Multi-omics integrated analysis of Metastatic Breast Cancer

Universitat Oberta de Catalunya

UNIVERSITAT DE BARCELONA

Núria Moragas Garcia, PhD

MU Bioinformàtica i Bioestadística

Àrea de treball 4

13/01/2023

FINAL WORKTABLE

Títol del treball:	Multi-omics integrated analysis of Metastatic Breast Cancer
Nom de l'autor:	Núria Moragas Garcia
Nom del consultor/a:	Helena Brunel Montaner
Nom del PRA:	David Merino Arranz
Data de lliurament (mm/aaaa):	01/2023
Titulació o programa:	Màster en Bioinformàtica i Bioestadística
Àrea del Treball Final:	Àrea 4, TFM- Bioinformàtica i Bioestadística
Idioma del treball:	Anglès
Paraules clau	Breast Cancer, metastasis, multi-omics

Resum del Treball

El càncer de mama (CM) és el tipus de càncer més freqüent i la primera causa de mort per càncer entre les dones (2020), i aproximadament el 35% de les dones diagnosticades amb CM invasiu desenvolupen metàstasi (MCM). Tot i amb els avenços, la taxa de supervivència a cinc anys amb un esdeveniment de MCM és del 29%, amb un temps de supervivència mitjà de 18-24 mesos. Actualment, el coneixement sobre els factors implicats en la progressió del CM cap a metàstasis encara és pobre i poc entesa.

Per tot això és fonamental millorar el coneixement sobre el perfil genètic i molecular del CM i la seva probabilitat de colonitzar altres òrgans, per poder predir amb precisió el risc de progressió de cada pacient i determinar el tractament adequat segons cada cas. Tenint en compte aquestes dades, aquest projecte s'ha centrat en l'ús de diverses tècniques òmiques per millorar el coneixement de les MCM, tenint com a objectiu principal definir un perfil multiòmic de MCM que diferenciï aquells casos de CM que progressaran a càncer metastàtic d'aquells que no ho faran.

Per aquest motiu, s'han analitzat i integrat 3 òmiques comparant mostres de tumor primari amb mostres de metàstasi. Aquestes anàlisis ha permès constatar diferències significatives a nivell d'expressió gènica, metilació i CNV, i determinar una llista de 10 gens implicats en processos clau en la metàstasi com l'adquisició de mobilitat de les cèl·lules tumorals i la inhibició de l'apoptosi.

Abstract

Breast cancer (BC) is the most frequently diagnosed cancer and the first cause of cancer death among women in 2022, and approximately 35% of women diagnosed with invasive BC

will develop metastasis (MBC). Even with all the advances, the five-year survival rate with an MBC event is 29%, having a median survival time of 18-24 months. Currently, the knowledge about the factors involved in breast cancer progression to a metastatic event is still poor.

Considering all this data about MBC it is crucial to improve the knowledge about the genetic and molecular profile of breast cancer and its probability to colonize other organs, to be capable of predicting with precision the progression risk of each patient, and to determine the appropriate treatment according to each case. Taking all this into consideration, this project has focused on using diverse omics techniques to improve the knowledge of MBC, having as the main objective to define an MBC multi-omics risk profile that characterizes those BC cases that will progress to metastatic cancer from the ones that will not.

For this reason, 3 omics have been analyzed and integrated comparing primary tumor samples against metastasis samples. These analyzes have made it possible to verify significant differences at the level of gene expression, methylation and CNV, and determine a list of 10 genes involved in key processes in metastasis such as the acquisition of tumor cell mobility and the inhibition of apoptosis.

INDEX

LIST OF TABLES and FIGURES

List of tables:

List of figures:

1. INTRODUCTION

1.1 Context and justification

Breast cancer (BC) is the most frequently diagnosed cancer and the first cause of cancer death among women in 2020 [1]. The advances, during the last decades, improve the detection, diagnosis, and treatments, which have greatly improved the global average survival [2]. Currently, the five-year survival rate after diagnosis is 99% for localized BC and 86% for Regional BC [3]. Nevertheless, and despite all the advances, the five-year survival rate for distant stage (with a metastatic event) is 29% [4], having a median survival time of 18-24 months [5].

There are two systems for breast cancer classification: molecular, which is based on the gene expression pattern of the tumors, and histological classification, which is based on the location, invasiveness, and histological appearance of the tumor. These classifications determine the prognosis and the type of treatment to apply to each patient [6] [7] [8]. Nonetheless, recent studies would indicate that this prediction and prognostic systems are insufficient, therefore the classification of breast cancer is a field in constant study and growth [9].

Albeit controversial, the most accepted molecular classification includes 4 subtypes: Luminal A, Luminal B, HER2 and Basal-like [6] [10]. The histological classification is divided into two classes, *in situ* and invasive carcinoma. As *in situ* carcinoma, invasive carcinomas are divided into two sub-types, the Invasive Ductal Carcinoma (IDC) and the Lobular Invasive Carcinoma (ILC).

IDC is the most common type of breast cancer, representing about 70-80% of the total number of cases [11], and is defined as a malignant proliferation of the epithelial cells inside the duct, with evolution and local stroma invasion through the duct wall [12]. Nevertheless, IDC is a very heterogeneous disease including tubular, mucinous, clear cell, and sebaceous carcinoma, with different clinical outcomes and histological and morphological phenotypes [13].

Metastasis Breast cancer (MBC) is the most advanced stage of the tumor (also known as stage IV) and is defined as the speeding of breast tumor cells to a secondary organ where a secondary tumor will grow. This secondary tumor is the result of a complex and dynamic cascade of steps. Emphasize, that approximately 35% of women diagnosed with invasive BC will develop a metastasis [14].

Taking into account all this data about metastasis it is crucial to improve the knowledge about the genetic and molecular profile of breast cancer and its probability to colonize other organs, to be capable of predicting with precision the progression risk of each patient, and to determine the appropriate treatment according to each case, as well as, to improve the treatment of BC so that it does not progress to metastasis.

High-throughput sequencing techniques are constantly growing, and have been essential in genomics, transcriptomics, and epigenetics. The combination of various types of omics data and its integration into multi-omics analyses is allowing a better study and interpretation of disease biology at multiple levels to improve prognosis, diagnosis, and treatments [15]. Cancer is a complex

disease and its hallmarks capabilities acquisition of the transition from normal to malignancy alterations are driven by molecular aberrations in the genome, epigenome, transcriptome, proteome, and metabolome of the cancer cell. Then it is obvious that multi-omics analyses would be needed (advantageous) to understand the cancer progression and develop new biomarkers and/or effective personalized therapies [16] [17] [18] [19]. The use of omics is wild extended in breast cancer investigation. One key application was its classification into molecular subtypes thanks to gene array technologies [6] [10], and multi-omics has confirmed it [20]. Recent studies using multi-omics analysis of invasive breast cancer have developed a machine learning model that can predict the patient's response to a determined therapy [21], which is a clear example of the potential of applying multi-omics and the direct effect it can have on patients' life.

Taking all this into consideration, this project will focus on using the integration of diverse omics techniques to improve the knowledge of MBC and define a predictive profile, that helps to choose an appropriate treatment for each patient and to develop new biomarkers and therapeutic opportunities.

1.2 Objectives

In this scenario, the principal objective of this project is:

1. **Define a multi-omics risk profile of MBC,** that characterizes those BC cases that will progress to metastatic cancer from the ones that will not. To achieve this overall objective, the following three sub-objectives are proposed:
    1.1. Analyze different omics through dedicated bioinformatics tools. (Transcriptomic, epigenetic, and genomic data analysis).
    1.2. Integrate the multi-omics data and characterize the MBC profile.
    1.3. Validate the MBC profile. Eventually, if the time allows it, an algorithm that will be capable of predicting the probability of a metastatic event on a diagnosed breast cancer through the multi-omics MBCs risk profile described in the main objective will be proposed.

1.3 Approach and method to follow

To achieve the purpose of this project, a linear work plan has been developed composed of several tasks described in section 1.4. The workflow is shown in Figure 1.

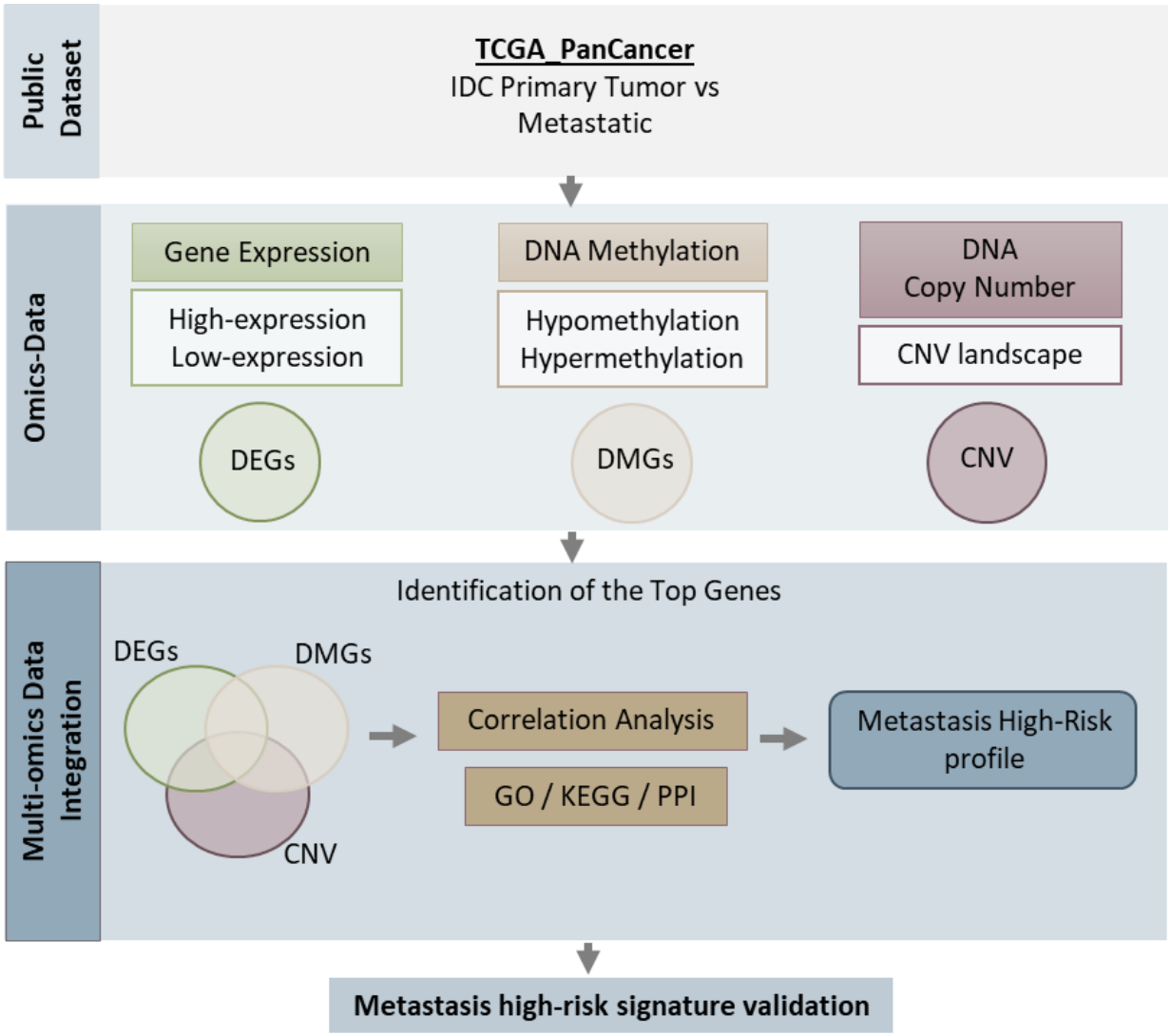


Figure 1. Workflow of multi-omics data analysis.

The following methods will be followed to complete these tasks. Most data processing and analysis will be done with R [22].

The first part was the **selection of a public dataset:** Bibliographic research has been done for public cohorts of Breast Cancer patients with several omics data available. TCGA_BRCA public cohort was selected, as that cohort contains several levels of omics data such as gene expression, DNA methylation, DNA copy number alterations, and clinic data (Table 1). The Metastatic Breast Cancer Project (February 2020) (MBCP), and the METABRIC [23] database will be used to validate the data obtained with TCGA_BRCA. After downloading the data, an analysis of the patient's cohort will be performed (age, type of cancer, survival, etc.).

REFERENCE	ORIGIN	DATA TYPE	SAMPLS to ANALYZE
TCGA-BRCA PanCancer Atlas [24]	cBioPortal	mRNA expression (RNA Seq V2 RSEM)	IDC simples: Primary tumor vs Metastatic
		Methylation (HM450)	
		Copy-number alterations	
		Clinical data	
The Metastatic Breast Cancer Project (February 2020)	cBioPortal	Clinical data	IDC samples.
METABRIC	cBioPortal	Clinical data	IDC samples

Table 1. Public data cohorts selected from the bibliography.

The following steps will be involved in the **analysis of the different omics.** Transcriptome, methylome and copy number alterations (CNA), from Primary tumor vs metastatic samples, will be analyzed first independently, and secondly, in an integrative analysis (Fig. 1). Before the analysis of the different omics, a preprocessing (Quality control and normalization) of the data will be performed.

RNA-seq data from TCGA-BRCA cohort will be used for differential expression analysis, and the differentially expressed genes (DEGs) obtained (upregulated or downregulated) from these analyses will be used to determine the genes related to the Metastatic progression.

Methylation profiling datasets will be used to detect differentially methylated regions (DMCs) and genes (DMGs). The objective is to detect the genes that are hypermethylated or hypomethylated.

The **copy number variations (CNV)** map will be downloaded from GDC and will be used to study what genes gains and deletions are associated with IDC (how many, where are they, and with which genes are associated). Secondly, CNA gene list will be used for the multi-omics data integration.

Following, tasks related to the **multi-omics data integration** will be done. Once gene expression, methylation, and CNA independent analysis have been completed, all these data will be integrated to determine an IDC Metastatic high-risk profile and then a deep analysis of these gene sets will be done analyzing pathways, networks, PPi interactions, and its relation with overall survival in DCIS patients. First, DEGs, MEGs, and CNA genes lists obtained in the individual analysis will be overlapped. Secondly, the triple intersected resulting list will be used for a person correlation analysis between the DNA methylation level and RNA expression. Third, genes with a significant negative correlation between DMGs and DEGs (hypermethylated lowly-expressed genes and hypomethylated highly expressed) will be chosen. Next, the top hypermethylated lowly-expressed and top hypomethylated highly expressed genes will be selected taking into consideration the data of the multi-omics integration and its function, pathways, and network study (GO, KEGG and PPI).

Lastly, The Metastatic Breast Cancer Project (MBCP), and METABRIC clinical information will be used to **validate the Metastatic high-risk profile**. The gene expression of the most robust genes (top Hypermethylation-low expression hub genes and top hypomethylation-high expression hub genes) will be validated in overall survival analyzed in IDC breast samples of MBCP, and METABRIC datasets.

Additionally, IDC Metastatic high-risk profile will be try integrated into a predictive model using machine learning. This point will be studied deeper if there is enough time.

1.4 Work Plan:

1.4.1 Task and Timeline

The tasks and milestones that must be done to complete the objectives are described in the Grantt diagram (Fig. 2). Briefly, the tasks are divided into four blocks (referred to de PECs deadline), the first bloc is the project planning (including the elaboration of this work plan). In the second block, the main project part will be done and includes the multi-omics data analysis, its integration, and validation. This block is divided into two PECs. The third bloc is the final manuscript redaction and the fourth block is the oral defense preparation.

0. WORK PLAN ELABORATION:

0.1. Bibliographic research
0.2. Public dataset research, and selection.
0.3. Planification and elaboration of PEC 1 Work Plan

PAC 1. Submit Work Plan

1. **TASKS CORRESPONDING TO THE FIRST SUB-OBJECTIVE:**

Analyze different omics through dedicated bioinformatic tools.

1.1. TASK 1: Determine a list of genes differentially expressed in MBC, through a transcriptomic data analysis.

a) RNA-seq raw data preprocessing and normalization
b) Differential Expression Analysis to determine differential gene expression (DEG) between primary tumor and metastatic samples.

1.2. TASK 2: Determine a list of differentially methylated sites and genes in MBC, through epigenetic data analysis.

a) DNA methylation raw data preprocessing and normalization
b) Differential Methylation CpG sites and genes identification between primary tumor and metastatic samples.

1.3. TASK 3: Determine a list of Copy Number Variations (CNV), through genomic data analysis.

a) Raw data preprocessing and normalization
b) Copy Number Variations (CNV) identification between primary tumor and metastatic samples.

PAC 2. Write and submit the first report.

2. TASKS CORRESPONDING TO THE SECOND SUB-OBJECTIVE:

Integrate the multi-omics data obtained in the first objective, and characterize the MBC profile.

2.1. **TASK 1:** Define a risk profile combining the obtained molecular markers in the transcriptomic, epigenomic, and genomic analysis.

2.2. **TASK 2:** Determine the differences between MBC and primary BC according to their profile.

2.3. **TASK 3:** Analyze biologically the biological value of the obtained markers through functional enrichment (GO, KEGG pathway, and PPI)

3. **TASK CORRESPONDING TO THE THIRD SUB-OBJECTIVE:**

Validation of the MBC risk profile is defined in the second objective**.**

3.1. **TASK 1:** Look at prognosis in TCGA and another patient's cohort.

3.2. **TASK 2:** Apply a prediction algorithm.

PAC 3. Write and submit a second report.

4. **MANUSCRIPT:**

4.1. **TASK 1:** Manuscript writing

PAC 4. Submit final manuscript.

5. **ORAL PROJECT DEFENSE:**

5.1. **TASK 1:** PowerPoint Preparation

PAC 5. Oral defense.

1.4.2 Risk analysis

Analysis of the problems that may arise during the development of this work, and their possible solutions.

- Problems in the breast cancer data set selection. There are a large number of breast cancer data sets, but generally, they are from primary tumor samples, or on the contrary only samples of metastasis. TCGA database has both types of samples, but the proportion between them is very large (200-400 primary tumor vs 4-6 metastasis samples, depending on the type of omics). In order to solve those problems, primary tumor samples can be filtered according to their stage (TNM Staging System). The numerical distance and variability within the primary tumor group would be reduced.
- Problems with computational resources. A large amount of data will be generated, which will require a large Ram memory. If the computer is not sufficient, access to a server will be requested or the scrip with the analysis will be sent to a third person.
- Problems with the timings. It is possible that some tasks take more time than planned, in this case, the sub-objective of developing a predictive algorithm will be eliminated.

- Problems in the bioinformatic tool selection. General R packages (that can be used in multiple types of data, like limma) will be used. However, if the analysis cannot be performed with these tools, it will be substituted by specific TCGA packages or by already developed application.

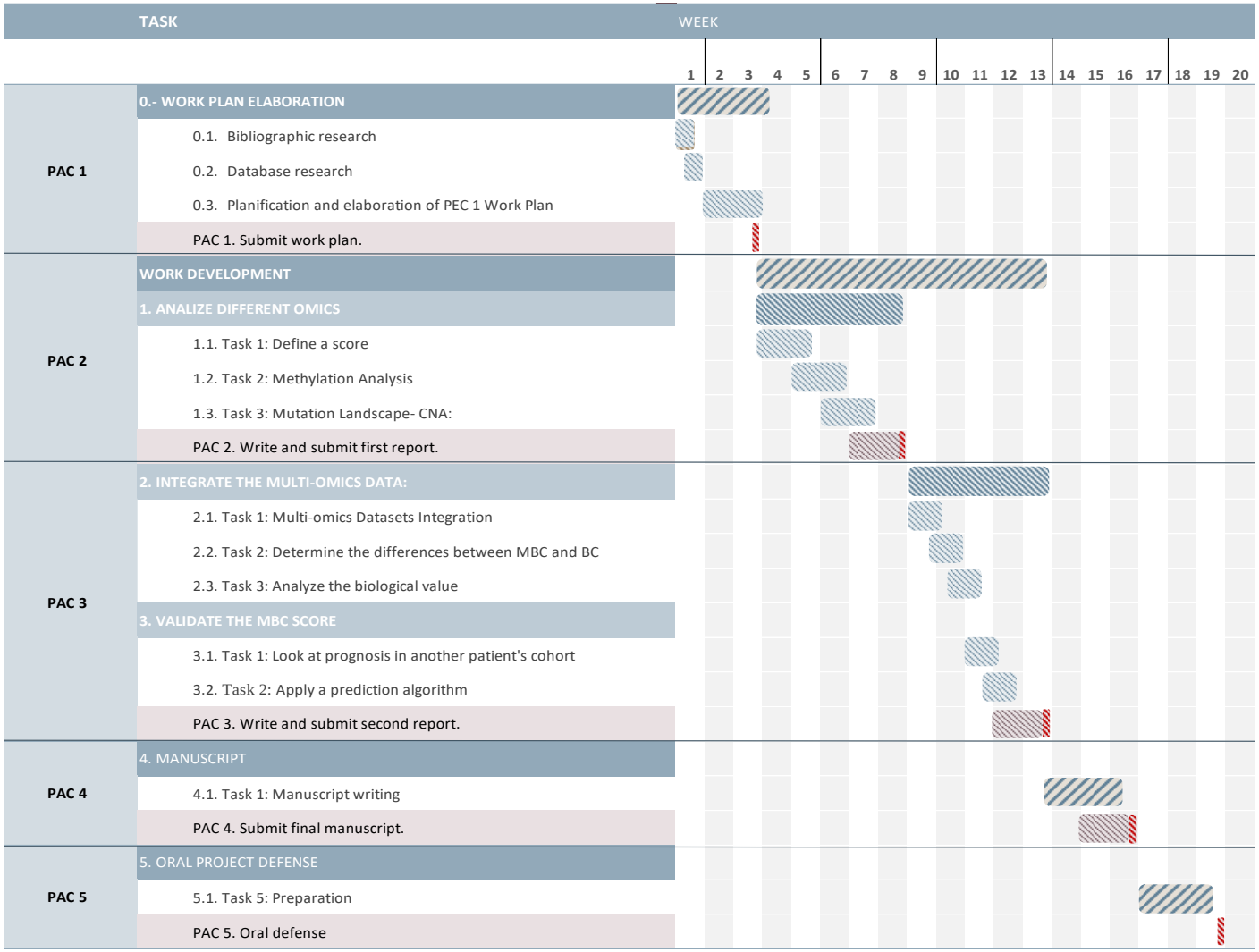


Figure 2. Gantt Diagram with the working plan defined.

1.5 Impact on sustainability, ethical-social, and diversity

No gender distinctions have been made when searching for bibliographic references. Despite the analysis carried out, only the samples from women have been taken into account, therefore the results obtained are only applicable to them, the reason is that breast cancer is the majority among this population and a very minority among men.

On the other hand, if it is possible to determine a profile of tumor biomarkers capable of predicting metastasis in breast cancer, it would have a great social impact, allowing for improved diagnosis, prognosis, and treatment of patients with this type of cancer.

The results of this TFM have no positive or negative impact on environmental, sustainability, and/or ecological footprint aspects. Memory's sections brief description

Below is a brief description of the section that will be part of the final report to f this TFM.

- **Section 2:** State of the art. This section will be revised on breast cancer, omics techniques, and their integration.
- **Section 3:** Materials and methods. The methodology used during the multi-omics analysis and its integration will be defined in this chapter. Mainly, it will describe the R tools and packages used.
- **Section 4:** Results. This section will be the main part of the TFM and will consist of the results obtained according to the objectives and tasks.
- **Section 5:** Discussion. The results obtained in the analysis will be compared and contrasted with the bibliography.
- **Section 6:** Conclusions. Conclusions will be drawn according to the results obtained and the bibliographic review. In addition, the difficulties and limitations of the present work will be exposed.
- **Section 7:** Bibliography.
- **Section 8:** Supplementary material.

2. STATE-OF-THE-ART

As mentioned in the introduction, breast cancer is the most common type of cancer among women, and metastasis is the main cause of death for diagnosed women, representing 90% of breast cancer deaths [25] [26]. Despite these overwhelming data, today there are no tools that distinguish those breast cancers with metastatic potential from those that do not. A large majority of omics studies focus on the study of primary tumors, and few on their metastases, and those that exist have focused on defining only mutations between primary tumors vs. metastatic samples [27] [28] [29]. The differences found in these studies cannot fully explain the heterogeneity in the evolution of breast cancer [30], nor do they present a clear translational goal, which is a clear indication of the need to expand the studies aimed at the deeper characterization of metastasis. In this context, it should be noted that recent studies have shown that tumors that presented similar genetic profiles, with the use of new technologies and the use of multi-omics approaches, are very different. It has been shown that studies involving the integration of multiple omics improve specificity and sensitivity, increasing the potential for the discovery of new biomarkers, prognostic factors, and therapies.

Additionally, the increase in new sequencing technologies and the cost reduction have led to an increase in databases that, beyond gene expression, also offer data from other omics such as methylation, metagenomics, CNV, SNPs among others. What databases exist for breast cancer and metastasis will be discussed later.

1.1 Omics integration

The continuous and fast advance of the high-throughput sequencing techniques and informatic tools is allowing the study of biological systems at multiple layers, which can also be called omics and which include RNA expression (transcriptomics), methylation profiles, chromatin remodeling, chromosomal conformation (epigenomics), DNA sequence (genomics), metabolites levels (metabolomic), protein profile expression (proteomics), the interaction between molecules (interactomics), etc [31] [32] [33] [34] [35].

In the same way, some omics study external components as metagenomics, which studies microorganisms that live in a specific niche inside another organism (Ex. sequencing the gut microbiome), or toxicogenomic (or pharmacogenomics), which studies how a determined drug affects an organism [32] [31].

Each of the omics described can identify key elements in a phenotype of interest or a certain disease, however, their study separately does not consider the complexity of biological systems. All the data generated by these omics can be combined, in a process known as data integration in a multi-omics study. Data integration allows a better understanding of organisms at different levels, as well as diseases such as cancer [2] [1].

Objectives that can be achieved with the integration of multi-omic data are detailed below.

- Classify and determine disease subtypes.

- Build prediction/risk models.
- Identify diagnostic, prognostic, and driver genes biomarkers candidates.
- Deriving insight into disease biology. Deciphering the molecular mechanisms and the interactions between them (Pathways and networks).

1.1.1 Integration approaches and methods

Currently, there are many methods and approaches for integrating multiple omics data, which can also be classified in several ways.

Two different approaches exist for the integration of multi-omics layers:

- **Simultaneous integration** (or Parallel integration): Use all the omics data at the same time in a single modeling step, without previous single-omics analysis. It considers the complementary data associated with each omic and the correlation between omics. Only applicable to samples that come from the same individuals/cohort.
- **Step-wise integration** (or sequential integration): consists of several steps. The first is single omics analysis or in a specific combination, and the second step implies integration. Applicable to samples from different cohorts. The integration of multi-omics layers between different cohorts limited the integration methods that it can use.

On the other hand, there are several methods of integration, here it's have been divided according to mathematical methods (as *Cavill et al* do) [36]. Additionally, some of the tools available to perform the integration are detailed in Table 2, some are exclusive to each method others can perform the integration by applying different methods [32] [37] [38] [39]. Moreover, these tools could also be classified according to the types of omics they accept.

1- **Conceptual integration.** First, the individual analysis of each omics is carried out. Second, a relationship is sought between the results obtained.

2- **Statistical integration**. It consists of looking for statistical relationships. They can be divided into four subgroups. Add that these methods are not mutually exclusive and that they can be used in the same study.
   a. **Correlation.** Look for correlation between elements of two omics, applying statistical methods such as Person and Spearman correlation, Goodman's range test, robust linear models, and partial correlations.
   b. **Concatenation.** Join the data from the different omics (concatenate), and then analyze the integrated data as a whole. Self-Organizing Maps [40] [41], K-means cluster analysis [42], or random forest [43] are standard techniques for concatenation data integration.
   c. Multivariate analysis. Simultaneous observation and analysis of more than one statistical variable. It allows analysis of more than three variables and is used in those sets of data that present high collinearity (similar profiles). The principal component analysis (PCA) [44]

and the partial least squares (PLS) [45] methods are the most used techniques in multivariate analysis.

d. **Pathways-base integration.** Based on the use of pre-existing biological data in the literature, i.e. the use of pathways and networks databases. This approach allows the determination and visualization of complex interactions between components such as genes, proteins, etc. Also, can be defined as a Knowledge-based approach.

3- **Machine learning techniques.** Methods that allow the integration of omic data automatically through the use of algorithms that can recognize patterns and use them to make predictions [48]. Machine learning models are divided into 3 methods: Supervised, Unsupervised, and Reinforcement learning [46].

Table 2. Summary of tools for the integration of multi-omics data.

Tool	Method	Application	Omics supported	Objective
CNAmet	Correlation	R	CNV, DNA methylation, gene expression (numerical and categorical)	Disease subtyping, Biomarker prediction
PFA	Concatenation	MATLAB	DNA methylation, miRNA, gene and protein expression (numerical)	Disease subtyping.
SNF	Concatenation Pathways-base integration	R/ MATLAB	DNA methylation, miRNA, gene and protein expression (numerical)	Disease subtyping.
MetaGeneAlyse	Concatenation	Web tool	Gene expresion, metabolite data	Disease subtyping.
PSDF	Concatenation	MATLAB	CNV, gene expression (Categorical)	Disease subtyping.
moCluster	Multivariate análisis	R	Multi-omics (numerical)	Disease subtyping.
MFA	Multivariate análisis	R	Multi-omics (numerical and categorical)	Disease subtyping.
rMKL-LPP	Multivariate análisis	Web tool	Multi-omics (numerical)	Disease subtyping.
iNMF	Multivariate análisis	Python	Multi-omics (numerical)	
FSMKL	Multivariate análisis Machine learning	MATLAB	Multi-omics (numerical and categorical)	Biomarker prediction
PMA	Multivariate análisis	R	Multi-omics (numerical and categorical)	Biomarker prediction
sMBPLS	Multivariate análisis	MATLAB	Multi-omics (numerical)	Desease insight
T-SVD	Multivariate análisis	R	Multi-omics (numerical)	Desease insight
mixOmics	Multivariate análisis	R	Multi-omics (numerical and categorical)	Disease subtyping and biomarker prediction
MCIA	Multivariate análisis	R	Multi-omics (numerical)	Disease subtyping and disease insight

iCluster		R	CNV, DNA methylation, gene expression (numerical)	Disease subtyping and Biomarker prediction
MethylMix	Multivariate análisis	R	Gene expresion and DNA methylation	
NetICS	Pathways-base integration	MATLAB	Multi-omics (numerical and categorical)	Biomarker predictor
PARADIGM	Pathways-base integration	Python	Multi-omics (numerical)	Disease subtyping and insight
Cytoscape	Pathways-base integration	Software	Multi-omics (numerical and categorical)	Biomarker predictor, disease insight
ConsensusPathDB	Pathways-base integration	Web tool	Protein and gene expression, metabolism data, gene regulation, and drug-target.	Biomarker predictor, disease insight
IMPaLA	Pathways-base integration	Web tool	Gene and protein expression and metabolomics	Biomarker predictor, disease insight, Disease subtyping
MetaCore	Pathways-base integration	Web tool	Gene and protein expresion, siRNA, microRNA	Disease insight
Ingenuity IPA	Pathways-base integration	Software	Gene expression, miRNA, SNP data	Biomarker predictor, disease insight
Paintomics	Pathways-base integration	Web tool	Gene and protein expression, metabolomics, miRNA	Biomarker predictor, desease insight
InCroMAP	Pathways-base integration	Software	Gene and protein expression, miRNA and DNA methylation,	Biomarker predictor, disease insight
INMEX	Pathways-base integration	Web tool	Gene expression and metabolomics	Biomarker predictor, disease insight

The 3 omics used in the present study are described in more detailed below, as well as a brief description of the techniques, methods, tools, and datasets that are available for their identification, study, and integration.

1.1.2 Genomics

Genomics is the study of whole DNA sequences (genome – WGS) of an organism, including the coding regions (<2%) and the non-coding regions (>98%) [47]. Looking inside an individual's genome can allow identifying mutations and/or variations that discriminate between health and disease. The variants that can be distinguished in a genome are [48]:

- Single Nucleotides Variations, single nucleotide changes in DNA sequence in contrast with the reference sequence. Common variations (frequency in the population >1%) are called Single Nucleotide Polymorphisms (SNPs) [49] [50] and are the more abundant class of genetic variants.

- Small/short insertions and deletions (indels) of more than one nucleotide. Indels are the second most abundant class of genetic variants.
- Structural variations (SVs) are large genomic alterations in chromosome structures (> 1kb) [39], they can be classified into unbalanced variants (Deletions, duplications, and insertions) and balanced variants [40] (Inversions and translocations). Balanced variants do not alter the copy number of a determined DNA sample vs reference, conversely the unbalanced variants yes, for this reason also are called Copy Number Variations (CNVs).
    - **Copy Number Variations (CNVs)** are an abnormal number of copies of a specific segment of DNA (between 1 kb and 5 Mb) in a reference genome. Deletions and duplications are widely extended in the human genome with a frequency between 4.8 and 9.5% [51] [52].

In the present TFM only CNV data from SNP-array has been used, for this reason, the next sections are related to this technique.

- **Technologies for the CNV detection**

Since 2003 the most used techniques for CNV detection were SNP-array and array-based comparative genomic hybridization (arrayCGH). The bases of SNP-array are the same as the DNA microarray [53].

Microarray-based techniques: Microarray has its origin in the Southern blot technique, which only allowed gene study one by one. In 1995 in Science journal was describe the microarray as it is known today, which allows the study of diverse genes at a time, a fact that meant a change of paradigm and era in science [54]. The basis of microarray consists of the hybridization (pairing of the complementary bases) between the samples of interest attached to a fluorescent dye (target), and a collection of known DNA fragments attached to a surface chip (probes). The hybridization between targets and probes is measurably applying a laser light, the emitted fluorescence is proportional to the quantity of DNA in the sample of interest. In SNP-array the probes are allele-specific oligonucleotide (ASO), and the target nucleic acid sequences also are labeled with fluorescent dye [55] [53].

SNP-array technique has some handicaps like limited coverage for genome, low resolution, and difficulty in detecting novel and rare mutations. Most recently high-throughput sequencing has merged as a potent substitute for microarray techniques [56].

Sequencing-based techniques: Next-generation sequencing history is long, and it starts with the discovery of the double-helix structure of DNA in 1953, after this discovery the race for sequencing began and several attempts were made. The first-generation sequencing method was developed by Frederick Sanger, who sequenced the first complete DNA genome of a bacteriophage in 1977 [57] [58], and was the bases for the Human Genome Project. The complete human genome sequencing, achieved in 2003, took 13 years and 3 billion USD [59] [60].

Sanger sequencing (or first-generation sequencing) is based on the random base-by-base (dNTPs) incorporation by a DNA polymerase in a single-stranded DNA template during in vitro DNA replication [58]. It was the main sequencing method used for 40 years, but recently it has been replaced by the next generation sequencing (NGS).

The NGS (or Second-generation sequencing), was developed 15 years ago and has become an essential tool in molecular biology since it allows for sequencing DNA and RNA much more cheaply and quickly than Sanger sequencing [61] [62]. It is based on DNA segmentation into several fragments, that are sequenced with 10-30 million reads per sample. The last step is computational, the sequenced reads are aligned to a reference sequence (genome or transcriptome) [63]. Although this technology has existed for a decade, it is still revolutionizing science, and it is expected that it will continue to do in the next years [64]. Some of the most used NGS platforms are Illumina, IonTOrrent, and BGI/MGI among others [65].

1.1.3 Transcriptomics

Transcriptomics is the study of all the RNA molecules (also called transcripts), including coding RNA (<4%) which is transduced into proteins, and non-coding (>95%) RNAs, in an organism, cell, or tissue in a concrete situation (example in disease), based in its gene expression profiles. Coding RNA is known as messenger RNA (mRNA)), and non-coding RNA includes ribosomic RNA (rRNA), transfer RNA (tRNA), micro-RNA (miRNA), and non-coding RNA [66] [67]**.**

The mRNA study allows, mainly, to know which part of the genome is transcribed and therefore which genes are expressed. Indirectly, also allows the study of post-transcriptional processes such as alternative splicing and the prediction of protein isoforms [68].

*In the present TFM only RNA data from RNA-sequencing, has been **analyzed and** integrated**,** for this reason, the next sections are related to this technique.*

Currently, transcriptomes are obtained by the same DNA detection techniques: microarray and sequencing, described in the previous section. Gene expression analysis using mRNA by DNA microarray or sequencing required a previous step, mRNA is extracted and converted to complementary DNA (cDNA), and then hybridized to microarray or sequenced.

1.1.4 Epigenomics

Epigenomics is the study of all the epigenetic changes in the genetic material in a cell or organism (epigenome). Understanding epigenetics as the study of gene expression regulation (when and how a determined gene is turned on or off) through chromatin remodeling, DNA packaging, histone modification, and DNA methylation [69] [70].

DNA methylation consists of the addition of a methyl group (CH_3) to carbon 5 of the cytokine residues of the DNA by the methyltransferase enzyme. The addition of this methyl group leads gene expression regulation by inhibiting the binding of transcription factors or by the recruitment

of proteins involved in gene repression. Methylation is a reversible modification mediated by demethylase enzymes, and it is estimated that 3% of the total DNA in humans is methylated [71].

Most methylations take place in cytokines that are followed by a guanine, these areas are known as CpG sites, non-CpG methylation sites are present in human embryonic cells, but this kind of methylation is lost in mature tissues. In addition, it is important to mention that the location of the CpG sites in the genes (intergenic regions, CpG islands, gene body) is important and has different effects on genes, it will go into more detail in the results section [72].

In the present TFM only methylation data from bisulfite sequencing has been analyzed and integrated, for this reason, the next sections are related to this technique.

- **Identifying DNA methylation - technologies**

There are several techniques to identify DNA methylation, it can be divided into three categories depending on the aim of the study. To determine the methylation status of a specific gene of interest, there are bead array, PCR, and pyrosequencing, among others. On the other hand, if the goal is to know the general methylation status of the entire genome there are high-performance liquid chromatography-ultraviolet (HPLC-UV), Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS), ELISA-based methods, etc. Lastly, for the identification of differential methylation: there are array or bead hybridization and Bisulfite Sequencing [73].

The last one, the Bisulfite Sequencing technique, is the "gold-standard" technology for the detection of DNA methylation, and consist of treating the samples with sodium bisulfite, which causes unmethylated cytosine residues to be converted to uracil, while methylated ones remain unchanged. Next, uracils are recognized as thymines after successive PCR amplification and sequencing [74].

1.2 Breast Cancer database

In the last decade and with the increase in the use of techniques such as sequencing, a large amount of data from different omics has been generated. Some of them are of public access, others with restrictions on demand. These databases contain sensitive information regulated under ethical and legal norms.

Some repositories are specific to one type of omic data, others to specific diseases with data from one or multiple omics as well as epidemiological data, while others are large consortia that combine data from all types of experiments (Multi-omics data repositories) [37].

Below are some examples related to metastatic breast cancer databases:

- **The Cancer Genome Atlas (TCGA)**
  https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
  TCGA is one of the largest multi-omic data repositories that exists, it consists of more than 20,000 samples of primary tumors from 33 different types of cancer, among them breast cancer, and associated with some of these samples there are blood, normal tissue and

metastasis (among others). The omic data you can find in it is multiple from CNV, mutations, metabolomics, RNA-seq, DNA-seq, miRNA, clinical information, and histological data.

- **Molecular Taxonomy of Breast Cancer International Consortium (METABRIC).** https://ega-archive.org/studies/EGAS00000000083
  METABRIC is an exclusive database of breast cancer, containing genomic, transcriptomic, promoter methylation (RRBS), and clinical data from more than 2000 primary tumor samples and 548 matched normal samples collected in the UK and Canada. It is the most extensive breast cancer database with different omics.

- **Gene Expression Omnibus (GEO).** https://www.ncbi.nlm.nih.gov/geo/
  GEO is a general public repository of omics data managed by the National Center for Biotechnology Information (NCBI), and gene expression, methylation, and DNA data from arrays and sequencing submitted by the research community.

- **The Metastatic Breast Cancer Project.** www.mbcproject.org
  Ongoing dataset with 379 primary and metastatic samples from 301 patients with breast cancer. This study includes whole-exome sequencing (Mutations, CNV, and structural variants), RNA-sequencing, and clinical data.

- **UK Biobank.** https://www.ukbiobank.ac.uk/
  UK Biobank contains more than half a million genetic and epidemiological data (lifestyle and diet). The data was collected in a prospective study that was carried out between 2006 and 2010, in total they have more than 10,000 samples from patients who had breast cancer when the study starts or who developed one during it. It is mainly aimed at the study of risk factors related to cancer and lifestyle.

- **cBioPortal.** https://www.cbioportal.org/
  cBioPortal is a cancer-specific repository that offers different services. First it is a gateway to many types of databases with a very simple finder, including TCGA and METABRIC databases. Second, it is a web tool that performs analysis with the data chosen and visualized in a very simple way. The data are from multiple projects and many types of cancer, for example, there are 27 different databases of breast cancer. It should be noted that there is an R package that facilitates downloading and analyzing the data from this portal (cBioPortalData – Bioconductor).

Of all these studies, TCGA dataset was selected to develop the present work, because TCGA is the only one that met the requirements of having data from 3 omics, and clinical survival data in primary tumor and metastasis samples.

3. MATERIALS AND METHODS

3.1 Breast Cancer dataset – sample selection

The breast cancer dataset used and analyzed in this work has been downloaded from the Cancer Genome Atlas Program (https:://cancergenome.nih.gov/) (TCGA-BRCA database) through the R package TCGAbiolinks (GDCquery, GDCdownload, and GDCprepare). TCGA-BRCA was composed of 1098 cases with a primary solid tumor, blood, normal tissue, and metastatic samples among others, and contains RNA-seq, DNA methylation, mutation, CNV, and clinical data.

Before downloading, **groups selection** was performed, the eligibility criteria include:

- **Sample type** (sample_type): primary solid tumors and metastasis sample
- **Samples subtype** (primary_diagnosis): only the Invasive Ductal Carcinoma (IDC) subtype
- **Pathological stat** (ajcc_pathologic_m): In Primary Tumor samples, only the pathologic M state equal to M0 was included to eliminate the Primary Tumor samples with a metastasis event. Metastatic samples do not include this data.
- **Stage** (ajcc_pathologic_stage): In Primary Tumor samples, only the tumor stage equal to I (I, IA, IB) and II (II, IIA, IIB) samples were included. Metastasis samples do not include this data.
- **Gender** (gender): only women have been included in this study.

Selection criteria have been applied to have two well-differentiated groups between primary and metastatic tumors. The barcodes of the resulting samples were saved for the subsequent download of the data from the different omics. A summary of downloaded data:

Table 3. Summary of downloaded data according to omics.

Type	Method	Nº Primary Tumor	Nº Metastasis
TCGA Biospecimen	N.A	2171	8
Gene Expression	Illumine HiSeq 2000 RNA-seq	529	6
DNA Methylation	Illumina Human Methylation 450 (HM450)	319	4
CNV	Affymetrix Genome-wide Human SNP array 6.0	781	6

The TCGA data are classified according to 4 levels of processing. Level 1: Raw data obtained from arrays or sequencing. Level 2: Processed data. Level 3: Processed and segmented data. Level 4: processed and annotated data. The downloaded levels of each omic are detailed below in the corresponding section.

The Metastatic Breast Cancer Project (MBCP) (Provisional, February 2020) and METABRIC datasets have been downloaded from cBioPortal and used as validation datasets.

3.2 Statistical analysis and code availability

R software (version 4.2.1) (https://www.r-project.org/) has been used for all the statistical analyses. The details of each experiment (statistical analysis and significance) are detailed in the corresponding Materials and Methods section and figure legends.

Detailed R scripts used in this study are available in GitHub repository at https://github.com/nmoragas/TFM_UOC.

3.3 RNA expression data analysis

TCGA-BRCA mRNA gene expression data measured by the Illumina HiSeq 2000 RNA sequencing have been obtained by *TCGAbiolinks* [75] R package using the following parameters:

- project = "TCGA-BRCA",
- barcode = *barcodes after group selection,*
- data.category = "Transcriptome Profiling",
- data.type = "Gene Expression Quantification",
- experimental.strategy = "RNA-Seq",
- workflow.type = "STAR - Counts",
- sample.type = c("Primary Tumor", "Metastatic"))

RNA-seq data download includes gene expression row counts (number of reads overlapping a given gene and sample without normalization), of 529 primary tumor and 6 metastasis samples.

Row data preprocessing includes the transformation to counts per million (CPM), filtering (counts >10), creation of a DGEList, and normalization using the EdgeR packages. Data normalization has been done, first using calcNormFactors R function, which calculates scaling factors to convert raw library sizes into effective library sizes, and then has been performing the scale normalization by voom function. Supplementary Figure 1A-B shows the distribution pre-normalized and normalized, and the trend of the mean-variance (log-cpm) before and after fitting to a linear model using voom [76], a step necessary for the differential gene expression analysis with limma (Supplementary Figure 1C-D). These data do not present missing values.

Differential Gene Expression (DEGs) analyses have been performed using limma [77] and EdgeR [78] packages. A gene has been considered significantly differentially expressed with a p-value < 0.05. DEGs obtained (upregulated or downregulated) from these analyses have been listed and plotted using ggplot2 package [79] [80] [81].

3.4 DNA methylation data analysis

TCGA-BRCA DNA methylation data measured by Illumina Human Methylation 450 (HM450) platform have been obtained with the *TCGAbiolinks* R package using the following parameters:

- project = "TCGA-BRCA",
- barcode = *barcodes after group selection*,
- data.category = "DNA methylation",
- data.type = "Methylation beta value",
- platform = "Illumina Human Methylation 450",
- sample.type = c("Primary Tumor", "Metastatic"))

DNA methylation data are level 3, which includes normalized beta values (methylation at known CpG sites) per probe (Probe IDs) of 319 primary tumor samples and 4 metastasis samples, with a total of 485577 CpG methylated sites.

Data preprocessing steps include deletion of missing values (N/A), probes containing SNPs (overlap), and probes that have been demonstrated to map multiple places in the genome [82], after preprocessing, 333315 methylation sites have remained. As a control, a density plot of β-values of the primary tumor and metastasis samples was made, which shows a similar pattern between them (Supplementary Figure 2).

The *EdgeR* package was used to determine the **differential methylated CpGs (DMCs)**. CpG sites have been considered significantly differentially methylated (hypermethylated or hypomethylated) with a p-value < 0.05. DMCs obtained from these analyses have been listed and plotted using the *ggplot2* package [83] [84] [85].

DMCs have been annotated with the genes to which they belong with the *IlluminaHumanMethylation450kanno.ilmn12.hg19* package [86].

3.5 Copy Number Variation data analysis

Copy Number Variations (CNVs) data measured by Affymetrix Genome-Wide Human SNP Array 6.0) have been obtained by TCGAbiolinks R package using the following parameters:

- project = "TCGA-BRCA",
- data. Category = barcodes after group selection,
- data.category = "Copy Number Variation",
- data.type = "Copy Number Segment",
- sample.type = c("Primary Tumor") or c("Metastatic"))

CNV data are level 3 and includes data from 781 primary tumors and 6 metastasis samples. First, CNV data have been checked, it did not contain missing values and a density plot have been made to see the mean segment of the samples, both types of samples have a correct distribution with the highest peak at 0 (Supplementary Figure 3). Ggplot2 package has been used to plot primary tumor and metastatic sample counts in an exploratory analysis.

Next, the *GAIA* (Genomic Analysis of Important Aberrations) package has been used to identify recurrent CNV only in metastatic samples [87]. This package first identifies deletions and amplifications with a corrected p-value <0.0001 comparing a sample matrix with a genomic probes matrix (SNP6 GRCh38 Liftover Probeset download from GDC website), and secondly annotates the CNVs to the corresponding gene using the *biomaRt* package [88].

3.6 Omics data integration

Once gene expression, methylation, and CNA independent analysis have been completed, all these data have been integrated to determine an IDC Metastatic gene set and then a deep analysis of these gene sets has been done analyzing pathways and networks where they are involved and the relationship between them.

Data integration has been done in several steps. First, DEGs, MEGs, and CNA selected in the individual analysis have been conceptually integrated through their overlap using *ggvenn* function based on *ggplot2* package [89]. Secondly, the triple intersected genes result of these overlapping (the gene set) have been analyzed in depth from several points of view: analyzing the relationship between genes and their methylation (methylation sites and correlation with gene expression), as well as their joint involvement in pathways and networks (Described in next section 3.7).

Correlation between gene expression and methylation has been calculated using a Pearson correlation ($r > 0.3$ and p-value < 0.05) analysis between the gene set normalized expression and the β values of CpGs, using a specific R function designed by Hamid Ghaedi [90], and *ggpubr* and *ggscatter* packages.

3.7 Functional Enrichment Analysis (GSEA - KEGG)

Different enrichment analyses have been performed using the following gene-set libraries: Ontology Gene Set (C5), Oncogenic Signature Gene set (C6), Hallmark Gene Set (H), downloaded directly from the MSigDB (The Molecular Signatures Database) database [91]. The defined metastatic gene set has been divided into up and down-regulated gene expression. Then, in each of the two groups created, has been studied which processes from the gene-set libraries are enriched using *enricher* R package (Bioconductor) with a p-value < 0.05.

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) database has been used to perform pathway enrichment analysis. The metastatic gene set has been divided into up and down-regulated gene expression and it has been studied using *enrichKEGG* R package (Bioconductor) with a p-value < 0.05.

The results have been plotted with *dotplot* function of *lattice* package (only the first 10 pathways results have been plotted).

To predict and analyze the functional network between the enriched pathways determined in the previous step, EnrichmentMap App [92] has been used and visualized in Cytoscape software (FDR q-value < 0.05).

3.8 Top Gens validation in MBCP and METABRIC database

To validate the top gens obtained from the data integration, the Metastatic Breast Cancer Project (MBCP) and the METABRIC database have been used. Clinical and gene expression data have been downloaded from cBioPortal web. In both cases and prior to the analysis data selection

have been made, only IDC samples has been selected, duplicate gene have been removed, and expression data have been transformed to z-score. MBCP data have been used to analyses top top gen relation to the clinical features: metastasis status and time to metastasis, using the Wilcoxon text with p-value < 0.05 and differences have been plotted with ggplot2 R package. METABRIC dataset have been used to analyses the top gene expression relation to the Overall Survival (OS) and to the Relapse Free Survival (RFS), the differences have been indicated with Kaplan-Maier plots using the *survival* and *survminer* R package, p-value <0.05.

4. RESULTS

The breast cancer dataset from TCGA (TCGA-BRCA) is a large cohort that includes data from diverse sample types and diverse molecular information. The current study has analyzed only the mRNA, DNA methylation, and copy number alteration (CNA) data from Invasive Breast Cancer (IDC), comparing metastatic (TM) vs Primary tumor (TP) samples. Taking into consideration the eligibility criteria it has been identified a different number of Primary Tumor and Metastatic samples depending on the molecular data. These patients and molecular samples are listed in Table 3. Data from the three omics were first analyzed separately and then in an integrative way.

4.1 Identification of differentially expressed genes in metastatic samples

After pre-processing RNA-seq data as mentioned in the Material and Methods section (3.3 RNA expression data analysis) (results in Supplementary Figure 1), an exploratory analysis was performed. The counts were transformed to log2 and analyzed by Principal Component Analysis (PCA) sorted by sample type, to look at whether the samples of metastases presented a distinctive pattern. PCA plot shows that the metastatic gene expression pattern does not present differences from primary tumors (Figure 3A), however, a differential profile has been seen on the left side of the plot, which corresponds to the basal subtype (Figure 3B).

Differential expression analysis was assessed with voom-limma method between read counts from primary and metastatic RNA-seq samples. A total of 1689 DEGs have been obtained, 869 down-regulated and 820 up-regulated in metastatic samples compared to primary samples, volcano plot in Fig 1C shows the expression for every gene in the comparison, and Table 4 the top 20 genes ranked by p-value (p-value >0.05). The top five most significantly up-regulated genes in metastatic samples are C7, CD22, MST1, GRIK5, and F10, all protein-coding genes. The top five down-regulated genes are AC005291.2, TBX5-AS1, SFRP2, C3orf80, and TBX5, the first two are lncRNA genes, and the other three are protein-coding genes.

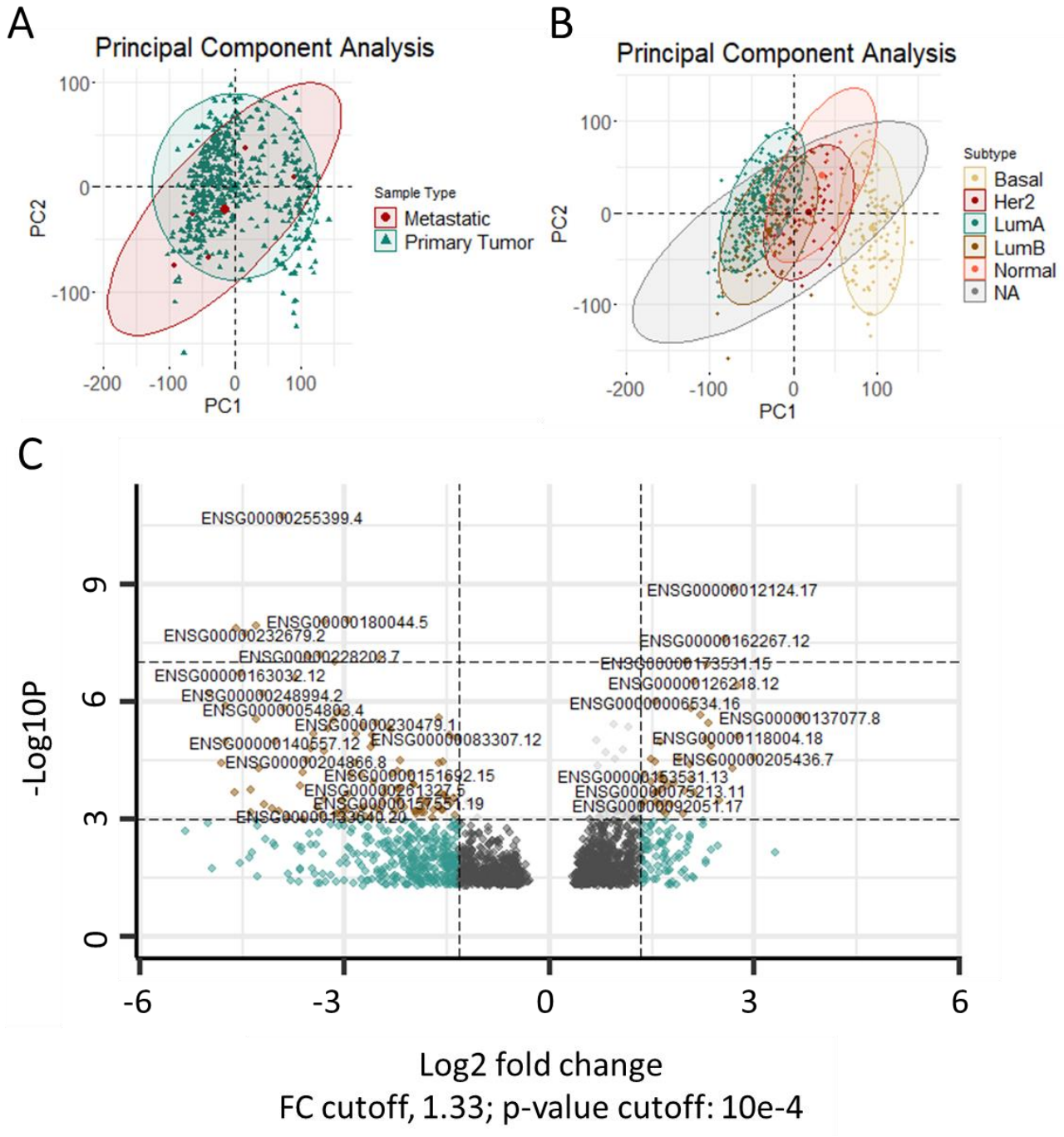


Figure 3. Differential Gene Expression analysis of metastatic samples. A-B) Principal component analysis (PCA) of log2 RNA-seq counts between sample type **(A),** and sample subtype **(B)** comparing metastatic vs tumor primary samples. Ellipse indicates a 90% confidence interval of groups. **C) Volcano plot of the difference expression genes** in metastatic samples vs primary tumor samples (p-value < 0.05). A total of 1689 DEGs, 869 down-regulated and 820 up-regulated, represented between the log2 Fold-Change and the log10 p-value. FC: Fold Change.

	gene_type	gene_name	level	logFC	AveExpr	t	P.Value	adj.P.Val	B
ENSG00000273388.1	lncRNA	AC005291.2	2	-5,43665	-0,94041	-8,30345	8,22E-16	1,49E-11	23,81866
ENSG00000112936.19	protein_coding	C7	1	5,559323	1,995661	7,303886	1,01E-12	9,18E-09	18,38124
ENSG00000255399.4	lncRNA	TBX5-AS1	1	-3,90759	0,894768	-6,85752	1,93E-11	1,16E-07	14,8048
ENSG00000145423.5	protein_coding	SFRP2	2	-6,43062	8,233741	-6,18982	1,19E-09	4,73E-06	11,3932
ENSG00000012124.17	protein_coding	CD22	1	2,683749	2,517614	6,175106	1,3E-09	4,73E-06	11,53916
ENSG00000180044.5	protein_coding	C3orf80	2	-2,94984	1,607013	-5,85536	8,28E-09	2,36E-05	9,423909
ENSG00000089225.20	protein_coding	TBX5	1	-3,31005	1,361162	-5,83844	9,12E-09	2,36E-05	9,318614
ENSG00000261039.3	lncRNA	LINC02544	2	-4,29904	0,744506	-5,79049	1,19E-08	2,52E-05	9,068855
ENSG00000149968.12	protein_coding	MMP3	2	-6,69726	2,717329	-5,76362	1,39E-08	2,52E-05	8,937667
ENSG00000105989.10	protein_coding	WNT2	1	-4,58471	2,978126	-5,7635	1,39E-08	2,52E-05	8,956835
ENSG00000232679.2	lncRNA	LINC01705	2	-4,44902	-1,15948	-5,70604	1,91E-08	3,13E-05	8,649624
ENSG00000167749.11	protein_coding	KLK4	2	-5,80277	0,375268	-5,69157	2,07E-08	3,13E-05	8,579178
ENSG00000162267.12	protein_coding	ITIH3	2	2,56458	-0,69369	5,6497	2,61E-08	3,63E-05	8,581422
ENSG00000134443.10	protein_coding	GRP	1	-6,22791	0,883065	-5,56151	4,22E-08	5,46E-05	7,947352
ENSG00000228203.7	lncRNA	GRASLND	2	-3,35698	0,39646	-5,47361	6,77E-08	8,18E-05	7,527034
ENSG00000006788.14	protein_coding	MYH13	2	-3,53291	-1,65126	-5,45872	7,33E-08	8,18E-05	7,455695
ENSG00000180785.10	protein_coding	OR51E1	1	-2,49928	0,704597	-5,45014	7,67E-08	8,18E-05	7,427125
ENSG00000173531.15	protein_coding	MST1	1	1,985869	1,515526	5,410875	9,45E-08	9,52E-05	7,492624
ENSG00000133110.15	protein_coding	POSTN	2	-3,13957	9,675425	-5,39846	1,01E-07	9,63E-05	7,401517
ENSG00000105737.9	protein_coding	GRIK5	2	2,302981	0,061809	5,378538	1,12E-07	0,000102	7,281366

Table 4. Top 20 significant up-regulated and down-regulated genes in metastatic comparing to primary tumor samples with its gene annotations.

4.2 Identification of differential DNA methylation regions in metastatic samples

DNA methylation data required a more complex preprocessing than RNA-seq data since it was necessary to eliminate rows with missing values, SNPs overlapping probes, and probes that have been demonstrated to map to multiple places in the genome (results in Supplementary Figure 2).

Next, an exploratory analysis was performed, first, the β-values were analyzed by PCA to see if there was a differential profile between metastatic and primary tumor samples. The PCA shows that metastatic samples do not present a differential profile (Figure 4A), both primary tumors and metastatic samples exhibit high epigenetic variability, however, as in the gene expression, there is a difference in the basal subtypes of the primary tumor samples (Figure 4B). In parallel, a box plot has been generated with the average methylation values between the metastatic and primary tumor samples, although not significant, the metastatic samples have a lower methylation average than the primary tumor ones (Figure 4C).

Following the exploratory analysis, normalized data were used to identify differentially methylated CpGs (DMCs) using *limma* R package applying p-value < 0.05. A total of 34,128 DMCs have been determined (it represents 10,2% of the total CpGs analyzed (333315)), of which 30802 are hypomethylated and 3326 are hypermethylated. It should be noted that the hypomethylated DMCs represent more than 90% of the total DMCs found compared to 10% of hypermethylated ones. The volcano plot in Figure 4D shows the gene-annotated DMRs, and in Table 4 the top 20 genes annotated DMCs ranked by p-value (p-value >0.05). The top five most significantly hypermethylated genes annotated DMCs in metastatic samples are IER2, SPINT2, MREG, PRR19_PAFAH1B3, and PRR7, three of which are promoter associated. The top five hypomethylated genes annotated DMCs are FKRP, HDAC4, NR5A1, COMT, and MASP2, the first two are lncRNA genes, and the other three are protein-coding genes.

Additionally, DMC chromosomic location was analyzed. DMCs are distributed throughout the genome and chromosomes 1, 2 and 6 contain a larger percentage, with 9.4%, 7.8%, and 7.3% respectively (Figure 5A). Second, DMCs locations were analyzed according to genes and CpG islands. Regarding the genes, the hypomethylated DMCs have been mainly found in the body of the genes and the intergenic zones (IGR) at a ratio of 39% and 31% respectively. On the other hand, the hypermethylated DMCs present a greater proportion in the gene promoter zones, TSS200 (18%), TSS1500 (16%), and 5'UTR (13%) (Figure 5B), which would agree with the fact that there are 46% of hypermethylated DMCs in the CpG island area (islands are usually in the promoters, 5'UTR and 1st exon) Hypomethylated DMC and respective CpG islands have been mainly in ope sea areas (39%) (Figure 5C).

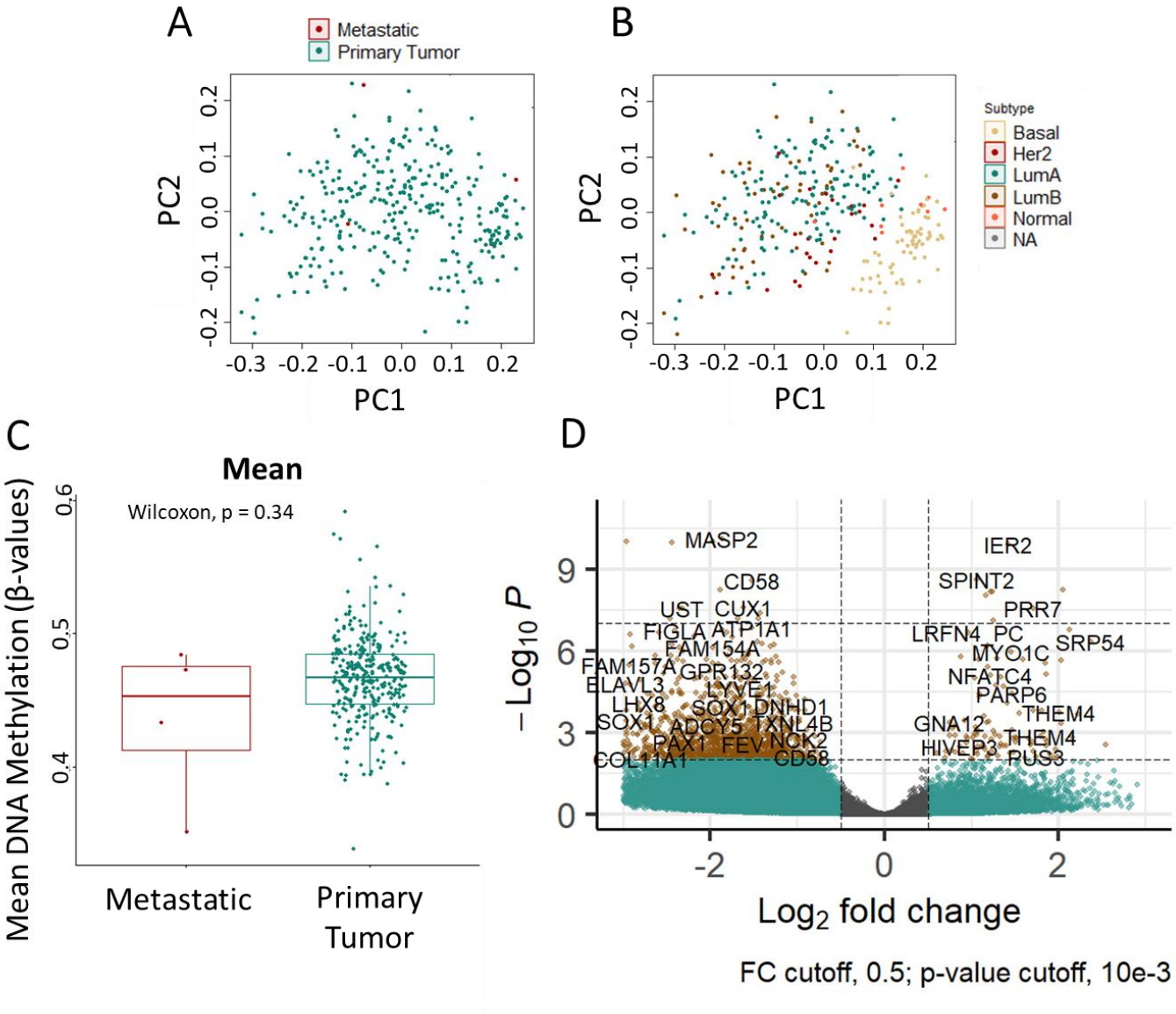


Figure 4. Differential DNA methylation profile in metastatic samples. A-B) Principal component analysis (PCA) of β-values between sample type **(A),** and sample subtype **(B)** comparing metastatic vs tumor primary samples. Ellipse indicates a 90% confidence interval of groups. **C) β-values mean**. Box plot of the differences in the methylation levels (β-values mean) between primary tumor vs metastatic samples. **D) Volcano plot of gene annotated DMRs** in metastatic samples vs primary tumor samples (p-value < 0.05). A total of 34128 CpG sites, 3326 hypermethylated and 30803 hypomethylated, were fond and represented between the log2 Fold-Change and the log10 p-value and annotated with the genes where they belong.

	chr	pos	strand	Relation_ to_Island	Gene	RefGene_ Group	logFC	AveExpr	t	P.Value	adj.P.Val	B
cg27297025	chr19	47259588	-	Island	FKRP	Body	-2,7253441	5,8984612	-8,9903424	2,1066E-17	7,0217E-12	27,6632066
cg18359321	chr2	240058950	-	N_Shelf	HDAC4	Body	-1,7217728	5,05139421	-8,8510504	5,7811E-17	9,6347E-12	26,7405803
cg13661129	chr9	127269876	+	S_Shelf	NR5A1	TSS200	-1,8140787	2,46610387	-8,2522686	3,9783E-15	3,845E-10	22,8725894
cg03724721	chr22	19939061	-	OpenSea	COMT	5'UTR	-1,9918643	3,80276154	-8,2308059	4,6143E-15	3,845E-10	22,7370347
cg00523012	chr19	13264324	+	Island	IER2	Body	1,41674946	-6,1866303	8,10289407	1,111E-14	6,8551E-10	21,933792
cg12888113	chr1	11107048	-	OpenSea	MASP2	Body;Body	-1,8408701	6,00937614	-8,0683213	1,4067E-14	6,8551E-10	21,71806
cg12528056	chr11	60619955	-	Island	GPR44	3'UTR	-2,4330618	5,97351641	-8,0649247	1,4397E-14	6,8551E-10	21,6968972
cg14451382	chr5	1876397	-	Island	NA		-2,9270374	0,82096889	-7,9491595	3,1573E-14	1,3155E-09	20,9790458
cg08231577	chr1	117077088	+	Island	CD58	Body	-1,533621	5,38560436	-7,3969946	1,206E-12	4,4666E-08	17,6501705
cg21435684	chr17	80255457	-	S_Shelf	NA		2,04553513	-4,8623527	7,33116881	1,8402E-12	6,1337E-08	17,2642084
cg15375239	chr19	38755287	-	Island	SPINT2	5'UTR	1,0579897	-6,4284235	7,29085012	2,3807E-12	6,8877E-08	17,0289872
cg08182975	chr2	7164527	+	OpenSea	RNF144A	Body	-1,8473419	4,40019047	-7,281425	2,5281E-12	6,8877E-08	16,9741307
cg15427886	chr8	55379663	+	Island	NA		-4,2703507	0,42815252	-7,271889	2,6864E-12	6,8877E-08	16,91868
cg09187505	chr6	106434429	-	Island	NA		-4,0217433	4,20290176	-7,144983	5,9953E-12	1,4274E-07	16,1855918
cg01778114	chr2	216877947	+	Island	MREG	Body	1,20525166	-5,220954	7,13275208	6,4742E-12	1,4386E-07	16,1154196
cg18517898	chr19	36435675	+	Island	LRFN3	Body	-1,8609022	6,18113809	-7,0993782	7,9811E-12	1,6626E-07	15,9243782
cg27324804	chr7	101509161	+	OpenSea	CUX1	Body	-1,6217908	5,71482482	-7,0015657	1,4679E-11	2,7177E-07	15,3681554
cg07493465	chr6	149301241	-	OpenSea	UST	Body	-2,3394407	3,79104967	-6,9947834	1,5309E-11	2,7177E-07	15,3297921
cg10848272	chr14	70653719	-	N_Shore	SLC8A3	5'UTR	-3,2797162	0,63163133	-6,9928658	1,5492E-11	2,7177E-07	15,31895
cg27320983	chr19	1230171	+	Island	C19orf26	3'UTR	-1,721587	6,77709189	-6,9613593	1,8824E-11	3,1372E-07	15,1411219

Table 5. Top 20 significant hypermethylated and hypomethylated CpG regions in metastatic samples comparing to primary tumor samples with its gene annotations.

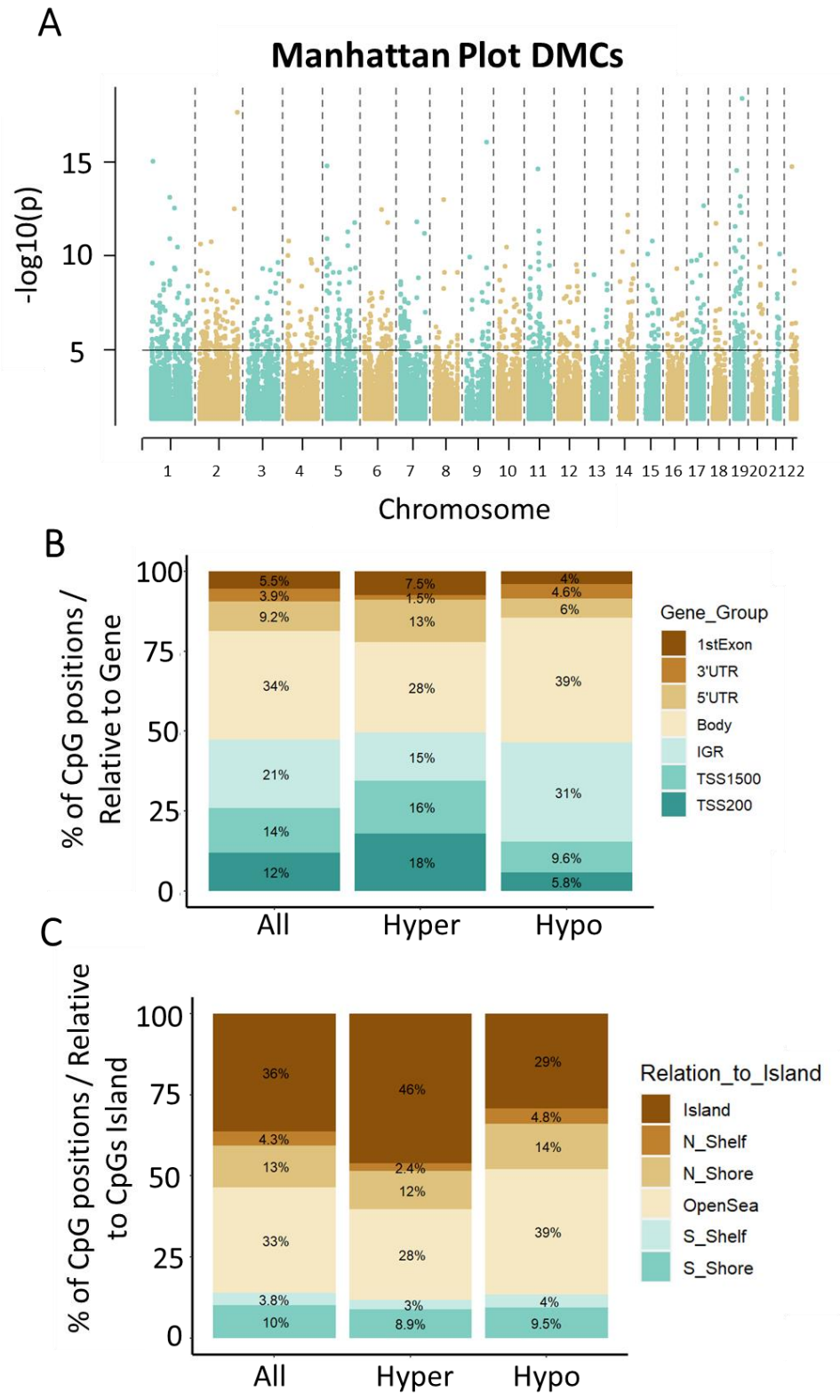


Figure 5.- DMCs distribution in metastatic samples. A) Manhattan plot of the DNA methylation sites (CpG sites) in metastatic samples based on p-values, organized according to their location on different chromosomes. **B) Genomic region of DMCs.** Percentage of the CpG sites (hypermethylated, hypomethylated, and both together (All)) referring to its distribution across gene regions, divided into promoters (1stExon, 3'UTR, 5'UTR, gene body, intra-genic regions (IGR), TSS1500, and TSS200). **C) Genomic localization of DMCs sites related to CpG islands.** Percentage of the CpG sites (hypermethylated, hypomethylated, and both together (All)) referring to its localization related to CpG islands, divided into Island, opensea, CpG self, and CpG shore.

4.3 Identification of significant Copy Number Variation (CNV) in metastatic samples

After download CNV data, exploratory analysis of Primary Tumor and Metastatic samples CNVs were performed. Table 6 shows the number of CNVs downloaded for each type of sample.

Sample Type	Count	Mean	sd
Metastatic	4296	-0.1216501	0.9150725
Primary Sample	611725	-0.1428994	1.0293197

Table 6. Summary table of downloaded CNV data. sd = standard deviation.

The density plot of the segment means of Primary Tumor and Metastatic samples shows that metastatic samples have an increase in the density (segment mean different from 0), this could indicate differences in the number of amplifications and deletions (Supplementary Figure 3). Figure 6A plot (segment mean values plot) confirms this fact, it can be observed that the segment means of the primary tumor samples are centered at 0, with no peaks. However, the metastatic samples, despite also presenting a peak centered at 0, also observe two peaks above and below 0.3 and -0.3, respectively, indicative of the presence of CNV. In addition, the average of metastatic samples is significantly different from the primary tumor samples (Figure 6A). In the same way, a bar plot was generated of the segment mean but divided between chromosomes, and according to whether they represent a gain or a loss in the primary tumor and metastatic samples separately to see how the aberrations are distributed along the genome (Figure 6B).

After the exploratory analysis, the identification of recurrent CNVs (amplifications and deletions) of metastatic samples were detected using GAIA as described in materials and methods. As a result, 4288 genes with CNVs have been obtained, including 2733 with amplifications and 1555 with deletions. After duplicate deletion, there is a total of 2221 genes with CNV, 1299 with amplifications, and 922 with deletions. The arrangement of this CNV according to its location on the chromosomes is shown in Figure 6C and Table 7.

Chromosomes	6	9	13	14	15	16	20	23
Amplification	2	463	0	0	0	372	0	462
Deletion	0	0	349	540	10	0	23	0

Table 7. Summary table of CNV (amplifications and deletions) in metastatic samples divided by chromosome localization.

Table 8 shows some of the genes that present CNVs. 10 with amplification: DDX39BP1, MCCD1P1, SLC1A1, SPATA6L, RP11, PPAPDC2, CDC37L1, AK3, RCL1, and MIR101, and 10 with deletion: STARD13, AL138999.1, RFC3, RNU5A, AL161891.1, VDAC1P12, SNORA25, LINC00457, GAMTP2, and NBEA. Also, the chromosome where the CNVs are found are shown, amplifications are only found on 4 chromosomes, 6, 9, 16, and 23, and deletions on 4 different ones, 13, 14, 15, and 20.

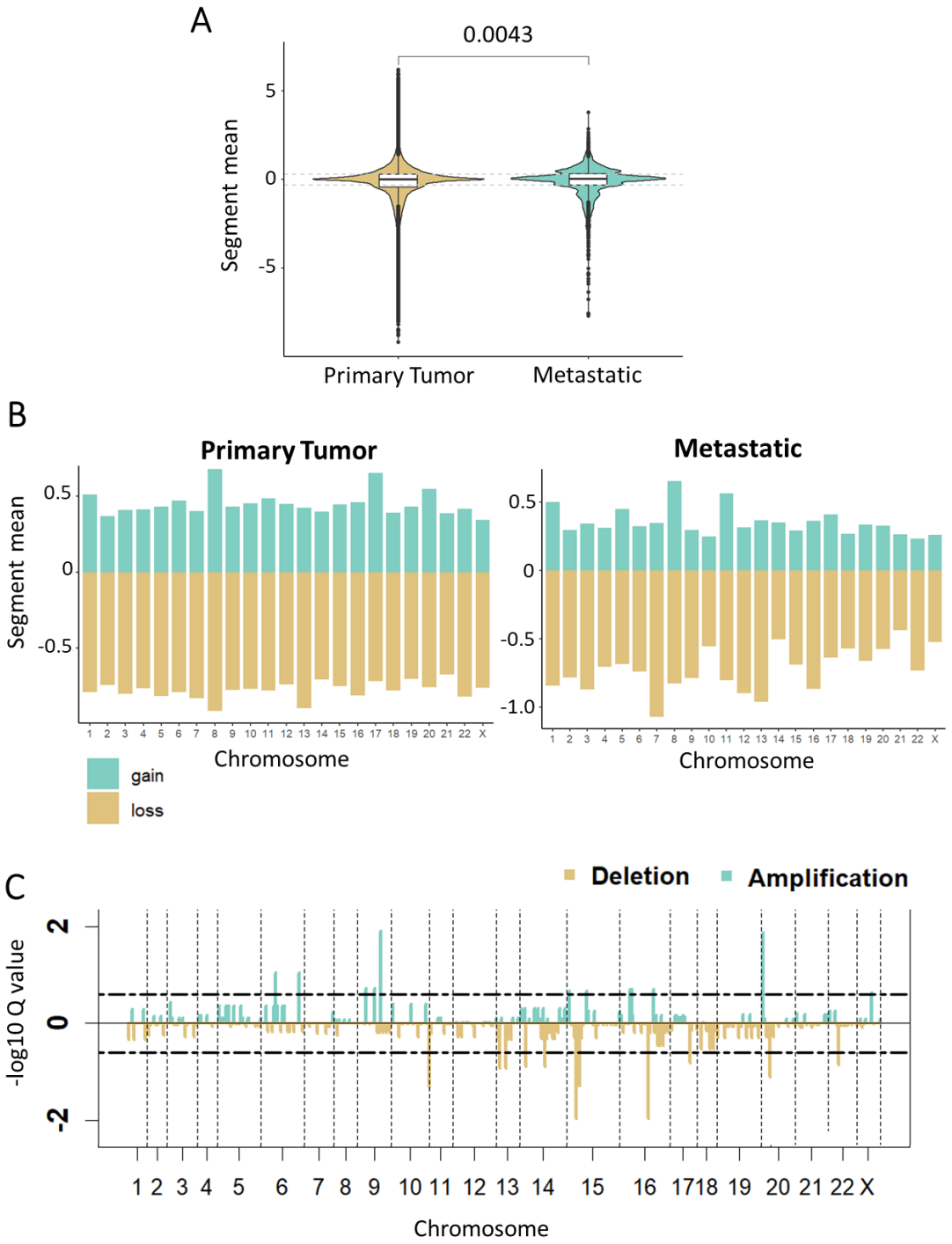


Figure 6. Identification of recurrent CNV in Metastatic samples. A) Violin plot of the segment mean distribution of primary tumor and metastatic samples. Discontinuous lines represent the values used for the detection of CNVs in the GAIA approach (>0.3 as amplification and < -0.3 as deletion). **B) Box plot representing the segment mean** divided by chromosomes, and by the gain (top) and the loss (bottom) of primary tumor samples (left), and metastatic samples (right). **C) Bar plot of the identification of the recurrent CNVs in metastatic samples** distributed by chromosomes, and by the amplifications (top) and the delations (bottom).

GeneSymbol	chr	Aberration	q-value	AberrantRegion	GeneRegion
DDX39BP1	6	Amp	8,71875E+14	6:29881490-1	6:29874320-29874686
MCCD1P1	6	Amp	8,71875E+14	6:29881490-1	6:29875560-29876422
SLC1A1	9	Amp	1,88822E+14	9:66189025-1	9:4490444-4587469
SPATA6L	9	Amp	1,88822E+14	9:66189025-1	9:4553386-4666674
RP11	9	Amp	1,88822E+14	9:66189025-1	9:4633027-4633756
PPAPDC2	9	Amp	1,88822E+14	9:66189025-1	9:4662298-4665256
CDC37L1	9	Amp	1,88822E+14	9:66189025-1	9:4679559-4708398
AK3	9	Amp	1,88822E+14	9:66189025-1	9:4711155-4742043
RCL1	9	Amp	1,88822E+14	9:66189025-1	9:4792869-4885917
MIR101	9	Amp	1,88822E+14	9:66189025-1	9:4850291-4850381
STARD13	13	Del	1,1811E+14	13:68677373-0	13:33677272-33924767
AL138999.1	13	Del	1,1811E+14	13:68677373-0	13:33911020-33911109
RFC3	13	Del	1,1811E+14	13:68677373-0	13:34392186-34540695
RNU5A	13	Del	1,1811E+14	13:68677373-0	13:34403676-34403786
AL161891.1	13	Del	1,1811E+14	13:68677373-0	13:34486129-34486218
VDAC1P12	13	Del	1,1811E+14	13:68677373-0	13:34656566-34657447
SNORA25	13	Del	1,1811E+14	13:68677373-0	13:34674848-34674975
LINC00457	13	Del	1,1811E+14	13:68677373-0	13:35009587-35214822
GAMTP2	13	Del	1,1811E+14	13:68677373-0	13:35148341-35148803
NBEA	13	Del	1,1811E+14	13:68677373-0	13:35516424-36247159

Table 8. 20 of the detected CNVs in metastatic samples with their gene annotations.

.

4.4 Identification of metastatic 3-omics gene set

4.4.1 3 - omics data conceptual integration

Once the annotation of the genes differentially expressed, methylated, and with differences in their CNV patterns has been obtained, this data was conceptually integrated through their comparison to obtain which genes are found in common in the 3 analyses. A metastatic gene set of 24 genes has been found (Figure 7A), which correspond to ADAMTSL1, ANGEL1, ATXN2L, CCL21, CCNA1, CLN3, COG3, DDX58, HS3ST2, IRX3, IRX6, LRFN5, MAZ, MMP2, NEK9, NETO2, PRRT2, SALL1, SLC25A15, STX4, TBX6, THSD1 and TMEM229B (Table 9).

In Table 9 apart from the genes resulting from the integration, there are the associated differentially methylated CpGs. A large majority of these genes have hypomethylated DMCs regardless of whether they are up or down-regulated. In total, 20 genes (83.3%) are hypomethylated and 3 hypermethylated (12.5%). Curiously STX4 is hyper/ hypomethylated in different DMRs (Table 9).

Taking into account that it is described that methylation leads to the cancer suppressor gene's inactivation in cancer, it was studied in more depth because most of the selected genes are hypomethylated regardless of their expression. For this reason, it has been looked at where the DMCs are concerning CpG islands and genes. Regarding the genes, the hypomethylated DMCs have been mainly found in the body of the genes and the TSS1500 promoter region at a ratio of 38% in both cases. On the other hand, the hypermethylated DMCs are present in two locations, in the body at 67% and in the TSS1500 promoter regions at 33% (Figure 7B). Regarding the CpG Island position, in both cases, hypomethylated and hypermethylated DMCs are mainly in the island with 42% and 67% respectively (Figure 7C).

In addition, the correlation between gene expression and methylation levels of DMCs were studied. Although not in a very strong way, broadly speaking, gene expression and methylation correlate positively, i.e. in upregulated genes hypermethylation is seen, and in downregulated genes hypomethylation (Figure 7D). However, this trend could be influenced by the previously mentioned fact that a large majority of genes are hypomethylated. Supplementary Figure 4 shows the correlation individually for the 24 genes of the metastasis gene set, where it can be observed that some of the genes have a negative correlation (In those cases where a gene has more than one methylated DMCs only a representative case is shown).

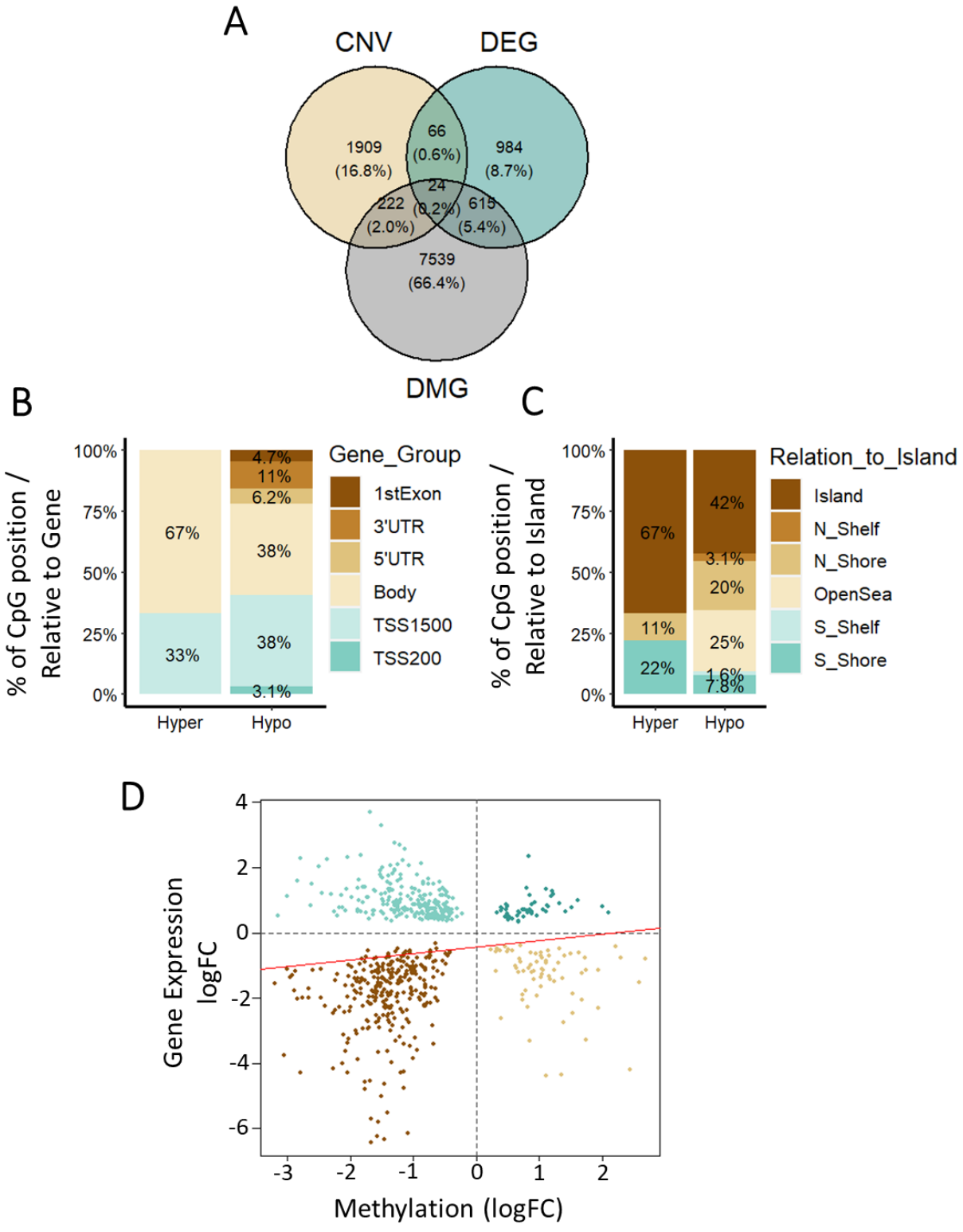


Figure 7. Integration of DEG, DMR and CNVs data of metastatic samples. A) Venn Diagram of significant DEG, DMR and CNVs gene annotation overlapping. **B) Genomic region of DMCs.** Percentage of the CpG sites (hypermethylated and hypomethylated) referring to its distribution across gene regions, divided into promoters (1stExon, 3'UTR, 5'UTR, gene body, intra-genic regions (IGR), TSS1500, and TSS200). **C) Genomic localization of DMCs sites related to CpG islands.** Percentage of the CpG sites (hypermethylated and hypomethylated) referring to its localization related to CpG islands, divided into Island, opensea, CpG self, and CpG shore. **D)** Scatter plot of gene expression and DNA methylation correlation. Each point represents a significant up/down methylated CpG site and its gene pair.

Gene	Gene_ expression	chr	Name	DNA_ Methylation	Relation_to_Island	Gene_ Group	Aberration
ADAMTSL1	Down	chr9	cg06294856	Hypo	N_Shore	Body	Amp
			cg13739410	Hypo	OpenSea	Body	
			cg06500883	Hypo	Island	Body	
			cg13468759	Hypo	OpenSea	Body	
			cg14523394	Hypo	Island	Body	
			cg13883376	Hypo	OpenSea	Body	
			cg14003978	Hypo	OpenSea	Body	
ANGEL1	Up	chr14	cg05349242	Hypo	OpenSea	3'UTR	Del
ATXN2L	Up	chr16	cg23051598	Hypo	Island	5'UTR	Amp
			cg05785839	Hypo	OpenSea	3'UTR	
CCL21	Up	chr9	cg27443224	Hypo	OpenSea	1stExon	Amp
CCNA1	Down	chr13	cg14089714	Hypo	N_Shore	TSS1500	Del
			cg26345046	Hypo	N_Shore	TSS1500	
			cg19866195	Hypo	N_Shore	TSS1500	
CLN3	Up	chr16	cg09900266	Hypo	OpenSea	3'UTR	Amp
COG3	Up	chr13	cg22419414	Hypo	OpenSea	Body	Del
DDX58	Down	chr9	cg14277298	Hypo	OpenSea	3'UTR	Amp
HS3ST2	Down	chr16	cg09869531	Hypo	OpenSea	Body	Amp
			cg19064258	Hypo	Island	1stExon	
			cg03132773	Hypo	Island	TSS1500	
			cg16399049	Hypo	Island	TSS200	
			cg08214995	Hypo	Island	1stExon	
IRX3	Down	chr16	cg13660279	Hypo	Island	Body	Amp
IRX6	Down	chr16	cg02787087	Hypo	N_Shore	Body	Amp
		chr16	cg05099576	Hypo	N_Shore	Body	
		chr16	cg02198701	Hypo	N_Shore	3'UTR	

		chr16	cg02602550	Hypo	S_Shore	Body	
		chr16	cg04834436	Hypo	Island	Body	
		chr16	cg06431877	Hypo	S_Shore	Body	
LRFN5	Down	chr14	cg04784672	Hypo	S_Shore	5'UTR	Del
		chr14	cg13526007	Hypo	S_Shore	TSS200	Del
MAZ	Up	chr16	cg12521167	Hyper	Island	Body	Amp
		chr16	cg16518772	Hyper	Island	Body	
		chr16	cg04588455	Hyper	Island	Body	
		chr16	cg00564759	Hyper	Island	Body	
		chr16	cg07675334	Hyper	Island	Body	
MMP2	Down	chr16	cg08133699	Hypo	OpenSea	Body	Amp
		chr16	cg09530163	Hypo	N_Shore	TSS1500	
		chr16	cg26795346	Hypo	S_Shelf	Body	
		chr16	cg12317456	Hypo	N_Shore	TSS1500	
		chr16	cg08318842	Hypo	N_Shore	TSS1500	
		chr16	cg01821058	Hypo	N_Shore	TSS1500	
NEK9	Up	chr14	cg08310116	Hypo	S_Shore	TSS1500	Del
NETO2	Down	chr16	cg05532446	Hypo	N_Shelf	Body	Amp
		chr16	cg03283929	Hypo	OpenSea	3'UTR	
PRRT2	Up	chr16	cg04203429	Hyper	N_Shore	TSS1500	Amp
RCBTB1	Up	chr13	cg11801959	Hypo	OpenSea	Body	Del
SALL1	Down	chr16	cg06724588	Hypo	Island	TSS1500	Amp
		chr16	cg06274671	Hypo	Island	TSS1500	
		chr16	cg06232807	Hypo	Island	TSS1500	
		chr16	cg05404010	Hypo	Island	TSS1500	
		chr16	cg05213609	Hypo	Island	TSS1500	
		chr16	cg02288754	Hypo	Island	Body	
		chr16	cg09016242	Hypo	Island	TSS1500	

		chr16	cg07498275	Hypo	Island	TSS1500	
		chr16	cg06653699	Hypo	Island	TSS1500	
		chr16	cg02864757	Hypo	Island	Body	
		chr16	cg00310215	Hypo	Island	TSS1500	
		chr16	cg27423760	Hypo	Island	TSS1500	
		chr16	cg01146232	Hypo	Island	TSS1500	
		chr16	cg01500945	Hypo	Island	TSS1500	
		chr16	cg08776356	Hypo	Island	TSS1500	
		chr16	cg04844564	Hypo	Island	Body	
SLC25A15	Up	chr13	cg16989646	Hypo	Island	5'UTR	Del
		chr13	cg00151919	Hypo	N_Shore	TSS1500	
STX4	Up	chr16	cg05916757	Hypo	N_Shelf	3'UTR	Amp
		chr16	cg09775103	Hyper	Island	Body	
TBX6	Up	chr16	cg05806717	Hypo	N_Shore	5'UTR	Amp
THSD1	Down	chr13	cg04549287	Hypo	OpenSea	Body	Del
		chr13	cg23498925	Hypo	OpenSea	Body	
TMEM229B	Up	chr14	cg25006823	Hyper	S_Shore	TSS1500	Del
		chr14	cg20454158	Hyper	S_Shore	TSS1500	

Table 9. Summary of metastatic gene set resulting from the integration of the DEG, DMR and CNV data of the metastatic samples against primary tumor samples.

4.4.2 Pathway characteristics of metastatic gene set

The metastatic gene set, defined in the conceptual data integration, was analyzed in more depth to determine how it could be influencing the evolution of breast cancer and its real importance in the metastatic state. For this reason, the possible involvement of these genes in several previously described biological mechanisms was analyzed through the analysis of several genes sets such as GO, Hallmarks, Cancer gene set, and KEGG.

It is necessary to mention that the number of genes that form the metastatic gene set is small, and this fact makes it difficult to see enriched processes. Therefore, having as an objective to practice the more technical aspects in the use of programs such as Cytoscape, the enrichment of processes involving only 1 gene of the set gene was considered valid.

The enrichment analysis has determined that the upregulated genes are positively involved (enriching) in processes related to membrane organization and fusion, as well as with several processes related to the SNARE family of proteins such as vesicle transport (Figure 8A, C). Likewise, the use of the Hallmarks database has determined that some of the upregulated genes are involved in apical junctions (Figure 8B). All these processes seem to be related to each other, a fact that will be analyzed later with the use of the Cytoscape tool. However, no enriched processes have been found when the cancer database was used. The upregulated genes involved in these processes are PRRT2, STX4, CLN3, and CCL21.

On the other hand, it has been possible to relate downregulation genes with cell junction assembly (Figure 8A) and with apoptosis (Figure 8B). The downregulated genes involved in these processes are IRX3, THSD1, LRFN5, MMP2, and CCNA1.

Besides, the enrichment analysis with KEGG database has determined that both upregulated and downregulated genes are implicated in NF-kappa B signaling pathway. The genes involved in these processes are CCL21 and DDX58 respectively.

It must also be stated that enriched processes have been observed that are not mentioned and are involved in topics as diverse as urogenital development or abnormal iris morphogenesis. Those have been discarded for not having relationship with cancer and have not been further analyzed. Notwithstanding, and considering that these are processes related to development, it would be necessary to analyze them in depth in the case of looking for treatment targets.

Considering the results obtained from the enrichment analysis with the GO database with the upregulated genes, the results were integrated using the EnrichmentMap app (Cytoscape) to visualize the relationship between them. This analysis has determined three large blocks, all of them related to vesicle transport and cellular mobility. Regarding the networks of downregulated genes, highlight those involved in gap junction assembly, key in maintaining the cell-cell junctions of the epithelial monolayer of the breast, and in avoiding the epithelial-mesenchymal transition (Figure 9).

Seeing that there is a pattern of functionality in 10 of the 24 genes (PRRT2, STX4, CLN3, and CCL21 as up-regulated genes and IRX3, THSD1, LRFN5, MMP2, CCNA1 and DDX58 as down-regulated), this set of genes has been selected as top genes set for validation.

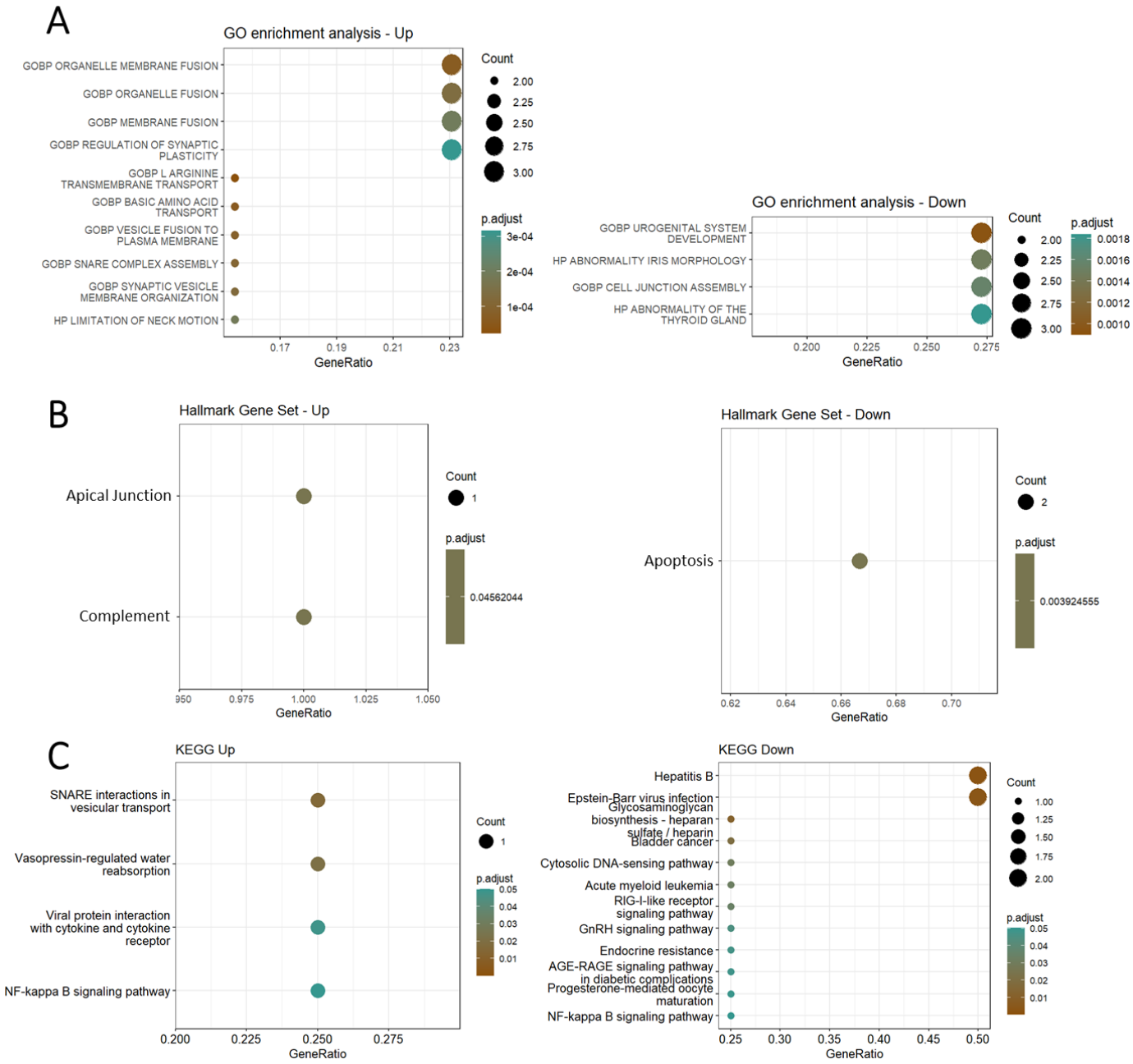


Figure 8. Enrichment analysis of metastatic gene set divided by up/downregulation gene expression. A) Gene Ontology (GO) enrichment analysis. **B)** Hallmark enrichment analysis. **C)** Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis. P-value < 0.05.

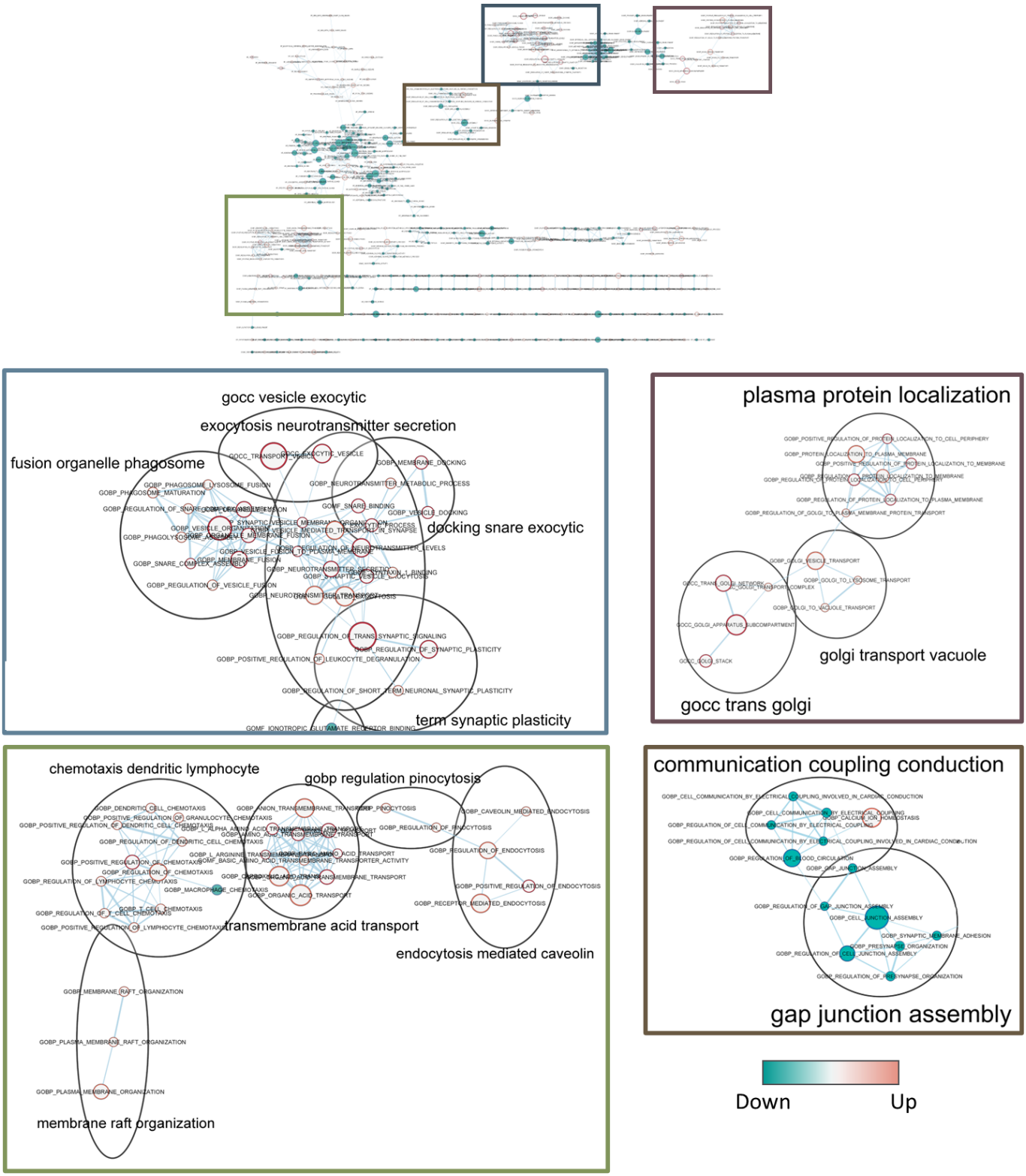


Figure 9. Enrichment Map of Metastatic Gene set corresponding to the GO Enrichment Analysis divided by up/downregulation gene expression. General visualization using EnrichmentMap app in Cytoscape. GO-terms have been represented by nodes, the size represents the number of genes assigned to the process, and the color represents the group. (Green represents the process enriched by downregulated genes, and brown represents the process enriched by up-regulated genes).

4.5 External Validation of metastatic top genes in multiple databases

Genes selected as top genes were validated in two independent databases, the Metastatic Breast Cancer Project (MBCP) and the METABRIC database to determine their potential as metastatic predictors.

First, using the MBCP database, the relationship between the top genes expression and the presence of metastasis or not has been analyzed, in this case, no significant difference has been observed.

Next, the relationship between top gene expression and the time to present a metastasis event (in days) was analyzed. To make this comparison, two groups have been generated, one with high expression and another with low expression, for each gene separately. In this case, only a significant difference has been observed in the case of CLN3, which determines that at higher expression, more time to present a metastasis event, which would go against the hypothesis presented in this work, since this gene in the TCGA database would be a marker of poor prognosis (Figure 10).

Considering that no positive results have been obtained in the MBCP validation, and therefore top genes would not be valid for predicting metastasis, it has been decided to see if, on the contrary, they could have some functionality in determining the prognosis of breast cancer regardless of the presence of metastasis. For this reason, a second database has been selected to carry out the validation. In this case, the METABRIC database was used to analyze the relationship between the expression of top genes with overall survival (OS) and relapse-free survival (RFS). In this case, the top genes were divided into up/down-regulated genes and in turn, these two groups were divided into two other groups according to high and low expression.

No significant differences have been obtained in the *Kaplan-Meier* analysis, but yes a trend that determines that upregulated genes could be decreasing OS and RFS. On the contrary, a high expression of the top genes found downregulated in TCGA would seem to correlate with a greater OS and RFS (Figure 11).

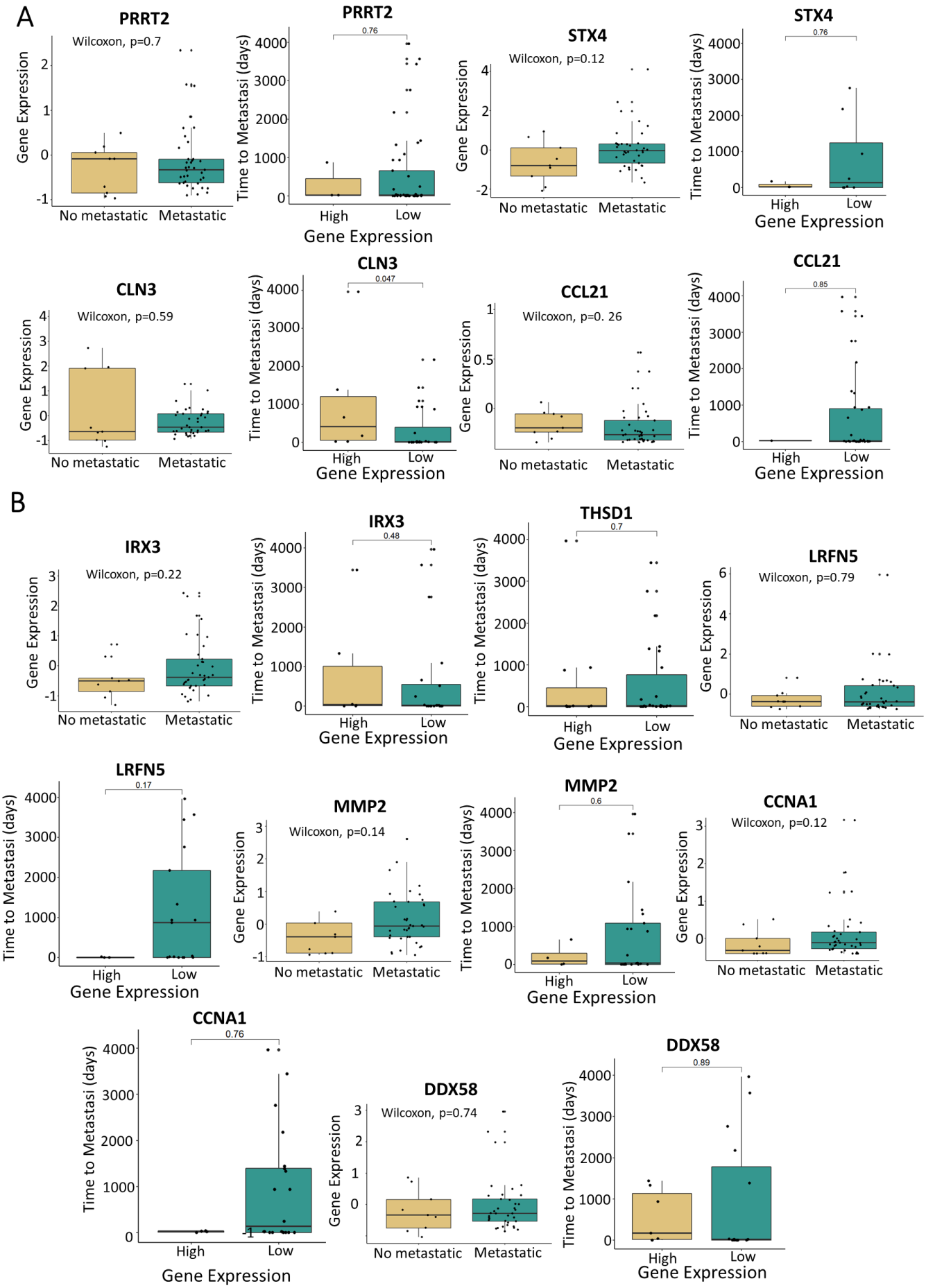

A
PRRT2
Wilcoxon, p=0.7
Gene Expression
No metastatic
Metastatic
PRRT2
0.76
Time to Metastasi (days)
High
Low
Gene Expression
STX4
Wilcoxon, p=0.12
Gene Expression
No metastatic
Metastatic
STX4
0.76
Time to Metastasi (days)
High
Low
Gene Expression
CLN3
Wilcoxon, p=0.59
Gene Expression
No metastatic
Metastatic
CLN3
0.047
Time to Metastasi (days)
High
Low
Gene Expression
CCL21
Wilcoxon, p=0. 26
Gene Expression
No metastatic
Metastatic
CCL21
0.85
Time to Metastasi (days)
High
Low
Gene Expression
B
IRX3
Wilcoxon, p=0.22
Gene Expression
No metastatic
Metastatic
IRX3
0.48
Time to Metastasi (days)
High
Low
Gene Expression
THSD1
0.7
Time to Metastasi (days)
High
Low
Gene Expression
LRFN5
Wilcoxon, p=0.79
Gene Expression
No metastatic
Metastatic
LRFN5
0.17
Time to Metastasi (days)
High
Low
Gene Expression
MMP2
Wilcoxon, p=0.14
Gene Expression
No metastatic
Metastatic
MMP2
0.6
Time to Metastasi (days)
High
Low
Gene Expression
CCNA1
Wilcoxon, p=0.12
Gene Expression
No metastatic
Metastatic
CCNA1
0.76
Time to Metastasi (days)
High
Low
Gene Expression
DDX58
Wilcoxon, p=0.74
Gene Expression
No metastatic
Metastatic
DDX58
0.89
Time to Metastasi (days)
High
Low
Gene Expression

Figure 10. Metastatic Gene set expression in the Metastatic BreastCancer Project (MBCP) validation clinical cohort. Box plots representing time to Metastasis in days and state of no metastasis vs metastasis depending on gene expression of gene set, each gene individual. **A)** Up-regulated genes, and **B)** Downregulated genes more significantly after the enrichment analysis.

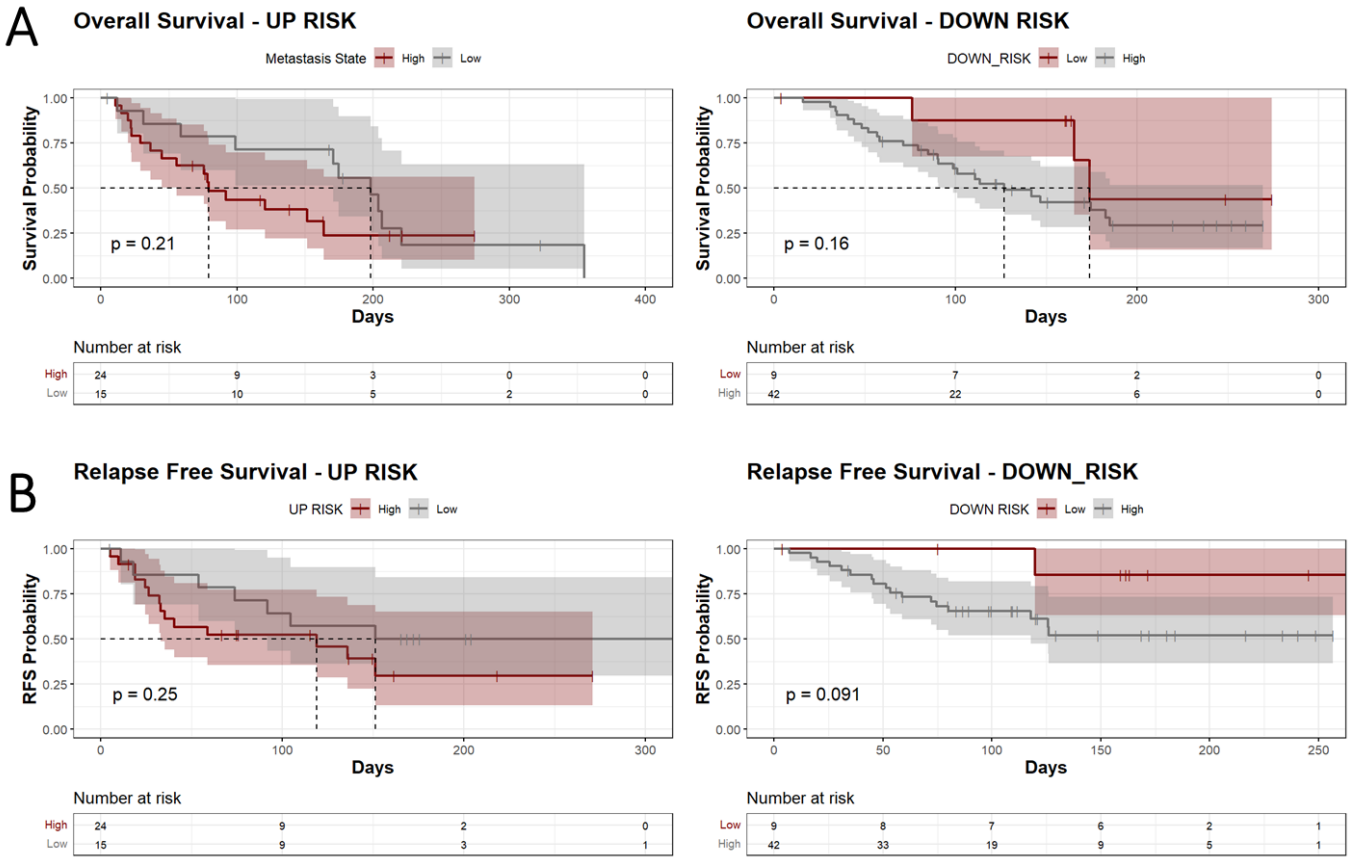


Figure 11. Metastatic Gene set expression in the METRABRIC validation clinical cohort. Kaplan-Meier analysis of **A)** Overall survival (OS) Probability between high and low expression of upregulated (left), and down-regulated (right) genes from the metastatic gene set. B) Relapse Free Survival (RFS) probability between high and low expression of upregulated (left), and down-regulated (right) genes form the metastatic gene set.

5. DISCUSSION

Tumor metastasis is the main cause of death in breast cancer patients, despite this fact, the mechanism by which a primary tumor evolves into metastasis remains poorly understood and currently, there is no treatment to prevent it. Most studies found in the literature focus on the study of primary tumors and few on the mechanism of metastasis. The study of metastasis progression presents great difficulty since it has been described that the differences between primary tumors and its metastasis are minimal, presenting similar genetic profiles [93] [94] [95].

Another major limitation of this study has been the number of metastasis samples used it has worked. Only 4-6 metastasis samples have been available in front of more than 1000 primary tumor samples, which did not allow for example to look at differences between IDC subtypes or to look at differences between the places where the metastasis is found. Fact that could be key in defining a good prediction model. There are indeed databases with more metastatic samples, but these databases only have RNA sequences and no other omics, for this reason, have been discarded.

Despite these difficulties, it has been possible to obtain some results, and more importantly, it has been possible to work with several bioinformatics tools and increase the knowledge in this field.

In this study, after individual omics analysis and data integration, 10 genes have been identified/proposed as top genes in the evolution of breast cancer towards a metastatic event, PRRT2, STX4, CLN3, CCL21, IRX3, THSD1, LRFN5, MMP2, CCNA1, and DDX58. The first 4 genes are upregulated, have CNV amplification, and are hypomethylated, except PRRT2 which is hypermethylated in metastatic samples vs primary tumor samples. On the other hand, the last 6 genes, are downregulated and hypomethylated (DDX58, IRX3, MMP2 have CNV amplification and THSD1, LRFN5, and CCNA1 deletions).

The results obtained with upregulated genes are in the same line as the fact that the hypomethylation of certain genes correlates with its upregulation and activation and with its role in cancer development, a process widely described in several types of cancer including breast cancer [96] [97] [98]. In the same way, it has also been described that CNV (amplification) of certain genes is involved in the development of cancer, altering the expression of certain genes in a positive way [99] [100] [101].

Contrary to what was expected, all the obtained downregulated genes are hypomethylated, when it is widely described that in a cancer context, generally, hypomethylated genes are genes that are upregulated, and those that are hypermethylated are downregulated [102] [103] [104]. However, it has also been described that DNA hypomethylation is associated with chromatin repression and gene silencing [105], as well as gene body hypomethylation, has been related to the loss of gene expression in various cancers [106] [107]. This would explain why the IRX3, THSD1, LRFN5, MMP2, CCNA1, and DDX58 genes have been hypomethylated and downregulated.

The enrichment study of all these genes has shown that they are involved in several processes, among which those involved in cell mobility and vesicle transport should be highlighted since they have been obtained for all the genes.

Transmembrane acid transport, pinocytosis, and endocytosis by caveolin are processes related to the Wnt pathway. This pathway is involved in various cellular functions such as migration and cell polarity. In addition, its involvement in cancer is widely described, specifically, it has been described that in breast cancer, it is mainly involved in proliferation and metastasis [108] [109]. These processes are also related to the internalization of membrane proteins such as cadherins, integrins among others involved in cell-cell adhesion, which favors the mesenchymal epithelium transition (EMT), the increase in cell motility and therefore metastasis [110] [111].

Above all, the enrichment of the pathways involving the SNARE protein family should be highlighted. Soluble N-ethylmaleimide-sensitive factor attachment protein receptors (SNAREs) are involved in the transport of proteins through transmembrane vesicles, including those involved in the formation of invadopodium and consequently cell mobility, invasion, and metastasis [112] [113]

On the other hand, it has been determined that some of the genes found to be downregulated are also involved in processes related to the membrane and mobility, such as gap junction assembly. Gap junctions are needed to maintain the mono epithelial cell layer that conforms breast duct, its dis-assembly is one of the first breast cancer steps [114]. This fact is consistent with the rest of the enriched pathways, if there is a downregulation of genes involved in maintaining and forming gap junctions, i.e. loss of these, it increases cell mobility.

Apart from the decrease in cell anchorage, the downregulation of genes related to apoptosis has been determined. Apoptosis pathways are typically inhibited in cancer processes [115].

However, top genes could not be validated as predictors of metastasis or as predictors of poor prognosis. This could be due to the limitations mentioned at the beginning of this section.

6. CONCLUSIONS

The following conclusions have been reached:

1.- Metastatic breast cancer samples present significant differences at the level of gene expression, methylation, and CNV concerning primary tumor samples.

2.- The metastatic samples present a greater proportion of hypomethylated genes than hypermethylated ones.

3.- The integration of the three omics data defined 10 top genes: PRRT2, STX4, CLN3, CCL21, IRX3, THSD1, LRFN5, MMP2, CCNA1, and DDX58.

4.- The key process in the evolution of a primary tumor towards a metastatic process is the acquisition of tumor cell mobility as well as the inhibition of apoptosis.

5.- However, the first objective proposed in this work, of defining a multi-omic risk profile of MBC, could not be completed. The list of genes resulting from the analysis is not able to predict metastasis.

Apart from the more biological conclusions, it is also necessary to define some conclusions at the learning level:

- The development of this work has allowed the acquisition of a wider knowledge of databases as well as of the various existing omics.
- Acquisition of experience in handling, preprocessing, and statistical analysis of various biological data.
- The acquisition of experience in omics data integration as well as in the use of R and computer tools such as Cytoscape.

6.1 Future Perspectives

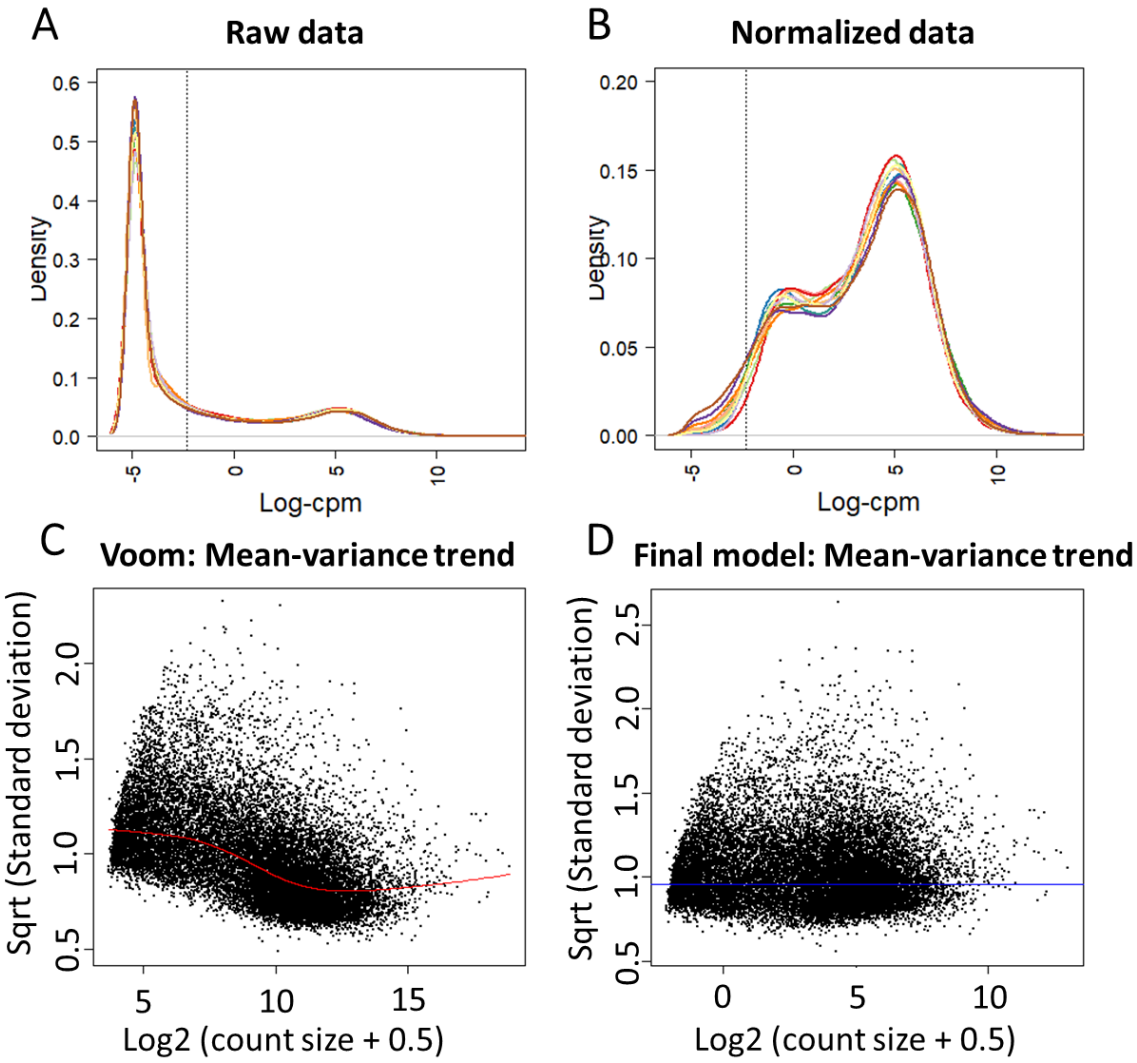
Bearing in mind that it has not been possible to validate top genes as predictors of metastasis, the following is proposed:

- As mentioned in the discussion, one of the limitations of this work is the size of the metastatic cohort, for this reason, the same analysis should be performed in a larger cohort. With a larger cohort, it would be possible to carry out the analysis independently for each subtype of IDC or divide by the site where the metastasis has occurred.
- Compare primary tumor without metastases (with a good prognosis) with primary tumor that have developed a metastasis (poor prognosis).
- Another approach would be to repeat the integration, applying various statistical methods of integration and comparing the results.
- To validate the results of the top genes, they could be compared with existing breast cancer risk scores.

- Finally, there has not been enough time and positive results to generate a risk prediction algorithm.

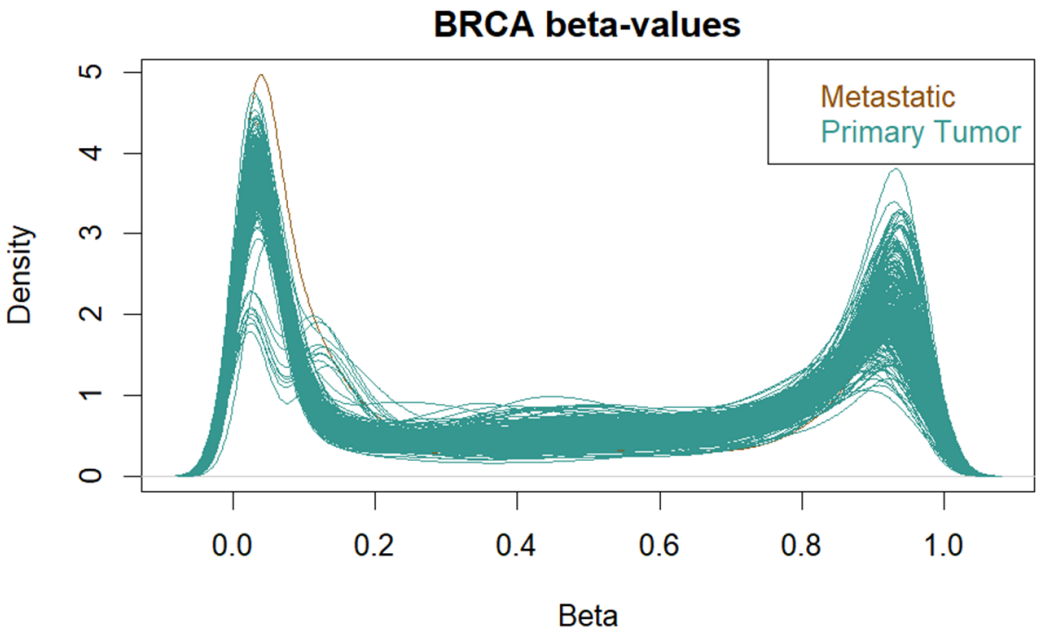
7. ANNEXES

7.1 Supplementary Figure 1



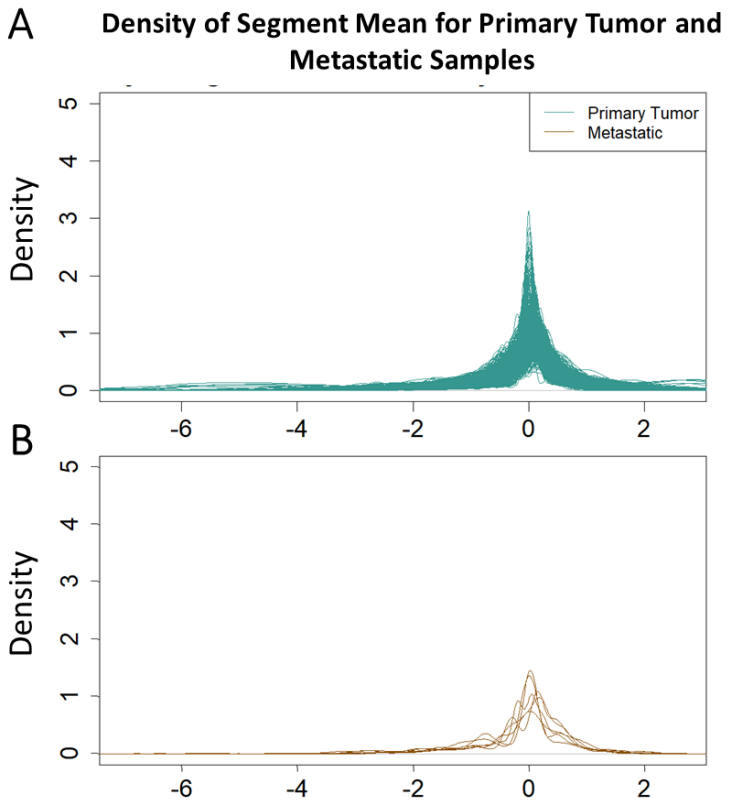
Supplementary Figure 1. Pre-processing of the RNA-seq data from primary and metastatic tumor samples. A) Density plot of row data before filtering and normalization. **B)** Density plot of the data after filtering and normalizing. **C)** Scatterplot of all the data showing the trend of the variance (log-cpm), the red curved line indicates that the samples have an overdispersion and that they need to be adjusted to a linear model to apply the *voom* function. **D)** Scatterplot of all the data showing the trend of the variance after applying the *lmFit* function. The straight line indicates that the variance no longer depends on the average level of expression and it is appropriate to proceed to the study of the differential expression.

7.2 Supplementary Figure 2



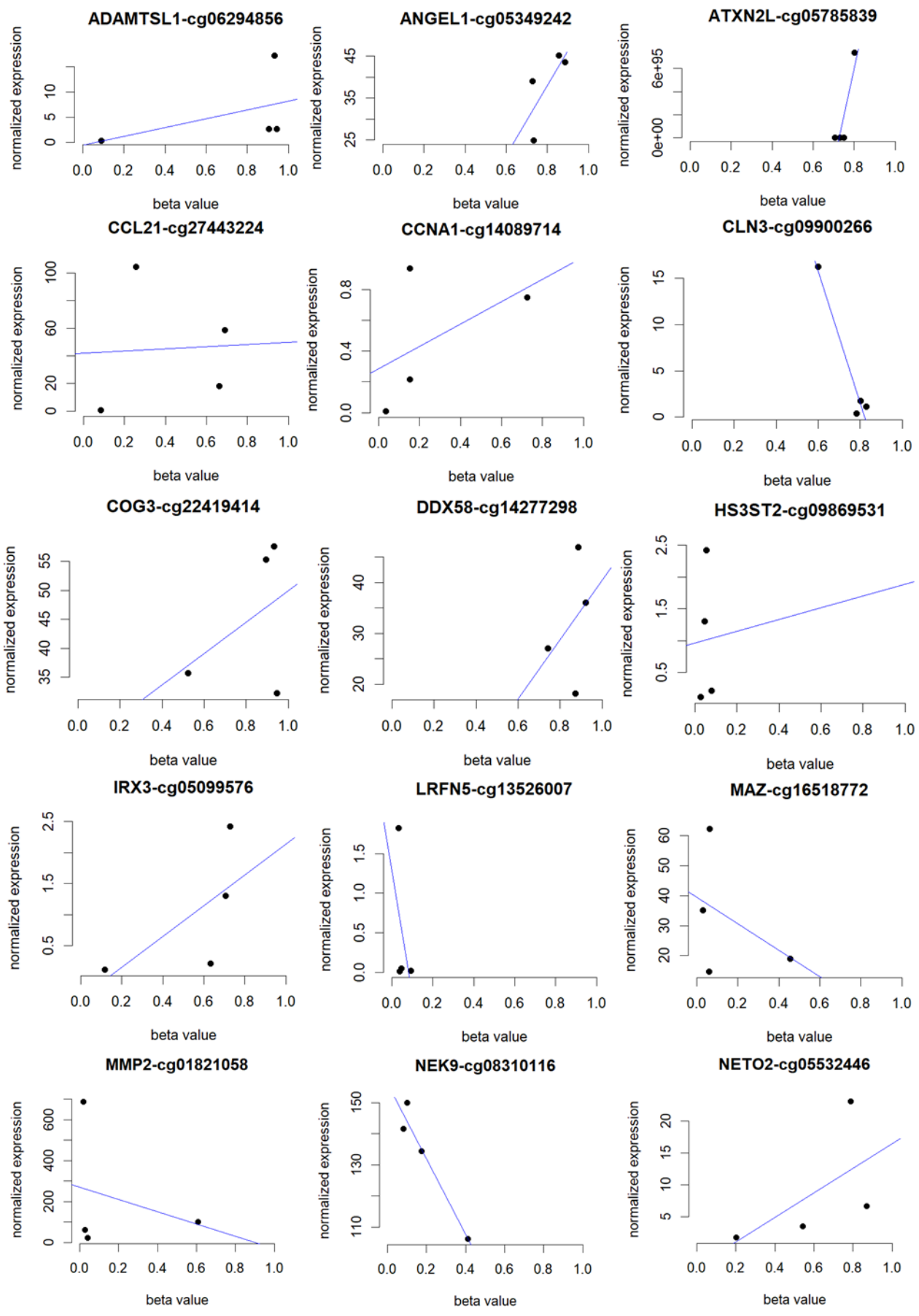
Supplementary Figure 2. B-values density plot of the primary tumor and metastatic samples.

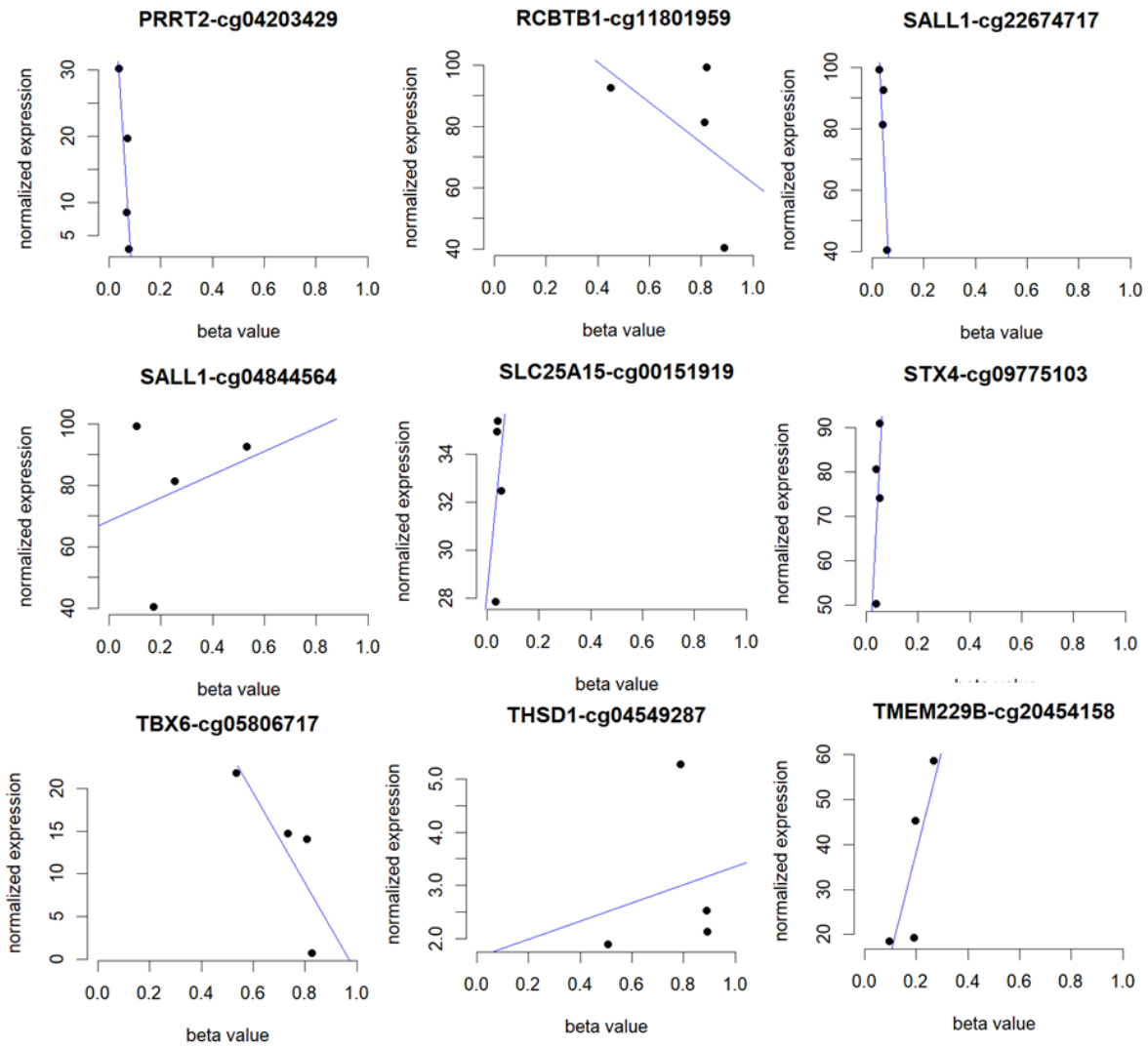
7.3 Supplementary Figure 3



Supplementary Figure 3. Density of segment means of the primary tumor (A) and metastatic (B) samples (separately because otherwise the density of the metastatic samples was not seen). In both cases, the density is the expected, and the peaks of the segment mean are 0.

7.4 Supplementary Figure 4





Supplementary Figure 4. Individual correlation between gene expression and DMCs methylation of the final metastatic gene set in the metastatic samples. Blue line represents the lineal correlation.

8. BIBLIOGRAPHY

[1] Siegel, R.L., Miller, K.D., Fuchs, H.E. & Jemal, A., "Cancer statistics," *Cancer J Clin,* pp. 7-33, 2022.

[2] Howlader, N.e.a. , "SEER Cancer Statistics Review (CSR) 1975-2016," *National Cancer Institute,* 2019.

[3] Howlader N, N.A., Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds), "SEER Cancer Statistics Review, 1975-2018,," *National Cancer Institute,* 2021.

[4] "Survival Rates for Breast Cancer. American Cancer Society.," 2022. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html..

[5] Eng, L.G., et al., "Ten-year survival in women with primary stage IV breast cancer.," *Breast Cancer Res Treat ,* vol. 160, pp. 145-152, 2016.

[6] Perou, C.M., et al, "Molecular portraits of human breast tumours.," *Nature ,* vol. 406, pp. 747-752, 2000.

[7] Prat, A., et al. , "Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.," *Breast Cancer Res ,* vol. 12 , p. R68, 2010.

[8] Sorlie, T., et al. , "Repeated observation of breast tumor subtypes in independent gene expression data sets.," *Proc Natl Acad Sci U S A ,* vol. 100, pp. 8418-8423 , 2003.

[9] Rakha, E.A. & Pareja, F.G., " New Advances in Molecular Breast Cancer Pathology.," *Semin Cancer Biol ,* Vols. 72,, pp. 102-113, 2021.

[10] Sorlie, T., et al. , "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.," *Proc Natl Acad Sci U S A ,* vol. 98, pp. 10869-10874 , 2001.

[11] Malhotra, G.K., Zhao, X., Band, H. & Band, V. , "Histological, molecular and functional subtypes of breast cancers.," *Cancer Biol Ther ,* vol. 10, pp. 955-960, 2010.

[12] Makki, J. , "Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance.," *Clin Med Insights Pathol ,* vol. 8, pp. 23-31, 2015.

[13] Cserni, G., "Histological type and typing of breast carcinomas and the WHO classification changes over time," *Pathologica,* vol. 112, pp. 25-41, 2020.

[14] Tao, Z., et al. , "Breast Cancer: Epidemiology and Etiology.," *Cell Biochem Biophys ,* vol. 72, pp. 333-338, 2015.

[15] Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. , "Multi-omics Data Integration, Interpretation, and Its Application.," *Bioinform Biol Insights ,* Vols. 14, , 2020.

[16] Chakraborty, S., Hosen, M.I., Ahmed, M. & Shekhar, H.U. , "Onco-Multi-OMICS Approach: A New Frontier in Cancer Research.," *Biomed Res Int 2018,* 2018.

[17] de Anda-Jauregui, G. & Hernandez-Lemus, E. , "Computational Oncology in the Multi-Omics Era: State of the Art.," *Front Oncol ,* vol. 10, p. 423, 2020.

[18] Heo, Y.J., Hwa, C., Lee, G.H., Park, J.M. & An, J.Y. , "Integrative Multi-Omics Approaches in Cancer

Research: From Biological Networks to Clinical Subtypes.," *Mol Cells,* vol. 44, pp. 433-443, 2021.

[19] Lu, M. & Zhan, X. , "The crucial role of multiomic approach in cancer research and clinically relevant outcomes.," *EPMA J ,* vol. 9, pp. 77-102 , 2018.

[20] "Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours.," *Nature,* vol. 490, pp. 61-70, 2012.

[21] Sammut, S.J., et al. , "Multi-omic machine learning predictor of breast cancer therapy response.," *Nature ,* vol. 601, pp. 623-629, 2022.

[22] "Computing, R.C.T.R.F.f.S. R: a language and environment for statistical computing.," 2017. [Online].

[23] Bernard Pereira, Suet-Feung Chin, Oscar M Rueda 1 2, Hans-Kristian Moen Vollan 3 4, Elena Provenzano 5 6, Helen A Bardwell 1, Michelle Pugh 7, Linda Jones 5 6, Roslin Russell 1, Stephen-John Sammut 1 2, Dana W Y Tsui 1, Bin Liu 2, Sarah-Jane Dawso, "The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes," *Nat Commun,* 2016.

[24] "TCGA-BRCA. Breast Invasive Carcinoma.," [Online]. Available: https://portal.gdc.cancer.gov/projects/TCGA-BRCA..

[25] G. Frisk, T. Svensson, L.M. Backlund, E. Lidbrink, P. Blomqvist, K.E. Smedby, "Incidence and time trends of brain metastases admissions among breast cancer patients in Sweden," *Br J Cancer,* vol. 106, pp. 1850-1853, 2012.

[26] E.M. Pelletier, B. Shim, S. Goodman, M.M. Amonkar, "Epidemiology and economic burden of brain metastases among patients with primary breast cancer: results from a US claims data analysis," *Breast Cancer Res Treat,* vol. 108, pp. 297-305, 2008.

[27] F. Meric-Bernstam, X. Zheng, M. Shariati, S. Damodaran, C. Wathoo, L. Brusco, et al., "Survival outcomes by TP53 mutation status in metastatic breast cancer," *JCO Precis Oncol,* 2018.

[28] L.R. Yates, S. Knappskog, D. Wedge, J.H.R. Farmery, S. Gonzalez, I. Martincorena, et al., "Genomic evolution of breast cancer metastasis and relapse," *Cancer Cell,* vol. 32, pp. 169-184, 2017.

[29] L. Angus, M. Smid, S.M. Wilting, J. van Riet, A. Van Hoeck, L. Nguyen, et al., "The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies," *Nat Genet,* vol. 51, pp. 1450-1458, 2019.

[30] Mario Cioce, Andrea Sacconi, Sara Donzelli, Claudia Bonomo, Letizia Perracchio, Mariantonia Carosi, Stefano Telera, Vito Michele Fazio, Claudio Botti, Sabrina Strano, Giovanni Blandino, "Breast cancer metastasis: Is it a matter of OMICS and proper ex-vivo models?," *Elsevier,* vol. 20, pp. 4003-4008, 2022.

[31] Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, Raffaele Ferrari, "Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences," *Brief Bioinform,* vol. 19, pp. 286-302, 2018.

[32] Maria A.Wörheidea, Jan Krumsiek, Gabi Kastenmüller, Matthias Arnold, "Multi-omics integration in biomedical research – A metabolomics-centric review," *Analytica Chimica Acta,* vol. 1141, pp. 144-162, 2021.

[33] Ryuji Hamamoto, Masaaki Komatsu, Ken Takasawa, Ken Asada and Syuzo Kaneko, "Epigenetics Analysis and Integrated Analysis of Multiomics Data, Including Epigenetic Data, Using Artificial

Intelligence in the Era of Precision," *biomolecules,* 2019.

[34] Bilal Aslam, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, Muhammad Hidayat Rasool, "Proteomics: Technologies and Their Applications," *J Chromatogr Sci,* vol. 55, pp. 182-196, 2017.

[35] "Next-generation Interactomics: Considerations for the Use of Co-elution to Measure Protein Interaction Networks," *Daniela Salas, R Greg Stacey, Mopelola Akinlaja, Leonard J Foster ,* vol. 19, pp. 1-10, 2020.

[36] Rachel Cavill, Danyel Jennen, Jos Kleinjans, Jacob Jan Briedé, "Transcriptomic and metabolomic data integration," *Briefings in Bioinformatics,* vol. 17, p. 891–901, 1016.

[37] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika, "Multi-omics Data Integration, Interpretation, and Its Application," *Bioinform Biol Insights,* vol. 14, 2020.

[38] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani & Luciano Milanesi , "Methods for the integration of multi-omics data: mathematical aspects," *BMC Bioinformatics,* vol. 15, 2016.

[39] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani & Luciano Milanesi , "Methods for the integration of multi-omics data: mathematical aspects," *BMC Bioinformatics,* vol. 17, 2016.

[40] Ji Zhang and Hai Fang, Using Self-Organizing Maps to Visualize, Filter and Cluster Multidimensional Bio-Omics Data, Intechopen, 2012.

[41] Hans Binder and Henry Wirth, Analysis of large-scale OMIC data using Self Organizing Maps, Encyclopedia of Information Science and Technology, Third Edition, 2015.

[42] Nimrod Rappoport and Ron Shamir, "Multi-omic and multi-view clustering algorithms: review and cancer benchmark," *Nucleic Acids Res.,* vol. 46, p. 10546–10562., 2018.

[43] Animesh Acharjee, Bjorn Kloosterman, Richard G F Visser, Chris Maliepaard, "Integration of multi-omics data for prediction of phenotypic traits using random forest," *BMC Bioinformatics,* 2016.

[44] Ian T Jolliffe 1, Jorge Cadima , "Principal component analysis: a review and recent developments," *Philos Trans A Math Phys Eng Sci,* vol. 13, 2016.

[45] M Henningsson 1, E Sundbom, B A Armelius, P Erdberg, "PLS model building: a multivariate approach to personality test data," *Scand J Psychol,* vol. 42, pp. 399-409, 2001.

[46] ZhaoxiangCai, Rebecca C.Poulos, JiaLiu, QingZhong, "Machine learning for multi-omics data integration in cancer," *iScience,* vol. 25, 2022.

[47] J Craig Venter, Hamilton O Smith, Mark D Adams, "The Sequence of the Human Genome," *Clin Chem,* vol. 61, pp. 1207-8, 2015.

[48] Can Alkan, Bradley P Coe, Evan E Eichler, "Genome structural variation discovery and genotyping," *Nat Rev Genet,* vol. 12, pp. 363-76, 2011.

[49] John M. Butler, "Single Nucleotide Polymorphisms and Applications," *Emery and Rimoin's Principles and Practice of Medical Genetics (Sixth Edition),* vol. Chapter 12, 2013.

[50] Francis Robert and Jerry Pelletier, "Exploring the Impact of Single-Nucleotide Polymorphisms on Translation," *Frontiers Genetics,* vol. 30, 2018.

[51] Jennifer L Freeman, George H Perry, Lars Feuk, Richard Redon, Steven A McCarroll, David M Altshuler, Hiroyuki Aburatani, Keith W Jones, Chris Tyler-Smith, Matthew E Hurles, Nigel P Carter, Stephen W Scherer, Charles Lee, "Copy number variation: new insights in genome diversity," *Genome Res.,* vol. 16, pp. 949-61, 2006.

[52] Mehdi Zarrei, Jeffrey R MacDonald, Daniele Merico, Stephen W Scherer, "A copy number variation map of the human genome," *Nat Rev Genet,* vol. 16, p. 172–183, 2015.

[53] Nigel P Carter, "Methods and strategies for analyzing copy number variation using DNA microarrays," *Nat Genet,* vol. 39, pp. 16-21, 2007.

[54] Mark Schenadari Shalonronald W. Davisand Patrick O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *SCIENCE,* vol. 270, pp. 467-470, 1995.

[55] Chiao-Feng Lin, Adam C Naj, and Li-San Wang, "Analyzing Copy Number Variation using SNP Array Data: Protocols for Calling CNV and Association Tests," *Curr Protoc Hum Genet.,* vol. 79, p. Unit–1.27, 2013.

[56] Danh V Nguyen, A Bulak Arpat, Naisyin Wang, Raymond J Carroll, "DNA microarray experiments: biological and technological aspects," *Biometrics,* vol. 58, pp. 701-17, 2002.

[57] F Sanger, G M Air, B G Barrell, N L Brown, A R Coulson, C A Fiddes, C A Hutchison, P M Slocombe, M Smith, "Nucleotide sequence of bacteriophage phi X174 DNA," *Nature,* vol. 265, pp. 687-95, 1977.

[58] F Sanger, S Nicklen, A R Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc Natl Acad Sci U S A,* vol. 74, pp. 5463-7, 1977.

[59] E S Lander 1, L M Linton, B Birren, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature,* vol. 409, pp. 860-921, 2001.

[60] J C Venter et al., "The sequence of the human genome," *Science,* vol. 291, pp. 1304-51, 2001.

[61] Scott J Emrich 1, W Brad Barbazuk, Li Li, Patrick S Schnable, "Gene discovery and annotation using LCM-454 transcriptome sequencing," *Genome Res,* vol. 17, pp. 69-73, 2007.

[62] Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, Joseph R Ecker, "Highly integrated single-base resolution maps of the epigenome in Arabidopsis," *Cell,* vol. 133, pp. 523-36, 2008.

[63] Erwin L van Dijk, Hélène Auger, Yan Jaszczyszyn, Claude Thermes, "Ten years of next-generation sequencing technology," *Trends Genet,* vol. 30, pp. 418-26, 2014.

[64] Min Zhao, Qingguo Wang, Quan Wang, Peilin Jia & Zhongming Zhao, "Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives," *BMC Bioinformatics,* vol. 14, p. S1 , 2013.

[65] Aquillah M. Kanzi*, James Emmanuel San, Benjamin Chimukangara, Eduan Wilkinson, Maryam Fish, Veron Ramsuran and Tulio de Oliveira, "Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance," *Front. Genet.,* p. Sec. Human and Medical Genomics, 2020.

[66] Berg, J.M., Tymoczko, J.L. and Stryer, L., Biochemistry 5th Edition, New York.: W. H. Freeman Publishing, 2002.

[67] John S Mattick, Igor V Makunin, "Non-coding RNA," *Hum Mol Genet,* vol. 15, pp. R17-29, 2006.

[68] Alex Sánchez-Pla, Ferran Revertera M. Carme Ruíz de Villa, Manuel Comabella, "Transcriptomics: mRNA and alternative splicing," *Journal of Neuroimmunology,* vol. 248, no. Issues 1–2, pp. 23-31, 2012.

[69] Kevin C. Wang and Howard Y. Chang, "Epigenomics," *Circulation Research,* vol. 112, no. Issue 9, p. 1191–1199, 2018.

[70] E R Gibney and C M Nolan, "Epigenetics and gene expression," *Heredity volume,* vol. 105, p. 4–13 , 2010.

[71] T M Nafee, W E Farrell, W D Carroll, A A Fryer, K M K Ismail, "Epigenetic control of fetal gene expression," *BJOG,* vol. 115, pp. 158-68, 2008.

[72] Lisa D Moore, Thuc Le, Guoping Fan, "DNA methylation and its basic function," *Neuropsychopharmacology,* vol. 38, pp. 23-38, 2013.

[73] Sergey Kurdyukov, and Martyn Bullock, "DNA Methylation Analysis: Choosing the Right Method," *Biology (Basel),* 2016.

[74] Yuanyuan Li and Trygve O. Tollefsbol, "DNA methylation detection: Bisulfite genomic sequencing analysis," *Methods Mol Biol. ,* vol. 791, pp. 11-21, 2011.

[75] Antonio Colaprico, Tiago C. Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S. Sabedot, Tathiane M. Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi, Houtan Noushmehr, "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data," *Nucleic Acids Research,,* vol. 44, no. 8, p. 71, 2016.

[76] Charity W Law, Yunshun Chen, Wei Shi & Gordon K Smyth, "voom: precision weights unlock linear model analysis tools for RNA-seq read counts," *Genome Biology,* vol. R29, 2014.

[77] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, Gordon K Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res,* vol. 43, p. e47, 2015.

[78] Mark D. Robinson,1,2,*† Davis J. McCarthy,corresponding author2,† and Gordon K. Smyth2, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics.,* vol. 26, p. 139–140, 2010.

[79] Levi Waldron, Sean Davis, Marcel Ramos, Lori Shepherd, Martin Morgan., "The Bioconductor 2018 Workshop Compilation. 6 200: RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR," 30 July 2018. [Online]. Available: https://bioconductor.github.io/BiocWorkshops/..

[80] Benjamin P. Berman, Tiago C. Silva, "Analyzing and visualizing TCGA data," 2019. [Online]. Available: https://bioconductor.org/packages/devel/bioc/vignettes/TCGAbiolinks/inst/doc/analysis.html.

[81] Tiago Maié, Martin Manolov, "Analysis of Cancer Genome Atlas in R," 20 November 2020. [Online]. Available: https://www.costalab.org/wp-content/uploads/2020/11/R_class_D3.html#2_TCGA_data.

[82] Yi-an Chen, Mathieu Lemire, Sanaa Choufani, Darci T. Butcher, Daria Grafodatskaya, Brent W. Zanke, Steven Gallinger, Thomas J. Hudson and Rosanna Weksberg, "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray," *Epigenetics,* vol. https://doi.org/10.4161/epi.23470, pp. 203-209, 2013.

[83] Yunshun Chen, Bhupinder Pal, Jane E. Visvader, Gordon K. Smyth, "Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR," 8 Oct 2018. [Online]. Available: https://f1000research.com/articles/6-2055/v2.

[84] Tiago Chedraoui Silva, Antonio Colaprico2,3, Catharina Olsen, Fulvio D'Angelo, Gianluca Bontempi, Michele Ceccarelli6,7 and Houtan Noushmehr, "TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages," 07 January 2022. [Online]. Available: http://bioconductor.org/packages/release/workflows/vignettes/TCGAWorkflow/inst/doc/TCGAWorkflow.html#Epigenetic_analysis.

[85] Jovana Maksimovic*, Belinda Phipson and Alicia Oshlack, "A cross-package Bioconductor workflow for analysing methylation array data," 29 April 2022. [Online]. Available: https://www.bioconductor.org/packages/release/workflows/vignettes/methylationArrayAnalysis/inst/doc/methylationArrayAnalysis.html.

[86] [Online]. Available: https://bioconductor.org/packages/release/data/annotation/html/IlluminaHumanMethylation450kanno.ilmn12.hg19.html.

[87] [Online]. Available: https://www.bioconductor.org/packages/release/bioc/vignettes/gaia/inst/doc/gaia.pdf.

[88] Tiago C. Silva, Antonio Colaprico, Catharina Olsen, Fulvio D'Angelo, Gianluca , Michele Ceccarelli, "TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages," F1000 Research, 2016. [Online]. Available: https://f1000research.com/articles/5-1542.

[89] [Online]. Available: https://github.com/yanlinlin82/ggvenn.

[90] [Online]. Available: https://github.com/hamidghaedi/Methylation_Analysis.

[91] [Online]. Available: https://www.gsea-msigdb.org/gsea/msigdb/.

[92] [Online]. Available: http://apps.cytoscape.org/apps/enrichmentmappipelinecollection.

[93] Lahdesmaki, H., Hao, X., Sun, B., Hu, L. et al.,, "Distinguishing key biological pathways between primary breast cancers and their lymph node metastases by gene function-based clustering analysis," *Int. J. Oncol. ,* vol. 24, p. 1589–1596, 2004.

[94] Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M. et al., "Molecular portraits of human breast tumours," *Nature,* vol. 406, p. 747–752., 2000.

[95] Feng, Y., Sun, B., Li, X., Zhang, L. et al.,, "Differentially expressed genes between primary cancer and paired lymph node metastases predict clinical outcome of node-positive breast cancer patients," *Breast Cancer Res,* vol. 103, pp. 319-329, 2007.

[96] Bernardo P. de Almeida, Joana Dias Apolónio, Alexandra Binnie & Pedro Castelo-Branco, "Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers," *BMC Cancer,* vol. 19, 2019.

[97] Yan-Ni Cao, Qian-Zhong Li, Yu-Xian Liu, Wen Jin, and Rui Hou, "Discovering the key genes and important DNA methylation regions in breast cancer," *Hereditas,* vol. 159, 2022.

[98] Wanxue Xu, Mengyao Xu, Longlong Wang, Wei Zhou, Rong Xiang, Yi Shi, Yunshan Zhang & Yongjun Piao , "Integrative analysis of DNA methylation and gene expression identified cervical cancer-

specific diagnostic biomarkers," *Nature,* vol. 13, 2019.

[99] Xin Shao, Ning Lv, Jie Liao, Jinbo Long, Rui Xue, Ni Ai, Donghang Xu & Xiaohui Fan, "Copy number variation is highly correlated with differential gene expression: a pan-cancer study," *BMC Medical Genetics,* vol. 20, 2019.

[100] Keiichi Ohshima, Keiichi Hatakeyama, Takeshi Nagashima, Yuko Watanabe, Kaori Kanto, Yuki Doi, Tomomi Ide, Yuji Shimoda, Tomoe Tanabe, Sumiko Ohnami, Shumpei Ohnami, Masakuni Serizawa, Koji Maruyama, Yasuto Akiyama, , "Integrated analysis of gene expression and copy number identified potential cancer driver genes with amplification-dependent overexpression in 1,454 solid tumors," *Sci Rep.,* vol. 641, 2017.

[101] Rameen Beroukhim etal., "The landscape of somatic copy-number alteration across human cancers," *Nature,* vol. 463, p. 899–905. , 2010.

[102] Susana Romero-Garcia, Heriberto Prado-Garcia and Angeles Carlos-Reyes, "Role of DNA Methylation in the Resistance to Therapy in Solid Tumors," *Frontiers,* 2020.

[103] Melanie Ehrlich, "DNA hypomethylation in cancer cells," *Future Medicine,* 2009.

[104] Chunxiao Liu, Yuting Xu, Xu Liu, Yingqiang Fu, Kaiyuan Zhu, Zhenbo Niu, Jiaxin Liu, and Cheng Qiancorresponding author, "Upregulation of LINC00511 expression by DNA hypomethylation promotes the progression of breast cancer," *Gland Surg,* vol. 10, p. 1418–1430., 2021.

[105] Gary C. Hon, R. David Hawkins, Otavia L. Caballero, Christine Lo, Ryan Lister, Mattia Pelizzola,4 Armand Valsesia,5 Zhen Ye,1 Samantha Kuan,1 Lee E. Edsall, Anamaria Aranha Camargo, Brian J. Stevenson, Joseph R. Ecker, Vineet Bafna, Robert L. St, "Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer," *Genome Res,* vol. 22, 2012.

[106] Ehrlich M, "DNA hypomethylation in cancer cells," *Epigenomics,* 2009.

[107] Hon G.C., Hawkins R.D., Caballero O.L., Lo C., Lister R., Pelizzola M., Valsesia A., Ye Z., Kuan S., Edsall L.E. et al., "Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer.," *Genome Res,* vol. 22, p. 246–258, 2012.

[108] Xiufang Xu, Miaofeng Zhang, Faying Xu, Shaojie Jiang, "Wnt signaling in breast cancer: biological mechanisms, challenges and opportunities," *Mol Cancer,* vol. 165, 2020.

[109] Pradip De, Jennifer H. Carlson, Hui Wu, Adam Marcus, Brian Leyland-Jones, Nandini Dey, "Wnt-beta-catenin pathway signals metastasis-associated tumor cell phenotypes in triple negative breast cancers," *Oncotarget,* vol. 7, 2016.

[110] Wenjun Guo and Filippo G. Giancotti, "Integrin Signaling During Tumour Progression," *Nature,* vol. 5, 2004.

[111] Takeshi Kawauchi, "Cell Adhesion and Its Endocytic Regulation in Cell Migration during Neural Development and Cancer Metastasis," vol. 13, pp. 4564-4590, 2012.

[112] Genya Gorshtein, Olivia Grafinger, and Marc G. Coppolino , "Targeting SNARE-Mediated Vesicle Transport to Block Invadopodium-Based Cancer Cell Invasion," *Front Oncol. ,* vol. 11, 2021.

[113] Jianghui Meng et al., "Role of SNARE proteins in tumorigenesis and their potential as targets for novel anti-cancer therapeutics," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer,* 2015.

[114] Nataly Naser Al Deen, Mounir G Abouhaidar, Rabih Talhouk, "Connexin43 as a Tumor Suppressor: Proposed Connexin43 mRNA-circularRNAs-microRNAs Axis Towards Prevention and Early Detection in Breast Cancer," *Frontiers in Medicine,* vol. 192, 2019.

[115] Pierina Cetraro, Julio Plaza-Diaz, Alex MacKenzie, and Francisco Abadía-Molina, "A Review of the Current Impact of Inhibitors of Apoptosis Protein and Their Repression in Cancer," *Cancers,* 2022.

[116] Benjamin D. Lee ,Anthony Gitter,Casey S. Greene,Sebastian Raschka,Finlay Maguire,Alexander J. Titus,Michael D. Kessler,Alexandra J. Lee,Marc G. Chevrette,Paul Allen Stewart,Thiago Britto-Borges,Evan M. Cofer,Kun-Hsing Yu,Juan Jose Carmona,Elana J. Fertig,, "Ten quick tips for deep learning in biology," *PLOS Computational biology,* vol. e1009803. https://doi.org/10.1371/journal., 2022.

[117] Charles Y. Chiu & Steven A. Miller , "Clinical metagenomics," *Nature Reviews Genetics,* vol. 20, p. 341–355, 2019.

[118] Stefania Morganti, Paolo Tarantino, Emanuela Ferraro, Paolo D'Amico, Bruno Achutti Duso , Giuseppe Curigliano, "Next Generation Sequencing (NGS): A Revolutionary Technology in Pharmacogenomics and Personalized Medicine in Cancer," *Adv Exp Med Biol,* vol. 1168, pp. 9-30, 20189.

[119] Shunichi Kosugi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo & Yoichiro Kamatani, "Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing," *Genome Biology ,* vol. 20, no. 117, 2019.

[120] Masahiro Nakatochi, Itaru Kushima & Norio Ozaki , "Implications of germline copy-number variations in psychiatric disorders: review of large-scale genetic studies," *Journal of Human Genetics,* vol. 66, p. 25–37, 2021.

[121] Malhotra, G.K., Zhao, X., Band, H. & Band, V. , "Histological, molecular and functional subtypes of breast cancers.," *Cancer Biol Ther,* vol. 10, pp. 955-960, 2010.

[122] Makki, J., "Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance," *Clin Med Insights Pathol,* vol. 8, pp. 23-31, 2015.

[123] Burstein, H.J., Polyak, K., Wong, J.S., Lester, S.C. & Kaelin, C.M. , "Ductal carcinoma in situ of the breast," *N Engl J Med ,* vol. 350, pp. 1430-1441, 2004.

[124] Polyak, K. et al., "Breast cancer: origins and evolution.," *J Clin Invest ,* vol. 117, pp. 3155-3163, 2007.