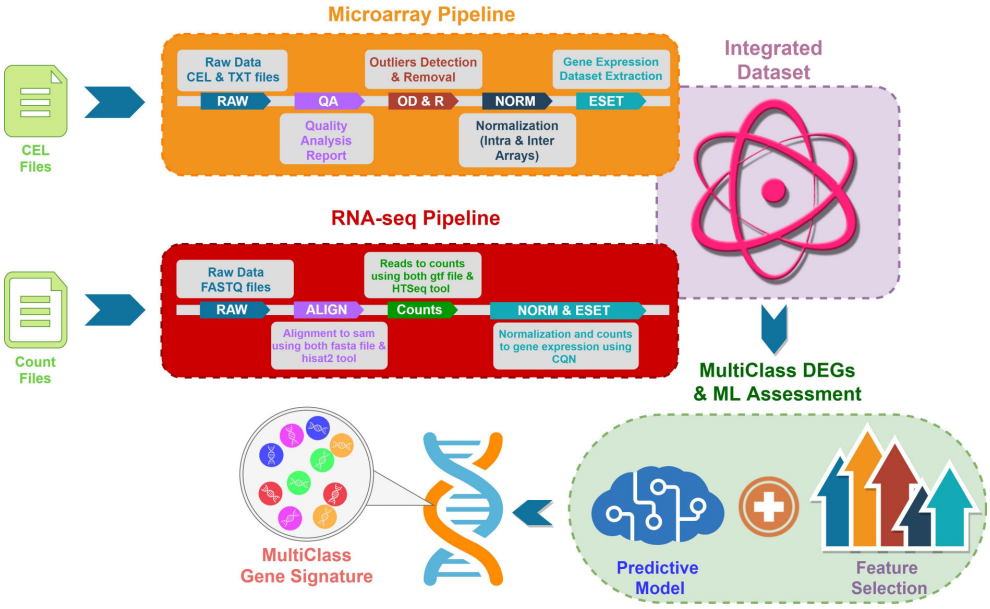

Figure 6.1: Proposed pipeline for the integration and classification of heterogeneous (Microarray and RNA-Seq) biological data, and the posterior Machine learning assessment.

6.2.2.1 *Microarray and RNA-Seq Integration*

As it was already mentioned, this pipeline is an improvement of the first integrated pipeline designed and tested at Chapter 5. Some indications about the pipeline is given herein in order to sum up the integration process.

Firstly, data are merged, requiring the correct individual pre-processing of all available series in the study. Afterwards, heterogeneous integration was carried out using the *merge* function from the base R package.

The second step in the integration is the normalisation of the integrated dataset. To perform this normalisation, the *normalizeBetweenArrays* function was used. This function is based on quantile normalisation and allows removing any possible deviation or variation among different arrays or datasets.

These two tasks are essential in order to achieve a correct normalisation of the biological data and its subsequent processing. The integration is indeed a critical process in the study as if the merging or normalisation are not properly done then the extracted DEGs would be erroneous. This would introduce confusing values in the research, that would lead irretrievably to a misleading selection of DEGs and subsequent erroneous outcome of the machine learning process.

6.2.2.2 *Multiclass DEGs Extraction*

Once the integration has been properly carried out, the next step is the DEGs extraction. There are some statistical toolboxes/routines that allow estimating if a gene has enough statistical significance for being considered a DEGs or not.

As mentioned before, a multiclass problem is addressed in this study, which means that DEGs have to be valid for discerning among more than two classes. Specifically, five classes are identified in this study, being necessary to find a group of DEGs with the capability of discerning among these classes in order to achieve multiclass classification. However, it is important to notice that limma, when dealing with a multiclass problem, takes into account only one value of the LFC between two classes, to identify if a gene is relevant or not, omitting the needed consideration of the rest of classes and class comparisons.

The total number of binary problems (taking one against one class comparisons) that take place in a problem with N classes can be defined as shown in Equation 6.1 and it will be called from now on COV_{max}. With the purpose of properly identifying the differences in gene expression among all classes involved in a multiclass problem, and thus identifying the best DEGs for the same purpose, the COV of a gene is defined as the number of class pair comparisons that a gene covers when a LFC restriction is imposed (furthermore, the possible DEGs also have to reach a P-value lower or equal to 0.001). In our present problem including 5 classes, COV_{max} takes the value of 10 (ten pairs of class combinations). The real potential of this parameter lies in the

ability to discover high coverage DEGs, thus allowing us to discern among the maximum possible number of classes.

$$COV_{max} = \frac{N^2 - N}{2} \tag{6.1}$$

Using COV as criteria to identify gene signatures implies choosing a coverage threshold, so that a certain biomarker will be selected only if it covers or differentiates at least a certain number of binary class comparisons. A large coverage threshold could be too restrictive, and a small coverage threshold could lead to the selection of too many biomarkers. The determination of an appropriate COV threshold is therefore critical in this problem. A medium size COV representing a trade-off between large and low differentiability in number of class pairs -coverage- (a fraction of COV_{max}, such as $COV_{max}/2$, $COV_{max}/3$, depending on the number of classes considered) may seem to be a reasonable threshold. The ANOVA test presented later will study the performance of different COV threshold values and their importance in the ML results.

ANOVA: ANalysis Of VAriance

6.2.2.3 *Machine Learning Assessment*

As it has been highlighted along this doctoral thesis, DEGs extraction is a very sensitive process because. In order to assess the DEGs selection, a ML process is performed, which is explained below.

Firstly, the DEGs dataset is normalised with median 0 and standard deviation 1. This step homogenises into the same type of distribution the gene expression values and can suppress the effect of possible remaining outliers due to the bounded range.

Although the use of limma for DEGs extraction already identifies candidate DEGs for the classification of the different diseases, the application of a specific FS process before the classification step would reduce the dimensionality of the problem, simplifying the classification while keeping the final accuracy. Precisely, the mRMR algorithm (Explained at Chapter 3) has been used with the aim of obtaining a ranking with the most appropriate combination of our DEGs, according to the operation of the algorithm.

When both, the normalisation and FS process are done, four different supervised learning algorithms are applied to an increasing number of selected DEGs, according to the mRMR algorithm. A CV step using

k-fold is performed in order to assess the results of the classifiers on the dataset. As a reminder, K-fold CV algorithm iteratively leaves out $1/k$ data from the training dataset, which are used to assess the classifier when the training is done. Finally, both accuracy and f1-score are calculated using the outcomes from the k assessment processes. This last measure takes into account the grade of classification of each class, not only the total amount of samples correctly classified. Equation 6.4 represents f1-score, that is used when a multiclass problem is tackled due to the relevance of this measure in this type of problems. It is calculated by using both the precision or accuracy (Equation 6.2) and the recall or sensitivity (Equation 6.3).

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{6.2}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{6.3}$$

$$f1\_score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6.4}$$

The whole machine learning process has been explained so far, but without going into detail of each of the four implemented classifiers (k-NN, SVM, RF and NB) due to they have been in-depth explained at Chapter 3, and two of them briefly reminded at Chapter 5.

6.2.3 *ANOVA test*

The ANOVA test plays a very important role in this experiment as it will analyse and compare the performance of the four classifiers, as well as of different combinations of the hyperparameters LFC and COV for optimal relevant biomarker identification. This step is very useful in order to determine if the classifiers have significant differences among them. Moreover, the test will provide valuable information about which combination of LFC and COV is better for our study and it will bring some light into the optimization of these parameters for further studies.

6.3 RESULTS

This section will be split into three subsections in order to expose the results in the most clearly and organised way. Firstly, the impact of both the LFC and the COV on the final classification results is evaluated performing an ANOVA test. Secondly, the first four steps of our pipeline are applied for the extraction of the DEGs. Finally, the results of the DEGs ranking process, and the assessment of those DEGs using a machine learning process will be shown.

6.3.1 *Statistical assessment through ANOVA test*

The ANOVA test is very useful at this point due to two main reasons. On one hand, it is important to optimise the parameters because there are several possible values that both the LFC and the COV could take. On the other hand, four different predictive models are assessed in this experiment. At sight of these considerations, the ANOVA test can decide if there are statistically significant differences among the classifiers and how the value of both the LFC and the COV could affect final results. The chosen classifier, LFC and COV are the parameters that a priori could cause more impact in the study. Nevertheless, the final number of selected genes after a feature ranking process (mRMR) could also affect final results, so it will also be considered.

Therefore, the possibles values that the four factors can take in the test are the following:

- Classifier: this variable represents the classifier used for the simulation. This classifier can be SVM, k-NN, NB or RF.
- LFC: this variable represents the Log-Fold Change used in order to extract the relevant genes, taking the values 1, 1.5, 2, 2.5 or 3.
- COV: this variable represents in a multiclass problem, the number of combination of classes in which a gene is truly relevant, taking the values 2, 3, 4 or 5.
- NR. GENES: this variable represents the number of genes finally selected with mRMR, which were used as input for the classifier, taking values 10, 20, 30 or 40.

From all the possible combinations of the previous factors, a wide range of simulations have to be addressed and evaluated in order to achieve an statistical interpretation of the results. Consequently, the test evaluates how both accuracy and f1-score are affected by the four chosen variables for the test (Classifier, LFC, COV, NR. GENES). Table 6.3 shows the relevance of these variables with respect to the accuracy. In this table is clearly seen how all the studied variables are relevant with regard to the accuracy, taking into account that P-value of each of them is less than 0.05. This means that it is important to take all of them into account for the research.

Table 6.3: Variance analysis for the accuracy - Sum of Squares type III

Source	Sum of Squares	Gl	Medium Square	F-value	P-value
Mainly effects					
A:Classifier	0.0802257	3	0.0267419	67.86	**0.0000**
B: LFC	0.00438299	4	0.00109575	2.78	**0.0256**
C: COV	0.00660603	3	0.00220201	5.59	**0.0008**
D: Nº GENES	0.0724379	3	0.024146	61.27	**0.0000**
Residuals	0.593102	1505	0.000394088		
Corrected total	0.757276	1518			

In the same way that the previous table shows how selected variables affect accuracy, Table 6.4 analyses the impact on the f1-score. As it happened for accuracy, selected variables are also relevant for the f1-score because the P-value for each analysed variable is less than 0.05.

Table 6.4: Variance analysis for the f1-score - Sum of Squares type III

Source	Sum of Square	Gl	Medium Square	F-value	P-value
Mainly effects					
A:Classifier	0.0177814	3	0.00592712	38.60	**0.0000**
B: LFC	0.010959	4	0.00273975	17.84	**0.0000**
C: COV	0.0014741	3	0.00491365	32.00	**0.0000**
D: Nº GENES	0.0398155	3	0.0132718	86.43	**0.0000**
Residuals	0.230789	1503	0.000153552		
Corrected total	0.311559	1516			

Previous tables showed how the studied variables affected both accuracy and f1-score. At this point, the impact of each value that the different variables could take will be described through a group of plots. In this sense, Fig 6.2 shows a graph for each of the studied variables (Classifier, LFC, COV, NR. GENES). The classifier variable chart clearly shows that the classifier attaining the highest accuracy is k-NN. Furthermore, both SVM and NB obtain the same results leaving RF as the worst classifier for this study.

Focusing on the LFC, the plot shows that an increase of LFC does not lead to an accuracy improvement. However, it is really important to note the behaviour of COV. As it can be seen, an increase in COV leads to an accuracy improvement, meaning that the new proposed measure is an important criterion in the selection of multiclass biomarkers. Among the values compared, ranging from 2 to 5, an improvement in the final recognition was shown due to the fact that as the value of COV increases more classes are covered by the selected genes.

Finally, it is straightforward to expect that larger gene signatures may attain higher accuracy in the identification of the different pathologies studied than shorter gene signatures. This is shown by the NR. GENES variable, whose values range from 10 to 40.

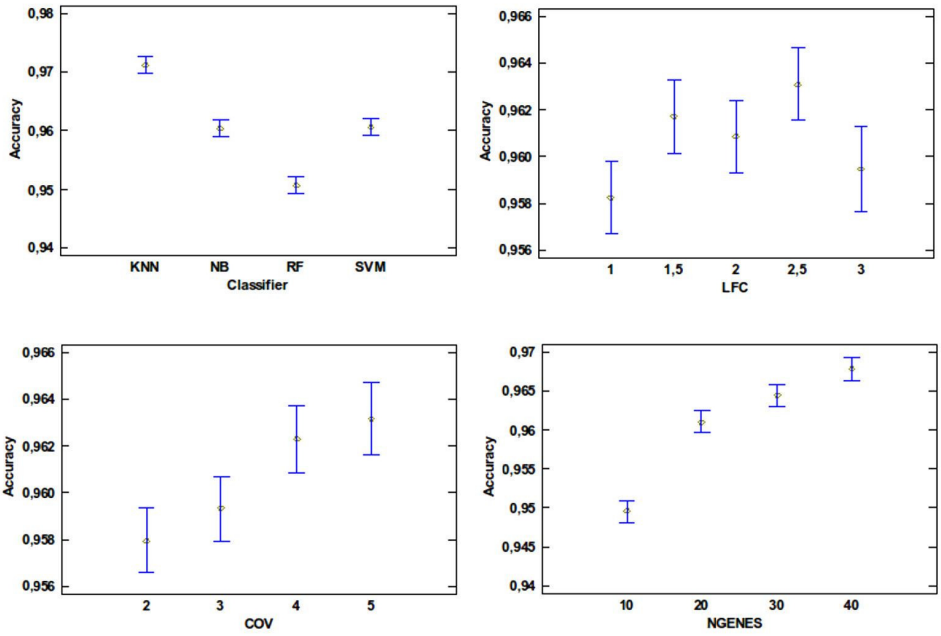

Figure 6.2: ANOVA results showing the impact on accuracy for each of the evaluated variables.

In the same way that the previous figure shows how the variables have a direct impact in the accuracy, Fig 6.3 represents the impact regarding the f1-score. The results for the LFC, COV and NR. GENES have the same behaviours in the f1-score than in the accuracy but for the classifier variable. In this case, SVM achieves better results than the NB. However, k-NN is still the best classifier and RF the worst for the study.

At sight of the ANOVA test results, only one of the possible combinations of LFC and COV will be taken into account to extract the

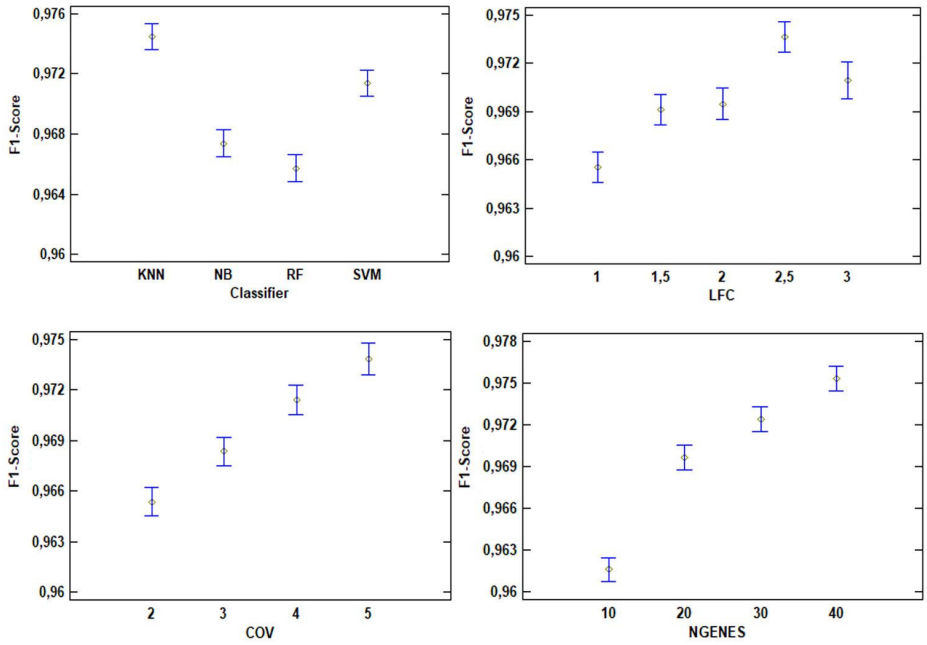

Figure 6.3: ANOVA results showing the impact on the f1-score for each of the evaluated variables.

biomarkers of this study. Specifically, the LFC equal or greater than 2.5 and the COV equal or greater than 5. These values are both the best combination in terms of achieving the best possible results, as can be seen in the ANOVA plots exposed before. Then, for these requirement settings, gene signatures with different NR. GENES will be finally studied, and the detailed results of the four classifiers will be shown.

6.3.2 *Applying Coverage for DEGs extraction*

Before presenting the extracted DEGs, the results of the series integration for this study will be shown. In order to perform an integrated analysis of these series, an individual analysis and correction of each series was done. Furthermore, it is necessary to correct the existing differences among the series due to the variety of technologies and platforms present in this study. In this sense, Fig 6.4 shows the normalisation and bit depth correction across the series once have been normalised separately.

In order to achieve the best coercing among the series after the integration, a joint normalisation with quantile normalisation is required,

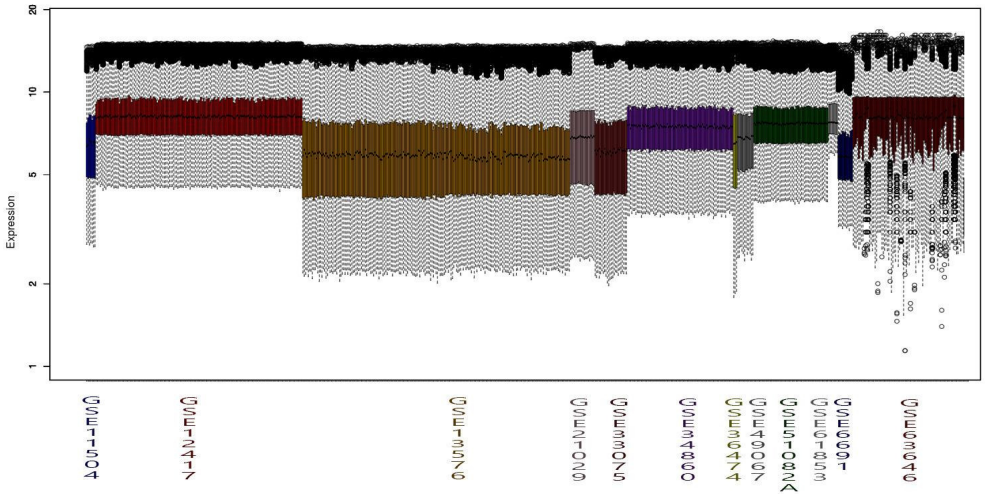

Figure 6.4: Expressions values comparison among leukemia series before the joint normalisation and integration steps.

with the purpose of obtaining the same dynamical range among the series (see Fig 6.5).

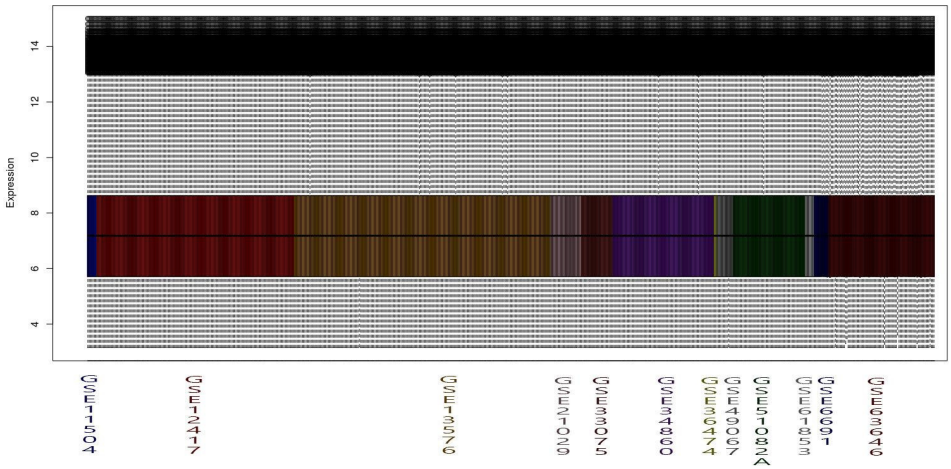

Figure 6.5: Expressions values comparison among leukemia series after the joint normalisation and integration steps.

Once normalisation using the *normalizationBetweenArrays* function has been accomplished, the dataset is completely ready to perform gene expression analysis. Thus, limma will be used in order to achieve this analysis. Nevertheless, it is important to note that this study is a multiclass gene expression problem. Therefore, it is necessary to use limma with this consideration, avoiding the classical biclass pipeline which limma implements by default.

Once multiclass limma was applied, a list of DEGs were reported. Those genes passed the three imposed restrictions: LFC equal or greater than 2.5, COV equal or greater than 5 and P-value equal or less than 0.001. A total of 42 genes that satisfy these restrictions were returned. Table 6.5 collects the 42 genes and shows statistical values about those DEGs. These statistical values are the T-statistic, the P-value, the LFC and the COV for each gene.

6.3.3 *Multiclass DEGs assessment using Machine Learning*

Once the DEGs extraction pipeline was done by using the integrated dataset, mRMR FS algorithm was applied, obtaining a ranking of DEGs in which the most relevant DEGs would be placed on the first positions within of this ranking. Furthermore, thanks to its operation taking into account the mutual information among the selected DEGs, this algorithm can also minimise the redundancy among them.

Fig 6.6 shows in an ordered way the 10 first genes returned by mRMR ranking, revealing for each of them, the expression levels of each class. Such genes will be more deeply commented both at bioinformatic and at biological level in the Discussion section. Reminding the COV parameter introduced in this study, it can be observed how the DEGs, concretely the 10 first of the mRMR ranking, present different expression levels not only for one class with regard to the others but also among several classes. For example, the first gene of the Figure (BLK) shows different expression levels for four of the types of leukemia, hence allowing us to discern among these classes. Indeed, due to the behaviour of the mRMR algorithm, this gene is the one with the highest level of Mutual Information with respect to the classifier variable.

Subsequently, the performance of the obtained ranking was evaluated. For that, four different classifiers were implemented and compared. Furthermore, this comparison has been performed for different number of genes (10, 20, 30 and 40) and for both the Accuracy and the f1-score. As for the simulations of the ANOVA test, a CV process (5-fold) was applied with the objective of providing an estimation of the performance on unseen samples, avoiding overfitting. This restrictive CV process guaranteed a significant representation of the lowest frequent classes (specially HBM and CLL, see Table 6.2) in all data folds.

The result of these comparisons can be seen at Table 6.6. This Table shows how k-NN reaches better results with respect to the rest of the classifiers in practically all the comparisons, reaching 96.40% of

Table 6.5: Table with the expressed genes that represents several statistical values of these genes.

Expressed Genes	μ $T-statistic$	μ $P-value$	μ LFC	COV
HIPK1	16.0875204	2.4713e-23	3.288872	6
SLC4A1	13.5735052	4.7010e-12	3.421740	6
FGD2	22.0274586	8.1678e-21	4.201152	6
DOCK2	24.4474942	1.9231e-24	3.678663	6
PADI2	22.1425831	1.3321e-30	3.050733	6
LAPTM4B	10.7437126	1.1139e-12	2.935885	6
C11orf58	12.5259082	6.9347e-10	4.240850	6
LITAF	12.7181709	2.1436e-10	3.927194	6
SF1	10.9802843	4.1383e-06	4.076528	6
RPS24	9.6354228	5.2593e-07	3.943997	6
CLEC2B	8.5098540	1.2060e-11	3.407249	6
EEF1A1	10.8653720	9.2231e-08	4.247919	7
LGALS3	11.2588096	1.2091e-10	3.999583	6
PLCG2	28.0170676	1.6955e-56	3.616019	6
BANK1	17.9033482	1.2248e-26	4.082236	6
H2AFY	13.7470887	2.1613e-09	5.151885	6
TSPAN3	18.6133699	1.8746e-18	3.606852	6
MKNK1	14.3396635	8.2579e-12	3.432306	6
PABPC1	9.7297018	2.8194e-06	4.177165	6
TAB2	15.0096557	3.7740e-18	3.071626	6
ROCK1	23.5206238	2.6058e-34	3.951989	6
RPS15	19.3210248	1.1963e-10	5.208190	7
GSN	8.5034573	3.6481e-10	2.994001	6
CMTM6	13.4600508	2.1206e-08	4.303823	6
FUS	22.7727271	1.5430e-33	3.524736	6
SEPT7	20.3301923	6.7129e-24	3.804574	6
ZNF160	14.9926210	5.0560e-20	3.280500	6
ANXA2	10.3112336	1.0576e-07	4.099294	6
EAF2	11.5559113	1.4391e-09	3.466531	6
TCF4	15.7648662	2.0821e-22	3.698245	6
CD22	34.1927437	1.1397e-64	3.522825	6
POU2AF1	31.2157033	8.9199e-59	5.590414	6
CFD	16.3828557	2.5946e-18	3.855234	6
BLK	34.6994030	2.5379e-69	5.076078	6
CD19	36.4443927	6.559e-108	4.806683	6
BLNK	25.7630631	1.7997e-37	5.563086	6
ACTN1	14.7867854	2.1587e-20	4.260679	6
CTGF	11.8350681	7.5825e-13	4.142035	6
TCL1A	24.0024506	1.1794e-34	4.774254	6
PPP1R16B	11.0936657	4.0266e-11	2.949562	6
AZU1	11.1121946	4.3126e-11	4.161520	6
ATP8B4	17.1102708	2.0803e-28	3.665865	6

Accuracy using only the 10 first DEGs chosen by mRMR algorithm, from the total of 42 DEGs. However, the f1-score reached by k-NN in this case is lower than the one reached by the rest of the classifiers for this number of genes. For 20 DEGs, k-NN reaches 98.56% of Accuracy and 98.75 of f1-score, being ahead of the rest of the classifiers. This behaviour is repeated for both 30 and 40 DEGs as can be seen at Fig 6.7 and at Fig 6.8. These figures show the evolution of the Accuracy and the f1-score, respectively, for the four implemented classifiers. Regarding the rest of classifiers, SVM reaches comparable although slightly worse results than k-NN for 30 and 40 genes. Finally, RF and NB present clearly worse results regardless the number of selected genes.

Table 6.6: Results of the four classifiers for both the accuracy and f1-score when using a different number of genes

	10 Genes		20 Genes		30 Genes		40 Genes	
Classifier	ACC	f1-score	ACC	f1-score	ACC	f1-score	ACC	f1-score
SVM	95,64%	97.13%	96,61%	98.27%	98,14%	98.75%	97,83%	98.59%
k-NN	96.40%	96.28%	98.56%	98.75%	98.78%	99.05%	98.87%	99.05%
NB	94.76%	97.29%	95.98%	97.79%	95.34%	97.79%	95.66%	97.61%
RF	95.51%	97.01%	95.05%	96.58%	95.35%	96.81%	95.42%	96.94%

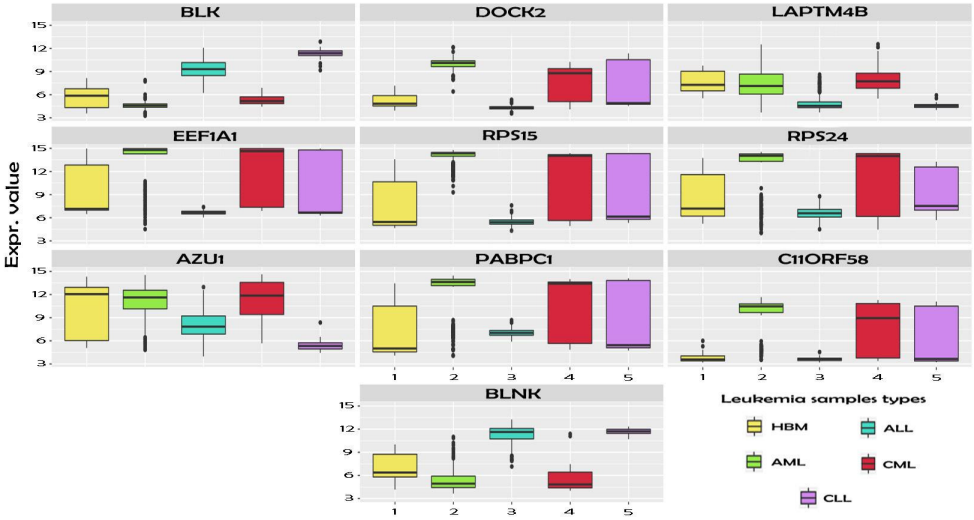

Figure 6.6: 0 first selected differentially expressed genes by mRMR algorithm (order from left to right and from top to bottom: BLK, DOCK2, LAPTM4B, EEF1A1, RPS15, RPS24, AZU1, PABPC1, C11ORF58 y BLNK), with the expression levels for each type of leukemia studied.

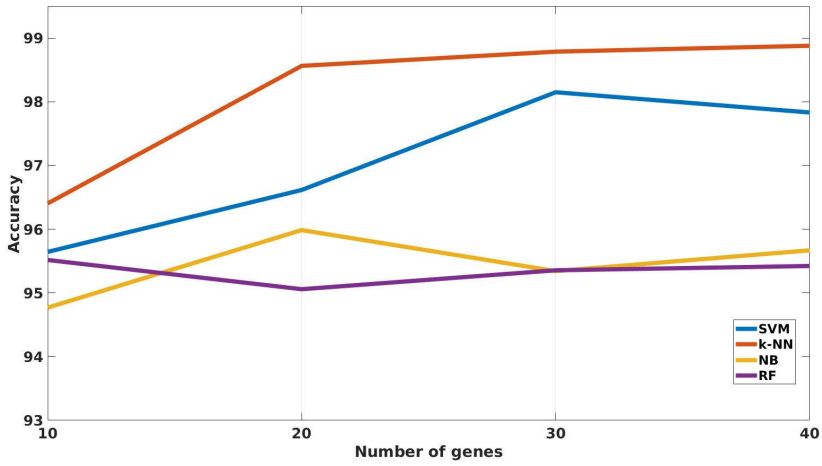

Figure 6.7: Plot that represents the accuracy achieved by each of the four classifiers used in the study.

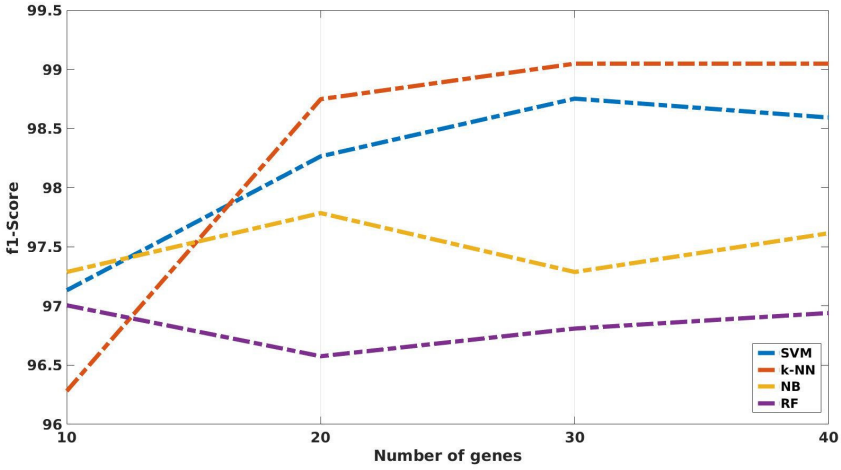

Figure 6.8: Plot that represents the f1-score achieved by each of the four classifiers used in the study.

6.4 RESULTS INTERPRETATION

Once the results of this study were exposed, the discussion of these will be performed with the purpose of giving an explanation both at biological and bioinformatic level.

6.4.1 *ANOVA interpretation*

ANOVA test is one of most important and crucial step carried out in this study due to the relevance of the parameters involved in the test. All the conclusions acquired emanated from the interpretation of both Table 6.3 and Table 6.4 and also from both the Figure 6.2 and the Figure 6.3.

Firstly, the best classifiers for both the Accuracy and the f1-score were k-NN and SVM. This behaviour is the same one observed in our previous study related to breast cancer but at bi-class level instead of at multiclass level. This result also coincides with the results obtained in other related works [183–185]. Furthermore, a Naive Bayes classifier was implemented in order to assess its performance with respect to the other classifiers, which were already compared in the literature. However, the performance obtained by the NB classifiers came out to be worse than SVM and k-NN classifiers. Both, SVM and k-NN work based on distances or similarity measures (kernels), while NB is based on bayesian probabilities and RF is based on the creation of decision trees to carry out the classification. This operation of both k-NN and SVM algorithms might be giving rise to these results.

Secondly, it was observed how the LFC parameter does not have a relevant impact on the optimal DEGs selection. For the evaluated values of this parameter, there is not a clear increase of the performance when the value of LFC increases too. Nevertheless, it is important to note that LFC is not used in a isolated way as this parameter is used in conjunction with the COV parameter, which showed to be the most important parameter for the DEGs selection.

Thirdly, the statistical evaluation of the COV parameter proves that it is the most important parameter in this study. COV parameter increases the classification performance as its value increases too. However, this parameter must be used carefully in a multiclass problem. A too high value of the COV parameter can lead to an aggressive and extreme reduction on the final number of selected DEGs. On the other hand, a too small value of this parameter would let too many genes, increasing the payload of the system and reducing the possibility of finding a reduced and thus useful genetic signature. In our case, the ANOVA test shows how an intermediate COV value of 5 was the optimal among the studied values and hence, this value was used for the definitive extraction of the DEGs.

Lastly, when the number of genes increases, the final classification rate increases too for both the Accuracy and the f1-score measures. Nevertheless, this rise significantly decreases when the number of genes exceeds the value of 20, showing either that it is not possible to provide more information to perform the classification, or that overfitting occurs.

To sum up, it can be seen how the ANOVA test shows the importance of the considered parameters, excluding the LFC, for the optimal recognition of possible genetic signatures. Moreover, thanks to the new parameter COV great classification results were obtained due to the multiclass DEGs selection achieved by this parameter.

6.4.2 *Differential Expressed Genes selection and assessment*

The ANOVA test allows defining an optimal parametrisation for this problem. By using a value for the COV greater or equal than 5 and a value for the LFC greater or equal than 2.5, a total of 42 multiclass DEGs were obtained. These genes were used as input variables for the classifiers with the main purpose of evaluating their potential as DEGs with the capability of discerning among the studied pathologies.

During the extraction process, an integration of the series from both RNA-Seq and Microarray has been carried out. Thanks to this, the number of samples available for this study increased considerably. The lack of public samples from RNA-Seq in comparison with Microarray was the motivation to perform the integration. Therefore, it was shown how Microarray still has a great potential in the field of gene quantification technologies and can longer be used in order to reinforce studies of this nature, increasing the number of available data and allowing to a more robust genetic signature discovery in diverse pathologies.

Moreover, thanks to the introduction of the COV parameter in the study, each DEG selected has the potential to discern among several of the 5 proposed classes. Concretely, if the maximum number of pair classes comparison are 10 and, the imposed restriction for each gene is that these DEGs discern minimum among 5, in the Table 6.5 it is shown how all the selected DEGs have the capability of discerning between 6 and 7 pair classes. Therefore, it makes sense to think that all the classes are discerned by any of the DEGs at any point with respect the other classes. In the validation process, theses genes have proven that they are strong candidates to be a possible gene signature that discern among the different types of leukemia studied in this article.

For the DEGs evaluation, 4 different classifiers were implemented. The results of these showed a high classification rate for the studied measures, Accuracy and f1-score. Furthermore, thanks to the feature selection process performed by mRMR, an optimal selection of subsets of genes for a reduced genetic signature was ensured.

Additionally, it is important to highlight that in order to ensure the correct validation of the data, a 5-fold validation has been implemented keeping an homogeneous distribution of samples of each state in each fold.

Hence, at sight of these results, with only 10 of the 42 DEGs, it is achieved a practically full discernment of the available samples under CV. With only 10 genes, k-NN classifier reaches a 96.40% of Accuracy and a 96.28% of f1-score for the five groups considered including healthy samples. This means that the DEGs found applying our methodology, which makes use of the COV parameter, properly works to discern, with a very high precision, among the 5 proposed classes.

6.4.3 *Biological relevance of the DEGs*

The top ten of the genes highlighted (see Figure 6.6) in our study were related in one way or another to peripheral blood leukocytes. Among them, DOCK2 gene which codifies by a protein involved in cytoskeleton remodelation and migration in response to chemokine signaling, has an especial relevance [186]. It has been reported that DOCK2 gene is overexpressed in chronic lymphocytic leukemia B-cells promoting their proliferation in response to Wnt5a [187]. This gene has been proposed as an drug target against leukemic cells since its expression is limited to hematopoietic tissues theoretically limiting side effects [188]. In addition, our results showed modulation of genes normally expressed in B-cells, such as BLK, BLNK and PABPC1. BLK is a proto-oncogene that encodes for a nonreceptor tyrosine-kinase involved in B-cell proliferation and differentiation. Signaling through BLK supports the pro-B to pre-B transition, growth arrest and apoptosis downstream of B-cell receptor [189]. Interestingly, this gene acts as a tumor suppressor in chronic myeloid leukemia stem cells and has been implicated in the progression of acute lymphoblastic leukemia [190, 191]. On the other hand, BLNK gene encodes for a cytoplasmic adaptor that plays a critical role in B-cell development [192]. Deficiency in this protein has been identified in some cases of pre-B acute lymphoblastic leukemia [193, 194]. In fact, the somatic loss of BLNK and concomitant mutations lead to a constitutive activation of Jak/STAT5 pathway,

resulting in the generation of pre-B-cell leukemia [195]. Finally, PABPC1 encodes for a poly(A) binding protein that regulates immunoglobulin secretion in these cells [196].

Modulation of genes normally expressed in others peripheral blood leukocytes such as T-cells (LAPTM4B and EEF1A1 genes) and neutrophils (Azurocidin 1 gene) were also detected in our study. LAPTM4B gene acts downregulating the TGFB1 production in regulatory T-cells [197]. Differential expression of LAPTM4B and MIR155HG was confirmed in a small cohort of young adult NPM1-mutated cytogenetically normal acute myeloid leukemia (CN-AML) patients. Although there is no direct evidence that links LAPTM4B to leukemia, its upregulation has been shown to implicate PI3K/AKT signaling and ubiquitination pathways, both with crucial roles in leukemogenesis [198]. The gene EEF1A1 also plays a key role on the proliferation inhibition and apoptosis induction of human acute T lymphocytic leukemia cells, contributing to cancer survival in haematopoietic malignancies [199]. On the other hand, Azurocidin 1 gene encodes for a preproprotein that matures into azurophil granule antibiotic protein, with monocyte chemotactic and antimicrobial activity [200]. In chronic myeloid leukemia, this gene has been included inside a set of six genes to discriminate between tyrosine kinase inhibitor therapy responders and non-responders [201]. Dunne et al. demonstrated that downregulation of this gene correlates with a poor treatment outcome in patients with acute myeloid leukemia [202].

Finally, both RPS15 and RPS24 genes encode for ribosomal proteins that are component of the 40S subunit. Interestingly, the first one has been found in different studies to appear mutated in chronic lymphocytic leukemia patients, as it lead to impaired p53 stability [203, 204]. The second one appears mutated in Diamond-Blackfan anemia, a congenital non-regenerative hypoplastic pathology, characterized by macrocytic anemia, erythroblastopenia, and an increased risk of developing leukemia [205]. Finally, C11orf58 gene encodes for the Chromosome 11 open reading frame 58, also be known as Small Acidic Protein (SMAP) [206].

The most common mutations associated with AML are in FLT3, NPM1, CEBPA, and TP53 [207]. However, the extensive work involving the sequencing of genomes and exomes of this malignancy has revealed a variety of recurrent gene mutations associated [208]. In the same way, it is well described that ALL is a multistep disease, caused by the accumulation of mutations involving cell growth, proliferation, survival, and differentiation [209]. The introduction of genome-wide technologies has contributed to elucidate the molecular mechanisms underlying

leukemic transformation in ALL and has allowed the identification of different subgroups [210].

Although the starting point of CML is well known, other genetic and cytogenetic changes play important roles in prognosis and treatment of this malignancy [211]. In this context, the mechanisms for insensitivity of CML stem remains unclear. Factors such as quiescence, high level of BCR-ABL expression, acquired mutations in the oncogene, and overexpression of membrane transporter proteins are very important [212]. CLL has the highest genetic predisposition of all hematologic neoplasms (approximately 5–10% of cases have a family history of CLL) [213]. It this disease, the genetic alterations have a great impact on the clinical course of the patients. Previous whole genome and exome sequencing studies have reveled recurrently mutated genes (such as NOTCH1, MYD88, TP53, ATM, SF3B1, FBXW7, POT1, CHD2, RPS15, IKZF3, ZNF292, ZMYM3, ARID1A, and PTPN11), but deletions of chromosome 13q14 is the most frequent aberration in CLL, occurring in 55% of cases [214].

6.5 CONCLUSIONS OF THE CHAPTER

Throughout this Chapter a new approach of the integrated pipeline presented at Chapter 5 has been designed. Datasets from different technologies and from different platforms have been integrated with the purpose of collecting a higher number of samples due to the lack of RNA-Seq samples of leukemia available at public databases, ensuring also the heterogeneity of the study. Furthermore, different types of leukemia series have been selected with the purpose of trying to find relevant biomarkers that allow to discern among the five classes. This study was not performed yet and both, the introduced pipeline at multiclass level and the metrics for the DEGs extraction, are a very novelty step in this field.

On one hand, in this multiclass study that considers different types of leukemia, an important new parameter called COV has been used for extracting the DEGs along with the classical LFC. This parameter extracts biomarkers that are able to discern one or different classes from the rest using paired combinations. Moreover, the ANOVA test performed has shown that this parameter has been crucial in the development of the study. Therefore, the combination of both the LFC and the COV for multiclass biomarkers selection is an important advance in this field with very promising results.

A set of 10 DEGs have been identified as possible candidate biomarkers and assessed through a set of machine learning classifiers. On the other hand, the classification results at the multiclass level using the extracted DEGs has shown a high percentage for both the Accuracy and the f1-score metrics, overcoming the 96% with only a small subset of ten genes. At sight of these results, our DEGs can discern among the five proposed classes and can shape a powerful tool that could be very useful for the clinicians in decision making.

Thereafter, the biological study of the small subset of ten genes reveals a strong relationship between nine of the ten genes with the leukemia disease. Concretely, these genes highlighted were related with relevant biological processes such as proliferation, apoptosis or migration among others in peripheral blood leukocytes including B-cells, T-cells and neutrophils. Furthermore, these genes have been previously related to the neoplastic process in the hematopoietic tissue being especially relevant the modulation of the DOCK2 gene which has shown therapeutic implications in leukemic cells.

In conclusion, the designed pipeline has allowed extracting multiclass DEGs, which are closely related with Leukemia and the different addressed sub-types.

7

KNOWSEQ: BEYOND THE TRADITIONAL GENE EXPRESSION PIPELINE

CONTENTS

This Chapter is a a reorganised and extended version of the under review in pre-print manuscript "KnowSeq R/bioc package: Beyond the traditional RNA-Seq pipeline. A breast cancer case study" [8], in which KnowSeq R/Bioc package is formally presented, as well as a breast cancer study performed using the package library to show its potential.

7.1 BACKGROUND

Along the experiments that support this doctoral thesis, a new pipeline has been designed and implemented to carry out complex transcriptomic analysis, involving several steps. Firstly, a raw samples treatment and a quality analysis is important to extract gene expression values. After that, a DEGs extraction and a subsequent gene enrichment can be performed. The development of intelligent predictive tools results essential in bioinformatics given that there exists a real need of assistance for decision-making systems towards precision medicine. Therefore, with this in mind, KnowSeq was born with the aim of encapsulating all those functionalities under the same scope. The main advantage of KnowSeq is the incorporation of ML steps such as feature selection and classifier design in the traditional transcriptomic pipeline. No tool exists in the research community that achieves this complete transcriptomic analysis, encapsulating all those steps in one single tool.

In order to show the functionalities provided by the general pipeline designed for the KnowSeq package, an application to a real problem is included in this Chapter together with a technical description of KnowSeq. Concretely, an analysis of a breast cancer set of patients collected from the controlled repository GDC portal is performed, keeping paired samples between tumour and control.

DNA sequencing studies are fundamental to win the battle against multifactorial and genetic diseases like cancer. Cancer is still the second cause of death worldwide, just behind cardiovascular disease. Although the survival rate is increasing gradually thanks to the medical researches and advances, the design of novel bioinformatic tools that allows processing and extracting multi-omics information from raw data is a crucial objective in this research area. Currently, there exist different tools that combine the different steps and technologies involved in this scope [215–217]. Nevertheless, to the best of our knowledge, there are no tools that integrate the traditional DEGs extraction steps with further, and nowadays essential, steps dealing with the intelligent predictive model design and biological enrichment processes. Those steps are focused on the design of decision-making system applied to precision medicine [5]. KnowSeq is thought to deal with the Homo Sapiens genome, but it is prepared to support any other species.

Specifically, the study addresses the application of KnowSeq for the search of relevant biomarkers for breast cancer detection as study case, together with their related biological information. This means that the

pipeline applied for breast cancer data here could be applied for any other study from genetic sources, no matter the pathology to address.

Although KnowSeq is focused on RNA-Seq as it is the most powerful and widespread genetic characterisation technology for transcriptome nowadays, Microarray can be analysed too. KnowSeq comprises a large part of the tools used in our previous studies/publications involving RNA-Seq data. Several cancer types were addressed such as breast cancer, skin cancer, leukemia and lung cancer and in all them relevant results were achieved [6, 7, 218, 219]. They widely confirm the validity of KnowSeq to carry out gene expression analysis.

In this scope, KnowSeq can be very helpful to perform these types of analysis to find and assess biomarkers. For that, this Chapter addresses a real application of KnowSeq to a set of raw breast cancer controlled data coming from GDC Portal [148]. Although KnowSeq allows the SRA/FASTQ alignment, GDC Portal does not supply those files, thus, for the analysis, 180 BAM files belonging to 90 breast cancer patients were used. For each patient, two samples were collected, a primary tumour sample and a solid tissue normal sample. Thanks to this, the experiment was designed with Tumour-Normal paired samples, which ensures the best experiment quality in terms of samples.

KnowSeq was published at Bioconductor in June 2019 and it has gradually achieved more downloads monthly. Figure 7.1 shows this increase in the number of downloads as well as the number of distinct IPs that have downloaded KnowSeq.

Although the methodology will be deeply explained in the next sections, a brief summary of KnowSeq giving basic information about its operation and possibilities is given herein: the download and alignment of the samples is performed automatically. Then the gene expression values are estimated and the quality analysis and batch effect removal is carried out. When the quality is checked, the DEGs between two or more conditions established by the user are extracted (e.g. treated vs non treated, normal vs control, etc). At this point, the traditional primary pipeline is over. Nevertheless, KnowSeq adds a set of steps to provide depth to the studies. In these new steps, a feature selection approach is included to estimate those genes that contain more information to discern between conditions (in our study case, normal vs tumoral tissue). Furthermore, it also includes a machine learning step with different algorithms and configurations to assess those DEGs. Finally, the most useful step at biological level added by KnowSeq is the DEGs enrichment step. In this sense, the tool allows retrieving information about the GOs of the DEGs, the involved pathways coloured

GOs: Gene Ontologies

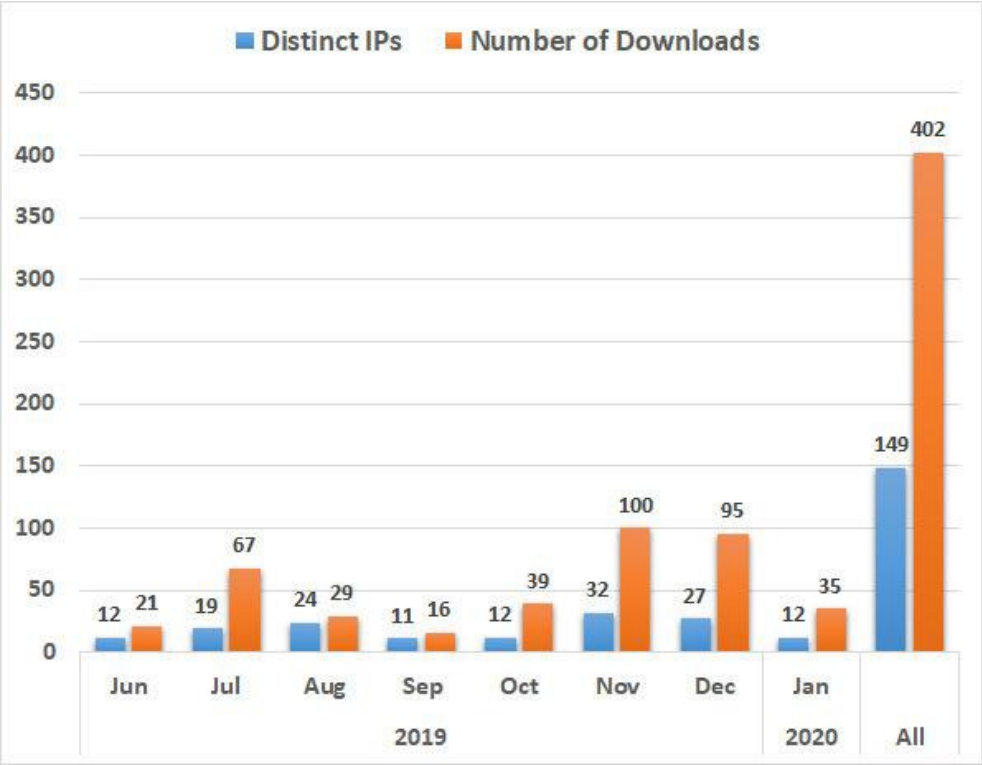

Figure 7.1: KnowSeq downloads statistics from its publication in June 2019 to January 2020. It can be seen how the number of total downloads it increasing monthly.

according to the gene expression level of the samples and a list of diseases related with the DEGs and different combination of those DEGs. Due to all of these reasons, KnowSeq is the only R/bioc package that allows performing a complete RNA-Seq study by using the same single tool and programming language during all the process.

7.2 IMPLEMENTATION

This section describes all the steps implemented by the KnowSeq pipeline, being also applied to this research. Figure 7.2 represents the whole pipeline and four different steps can be clearly distinguished: Webdata Resources gathering, RNA-Seq RAW data processing, Biomarkers identification and assessment and DEGs enrichment methodology. On this basis, each step is presented in one subsection with the purpose of giving a deeper explanation for each of them. It is to be highlighted that KnowSeq is designed to achieve a high modularity. This means that each of the steps and sub-steps conforming KnowSeq can be perfectly replaced, provided that the inputs maintain

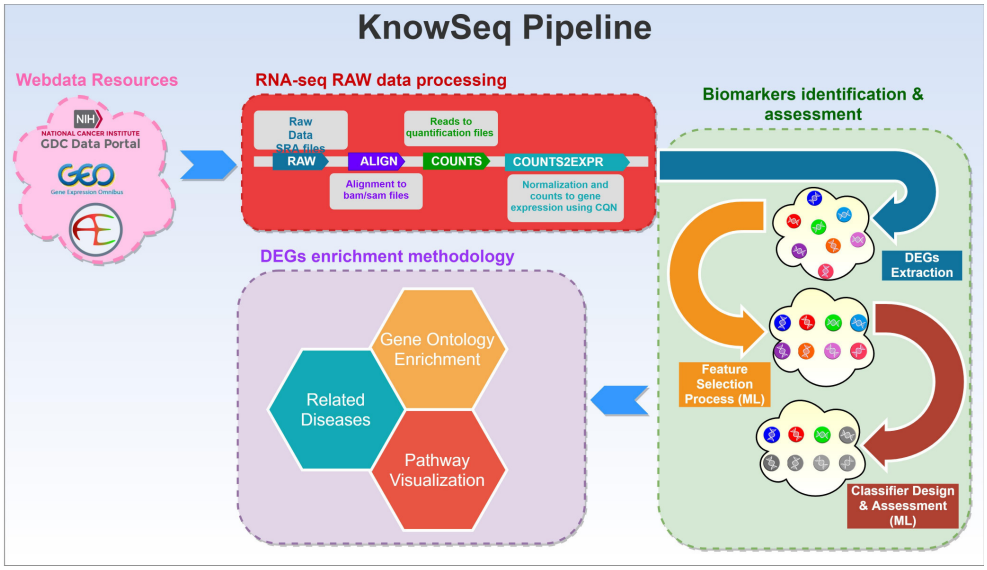

Figure 7.2: Pipeline implemented by KnowSeq R/bioc package. In the pipeline are the traditional steps in the RNA-Seq data pipeline together with the new steps added by KnowSeq.

the same data type. Because of this, KnowSeq can be easily adapted even for different species and biological data types not explicitly addressed in this first version of our tool. Furthermore, the pipeline can be launched from different steps depending on the type of input files (e.g. SRA/FASTQ, BAM or counts). In order to summarise the different functions available in the package, Table 7.1 shows for each function the name, the pipeline step where this function is used, the description of the functionality and the different options implemented inside the function. Furthermore, the functions inside the table are ordered by the steps in Figure 7.2.

7.2.1 *Webdata Resources*

One of the hardest step in any biological study is the data gathering. KnowSeq allows to automatize the download of public and controlled samples from the most renowned web platform databases: NCBI/GEO, ArrayExpress and GDC Portal. The data from NCBI/GEO and ArrayExpress are publicly available and from these web platforms KnowSeq only requires the series ID to download. However, the raw data in GDC Portal are under restricted access and an authorisation is required via token file. If the user has this token file, GDC Portal raw data could be automatically downloaded by calling the function *gdcClientDownload*.

Table 7.1: Table that contains the most important functions in KnowSeq. For each function, the name, the pipeline step where this function is, the description and the options inside the function are showed.

Function Name	Pipeline step	Description (options)
downloadPublicSeries	Webdata resources	Download series from GEO and AE
gdcClientDownload	Webdata resources	Download data from GDC-Portal
rawAligment	RNA-seq RAW data processing	Raw data aligment with different algorithms (tophat2, salmon, hisat2, kallisto)
countsToMatrix	RNA-seq RAW data processing	Convert count files to matrix
calculateGeneExpressionValues	RNA-seq RAW data processing	Gene expression values calculation
RNAseqQA	Biomarkers Identification & Assessment	Expression matrix QA
getAnnotationFromEnsembl	Biomarkers Identification & Assessment	Retrieve information for a DEGs list
batchEffectRemoval	Biomarkers Identification & Assessment	Batch effect treatment (Combat, SVA)
limmaDEGsExtraction	Biomarkers Identification & Assessment	Biclass and multiclass DEGs extraction
dataPlot	Biomarkers Identification & Assessment	Plots different data information and results (boxplot, orderedBoxplot, genesBoxplot, heatmap, confusionMatrix, classResults)
featureSelection	Biomarkers Identification & Assessment	Feature selection for a DEGs matrix (mRMR,RF)
knn_CV	Biomarkers Identification & Assessment	Run a knn-CV for a DEGs matrix
knn_test	Biomarkers Identification & Assessment	Run a knn-test
rf_CV	Biomarkers Identification & Assessment	Run a rf-CV for a DEGs matrix
rf_test	Biomarkers Identification & Assessment	Run a rf-test
svm_CV	Biomarkers Identification & Assessment	Run a svm-CV for a DEGs matrix
svm_test	Biomarkers Identification & Assessment	Run a svm-test
DEGsToDiseases	DEGs Enrichment methodology	Related diseases for a DEGs list (targetValidation, genes2Diseases)
geneOntologyEnrichment	DEGs Enrichment methodology	Gene ontology for a DEGs list
DEGsPathwayVisualization	DEGs Enrichment methodology	Pathway visualization for a DEGs list

Users need then to construct or download a CSV/TSV files with the information of each data/series. Using the specific function *downloadPublicSeries*, an automatic download of these supporting files are made with series that belong to NCBI/GEO and ArrayExpress. Thanks to this, it is very simple to specify and gather the series and samples required to perform an analysis. All the data used to carry out our study, were downloaded by using this method.

CSV: Comma-Separated Values
TSV: Tab-Separated Values

7.2.2 *RNA-Seq RAW data processing*

When working with RNA-Seq data, an alignment process is required by using the human reference genome in order to obtain the count files to perform the DEGs analysis. KnowSeq allows to download the Human Reference Genome GRCh37 and GRCh38 from Ensembl, although whichever reference genome can be used if the user indicates the path to the file. In this step, the raw files in SRA/FASTQ [220] formats are processed to obtain the BAM files. This is performed through the use of *rawAlignment* KnowSeq function. For this process, KnowSeq counts on the samtools [221] and four of the most well-known aligners with the purpose of giving to the academics not only one option to apply. The aligners are tophat2, hisat2, salmon and kallisto [152, 153, 222, 223]. Furthermore, the Htseq-count tool extracts the count files for each samples [154].

GRCh: Genome Reference Consortium human

Finally, through the function *countsToMatrix*, all the count files are merged in one aggregated matrix with edgeR. By the use of the function *calculateGeneExpressionValues*, the equivalent gene expression values are calculated with cqn R package [224, 225]. By applying this step to the counts files, the desired number of samples can be automatically processed. It is highly recommended to run the raw data alignment in a computer cluster as the use of the tools involves high computational cost for this task.

7.2.3 *Biomarkers identification & assessment*

To achieve the DEGs extraction, KnowSeq implements a step that allows to do that task for any specie and disease with genetic relation. Moreover, our tool incorporates mechanisms to study the quality of the samples and the batch effects. It also includes the possibility of plotting all the required charts for the graphical assessment of the samples

(e.g. boxplots by samples, boxplots by genes, heatmaps...) in a unique function named as *dataPlot*.

Although the output of the KnowSeq aligner step can be used as input of this step, the user can also introduces its own samples matrix. KnowSeq has been designed as a modular tool, this meaning that the user can carry out all the study by using KnowSeq or can use only the steps in which the user has interest.

DEGs extraction is a very delicate process because the samples must pass a strong quality analysis and batch effect removal steps. KnowSeq has a quality analysis step by using arrayQualityMetrics package adapted to RNA-Seq by running the function *RNAseqQA*. This package counts on several statistical analysis to detect possible outliers in the samples [155]. Furthermore, our tool also has graphical representation such as gene expression boxplots disordered and ordered by class or label, heatmaps and gene by gene boxplot even allowing multiclass representation. It is very crucial to perform the quality analysis in a rigorous manner to ensure the correct development within the rest of the study. Even though the quality analysis is well done, there still exists the possibility of having batch effect among the chosen samples or series. The batch effect is a deviation effect in the gene expression values due to several external technical factors (origin, sequencing hour, lab technician, among others) and it is very hard to treat [159]. KnowSeq allows to use two of the most relevant algorithms to treat batch effect such as ComBat for predefined batch groups and sva for unknown batch groups [226] through the function *batchEffectRemoval*.

In order to perform DEGs extraction, limma R package is used, with the peculiarity that KnowSeq automatically detects the number of different classes or labels and consequently applies limma biclass or multiclass [163]. For the multiclass, the coverage parameter introduced at Chapter 6 that allows detecting DEGs that are expressed for more than one biclass comparison has been added to KnowSeq. This DEGs extraction is carried out by using the function *limmaDEGsExtraction*.

Next, a FS process is highly recommended for precision medicine to reduce the system complexity, diminishing the number of genes and, helping to make clinical decisions [227–229]. For that, KnowSeq allows to apply with the function *featureSelection*, two different feature selection algorithms, mRMR [230] and RF as feature selector [139]. These algorithms create a ranking of DEGs in order to increase the classification rate by putting the DEGs with more information for the classifier listed at the top.

Finally, for the supervised ML process, KnowSeq allows using three of the most relevant classifiers: SVM [126, 231], k-NN [232] and RF [233]. There exist two versions for each classifier in KnowSeq, one version with CV (knn_CV, svm_CV & rf_CV) in which the user decides the number of fold and the data partitions always considering the representation of all the classes, and other version for testing (knn_test, svm_test & rf_test), by using a test dataset without CV independent from the dataset used for the DEGs extraction and CV assessment. Furthermore, for the three algorithms the hyperparameters are optimised, searching the acquisition of the best model for each analysis. Moreover, KnowSeq allows plotting the graphical representation of the results, including the confusion matrix, the sensitivity, the specificity and the f1-score. This gives to the user the possibility to perform a complete analysis and assessment of the addressed problem in a very simple and quick way.

7.2.4 *DEGs enrichment methodology*

This tool is designed to automatise the knowledge extraction whatever is being the disease and for that, the last step of KnowSeq pipeline attains biological knowledge related to the final DEGs candidates. This knowledge must be interpreted by a clinician or a person with biological profile. In this sense, KnowSeq can retrieve information from three different sources to help with that interpretation. One of these sources is the GOs enrichment with information about the biological functions and locations of the DEGs [234, 235]. The three available GO domains are queried by our tool: the BPr, the MF and the CC. Thanks to this, the biological functions related with the DEGs can be acquired in order to perform a more deeply study trying to find connections with the addressed disease. For the GOs enrichment, the topGO R package is used [236] and, in order to carry out the GOs retrieval, KnowSeq has the function *geneOntologyEnrichment*.

BPr: Biological Process
MF: Molecular Function
CC: Cellular Component

The second source of biological information is the pathway visualisation. Nowadays it is well known that the interaction of several genes whether can lead to a genetic disorder or not. Genes interacting among them in the same biological function are distributed in the same pathway. For that reason, it is important to know not only the expression of the DEGs but also their interactions with genes that belong to the sames pathways of those DEGs. The pathview package allows to colour the pathways depending on the expression values of the genes inside the pathways [237]. KnowSeq has kept this idea to automatically retrieve and colour all the pathways related with the final DEGs candidates and

KEGG: Kyoto Encyclopedia of Genes and Genomes

listed in the KEGG database [238]. With this implementation, it is easy to know if the expression of the DEGs and the surrounding genes are affecting a critical function in the disease development. The function *DEGsPathwayVisualization* takes care of this process.

The last source of biological information implemented in KnowSeq is the related diseases retrieval and it is performed executing the function *DEGsToDiseases*. In this step, all the related diseases of the DEGs candidates listed in the literature are obtained with the purpose of finding possible relation with the pathology addressed and with other possible precursor pathologies. Furthermore, the diseases related with a set of DEGs are also obtained in order to find possible DEGs that are related with the same pathology. This information can be attained from two different sources: the first one is the Gene Set to Diseases web platform [239] and the second one is the targetValidation [240] web platform. Then, the acquired diseases are correctly formatted by KnowSeq to do more readable this information for the user.

With the information collected by KnowSeq automatically from the three different sources, a strong biological enrichment process is done in order to build a biological profile for each of the DEGs without requiring external tools.

7.3 BREAST CANCER APPLICATION

This section is divided in four subsections, one for the information about the data acquisition and three representing categories of results that were obtained for this study. For the last three subsections, the first one is focused on the final candidate DEGs extraction and the restrictions imposed to achieve them. The second one shows the assessment of those DEGs by using machine learning techniques with the main goal of finding a smaller sub-set of DEGs. Finally, the last one describes the enrichment of the sub-set of DEGs in order to find relevant biological information about them in an easy way by using KnowSeq.

7.3.1 *Data preparation & description*

TCGA: The Cancer Genome Atlas

All the data or samples used in this case of study come from TCGA and have been acquired through GDC Portal platform. For this breast cancer

study, 90 patients were selected with the condition of having BAM files from both solid normal and primary tumour tissues for each patient. With this condition the paired datasets are ensured, achieving the best quality conditions in terms of samples for the study. Primary breast cancer is a tumour that still remains inside the breast or the lymph nodes (glands) under the arm. On the other hand, the solid normal tissue is collected from the adjacent healthy tissue to the primary breast tumour. A table with all patient data to replicate the study is available at Table 7.2. In order to perform a more robust study, two different datasets will be taken into account. The first dataset is formed by 80 patients and will be used to extract the DEGs. The second dataset is conformed for the 10 remaining patients and will be only used for testing those DEGs in a machine learning step. Thanks to this division, the DEGs extracted will be independent of the samples used to assess them.

7.3.2 *DEGs extraction and analysis*

The importance of achieving robust biomarkers is crucial for this type of problems thus, it is important to correctly select the imposed restriction to extract the set of DEGs. To find them, as mentioned before, 80 patients were used and the 10 remaining patients were kept only for testing those DEGs in the machine learning step. This separation is very important to test the DEGs in patient never seen before in the process, bringing robustness to the results and avoiding overfitting. The quality analysis was first performed to the 80 patients and no outlier was detected among them. Then, the batch effect removal step was applied taking into account that the possible batches were unknown. The SVA algorithm [161] was performed to find the surrogate variables in order to create a model considering those variable to remove the deviations. It is critical to remove the batch effect in order to correct the data but without removing any possible deviations caused by biological processes. After the quality analysis and the batch effect correction steps, DEGs candidates can now be extracted. To carry out this extraction, the thresholds imposed were very restrictive, using two well-known statistics values for filtering: the LFC greater or equal than 3 and the P-value less or equal than 0.001. Applying these restrictions, a total amount of 50 DEGs candidates ordered by LFC were extracted.

Table 7.3 shows those DEGs with several statistical values that describe at numerical level why those genes have been selected as DEGs. Those values are the five statistics seen at Chapters 5 and 6. As it can be seen,

all the DEGs candidates pass the imposed restrictions and they will be assessed to corroborate their validity.

Table 7.2: Patients Samples IDs from GDC used for the development of this research.

BRCA Project GDC ID		
TCGA-BH-A0AU	TCGA-BH-A0DZ	TCGA-BH-A1ET
TCGA-BH-A0AY	TCGA-BH-A0E0	TCGA-BH-A1EU
TCGA-BH-A0AZ	TCGA-BH-A0E1	TCGA-BH-A1EV
TCGA-BH-A0B3	TCGA-BH-A0H5	TCGA-BH-A1EW
TCGA-BH-A0B5	TCGA-BH-A0H7	TCGA-BH-A1F0
TCGA-BH-A0B7	TCGA-BH-A0H9	TCGA-BH-A1F2
TCGA-BH-A0BA	TCGA-BH-A0HA	TCGA-BH-A1F6
TCGA-BH-A0BC	TCGA-BH-A0HK	TCGA-BH-A1F8
TCGA-BH-A0BJ	TCGA-A7-A13E	TCGA-BH-A1FC
TCGA-BH-A0BM	TCGA-A7-A13F	TCGA-BH-A1FD
TCGA-BH-A0BQ	TCGA-A7-A13G	TCGA-BH-A1FE
TCGA-BH-A0BT	TCGA-E2-A153	TCGA-BH-A1FG
TCGA-BH-A0BV	TCGA-E2-A158	TCGA-BH-A1FJ
TCGA-BH-A0BW	TCGA-E2-A15I	TCGA-BH-A1FM
TCGA-BH-A0BZ	TCGA-E2-A15K	TCGA-BH-A1FN
TCGA-BH-A0C0	TCGA-E2-A15M	TCGA-BH-A1FR
TCGA-BH-A0C3	TCGA-BH-A18J	TCGA-BH-A1FU
TCGA-A7-A0CE	TCGA-BH-A18K	TCGA-E2-A1LB
TCGA-A7-A0D9	TCGA-BH-A18L	TCGA-E2-A1LH
TCGA-A7-A0DB	TCGA-BH-A18M	TCGA-E2-A1LS
TCGA-BH-A0DD	TCGA-BH-A18N	TCGA-BH-A203
TCGA-BH-A0DG	TCGA-BH-A18P	TCGA-BH-A204
TCGA-BH-A0DH	TCGA-BH-A18Q	TCGA-BH-A208
TCGA-BH-A0DK	TCGA-BH-A18R	TCGA-BH-A209
TCGA-BH-A0DL	TCGA-BH-A18S	TCGA-AC-A23H
TCGA-BH-A0DO	TCGA-BH-A18U	TCGA-GI-A2C8
TCGA-BH-A0DP	TCGA-BH-A18V	TCGA-GI-A2C9
TCGA-BH-A0DQ	TCGA-E2-A1BC	TCGA-AC-A2FB
TCGA-BH-A0DT	TCGA-BH-A1EN	TCGA-AC-A2FF
TCGA-BH-A0DV	TCGA-BH-A1EO	TCGA-AC-A2FM

Table 7.3: Table with the 50 DEGs candidates extracted for this study and several statistical values for those DEGs.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
COL10A1	-7.1720	14.8062	-23.9116	1.3885e-20	2.0014e-18	37.0692
CST1	-6.8780	12.0813	-15.2775	2.2569e-15	4.2166e-14	24.9740
MMP13	-6.6652	12.6433	-23.5212	2.1842e-20	2.9420e-18	36.6168
LINC01614	-6.5380	14.5261	-20.4209	1.0349e-18	6.4411e-17	32.7475
SLC24A2	-6.2606	10.8512	-19.2851	4.8451e-18	2.2617e-16	31.1924
COL11A1	-5.8136	16.1869	-26.8715	5.4897e-22	2.0147e-19	40.2812
MMP11	-5.5748	15.7695	-27.1724	4.0272e-22	1.6695e-19	40.5878
CA4	5.4495	12.3041	22.7570	5.4093e-20	5.8144e-18	35.7099
IBSP	-5.4158	9.7764	-17.3262	8.4221e-17	2.4182e-15	28.3072
PLPP4	-5.4069	11.0745	-20.3021	1.2119e-18	7.3562e-17	32.5887
MMP1	-5.2966	12.5466	-17.7029	4.7652e-17	1.4990e-15	28.8834
LEP	5.2911	15.4111	12.7145	2.3314e-13	2.3836e-12	20.2601
MYOC	5.2690	11.2076	15.8110	9.2770e-16	1.9577e-14	25.8761
NPY2R	5.1723	12.0594	14.2605	1.3176e-14	1.9586e-13	23.1819
EPYC	-5.1495	9.8030	-16.9509	1.5009e-16	4.0510e-15	27.7222
LINC00922	-4.9341	9.0520	-19.3696	4.3080e-18	2.0622e-16	31.3109
CST2	-4.8586	11.9503	-15.2360	2.4212e-15	4.4841e-14	24.9026
CST4	-4.7626	10.3939	-14.0907	1.7854e-14	2.5427e-13	22.8731
CD300LG	4.7297	15.0205	27.8248	2.0795e-22	1.1035e-19	41.2410
ANGPTL7	4.6405	13.1176	14.4855	8.8454e-15	1.3878e-13	23.5869
SCARA5	4.6061	14.4664	23.1422	3.4131e-20	4.0065e-18	36.1706
ADGRD2	4.5781	9.9209	19.5686	3.2720e-18	1.6478e-16	31.5882
AC044784.1	-4.5405	12.4070	-14.0387	1.9606e-14	2.7528e-13	22.7780
OPRPN	4.4570	11.9701	10.6963	1.4700e-11	9.6064e-11	16.0423
GLYAT	4.4221	11.4900	14.8858	4.4038e-15	7.5592e-14	24.2953
LINC01705	-4.4206	8.2294	-21.5106	2.5216e-19	2.0538e-17	34.1667
AC093895.1	-4.4142	7.2190	-13.6822	3.7519e-14	4.8207e-13	22.1182
DLK1	4.4100	11.1670	11.1259	5.8440e-12	4.1775e-11	16.9813
DSCAM-AS1	-4.3949	12.0297	-8.8010	1.1385e-09	5.0128e-09	11.6176
PLAC1	-4.3843	8.8461	-15.0985	3.0582e-15	5.4733e-14	24.6655
COMP	-4.3195	15.8240	-17.6860	4.8879e-17	1.5286e-15	28.8577
LINC02408	-4.3143	7.0213	-13.9943	2.1243e-14	2.9515e-13	22.6965
AC104407.1	4.2835	13.0120	13.0863	1.1424e-13	1.2825e-12	20.9859
PITX1	-4.2759	14.3922	-18.0804	2.7203e-17	9.3295e-16	29.4503
CXCL2	4.2568	17.2496	17.7874	4.1995e-17	1.3449e-15	29.0112
WIF1	4.2474	12.1155	13.5579	4.7191e-14	5.8834e-13	21.8850
PLIN4	4.2328	18.3265	16.4032	3.5554e-16	8.5226e-15	26.8485
CCL11	-4.2158	12.6651	-16.2890	4.2684e-16	9.9858e-15	26.6633
VEGFD	4.1739	13.6021	21.6667	2.0704e-19	1.7581e-17	34.3646
CSN1S1	4.1379	11.1700	5.9985	1.6242e-06	4.1544e-06	4.2666
LRRC15	-4.1185	16.6620	-18.0938	2.6671e-17	9.2062e-16	29.4702
CIDEC	4.0943	16.1264	12.6589	2.5971e-13	2.6166e-12	20.1503
AC112721.2	-4.0904	9.5582	-19.0034	7.1918e-18	3.1717e-16	30.7939
CNTNAP2	-4.0853	17.0082	-12.7531	2.1636e-13	2.2371e-12	20.3361
S100P	-4.0752	15.6872	-10.5751	1.9148e-11	1.2194e-10	15.7732
ADIPOQ	4.0540	17.3106	10.6432	1.6498e-11	1.0663e-10	15.9248
WT1	-4.0456	9.7736	-9.5566	1.9007e-10	9.7380e-10	13.4377
GPD1	4.0364	18.0172	13.6989	3.6389e-14	4.7058e-13	22.1493
CHRNA6	-4.0287	8.9468	-13.2380	8.5741e-14	9.9333e-13	21.2777
TRHDE-AS1	4.0243	12.2240	12.4349	4.0261e-13	3.8499e-12	19.7042

Furthermore, the Figure 7.3 represents an expression heatmap that graphically shows important differences. It can be seen how the DEGs expression levels are different between both groups (normal and tumour). Due to that expression differences, ML models could learn the way of discerning among the addressed groups in order to determine the validity of these DEGs for this problem.

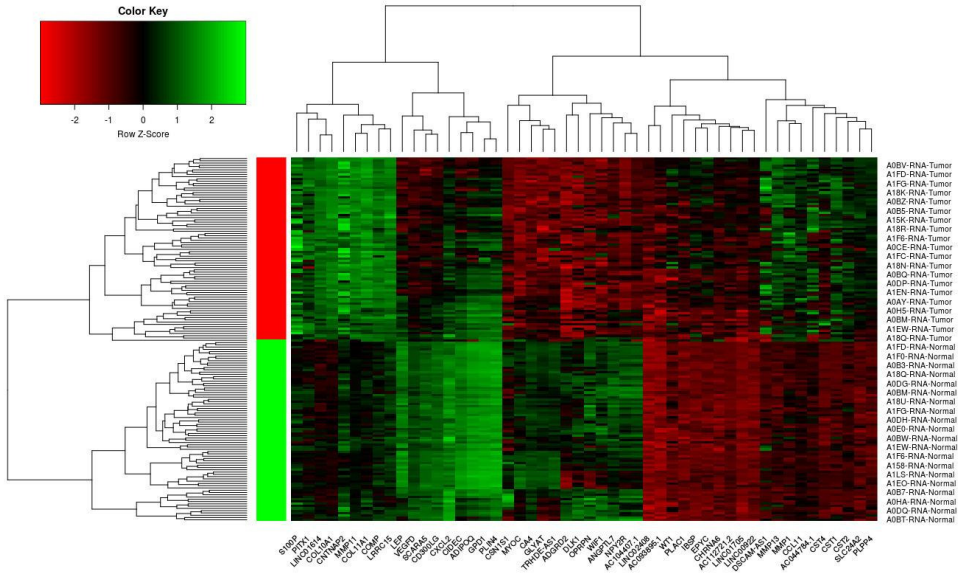

Figure 7.3: Heatmap of the 50 DEGs candidates clearly showing differences between tumour and normal samples.

7.3.3 *Machine Learning assessment*

KnowSeq includes a ML step to assess those DEGs and their capability to discern among the considered pathologies. Through this process, a smaller subset of DEGs can be achieved with the purpose of finding a more reduced gene signature candidate. For that, KnowSeq has three different supervised ML algorithms and two different FS methods as was explained before. This ML step has two different approaches. The first one is the application of a CV process to assess the DEGs with the training patients. The second one is the test process in which our DEGs are evaluated by using the 10 test patients previously chosen only for this purpose.

Firstly, a 10-CV step was applied in order to see the behaviour of the classifier with the 80 patients training dataset when those DEGs are used for classify. Thereupon, all the different combination of classifiers with FS algorithms reached better results than without applying FS, recognising all the training samples with a few number of genes. SVM

and RF acquired outstanding results, but k-NN had slightly better results than the other two algorithms.

However, it is important to know how the classifier behaves with samples never seen before in order to simulate a real clinical case. This is the reason to create a test process with the 10 patients (20 samples) datasets. These patients were left out at the beginning for all the study to be used now to assess the DEGs. Different matches, or combinations between classifiers and FS algorithm, were executed with the purpose of searching the combination with the best results. Those combinations are the possible permutations resulting from the different classifiers (SVM, k-NN and RF) with none feature selection (No F.S.) and with the different feature selection algorithms (mRMR and RF f.s.). Table 7.4 contains the results for all these combinations depending on the number of genes used to classify. It is important to highlight that with only 3 genes, k-NN reached 100% of accuracy when mRMR and RF f.s. were applied. SVM also reached 100% with RF f.s. but no with mRMR. For its part, RF only achieved 100% with 10 genes and by using RF f.s. algorithm. Although all of them achieved prominent results, k-NN obtained the best results whatever being the FS algorithm and the number of genes used. As it can be seen, with only 3 genes selected by the FS process from all the DEGs, all the test patients were perfectly recognised for the machine learning designed models. This means that KnowSeq brings the support to create intelligent systems with the capability of extracting relevant biomarkers that are useful to discern among the addressed diseases or states.

Once the classification is done, it is very helpful to see graphically the gene expression differences that exist between the tumour samples and the normal samples for the three 3 DEGs that discriminate perfectly the test patients. In order to carry out this representation, KnowSeq counts contains the *dataPlot* function in mode genesBoxplot. Figure 7.4 represents the genes Boxplots for the top 3 DEGs without apply FS (ordered by LFC), applying mRMR and applying RF f.s. respectively. In this figure, the first gene selected by the three methods (No F.S, mRMR and RF f.s.) is the same (COL10A1). However, the second gene selected by mRMR and RF f.s.(VEGFD & MMP11), both are different than the second gene with more LFC (CST1). The third and last gene selected by mRMR and RF f.s. (PITX1, LINC01614), are also different again than the third gene with higher LFC (MMP13). Nevertheless, even though the genes selected by LFC have more differences in average expression between the states, the genes selected by the FS algorithms discern better between such states, thus reaching better classification results as can be seen in Table 7.4. Consequently, adding a refined FS as well as a classification algorithm based on ML technology proved that the

Table 7.4: Table that contains the test results for the different combinations of feature selection algorithms with the classifiers depending on the number of DEGs selected.

	No F.S.			*mRMR*			*RF f.s.*		
n. Genes	3	5	10	3	5	10	3	5	10
SVM	85%	90%	95%	95%	95%	100%	100%	95%	100%
k-NN	90%	85%	100%	100%	100%	100%	100%	100%	100%
RF	85%	90%	95%	90%	70%	95%	85%	85%	100%

selected DEGs potentially improve the differentiation of states against classical metrics like LFC with a few number of DEGs.

It is a priority in this research to minimise the number of genes and maximise the final achieved accuracy. This way, a very small sub-set of DEGs can be found to have the capability of discerning among the studied states. Nevertheless, KnowSeq is flexibly prepared to use and analyse as many genes or DEGs as the user requires. Also, it is important to highlight that, even though a bi-class problem was taken into account for this study, KnowSeq is designed to analyse any multiclass problem. In this sense, the confusion matrix, the f1-score, the sensitivity and the specificity metrics calculation are considered by our package.

7.3.4 *DEGs enrichment*

At this point of the study, our DEGs have been assessed by applying a machine learning process. Nevertheless, those DEGs must be interpreted at biological level by experts in the field. In order to help with the biological interpretation, KnowSeq has a last step in its pipeline created solely and exclusively to this purpose (DEGs Enrichment). Although this study searches a very small subset of DEGs, the enrichment step in KnowSeq does not depend on the number of DEGs, because the package can compute all of them.

Previously, in the machine learning results, the 10 test patients were totally recognised with only 3 genes selected by both RF f.s. and mRMR in conjunction with the k-NN classifier. Firstly, the relationship between those 6 DEGs and breast cancer will be searched by using the function *DEGsToDiseases* with the targetValidation platform selected. This platform has several scores to determine if a gene is related with the different possible diseases based on the information collected by

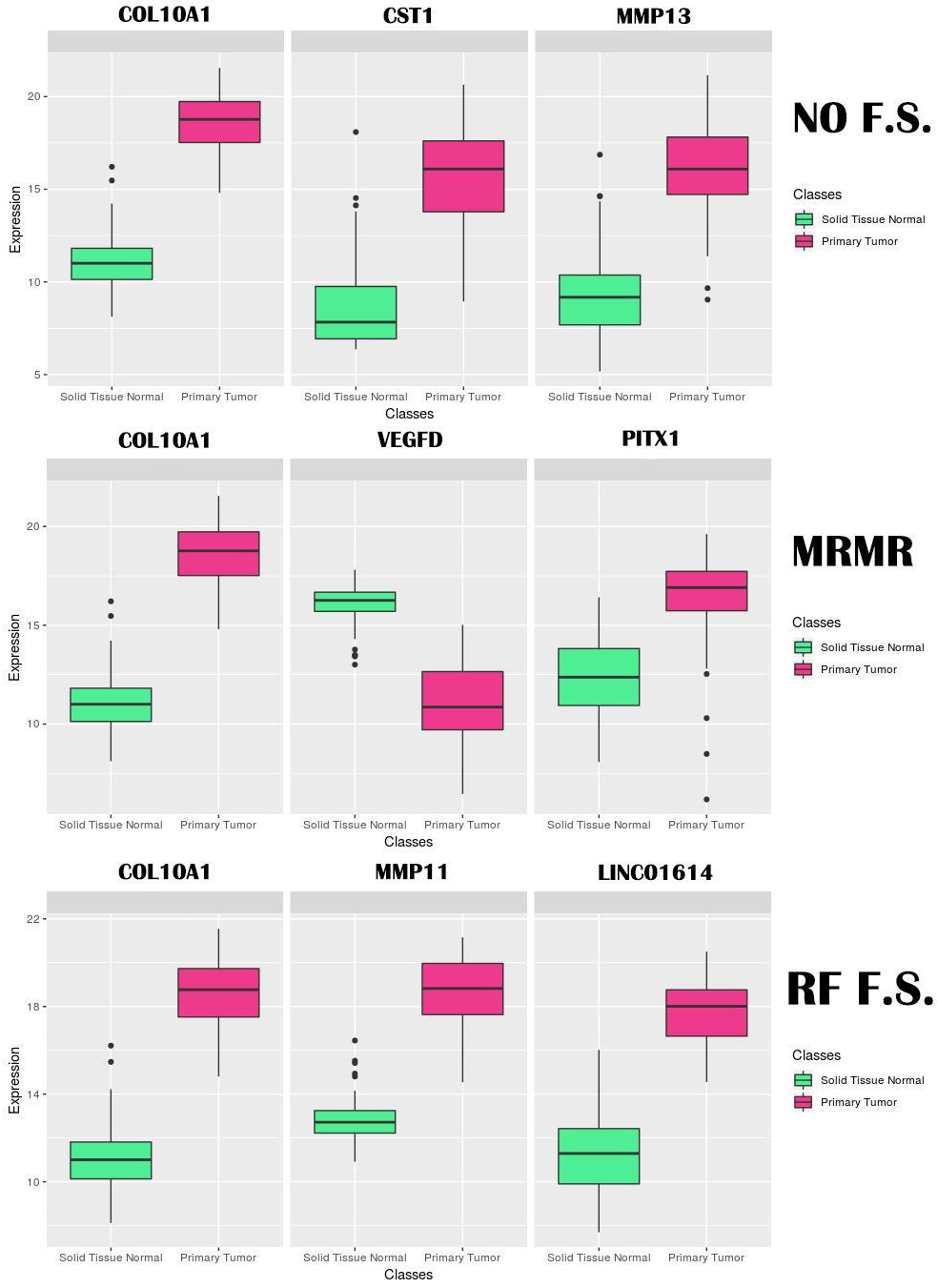

Figure 7.4: Boxplots of the 3 first DEGs selected by KnowSeq without feature selection algorithm and with mRMR and RF.

the web platform. Those scores increase when the association increases too, that meaning a strong association with the selected disease.

From the 9 DEGs commented before, only two DEGs from RF f.s. (COL10A1 & MMP11) and two DEGs from mRMR (COL10A1 & VEGFD) have a strong reported relation with breast cancer and one of them is common to No F.S., mRMR and RF f.s. (COL10A1). The 6 remaining DEGs have no relation or the relation is poor (a very low association score). It is very interesting to note that only the first gene of the top 3 DEGs without FS has important relations with breast cancer, although they are the DEGs with the higher LFC or P-value. Therefore, the use of a FS step in this case has remarkably supposed the determination of DEGs in the first positions more related with breast cancer. This fact clearly improves the classification accuracy as shown in the previous sub-section. Hence, the 3 breast cancer reported DEGs will be used for the enrichment.

As can be seen at the heatmap in Figure 7.5, the 3 final DEGs (COL10A1, MMP11 and VEGFD) clearly distinguish between tumour and control samples. In the case of COL10A1 and MMP11 are inhibited in tumour samples with regards to normal samples. Otherwise, VEGFD are over-expressed in tumour samples in comparison with normal samples. It is very important to know about these differences in order to find drugs or treatment that can correct them.

For these DEGs, a set of scores are showed in Table 7.5. These 4 scores acquire values between 0 and 1: the Literature score is calculated based on the evidences in the literature of the involvement of a gene with the corresponding cancer (breast cancer in this case); the RNA Expression score uses data from Expression Atlas to see if a gene has differences at expression level for a disease; the Affected Pathways score evaluates from the reactome platform if the gene is involved in relevant pathways for the disease. Lastly, the final association score is calculated from the previous scores. As can be seen in the table, the three genes have a strong final association, so they are highly involved in breast cancer. From this point, the experts in the field have an important overview of the genes to continue investigating them.

Once the disease relationship process has been carried out, the next step is the GOs enrichment. For this process the same 3 DEGs are used and the five most important GOs for the three DEGs and for the three different ontologies (BPr, MF & CC) will be retrieved with the function *geneOntologyEnrichment*. Table 7.6 shows the top 5 GOs for our 3 DEGs. As it can be seen, the VEGFD gene does not appear for any GOs terms in the top 5, but only GOs related with COL10A1 and MMP11 genes.

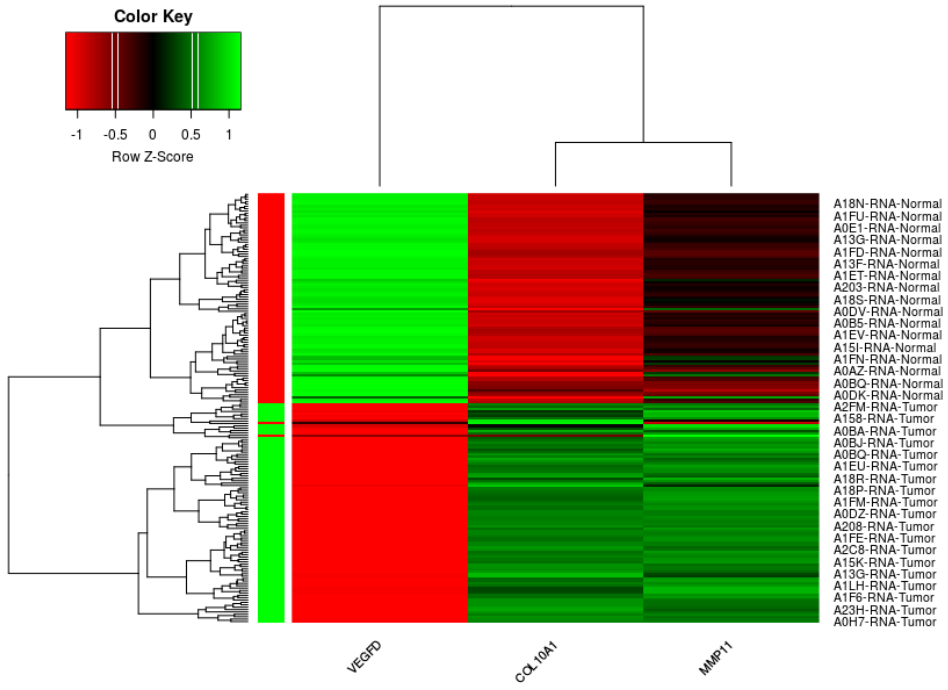

Figure 7.5: Heatmap for the 3 breast cancer related DEGs selected the feature selection algorithms.

Table 7.5: Table with the information about the association scores for the final 3 DEGs to study.

Gene	Liter. Score	RNA Exp. Score	Affected Paths. Score	Final Score
COL10A1	0.0372	0.1787	0.6835	0.7323
MMP11	0.1935	0.1094	0.6065	0.6670
VEGFD	0.1169	0.1400	0.6948	0.7428

Only increasing the maximun number of retrieved GOs, GOs related to the VEGFD were retrieved. Thanks to this step, the BPr, the MF and the CC of the DEGs are stored by KnowSeq to help users knowing the biological domain of each DEGs and studying possible relations with processes that could lead to develop cancer.

Table 7.6: Table that contains top 5 GOs for the three different ontologies for the 3 final DEGs

Ontology	GO.ID	Term	GO_Genes	Description
BP	GO:0001501	skeletal system development	COL10A1	The process whose specific outcome is the progression of the skeleton over time, from its formation to the mature structure. The skeleton is the bony framework of the body in vertebrates (endoskeleton) or the hard outer envelope of insects (exoskeleton or dermoskeleton).
	GO:0016043	cellular component organization	COL10A1,MMP11	A process that results in the assembly, arrangement of constituent parts, or disassembly of a cellular component.
	GO:0030198	extracellular matrix organization	COL10A1,MMP11	A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of an extracellular matrix.
	GO:0043062	extracellular structure organization	COL10A1,MMP11	A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of structures in the space external to the outermost structure of a cell. For cells without external protective or external encapsulating structures this refers to space outside of the plasma membrane, and also covers the host cell environment outside an intracellular parasite.
	GO:0071840	cellular component organization or bioge...	COL10A1,MMP11	A process that results in the biosynthesis of constituent macromolecules, assembly, arrangement of constituent parts, or disassembly of a cellular component.
MF	GO:0005198	structural molecule activity	COL10A1	The action of a molecule that contributes to the structural integrity of a complex or its assembly within or outside a cell.
	GO:0005201	extracellular matrix structural constitu...	COL10A1	The action of a molecule that contributes to the structural integrity of the extracellular matrix.
	GO:0030020	extracellular matrix structural constitu...	COL10A1	A constituent of the extracellular matrix that enables the matrix to resist longitudinal stress.
	GO:0043167	ion binding	COL10A1,MMP11	Interacting selectively and non-covalently with ions, charged atoms or groups of atoms.
	GO:0043169	cation binding	COL10A1,MMP11	Interacting selectively and non-covalently with cations, charged atoms or groups of atoms with a net positive charge.
CC	GO:0005581	collagen trimer	COL10A1	A protein complex consisting of three collagen chains assembled into a left-handed triple helix. These trimers typically assemble into higher order structures.
	GO:0005783	endoplasmic reticulum	COL10A1	The irregular network of unit membranes, visible only by electron microscopy, that occurs in the cytoplasm of many eukaryotic cells. The membranes form a complex meshwork of tubular channels, which are often expanded into slitlike cavities called cisternae. The ER takes two forms, rough (or granular), with ribosomes adhering to the outer surface, and smooth (with no ribosomes attached).
	GO:0005788	endoplasmic reticulum lumen	COL10A1	The volume enclosed by the membranes of the endoplasmic reticulum.
	GO:0031012	extracellular matrix	COL10A1,MMP11	A structure lying external to one or more cells, which provides structural support, biochemical or biomechanical cues for cells or tissues.
	GO:0032991	protein-containing complex	COL10A1	A stable assembly of two or more macromolecules, i.e. proteins, nucleic acids, carbohydrates or lipids, in which at least one component is a protein and the constituent parts function together.

Finally, the last biological enrichment step included in KnowSeq is the pathways enrichment. Pathways involving our DEGs are interesting to understand how the expression changes are affecting other genes and biological processes as well as how theses changes can turn into cancer (breast cancer, in this case). To achieve that, KnowSeq includes the function *DEGsPathwayVisualization*. This function makes use of KEGG database to acquire the pathways information. For the COL10A1, there is one reported pathway affected. For the MMP11 there exists no affected pathways in KEGG. Finally, for the VEGFD gene there are nine reported pathways. Figure 7.6 shows the pathway hsa04974 that is related with the colagen gene (COL10A1) and performs the Protein digestion and absorption process. In the figure, the collagen box shows a clearly difference between the tumour samples (red) and the normal samples (green). This means that the COL10A1 gene could activate erroneous processes inside the pathway depending on its expression. Table 7.7 shows the nine VEGFD related pathways as well as the pathway related with COL10A1 gene.

Table 7.7: Table that contains the retrieved pathways with their description for the final DEGs.

DEGs	ID	Description
COL10A1	hsa04974	Protein digestion and absorption
VEGFD	hsa04010	MAPK Signaling Pathway
	hsa04014	RAS Signaling Pathway
	hsa04015	RAP1 Signaling Pathway
	hsa04151	PI3K-AKT Signaling Pathway
	hsa04510	Focal Adhesion
	hsa04668	TNF Signaling Pathway
	hsa04926	Relaxing Signaling Pathway
	hsa04933	Age-Rage Signaling Pathway in diabetic complications
	hsa05200	Pathway in Cancer

The gene VEGFD is involved in several pathways (Pathway in cancer included). The changes in its expression could produce disorders in those pathways which could end up in the development of breast cancer and other diseases.

When all the enrichment pipeline of KnowSeq is over, this information is used to find and learn more about those DEGs and their relation with breast cancer. For that, it is very important that all these details will be studied and analysed by experts in bioinformatics and biology. KnowSeq expects to provide a very powerful and useful tool for those

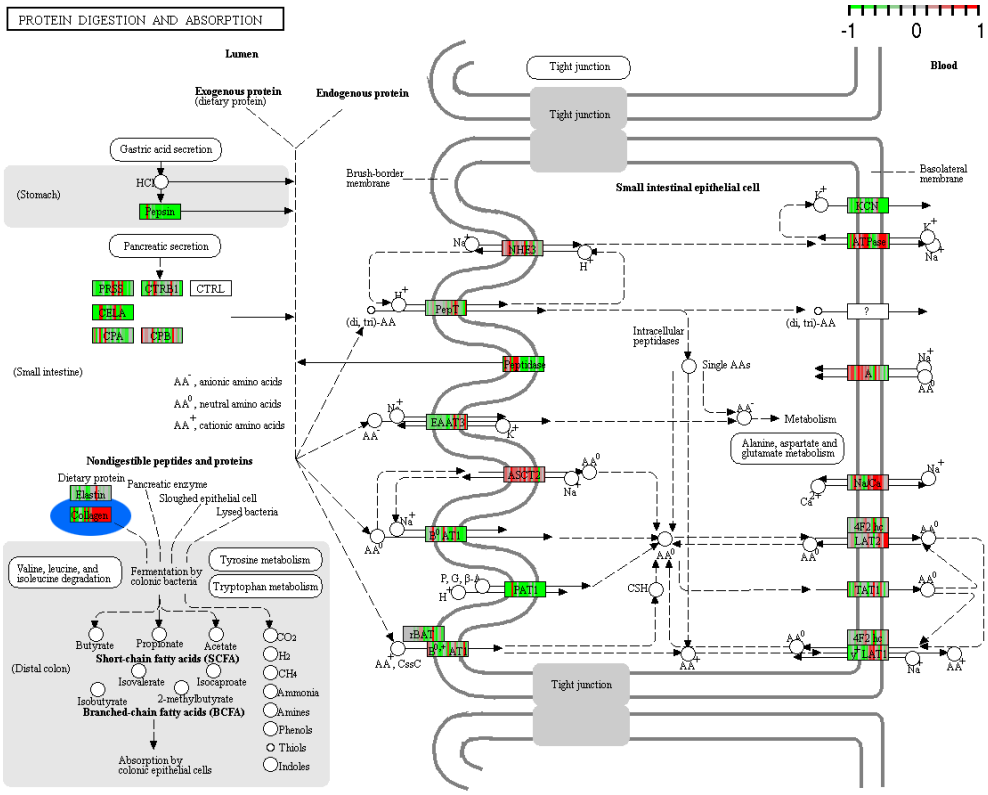

Figure 7.6: Pathway hsa04974 in which the COL10A1 gene is involved. As can be seen in the pathway, the collagen box indicates a strong expression change in the tumour samples in comparison to the normal samples.

experts that could retrieve the most crucial information for the DEGs based on their expression in an easy and adaptable way.

7.4 CONCLUSIONS OF THE CHAPTER

Along this Chapter, the specific objective followed in this doctoral thesis to design and encapsulate in a new tool the automatic and intelligent pipeline has been completed. KnowSeq includes the traditional steps in gene expression studies but also implements a FS and a ML step, as well as an enrichment step. Thanks to this, complete analyses can be done from RAW data up to the biological knowledge extraction in a easy, modular and flexible way.

Furthermore, in order to present a case of study with the tool, a breast cancer problem has been addressed with BAM files automatically downloaded with KnowSeq from GDC Portal. A total of 80 patients

with paired samples (Normal-Tumour) were used to extract the DEGs candidates. Those DEGs candidates were assessed through a ML step with the purpose of finding a very reduced sub-set of DEGs with the capability to discern between normal and tumour samples. Furthermore, different FS algorithms were applied in order to find a better order of those DEGs to achieve outstanding classification results rate with less DEGs. Finally, the DEGs were assessed by using 10 patients never seen before, achieving remarkable results since all the patients were totally recognised with only three DEGs for several combinations of classifiers and FS algorithms.

Then, a final sub-set of three DEGs were enriched by using the KnowSeq functions designed with this purpose. Those DEGs have a strong relation with breast cancer, there exist evidences at gene expression level, in the literature and in affected pathways that link the final three enriched DEGs with the disease.

At sight of these considerations and by way of conclusion of this Chapter, KnowSeq is an R package that gives the possibility to carry out RNA-Seq and Microarray analyses in an easy way with all the required steps included in the pipeline. KnowSeq expects to serve as a novel tool to help to the experts in the field to acquire robust knowledge and conclusions for the data and diseases to study. KnowSeq has three clear strengths: the first one is the modular design, because the analyses can be started from different points (FASTQ, BAM, count and even a custom expression matrix); the second one is the versatility due to the different algorithms for ML and FS and the different databases implemented in KnowSeq; and the last one is the adaptability of the analyses, because KnowSeq allows to use data from different sources and, even select different parameters that give to the user a real control of the pipeline.

8

CONCLUSIONS & FUTURE WORKS

CONTENTS

Along this doctoral thesis an existing problem dealing with the processing of heterogeneous transcriptomic data has been exposed, tackled and justified. With the aim of proposing a methodology to deal with such data in different cancers, the design of an intelligent automatic pipeline for the integration and analysis of heterogeneous transcriptomic data has been carried out. Furthermore, different cancer types have been studied and DEGs related to them have been extracted and assessed through ML techniques. Finally, a new tool, an R package named (KnowSeq) now publicly available at Bioconductor, was designed to bring the researchers a way of automatically performing those analysis in an easy and quick form. At sight of these consideration, this last chapter presents the final conclusions that support this thesis, taking into account the main objectives proposed in Chapter 1. In addition, the future work, which is intended to be addressed at soon as possible in order to continue with the high quality contributions to the scientist community of the research line proposed, is also detailed here

8.1 FINAL CONCLUSIONS

Considering the results and conclusions for the researches presented at Chapters 5, 6, 7 that support this thesis, the final conclusions for this doctoral thesis will be in-depth detailed, taking into account the specific objectives proposed at Chapter 1. For that, the conclusions will be sub-divided into one conclusion per objective along with a final thesis conclusion.

The first proposed specific objective was the design and implementation of an automatic pipeline for the integration of heterogeneous transcriptomic data sources regardless the genetic disease addressed. This idea emerged with the aim of taking advantages of the different heterogeneous data existing in public databases. There is a huge number of Microarray data that have not been analysed so far, containing in any case useful information about genetic diseases. Those Microarray data together with new RNA-Seq series may allow creating larger datasets than never before. As evidenced by the results of the presented researches, the integration has been successfully carried out for different types of cancer by combining Microarray and RNA-Seq samples coming from different sources. Furthermore, when the integration was done, related DEGs to the studied cancer were correctly retrieved. Thereby as a conclusion for the first proposed objective, the implementation of an automatic pipeline for integrating heterogeneous transcriptomic data sources has been satisfactorily completed, even for different types and sub-types of cancer (Breast cancer, Leukemia and Lung cancer -see Appendix A-), achieving a complete and general pipeline.

On the basis of the first specific objective, the second proposed specific objective was the assessment of the extracted DEGs through the use of ML techniques. In this sense, different FS and classification algorithms were proposed and evaluated for the different tackled types of cancers. In all of the cases in which classification algorithms were applied, outstanding results were achieved, even for unseen samples. In addition, the use of FS techniques has allowed to select a reduced sub-sets of DEGs that allow discerning among the different states, achieving practically the same results than with all of DEGs. This confirms the validity of the integrated pipeline to extract DEGs related to the addressed cancers, as well as the proper application and implementation of the different proposed predictive models. In view of these considerations, the second proposed objective has been also successfully carried out, ushering the possibility of finding new robust gene signatures for the addressed types of cancer.

What's more, it is important to note that many researches are very difficult to reproduce, because the code is not available or the explanation about the implemented methods is confusing. For that reason, the third specific objective of this thesis tries to bring to the scientist community a way of carrying out differential expression analysis together with ML assessment, encapsulated in an advanced software R package. KnowSeq is the first R/Bioc package that allows executing the necessary steps to perform analysis from transcriptomic RAW file to ML biomarkers assessment and biological enrichment under the same tool and language. Moreover, KnowSeq is not only available at Bioconductor, the most important bioinformatics repository, but also at Docker and Github with the purpose of reaching as many scientists as possible. To sum up, this third specific objective has been accomplished and KnowSeq has kept an important number of downloads in the last months and is climbing up in Bioconductor packages ranking.

Finally, the last proposed objective was the application of HPC approaches for the optimisation of the heaviest steps in this type of analysis. Nowadays, with the massive available transcriptomic data and the size per sample, it is becoming impossible to process and analyse them without HPC architectures or dedicated hardware. In this thesis, two different HPC clusters have been used for the parallelization of the RNA-Seq RAW data alignment, as the alignment is the most computational exhaustive process in the whole pipeline. Although, the use of GPUs could not be implemented and included in this doctoral thesis, partially due to the current time limitations of the PhD program, the implementation with CUDA of ML algorithms has already begun in the search for optimization of the intelligent biomarkers assessment process.

CUDA: Compute Unified Device Architecture

This doctoral thesis was proposed with the end of designing a new unified and automatic pipeline to analysis and integrate heterogeneous transcriptomic sources. For that, an early complete experimentation using Microarray and RNA-Seq series from Breast cancer was done, taking into account only two classes (Cancer vs Control). Then, performing a more advanced version of the pipeline, a new experiment that involved a larger amount of heterogeneous data and different types of Leukemia was addressed. In this experiment, a new parameter named "Coverage" for an efficient selection of multiclass biomarkers was also proposed. At the end of the Leukemia study, the pipeline was sufficiently mature and tested for its encapsulation in a public tool. At that point, the pipeline was expanded with new functionalities and KnowSeq was born as a new public R/Bioc package for the scientist community. Finally in Appendix A, KnowSeq was applied to a Lung Cancer multiclass study containing only Microarray data in order to

show the validity of KnowSeq for Microarray analysis too. Furthermore, a first version of a new Biologically-based FS method has been proposed. It is named DA-FS and tries to provide to the scientists a FS method with biological sense. This overview has showed how the main milestones accomplished in this thesis, successfully achieving all the established objectives.

8.2 LOOKING TO THE FUTURE

Bioinformatics is an emerging field with a very promising future ahead. In this sense, there is a countless number of new experiments and of important advances to discover in the fight against genetic diseases in general, and against cancer in particular. The contributions of this thesis to the trancriptomic heterogeneous data analysis at gene expression level, will help the development of new studies in a more automatic and straightforward way .

It is well-known however, that there are many factors that can alter gene expression, rising the possibility of producing cancer. For that, as a future work, the integration of new biological sources different from transcriptomic sources will be addressed. Besides, the creation of a set of synchronised predictive models, working together to predict and find hidden relationships among the integrated data will be performed too. In the stay at Institute of Bioinformatics WWU Muenster, a prototype for the integration of Copy Number Variation information and Gene Expression from the same patients was carried out. This prototype will be continued and finished as soon as possible. Furthermore, this new integration will be added as a new process in KnowSeq R/Bioc package.

Other important future works will be the continuous development and update of KnowSeq with the aim of improving and adding new functionalities to the software package. For example, optimised version of the ML classification techniques via CUDA are now under development and is intended to be added to KnowSeq during the next year. Furthermore, a new biological enrichment function to retrieve drugs related to DEGs is expected to be added soon.

To conclude, the integration of slides tissue images and clinical information with the gene expression information is now under development. This integration has the aim of merging as many biological and histopathological information sources as possible to improve the diagnosis, taking advantages of ML techniques.

Part III

APPENDICES & BIBLIOGRAPHY

A

IMPACT OF FEATURE SELECTION FOR BIOMARKER DETECTION IN MULTICLASS LUNG CANCER

CONTENTS

The present study is a reorganised and extended version of the international conference published manuscript "Feature Selection and Assessment of Lung Cancer Sub-types by Applying Predictive Models" [219]. This study was designed to evaluate the impact of FS techniques in DEGs selection and assessment. This extension introduces several changes with respect the original published paper, and it is intended to be sent to a Journal with Impact factor when RNA-Seq data is also added to the study. The whole study was repeated using KnowSeq R/Bioc package and the selection of DEGs is different from the original paper. The most important improvement in this presented work is the introduction of a new Biologically-based FS method (DA-FS), which allows including biological information of the DEGs for creating a new feature ranking.

A.1 BACKGROUND

This first Appendix emerged with the aim of showing a more complex study designed and implemented with KnowSeq, drawn from a previous published manuscript [219]. Furthermore, all the FS and classifiers implemented in KnowSeq are evaluated with the purpose of measuring their impact in the discernment capability of the candidate DEGs.

Along this study, a huge number of Microarray samples have been used to identify a robust set of DEGs, having the capability of discerning among the different sub-types of lung cancer: SCLC, ACC, SCC and LCLC. To achieve this goal, an overall DEGs analysis was performed by using data from gene expression microarrays publicly stored at NCBI/GEO platform and the usage of the KnowSeq R/Bioc package. Furthermore, a novel Biological-Based FS method named as Diseases Association feature selection (DA-FS) is proposed and included into KnowSeq.

SCLC: Small Cell Lung Cancer
ACC: Adenocarcinoma
SCC: Squamous Cell Cancer
LCLC: Large Cell Lung Carcinoma

As a reminder, a gene signature is a single or combined group of genes in a cell with a uniquely characteristic pattern of gene expression that occurs as a result of an altered or unaltered biological process or pathogenic medical condition [241]. Discovering these gene signatures can lead to an early diagnose and to understand the root cause for developing a multifactorial disease such as cancer. Henceforth, by their usage, a discrimination between a patient suffering from cancer and a healthy one can be performed. This discrimination can not only be performed within healthy or cancer patients but also between different states or sub-types of the same cancer disease.

Lung cancer is the most common cancer and the main cause of death by cancer in men, followed by prostate cancer and colorectal. In women, lung cancer has the second and third position in mortality and incidence, respectively [242]. There are two main types of lung cancer:

- SCLC: In SCLC, the cells contain dense neurosecretory granules (vesicles containing neuroendocrine hormones), which give this tumor an endocrine or paraneoplastic syndrome association [243]. Most cases arise in the larger airways (primary and secondary bronchi).
- NSCLC: NSCLC has three differentiated sub-types, namely:
  - ACC: The signs and symptoms of this specific type of lung cancer are similar to other forms of lung cancer, and patients

NSCLC: Non Small Cell Lung Carcinoma

most commonly complain of persistent cough and shortness of breath. Adenocarcinoma is more common in patients with a history of cigarette smoking, and is the most common form of lung cancer in younger women and Asian populations. The pathophysiology of adenocarcinoma is complicated, but generally follows a histologic progression from cells found in healthy lungs to distinctly dysmorphic, or irregular cells [244].

- SCC: It is the second most prevalent type of lung cancer after lung adenocarcinoma and it originates in the bronchi. Its tumor cells are characterized by a squamous appearance, similar to the one observed in epidermal cells. Squamous-cell carcinoma of the lung is strongly associated with tobacco smoking, more than any other form of NSCLC [244].

- LCLC: LCLC is a heterogeneous group of undifferentiated malignant neoplasms that lack the cytologic and architectural features of small cell carcinoma and glandular or squamous differentiation [245].

The identification of genetic biomarkers associated with lung cancer allows the early prognostic and the right treatment. This is critical nowadays, as this could be the difference between the recovery of the patient or his decease. For that, it is crucial to know what genes could be promoting disorders in one or more biological process that finally cause, in this case, any of the different sub-types of lung cancer.

For several decades, Microarray technology has allowed studying the alteration at gene expression level with the purpose of finding genes involved in pathologies of genetic source. This technology is highly widespread and known and is based on the capability of the complementary molecules to hybridise among themselves to determine the gene expression values of each studied gene in the analysed samples [85]. Through this process, the over-expressed or inhibited genes can be identified in tumor samples when comparing to normal samples.

Previous studies performed by Sanchez-Palencia A. et al. have used this technique in the identification of biomarkers for the sub-types ACC and SCC, for a reduced number of patients [246]. Others like Yanaihara N. et al. have studied the molecular profiles of lung cancer by using microRNA data [247].

This work is aimed to identify a reduced set of genes that has the ability to discern among the five contemplated states (ACC, SCC, LCLC, SCLC

and Control). This set was later used to design and compare a number of predictive models. These models can perform the prediction of the state of samples not seen before in the learning process, with a great reliability.

A.2 DATA GATHERING

All lung cancer samples have been obtained from the NCBI/GEO public repository [248]. Specifically, 13 series has been used which are publicly available in this repository. These series have samples from the different lung cancer sub-types described above as well as healthy ones. A total of 851 samples from the different series constitute the final dataset. In Table A.1, information from each of the used series is presented. The NCBI/GEO ID Microarray platform is presented in order to identify the series used in this work. The index of the exclude outliers and the number of samples from the different lung cancer types for each series are shown. The total of samples within each class in the final dataset is also indicated.

The total amount of lung cancer samples have been obtained from the public repository . Concretely, 13 series stored and publicly available in this repository have been used. These series have samples from the different sub-types of lung cancer addressed in the study, as well as healthy samples. A total of 851 samples from the different series conform the final dataset. Table A.1 shows certain information about each of the 13 series. For each series, the information about the NCBI/-GEO ID Microarray platform, the excluded outliers and the number of samples from the different lung cancer types is shown. Finally, the table includes the total sum of samples that conform each lung cancer and control states.

A.3 METHODOLOGY

The methodology followed to carry out the study can be sub-divided and presented in two subsections. The first one is focused on the DEGs extraction that will help discern later among the addressed pathological states (lung cancer sub-types and control). The second one comprehends the application and explanation of the Computer Intelligence-based predictive models proposed for this research. Figure A.1 shows the whole pipeline. All the study has been carried out

Table A.1: Table with the information about the 13 series used in this study. For each series, the information about the GEO ID, the platform used, the removed outliers and the number of samples from different subtypes that each series has, is shown.

Serie	Platform	Index Outliers	SCLC	LCLC	SCC	ACC	Control	Total
GSE7670	Affymetrix HG-U133A	61,46,34	-	2	0	29	2	33
GSE99316	Affymetrix HG-U133B	36	23	-	-	-	-	23
GSE43580	Affymetrix HG-U133_Plus_2	24,30,43,46,58,59, 82,87,124,128,133	-	-	69	70	-	139
GSE73160	Affymetrix GPL11028	1,10,21,32,42,63	62	-	-	-	-	62
GSE3268	Affymetrix HG-U133A	–	-	-	10	-	-	10
GSE40275	Human Exon 1.0 ST Array GPL15974	10,18	15	3	5	14	41	78
GSE18842	Affymetrix HG-U133_Plus_2	60,63,74,84,85	-	-	-	-	42	42
GSE37745	Affymetrix HG-U133_Plus_2	39,52,55,57,58,71, 82,79,95,155	-	21	64	-	-	85
GSE41271	Illumina HumanWG-6 v3.0	83,84,88,138,145, 185,201,202,203, 223,253	-	3	74	180	-	257
GSE12771_1	Illumina GPL6097	5,8,13	-	-	-	-	24	24
GSE12771_2	Illumina GPL6102	112,152,237,238, 239,240,241,242	-	-	-	-	40	40
GSE39345	Illumina GPL6104	21,23,24,25,39	-	-	-	-	20	20
GSE21933	Phalanx Human GPL6254	2,4,22,30	-	-	9	10	19	38
		Total	100	29	231	303	188	**851**

by using KnowSeq R/Bioc package. A new tool to address complex RNA-seq analyses that its also prepared to tackle Microarray data [8].

A.3.1 *Pre-processing and DEGs Extraction*

The aim of this study is to gather/put together the information from a large number of Microarray samples with the purpose of having statistical significance in the lung cancer DEGs detection and assessment. To this end, different Microarray series are needed, becoming the pre-processing step very sensitive and crucial. For that, each series must be treated with the utmost care in order to ensure the right harmony among them. Taking this into account, a series integration without losing biological information can be achieved.

This pre-processing step, as well as the posterior integration and DEGs extraction, correspond to the first three steps in Figure A.1. The pre-processing step comprises a set of sub-steps that are explained herein.

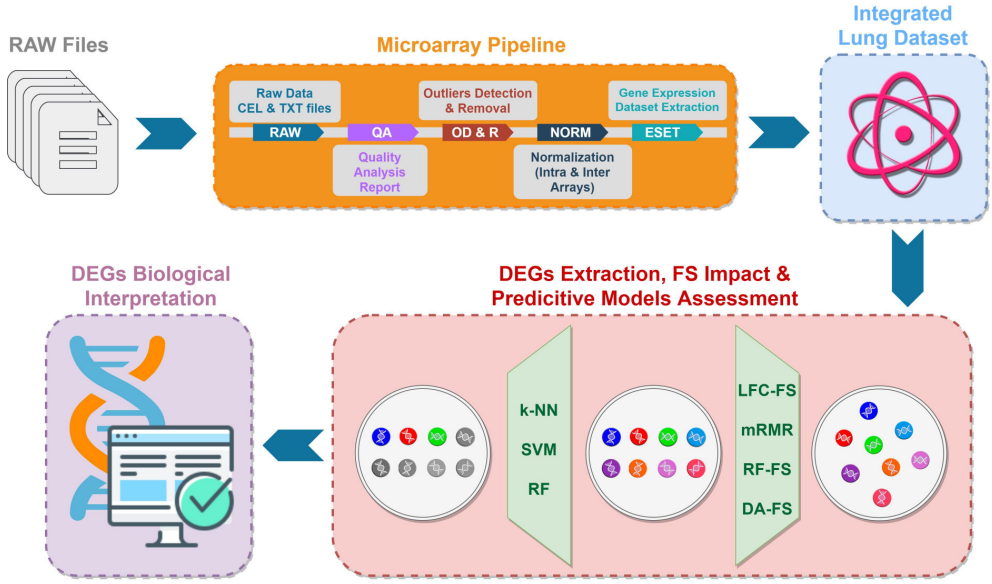

Figure A.1: Pipeline designed for this study, to first analyse and integrate the addressed series and to finally evaluate the extracted DEGs by using predictive models.

Firstly, it is required to verify the existence of outliers in the series in order to remove those samples that could distort the final results. Then, it is very important to apply the same logarithmic transformation to reach the right cohesion of the data. Furthermore, a correction of the bit depth of the data is performed in order to equalise the series. The last consideration for an appropriate pre-processing is the batch effect study and correction. The batch effect is a deviation effect in the gene expression values due to several external technical factors (origin, sequencing hour, lab technician, among others) and it is very hard to treat [159]. For the batch effect treatment in this study, the SVA algorithm was applied in order to remove batch effect when the batches are unknown [226]. Once the pre-processing step is finished, all the series are merged into one integrated series.

Once the integrated series is obtained, DEGs analysis step can be carried out. It will determine which genes are differentially expressed in lung cancer samples in contrast with control samples. For DEGs extraction, *KnowSeq* makes use of limma package to statistically compare the expression of the selected samples to detect DEGs among the compared classes. This process is performed by leaving the 20% of the data out in order to have a test dataset to verify the suitability of the DEGs with data no seen before. DEGs extraction for this multi-class problem was carried out by performing different bi-class DEGs extraction considering the following 5 comparisons:

- SCLC vs. Rest
- ACC vs. Rest
- SCC vs. Rest
- LCLC vs. Rest
- Control vs. Rest

For each of the comparisons, DEGs extraction is carried out fixing the restriction values in limma package. The first restrictive parameter that allows to know if a gen is or not expressed, is the P-value (the value of statistic significance according to the statistical test t-Student). Moreover, the LFC is used as restrictive parameter (existing differences among the mean expression of the analysed conditions) in conjunction with the P-value to decide the DEGs candidates. Applying a P-value lower or equal than 0.05 and a LFC greater or equal than $,log2(1.1)$, an amount of DEGs between 50 and 60 for each of one the comparisons proposed before were extracted.

Finally, with the aim of identifying the common DEGs among the five comparisons, an intersection was carried out and 37 common DEGs were detected. Those DEGs are selected no matter the combination chosen, so in theory they have the capability to discern among all the addressed lung cancer sub-types. However, the 37 final DEGs will be assessed afterwards to corroborate their discerning potential.

A.3.2 *Predictive Models development & Assessment*

The genes identified as relevant in the previous comparisons, determine the features in the dataset used in the development of the predictive models. Due to the randomly generated global splitting implemented in the preprocessing step, there is a training dataset, formed by the 80% of the total amount of samples and a test dataset with the 20% of the remaining samples.

A.3.2.1 *Feature selection*

Feature selection techniques are widely applied in the machine learning scope to reduce the curse of dimensionality, specially in those

presenting a large number of features in comparison with the number of samples [249]. Concretely, after applying the two previous steps, feature selection can be used to identify a more reduced gene signature. Also, a removal of redundant information and a preservation of the one that allows discerning among the classes is achieved in this new reduced gene signature. In the feature selection process carried out, two stages can be clearly distinguished:

- Feature selection: In general these algorithms measure the relationship among all the features, generating a ranking of these in function of their relevance with the objective class. In this study, four methods to select features are implemented to achieve four rankings. The first method simply sorts the final DEGs by their LFC, placing the DEGs with the greater LFC in the first positions (LFC-FS). Then, two widely known feature selection algorithms in the literature were used. These algorithms are mRMR, and RF as FS (RF-FS) [139, 230]. Furthermore, the study has a novel biological feature selection method explained in detail in the next subsection.

- Wrapper selection: SVM, k-NN and RF classifiers are used in an incremental manner, in the search for the optimal set of genes [126, 231–233]. This technique consists in creating as many models as number of features in the dataset, following the ranking given by the feature selection algorithm, in order to identify the subset bringing the best performance.

The identification of the optimal number of genes is thus supported by the performance of the classifiers in the training dataset, trying always to achieve a reduced number of genes without losing relevant information. If the feature selection is carried out right, the DEGs would be optimal for the approached problem and, they are expected to be related to a greater or lesser extent with the pathology.

A.3.2.2 *Disease Association Feature Selection*

For this study, a novel feature selection method was designed. This method makes use of targetValidation webplatform to acquire an association score for each DEGs with the required disease, lung cancer in our case [240]. This score takes values between 1 and 0, meaning 1 a total association and 0 no association. Therefore, the DEGs are sorted by this score, achieving a ranking in which the first places are occupied by those DEGs with more biological relation to lung cancer.

This novel feature selection method is named as Disease Association feature selection (DA-FS), and it is already included inside *KnowSeq* R/Bioc package.

Benefits can be obtained from using this biological based feature selection algorithm upon those based on heuristics. For clinicians, it would be easier to understand a ranking based on the evidence in literature of the relation between those genes and the studied cancer than those based on information theory.

A.3.2.3 *Predictive Models Validation*

To perform the assessment of the DEGs, 12 alternative approaches were compared. Each approach was formed by the combination of one of the three classifiers with one of the four feature selection algorithms, performing all the possible combinations. The test dataset was used for the predictive model testing. For each combination the accuracy, the f1-score, the sensitivity and the specificity were measured. These metrics were chosen based on their importance when evaluating the performance of a model in a multiclass classification problem [250].

Moreover, through the use of the web platform *targetValidation*, which provides information about the biological background of the DEGs, the relationship between those DEGs and the pathology addressed can be inspected [251]. As a reminder, the whole methodology has been completed with *KnowSeq* R/Bioc package, including the DEGs biological relations with lung cancer.

A.4 RESULTS & DISCUSSION

As previously stated, different combinations of FS and predictive models algorithms were taken into account and analysed by using the *KnowSeq* ML module. Firstly, results using a 10-fold CV was implemented in order to analyse the behaviour of the DEGs for the training dataset and the different combinations. Although all the executions revealed outstanding results, the combination of k-NN with mRMR FS was the best one. In this context, Figure A.2 shows the results obtained by 10-folds CV k-NN and mRMR on the training dataset. For every fold, each FS algorithm was used in order to determine the best number of DEGs for classification. With only 5 genes selected by mRMR, the classifier reached almost 90% of accuracy in several folds.

Given this behaviour in the training dataset, 5 genes were selected as candidate size for create the models and testing all of them in the search of a possible multiclass gene signature.

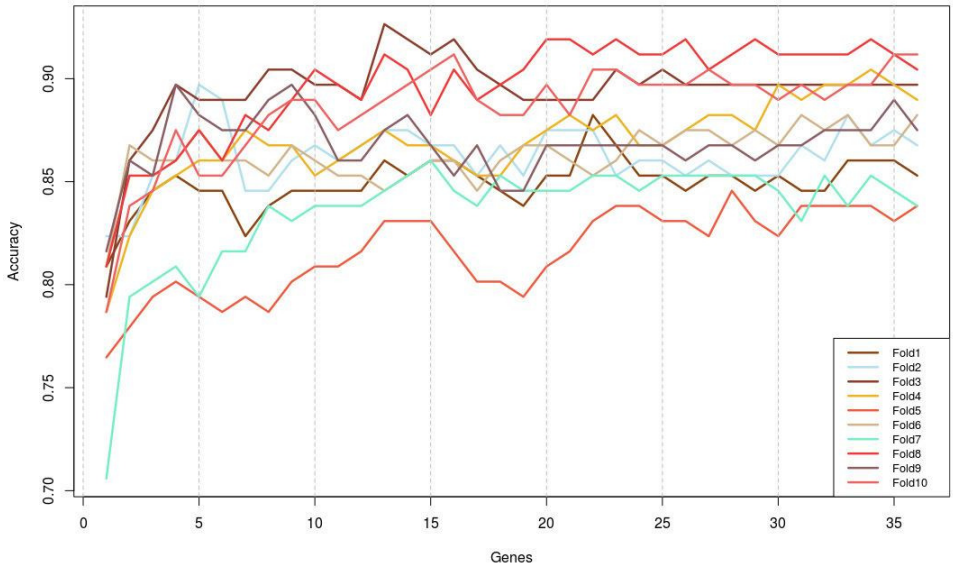

Figure A.2: 10 folds CV k-NN results by using mRMR and the final 37 DEGs candidates. The results show the potential of those DEGs as DEGs with a strong discerning capability for the addressed lung cancer types.

Once the validation was carried out, the test dataset was assessed in order to corroborate if the models maintain the same trend than in CV, facing an unused data in the process. For that purpose, Figure A.3 exposes a classification plot that contains the results of k-NN in conjunction with each of the feature selection methods in the study. In the plot, mRMR clearly achieves better results with 5 or less genes than the rest of the FS methods. Thus confirm the validity of the mRMR process previously performed. Furthermore, the DA-FS attains significant better results than LFC-FS and RF-FS with 5 genes. This is, therefore, very interesting because DA-FS reorder the DEGs in function of their direct correlation with the disease. Consequently, the top 5 DEGs for DA-FS are the most related DEGs in the list with lung cancer. This novel feature selection methods achieves lower precision than mRMR but, DA-FS uses biological information instead heuristic approaches, which can bring to the clinicians and experts a very powerful and comprehensible feature selection method. LFC-FS and RF-FS obtain worse results with 5 DEGs.

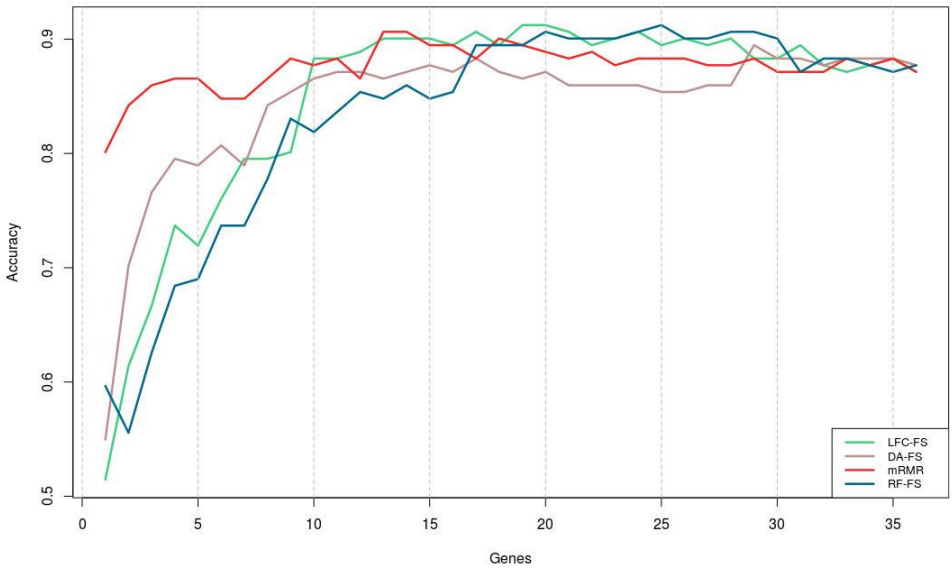

Figure A.3: k-NN test results achieved by using all the DEGs and the four different feature selection algorithms. The figure shows the mRMR gains with a lower number of genes in comparison with the other three algorithms.

With the purpose of comparing k-NN results with both SVM and RF results, Table A.2 collects the numerical outcome for all the combinations addressed in the study selecting always the top 5 DEGs. This table not only shows the accuracy but also the sensitivity, the specificity and the f1-score. In view of these results, a two behaviours can be clearly distinguished. Firstly, the best results for all the measures are always reached in those combination with mRMR whichever been the classifier implemented. As with the accuracy achieved by k-NN and DA-FS, this feature selector slightly gets worse results than mRMR in all the combinations and measures. Although DA-FS slightly losses information in classification in comparison with mRMR, on the contrary gains biological relevance with regards to the top selected DEGs. Secondly, while the combination of k-NN and mRMR is the best one, in general k-NN achieves worse results than those obtained with RF. However, it still outperform SVM results in general. To sum up, the best combination (k-NN with mRMR) achieves an accuracy equal to 86.5%, a sensitivity equal to 79.3%, a specificity equal to 96.1% and a f1-score equal to 80.3%.

At this point, the test dataset was assessed and the classification metrics calculated. Metrics show a great discerning capability of the selected DEGs by mRMR and DA feature selector. However, it is interesting

Table A.2: Multi-class test classification results for each combination of the feature selection algorithms with the classifiers, by using the top 5 DEGs. The table shows the accuracy, the mean sensitivity, the mean specificity and f1-score.

	Accuracy			Sensitivity			Specificity			F1-score		
	k-NN	SVM	RF	k-NN	SVM	RF	k-NN	SVM	RF	k-NN	SVM	RF
LFC-FS	73.6%	70.1%	74.2%	64.3%	60.8%	66.9%	92.5%	91.6%	92.7%	62.0%	58.2%	67.5%
DA-FS	79.2%	77.7%	76.6%	66.7%	67.0%	71.8%	94.0%	93.5%	93.3%	66.7%	65.9%	65.5%
mRMR	86.5%	84.7%	84.2%	79.3%	72.3%	77.1%	96.1%	95.5%	95.5%	80.3%	70.8%	78.1%
RF-FS	68.4%	71.3%	72.5%	59.5%	60.5%	62.6%	90.8%	91.7%	92.1%	58.2%	59.2%	61.6%

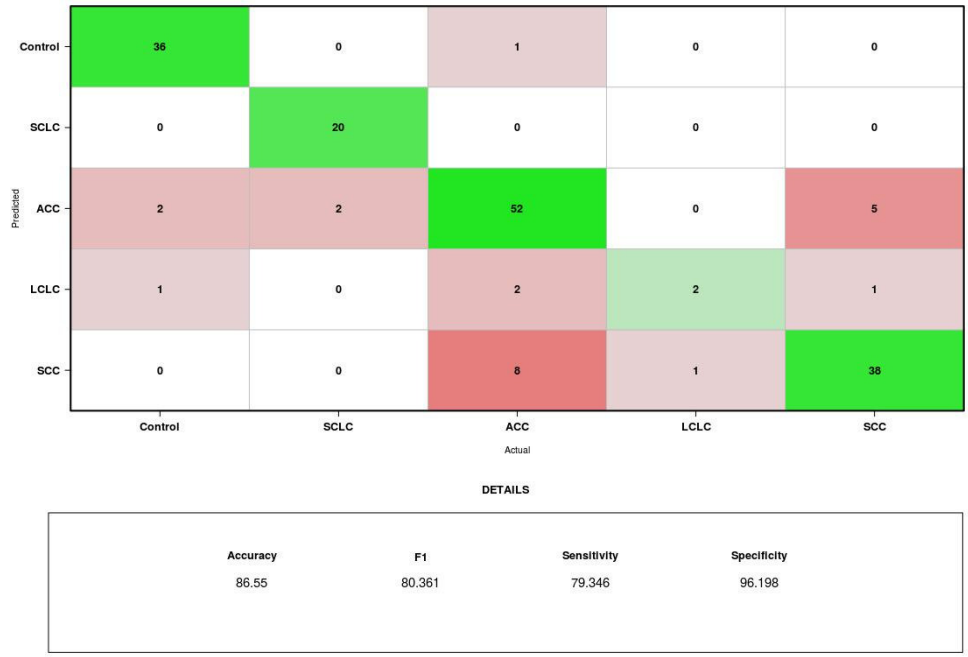

Figure A.4: 5-class Confusion matrix that shows the test results attained by k-NN with the top 5 DEGs of mRMR. Moreover, the accuracy, sensitivity, specificity and f1-score are listed. LCLC, ACC and SCC are confounded among them.

to observe between which classes were more difficult to discern and those that were easily identifiable. Figure A.4 shows the confusion matrix for the five classes for the samples in the test dataset. In the Figure, LCLC, ACC and SCC are not easily discernible. The origin of this lack of discernment could the fact that the three classes belong to the super-class NSCLC and, the differences at gene level among these classes seem not to be enough to attain a complete separation and recognition of them. The rest of the classes are well classified achieving outstanding results for both the accuracy and the specificity (86.55% and 96.19% respectively). The f1-score and the sensitivity are lower than the other two metrics due to the bad classification results for the

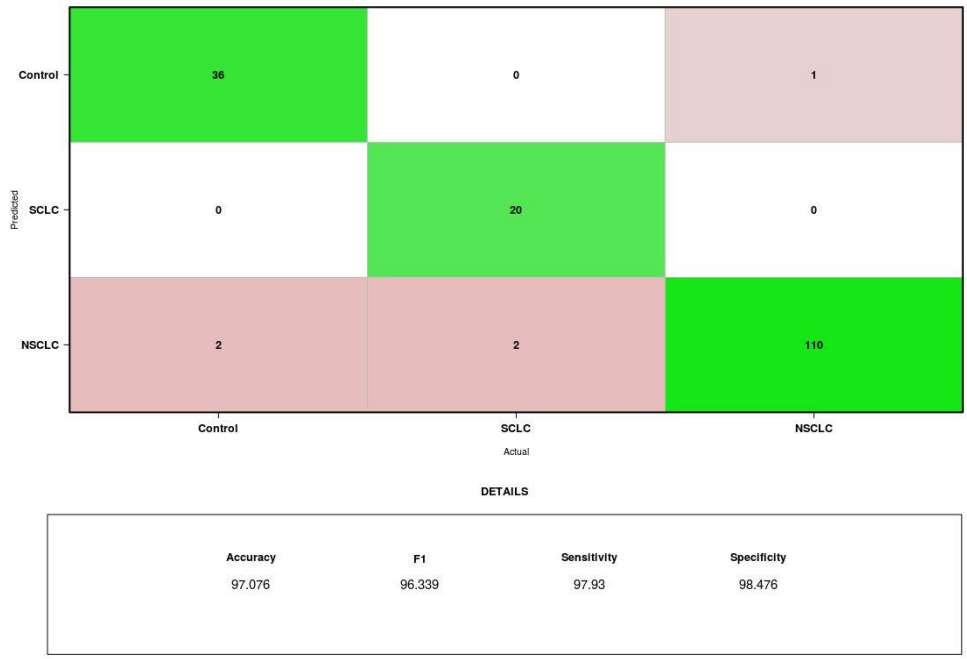

Figure A.5: 3-class Confusion matrix that shows the test results attained by k-NN with the top 5 DEGs of mRMR. This matrix joins LCLC, ACC and SCC in the super-class NSCLS. The accuracy, sensitivity, specificity and f1-score significantly improve due to this classes fusion.

confounded classes, achieving a 80.36% the f1-score and a 79.34% the sensitivity.

Relying on the results of the previous figure, the LCLC, ACC and SCC classes were joined in the NSCLC super-type in order to corroborate if classification metrics are increased. Figure A.5 shows the confusion matrix for 3 classes classification (Control, NSCLC and SCLC) by using k-NN with the top 5 DEGs from mRMR. As was mentioned above, LCLC, ACC and SCC were confounded among them. When these classes were joined and the predictive model applied again, the stats significantly raised, failing only in 4 samples. With three classes, the accuracy is equal to 97.07%, the f1-score achieves 96.33%, the sensitivity obtains 97.93% and the specificity attains a total of 98.47% of recognition rate.

Finally, as a last simulation, all the sub-classes that belong to Lung Cancer (ACC, SCC, LCLC, SCLC) were joined under the same class (LC) with the aim of corroborating the behaviour of those DEGs to discern among tumour or control. Figure A.6 shows how these DEGs improve the results acquired classifying with 5 and 3 classes. An accuracy equal

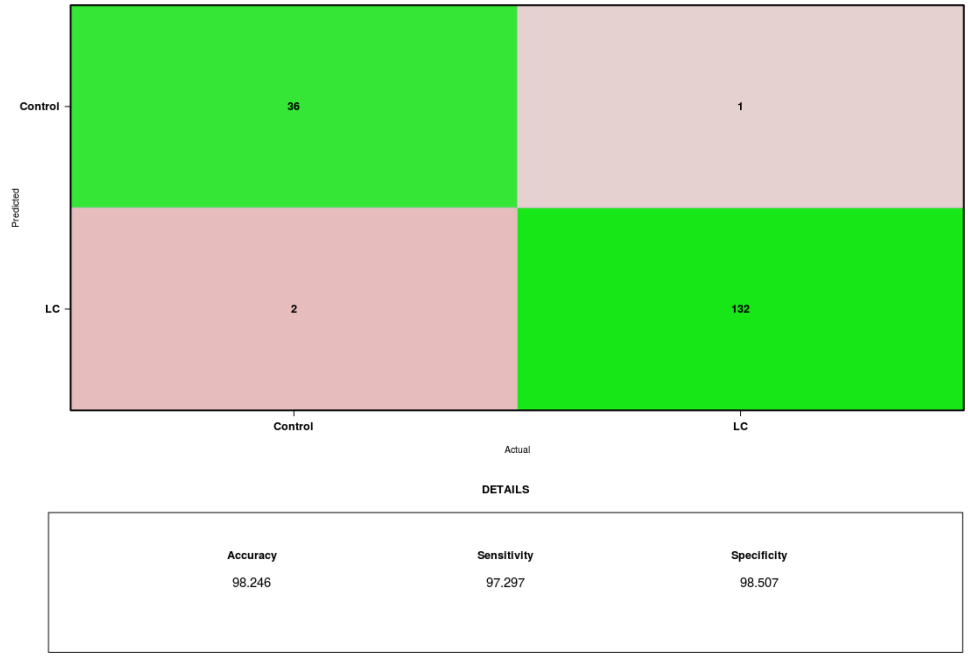

Figure A.6: 2-class Confusion matrix that shows the test results attained by k-NN with the top 5 DEGs of mRMR. This matrix joins SCLC, LCLC, ACC and SCC in the super-class LS. The classifier only fails in 3 predicitions.

to 98.24% is obtained together with a sensitivity equal to 97.29% and a specificity equal to 98.50%. It is to be highlighted that only three samples were confounded.

At sight of these results, there is clear it exists relationship of those genes with lung cancer to the point of discerning at 5, 3 and 2 levels the different addressed states with a great precision.

Moreover, with the aim of seeing what classes each DEGs can discern, Figure A.7 contains 5 genes boxplots, which 5 boxplot (One per class) inside each of them. These boxplot shows how each gene has the capability to discern among different but not all the lung cancer sub-types. It is seems to be clear that thanks to the 5 DEGs, the samples are correctly classified because each gene contributes to discern concrete sub-types or classes.

Once the predictive models and DEGs assessment process has been completed, it is of utmost importance to corroborate if those DEGs are truly related with lung cancer. For that, Table A.3 shows the brief biological information for the top 5 DEGs from mRMR and DA fea-

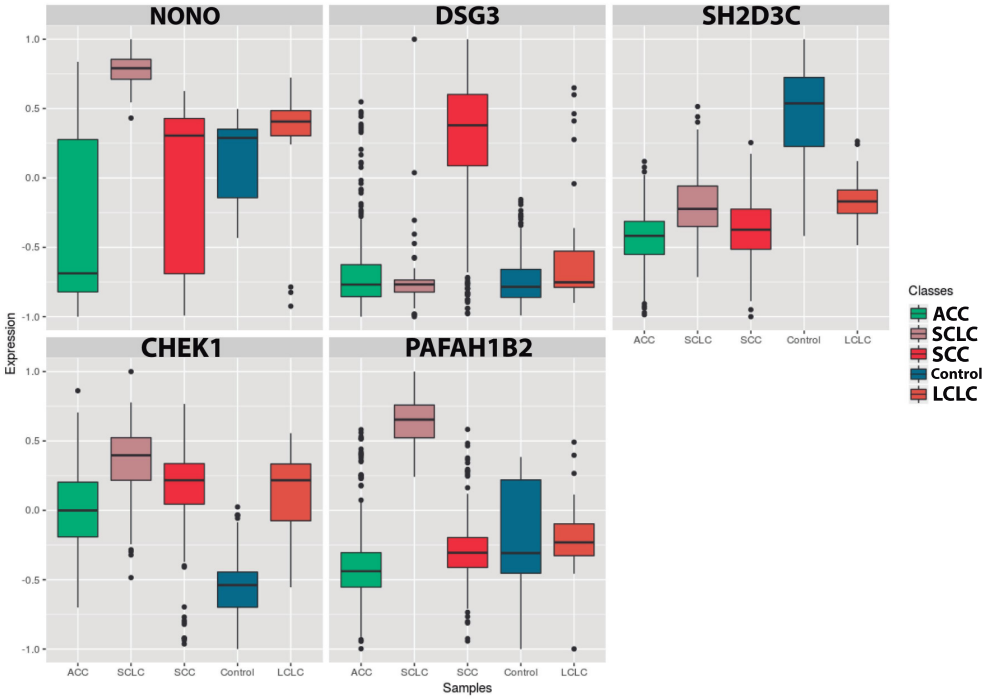

Figure A.7: 5 first selected differentially expressed genes by mRMR algorithm (order from left to right and from top to bottom: NONO, DSG3, SH2D3C, CHEK1, PAFAH1B2), with the expression levels for each sub-type of lung cancer and for control.

ture selector. Although, according to the results previously presented, mRMR attains the best results, DA-FS achieves great performance, including in its operation biological information with the related disease. Henceforth, an overview about the importance of those DEGs in lung cancer from both selection methods is of great interest. As would be expected, the association score in the table is higher in the top 5 DEGs selected by DA-FS than in those selected by mRMR. However, both have the CHEK1 gene in common. According to the literature, the top 5 DEGs selected by DA-FS are strongly related with lung cancer and its development. Concretely, they are associated with NSCLC and SCLC. Conversely, DEGs selected by mRMR do not have a clear or direct relation with lung cancer, excepting CHEK1 gene. In the light of the results of this section, the ranking proposed by DA-FS achieves results very close to mRMR, being the DEGs in the first positions are robustly related with the disease (Lung cancer in our case).

Table A.3: mRMR and DA-FS top 5 DEGs related with lung cancer. For each DEGs, its name, brief description and targetValidation Association Score are showed.

F.S. Method	Gene	Description	Asso. Score
MRMR	NONO	No exist evidences that probe the relationship of this gene with lung cancer. However, the protein codified by this gene interacts with a robust biomarker of NSCLC [252].	0.3213
	DSG3	Exists studies that correlate SCC with a differential expression of this gene in both miRNA and RNA [253, 254].	0.3705
	SH2D3C	There are evidences that this gene is a gene regulator of the transcription factor ELF5, that is related with lung cancer [255].	0.0450
	CHEK1	High expression of CHEK1 in lung tumors was associated with poor overall survival [256].	0.8983
	PAFAH1B2	There are no evidences that related this gene with lung cancer. However, recent profiling studies have revealed that these enzyme may be dysregulated broadly across many types of cancers [257].	0.3768
DA-FS	CA12	CA12 is an important clinical prognostic Serum Tumor Biomarker. Studies correlate that gene with NSCLC and its sensitivity to chemotherapy [258, 259].	1
	PPIA	This gene is related with lung cancer. In a study about seven endogenous control genes in NCSLC, this gene was the gene with higher expression in tumours [260].	0.9311
	CHEK1	High expression of CHEK1 in lung tumors was associated with poor overall survival [256].	0.8983
	HTRA2	This gene might promote the apoptosis of NSCLC cells, and serve as a target for NSCLC's treatment [261].	0.7337
	FLI1	The gene is closely related to lung cancer. FLI1 promoted tumorigenesis of small cell lung cancer cells. The gene may serve as target for therapeutic intervention of SCLC [262].	0.6762

A.5 CONCLUSIONS OF THE STUDY

Along this Appendix, a new study in which four different feature selection methods and three classifiers, for DEGs extraction and assessment respectively, have been addressed and compared. All the possible combinations between the feature selection methods (LFC-FS, mRMR, RF-FS and DA-FS) with the classification algorithms (k-NN, SVM and RF) were assessed. Also, an integration of heterogeneous microarray datasets belonging to 5 different lung sub-type tissues was performed, leading to the identification of 37 multiclass DEGs.

Then, when using a 10-fold cross validation, k-NN with mRMR in validation achieved the best results, using only 5 genes and nearly a 90% of accuracy. The k-NN with mRMR test results support the selection of these 5 DEGs by reaching similar ones to those obtained in training assessment but on unseen data. Our novel Disease Association feature selection method (DA-FS) reaches very similar results than mRMR but with the novelty of using DEGs with a strong biological relation with lung cancer, in this case. DA-FS ranking can provide clinicians and experts with a way of selecting DEGs based on their biological relevance instead of an heuristic. The top 5 DEGs selected by DA-FS keep strong relation with lung cancer and its development based on the results presented in literature.

To sum up, this study addresses a multiclass lung cancer problem from different machine learning methodologies. Furthermore, a novel biological-based feature selection method is proposed. Finally, it is important to highlight than all the study has been performed by using *KnowSeq* R/Bioc package and, DA-FS is already included in the tool.

B

KNOWSEQ USER DOCUMENTATION

CONTENTS

The KnowSeq user guide is presented in this Appendix together with example code to carry out a complete experiment with two public series from NCBI/GEO. This documentation is also available at Bioconductor.

B.1 INSTALLATION

To install and load KnowSeq package in R, it is necessary the previous installation of BiocManager from Bioconductor. The next code shows how KnowSeq installation can be performed:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("KnowSeq")
library(KnowSeq)
```

Furthermore, KnowSeq is also available through Docker, removing the necessity of dependencies installation just by running the next command in a terminal:

```
Docker run -it casedugr/knowseq
```

B.2 INTRODUCTION

KnowSeq proposes a whole pipeline that comprises the most relevant steps in the RNA-Seq gene expression analyses, with the main goal of extracting biological knowledge from raw data (DEGs, GOs enrichment, pathway visualisation and related diseases). In this sense, KnowSeq allows aligning raw data from the original fastq or sra files, by using the most renowned aligners such as tophat2, hisat2, salmon and kallisto. Nowadays, there is no package that only from the information of the samples to align -included in a text file-, performs automatically the download and alignment of all of the samples. Furthermore, the package includes functions to: calculate the gene expression values; remove batch effect; DEGs extraction; plot different graphs; and perform the DEGs biological enrichment with the GOs information, pathways visualisation and related diseases information retrieval. Moreover, KnowSeq is the only package that allows applying both a ML and DEGs biological enrichment processes just after the DEGs extraction. To achieve these objectives, there are functions that allows performing a FS process as well as a ML process using well-known supervised classifiers algorithms as k-NN, RF or SVM. Similarly, there are functions allowing the retrieval of biological knowledge of the DEGs candidates. This idea

emerged with the aim of proposing a complete tool to the research community containing all the necessary steps to carry out complete studies in a simple and fast way. To achieve this goal, the package uses the most relevant and widespread tools in the scientific community for the aforementioned tasks. The current version of the aligner functions works under Unix, but further version will be extended to MAC_OS and to Windows (if the tools were available).

The whole pipeline included in KnowSeq has been designed carefully with the purpose of achieving a great quality and robustness for each of the steps that conform the pipeline. Thus, the pipeline has four fundamental processes:

- Automatic data gathering
- RNA-Seq RAW data processing
- Biomarkers identification & assessment
- DEGs enrichment methodology

B.3 AUTOMATIC DATA GATHERING

The first step in the pipeline is the automatic data gathering process. KnowSeq allows automatically downloading series and datasets from three databases (NCBI/GEO, ArrayExpress and GDC Portal). To carry out the automatic downloads, the function downloadPublicSeries has to be executed with the Series ID for NCBI/GEO or Arrayexpress or with a samples manifest for GDC Portal. It is important to know the format of the downloaded CSV file, which will be used in next step for the RAW alignment. Each of the repositories has its own format in the CSV file that contains the information to download and process the desired samples. The format for each repository is explained herein.

B.3.0.1 *NCBI/GEO CSV format*

Series belonging to RNA-Seq have a SRA identifier. By calling the function downloadPublicSeries with the NCBI/GEO GSE ID of the wanted series to automatically, the CSV with the required information about the serie will be downloaded.

This CSV file contains a number of columns with information about the samples. However, in order to running the alignment step only the three columns shown at Table B.1 are necessaries (although the rest of the columns can be kept).

Table B.1: NCBI/GEO CSV Format

Run	download_path	LibraryLayout
SRR2753177	sra-download.ncbi.nlm.nih.gov/traces/sra21/SRR/0026...	SINGLE
SRR2753178	sra-download.ncbi.nlm.nih.gov/traces/sra21/SRR/0026...	SINGLE
SRR2753179	sra-download.ncbi.nlm.nih.gov/traces/sra21/SRR/0026...	SINGLE

B.3.0.2 *ArrayExpress CSV format*

The process for ArrayExpress is very similar to NCBI/GEO. The only change is that the IDs for the series from ArrayExpress are different than the IDs from NCBI/GEO. As with the NCBI/GEO CSV, the CSV of ArrayExpress requires only three columns as is shown at Table B.2.

Table B.2: ArrayExpress CSV Format

Comment[ENA_RUN]	Comment[FAST_URI]	Comment[LIBRARY_LAYOUT]
ERR1654640	ftp.sra.ebi.ac.uk/vol1/fastq/...	PAIRED
ERR1654640	ftp.sra.ebi.ac.uk/vol1/fastq/...	PAIRED

B.3.0.3 *GDC Portal CSV format*

GDC portal has the BAM files access restricted or controlled for the user who has access to them. However, the count files are open and can be used directly in this package as input of the function countsToMatrix. If there exist the possibility to download the controlled BAM files, the TSV file that this package uses to convert them into count files is the TSV file generated when the button Sample Sheet is clicked in the cart.

As in the other two repositories, there are a lot of columns inside the TSV files but this package only needs two of them. Furthermore, if the BAM download is carried out by the gdc-client or the web browser, the BAM has to be moved to the path ReferenceFiles/Samples/RNAseq/BAMFiles/Sample.ID/File.Name/ where Sample.ID and File.Name are the columns with the samples information in the TSV file. However, GDC portal has public access to count files that can be used in a posterior step of the KnowSeq pipeline to merge and analyse them.

B.3.0.4 *Downloading automatically GDC Portal controlled files (GDC permission required)*

It exists the possibility to download automatically the raw data from GDC portal. In order to carry this out, the function needs the parameters downloadSamples and fromGDC set to *TRUE*, the path to the token in order to obtain the authentication to download the controlled data and the path to the manifest that contains the information to download the samples. This step needs the permission of GDC portal to the controlled data.

```
# GDC portal controlled data processing with automatic raw
    data download
rawAlignment(x, downloadRef=TRUE, downloadSamples=TRUE,
    fromGDC = TRUE, tokenPath = "~/pathToToken", manifestPath
     = "~/pathToManifest")
```

B.4 RNA-SEQ PROCESSING

The RNA-Seq RAW data treatment step has the purpose of extracting a set of count files from raw files stored in the repositories supported by our package (NCBI/GEO, ArrayExpress and GDC Portal). The second one comprises the DEGs identification and extraction, and the assessment of those DEGs by applying advanced ML techniques (FS process and supervised classification). The last process, once the DEGs were assessed, is the DEGs enrichment methodology which allows retrieving biological information from the DEGs. In this process, relevant information (such as related diseases, biological processes associated and pathways) about the DEGs is retrieved by using very well-known tools and databases. The three types of enrichment are GOs study, pathways visualisation taking into account the gene expression, and DEGs related diseases.

With the pipeline designed and addressed by KnowSeq, researchers can convert RAW RNA-Seq data into real knowledge, helping to the identification of possible gene signatures related to the studied diseases.

B.4.1 *Aligners Preparation*

In order to avoid version incompatibilities with the aligners and the required external tools, pre-compiled versions of them will be used to run the R functions. Consequently, all the tools were compressed and stored in an external server to be downloaded whenever it is required by the users (`http://iwbbio.ugr.es/utils/unixUtils.tar.gz`). If the tools are directly downloaded from the link, the compressed files must be decompressed in the current project folder in R or RStudio. The name of the resultant folder must be utils. Nevertheless, this file will be automatically downloaded just by calling the function rawAlignment, in case the utils folder will not be detected in the project folder. This is all needed to run the different aligners through the function rawAlignment. It is not possible to run the alignment without the utils folder. It is also important to note that the different files included in the compressed .tar.gz are not only the aligners but also functions needed in the raw alignment process. The tools included are the following:

- Bowtie2
- Hisat2
- Htseq-count
- Kallisto
- Salmon
- Samtools
- Sratoolkit
- Tophat2
- GDC-client

B.4.2 *Launching Raw Alignment step*

The rawAlignment function allows running different aligners, chosen by the user. The function takes as single input a CSV from GEO or ArrayExpress. There is the possibility to process data from GDC portal,

but a previous authorization (token file) from this platform is required. Then, the user has to select with the seq parameter which aligner he wants to run (tophat2 by default). Furthermore, there is a set of logical parameters to edit the default alignment pipeline. With these parameters, users can select if the BAM/SAM/Count files are created or not.Users can also choose if they want to download the reference genome, the GTF, and which version. Even if the users have custom FASTA and GTF files, they can specify them by setting the parameter referenceGenome to custom and using the parameters customFA and customGTF to indicate the paths to the custom files. Other functionality is the possibility to process BAM files from the GDC Portal database by setting to TRUE the parameter fromGDC. Then, the function will download the specific genome reference from GDC, processing the BAM files to Count files. Furthermore, if users have access to the controlled data, with the token and the manifest acquired from GDC Portal web platform, samples can be automatically downloaded. An example to run the function with hisat2 aligner is showed herein:

```
# Downloading one series from NCBI/GEO and one series from
    ArrayExpress
downloadPublicSeries(c("GSE74251"))

# Using read.\gls{acr:csv} for NCBI/GEO files (read.\gls{acr
    :csv}2 for ArrayExpress files)
GSE74251\gls{acr:csv} <- read.\gls{acr:csv}("ReferenceFiles/
    GSE74251.\gls{acr:csv}")

# Performing the alignment of the samples by using hisat2
    aligner
rawAlignment(GSE74251\gls{acr:csv},seq="hisat2",downloadRef=
    TRUE,downloadSamples=TRUE,BAMfiles = TRUE, SAMfiles =
    TRUE,countFiles = TRUE,referenceGenome = 38, fromGDC =
    FALSE, customFA = "",  customGTF = "", tokenPath = "",
    manifest = "",tx2Counts = "")
```

To run the function with salmon or kallisto, it is necessary to use the parameter tx2Counts. The quantification files of these aligners contain the identification of the transcriptions, but for the count files it is necessary to convert these transcriptions IDs to gene IDs. To perform that, the tx2Counts parameter needs a matrix with two columns. One column with the transcription IDs and a second column with the correspondent gene IDs for each transcription. The package tximportData has a set of files that contain different transcript conversion that can be used to

achieve the tx2Counts matrix. An example to run the function with kallisto aligner is showed below:

```
# Downloading one series from NCBI/GEO and one series from
    ArrayExpress
downloadPublicSeries(c("GSE74251"))

# Using read.\gls{acr:csv} for NCBI/GEO files (read.\gls{acr
    :csv}2 for ArrayExpress files)
GSE74251\gls{acr:csv} <- read.\gls{acr:csv}("ReferenceFiles/
    GSE74251.\gls{acr:csv}")

# Loading the transcripts to genes converter variable
dir <- system.file("extdata", package="tximportData")
tx2gene <- read.\gls{acr:csv}(file.path(dir, "tx2gene.
    ensembl.v87.\gls{acr:csv}"))

# Performing the alignment of the samples by using kallisto
    aligner
rawAlignment(GSE74251\gls{acr:csv},seq="kallisto",
    downloadRef=TRUE,downloadSamples=TRUE,BAMfiles = TRUE,
    SAMfiles = TRUE,countFiles = TRUE,referenceGenome = 38,
    fromGDC = FALSE, customFA = "", customGTF = "", tokenPath
     = "", manifest = "",tx2Counts = tx2gene)
```

RawAlignment function creates a folder structure in the current project folder which will store all the downloaded and created files. The main folder of this structure is the folder ReferenceFiles but inside of it there are more folders that allows storing the different files used by the process in an organised way.

B.4.3 *Preparing count files*

From now on, the data that will be used for the documentation are real count files, but with a limited number of genes (around 1000). Furthermore, to reduce the computational cost of this example, only 5 samples from each of the two selected series will be taken into account. With the next code, two RNA-Seq series from NCBI/GEO are automatically downloaded and the existing count files prepared to be merged in one matrix with the purpose of preparing the data for further steps:

From now on, the data that will be used for the documentation are real count files, but with a limited number of genes (around 1000). Furthermore, to reduce the computational cost of this example, only 5 samples from each of the two selected series will be taken into account. With the next code, two RNA-Seq series from NCBI/GEO are downloaded automatically and the existing count files prepared to be merged in one matrix with the purpose of preparing the data for further steps:

```
# Downloading one series from NCBI/GEO and one series from ArrayExpress
downloadPublicSeries(c("GSE74251","GSE81593"))

# Using read.csv for NCBI/GEO files and read.csv2 for ArrayExpress files
GSE74251 <- read.csv("ReferenceFiles/GSE74251.csv")
GSE81593 <- read.csv("ReferenceFiles/GSE81593.csv")

GSE74251 <- GSE74251[1:5,]
GSE81593 <- GSE81593[8:12,]

dir <- system.file("extdata", package="KnowSeq")

# Creating the CSV file with the information about the counts files location and the labels
Run <- GSE74251$Run
Path <- paste(dir,"/countFiles/",GSE74251$Run,sep = "")
Class <- rep("Tumor", length(GSE74251$Run))
GSE74251CountsInfo <-  data.frame(Run = Run, Path = Path, Class = Class)

Run <- GSE81593$Run
Path <- paste(dir,"/countFiles/",GSE81593$Run,sep = "")
Class <- rep("Control", length(GSE81593$Run))
GSE81593CountsInfo <-  data.frame(Run = Run, Path = Path, Class = Class)

mergedCountsInfo <- rbind(GSE74251CountsInfo, GSE81593CountsInfo)

write.csv(mergedCountsInfo, file = "ReferenceFiles/mergedCountsInfo.csv")
```

However, the user can run a complete example by coding the following code:

```
dir <- system.file("script", package="KnowSeq")

# Code to execute the example script
source(paste(dir,"/KnowSeqExample.R",sep=""))

# Code to edit the example script
file.edit(paste(dir,"/KnowSeqExample.R",sep=""))
```

B.4.3.1 *Processing count files*

After the raw alignment step, a list of count files of the samples is available at ReferenceFiles/Samples/RNAseq/CountFiles. The next step in the pipeline implemented in this package is the processing of those count files in order to obtain a gene expression matrix by merging all of them.

B.4.3.2 *Merging all count files*

After the alignment, there has to be as many count files as samples in the CSV used for the alignment. In order to prepare the data for the DEGs analysis, it is important to merge all these files in one matrix that contains the genes Ensembl ID (or other IDs) in the rows and the name of the samples in the columns. To carry this out, the function countsToMatrix is available. This function reads all count files and joints them in one matrix by using edgeR package. To call the function it is only necessary a CSV with the information about the count files paths. The required CSV has to have the format shown at Table B.3.

Table B.3: Counts information CSV Format

Run	Path	Class
SRR2753159	/ReferenceFile/Count/SRR2753159/	TUMOUR
SRR2753162	/ReferenceFile/Count/SRR2753162/	TUMOUR
SRR2827426	/ReferenceFile/Count/SRR2827426/	HEALTHY
SRR2827427	/ReferenceFile/Count/SRR2827427/	HEALTHY

The column Run is the name of the sample without .count, the column Path is the Path to the count file and the Class column is the labels of the samples. Furthermore, an example of this function is shown below:

```
# Merging in one matrix all the count files indicated inside
     the CSV file
countsInformation <- countsToMatrix("ReferenceFiles/
    mergedCountsInfo.\gls{acr:csv}")

# Exporting to independent variables the counts matrix and
    the labels
countsMatrix <- countsInformation$countsMatrix
labels <- countsInformation$labels
```

The function returns a list that contains the matrix with the merged counts and the labels of the samples. It is very important to store the labels in a new variable because as it will be required in several functions of KnowSeq.

B.4.3.3 *Getting the annotation of the genes*

This step is only required if the user wants to get the gene names and the annotation is retrieved with ensembl biomaRt package. Normally, the counts matrix has the Ensembl Ids as gene identifier, but with this step, the Ensembl Ids are change by the gene names. However, the user can decide to keep its own annotation or the Ensembl Ids. For example, to achieve the gene names the function needs the current Ensembl Ids and the number of the reference genome to use for the annotation (37 or 38). If the user wants a different annotation than the human annotation, the parameter notHSapiens has to be set to TRUE and the desired specie dataset from ensembl indicated in the parameter notHumandataset (i.e. mmusculus_gene_ensembl). An example can be seen below:

```
# Downloading human annotation or MusMusculus
myAnnotation <- getAnnotationFromEnsembl(rownames(
    countsMatrix),referenceGenome=37)

myAnnotationMusMusculus <- getAnnotationFromEnsembl(rownames
    (countsMatrix), notHSapiens = TRUE,notHumandataset = "
    mmusculus_gene_ensembl")
```

B.4.3.4 *Converting to gene expression matrix*

Finally, once both the countsMatrix and the annotation are ready, it is time to convert those counts into gene expression values. For that, the function calculateGeneExpressionValues uses the cqn package to calculates the equivalent gene expression. This function performs a conversion of counts into gene expression values, and changes the Ensembl Ids by the gene names if the parameter geneNames is equal to TRUE. An example of the use of this function is showed next:

```
# Calculating gene expression values matrix using the counts
    matrix

expressionMatrix <- calculateGeneExpressionValues(
    countsMatrix,myAnnotation, genesNames = TRUE)
```

At this time of the pipeline, there is a function that plots the expression data and allows verifying if the data is well normalised. This function has the purpose of join all the important graphical representation of the pipeline in the same function and is called dataPlot. It is very easy to use because as only by changing the parameter method many different representations can be achieved. In this case, in order to see the expression boxplot of each sample, the function has to be called with the parameter mode equal to boxplot. The labels are necessary to colour the different samples depending on the class of the samples. These colours can be selected by the user, by introducing in the parameter colours a vector with the name of the desired colours. The function also allows exporting the plots as PNG and PDF files.

```
# Plotting the boxplot of the expression of each samples for
    all the genes

dataPlot(expressionMatrix,labels,mode = "boxplot", toPNG =
    TRUE, toPDF = TRUE)
```

B.5 BIOMARKERS IDENTIFICATION & ASSESSMENT

B.5.1 *Quality analysis and batch effect removal*

Before the DEGs extraction process, it is important to detect and removes any possible outlier that can be present in the samples. The outliers are samples numerically different with respect to the rest of samples, introducing noise in the study. In order to achieve that, the function RNAseqQA performs different statistical test by using arrayQualityMetrics bioc package. This package was designed for microarrays but it has been adapted in our function to allow RNA-Seq data as input. The output of this function is the same as the output of the arrayQualityMetrics package, creating a new folder with an index.html file including a report about the results of the different statistical tests and the possible detected outliers.

```
# Performing the quality analysis of the samples
RNAseqQA(expressionMatrix)
```

The other important step in this section is the batch effect treatment. It is widely known that this is a crucial step in the omics data processing due to the intrinsic deviations that the data can present due to its origin, sequencing design, i.e. Besides, when working with public data it is very difficult to know if exists a real batch effect among the selected datasets. This package provides a way of detecting possible clusters implying possible batch effect groups and correcting them. If there are batch effects in the data, it will present clusters formed because of the batch effect influence. For that, first, the function dataPlot with the parameter mode equal to optimalClusters has to be run with the purpose of detecting the optimal number of clusters existing in the samples. Furthermore, this clusters can be represented graphically by calling the function dataPlot again but this time with the parameter method equal to knnClustering. Once the optimal number of clusters is calculated, the second and final step to remove the batch effect is by calling the function batchEffectRemoval, that makes use of SVA package, with the parameter mode equal to combat and the parameter clusters equal to the optimal number of clusters calculated before. This step allows obtaining an expression matrix with the batch effect treated by combat method. An example to do this is below:

```
# Calculating the optimal number of clusters presented in
    the samples in order to
```

```
# try to identificate the batch effect groups to remove it by combat method
dataPlot(expressionMatrix,labels,mode = "optimalClusters", toPNG = TRUE,toPDF = TRUE)

dataPlot(t(expressionMatrix),labels,mode = "knnClustering", clusters = 9,toPNG = TRUE,toPDF = TRUE)

expressionMatrixCorrected <- batchEffectRemoval( expressionMatrix, labels, clusters = 9, method = "combat" )
```

There is another method in the function that removes the batch effect and it is by using surrogate variable analysis or SVA. To use this method, it is not necessary to calculates the optimal number of clusters, the only requirement to use it is to set the parameter method equal to SVA. This method does not return a matrix with the batch effect corrected, instead of this, the function returns a model that has to be used as single input parameter of the function limmaDEGsExtraction.

```
# Calculating the surrogate variable analysis to remove batch effect
svaMod <- batchEffectRemoval(expressionMatrix, labels, method = "sva")
```

B.5.2 *Differential Expressed Genes extraction and visualisation*

There is a long way between the raw data and the DEGs extraction, for that in this step the samples have to have had a strong pre-processing step applied. At this point of the pipeline the DEGs existing among two or more classes will be extracted using the most extended library for that called limma. The function limmaDEGsExtraction receives an expression matrix, the labels of the samples and the restriction imposed for considering a gene as differential expressed gene. The function returns a list containing the table with statistical values of each DEGs and the expression matrix of the DEGs instead all of the genes. The well-known limma package is used internally to perform the DEGs extraction. The call to the function is listed below:

```
# Extracting DEGs that pass the imposed restrictions
DEGsInformation <- limmaDEGsExtraction(
    expressionMatrixCorrected, labels,
lfc = 1.0, pvalue = 0.01, number = 100)

topTable <- DEGsInformation$Table

DEGsMatrix <- DEGsInformation$DEGsMatrix
```

Furthermore, if in the batch effect step the method used was SVA, this function has two parameters to indicate that the model of limma would take into account the SVA model calculated previously for the expression matrix. To achieve this, svaCorrection parameter has to be setted to TRUE and the SVA model has to be passed in the parameter svaMod. An example of this is the following:

```
# Extracting DEGs that pass the imposed restrictions but
    using sva model  calculated before to remove batch effect

DEGsInformation <- limmaDEGsExtraction(expressionMatrix,
    labels, lfc = 2.0,
pvalue = 0.01, number = Inf, svaCorrection = TRUE, svaMod =
    svaMod)

topTable <- DEGsInformation$Table

DEGsMatrix <- DEGsInformation$DEGsMatrix
```

The function also detects automatically if the labels have more than two classes and calculates the limma multiclass DEGs extraction in this case. In order to do that correctly, there is a parameter called COV that represents the number of different pathologies that a certain gen is able to discern. By default, the parameter is set to 1, so all genes that has the capability to discern among the comparison of two classes would be selected as DEGs. To understand better this parameter, our multiclass study applied to different leukemia sub-types introduces it, and it is publicly available.

DEGs are genes that have a truly different expression among the studied classes, for that it is important to try to see graphically if those DEGs comply with this requirement. In order to provide a tool to perform

this task, the function dataPlot encapsulate a set of graphs that allows plotting in different ways the expression of the DEGs.

dataPlot function also allows representing an ordered boxplot that internally orders the samples by class and plots a boxplot for each samples and for the first top 12 DEGs in this example. With this plot, the difference at gene expression level between the classes can be seen graphically. The code to reproduce this plot is the following:

```
# Plotting the expression of the first 12 DEGs for each of
    the samples in an ordered way
dataPlot(DEGsMatrix[1:12,],labels,mode = "orderedBoxplot",
    toPNG = FALSE,toPDF = FALSE)
```

In the previous boxplot the expression of a set of DEGs for each sample its showed, however it is interesting to see the differentiation at gene expression level for each of the top 12 genes used before separately. It is recommendable to use this function with a low number of genes, because with a larger number the plot it is difficult to distinguish the information provided and R would not have enough memory to calculate the plot. For that, the function dataPlot with the mode genesBoxplot allows to do that by executing the next code:

```
# Plotting the expression of the first 12 DEGs separatelly
    for all the samples
dataPlot(DEGsMatrix[1:12,],labels,mode = "genesBoxplot",
    toPNG = FALSE,toPDF = FALSE)
```

Finally, it is possible to plot one of the most widespread visualization methods in the literature, the heatmap. By setting the parameter method to heatmap, the function calculates the heatmap for the given samples and classes. The code to do this is the same than for the previous boxplot but changing the method parameter:

```
# Plotting the heatmap of the first 12 DEGs separatelly for
    all the samples
dataPlot(DEGsMatrix[1:12,],labels,mode = "heatmap",toPNG =
    FALSE,toPDF = FALSE)
```

B.5.3 *Performing the machine learning processing: classifier design and assessment and gene selection*

Normally, in the literature, the last step in the pipeline for differential gene expression analysis is the DEGs extraction step. However, in this package a novel machine learning step is implemented with the purpose of giving to the user an automatic tool to assess the DEGs, and evaluate their robustness in the discernment among the studied pathologies. This library has three possible classification methodologies to take into account. These options are k-NN, SVM and RF, three of the most popular classifiers in the literature. Furthermore, it includes two different working procedures for each of them. The first one implements a CV process, in order to assess the expected accuracy with different models and samples the DEGs with a specific number of folds. The second one is to assess a specific test dataset by using a classifier trained using the training dataset separately. Moreover, the function featureSelection allows performing a FS process by using either mRMR or RF (as feature selector instead of classifier) algorithms with the purpose of finding the best DEGs order to assess the data. The functions return a list with 4 objects that contain the confusion matrices, the accuracy, the sensitivity and the specificity.

To invoke these functions, it is necessary an expression matrix with the samples in the rows and the genes in the columns and the labels of the samples, the genes that will be assessed and the number of fold in the case of the cross-validation function. In the case of the test functions, it is necessary the matrix and the labels for both the training and the test datasets:

```
1  DEGsMatrixML <- t(DEGsMatrix)
2
3  # Feature selection process with mRMR and RF
4  mrmrRanking <- featureSelection(DEGsMatrixML,labels,colnames
       (DEGsMatrixML), mode = "mrmr")
5  rfRanking <- featureSelection(DEGsMatrixML,labels,colnames(
       DEGsMatrixML), mode = "rf")
6
7  # CV functions with k-NN, SVM and RF
8  results_cv_knn <- knn_CV(DEGsMatrixML,labels,colnames(
       DEGsMatrixML)[1:10],5)
9
10 results_cv_svm <- svm_CV(DEGsMatrixML,labels,rfRanking
       [1:10],5)
11
```

```
results_cv_rf <- rf_CV(DEGsMatrixML,labels,names(mrmrRanking)[1:10],5)
```

It is important to show graphically the results of the classifiers and for that purpose, the function dataPlot implements some methods. Concretely, to plot the accuracy, the sensitivity or the specificity reached by the classifiers, the function dataPlot has to be run with the parameter method equal to classResults. This method generated as many random colours as folds or simulations in the rows of the matrix passed to the function but, through the parameter colours a vector of desired colours can be specified. For the legend, the function uses the rownames of the input matrix but these names can be changed with the parameter legend. An example of this method is showed below:

```
# Plotting the accuracy of all the folds evaluated in the CV process
dataPlot(results_cv_knn$accMatrix,mode = "classResults", main = "Accuracy for each fold with k-NN", xlab = "Genes", ylab = "Accuracy")

# Plotting the sensitivity of all the folds evaluated in the CV process
dataPlot(results_cv_knn$sensMatrix,mode = "classResults", main = "Sensitivity for each fold with k-NN", xlab = "Genes", ylab = "Sensitivity")

# Plotting the specificity of all the folds evaluated in the CV process
dataPlot(results_cv_knn$specMatrix,mode = "classResults", main = "Specificity for each fold with k-NN", xlab = "Genes", ylab = "Specificity")
```

Furthermore, the function dataPlot counts with another similar mode to the previous but this time to represents confusion matrices. This mode is called confusionMatrix and allows creating graphically a confusion matrix with the most important statistical measures. The following code allows doing this:

```
# Plotting the confusion matrix with the sum of the confusion matrices of each folds evaluated in the CV process
```

```
allCfMats <- results_cv_knn$cfMats[[1]]$table + results_cv_knn$cfMats[[2]]$table +
results_cv_knn$cfMats[[3]]$table + results_cv_knn$cfMats[[4]]$table +
results_cv_knn$cfMats[[5]]$table

dataPlot(allCfMats,labels,mode = "confusionMatrix")
```

Once the validation is done, a test process can be carried out and the results plotted by executing the code herein:

```
# Test functions with k-NN, SVM and RF
trainingMatrix <- DEGsMatrixML[c(1:4,6:9),]
trainingLabels <- labels[c(1:4,6:9)]
testMatrix <- DEGsMatrixML[c(5,10),]
testLabels <- labels[c(5,10)]

results_test_knn <- knn_test(trainingMatrix, trainingLabels, testMatrix,
testLabels, names(mrmrRanking)[1:10])

results_test_svm <- svm_test(trainingMatrix, trainingLabels, testMatrix,
testLabels, rfRanking[1:10])

results_test_rf <- rf_test(trainingMatrix, trainingLabels, testMatrix,
testLabels, colnames(DEGsMatrixML)[1:10])

# Plotting the accuracy achieved in the test process
dataPlot(results_test_knn$accVector,mode = "classResults", main = "Accuracy with k-NN", xlab = "Genes", ylab = "Accuracy")

dataPlot(results_test_svm$accVector,mode = "classResults", main = "Accuracy with SVM", xlab = "Genes", ylab = "Accuracy")

dataPlot(results_test_rf$accVector,mode = "classResults", main = "Accuracy with RF", xlab = "Genes", ylab = "Accuracy")
```

B.6 DEGS ENRICHMENT METHODOLOGY

The main goal of the previous pipeline is the extraction of biological relevant information from the DEGs. For that, this package provides a set of tools that allows doing it. The last step of the pipeline conformed by all the available tools in KnowSeq is the DEGs enrichment and this enrichment has three different points of view. The GOs information, the pathway visualisation and the relationship between the DEGs and diseases related to the studied pathologies.

B.6.1 *Gene Ontology*

GOs provide information about the biological functions of the genes. In order to complete this pipeline, it is important to know if the DEGs have functions related with the studied pathologies. In this sense, this package brings the possibility to know the GOs from the three different ontologies (BPr, MF and CC) by using the function geneOntologyEnrichment that internally used the packaged topGO. The only requirement is to put the label of first class to 1 and the label of the second class to 0. Furthermore, with the parameter nGOs, the number of resultant GOs that are returned can be modified. The function returns a list that contains a matrix for each ontology and a matrix with the GOs of the three ontologies together. Moreover, the matrices have different statistical measures and the description of the functionality of each GO.

```
# Retrieving the GO information from the three different
    ontologies
labelsGo <- gsub("Control",0,labels)

labelsGo <- gsub("Tumor",1,labelsGo)

GOsMatrix <- geneOntologyEnrichment(DEGsMatrix,labelsGo,nGOs
     = 20)
```

B.6.2 *Pathways Visualisation*

Another important step in the enrichment methodology in this pipeline is the pathway visualisation. The function uses the DEGs to show graphically the expression of the samples in the pathways in which those genes appear. For that, the function makes use of a DEGsMatrix with the expression of the DEGs and the annotation of those DEGs in which appear the pathway or pathways of each DEGs. Internally, the function DEGsPathwayVisualization uses pathview package to retrieve and colour the pathways, but a maximum number of 24 samples can be used, for that, if the input matrix has more than 24 samples, only the first 24 will be used by the operation. Furthermore, the function needs the expression matrix with all the genes in order to use them to colour the rest of the elements in the pathways. It is important to retrieve the annotation from Ensembl for both the DEGsMatrix and the expressionMatrix because the entrezgene IDs and the KEGG enzyme of each gene are necessary.

```
1 # Downloading and filling with expression the pathways of
      the DEGs
2 myDEGsAnnotation <- getAnnotationFromEnsembl(rownames(
      DEGsMatrix)[1:3], referenceGenome=38,attributes = c("
      external_gene_name","entrezgene_id"), filters = "external
      _gene_name")
3
4 allMyAnnotation <-  getAnnotationFromEnsembl(rownames(
      expressionMatrix), referenceGenome=38,attributes =  c("
      external_gene_name","entrezgene_id"), filters = "external
      _gene_name")
5
6 DEGsPathwayVisualization(DEGsMatrix[1:3,], myDEGsAnnotation,
       expressionMatrix, allMyAnnotation, labels)
```

B.6.3 *Related Diseases*

Finally, the last enrichment method implemented is the related diseases enrichment. In this step, the function DEGsToDisease searches the diseases related to a list of genes or DEGs indicated as parameter. The function returns a list of diseases only for genes and also for group of genes with several statistical values to know the relation between

the diseases and the gene or group of genes. This information can be retrieved from two different web platforms: the first one is the Gene Set to Diseases and the second one targetValidation. The web platform to use can be chosen by changing the method parameter.

```
# Downloading the information about the DEGs related
    diseases
diseasesGenes2Diseases <- DEGsToDiseases(rownames(DEGsMatrix
    ), method = "genes2Diseases", minCitation = 2)

diseasesTargetValidation <- DEGsToDiseases(rownames(
    DEGsMatrix), method = "targetValidation", size = 5)
```

C

PUBLICATIONS

C.1 INTERNATIONAL JOURNALS WITH IMPACT FACTOR

[1] D. Castillo, J.M. Gálvez, L.J. Herrera, B.S. Román, F. Rojas, and I. Rojas. "Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling". In: *BMC Bioinformatics* 18.1 (2017). Impact Factor 2.213, Q1 in Mathematical & Computational Biology, Cited by 21. DOI: 10.1186/s12859-017-1925-0. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1925-0.

[2] D. Castillo, J.M. Galvez, L.J. Herrera, F. Rojas, O. Valenzuela, O. Caba, J. Prados, and I. Rojas. "Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level". In: *PLoS One* 14.2 (2019). Impact Factor 2.776 (2018), Q2 in Multidisciplinary Sciences (2018), Cited by 7. DOI: 10.1371/journal.pone.0212127. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0212127.

[3] J.M. Gálvez, D Castillo, L.J. Herrera, B. San Roman, O. Valenzuela, F.M. Ortuno, and I. Rojas. "Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series". In: *PloS One* 13.5 (2018). Impact Factor 2.776, Q2 in Multidisciplinary Sciences, Cited by 11, e0196836. DOI: 10.1371/journal.pone.0196836. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196836.

[4] D. Castillo-Secilla, J.M. Gálvez, F.M. Ortuno, L.J. Herrera, and I. Rojas. "KnowSeq R-bioc package: Beyond the traditional gene expression pipeline. A breast cancer RNA-seq study case. PrePrint (Version 1) available at Research Square". In: *BMC Bioinformatics, Under Review.* (2019). DOI: 10.21203/rs.2.16962/v1+. URL: https://www.researchsquare.com/article/ca3e9617-18e4-4dbe-8af7-2e2e6d4f013c/v1.

[5] J.M. Galvez, D. Castillo, L.J. Herrera, O. Valenzuela, O. Caba, J.C. Prados, F.M. Ortuno, and I. Rojas. "Towards Improving

Skin Cancer Diagnosis by Integrating Microarray and RNA-seq Datasets." In: *IEEE Journal of Biomedical and Health Informatics* (2019). Impact Factor 4.217 (2018), Q1 in Mathematical & Computational Biology (2018). DOI: 10.1109/JBHI.2019.2953978. URL: https://ieeexplore.ieee.org/document/8939388.

C.2 INTERNATIONAL CONFERENCES

[1] D. Castillo, J.M. Gálvez, L.J. Herrera, and I. Rojas. "Breast cancer microarray and RNASeq data integration applied to classification". In: *International Work-Conference on Artificial Neural Networks*. Vol. 10305 LNCS. cited by 1. 2017, pp. 123–131. DOI: 10.1007/978-3-319-59153-7_11.

[2] S. González, D. Castillo, J.M. Galvez, I. Rojas, and L.J. Herrera. In: *International Work-Conference on Artificial Neural Networks*. cited by 2. 2019, pp. 883–894. DOI: 10.1007/978-3-030-20518-8_73.

[3] D. Castillo, J.M. Gálvez, O. Valenzuela, and I. Rojas. "A NSGA-II application with different gene expression technologies integration". In: *Sixth International Conference on Advances in BioInformatics, BioTechnology and Environmental Engineering (ABBE)*. Vol. 1. 2018, pp. 21–25. DOI: 10.15224/978-1-63248-148-1-05.

[4] O. Valenzuela, A. Carrillo, D. Castillo, J.M. Gálvez, and I. Rojas. "Computer aided diagnosis of Alzheimer's disease by automatically obtaining the best coronal slices for multi-classification recognition". In: *Sixth International Conference on Advances in BioInformatics, BioTechnology and Environmental Engineering (ABBE)*. Vol. 1. 2018, pp. 16–20. DOI: 10.15224/978-1-63248-148-1-04.

D

GRANTS AND SPECIAL ACKNOWLEDGEMENTS

This thesis has been supported by the following projects and grants. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscripts that support the doctoral thesis:

- National research project **TIN2015-71873**, funded by the Spanish Ministry of Economy and Competitiveness: *Advances in Computer Architectures for machine Learning from Heterogeneous Sources: Health and Well-Being Applications*.

- Regional project **P12–TIC–2082**, funded by the Government of Andalusia: *Advanced Computer Systems for Applications in the field of Biotechnology and Bioinformatics*.

- National research project **RTI2018-101674-B-I00**, funded by the Spanish Ministry of Economy and Competitiveness: *Computer Architectures and Machine Learning-based solutions for complex challenges in Bioinformatics, Biotechnology and Biomedicine*.

I want to acknowledge the following workmates for their essential and irreplaceable collaborations in the successful development of this thesis:

- Special thanks to **Mr. Francisco M. Illeras**, from Department of Computer Architecture and Technology of University of Granada, for the constant technical support on the *BioAtc and AtcBioSimul clusters*.

- Special thanks also to **Dr. Jose C. Prados Salazar & Dr. Octavio Caba Perez**, from Department of Anatomy and Human Embryology of University of Granada, for their support with the Biological DEGs analysis in the experimental studies presented in this thesis.

- Thanks to the Institute of Bioinformatics WWU Muenster, especially to **Prof. Wojciech Makałowski** and **Priv. Doz. Dr. Eberhard Korsching**, for allowing me to perform an internship with

them, for their hospitable welcome and for sharing their valuable knowledge with me.

BIBLIOGRAPHY

[1] WHO. *2018 Cancer Statistics*. URL: https://www.who.int/news-room/fact-sheets/detail/cancer (visited on 06/11/2019) (Cited on page 4).

[2] Mark Caulfield et al. "The National Genomics Research and Healthcare Knowledgebase". In: (August 2019). DOI: 10.6084/m9.figshare.4530893.v5. URL: https://figshare.com/articles/GenomicEnglandProtocol_pdf/4530893 (Cited on page 4).

[3] European Commission. *The 1 Million Genomes Declaration*. URL: https://ec.europa.eu/digital-single-market/en/news/norway-signs-1-million-genomes-declaration (visited on 07/11/2019) (Cited on page 4).

[4] National Institute of Health. *All of Us Research Program*. URL: https://allofus.nih.gov (visited on 07/11/2019) (Cited on page 4).

[5] Gonzalo Gómez-López, Joaquın Dopazo, Juan C Cigudosa, Alfonso Valencia, and Fátima Al-Shahrour. "Precision medicine needs pioneering clinical bioinformaticians". In: *Briefings in Bioinformatics* 20.3 (2017), pp. 752–766 (Cited on pages 5, 124).

[6] D. Castillo, J.M. Gálvez, L.J. Herrera, B.S. Román, F. Rojas, and I. Rojas. "Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling". In: *BMC Bioinformatics* 18.1 (2017). Impact Factor 2.213, Q1 in Mathematical & Computational Biology, Cited by 21. DOI: 10.1186/s12859-017-1925-0. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1925-0 (Cited on pages 9, 45, 79, 99, 125).

[7] D. Castillo, J.M. Galvez, L.J. Herrera, F. Rojas, O. Valenzuela, O. Caba, J. Prados, and I. Rojas. "Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level". In: *PLoS One* 14.2 (2019). Impact Factor 2.776 (2018), Q2 in Multidisciplinary Sciences (2018), Cited by 7. DOI: 10.1371/journal.pone.0212127. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0212127 (Cited on pages 9, 45, 99, 125).

[8] D. Castillo-Secilla, J.M. Gálvez, F.M. Ortuno, L.J. Herrera, and I. Rojas. "KnowSeq R-bioc package: Beyond the traditional gene expression pipeline. A breast cancer RNA-seq study case. Pre-Print (Version 1) available at Research Square". In: *BMC Bioinformatics, Under Review.* (2019). DOI: 10.21203/rs.2.16962/v1+. URL: https://www.researchsquare.com/article/ca3e9617-18e4-4dbe-8af7-2e2e6d4f013c/v1 (Cited on pages 9, 45, 123, 157).

[9] D. Castillo-Secilla, Gálvez J.M., F.M. Ortuno, L.J. Herrera, and I. Rojas. "KnowSeq: A R package to extract knowledge by using RNA-seq raw files". In: *Bioconductor, R package version 1.1.10.* (2019). URL: http://bioconductor.org/packages/devel/bioc/html/KnowSeq.html (Cited on page 9).

[10] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. "Bioconductor: open software development for computational biology and bioinformatics". In: *Genome Biology* 5.10 (2004), R80 (Cited on page 9).

[11] Mehdi Kchouk, Jean-François Gibrat, and Mourad Elloumi. "Generations of sequencing technologies: From first to next generation". In: *Biology and Medicine* 9.3 (2017) (Cited on pages 12, 15).

[12] Rory Stark, Marta Grzelak, and James Hadfield. "RNA sequencing: the teenage years". In: *Nature Reviews Genetics* 20.11 (2019), pp. 631–656 (Cited on pages 12, 39).

[13] Miaolong Lu and Xianquan Zhan. "The crucial role of multiomic approach in cancer research and clinically relevant outcomes". In: *EPMA Journal* 9.1 (2018), pp. 77–102 (Cited on page 12).

[14] Michal Janitz. *Next-generation genome sequencing: towards personalized medicine*. John Wiley & Sons, 2011 (Cited on page 12).

[15] Michael B Shimkin. *Contrary to Nature: Being an Illustrated Commentary on Some Persons and Events of Historical Importance in the Development of Knowledge Concerning... Cancer*. Vol. 76. 720. US Department of Health, Education, and Welfare, Public Health Service ..., 1977 (Cited on page 12).

[16] Dennis A Carson and Joao M Ribeiro. "Apoptosis and disease". In: *The Lancet* 341.8855 (1993), pp. 1251–1254 (Cited on page 12).

[17] American Society of Clinical Oncology. *Stages of Cancer.* URL: https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/stages-cancer (visited on 12/11/2019) (Cited on page 13).

[18] Max Roser and Hannah Ritchie. "Cancer". In: *Our World in Data* (2019). https://ourworldindata.org/cancer (Cited on page 14).

[19] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III". In: *Journal of Experimental Medicine* 79.2 (1944), pp. 137–158 (Cited on page 15).

[20] Alfred D Hershey and Martha Chase. "Independent functions of viral protein and nucleic acid in growth of bacteriophage". In: *The Journal of General Physiology* 36.1 (1952), pp. 39–56 (Cited on page 15).

[21] W Min Jou, G Haegeman, M Ysebaert, and W Fiers. "Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein". In: *Nature* 237.5350 (1972), p. 82 (Cited on page 15).

[22] Frederick Sanger, Steven Nicklen, and Alan R Coulson. "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the National Academy of Sciences* 74.12 (1977), pp. 5463–5467 (Cited on pages 15, 16).

[23] International Human Genome Sequencing Consortium et al. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011 (2004), p. 931 (Cited on page 15).

[24] Fred Sanger and Alan R Coulson. "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *Journal of Molecular Biology* 94.3 (1975), pp. 441–448 (Cited on page 16).

[25] Allan M Maxam and Walter Gilbert. "A new method for sequencing DNA". In: *Proceedings of the National Academy of Sciences* 74.2 (1977), pp. 560–564 (Cited on page 17).

[26] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, et al. "The complete genome of an individual by massively parallel DNA sequencing". In: *Nature* 452.7189 (2008), p. 872 (Cited on page 20).

[27] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. "Accurate whole human genome sequencing using reversible terminator chemistry". In: *Nature* 456.7218 (2008), p. 53 (Cited on page 20).

[28] Vicki Pandey, Robert C Nutter, and Ellen Prediger. "Applied biosystems solid™ system: ligation-based sequencing". In: *Next Generation Genome Sequencing: Towards Personalized Medicine* (2008), pp. 29–42 (Cited on page 22).

[29] Nicole Rusk. "Torrents of sequence". In: *Nature Methods* 8.1 (2010), p. 44 (Cited on page 22).

[30] Anthony Rhoads and Kin Fai Au. "PacBio sequencing and its applications". In: *Genomics, Proteomics & Bioinformatics* 13.5 (2015), pp. 278–289 (Cited on page 25).

[31] Anika Cheerla and Olivier Gevaert. "Deep Learning with Multimodal Representation for Pancancer Prognosis Prediction". In: *bioRxiv* (2019), p. 577197 (Cited on page 28).

[32] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. "Multi-omics approaches to disease". In: *Genome Biology* 18.1 (2017), p. 83 (Cited on page 28).

[33] Konrad J Karczewski and Michael P Snyder. "Integrative omics for health and disease". In: *Nature Reviews Genetics* 19.5 (2018), p. 299 (Cited on page 28).

[34] Dan R Robinson, Yi-Mi Wu, Robert J Lonigro, Pankaj Vats, Erin Cobain, Jessica Everett, Xuhong Cao, Erica Rabban, Chandan Kumar-Sinha, Victoria Raymond, et al. "Integrative clinical genomics of metastatic cancer". In: *Nature* 548.7667 (2017), p. 297 (Cited on page 28).

[35] Elizabeth A Worthey, Alan N Mayer, Grant D Syverson, Daniel Helbling, Benedetta B Bonacci, Brennan Decker, Jaime M Serpe, Trivikram Dasu, Michael R Tschannen, Regan L Veith, et al. "Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease". In: *Genetics in Medicine* 13.3 (2011), p. 255 (Cited on page 28).

[36] Euan A Ashley, Atul J Butte, Matthew T Wheeler, Rong Chen, Teri E Klein, Frederick E Dewey, Joel T Dudley, Kelly E Ormond, Aleksandra Pavlovic, Alexander A Morgan, et al. "Clinical assessment incorporating a personal genome". In: *The Lancet* 375.9725 (2010), pp. 1525–1535 (Cited on page 28).

[37] Monya Baker. *Structural variation: the genome's hidden architecture*. 2012 (Cited on page 29).

[38] Piotr J Wysocki, Konstanty Korski, Katarzyna Lamperska, Jerzy Zaluski, and Andrzej Mackiewicz. "Primary resistance to docetaxel-based chemotherapy in metastatic breast cancer patients correlates with a high frequency of BRCA1 mutations". In: *Medical Science Monitor* 14.7 (2008), SC7–SC10 (Cited on page 30).

[39] JRM Oliveira, RM Gallindo, LGS Maia, PR Brito-Marques, Paulo Alberto Otto, Maria Rita Passos-Bueno, MA Morais Jr, and Mayana Zatz. "The short variant of the polymorphism within the promoter region of the serotonin transporter gene is a risk factor for late onset Alzheimer's disease". In: *Molecular Psychiatry* 3.5 (1998), p. 438 (Cited on page 30).

[40] Alfred G Knudson. "Mutation and cancer: statistical study of retinoblastoma". In: *Proceedings of the National Academy of Sciences* 68.4 (1971), pp. 820–823 (Cited on page 30).

[41] Yoshio Makita, Yoko Narumi, Makoto Yoshida, Tetsuya Niihori, Shigeo Kure, Kenji Fujieda, Yoichi Matsubara, and Yoko Aoki. "Leukemia in Cardio-facio-cutaneous (CFC) syndrome: a patient with a germline mutation in BRAF proto-oncogene". In: *Journal of Pediatric Hematology/Oncology* 29.5 (2007), pp. 287–290 (Cited on page 30).

[42] Ségolène Aymé, Anna Kole, and Stephen Groft. "Empowerment of patients: lessons from the rare diseases community". In: *The Lancet* 371.9629 (2008), pp. 2048–2051 (Cited on page 30).

[43] Suvi Douglas, Atte Lahtinen, Jessica Koski, Lilli Leimi, Mikko A Keränen, Kimmo Porkka, Caroline A Heckman, Kirsi Jahnukainen, Outi Kilpivaara, and Ulla Wartiovaara-Kautto. *Germline Gene Aberrations Are Common in High-Risk Adult and Pediatric Acute Lymphoblastic Leukemia Patients*. 2019 (Cited on page 30).

[44] Olga Kovalchuk, Yuri E Dubrova, Andrey Arkhipov, Barbara Hohn, and Igor Kovalchuk. "Germline DNA: Wheat mutation rate after Chernobyl". In: *Nature* 407.6804 (2000), p. 583 (Cited on page 30).

[45] Iñigo Martincorena and Peter J Campbell. "Somatic mutation in cancer and normal cells". In: *Science* 349.6255 (2015), pp. 1483–1489 (Cited on page 31).

[46] Christopher Greenman, Philip Stephens, Raffaella Smith, Gillian L Dalgliesh, Christopher Hunter, Graham Bignell, Helen Davies, Jon Teague, Adam Butler, Claire Stevens, et al. "Patterns of somatic mutation in human cancer genomes". In: *Nature* 446.7132 (2007), p. 153 (Cited on page 31).

[47] Annapurna Poduri, Gilad D Evrony, Xuyu Cai, and Christopher A Walsh. "Somatic mutation, genomic variation, and neurological disease". In: *Science* 341.6141 (2013), p. 1237758 (Cited on page 31).

[48] Zhouhuan Dong, Linghong Kong, Zhiyi Wan, Fengwei Zhu, Mei Zhong, Yali Lv, Po Zhao, and Huaiyin Shi. "Somatic mutation profiling and HER2 status in KRAS-positive Chinese colorectal cancer patients". In: *Scientific Reports* 9.1 (2019), pp. 1–7 (Cited on page 31).

[49] Reiner A Veitia. "MIRAGE Syndrome: Phenotypic Rescue by Somatic Mutation and Selection". In: *Trends in Molecular Medicine* 25.11 (2019), pp. 937–940 (Cited on page 31).

[50] Bernard Pereira, Suet-Feung Chin, Oscar M Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, et al. "The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes". In: *Nature Communications* 7 (2016), p. 11479 (Cited on page 31).

[51] Martin C Abba, Jeffrey A Drake, Kathleen A Hawkins, Yuhui Hu, Hongxia Sun, Cintia Notcovich, Sally Gaddis, Aysegul Sahin, Keith Baggerly, and C Marcelo Aldaz. "Transcriptomic changes in human breast cancer progression as determined by serial analysis of gene expression". In: *Breast Cancer Research* 6.5 (2004), R499 (Cited on page 31).

[52] Lei Zhang, James J Farrell, Hui Zhou, David Elashoff, David Akin, No–Hee Park, David Chia, and David T Wong. "Salivary transcriptomic biomarkers for detection of resectable pancreatic cancer". In: *Gastroenterology* 138.3 (2010), pp. 949–957 (Cited on page 31).

[53] Beatriz Andrea Otálora-Otálora, Mauro Florez, Liliana Lopez-Kleine, Diana Grajales Urrego, Alejandra Canas Arboleda, and Adriana Rojas. "Joint transcriptomic analysis of lung cancer and other lung diseases". In: *Frontiers in Genetics* 10 (2019), p. 1260 (Cited on page 31).

[54] Derek Wong, Kohl Lounsbury, Amy Lum, Jungeun Song, Susanna Chan, Veronique LeBlanc, Suganthi Chittaranjan, Marco Marra, and Stephen Yip. "Transcriptomic analysis of CIC and ATXN1L reveal a functional relationship exploited by cancer". In: *Oncogene* 38.2 (2019), p. 273 (Cited on page 31).

[55] Jordi Rodon, Jean-Charles Soria, Raanan Berger, Wilson H Miller, Eitan Rubin, Aleksandra Kugel, Apostolia Tsimberidou, Pierre Saintigny, Aliza Ackerstein, Irene Braña, et al. "Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial". In: *Nature Medicine* 25.5 (2019), p. 751 (Cited on page 31).

[56] Balázs Győrffy, Pawel Surowiak, Jan Budczies, and András Lánczky. "Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer". In: *PloS One* 8.12 (2013), e82241 (Cited on page 31).

[57] A Marcell Szász, András Lánczky, Ádám Nagy, Susann Förster, Kim Hark, Jeffrey E Green, Alex Boussioutas, Rita Busuttil, András Szabó, and Balázs Győrffy. "Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic

data of 1,065 patients". In: *Oncotarget* 7.31 (2016), p. 49322 (Cited on page 31).

[58] Iñigo Landa, Tihana Ibrahimpasic, Laura Boucai, Rileen Sinha, Jeffrey A Knauf, Ronak H Shah, Snjezana Dogan, Julio C Ricarte-Filho, Gnana P Krishnamoorthy, Bin Xu, et al. "Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers". In: *The Journal of Clinical Investigation* 126.3 (2016), pp. 1052–1066 (Cited on page 31).

[59] Tom W Muir, Dolan Sondhi, and Philip A Cole. "Expressed protein ligation: a general method for protein engineering". In: *Proceedings of the National Academy of Sciences* 95.12 (1998), pp. 6705–6710 (Cited on page 32).

[60] Ping-Zhen Xu, Mei-Rong Zhang, Li Gao, Yang-Chun Wu, He-Ying Qian, Gang Li, and An-Ying Xu. "Comparative Proteomic Analysis Reveals Immune Competence in Hemolymph of Bombyx mori Pupa Parasitized by Silkworm Maggot Exorista sorbillans". In: *Insects* 10.11 (2019), p. 413 (Cited on page 32).

[61] Effendi Leonard, Parayil Kumaran Ajikumar, Kelly Thayer, Wen-Hai Xiao, Jeffrey D Mo, Bruce Tidor, Gregory Stephanopoulos, and Kristala LJ Prather. "Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control". In: *Proceedings of the National Academy of Sciences* 107.31 (2010), pp. 13654–13659 (Cited on page 32).

[62] Julia M Marchingo, Linda V Sinclair, Andrew JM Howden, and Doreen A Cantrell. "Quantitative analysis of how Myc controls T cell proteomes and metabolic pathways during T cell activation". In: *bioRxiv* (2019), p. 846998 (Cited on page 32).

[63] Susanne Gräslund, Pär Nordlund, Johan Weigelt, B Martin Hallberg, James Bray, Opher Gileadi, Stefan Knapp, Udo Oppermann, Cheryl Arrowsmith, Raymond Hui, et al. "Protein production and purification". In: *Nature Methods* 5.2 (2008), p. 135 (Cited on page 32).

[64] Alfred L Goldberg. "Protein degradation and protection against misfolded or damaged proteins". In: *Nature* 426.6968 (2003), p. 895 (Cited on page 32).

[65] Arren Bar-Even, Johan Paulsson, Narendra Maheshri, Miri Carmi, Erin O'Shea, Yitzhak Pilpel, and Naama Barkai. "Noise in protein expression scales with natural protein abundance". In: *Nature Genetics* 38.6 (2006), p. 636 (Cited on page 32).

[66] Andrei A Ivanov, Fadlo R Khuri, and Haian Fu. "Targeting protein–protein interactions as an anticancer strategy". In: *Trends in Pharmacological Sciences* 34.7 (2013), pp. 393–400 (Cited on page 33).

[67] Shaomeng Wang, Yujun Zhao, Denzil Bernard, Angelo Aguilar, and Sanjeev Kumar. "Targeting the MDM2-p53 protein-protein interaction for new cancer therapeutics". In: *Protein-protein interactions*. Springer, 2012, pp. 57–79 (Cited on page 33).

[68] Melanie Spears, Karen J Taylor, Alison F Munro, Carrie A Cunningham, Elizabeth A Mallon, Chris J Twelves, David A Cameron, Jeremy Thomas, and John MS Bartlett. "In situ detection of HER2: HER2 and HER2: HER3 protein–protein interactions demonstrates prognostic significance in early breast cancer". In: *Breast Cancer Research and Treatment* 132.2 (2012), pp. 463–470 (Cited on page 33).

[69] Macdonald Morris and Steven M Watkins. "Focused metabolomic profiling in the drug development process: advances from lipid profiling". In: *Current Opinion in Chemical Biology* 9.4 (2005), pp. 407–412 (Cited on page 34).

[70] Matthias S Klein and Jane Shearer. "Metabolomics and type 2 diabetes: translating basic research into clinical application". In: *Journal of Diabetes Research* 2016 (2016) (Cited on page 34).

[71] Vicente Bodi, Vannina G Marrachelli, Oliver Husser, Francisco J Chorro, Juan R Viña, and Daniel Monleon. "Metabolomics in the diagnosis of acute myocardial ischemia". In: *Journal of Cardiovascular Translational Research* 6.5 (2013), pp. 808–815 (Cited on page 34).

[72] Emily G Armitage and Coral Barbas. "Metabolomics in cancer biomarker discovery: current trends and future perspectives". In: *Journal of Pharmaceutical and Biomedical Analysis* 87 (2014), pp. 1–11 (Cited on page 34).

[73] Kyoungmi Kim, Pavel Aronov, Stanislav O Zakharkin, Danielle Anderson, Bertrand Perroud, Ian M Thompson, and Robert H Weiss. "Urine metabolomics analysis for kidney cancer detection and biomarker discovery". In: *Molecular & Cellular Proteomics* 8.3 (2009), pp. 558–570 (Cited on page 34).

[74] Tao Zhang, Xiaoyan Wu, Chaofu Ke, Mingzhu Yin, Zhenzi Li, Lijun Fan, Wang Zhang, Haiyu Zhang, Falin Zhao, Xiaohua Zhou, et al. "Identification of potential biomarkers for ovarian cancer by urinary metabolomic profiling". In: *Journal of Proteome Research* 12.1 (2012), pp. 505–512 (Cited on page 34).

[75] Vanessa Neveu, Geneviève Nicolas, Reza M Salek, David S Wishart, and Augustin Scalbert. "Exposome-Explorer 2.0: an update incorporating candidate dietary biomarkers and dietary associations with cancer risk". In: *Nucleic Acids Research* (2019) (Cited on page 34).

[76] Amelia McCartney, Alessia Vignoli, Laura Biganzoli, Richard Love, Leonardo Tenori, Claudio Luchinat, and Angelo Di Leo. "Metabolomics in breast cancer: A decade in review". In: *Cancer Treatment Reviews* 67 (2018), pp. 88–96 (Cited on page 34).

[77] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. "Machine learning applications in cancer prognosis and prediction". In: *Computational and structural biotechnology journal* 13 (2015), pp. 8–17 (Cited on page 36).

[78] Frederic Chibon. "Cancer gene expression signatures–the rise and fall?" In: *European Journal of Cancer* 49.8 (2013), pp. 2000–2009 (Cited on page 36).

[79] Yotam Drier and Eytan Domany. "Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes?" In: *PloS One* 6.3 (2011), e17795 (Cited on page 36).

[80] Ru He and Shuguang Zuo. "A robust 8-gene prognostic signature for early-stage non-small cell lung cancer". In: *Frontiers in Oncology* 9 (2019) (Cited on page 36).

[81] Fatima Cardoso, Laura J van't Veer, Jan Bogaerts, Leen Slaets, Giuseppe Viale, Suzette Delaloge, Jean-Yves Pierga, Etienne Brain, Sylvain Causeret, Mauro DeLorenzi, et al. "70-gene signature as an aid to treatment decisions in early-stage breast cancer". In: *New England Journal of Medicine* 375.8 (2016), pp. 717–729 (Cited on page 36).

[82] Thomas F Gajewski, Jamila Louahed, and Vincent G Brichard. "Gene signature in melanoma associated with clinical activity: a potential clue to unlock cancer immunotherapy". In: *The Cancer Journal* 16.4 (2010), pp. 399–403 (Cited on page 37).

[83] Heather A Hirsch, Dimitrios Iliopoulos, Amita Joshi, Yong Zhang, Savina A Jaeger, Martha Bulyk, Philip N Tsichlis, X Shirley Liu, and Kevin Struhl. "A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases". In: *Cancer Cell* 17.4 (2010), pp. 348–361 (Cited on page 37).

[84] Tomas Bonome, Douglas A Levine, Joanna Shih, Mike Randonovich, Cindy A Pise-Masison, Faina Bogomolniy, Laurent Ozbun, John Brady, J Carl Barrett, Jeff Boyd, et al. "A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer". In: *Cancer Research* 68.13 (2008), pp. 5478–5486 (Cited on page 37).

[85] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". In: *Science* 270.5235 (1995), pp. 467–470 (Cited on pages 37, 155).

[86] Illumina. *Illumina Genes Expression arrays*. 2009. URL: http://www.exiqon.com/microrna-microarray-analysis (Cited on page 37).

[87] Hinrich Gohlmann and Willem Talloen. *Gene expression studies using Affymetrix microarrays*. CRC Press (Cited on page 37).

[88] Michael Barnes, Johannes Freudenberg, Susan Thompson, Bruce Aronow, and Paul Pavlidis. "Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms". In: *Nucleic Acids Research* 33.18 (2005), pp. 5914–5923 (Cited on pages 38, 44).

[89] Arran K Turnbull, Robert R Kitchen, Alexey A Larionov, Lorna Renshaw, J Michael Dixon, and Andrew H Sims. "Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis". In: *BMC Medical Genomics* 5.1 (2012), p. 35 (Cited on page 38).

[90] Debora Fumagalli, Alexis Blanchet-Cohen, David Brown, Christine Desmedt, David Gacquer, Stefan Michiels, Françoise Rothé, Samira Majjaj, Roberto Salgado, Denis Larsimont, et al. "Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology". In: *BMC Genomics* 15.1 (2014), p. 1008 (Cited on page 38).

[91] Marianna Zahurak, Giovanni Parmigiani, Wayne Yu, Robert B Scharpf, David Berman, Edward Schaeffer, Shabana Shabbeer, and Leslie Cope. "Pre-processing Agilent microarray data". In: *BMC Bioinformatics* 8.1 (2007), p. 142 (Cited on page 38).

[92] Exiqon. *Exiqon Genes Expression arrays*. 2009. URL: http://www.illumina.com/techniques/microarrays/gene-expression-arrays.html (Cited on page 38).

[93] Taqman. *Taqman Genes Expression arrays*. 2009. URL: https://www.thermofisher.com/es/es/home/life-science/pcr/real-time-pcr/real-time-pcr-assays.html (Cited on page 38).

[94] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1 (2009), pp. 57–63 (Cited on page 40).

[95] Bogumil Kaczkowski, Yuji Tanaka, Hideya Kawaji, Albin Sandelin, Robin Andersson, Masayoshi Itoh, Timo Lassmann, Yoshihide Hayashizaki, Piero Carninci, and Alistair RR Forrest. "Transcriptome analysis of recurrently deregulated genes across multiple cancers identifies new pan-cancer biomarkers". In: *Cancer Research* 76.2 (2016), pp. 216–226 (Cited on page 43).

[96] Jun Liang, Jing Lv, and Zimin Liu. "Identification of stage-specific biomarkers in lung adenocarcinoma based on RNA-seq data". In: *Tumor Biology* 36.8 (2015), pp. 6391–6399 (Cited on page 43).

[97] Laura S Kremer, Daniel M Bader, Christian Mertes, Robert Kopajtich, Garwin Pichler, Arcangela Iuso, Tobias B Haack, Elisabeth Graf, Thomas Schwarzmayr, Caterina Terrile, et al. "Genetic diagnosis of Mendelian disorders via RNA sequencing". In: *Nature Communications* 8 (2017), p. 15824 (Cited on page 43).

[98] Greg T Sutherland, Michal Janitz, and Jillian J Kril. "Understanding the pathogenesis of Alzheimer's disease: will RNA-Seq realize the promise of transcriptomics?" In: *Journal of Neurochemistry* 116.6 (2011), pp. 937–946 (Cited on page 43).

[99] Lei Xu, Aik Choon Tan, Daniel Q Naiman, Donald Geman, and Raimond L Winslow. "Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data". In: *Bioinformatics* 21.20 (2005), pp. 3905–3911 (Cited on page 44).

[100] Cosmin Lazar, Stijn Meganck, Jonatan Taminau, David Steenhoff, Alain Coletta, Colin Molter, David Y Weiss-Solıs, Robin Duque, Hugues Bersini, and Ann Nowé. "Batch effect removal methods for microarray gene expression data integration: a survey". In: *Briefings in Bioinformatics* 14.4 (2012), pp. 469–490 (Cited on page 44).

[101] Andreas Heider and Rüdiger Alt. "virtualArray: a R/bioconductor package to merge raw data from different microarray platforms". In: *BMC Bioinformatics* 14.1 (2013), p. 75 (Cited on page 44).

[102] Michael Dondrup, Stefan P Albaum, Thasso Griebel, Kolja Henckel, Sebastian Jünemann, Tim Kahlke, Christiane K Kleindt, Helge Küster, Burkhard Linke, Dominik Mertens, et al. "EMMA 2–a MAGE-compliant system for the collaborative analysis and integration of microarray data". In: *BMC Bioinformatics* 10.1 (2009), p. 50 (Cited on page 44).

[103] Fatima Al-Shahrour, Pablo Minguez, Joaquın Tárraga, Ignacio Medina, Eva Alloza, David Montaner, and Joaquın Dopazo. "FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction

data with microarray experiments". In: *Nucleic Acids Research* 35.suppl_2 (2007), W91–W96 (Cited on page 44).

[104] Intawat Nookaew, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlén, and Jens Nielsen. "A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae". In: *Nucleic Acids Research* 40.20 (2012), pp. 10084–10097 (Cited on pages 44, 81).

[105] J.M. Gálvez, D Castillo, L.J. Herrera, B. San Roman, O. Valenzuela, F.M. Ortuno, and I. Rojas. "Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series". In: *PloS One* 13.5 (2018). Impact Factor 2.776, Q2 in Multidisciplinary Sciences, Cited by 11, e0196836. DOI: 10.1371/journal.pone.0196836. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196836 (Cited on page 45).

[106] J.M. Galvez, D. Castillo, L.J. Herrera, O. Valenzuela, O. Caba, J.C. Prados, F.M. Ortuno, and I. Rojas. "Towards Improving Skin Cancer Diagnosis by Integrating Microarray and RNA-seq Datasets." In: *IEEE Journal of Biomedical and Health Informatics* (2019). Impact Factor 4.217 (2018), Q1 in Mathematical & Computational Biology (2018). DOI: 10.1109/JBHI.2019.2953978. URL: https://ieeexplore.ieee.org/document/8939388 (Cited on page 45).

[107] Shweta S Chavan, Michael A Bauer, Erich A Peterson, Christoph J Heuck, and Donald J Johann. "Towards the integration, annotation and association of historical microarray experiments with RNA-seq". In: *BMC Bioinformatics*. Vol. 14. 14. BioMed Central. 2013, S4 (Cited on page 45).

[108] Alina Sırbu, Gráinne Kerr, Martin Crane, and Heather J Ruskin. "RNA-Seq vs dual-and single-channel microarray data: sensitivity analysis for differential expression and clustering". In: *PloS One* 7.12 (2012), e50986 (Cited on page 45).

[109] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. *Data integration in the era of omics: current and future challenges*. 2014 (Cited on page 45).

[115] Thomas Bayes. "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S". In:

Philosophical Transactions of the Royal Society of London 53 (1763), pp. 370–418 (Cited on page 48).

[116] Jingnian Chen, Houkuan Huang, Shengfeng Tian, and Youli Qu. "Feature selection for text classification with Naıve Bayes". In: *Expert Systems with Applications* 36.3 (2009), pp. 5432–5435 (Cited on page 50).

[117] Nicu Sebe, Michael S Lew, Ira Cohen, Ashutosh Garg, and Thomas S Huang. "Emotion recognition using a cauchy naive bayes classifier". In: *Object recognition supported by user interaction for service robots*. Vol. 1. IEEE. 2002, pp. 17–20 (Cited on page 50).

[118] Mrutyunjaya Panda and Manas Ranjan Patra. "Network intrusion detection using naive bayes". In: *International Journal of Computer Science and Network Security* 7.12 (2007), pp. 258–263 (Cited on page 50).

[119] Malik Yousef, Michael Nebozhyn, Hagit Shatkay, Stathis Kanterakis, Louise C Showe, and Michael K Showe. "Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier". In: *Bioinformatics* 22.11 (2006), pp. 1325–1334 (Cited on page 50).

[120] Evelyn Fix and Joseph L Hodges Jr. *Discriminatory analysis-nonparametric discrimination: consistency properties*. Tech. rep. California Univ Berkeley, 1951 (Cited on page 50).

[121] Thomas Cover and Peter Hart. "Nearest neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27 (Cited on page 50).

[122] Ying Huang and Yanda Li. "Prediction of protein subcellular locations using fuzzy k-NN method". In: *Bioinformatics* 20.1 (2004), pp. 21–28 (Cited on page 52).

[123] Claudia Nickel, Tobias Wirtl, and Christoph Busch. "Authentication of smartphone users based on the way they walk using k-nn algorithm". In: *2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE. 2012, pp. 16–20 (Cited on page 52).

[124] Fabio Maselli, Gherardo Chirici, Lorenzo Bottai, Piermaria Corona, and Marco Marchetti. "Estimation of Mediterranean forest attributes by the application of k-NN procedures to multitemporal Landsat ETM+ images". In: *International Journal of Remote Sensing* 26.17 (2005), pp. 3781–3796 (Cited on page 52).

[125] Hoai-Vu Nguyen and Yongsun Choi. "Proactive detection of DDoS attacks utilizing k-NN classifier in an anti-DDoS framework". In: *International Journal of Electrical, Computer, and Systems Engineering* 4.4 (2010), pp. 247–252 (Cited on page 52).

[126] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine Learning* 20.3 (1995), pp. 273–297 (Cited on pages 53, 131, 160).

[127] Christina Leslie, Eleazar Eskin, and William Stafford Noble. "The spectrum kernel: A string kernel for SVM protein classification". In: *Biocomputing 2002*. World Scientific, 2001, pp. 564–575 (Cited on page 55).

[128] Christian Schuldt, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach". In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Vol. 3. IEEE. 2004, pp. 32–36 (Cited on page 55).

[129] Tony Thomas, Athira P Vijayaraghavan, and Sabu Emmanuel. "Support Vector Machines and Malware Detection". In: *Machine Learning Approaches in Cyber Security Analytics*. Springer, 2020, pp. 49–71 (Cited on page 55).

[130] Naoki Shibata, Hiroto Shinoda, Hidetsugu Nanba, Aya Ishino, and Toshiyuki Takezawa. "Classification and Visualization of Travel Blog Entries Based on Types of Tourism". In: *Information and Communication Technologies in Tourism 2020*. Springer, 2020, pp. 27–37 (Cited on page 55).

[131] Leo Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32 (Cited on page 56).

[132] Alexandre Bureau, Josée Dupuis, Kathleen Falls, Kathryn L Lunetta, Brooke Hayward, Tim P Keith, and Paul Van Eerdewegh. "Identifying SNPs predictive of phenotype using random forests". In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 28.2 (2005), pp. 171–182 (Cited on page 58).

[133] Xue-Wen Chen and Mei Liu. "Prediction of protein–protein interactions using random decision forest framework". In: *Bioinformatics* 21.24 (2005), pp. 4394–4400 (Cited on page 58).

[134] J Albert, E Aliu, H Anderhub, P Antoranz, A Armada, M Asensio, C Baixeras, JA Barrio, H Bartko, D Bastieri, et al. "Implementation of the random forest method for the imaging atmospheric Cherenkov telescope MAGIC". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 588.3 (2008), pp. 424–432 (Cited on page 58).

[135] Quanlong Feng, Jiantao Liu, and Jianhua Gong. "UAV remote sensing for urban vegetation mapping using random forest and texture analysis". In: *Remote Sensing* 7.1 (2015), pp. 1074–1094 (Cited on page 58).

[136] Kenji Kira and Larry A Rendell. "A practical approach to feature selection". In: *Machine Learning Proceedings 1992*. Elsevier, 1992, pp. 249–256 (Cited on page 59).

[137] Hanchuan Peng, Fuhui Long, and Chris Ding. "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 8 (2005), pp. 1226–1238 (Cited on pages 60, 64).

[138] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection". In: *Journal of Machine Learning Research* 13.Jan (2012), pp. 27–66 (Cited on page 61).

[139] Ramón Dıaz-Uriarte and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest". In: *BMC Bioinformatics* 7.1 (2006), p. 3 (Cited on pages 62, 130, 160).

[140] Anna Louise Swan, Ali Mobasheri, David Allaway, Susan Liddell, and Jaume Bacardit. "Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology". In: *Omics: A Journal of Integrative Biology* 17.12 (2013), pp. 595–610 (Cited on page 64).

[141] Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods". In: *Bioinformatics* 26.3 (2009), pp. 392–398 (Cited on page 64).

[142] Choong-Wan Woo and Tor D Wager. "Neuroimaging-based biomarker discovery and validation". In: *Pain* 156.8 (2015), p. 1379 (Cited on page 64).

[143] Francisco Azuaje, Yvan Devaux, and Daniel Wagner. "Computational biology for cardiovascular biomarker discovery". In: *Briefings in Bioinformatics* 10.4 (2009), pp. 367–377 (Cited on page 64).

[144] Jorng-Tzong Horng, Li-Cheng Wu, Baw-Juine Liu, Jun-Li Kuo, Wen-Horng Kuo, and Jin-Jian Zhang. "An expert system to classify microarray gene expression data using gene selection by decision tree". In: *Expert Systems with Applications* 36.5 (2009), pp. 9072–9081 (Cited on page 64).

[145] Alexander Statnikov, Ioannis Tsamardinos, Yerbolat Dosbayev, and Constantin F Aliferis. "GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data". In: *International Journal of Medical Informatics* 74.7-8 (2005), pp. 491–503 (Cited on page 64).

[146] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. "Applications of machine learning in drug discovery and development". In: *Nature Reviews Drug Discovery* 18.6 (2019), pp. 463–477 (Cited on page 64).

[147] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. "NCBI GEO: archive for functional genomics data sets—update". In: *Nucleic Acids Research* 41.D1 (2012), pp. D991–D995 (Cited on page 70).

[148] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. "Toward a shared vision for cancer genomic data". In: *New England Journal of Medicine* 375.12 (2016), pp. 1109–1112 (Cited on pages 70, 125).

[149] Laurent Gautier, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry. "affy—analysis of Affymetrix GeneChip data at the probe level". In: *Bioinformatics* 20.3 (2004), pp. 307–315 (Cited on page 71).

[150] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data". In: *Biostatistics* 4.2 (2003), pp. 249–264 (Cited on page 71).

[151] Simon M Lin, Pan Du, Wolfgang Huber, and Warren A Kibbe. "Model-based variance-stabilizing transformation for Illumina microarray data". In: *Nucleic Acids Research* 36.2 (2008), e11–e11 (Cited on page 71).

[152] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biology* 14.4 (2013), R36 (Cited on pages 71, 129).

[153] Daehwan Kim, Ben Langmead, and Steven L Salzberg. "HISAT: a fast spliced aligner with low memory requirements". In: *Nature Methods* 12.4 (2015), p. 357 (Cited on pages 71, 129).

[154] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq—a Python framework to work with high-throughput sequencing data". In: *Bioinformatics* 31.2 (2015), pp. 166–169 (Cited on pages 72, 129).

[155] Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. "arrayQualityMetrics—a bioconductor package for quality assessment of microarray data". In: *Bioinformatics* 25.3 (2008), pp. 415–416 (Cited on pages 72, 130).

[156] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. "Entrez Gene: gene-centered information at NCBI". In: *Nucleic Acids Research* 33.suppl_1 (2005), pp. D54–D58 (Cited on page 73).

[157] Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, James Allen, M Ridwan Amode, Irina M Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, et al. "Ensembl 2019". In: *Nucleic Acids Research* 47.D1 (2018), pp. D745–D751 (Cited on page 73).

[158] Tina A Eyre, Fabrice Ducluzeau, Tam P Sneddon, Sue Povey, Elspeth A Bruford, and Michael J Lush. "The HUGO gene nomenclature database, 2006 updates". In: *Nucleic Acids Research* 34.suppl_1 (2006), pp. D319–D321 (Cited on page 73).

[159] Wilson Wen Bin Goh, Wei Wang, and Limsoon Wong. "Why batch effects matter in omics data, and how to avoid them". In: *Trends in Biotechnology* 35.6 (2017), pp. 498–507 (Cited on pages 73, 130, 158).

[160] W Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127 (Cited on page 74).

[161] Jeffrey T Leek and John D Storey. "Capturing heterogeneity in gene expression studies by surrogate variable analysis". In: *PLoS Genetics* 3.9 (2007), e161 (Cited on pages 74, 133).

[162] Laurent Jacob, Johann A Gagnon-Bartsch, and Terence P Speed. "Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed". In: *Biostatistics* 17.1 (2015), pp. 16–28 (Cited on page 74).

[163] Gordon K Smyth. "Limma: linear models for microarray data". In: *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, 2005, pp. 397–420 (Cited on pages 75, 130).

[164] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015), e47–e47 (Cited on page 75).

[165] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. "Bioconductor: open software development for computational biology and bioinformatics". In: *Genome Biology* 5.10 (2004), R80 (Cited on page 78).

[166] T. H. Kim, J. S. Chang, K. S. Park, J. Park, N. Kim, J. I. Lee, and I. D. Kong. "Effects of exercise training on circulating levels of Dickkpof-1 and secreted frizzled-related protein-1 in breast cancer survivors: A pilot single-blind randomized controlled trial". In: *PLoS One* 12.2 (2017), e0171771. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0171771. URL: http://www.ncbi.nlm.nih.gov/pubmed/28178355 (Cited on page 94).

[167] L. Y. Kong, M. Xue, Q. C. Zhang, and C. F. Su. "In vivo and in vitro effects of microRNA-27a on proliferation, migration and invasion of breast cancer cells through targeting of SFRP1 gene via Wnt/beta-catenin signaling pathway". In: *Oncotarget* (2017). ISSN: 1949-2553 (Electronic) 1949-2553 (Linking). DOI: 10.18632/oncotarget.14662. URL: http://www.ncbi.nlm.nih.gov/pubmed/28099945 (Cited on page 94).

[168] K. Mitrunen, N. Jourenkova, V. Kataja, M. Eskelinen, V. M. Kosma, S. Benhamou, H. Vainio, M. Uusitupa, and A. Hirvonen. "Glutathione S-transferase M1, M3, P1, and T1 genetic polymorphisms and susceptibility to breast cancer". In: *Cancer Epidemiol Biomarkers Prev* 10.3 (2001), pp. 229–36. ISSN: 1055-9965 (Print) 1055-9965 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/11303592 (Cited on page 94).

[169] J. Y. Choi et al. "Genetic polymorphisms of SULT1A1 and SULT1E1 and the risk and survival of breast cancer". In: *Cancer Epidemiol Biomarkers Prev* 14.5 (2005), pp. 1090–5. ISSN: 1055-9965 (Print) 1055-9965 (Linking). DOI: 10.1158/1055-9965.EPI-04-0688. URL: http://www.ncbi.nlm.nih.gov/pubmed/15894657 (Cited on page 95).

[170] Y. Xu, X. Liu, F. Guo, Y. Ning, X. Zhi, X. Wang, S. Chen, L. Yin, and X. Li. "Effect of estrogen sulfation by SULT1E1 and PAPSS on the development of estrogen-dependent cancers". In: *Cancer Sci* 103.6 (2012), pp. 1000–9. ISSN: 1349-7006 (Electronic) 1347-9032 (Linking). DOI: 10.1111/j.1349-7006.2012.02258.x. URL: http://www.ncbi.nlm.nih.gov/pubmed/22380844 (Cited on page 95).

[171] S. E. Flonta, S. Arena, A. Pisacane, P. Michieli, and A. Bardelli. "Expression and functional regulation of myoglobin in epithelial cancers". In: *Am J Pathol* 175.1 (2009), pp. 201–6. ISSN: 1525-2191

(Electronic) 0002-9440 (Linking). DOI: 10.2353/ajpath.2009.081124. URL: http://www.ncbi.nlm.nih.gov/pubmed/19541931 (Cited on page 95).

[172] G. Kristiansen et al. "Endogenous myoglobin in breast cancer is hypoxia-inducible by alternative transcription and functions to impair mitochondrial activity: a role in tumor suppression?" In: *J Biol Chem* 286.50 (2011), pp. 43417–28. ISSN: 1083-351X (Electronic) 0021-9258 (Linking). DOI: 10.1074/jbc.M111.227553. URL: http://www.ncbi.nlm.nih.gov/pubmed/21930697 (Cited on page 95).

[173] A. Bicker, A. M. Brahmer, S. Meller, G. Kristiansen, T. A. Gorr, and T. Hankeln. "The Distinct Gene Regulatory Network of Myoglobin in Prostate and Breast Cancer". In: *PLoS One* 10.11 (2015), e0142662. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0142662. URL: http://www.ncbi.nlm.nih.gov/pubmed/26559958 (Cited on page 95).

[174] L. Ai et al. "TRIM29 suppresses TWIST1 and invasive breast cancer behavior". In: *Cancer Res* 74.17 (2014), pp. 4875–87. ISSN: 1538-7445 (Electronic) 0008-5472 (Linking). DOI: 10.1158/0008-5472.CAN-13-3579. URL: http://www.ncbi.nlm.nih.gov/pubmed/24950909 (Cited on page 95).

[175] E Karjalainen and GA Repasky. "Molecular changes during acute myeloid leukemia (AML) evolution and identification of novel treatment strategies through molecular stratification". In: *Progress in molecular biology and translational science*. Vol. 144. Elsevier, 2016, pp. 383–436 (Cited on page 101).

[176] Nicholas J Short, Michael E Rytting, and Jorge E Cortes. "Acute myeloid leukaemia". In: *The Lancet* 392.10147 (2018), pp. 593–606 (Cited on page 101).

[177] Shilpa Paul, Hagop Kantarjian, and Elias J Jabbour. "Adult acute lymphoblastic leukemia". In: *Mayo Clinic Proceedings*. Vol. 91. 11. Elsevier. 2016, pp. 1645–1666 (Cited on page 101).

[178] Sabina Chiaretti, Monica Messina, and Robin Foà. "BCR/ABL1–like acute lymphoblastic leukemia: How to diagnose and treat?" In: *Cancer* 125.2 (2019), pp. 194–204 (Cited on page 101).

[179] Junia V Melo and David J Barnes. "Chronic myeloid leukaemia as a model of disease evolution in human cancer". In: *Nature Reviews Cancer* 7.6 (2007), p. 441 (Cited on page 101).

[180] Ali Amin Asnafi, Zeinab Deris Zayeri, Saeid Shahrabi, Kazem Zibara, and Tina Vosughi. "Chronic myeloid leukemia with complex karyotypes: Prognosis and therapeutic approaches". In: *Journal of Cellular Physiology* 234.5 (2019), pp. 5798–5806 (Cited on page 101).

[181] Muhammad Haseeb, Muhammad Ayaz Anwar, and Sangdun Choi. "Molecular interactions between innate and adaptive immune cells in chronic lymphocytic leukemia and their therapeutic implications". In: *Frontiers in Immunology* 9 (2018), p. 2720 (Cited on page 101).

[182] Maurizio Cavallari, Francesco Cavazzini, Antonella Bardi, Eleonora Volta, Aurora Melandri, Elisa Tammiso, Elena Saccenti, Enrico Lista, Francesca Maria Quaglia, Antonio Urso, et al. "Biological significance and prognostic/predictive impact of complex karyotype in chronic lymphocytic leukemia". In: *Oncotarget* 9.76 (2018), p. 34398 (Cited on page 101).

[183] Alexander Statnikov and Constantin F Aliferis. "Are random forests better than support vector machines for microarray-based cancer classification?" In: *AMIA annual symposium proceedings*. Vol. 2007. American Medical Informatics Association. 2007, p. 686 (Cited on page 117).

[184] Alexander Statnikov, Lily Wang, and Constantin F Aliferis. "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification". In: *BMC Bioinformatics* 9.1 (2008), p. 319 (Cited on page 117).

[185] Sung-Bae Cho and Hong-Hee Won. "Machine learning in DNA microarray analysis for cancer classification". In: *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19*. Australian Computer Society, Inc. 2003, pp. 189–198 (Cited on page 117).

[186] Karin Reif and Jason G Cyster. "The CDM protein DOCK2 in lymphocyte migration". In: *Trends in Cell Biology* 12.8 (2002), pp. 368–373 (Cited on page 119).

[187] Md Kamrul Hasan, Jian Yu, George F Widhopf, Laura Z Rassenti, Liguang Chen, Zhouxin Shen, Steven P Briggs, Donna S Neuberg, and Thomas J Kipps. "Wnt5a induces ROR1 to recruit DOCK2 to activate Rac1/2 in chronic lymphocytic leukemia". In: *Blood* 132.2 (2018), pp. 170–178 (Cited on page 119).

[188] Min Wu, Donald Small, and Amy S Duffield. "DOCK2: A novel FLT3/ITD leukemia drug target". In: *Oncotarget* 8.51 (2017), p. 88253 (Cited on page 119).

[189] Sabrina Crivellaro, Giovanna Carrà, Cristina Panuzzo, Riccardo Taulli, Angelo Guerrasio, Giuseppe Saglio, and Alessandro Morotti. "The non-genomic loss of function of tumor suppressors: an essential role in the pathogenesis of chronic myeloid leukemia chronic phase". In: *BMC Cancer* 16.1 (2016), p. 314 (Cited on page 119).

[190] Haojian Zhang, Cong Peng, Yiguo Hu, Huawei Li, Zhi Sheng, Yaoyu Chen, Con Sullivan, Jan Cerny, Lloyd Hutchinson, Anne Higgins, et al. "The Blk pathway functions as a tumor suppressor in chronic myeloid leukemia stem cells". In: *Nature Genetics* 44.8 (2012), p. 861 (Cited on page 119).

[191] Ekaterina Kim, Christian Hurtz, Stefan Koehrer, Zhiqiang Wang, Sriram Balasubramanian, Betty Y Chang, Markus Müschen, R Eric Davis, and Jan A Burger. "Ibrutinib inhibits pre-BCR+ B-cell acute lymphoblastic leukemia progression by targeting BTK and BLK". In: *Blood* 129.9 (2017), pp. 1155–1165 (Cited on page 119).

[192] Kai Xue, Jiazhe Song, Yan Yang, Zhi Li, Chunhua Wu, Jinhua Jin, and Wenzhe Li. "PAX5 promotes pre-B cell proliferation by regulating the expression of pre-B cell receptor and its downstream signaling". In: *Molecular Immunology* 73 (2016), pp. 1–9 (Cited on page 119).

[193] Joji Nakayama, Mutsumi Yamamoto, Katsuhiko Hayashi, Hitoshi Satoh, Kenji Bundo, Masato Kubo, Ryo Goitsuka, Michael A Farrar, and Daisuke Kitamura. "BLNK suppresses pre–B-cell leukemogenesis through inhibition of JAK3". In: *Blood* 113.7 (2009), pp. 1483–1492 (Cited on page 119).

[194] Naoto Imoto, Fumihiko Hayakawa, Shingo Kurahashi, Takanobu Morishita, Yuki Kojima, Takahiko Yasuda, Keiki Sugimoto, Shinobu Tsuzuki, Tomoki Naoe, and Hitoshi Kiyoi. "B cell linker protein (BLNK) is a selective target of repression by PAX5-PML protein in the differentiation block that leads to the development of acute lymphoblastic leukemia". In: *Journal of Biological Chemistry* 291.9 (2016), pp. 4723–4731 (Cited on page 119).

[195] Juan Carlos Núñez-Enrıquez, Diego Alberto Bárcenas-López, Alfredo Hidalgo-Miranda, Elva Jiménez-Hernández, Vilma Carolina Bekker-Méndez, Janet Flores-Lujano, Karina Anastacia Solis-Labastida, Gabriela Bibiana Martınez-Morales, Fausto Sánchez-Muñoz, Laura Eugenia Espinoza-Hernández, et al. "Gene expression profiling of acute lymphoblastic leukemia in children with very early relapse". In: *Archives of Medical Research* 47.8 (2016), pp. 644–655 (Cited on page 120).

[196] Yuanzheng Peng, Juanjuan Yuan, Zhenchao Zhang, and Xing Chang. "Cytoplasmic poly (A)-binding protein 1 (PABPC1) interacts with the RNA-binding protein hnRNPLL and thereby regulates immunoglobulin secretion in plasma cells". In: *Journal of Biological Chemistry* 292.29 (2017), pp. 12285–12295 (Cited on page 120).

[197] Caroline Huygens, Stéphanie Liénart, Olivier Dedobbeleer, Julie Stockis, Emilie Gauthy, Pierre G Coulie, and Sophie Lucas. "Lysosomal-associated transmembrane protein 4B (LAPTM4B) decreases transforming growth factor β1 (TGF-β1) production in human regulatory T cells". In: *Journal of Biological Chemistry* 290.33 (2015), pp. 20105–20116 (Cited on page 120).

[198] Liang Huang, Kuangguo Zhou, Yunfan Yang, Zhen Shang, Jue Wang, Di Wang, Na Wang, Danmei Xu, and Jianfeng Zhou. "FLT3-ITD-associated gene-expression signatures in NPM1-mutated cytogenetically normal acute myeloid leukemia". In: *International Journal of Hematology* 96.2 (2012), pp. 234–240 (Cited on page 120).

[199] Yi Huang, Jian-Da Hu, Yuan-Lin Qi, Yan-An Wu, Jing Zheng, Ying-Yu Chen, and Xiao-Li Huang. "Effect of knocking down eEF1A1 gene on proliferation and apoptosis in Jurkat cells and its mechanisms". In: *Zhongguo shi yan xue ye xue za zhi* 20.4 (2012), pp. 835–841 (Cited on page 120).

[200] Kenji Daigo, Naotaka Yamaguchi, Takeshi Kawamura, Koichi Matsubara, Shuying Jiang, Riuko Ohashi, Yukio Sudou, Tatsuhiko Kodama, Makoto Naito, Kenji Inoue, et al. "The proteomic profile of circulating pentraxin 3 (PTX3) complex in sepsis demonstrates the interaction with azurocidin 1 and other components of neutrophil extracellular traps". In: *Molecular & Cellular Proteomics* 11.6 (2012), pp. M111–015073 (Cited on page 120).

[201] Kihoon Cha, Yi Li, and Gwan-Su Yi. "Discovering gene expression signatures responding to tyrosine kinase inhibitor treatment in chronic myeloid leukemia". In: *BMC Medical Genomics* 9.1 (2016), p. 29 (Cited on page 120).

[202] J Dunne, C Cullmann, M Ritter, N Martinez Soria, B Drescher, S Debernardi, S Skoulakis, O Hartmann, M Krause, J Krauter, et al. "siRNA-mediated AML1/MTG8 depletion affects differentiation and proliferation-associated gene expression in t (8; 21)-positive cell lines and primary AML blasts". In: *Oncogene* 25.45 (2006), p. 6067 (Cited on page 120).

[203] Dan A Landau, Eugen Tausch, Amaro N Taylor-Weiner, Chip Stewart, Johannes G Reiter, Jasmin Bahlo, Sandra Kluth, Ivana Bozic, Mike Lawrence, Sebastian Böttcher, et al. "Mutations driving CLL and their evolution in progression and relapse". In: *Nature* 526.7574 (2015), p. 525 (Cited on page 120).

[204] Viktor Ljungström, Diego Cortese, Emma Young, Tatjana Pandzic, Larry Mansouri, Karla Plevova, Stavroula Ntoufa, Panagiotis Baliakas, Ruth Clifford, Lesley-Ann Sutton, et al. "Whole-exome

sequencing in relapsing chronic lymphocytic leukemia: clinical impact of recurrent RPS15 mutations". In: *Blood* 127.8 (2016), pp. 1007–1016 (Cited on page 120).

[205] Hanna T Gazda, Agnieszka Grabowska, Lilia B Merida-Long, Elzbieta Latawiec, Hal E Schneider, Jeffrey M Lipton, Adrianna Vlachos, Eva Atsidaftos, Sarah E Ball, Karen A Orfali, et al. "Ribosomal protein S24 gene is mutated in Diamond-Blackfan anemia". In: *The American Journal of Human Genetics* 79.6 (2006), pp. 1110–1118 (Cited on page 120).

[206] Toshio Ota, Yutaka Suzuki, Tetsuo Nishikawa, Tetsuji Otsuki, Tomoyasu Sugiyama, Ryotaro Irie, Ai Wakamatsu, Koji Hayashi, Hiroyuki Sato, Keiichi Nagai, et al. "Complete sequencing and characterization of 21,243 full-length human cDNAs". In: *Nature Genetics* 36.1 (2004), p. 40 (Cited on page 120).

[207] Gregory W Roloff and Elizabeth A Griffiths. "When to obtain genomic data in acute myeloid leukemia (AML) and which mutations matter". In: *Hematology 2014, the American Society of Hematology Education Program Book* 2018.1 (2018), pp. 35–44 (Cited on page 120).

[208] Hubert Hackl, Ksenia Astanina, and Rotraud Wieser. "Molecular and genetic alterations associated with therapy resistance and relapse of acute myeloid leukemia". In: *Journal of Hematology & Oncology* 10.1 (2017), p. 51 (Cited on page 120).

[209] Pieter Van Vlierberghe and Adolfo Ferrando. "The molecular basis of T cell acute lymphoblastic leukemia". In: *The Journal of Clinical Investigation* 122.10 (2012), pp. 3398–3406 (Cited on page 120).

[210] David Grimwade, Adam Ivey, and Brian JP Huntly. "Molecular landscape of acute myeloid leukemia in younger adults and its clinical relevance". In: *Blood, The Journal of the American Society of Hematology* 127.1 (2016), pp. 29–41 (Cited on page 121).

[211] Ali Amin Asnafi, Zeinab Deris Zayeri, Saeid Shahrabi, Kazem Zibara, and Tina Vosughi. "Chronic myeloid leukemia with complex karyotypes: Prognosis and therapeutic approaches". In: *Journal of Cellular Physiology* 234.5 (2019), pp. 5798–5806 (Cited on page 121).

[212] Yaoyu Chen, Cong Peng, Con Sullivan, Dongguang Li, and Shaoguang Li. "Critical molecular pathways in cancer stem cells of chronic myeloid leukemia". In: *Leukemia* 24.9 (2010), p. 1545 (Cited on page 121).

[213] Lynn R Goldin, Ruth M Pfeiffer, Xinjun Li, and Kari Hemminki. "Familial risk of lymphoproliferative tumors in families of patients with chronic lymphocytic leukemia: results from the Swedish Family-Cancer Database". In: *Blood* 104.6 (2004), pp. 1850–1854 (Cited on page 121).

[214] Valeria Spina and Davide Rossi. "Overview of non-coding mutations in chronic lymphocytic leukemia". In: *Molecular Oncology* 13.1 (2019), pp. 99–106 (Cited on page 121).

[215] Bastian Seelbinder, Thomas Wolf, Steffen Priebe, Sylvie McNamara, Silvia Gerber, Reinhard Guthke, and Joerg Linde. "GEO2RNAseq: An easy-to-use R pipeline for complete pre-processing of RNA-seq data". In: *bioRxiv* (2019), p. 771063 (Cited on page 124).

[216] Marc Lohse, Anthony M Bolger, Axel Nagel, Alisdair R Fernie, John E Lunn, Mark Stitt, and Björn Usadel. "R obi NA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics". In: *Nucleic Acids Research* 40.W1 (2012), W622–W627 (Cited on page 124).

[217] Kuan-Hao Chao, Yi-Wen Hsiao, Yi-Fang Lee, Chien-Yueh Lee, Liang-Chuan Lai, Mong-Hsun Tsai, Tzu-Pin Lu, and Eric Y Chuang. "RNASeqR: an R package for automated two-group RNA-Seq analysis workflow". In: *arXiv preprint arXiv:1905.03909* (2019) (Cited on page 124).

[218] Juan Manuel Gálvez, Daniel Castillo, Luis Javier Herrera, Belen San Roman, Olga Valenzuela, Francisco Manuel Ortuno, and Ignacio Rojas. "Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series". In: *PloS One* 13.5 (2018), e0196836 (Cited on page 125).

[219] S. González, D. Castillo, J.M. Galvez, I. Rojas, and L.J. Herrera. In: *International Work-Conference on Artificial Neural Networks*. cited by 2. 2019, pp. 883–894. DOI: 10.1007/978-3-030-20518-8_73 (Cited on pages 125, 153, 154).

[220] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". In: *Nucleic Acids Research* 38.6 (2009), pp. 1767–1771 (Cited on page 129).

[221] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079 (Cited on page 129).

[222] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods* 14.4 (2017), p. 417 (Cited on page 129).

[223] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34.5 (2016), p. 525 (Cited on page 129).

[224] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140 (Cited on page 129).

[225] Kasper D Hansen, Rafael A Irizarry, and Zhijin Wu. "Removing technical variability in RNA-seq data using conditional quantile normalization". In: *Biostatistics* 13.2 (2012), pp. 204–216 (Cited on page 129).

[226] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". In: *Bioinformatics* 28.6 (2012), pp. 882–883 (Cited on pages 130, 158).

[227] Verónica Bolón-Canedo, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benıtez, and Francisco Herrera. "A review of microarray datasets and applied feature selection methods". In: *Information Sciences* 282 (2014), pp. 111–135 (Cited on page 130).

[228] Tao Li, Chengliang Zhang, and Mitsunori Ogihara. "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression". In: *Bioinformatics* 20.15 (2004), pp. 2429–2437 (Cited on page 130).

[229] Huiqing Liu, Jinyan Li, and Limsoon Wong. "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns". In: *Genome Informatics* 13 (2002), pp. 51–60 (Cited on page 130).

[230] Chris Ding and Hanchuan Peng. "Minimum redundancy feature selection from microarray gene expression data". In: *Journal of Bioinformatics and Computational Biology* 3.02 (2005), pp. 185–205 (Cited on pages 130, 160).

[231] William S Noble. "What is a support vector machine?" In: *Nature Biotechnology* 24.12 (2006), p. 1565 (Cited on pages 131, 160).

[232] RM Parry, W Jones, TH Stokes, JH Phan, RA Moffitt, H Fang, L Shi, A Oberthuer, M Fischer, W Tong, et al. "k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction". In: *The Pharmacogenomics Journal* 10.4 (2010), p. 292 (Cited on pages 131, 160).

[233] Tin Kam Ho. "Random decision forests". In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282 (Cited on pages 131, 160).

[234] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. "Gene ontology: tool for the unification of biology". In: *Nature Genetics* 25.1 (2000), p. 25 (Cited on page 131).

[235] Gene Ontology Consortium. "The gene ontology resource: 20 years and still GOing strong". In: *Nucleic Acids Research* 47.D1 (2018), pp. D330–D338 (Cited on page 131).

[236] A Alexa and J Rahnenfuhrer. "topGO: Enrichment Analysis for Gene Ontology. R package version 2.36.0." In: *Bioconductor* (2019) (Cited on page 131).

[237] Weijun Luo and Cory Brouwer. "Pathview: an R/Bioconductor package for pathway-based data integration and visualization". In: *Bioinformatics* 29.14 (2013), pp. 1830–1831 (Cited on page 131).

[238] Minoru Kanehisa and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic Acids Research* 28.1 (2000), pp. 27–30 (Cited on page 132).

[239] Jean Fred Fontaine and Miguel A Andrade-Navarro. "Gene set to diseases (gs2d): Disease enrichment analysis on human gene sets with literature data". In: *Genomics and Computational Biology* 2.1 (2016), e33–e33 (Cited on page 132).

[240] Gautier Koscielny, Peter An, Denise Carvalho-Silva, Jennifer A Cham, Luca Fumis, Rippa Gasparyan, Samiul Hasan, Nikiforos Karamanis, Michael Maguire, Eliseo Papa, et al. "Open Targets: a platform for therapeutic target identification and validation". In: *Nucleic Acids Research* 45.D1 (2016), pp. D985–D994 (Cited on pages 132, 160).

[241] Jiangang Liu, Andrew Campen, Shuguang Huang, Sheng-Bin Peng, Xiang Ye, Mathew Palakal, A Keith Dunker, Yuni Xia, and Shuyu Li. "Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data". In: *BMC Medical Genomics* 1.1 (2008), p. 39 (Cited on page 154).

[242] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* 68.6 (2018), pp. 394–424 (Cited on page 154).

[243] G. Rosti, G. Bevilacqua, P. Bidoli, L. Portalone, A. Santo, and G. Genestreti. "Small cell lung cancer". In: *Annals of Oncology* 17.suppl$_2$ (April 2006), pp. ii5–ii10. ISSN: 0923-7534. DOI: 10.1093/annonc/mdj910. eprint: http://oup.prod.sis.lan/annonc/article-pdf/17/suppl_2/ii5/305111/mdj910.pdf. URL: https://doi.org/10.1093/annonc/mdj910 (Cited on page 154).

[244] Roy S Herbst, Daniel Morgensztern, and Chris Boshoff. "The biology and management of non-small cell lung cancer". In: *Nature* 553.7689 (2018), pp. 446–454 (Cited on page 155).

[245] Wendy A Cooper, David CL Lam, Sandra A O'Toole, John D Minna, Morgan R Davidson, Adi F Gazdar, Belinda E Clarke, Raymond U Osarogiagbon, Gail E Darling, Timothy G Whitsett, et al. "The textbook on Lung Cancer: time for personalized medicine". In: *Annals of Palliative Medicine* 4.2 (2015), p. 81 (Cited on page 155).

[246] Abel Sanchez-Palencia, Mercedes Gomez-Morales, Jose Antonio Gomez-Capilla, Vicente Pedraza, Laura Boyero, Rafael Rosell, and Mª Esther Fárez-Vidal. "Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer". In: *International Journal of Cancer* 129.2 (2011), pp. 355–364 (Cited on page 155).

[247] Nozomu Yanaihara, Natasha Caplen, Elise Bowman, Masahiro Seike, Kensuke Kumamoto, Ming Yi, Robert M Stephens, Aikou Okamoto, Jun Yokota, Tadao Tanaka, et al. "Unique microRNA molecular profiles in lung cancer diagnosis and prognosis". In: *Cancer Cell* 9.3 (2006), pp. 189–198 (Cited on page 155).

[248] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. "NCBI GEO: mining tens of millions of expression profiles—database and tools update". In: *Nucleic Acids Research* 35.suppl_1 (2006), pp. D760–D765 (Cited on page 156).

[249] David L Donoho et al. "High-dimensional data analysis: The curses and blessings of dimensionality". In: *AMS Math Challenges Lecture* 1.2000 (2000), p. 32 (Cited on page 160).

[250] Mohammad Hossin and MN Sulaiman. "A review on evaluation metrics for data classification evaluations". In: *International Journal of Data Mining & Knowledge Management Process* 5.2 (2015), p. 1 (Cited on page 161).

[251] Denise Carvalho-Silva, Andrea Pierleoni, Miguel Pignatelli, ChuangKee Ong, Luca Fumis, Nikiforos Karamanis, Miguel Carmona, Adam Faulconbridge, Andrew Hercules, Elaine McAuley, et al. "Open Targets Platform: new developments and updates two years on". In: *Nucleic Acids Research* 47.D1 (2018), pp. D1056–D1065 (Cited on page 161).

[252] Zirong Chen, Jian-Liang Li, Shuibin Lin, Chunxia Cao, Nicholas T Gimbrone, Rongqiang Yang, Dongtao A Fu, Miranda B Carper, Eric B Haura, Matthew B Schabath, et al. "cAMP/CREB-regulated LINC00473 marks LKB1-inactivated lung cancer and mediates tumor growth". In: *The Journal of Clinical Investigation* 126.6 (2016), pp. 2267–2279 (Cited on page 168).

[253] Cemile Dilara Savci-Heijink, Farhad Kosari, Marie-Christine Aubry, Bolette L Caron, Zhifu Sun, Ping Yang, and George Vasmatzis. "The role of desmoglein-3 in the diagnosis of squamous cell carcinoma of the lung". In: *The American Journal of Pathology* 174.5 (2009), pp. 1629–1637 (Cited on page 168).

[254] Friederike Saaber, Yuan Chen, Tiantian Cui, Linlin Yang, Masoud Mireskandari, and Iver Petersen. "Expression of desmogleins 1–3 and their clinical impacts on human lung cancer". In: *Pathology-Research and Practice* 211.3 (2015), pp. 208–213 (Cited on page 168).

[255] Feng Zhang, Xia Chen, Ke Wei, Daoming Liu, Xiaodong Xu, Xing Zhang, and Hong Shi. "Identification of key transcription factors associated with lung squamous cell carcinoma". In: *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* 23 (2017), p. 172 (Cited on page 168).

[256] Ben Liu, Jinli Qu, Fangxiu Xu, Yan Guo, Yu Wang, Herbert Yu, and Biyun Qian. "MiR-195 suppresses non-small cell lung cancer by targeting CHEK1". In: *Oncotarget* 6.11 (2015), p. 9445 (Cited on page 168).

[257] Rebecca A Kohnz, Melinda M Mulvihill, Jae Won Chang, Ku-Lung Hsu, Antonio Sorrentino, Benjamin F Cravatt, Sourav Bandyopadhyay, Andrei Goga, and Daniel K Nomura. "Activity-based protein profiling of oncogene-driven changes in metabolism reveals broad dysregulation of PAFAH1B2 and 1B3 in cancer". In: *ACS Chemical Biology* 10.7 (2015), pp. 1624–1630 (Cited on page 168).

[258] De-Hai Yu, Jin-Hui Li, Yi-Chun Wang, Jian-Guo Xu, Peng-Tao Pan, and Li Wang. "Serum anti-p53 antibody detection in carcinomas and the predictive values of serum p53 antibodies, carcino-embryonic antigen and carbohydrate antigen 12–5 in the neoadjuvant chemotherapy treatment for III stage non-small cell lung cancer patients". In: *Clinica Chimica Acta* 412.11-12 (2011), pp. 930–935 (Cited on page 168).

[259] R Salgia, D Harpole, Evan Pisick, Anthony Elias, Arthur T Skarin, et al. "Role of serum tumor markers CA 125 and CEA in non-small cell lung cancer." In: *Anticancer Research* 21.2B (2001), pp. 1241–1246 (Cited on page 168).

[260] R Salgia, D Harpole, Evan Pisick, Anthony Elias, Arthur T Skarin, et al. "Role of serum tumor markers CA 125 and CEA in non-small cell lung cancer." In: *Anticancer Research* 21.2B (2001), pp. 1241–1246 (Cited on page 168).

[261] Guoxin Mao, Liting Lv, Yifei Liu, Buyou Chen, Mei Li, Tingting Ni, Dunpeng Yang, Hongzhen Zhu, Qun Xue, and Runzhou Ni. "The expression levels and prognostic value of high temperature required A2 (HtrA2) in NSCLC". In: *Pathology-Research and Practice* 210.12 (2014), pp. 939–943 (Cited on page 168).

[262] Lingyu Li, Wei Song, Xu Yan, Ailing Li, Xiaoying Zhang, Wei Li, Xue Wen, Lei Zhou, Dehai Yu, Ji-Fan Hu, et al. "Friend leukemia virus integration 1 promotes tumorigenesis of small cell lung cancer cells by activating the miR-17-92 pathway". In: *Oncotarget* 8.26 (2017), p. 41975 (Cited on page 168).