SAPIENZA
UNIVERSITÀ DI ROMA

PhD Thesis

An intelligent management of integrated biomedical data for digital health via Network Medicine and its application to different human diseases

Department of Translational and Precision Medicine

PhD course in Innovative Biomedical Technologies in Clinical Medicine

Candidate

Pasquale Sibilio

Advisor

Prof. Paola Paci

Cycle XXXV

Acknowledgments

I would like to thank my mentor, professor Paola Paci, that gave me the possibility to spend my PhD training in her laboratory and to make me improve as scientist and as a person. I would like to thank my colleagues Federica Conte and Giulia Fiscon that gave me precious help during all these three years. Moreover, I want to thank Professor Lorenzo Farina for our collaborations and for communicating with me his brilliant view about science. I'm so grateful to Professor Edwin Silverman and Professor Dawn DeMeo that hosted me during my visiting at the Channing division of Network Medicine Division at Brigham and Women's Hospital, it was an outstanding experience to work with them. Markedly, I would like to thank professor Giuseppe Giannini, professor Sebastiano Filetti, and professor Cosimo Durante that gave me the possibility to work with them. Further, I'm grateful for the fruitful collaboration with Doctor Antonella Verrienti, Doctor Francesca Belardinilli, and Doctor Valerio Licursi. Then, I want to thank the PhD coordinator professor Marcello Arca, the PhD committee and the Department of Translational and Precision Medicine for their precious work over these three years to organize the PhD.

Abstract

Personalized medicine aims to tailor the health care to each person's unique signature leading to better distinguish an individual patient from the others with similar clinical manifestation. Many different biomedical data types contribute to define this patient's unique signature, such as omics data produced trough next generation sequencing technologies. The integration of single-omics data, in a sequential or simultaneous manner, could help to understand the interplay of the different molecules thus helping to bridge the gap between genotype and phenotype. To this end, Network Medicine offers a promising formalism for multi-omics data integration by providing a holistic approach that look at the whole system at once rather than focusing on the single entities. This thesis regards the integration of various omics data following two different procedures within the framework of Network Medicine: A procedural multi-omics data integration, where a single omics was first selected to perform the main analysis, and then the other omics were used in cascade to molecularly characterize the results obtained in the main analysis. A parallel multi-omics data integration, where the result was given by the intersection of the results of each single-omics. The procedural multi-omics data integration was leveraged to study Colorectal and Breast Cancer. In the Colorectal Cancer case study, we defined the molecular signatures of a new subgroup of Colorectal Cancer possibly eligible for immune-checkpoint inhibitors therapy. Moreover, in the Breast Cancer case study we defined 11 prognostic biomarkers specific for the Basal-like subtype of Breast Cancer. Instead, the parallel multi-omics data integration was exploited to study COVID-19 and Chronic Obstructive Pulmonary Disease. In the COVID-19 case study, we defined a pool of drugs potentially repurposable for COVID-19. Whereas, in the Chronic Obstructive Pulmonary Disease case study, we discovered a group of differentially expressed and methylated genes that have a considerable biological specificity and could be related to the inflammatory pathological mechanism of Chronic Obstructive Pulmonary Disease.

Contents

Introduction

The application of digital technologies for improving public health remains a largely unexplored territory yet, making the discovery of novel digital health solutions a mandatory task. A big challenge in this perspective is the personalization of medicine that should be aimed to improve the patients' clinical outcomes and to reduce the drugs' side effects by using a "bench-to-bedside" approach [1]. This new practice of medicine should tailor the health care to each person's unique signature leading to better distinguish an individual patient from the others with similar clinical manifestation. Many different biomedical data types contribute to define this patient's unique signature, including, but not limited to genomics, transcriptomics, proteomics, and metabolomics data. The integration of these individual omics data, in a sequential or simultaneous manner, could help to understand the interplay of the different molecules thus helping to bridge the gap between genotype and phenotype. Despite their power and promise, a variety of challenges must be considered in the successful design and execution of a multi-omics study, including the complexity of the biological systems as well as the ever-increasing amount of the biological data available from the quickly maturing field of the next generation sequencing (NGS) technologies. By providing an holistic approach that look at the whole system at once rather than focusing on the single entities, network theory offers a promising formalism for multi-omics data integration [2].

Network Medicine is a quickly maturing discipline that studies holistic relationships between various biological components by combining network theory and systems biology. The basic premise of this exercise is that no gene or gene product exerts its effect on phenotype in isolation. Investigating the molecular context (i.e., the network of the functional and molecular interactions within a cell) is essential for understanding the true bases for phenotype and pathophenotype [3].

This thesis regards the integration of various omics data following two different procedures within the framework of Network Medicine:

- a sequential approach (called procedural multi-omics data integration), where a single-omics was first selected to perform the main analysis, and then the other omics were used in cascade to molecularly characterize the results obtained in the main analysis;

- a simultaneous approach (called parallel multi-omics data integration), where the single-omics were analyzed in parallel, and the result was given by the intersection of the results of each single-omics.

For what concerns the procedural multi-omics data integration, we applied the analysis on two different human diseases:

- Colorectal Cancer (CRC), where we started from Copy Number Variations (CNVs) and Tumor Mutational Burden (TMB) data available on The Cancer Genome Atlas (TCGA) [4] and we applied an unsupervised learning technique to better classify CRC patients. Afterwards, we exploited DNA methylation, Single Nucleotide Variation (SNVs), and transcriptomic data from TCGA to define the molecular signatures of CRC subgroups possibly eligible for immune-checkpoint inhibitors therapy.

- Breast Cancer (BC), where we started from the survival data available on TCGA and we applied a Kaplan-Meier survival analysis to narrow the list of the predicted Basal-like specific biomarkers obtained from [5], to those that showed a statistically significant prognostic value. Eventually, we exploited the CNVs, DNA methylation, and transcriptional regulatory data to investigate whether variations in the expression of identified prognostic genes could be related to genetic (CNVs), epigenetic (DNA methylation differences), and transcription factor activities.

For what concerns the parallel multi-omics data integration, we applied the analysis on two different human diseases:

- COVID-19, where we used transcriptomics and interactomics data, and we applied a network-based drug repurposing analysis to identify novel uses for existing drugs that

can be repurposed outside the scope of their original medical indication for treating COVID-19 [6]. Transcriptomics data were obtained from GEO repository [7] and interactomics data from the supplementary material of [8].

- Chronic Obstructive Pulmonary Disease (COPD), where we used transcriptomics and DNA methylation data, and we applied a network-based integration analysis to build a *consensus network* of genes that are differentially modulated both in their expression and methylation profile. Data were retrieved from a lung tissue cohort of the Lung Tissue Research Consortium at the Channing division of Network Medicine Division at Brigham and Women's Hospital.

This thesis is structured in the following chapters:

- Chapter 1: Multi-omics data integration. This chapter describes the state of art of methods that perform multi-omics data integration.

- Chapter 2: Network Medicine: a new paradigm for personalized medicine. This chapter summarizes the main hypotheses, the biological networks, and the tools widely used in Network Medicine field.

- Chapter 3: Procedural multi-omics data integration for studying Colorectal Cancer. This chapter details the computational analysis developed to discover a novel molecular subset of CRC patients possibly eligible for immune checkpoint immunotherapy.

- Chapter 4: Procedural multi-omics data integration for studying Breast Cancer. This chapter details the computational analysis developed to discover new putative prognostic biomarkers for the Basal-like subtype of BC.

- Chapter 5: Parallel multi-omics integration to identify repurposable drugs for COVID-19. This chapter details the network-based drug repurposing analysis implemented to discover a pool of drugs potentially repurposable for COVID-19.

- Chapter 6: Parallel multi-omics data integration for studying COPD. This chapter details the network-based integration analysis developed to gain insights about the pathobiological mechanism of the disease.

Chapter1: Multi-omics data integration

1.1. Introduction

The comprehension of molecular mechanisms of human diseases requires the study of the biomolecular complexity at multiple levels such as genome, epigenome, transcriptome, proteome, and metabolome. The advent of NGS high-throughput technologies made biology more quantitative and strongly dependent on data generated at these levels, which together is called "multi-omics" data. Nowadays, the integration of these multi-omics data seems to be an unavoidable step to a better understanding of the molecular mechanisms underlying the diseases that can eventually aid in better treatment and prevention. The largest publicly available database providing multi-omics data is The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types [4]. The high-throughput data stored in TCGA include RNA-seq, DNA-seq, miRNA-seq, single-nucleotide variant (SNV), copy number variation (CNV), DNA methylation arrays, and reverse phase protein array (RPPA) data. Further, it contains also clinical and histological data, which enlarge the possibility to integrate omics and clinical information.

1.2. Methodologies for multi-omics data integration

In a recent study [2], different methods that allow integration of multi-omics data are discussed and organized based on their ability to address biological question of interest, broadly categorized into three main case studies:

- **Disease subtyping and classification based on multi-omics profiles**

  Complex diseases are characterized by heterogeneity in the etiology, disease progression, and therapeutic response in affected individuals. Additionally,

environment and life factors can be an important contribution to disease heterogeneity. Hence, it is crucial to improve disease classification, identifying novel subtypes or improving the samples classification into known subgroups, to offer different therapeutic strategies for patients belonging to different subtypes, moving toward a more personalized medicine. There exist diverse methodologies to identify disease subtypes or classify patients into subgroups based on their multi-omics profiles. For instance, Multiple co-inertia analysis (MCIA) [9], which employs a multivariate approach. MCIA performs a covariance optimization criterion to transform diverse omics datasets onto the same scale, projecting them into the same dimensional space. Employing graphical representations of the sample space is possible to identify disease subtypes, meanwhile studying the features space graph is possible to select features that are relevant for samples cluster. However, after MCIA analysis is needed additional molecular characterization employing integrative multi-omics analysis to define the molecular properties of the subgroups identified. In chapter 3 of this thesis, an integrative multi-omics approach based on unsupervised hierarchical clustering methodology is described to stratify colorectal cancer patients based on their genomics, epigenomics and transcriptomics profiles. The results of this study can pave the way for an expansion of immunotherapy therapeutic strategies to a novel identified subgroup of colorectal cancer.

- **Prediction of biomarkers for various applications including diagnostics and driver genes for diseases**

Biomarkers are molecular footprints that help to define a specific condition in human diseases. In a systemic view, biomarkers can be interpreted as important actors in crucial biological pathways affected in disease conditions, thus revealing the underlying pathobiology, and helping to guide new therapeutic strategies. Multi-variates and non-parametric statistics are statistical methods widely used to define biomarkers [2]. For example, Multi-Omics Factor Analysis (MOFA) [10], is a multivariate approach for integrating multi-omics data of the same or partially overlapping samples in an

unsupervised fashion. As first step, MOFA infers an interpretable low-dimensional data representation in terms of (hidden) factors, which can be viewed as generalization of principal component analysis (PCA) to multi-omics data. The hidden factors capture major sources of variation across data modalities, which can help the identification of continuous molecular gradient or discrete subgroup of samples. In the downstream analysis MOFA implements the analysis of the absolute loading of the top features of the different omics data on each factor to define the biomarkers of the disease. However, MOFA doesn't exploit survival data to define biomarkers. In chapter 4 of this thesis, a Kaplan-Meier estimator is used to infer prognostic biomarkers of basal-like subtype of breast cancer. Kaplan-Meier is a univariate non-parametric statistical method that allow to calculate how long a studied event (e.g., death, disease progression, etc.) occurred after starting a particular treatment in subjects that not experience the event before the end of the study [11].

- **Deriving insights into disease biology**

The comprehension of the mechanism of the disease is crucial for developing new diagnostic and therapeutic strategies. One of the main challenges in determining the mechanisms of the disease is to infer the regulatory relationship or the coordination between the elements of different omics. Correlation-based approach has been extensively used to address this issue, and a promising tool belonging to this category is CNAmet [12]. The authors have developed a method that integrates Copy Number Variations (CNVs), DNA methylation and mRNA expression data. The primary goal of the tool is to identify genes that are amplified, hypomethylated and upregulated or deleted, hypermethylated and downregulated. The CNAmet algorithm consists of three major steps: 1) the weight calculation step in which expression values are linked to copy number and methylation aberrations; 2) the score calculation step, in which the weight values are combined to a score that indicates genes whose expression alterations are due to changes in DNA methylation and copy number levels; 3) the significance evaluation step, in which corrected p-values of the scores are calculated with a

permutation test [12]. In chapter 6 of this thesis, we designed a network-based approach that builds the correlation networks for each biological layers considered (e.g., DNA-methylation and RNA-seq layer) for the disease of interest. Then, the networks are merged to build a "consensus network", where nodes are gene products that are both differentially methylated and expressed in the pathological condition and highly positive or negative correlated in both layers. The consensus network allows to study the coordination of genes across different biological layers. In chapter 6 of this thesis, we designed a network-based approach that builds the correlation networks for each biological layers considered (e.g., DNA-methylation and RNA-seq layer) for the disease of interest. Then, the networks are merged to build a "consensus network", where nodes are gene products that are both differentially methylated and expressed in the pathological condition and highly positive or negative correlated in both layers. The consensus network allows to study the coordination of genes across different biological layers which could help to define a new mechanism of regulation.

Chapter 2: Network medicine: a new paradigm for personalized medicine

2.1. Network medicine hypotheses and organizing principles

Network Medicine is considered as a promise formalism able to depict the complexity of biological systems and to manage big volume of biological data. It is a discipline that studies holistic relationships between molecular components by combining elements of graph theory, systems biology, and statistical analyses. The main purpose of Network Medicine is to understand the true bases phenotype and pathophenotype of diseases based on the principle that a disease is rarely a consequence of an abnormality in a single gene but reflects the perturbations of the complex intra-cellular and inter-cellular network that links tissue and organ systems [3]. Further, the network-based approaches to human disease can lead to a better understanding of the effects of cellular interconnectedness on disease progression that may lead to the identification of molecular determinant of disease (disease genes) and disease pathways, which, in turn, may offer better targets for drug development. Disease genes refer to genes with mutations that are known to have a phenotypic impact, e.g., sequence alterations that are causal for Mendelian diseases or variants that increase the susceptibility to complex diseases or cancer. As broadly established [3, 13], disease genes have unique, quantifiable characteristics that distinguish them from other genes. From a network perspective, this observation translates into the verification that disease genes are not randomly scattered in the interactome, but, rather, co-localize in specific subnetworks (disease modules). This concept lead to a series of widely used hypotheses and organizing principles that link network structure to biological function and disease [14] that can be summarized as follows:

- The local hypothesis, according to which proteins involved in the same disease have and increased tendency to interact with each other.

- The disease module hypothesis, according to which proteins involved in the same disease show a tendency to cluster in connected subnetworks (of connected components), within which one of them is often much larger than the others (largest connected component).

- The functional coherence hypothesis, according to which genes in a disease module show a tendency to be involved in closely disease – related cellular functions or casual molecular pathways.

- The shared components hypothesis, according to which related diseases are in the same interactome neighborhood from which unrelated diseases are separated.

In this new conceptualization of medicine, the human interactome is viewed as a map and a disease is a local perturbation of this map.

2.2. Biological networks and interaction resources

Cells can be considered as a complex network of macromolecular interactions, whose understanding requires appropriate network selection and analysis. The biological networks and interaction resources broadly used in my studies are:

- **Human interactome**

  The protein-protein interactions (PPIs) network, also known as interactome, is a network where nodes are the proteins, and the links represent the physical molecular interaction occurring between them. This data type is called interactomics. The interactions can be obtained from yeast-2-hybrid assays [15], co-immunoprecipitation [16], literature text-mining [17], 3D structure [18], sequence homology [19] and other sources.

- **Gene Co-expression Networks (GCNs)**

GCNs are networks where nodes are genes and an edge between two genes is drawn based on the calculation of pairwise correlation coefficients between the transcriptomic profile of the two genes (usually Pearson or Spearman coefficients). Usually, a threshold on the correlation is set to highlight highly correlated gene pairs, following two procedures, called hard and soft thresholding approach, respectively. Hard thresholding is used to build an unweighted networks, where the correlations coefficient between nodes below the threshold are suppressed (edge values set to 0), and the correlations coefficient above the threshold are considered (edges values set to 1). This approach encodes gene co-expression using binary information (connected=1, unconnected=0). On the contrary, the soft thresholding approach weighs each connection by a number in [0,1] and thus it is more suitable for building weighted networks.

- **Gene regulatory networks (GRNs)**

GRNs involve the complex interplay of multiple regulatory molecules including Transcription Factors (TFs), miRNA, epigenetic modifiers that model the transcriptional process of encoding genes. In these kinds of networks, nodes are usually both the regulators and their targets, and the links are the interactions occurring between them. Many databases exist that collect experimental evidences or computational predictions about the regulators-target interactions.

- **Drug-Targets interaction networks (DTNs)**

DTNs are bipartite networks where nodes are both drugs and protein targets, and a link occurs between two nodes if the corresponding drug-target interaction has been experimental validated or computational predicted. Many databases exist that collects these kind of informations.

- **Drug-Disease Interaction networks (DDIs)**

DDIs are bipartite networks where nodes are both drugs and diseases, and a link is placed between two nodes based on some association properties that can be computationally predicted and/or experimentally validated.

2.3. Network-based tools

The main network-based approaches exploited in the studies presented in this thesis are summarized as follows:

- **Weighted Gene Co-expression Network Analysis (WGCNA)**

Weighted gene co-expression network analysis (WGCNA) is one of the most commonly employed tool to construct gene co-expression networks across gene expression data, exploring the association between gene networks and phenotypic/clinical traits of interest [20, 21]. It defines co-expression networks as undirected, weighted gene networks, using a soft thresholding approach. To properly define the values of the parameters needed to build the correlation network, a reasonable choice is to select those that guarantees an approximately scale-free network topology, which is the typical topological structure of most biological networks [22, 23]. A scale-free network is a graph characterizing by many low-degree nodes (peripheral nodes) and few high-degree nodes (hub nodes), where the degree has generally been extended to the sum of weights when analyzing weighted networks. Then, WGCNA identifies modules of highly interconnected, or co-expressed, genes within the weighted network by grouping together the most similar nodes. The similarity measure between two nodes is expressed in terms of their direct connection strength as well as connection strengths "mediated" by shared neighbors. The relationship between modules can be studied by correlating the corresponding module eigengenes (MEs). The ME is defined as the first principal component of a given

module and can be considered a representative of the gene expression profiles in that module. For each gene, a measure called module membership (MM) is defined by correlating its gene expression profile with the module eigengene of a given module and can be computed for all input genes (irrespective of their original module membership). If MM of a given gene with respect to a given module is close to 0, that gene is not part of that module. On the other hand, if MM is close to 1 or -1, the gene is highly connected to the genes of that module. The sign of MM encodes whether the gene has a positive or a negative relationship with the module eigengene. Finally, to incorporate external information into the co-expression network, WGCNA makes use of gene significance (GS) measures computed as the correlations between gene expressions and external sample traits. Abstractly speaking, the higher the absolute value of GS of a given gene, the more biologically significant is that gene. The gene significance of 0 indicates that the gene is not significant regarding the biological question of interest. The gene significance can take on positive or negative values.

- **SWIM**

SWIM (SWItch Miner) is a freely downloadable network-based tool, developed both in MATLAB [24] and in R language [25], which predicts important (switch) genes that are strongly associated with drastic changes in cell phenotype. SWIM first computes the differentially expressed genes (DEGs) between two conditions of interest (e.g., normal state versus tumor state) and then builds a GCN by calculating correlations (positive and negative) between the expression profiles of each gene pair. Specifically, SWIM implements a hard thresholding approach to build a GCN where nodes are DEGs, and a link occurs if their expression profiles are highly correlated or anti-correlated (according to a defined threshold). Then, SWIM classifies each network hub (i.e., nodes with degree at least equal to 5 [22]) as date, party, or fight-club on the basis of the Average Pearson Correlation Coefficient (APCC) between its expression profile and that of its first nearest neighbors. Date hubs show a positive and mild APCC value; party hubs show a positive and high APCC value; fight-clubs hubs show a negative

APCC value. To date, the left tail (i.e., negative correlation between gene pairs) of the correlation distribution, and the interpretation of negative edges within a complex network representation of functional connectivity has largely been ignored, apart from the SWIM methodology. To assign a role to each node in the GCN, SWIM firstly searches for clusters (or modules) using the k-means algorithm and evaluates the quality of clusters by minimizing the Sum of the Squared Error (SSE), depending on the distance of each object to its closest centroid. The choice of the number of clusters to be selected can be done referring to the SSE plot (scree plot) computed as a function of the number of clusters. Particularly, a reasonable choice of the number of clusters is suggested by the position of an elbow in the scree plot. Then, SWIM draws the heat cartography map by evaluating two coordinates related to their intra- and inter-modular connections: the clusterphobic coefficient, which measures the links of each node to nodes outside its own cluster; the within-module degree, which measures how "well-connected" each node is within its own cluster. Nodes having much more external than internal links present high values of the clusterphobic coefficient and are called connectors, whereas high values of the within-module degree denote nodes that are hubs within their community and are called local hubs. Switch genes are defined as a subset of fight-club nodes with the following features:

- they are network connectors that mainly interact outside their own cluster
- they are not local hubs
- they are mainly anti-correlated with their interaction partners

Up to now, SWIM has sparked a widespread interest within the scientific community thanks to the promising results obtained through its application in a broad range of phenotype-specific scenarios, spanning from complex diseases [26–31] to grapevine berry maturation [32].

- **SAveRUNNER**

SAveRUNNER (Searching off-lAbel dRUg aNd NEtwoRk) is a freely downloadable network-based tool, developed in R language [33, 34], which generates predictions of drugs that can be used outside their original medical indication for a disease of interest (Drug repurposing) and for optimizing the efficacy of putative validation experiments. The hypothesis underlying SAveRUNNER's methodology is that for a drug to be effective against a disease, its associated targets (drug module) and disease-associated genes (disease module) should be topologically close to each other in the human interactome [8]. SAveRUNNER takes as inputs the human interactome network, the list of disease-associated genes and drug–target interactions, and predicts drug–disease associations by quantifying the interplay between the drug targets and disease-associated proteins in the human interactome via a novel network-based similarity measure (denoted adjusted similarity), which rewards associations between drugs and diseases located in the same network neighborhood. The idea behind is based on the assumption that if a drug and a disease group together it is more likely that the drug can be effectively repurposed for that disease [33, 34]. SAveRUNNER provides a list of predicted/prioritized associations among drugs and diseases in the form of a weighted bipartite drug–disease network, where one set of nodes represents drugs and the other represents diseases. A link between a drug and a disease is made if the corresponding drug targets and disease genes are closer in the interactome than is expected by chance, with an interaction weight based on the adjusted similarity value [33].

Up to now, SAveRUNNER was successfully applied to predict candidate repurposable drugs for COVID-19 [6, 33], Alzheimer's Disease (AD) [35], Amyotrophic Lateral Sclerosis (ALS) [36], Multiple Sclerosis (MS) [37], and BC subtypes [38].

Chapter 3: Procedural multi-omics data integration for studying Colorectal Cancer

An integrative in-silico analysis discloses a novel molecular subset of Colorectal Cancer possibly eligible for immune checkpoint immunotherapy

3.1. Introduction

Colorectal Cancer (CRC) is a major cause of cancer related death worldwide, accounting for approximately 8% of all annually diagnosed cancers [39]. Historically, the molecular classification of CRC was based on the global genomic status, which identified three major groups: tumors with microsatellite instability (MSI; ~ 15% of all CRCs), tumors with chromosomal instability (CIN; ~ 85% of all CRCs) and tumors with a CpG island methylator phenotype (CIMP; ~ 20% of all CRCs) [40]. In MSI tumors, defects of the mismatch repair (MMR) pathway are the leading cause of genetic instability. It can be due to inactivating mutations or to epigenetic silencing by promoter hypermethylation of DNA MMR genes [40], a condition frequently associated to high levels of CpG island methylation and referred to as CIMP-High (CIMP-H, ~ 70–85% of MSI CRCs). Defective DNA MMR (dMMR) leads to reduced restoration of replication errors resulting in the introduction of a high rate of mismatches in microsatellites. The consequent changes in microsatellite lengths may be monitored to classify different phenotypes as microsatellite stable (MSS) or unstable (MSI), which can be further subdivided MSIHigh (MSI-H) or MSI-Low (MSI-L) [40, 41]. Tumors with MSI-H typically display a high rate of point mutations [42, 43], a state referred to as hypermutation (HM). Besides dMMR, the HM phenotype is also related to somatic or germline mutations of POLE and POLD1 genes encoding DNA polymerase epsilon and delta, respectively [41]. CIN tumors instead bear high frequency of copy number variations (CNVs). In almost all cases they are MSS or MSI-L, usually share low mutation rate, and null or low level of CIMP (non-CIMP or CIMP-L) [40, 44]. Over the years, additional molecular classifications beyond CIN, MSI and CIMP were proposed with the aim to dissect the

heterogeneity of CRC for prognostic and predictive intents [45–49]. In example, the Consensus Molecular Subtypes (CMS) Consortium, analyzing CRC expression profiling data from multiple studies, converged on the definition of four main CMSs [48]. Although CMSs have prognostic and therapeutic implications, they have not been translated into clinical routine, yet. With the introduction of immune checkpoint inhibitors (ICIs) for the treatment of metastatic CRC (mCRC), MSI/CIN classification regained momentum as the dMMR/MSI-H condition (~ 2–4% of mCRCs) predicted sensitivity to ICIs in clinical trials, possibly due to both high rate of tumor mutational burden (TMB-H) and high levels of infiltrating lymphocytes typically present in these tumors [41, 50, 51]. Conversely, pMMR-MSS/MSI-L a group, appears resistant to ICIs therapies.

Here, we performed a procedural multi-omics data integration of genomic, epigenomic and transcriptomic data of 520 CRC samples downloaded from The Cancer Genome Atlas (TCGA) data portal. Our analysis provided a step forward toward a better understanding of the differences between MSI/ CIN status by discovering a novel CRC subgroup of patients relevant for therapeutic decisions. They are non-CIN, non-MSI and CIMP-L, and are characterized by KRAS-high/TP53-low mutation rate, distinct mutational signatures, and an inflamed tumor microenvironment.

3.2. Materials and Methods

Data collection and processing

We downloaded genomic, transcriptomic and epigenomic data from TCGA-COAD and READ projects stored on TCGA data portal (https://portal.gdc.cancer.gov/), accessed in November 2020. We performed meta-analysis on 520 TCGA-COAD and READ patients of which copy number variations (CNVs), whole exome sequencing (WES), transcriptomic (RNA-seq), DNA methylation and MSI status data were available. We developed a computational pipeline that includes molecular integrative analysis at genomic, epigenomic and transcriptomic level to

better classify patients affected by CRC. The pipeline is subdivided in steps, described in Fig. 3.1.

Step1: molecular-based CRC subgroups stratification

Tumor mutational burden (TMB) Analysis: Tumor mutational burden (TMB) was calculated dividing the total number of nonsynonymous mutations of every patient per 30 Megabase, which is the average size of the exome. Numbers of nonsynonymous mutations are derived from MAF files retrieved from TCGA resulting from variant analysis of WES experiments on 520 TCGA-COAD and READ patients. According to [43] patients with a TMB higher and lower than 20 per Megabase were classified as HM or non-HM, respectively.

CNV calling and analysis: We performed CNVs calling from segmented mean data employing GISTIC 2.0 which identifies genomic regions that are significantly gained or lost across the 520 TCGA-COAD and READ tumors [52]. The R package copynumber [53] was used to visualize the frequency of gain/loss in the chromosome regions among the CRC's subgroups identified. The association between frequency of CNVs events in the chromosome regions and the CRC's subgroups identified was evaluated using Fisher's exact test.

Step 2: molecular characterization of the subgroups identified

Single Nucleotide Variations (SNVs) data analysis:

The R package maftools [54], which contains functions to perform most used analyses in cancer genomics and to create feature rich customizable visualizations, were used to analyze MAF files of the 520 TCGA-COAD and READ tumors and to address the mutational signatures. We studied top frequently mutated genes discovered in our cohort plus recurrently mutated genes defined in the COSMIC database [55]. The association between different mutational rates in the genes analyzed and HM, HM-like and non-HM groups was evaluated using Fisher's exact test. Further, we performed the analysis of the non-silent mutations existing in POLE exonuclease domain from exon 9–14 in the three subgroups. Other algorithms implemented in the maftools package allowed the extraction of mutational signatures from

MAF files and to compare them with the validated signature present in the COSMIC curated database.

DNA methylation analysis

Data containing β-values from the Illumina Infinium HumanMethylation450 Array were available for 382/520 of the patients enrolled in the study. In pre-processing steps, we filtered out probes containing Single Nucleotide Polymorphisms (SNPs) and designed on X and Y chromosomes. To determine CpG Island Methylator Phenotype (CIMP) status, we first identified the 1000 differentially methylated CpGs between the three groups (ANOVA-like test using limma package) [56]. Afterwards, we computed an unsupervised hierarchical clustering that identified 3 clusters and considered the methylome patterns of the clusters we could assign to cluster 1 to CIMP-Low (CIMP-L), cluster 2 to CIMP-High (CIMP-H) and cluster 3 to non-CIMP (Fig. 3.3). The hierarchical clustering analysis was performed by using "maximum" as clustering distance and "ward.D2" as clustering method.

Step 3: tumor microenvironment inflammation assessment

The RNA-seq data of the 520 CRC patients was leveraged to perform a Weighted gene co-expression network analysis (WGCNA) by using the R package WGCNA [57, 58], and a deconvolution analysis of the quality and quantity of immune infiltrate in the tumoral environment by using the R package ImSig, [59]. ImSig incorporated immune/ inflammatory cells in 7 major classes (B cells, Interferon, Macrophages, Monocytes, Neutrophils, NK cells, T cells) plus 3 additional signatures (Plasma cells, Proliferation and Translation). A correlation cut-off of 0.8 was used, to remove genes that did not exhibit a strong correlation with the ImSig signatures. Furthermore, to assess the statistical significance of the difference of the mean expression of each immune signature in the multiple comparison of the three groups the Tukey's test was used, which is a post-hoc test after ANOVA analysis. In addition, we studied the expression of 79 Immune Checkpoint Genes (ICGs) curated by [60], in our cohort. A differentially expression analysis was performed using the multiple comparison of the three subgroups using Wald test and p-value was adjusted according to the Benjamini–Hochberg

method. Thresholds for FDR < 0.1 and Log2 Fold Change > 0.4 were used to select significant differentially expressed genes.

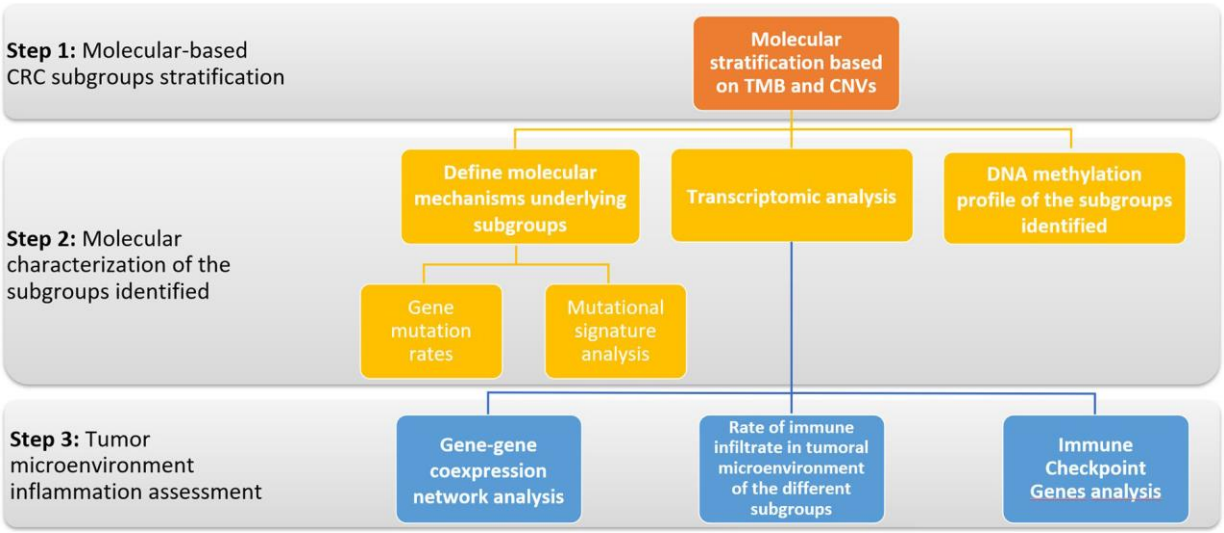

Figure 3.1 Computational pipeline flowchart

3.3. Results

Classification of the CRC samples according to TMB and CNVs

We subjected 520 CRC tumor samples of COAD/READ projects to TMB analysis. 78/520 (15%) samples were classified as hypermutated (HM: TMB > 20 per 106 bases) with a median value of 44.9 mutations per 106 bases (range: 26–347 per 106 bases), while 442/520 (85%) samples were classified as non-HM with a median value of 3.5 mutations per 106 bases (range: 0.1–24 per 106 bases). The CNV calling analysis resulted in 29 amplified and 41 deleted focal regions significantly altered through all sets of tumor samples. We then subjected the 520 tumor samples to an unsupervised hierarchical clustering analysis of the CNVs which identified two main clusters (Fig. 3.2): Cluster A (ClA) characterized by few CNV events and Cluster B (ClB) with a high number of CNVs events. ClA was enriched in HM samples (n = 76/117; 65%; Fig. 3.2, yellow bars), while ClB mostly contained non-HM samples (n = 401/403; 99.5%; Fig. 3.2, blue bars). Within ClA, we noted a group of 41 samples with low CNV profile

and very low TMB (median value of 3.9 mutations per 106 bases, range: 0.1–23 mutations per 106 bases). Based on their clinical-pathological similarities with HM CRCs (as described below) this subset will be referred to as HM-like (Fig. 3.2) and accounted for 7.8% of the entire dataset.

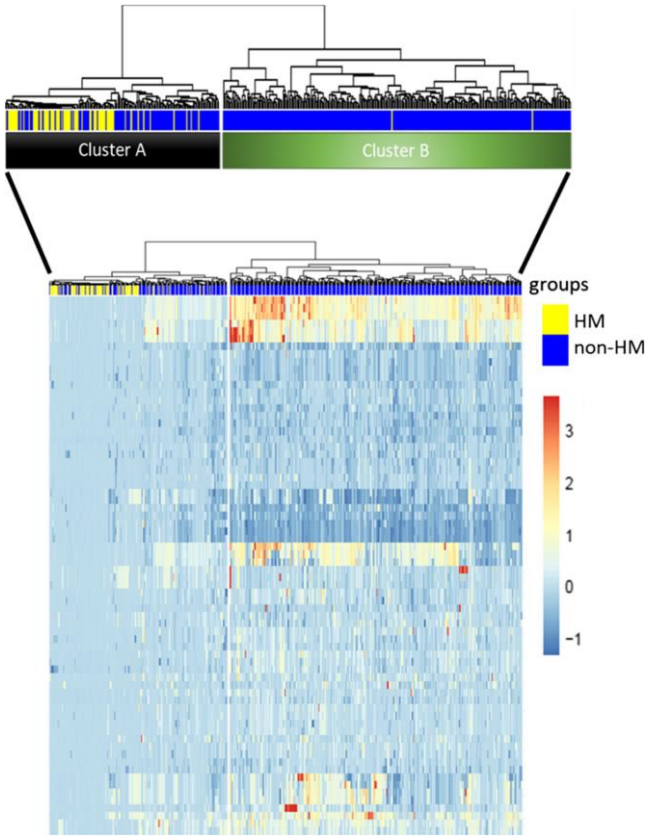

Figure 3.2. Unsupervised hierarchical clustering analysis based on CNVs data of the 520 CRC patients selected from TCGA-COAD and READ projects. The lines in the heatmap represent significant focal alteration. The columns correspond to the 520 patients. HM and non-HM samples are indicated in yellow and blue colors, respectively. This analysis identified two main clusters: cluster a (ClA) and cluster b (ClB). ClA (117/520; 22.5%) is characterized by a few events of CNVs along the chromosome regions and was enriched in HM samples (n = 76/117; 63.9%). ClB contains samples with a high number of CNVs events and it mostly consists of non-HM samples (n = 401/403; 99%). Among ClA, we identified a sub-group of tumors (called HM-like; n = 41/520; 7.8%) with a similar CNV profile of ClA, also characterized by a low TMB. To the right-hand side of the figure, a scale indicates the color code relative to the log2 segment mean value of CNVs (ranging from − 1 up to 3)

The profiles of CNVs amount and distribution among chromosomes were clearly distinct between the three subgroups. Overall, the HM-like group was characterized by a CNV profile more similar to the HM group than to the non-HM group (Fig. 3.3). However, these tumors also showed recurrence of gains (chromosomes 7, 9p and 19q) and losses (chromosome 8p, 10, 11, 15q, 17p and 18) more typical of non-HM samples (Fig. 3.3). As expected, most HM tumors were classified as MSI-H (n = 61/78; 78.2%), while non-HM and HM-like patients were much

more frequently MSS stable (n = 377/401; 94.0% and n= 34/41; 82.9%, respectively; Table 3.1). Consistent with the results of population studies [43], POLE exonuclease domain mutation rate was 2.9% (15/520) in our cohort, and all mutations fell in HM-group (15/78; 19.2%), while non-HM and HM-like patients showed no POLE alteration (Table 3.1). Overall, this analysis suggests that non-HM and HM subsets largely comprise CRCs associated with typical CIN and MSI/hypermutated phenotypes, respectively, while HM-like tumors appear as a distinct entity, with rather low CNVs and mutation rates.

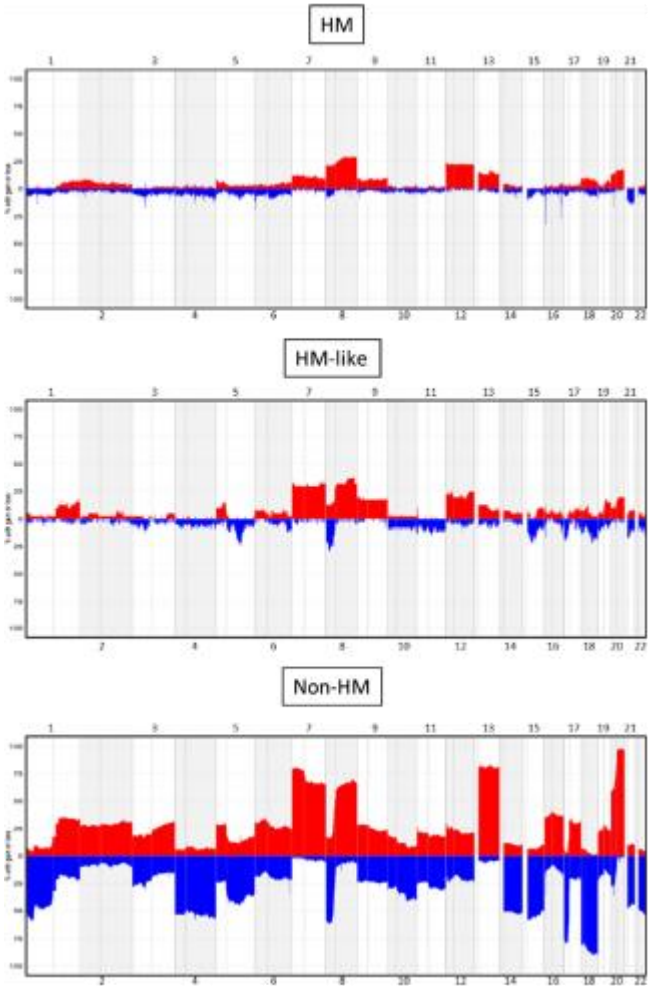

Figure 3.3 Frequency of CNV events along the genome identified in HM, HM-like and non-HM samples. Frequency of CNV events along the genome identified in HM, HM-like and non-HM samples. Frequencies (vertical axis, 0–100%) are plotted as a function of the chromosome location. Copy number gains and losses are highlighted in red and blue, respectively.

Clinical-pathological features and gene mutation rates in HM, HM-like and non-HM samples

Clinical-pathological features of HM, HM-like and nonHM samples are reported in Table 3.1. No significant associations were found with age or gender. As expected, HM patients were significantly enriched in early stages and in ascending colon localization compared to non-HM patients, which were more associated with stage 4 and in descending colon localization [49, 61]. Intriguingly, HM-like patients shared with HM subset a similar enrichment in early stages, with only 2.4% (1/41) and 3.8% (3/78) of the patients with HM-like and HM profiles in stage 4, against a rate of 18.0% (72/401) for non-HM patients ($P < 0.0001$, Fisher's exact test). Moreover, HM-like tumors were more frequently associated with ascending colon location (21/41; 51.2%) similar to HM (50/78; 64.1%), in contrast to non-HM tumors which were associated with descending colon location (253/401; 63.1%) ($P < 0.0001$, Fisher's exact test). To further compare the overall molecular features of HM-like versus HM and non-HM subsets we examined SNV data. As expected from the literature and according to their CIN profile [62] non-HM tumors had higher mutation rate in APC (84%), TP53 (69%) and KRAS (41%) compared to HM tumors (Table 3.2). In contrast, HM tumors had high mutation rates in genes of the WNT signaling, TGF-β, PI3K-AKT and MAPK/ERK pathways as well as in ATM, KMT2D and LRP1D [63]. Interestingly, HM-like tumors had the highest frequency in KRAS (59%) and SOX9 (27%) gene mutations compared to the other groups. Also, they showed the lowest TP53 mutation rate (15%) and a rate of APC mutations similar to HM samples and significantly lower than nonHM samples (Table 3.2). The pattern of mutational targets and rates support the hypothesis that HM-like tumors may represent a distinct subgroup of CRCs, which may develop and progress through a different sequence of genetic events compared to the well-known MSI/hypermutated and MSS/ CIN subsets, while sharing prevalence of early stages and ascending colon localization with the HM subset.

Table 3.1. Clinical-Pathological features of HM, HM-like and non-HM groups

		HM (n=78)	HM-like (n=41)	Non-HM (n=401)	P-value
Stage	I	14 (17.9%)	10 (24.4%)	62 (15.5%)	NS
	II	45 (57.7%)	16 (39.0%)	123 (30.9%)	***
	III	14 (17.9%)	13 (31.7%)	125 (31.2%)	*
	IV	3 (3.8%)	1 (2.4%)	72 (18.0%)	***
Location	Ascending	50 (64.1%)	21 (51.2%)	115 (28.7%)	***
	Transverse	10 (12.8%)	8 (19.5%)	16 (4.0%)	***
	Descending	12 (15.4%)	11 (26.8%)	253 (63.1%)	***
	No Data	6 (7.7%)	1 (2.4%)	17 (4.2%)	
Mutational Burden	Median of mutations/Megabase	44.9	3.9	3.5	
MSI-status	MSI-H	61 (78.2%)	6 (14.6%)	3 (0.7%)	***
	MSS/MSI-L	11 (14.1%)	34 (82.9%)	377 (94.0%)	***
	Indeterminate	6 (7.7%)	1 (2.4%)	21 (5.2%)	
Pol-ε exonuclease domain mutation		15 (19.2%)	0	0	-

Table 3.2. Mutational rate of most frequently altered genes in CRC in HM, HM-like and non-HM group

Genes	Pathway	HM	HM-like	Non-HM	P-value
APC	WNT signaling	49%	59%	84%	***
AMER1	WNT signaling	27%	15%	9%	***
CTNNB1	WNT signaling	24%	12%	3%	***
TCF7L2	WNT signaling	24%	0%	7%	***
FBXW7	WNT signaling	40%	32%	11%	***
ARID1A	WNT signaling	45%	5%	6%	***
SOX9	WNT signaling	15%	27%	11%	*
TGFBR2	TGF-β signaling	12%	7%	1%	NS
ACVR2A	TGF-β signaling	37%	15%	1%	**
SMAD4	TGF-β signaling	15%	17%	12%	NS
PIK3CA	PIK3 signaling	40%	41%	21%	***
PTEN	PIK3 signaling	22%	10%	3%	**
FAT4	Hippo signaling pathway	76%	24%	15%	***
ERBB2	MAPK signaling	15%	5%	2%	***
ERBB3	MAPK signaling	22%	5%	2%	***
KRAS	MAPK signaling	26%	59%	41%	***
NRAS	MAPK signaling	4%	7%	7%	NS
BRAF	MAPK signaling	62%	12%	3%	***
ATM	DNA damage response	50%	10%	7%	***
TP53	DNA damage response	29%	15%	69%	***
LRP1B	Membrane trafficking	53%	5%	13%	***
KMT2D	Histone methyl transferase	64%	15%	3%	***

Fingerprints of base substitutions in HM, HM-like and non-HM groups reveals unique mutational signature for each group

To further question whether HM-like CRCs are distinct from MSI/hypermutated and MSS/CIN subsets, we searched for the emergence of specific mutational signatures in the three subgroups. Indeed, different mutational processes generate unique combinations of base changes, termed "Mutational Signatures" which can be used as a readout of the biological history of a cancer [64]. To define the mutational signatures associated with HM, HM-like and non-HM groups we performed a classification of base substitutions to include the 3′ and 5′ flanking bases at the mutated site [54]. Thus, we extracted 3 mutational signatures from each group and compared them to COSMIC Single Base Substitution (SBS) Signatures database, a catalog of known mutational signatures identified from > 12,000 samples derived from 40 types of human cancer in which additional information for each signature were also provided. The top three signatures extracted from the HM group were the most similar to COSMIC SBS6, SBS10b and SBS44 signatures (Table 3.3) and that is consistent with the "hypermutated" phenotype defining the HM group, since SBS10b signature is associated with POLE mutations, which outbreaks in a high mutational rate, and COSMIC SBS6 and SBS44 are typically associated with dMMR. The three signatures extracted from non-HM samples had the highest similarity with COSMIC SBS1, SBS6 and SBS40 signatures (Table 3.3). SBS1, which was also noted in the other subgroups, is related to the spontaneous or enzymatic deamination of 5-methylcytosine to thymine and is widespread in many tumors. SBS40 signature is not clearly associated with a specific etiology, but like SBS1 it is widespread in most cancers and shows some relationships with the age of patients [65]. The three signatures extracted from HM-like samples showed high similarities with SBS1, SBS6 and SBS30. Similarity to SBS30 represents a feature unique to HM-like samples (Table 3.3). This signature was recently associated with deficiency in the base excision repair and with inactivation of the NTHL1 gene [66]. Despite some similarities in the mutational signatures were shared by two or even all three subgroups (i.e., SBS1, SBS6 and SBS15), this analysis further evidenced distinct mutational profiles between the HM, HM-like and non-HM subgroups.

Table 3.3. Records of the cosine similarity between the three mutational signatures extracted from each group from the MAF files and the three most similar COSMIC mutational signatures.
In the table are reported the best matches between the three mutational signatures extracted from the three groups and the COSMIC SBS Signatures database

		SBS Best match	Aetiology	Cosine similarity
HM	Signature 1	SBS44	Defective DNA mismatch repair	0.81
	Signature 2	SBS10b	Polymerase epsilon exonuclease domain mutations	0.78
	Signature 3	SBS6	Defective DNA mismatch repair	0.90
HM-like	Signature 1	SBS1	Spontaneous or enzymatic deamination of 5-methylcytosine	0.94
	Signature 2	SBS30	Deficiency in base excision repair due to inactivating mutations in NTHL1	0.83
	Signature 3	SBS6	Defective DNA mismatch repair	0.93
Non-HM	Signature 1	SBS1	Spontaneous or enzymatic deamination of 5-methylcytosine	0.96
	Signature 2	SBS40	Unknown	0.89
	Signature 3	SBS6	Defective DNA mismatch repair	0.77

Different CpG methylation patterns occur in the three CRC subgroups

Next, we performed an unsupervised hierarchical clustering analysis for 382 of the TCGA-COAD/READ samples for which CpGs methylation data were available. The hierarchical clustering dendrogram defined three distinct tumor groups: CIMP-H (n = 57/382; 14.9%) with a high rate of CpGs probes methylated; CIMP-L (n = 107/382; 28.0%) with low rate of CpGs probes methylated and non-CIMP (n = 218/382; 57.1%) characterized by the absence of CpGs methylated probes (Fig. 3.4). As expected, most of the HM patients belong to the CIMP-H cluster (41/57; 71.9%) and most non-HM tumors belong to the non-CIMP cluster (214/294; 72.8%), while a small number of HM and nonHM tumors clustered in the CIMP-L group. Interestingly, we revealed that the HM-like samples were mainly associated with CIMP-L phenotype (24/31; 77.4%) (Fig. 3.4). These results highlighted a different methylome pattern of HM-like tumors compared HM and non-HM.

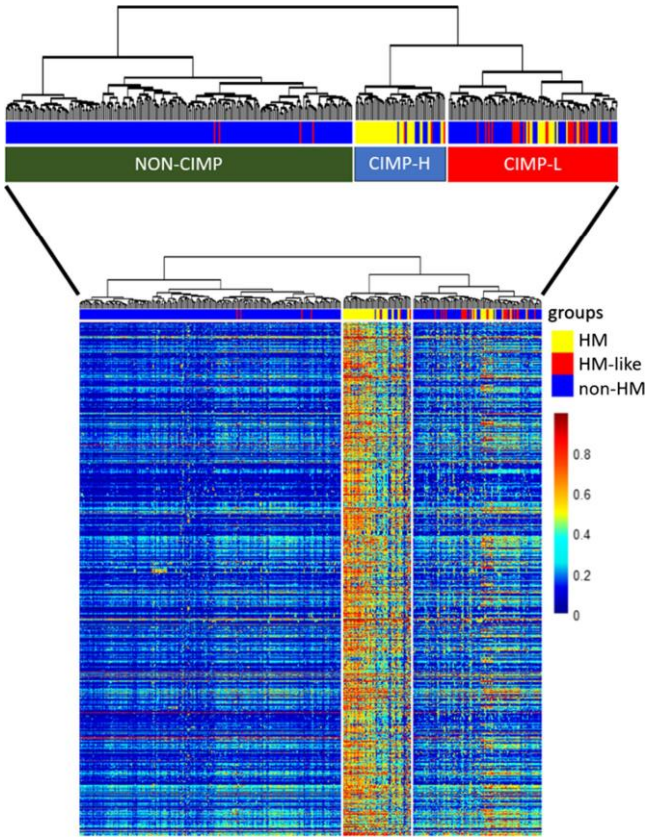

Figure 3.4 Unsupervised hierarchical clustering analysis based on CpGs methylation data of the 382 patients selected from TCGA-COAD and READ projects. The lines on the heatmap represent the 1000 most differentially methylated CpGs probes between HM, HM-like and non-HM groups. The columns correspond to the 382 patients. Inside the cells of the heatmap are reported the β-values which represent the methylation rate of the probes. The HM patients are reported in yellow; the HM-like patients in red while the non-HM patients in blue. The hierarchical clustering dendrogram supported three distinct tumor groups: CIMP-H (n = 57) defined by an high rate of CpGs probes methylated; CIMP-L (n = 107) with low rate of CpGs probes methylated and non-CIMP (n = 218) characterized by the absence of CpGs probes methylated

WGCNA analysis supports HM, HM-like and non-HM tumors as three distinct CRC subgroups

We performed the WGCNA network-based methodology on the transcriptomic data of 520 TCGA-COAD/ READ patients. This analysis revealed 12 highly correlated modules within the gene correlation network, which encompassed genes that were more correlated among each other than with other nodes in the network. For each module, through the WGCNA analysis,

we computed the module eigengene defined as the first principal component of that module. By considering as external clinical traits the HM, HM-like, and non-HM status, we then computed the Pearson correlation coefficient between the module eigengene of each module and these external traits (Fig. 3.5a). We found (1) three modules with statistically significant positive correlations with the HM status, meaning that genes belonging to these three modules were highly expressed in HM patients; (2) two modules with statistically significant positive correlations with the HM-like status, whose genes were highly expressed in HM-like patients; (3) one module with a statistically significant positive correlation with the non-HM status, whose genes were highly expressed in non-HM patients. All these modules did not overlap among the patient status (i.e., HM, HM-like, non-HM), suggesting that these three classes of CRC patients were different also with respect to the gene expression data. In order to identify specific gene signatures of the three subgroups, for each gene we computed the module membership (MM) as the correlation between its gene expression profile and the module eigengene and sorted genes within their own modules according to the MM. Yet, we considered as representative genes of a given module the ones whose MM was greater than 0.7. Then, for each patient status, we grouped together the representative genes of the modules with the highest correlation and performed a functional enrichment analysis. Via this process, we associated putative biomarkers and functional pathways to each status (i.e., HM, HM-like, non-HM). Also, this analysis confirmed relevant differences among the three subgroups. In detail, the HM status was characterized by high expression of genes mainly involved in the inflammatory bowel disease, Toll-like receptor signaling pathway, PI3K-Akt signaling pathway and JAK-STAT signaling pathways (Fig. 3.5b). The HM-like status was characterized by high expression of genes mainly involved in estrogen signaling and pathways related to the immune/inflammatory response (Fig. 3.5c). The non-HM status was characterized by high expression of genes mainly involved in RNA processing, DNA repair and VEGF signaling pathway (Fig. 3.5d).

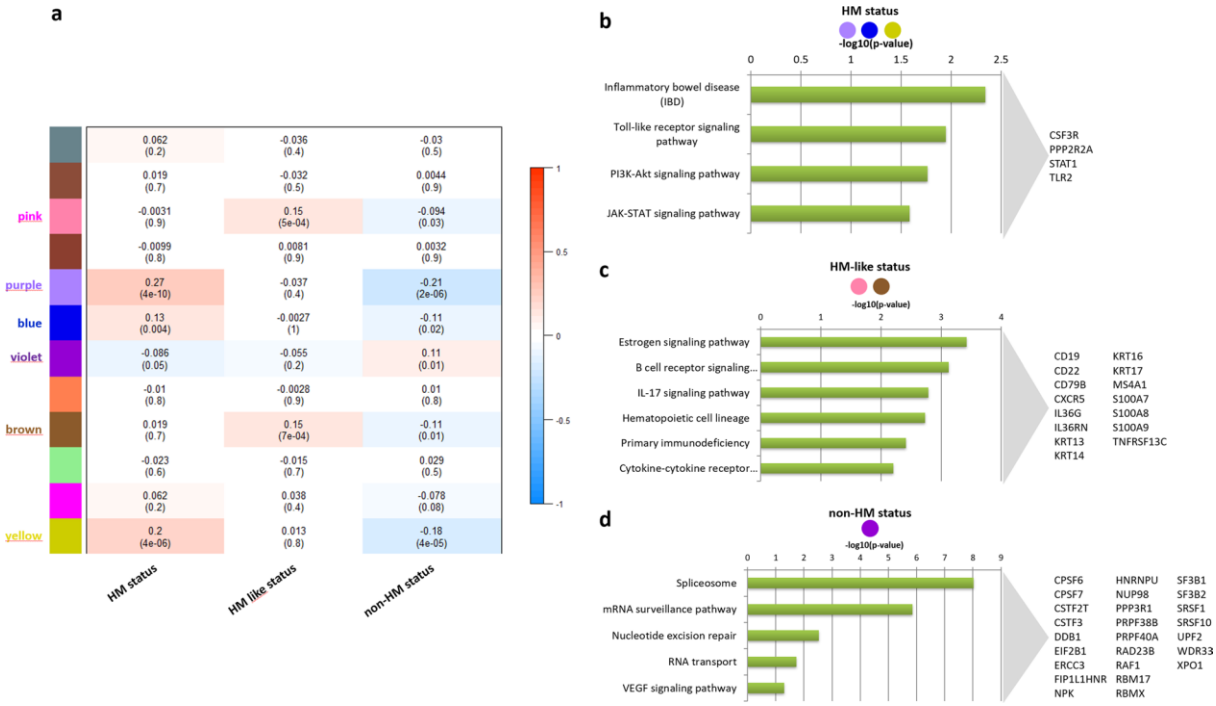

Figure 3.5. WGCNA analysis. (a) Heatmap of module-trait associations. In the heatmap, each row corresponds to a module eigengene and each column to a trait. Each cell contains the corresponding correlation and P value. The table is color-coded by correlation according to the color legend. The traits along the columns were numerically encoded as follows: HM status (no = 1, yes = 2); HM-like status (no = 1, yes = 2); non-HM status (no = 1, yes = 2). The color labels of modules with at least one statistically significant correlation were highlighted. b, c KEGG pathways. Results of KEGG pathways enrichment analysis for the most representative genes (module membership > 0.9) falling within the modules statistically significant correlated with the HM status (b), HM-like status (c), and non-HM status (d). The names of genes annotated for the enriched KEGG pathways were reported.

Rate of immune infiltrate in tumoral microenvironment of the three CRC subgroups

dMMR CRC, largely clustering in the HM subgroups, are typically associated with immune infiltration and good response to ICB therapy [41]. WCGNA analysis indicated activation of inflammatory/immune response genes in HM and HM-like tumors. Therefore, we set out to determine the rate of immune/inflammatory infiltration more specifically in the three subsets by a computational analysis of tumor transcriptomic data, using the R package ImSig [59]. By this mean, 10 signatures describing the relative abundance and statistical analysis of 7 inflammatory/immune cells plus 3 additional signatures were analyzed. Concerning T and NK lymphocytes, as expected, we observed the highest signature representation in the HM group, while non-HM have a significantly lower degree of immune cell infiltration ($P < 0.01$) (Fig.

3.6). Interestingly, HM-like tumors showed T and NK cell signatures similar to HM samples and significantly different than non-HM group ($P < 0.01$). On the other side, HM-like tumors had proliferation, macrophage, and interferon signatures more similar to non-HM than HM tumors. However, some genes belonging to interferon signatures and involved in inflammatory/immune responses shared a similar expression between HM-like and HM tumors, while being different from non-HM tumors .Next, we performed a differential expression analysis of the 79 ICGs described by [60] expressed in the series. The comparison between HM and non-HM samples revealed that 46 ICGs were differentially expressed and as expected, most of these (29/46; 63.0%) were more expressed in HM group than non-HM. These included KIR and HLA genes, possibly suggestive of NK and antigen presenting cells infiltration, as well as multiple genes directly involved in immune checkpoint regulation, including the well-known PD-L1, PD1, CTLA4, LAG3, TIM3 and TIGIT (Fig. 3.7). Interestingly, 17 genes were significantly less expressed in HM compared to non-HM samples. HM-like tumors profoundly differed from HM and non-HM samples. They showed 13 ICGs significantly more expressed compared to non-HM tumors. KIR genes, TIGIT, PD1 and CTLA4 show a similar trend compared to HM samples. Differences in the expression of HLA genes did not reach statistical significance, while CD96 appears even more differentially expressed in this subgroup than in HM tumors, comparing with non-HM subset. Remarkably, we noticed that 4 genes whose role in immune checkpoint regulation is emerging (VTCN1, BTNL9, BTLA and CD28) were specifically more expressed in HM-like group compared to non-HM samples (Fig. 3.7). Also, in this comparison we found repressed genes (i.e., SIRPA, BTN2A1 and PVR), some of which followed the same trend of HM tumors, while others where rather specific for this subset (i.e., CD70, CD40). Conversely, IDO1, TDO2, and CD40LG expression trends were completely opposite in HM versus HM-like subgroups.

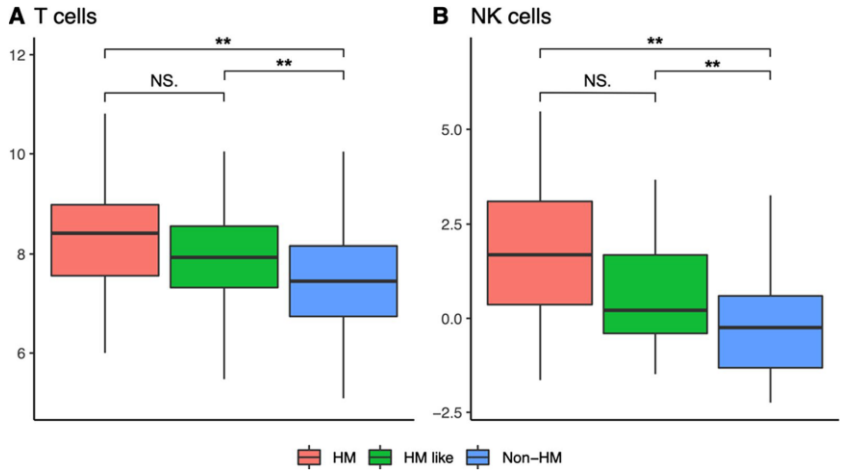

Figure 3.6. Results of immune signatures analysis performed by ImSig. The boxplots (A and B) show the gene expression of T and NK signature genes (estimated relative abundance) across the HM, non-HM and HM-like groups. Statistical analysis of data was performed using analysis of variance (ANOVA) followed by multiple comparison Tukey's test. **P < .01, *P < .05

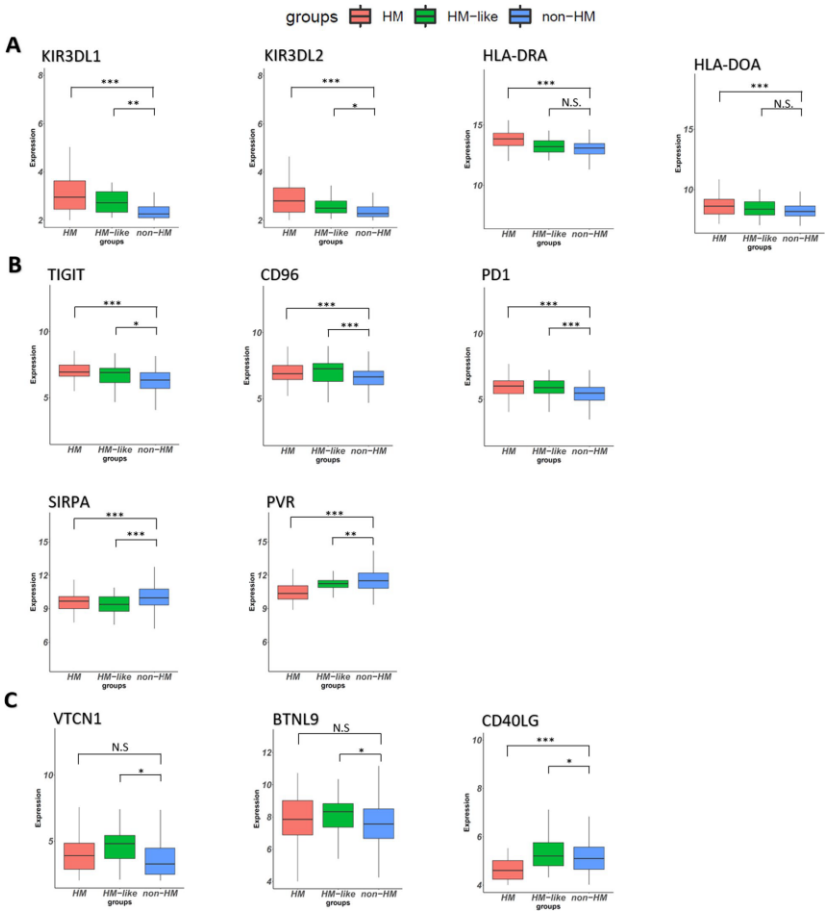

Figure 3.7. Example of ICGs differentially expressed in HM, non-HM and HM-like groups. The Box Plots show A ICGs more expressed in HM group versus non-HM, which may (KIR genes) or may not (HLA genes) be significantly more expressed in HM-like vs non-HM tumors; B gene sharing a similar trend of expression between HM and HM-like; C gene specifically more expressed in HM-like group (VTCN1 and BTNL9) or with an opposite trend of expression in HM versus HM-like (CD40LG). The analysis was performed using the multiple comparison of the three subgroups using Wald test and P value was adjusted according to the Benjamini–Hochberg method. Thresholds for FDR < 0.1 and Log2 Fold Change> 0.4 were used to select significant differentially expressed genes

3.4. Discussion

The comprehension of the biological processes underlying cancer evolution and the molecular stratification of tumors is extremely relevant for prognostic and therapeutic purposes. For what concerns ICI therapy, tumor with high mutation load may lead to generation of a high number of immunogenic neoantigens [67], which in turn can facilitate immune responses against

cancer cells. In this therapeutic framework, CIN (non-HM) CRCs usually bear low TMB and are mostly resistant to ICIs, meanwhile HM CRCs bear high TMB and are prone to ICIs therapy. By performing hierarchical clustering analysis of CNVs versus hypermutation status exploiting TCGA CRC datasets, we identified a third cluster of CRCs (7.8%) characterized by low CNVs and low TMB, distinct from the HM and non-HM subsets, which largely matched the MSI and CIN groups, respectively. Since this new cluster shared clinical-pathological features with HM CRCs, it was named HM-like subset. Interestingly, HM-like tumors also showed a distinct mutational profile compared with HM and non-HM tumors, for which we highlighted profiles essentially in line with the literature [49, 68]. In example, the rate of APC mutations in HM-like tumors was similar to HM samples and significantly lower than non-HM samples, while mutations in alternative targets of the WNT and TGF-beta pathways were much lower than those occurring in HM samples, suggesting that this tumor subset is probably less dependent from WNT activation than the other groups. Most importantly, HM-like tumors were characterized by the highest rate of KRAS mutation, a feature that was previously noted in CIMP-L CRCs [69]. This is a CRC subset with a yet poorly defined clinical relevance, often grouped with the non-CIMP tumors in various studies [70] and sharing the majority of methylation targets with CIMP-H tumors [71]. By methylation analysis, we found that HM-like tumors had mainly a CIMP-L phenotype, at variance with HM and non-HM tumors, which were mostly associated with CIMP-H and non-CIMP phenotype, respectively [49]. Therefore, our data confirm a particularly high recurrence of KRAS mutations in a specific subset of CRCs, associated with CIMP-L phenotype. While the molecular background for this association is not understood yet, recent studies seem to indicate that the strong association between BRAF mutations and CIMP-H phenotype might be due to the need to suppress a senescence-inducing gene expression program promoted by mutant BRAF [72]. Oncogenic RAS molecules are also known to activate senescence in untransformed cells [73, 74]. It is tempting to speculate that also the relevant overlap between KRAS mutation and the CIMP-L phenotype in the HM-like subgroup could be related to the repression of a similar senescence-inducing gene expression program. Further efforts will be required to formally prove this hypothesis. HM-like CRCs also showed the highest frequency of SOX9 gene mutations and

the lowest rate of TP53 mutations. This association was previously recognized, but its functional significance remains ununderstood [75]. Overall, the genetic marks of HM-like supports the hypothesis that may represent a distinct subgroup of CRCs, which may arise and progress through a different sequence of genetic events compared to the well-known MSI/hypermutated and MSS/CIN subsets. This is further supported by the analysis of mutational signatures, which indicate their unique similarity to the SBS30 pattern. This was recently associated with deficiency in the base excision repair and with inactivation of the NTHL1 gene [66]. Biallelic NTHL1 mutations are responsible for the NTHL1-tumor syndrome, a cancer-predisposing disease characterized by the occurrence of adenomatous polyposis and cancer at different sites, in addition to CRC [76]. This specific genetic fingerprint indicates that also the pathogenic mechanisms and the etiology underlying HM-like CRCs might be distinct from those leading to HM and non-HM CRCs. So far, we were unable to pull out genomic or transcriptomic alterations in the NTHL1 gene specifically occurring in the HM-like group, suggesting that functional inactivation of its pathway perhaps associated to the specific CIMP-L pattern might be involved in this respect. Additional studies should be implemented to highlight possible genetic/epigenetic hits or alternative/parallel pathways to NTHL1 inactivation, which might end up in eliciting the same molecular fingerprints. The existence of a small group of pMMR/MSS CRCs (~ 10%) responsive to ICIs therapies was inferred in several clinical studies [77–79]. Pagès and collaborators observed a high immunoscore in 21% of MSS compared to 45% of MSI [80]. Similar findings were reported by Kikuchi et al. which identified a subset of MSI-L/MSS CRCs within the TCGA COAD/READ dataset showing upregulation of the IFN-γ and CD8 T effector gene signatures [81]. They also confirmed the presence of a small fraction (~ 12%) of pMMR CRCs positive for PDL1 and p-STAT1 showing increasing grades of infiltrating CD4(+) or CD8(+) TILs on a population of 219 CRC samples. Our work raised the question whether the HM-like group identifies the same CRC subset. Indeed, not only WCGNA analysis of the transcriptome evidenced relevant differences among the three groups, but also indicated that the HM-like tumors bore high expression of genes associated with immune/inflammatory response. To better investigate this latter aspect, we defined the immune/inflammatory infiltration signature

in the three subsets, according to [59]. Intriguingly, we confirmed that HM-like tumors showed T and NK cells signatures similar to HM samples which, as widely known, are inflamed tumors well responsive to ICI therapies. In contrast, proliferation, macrophage and interferon signatures in HM-like tumors were on average more similar to the non-HM than to HM group. Data on the differential expressions of the ICGs curated by [60] further confirmed the outstanding differences among the three groups. Coherently with the immune infiltration analysis, HM samples showed a high expression of multiple ICGs, confirming the presence of an immune/inflammatory infiltrate (KIR and HLA genes) and differential expression of immune response modulators, including those targeted by established ICI therapies. KIR genes and ICGs (i.e., PD1, CTLA4, CD96 and TIGIT) for which specific targeting therapies were introduced in the clinical practice [41, 82] also showed a higher expression in HM-like tumors compared to non-HM samples, supporting their immune/inflammatory infiltration. Moreover, our analysis highlighted ICGs exclusively expressed in HM-like, e.g., VTCN1, PCDCD1, CD96, BTNL9 and BTLA, encoding for important immune regulators of both stimulatory and inhibitory pathways, some of which are emerging as new promising targets for immunotherapy [83, 84]. While these data confirm the presence of an immune/inflammatory infiltrate in HM-like tumors showing modulation of established and potentially new immune checkpoint targets to consider for ICI therapies, remarkable differences emerged between HM-like and HM group. Among them, the relatively lower expression of HLA genes in HM-like samples is in line with the poor macrophage signature observed in this subgroup compared to HM samples. The significance of a potentially lower infiltration by antigen presenting/ dendritic cells and the relevant differences in the pattern of immunomodulating molecules expressed in HM and HM-like tumors cannot be easily interpreted at the time being and requires further investigations. These differences, however, do not contrast with our hypothesis that HM-like CRCs might be responsive to ICI. Of relevance, the strong negative regulation of IDO and TDO2 in HM-like compared to both HM and non-HM tumors suggest that the formers are possibly characterized by a less immunosuppressive microenvironment caused by the release of tryptophan metabolites. Perhaps this condition might also be related to the more frequent association of HM-like tumors with early stages CRC and may eventually make them more prone to immune

reactivation. Unfortunately, a major limitation of this study is represented by the lack of a univocal specific molecular biomarker/s facilitating the identification of HM-like CRC, in clinical settings. To this end, the possibility to use CIMP-L phenotype needs to be explored.

Chapter 4: Procedural multi-omics data integration for studying Breast Cancer

In silico recognition of a prognostic signature in basal-like breast cancer patients

4.1. Introduction

Breast Cancer (BC) is the most common female cancer and despite important advances in early detection and research development, it continues to be the second leading cause of death in women worldwide [85]. BC is a heterogeneous pathology as witnessed by the existence of different subtypes with distinct morphologies and clinical implications [86]. These subtypes are usually defined by using immunohistochemical (IHC) [87] and genetic (PAM50) [88, 89] classifications. According to the IHC classification, the different BC subtypes are: Luminal A, Luminal B, Her2 positive and Triple negative. According to the PAM50 classification, the different BC subtypes are: Luminal A, Luminal B, Her2 positive and Basal-like. The most aggressive BC pathophenotypes are the triple-negative BC (TNBC) and the Basal-like, respectively. Triple-negative BC (TNBC) accounts for a minority of all diagnosed BCs (15–20%) [5]. It is a subtype with a heterogeneous nature, defined by the low or absent expression of estrogen (ER), progesterone (PR) receptors and the lack of expression of the human epidermal growth factor (EGF) receptor-2 (HER2) receptors [90]. These cancers differ from other BC subtypes in that they grow and spread faster, have limited treatment options (typically treated with chemotherapy) and their metastatic pattern spread with a higher likelihood of brain and lung involvement and less frequently with bone lesions. Relapse is common in TNBC, usually in the first 5years, leading to the poorest survival outcomes between all BC subtypes [91]. Currently, there are not available widely accepted prognostic markers to predict outcomes in TNBC patients. TNBC is often used as a surrogate for identifying the aggressive basal-like BC subtype. Although the two patterns share many similarities, biologically they are not the same, but both are associated with poor clinical outcomes. Therefore, the development of new prognostic indicators for basal-like subtype represents an unmet clinical challenge that might

be of benefit to the clinical management of this disease. In a recent study [5], the SWIM methodology was applied to the transcriptomic profile of a total of 505 BC patients stratified according to PAM50 subtypes classification to identify switch genes shared among four subtypes and specific for each subtype. In this study, the authors focused on the common switch genes and performed several in silico analysis and in vitro and ex vivo experiments to highlight molecular signatures shared among all BC subtypes.

Here, we performed a procedural multi-omics data integration of TCGA-BC transcriptomic, genomic, epigenomic and clinical data for the basal-like specific switch genes identified in [5], in order to identify a basal-like prognostic gene signature. Our study showed that 11 basal-like specific switch genes were overexpressed in BC tissues compared to normal counterpart and associated with BC patient's prognosis acting as unfavorable prognostic markers. Also, their highest expression was found in the basal-like subtype and this overexpression could be putatively related to genetic and epigenetic alterations as well as the action of important transcription factors. These 11 basal-like specific switch genes can constitute a gene signature to evaluate the prognosis of basal-like BC patients independently from the therapeutic intervention.

4.2. Materials and Methods

Data collection and processing

We exploited TCGA to obtain transcriptomic and clinicopathological data of the entire cohort of 1049 BC samples, and Copy Number Variations (CNVs) data of 317 TCGA-BC patients (92 basal-like and 225 luminal A). DNA methylation data of 152 TCGA-BC patients (37 basal-like and 152 luminal A) were retrieved from the Firehorse TCGA GDAC browser (https://gdac.broadinstitute.org/). The BC patients considered are female not subjected to a neoadjuvant treatment. Moreover, the Human Protein Atlas website (https://www.proteinatlas.org) was leveraged to identify tumor-type specific proteins expression patterns and to perform immunohistochemistry image a direct comparison of the protein expression of selected prognostic indicators between normal and tumor breast tissues.

SWIM software

An explanation of SWIM software is given in the Chapter 2 section 3 of this thesis.

Kaplan-Meier survival analysis

To analyze the correlation between the expression level of the 108 basal-like specific switch genes and patient overall survival (OS) and therefore to evaluate their prognostic value, we used the RNA-sequencing data from TCGA to split the entire cohort of BC patients (1049 samples) into two groups (called low-expression and high-expression group) according to the upper and lower expression quartile. Low- and high-expression groups refer to patients with expression levels of the given switch gene lower and greater than the 50th percentile (i.e., median), respectively. For each patient cohort, the cumulative survival rates were computed for each switch gene according to the Kaplan-Meier (KM) method [11] on the clinical metadata provided by TCGA. For each switch gene, the survival outcomes of the two patients' groups were compared by the log-rank test. Switch genes with log-rank p-values less than 0.05 were suggested as candidate prognostic biomarkers. In particular, the lower the p-value, the better the separation between the two prognosis groups. If the group of patients with high expression

of the selected prognostic gene has a higher observed event than expected event (worst prognosis), it is defined as an unfavorable prognostic gene; otherwise, if its high expression is associated with the best prognosis, it is a favorable prognostic gene. To confirm the prognostic value of the basal-like specific switch genes points out from the KM survival analysis on the TCGA Breast Cancer patients, we performed the KM analysis on different BC dataset. To do this, we exploited the Kaplan-Meier plotter website (http://kmplot.com/analysis/), which integrates gene expression data and OS information downloaded from GEO, EGA and TCGA for several types of cancer [92]. We ran KaplanMeier plotter by considering the entire BC database including 7,830 unique samples from 55 independent affymetrix datasets [93] and by dividing patients into high and low expression group based on the auto selected best cutoff computed between the lower and upper quartiles of switch genes expression.

Statistical methods

The one-way analysis of variance (ANOVA) is an extension of independent two-samples t-test for comparing means in a situation where there are more than two groups. In one-way ANOVA, the data is organized into several groups based on one single grouping variable (also called factor variable). In this study, the one-way ANOVA test was used to compare the means of selected genes in patients grouped based on the PAM50 BC subtypes. A p-value ≤ 0.05 indicated that at least two groups significantly differ from each other and multiple pairwise-comparisons exploiting the t-test method were performed to identify which ones.

Gene regulatory network

The gene regulatory network of the selected switch genes was constructed by integrating information from Pscan [94], TRRUST [95] and the human interactome (i.e., that is the network of all physical interactions within a cell, from protein-protein to regulatory protein–DNA and metabolic interactions [13]). Pscan is a web tool designed to computationally predict TF-target regulatory relationships [94]. It scans the sequence of the promoter regions from an input gene list with motifs describing the binding specificity of known transcription factors and assesses which motifs are significantly over-or under-represented, suggesting which

transcription factors could be common regulators of the input genes. In this study, the promoter regions were identified as the genomic regions spanning from -450 to +50 nucleotides to transcription start sites and the TF binding profiles were retrieved from JASPAR 2018 database [96] .TRRUST is a freely available and manually curated database containing 8,444 TF-target regulatory relationships of 800 human transcription factors. These relationships were derived from PubMed articles describing small-scale experimental studies of transcriptional regulations by using a sentence-based text mining approach [95]. The human interactome, also called protein-protein interaction (PPI) network, was downloaded from Cheng and coauthors [8], where the authors assembled their in-house systematic human interactome with 15 commonly used databases with several types of experimental evidence (e.g., binary PPIs from three-dimensional protein structures; literature-curated PPIs identified by affinity purification followed by mass spectrometry, Y2H, and/or literature derived low-throughput experiments; signaling networks from literature-derived low throughput experiments; kinase-substrate interactions from literature-derived low-throughput and high-throughput experiments). This version of the interactome is composed of 217,160 protein-protein interactions connecting 15,970 unique proteins.

Copy Number Variations (CNVs) data analysis

Copy Number Variations (CNVs) data reported contiguous chromosome regions with log2 ratio segment means in a tab-delimited format. To obtain segment means values of CNVs of the selected genes for the enrolled patients, we employed GISTIC 2.0 software [52]. Gistic's parameters used in this study are the following:

```
-b path_file;
-seg  filename;
-refgene refgenefiles/hg19.UCSC.add_miR.140312.refgene.mat;
-mk genome.info.6.0_hg19.na31_minus_frequent_nan_probes_sorted_2.1.txt;
-maxspace 2000;
-ta0.3;
-td0.3;
-js4;
-qvt 0.01;
-conf 0.99;
-genegistic 1;
```

```
-armpeel 1;
-savegene 1
```

The hierarchical clustering analysis was performed by using “Canberra” as clustering distance and “ward.D2” as clustering method. The association between the CNVs status of the selected genes and the BC subtypes was evaluated using Fisher’s exact test.

DNA methylation data analysis

The level of DNA methylation for more than 450 000 CpG sites in the human genome was represented as beta value To make available and pre-process methylation data in R environment, we used minfi package [97]. Pre-processing was performed using an in-house R script that eliminated probes with no methylation level detectable, removed all known single-nucleotide polymorphism (SNP)-associated CpG sites, associated CpG sites with known genes and matched patients and genes selected in our study. The hierarchical clustering analysis was performed by using “Euclidean” as clustering distance and “ward.D2” as clustering method.

4.3. Results

Study design

In a recent paper by Grimaldi and colleagues [5], a total of 505 BC subjects (229 Luminal A, 120 Luminal B, 58 HER2-enriched, and 98 Basal-like) were analyzed and 108 switch genes were identified as specific for the most aggressive BC subtype, i.e., the basal-like subtype [92–94]. In the present study, we aim to predict important prognostic biomarkers among these basal-like specific switch genes. A schematic for our study design is depicted in Fig. 4.1.

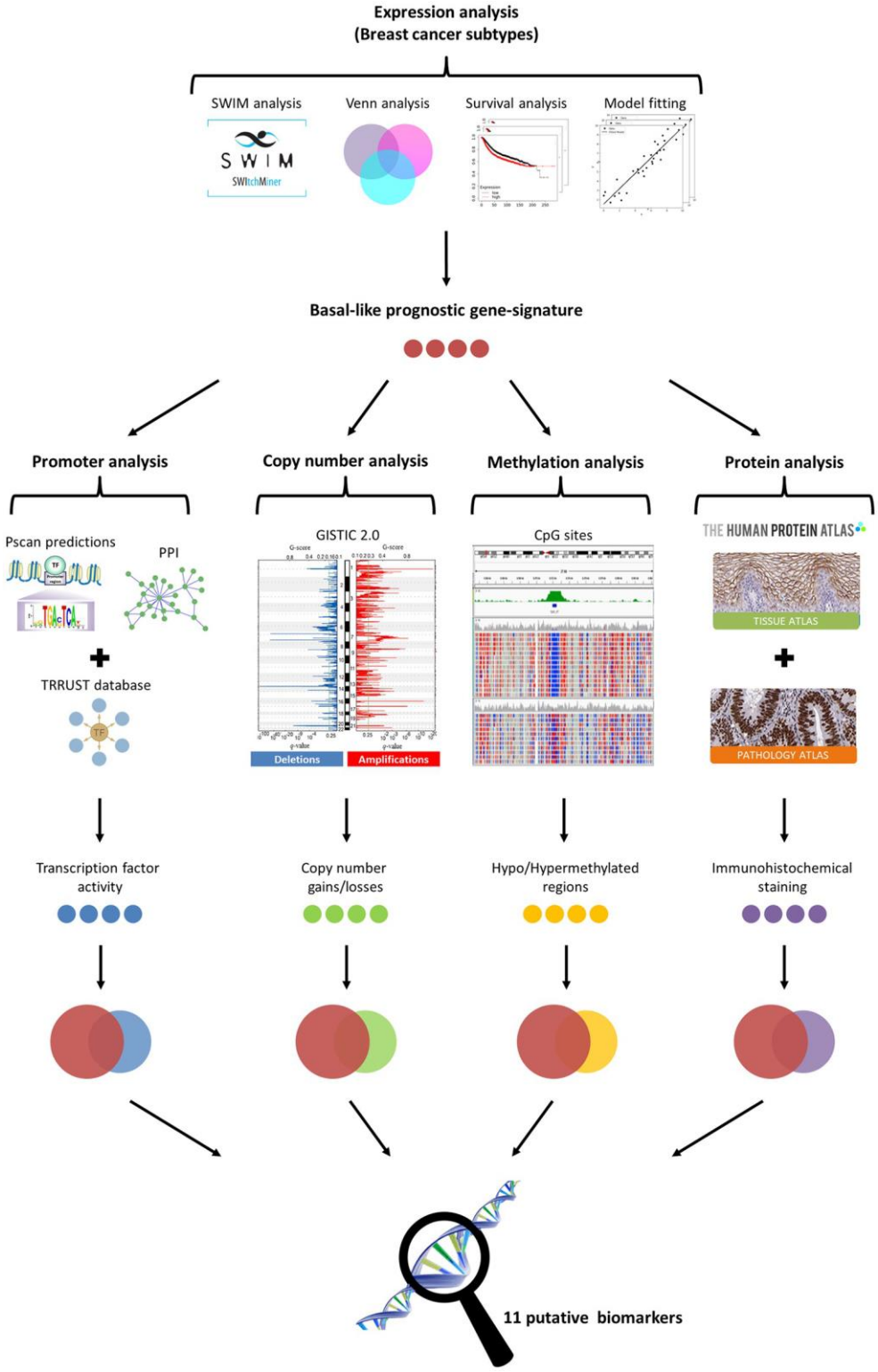

Figure 4 1. Study design. The figure depicts the schematic of the methodology applied in this study.

Prognostic value of basal-like specific switch genes

To study the clinical relevance of the basal-like specific switch genes with respect to the patients' survival, we exploited their expression profiles to perform the Kaplan-Meier analysis. We used the RNA-sequencing data available on the TCGA to stratify BC patients in two groups according to the expression levels of the 108 basal-like specific switch genes. Thus, for each switch genes, low (high)-expression groups refer to patients with the expression level of that gene lower (greater) than the median of its expression values across all BC patients. Then, a log-rank test was performed to assess a statistical significance (p-value) to each gene: the lower the p-value, the better the separation between the two prognosis groups. Switch genes with log-rank p-values less than 0.05 were candidate as potential biomarkers for predicting the survival rate of BC patients. We found a total of 15 basal-like specific switch genes that were significantly associated (p-values < 0.05) with BC patients' prognosis. Among them, 11 switch genes (i.e., CENPN, LRP8, DSCC1, CTPS, RCOR2, GINS4, TUBA1C, PRAME, SLC7A11, CDCA7, GSDMC) appeared to be an unfavorable prognostic gene (Fig. 4.2), suggesting that their higher expression could be associated with poorer BC patients' overall survival (OS). The other four switch genes (i.e., NXNL2, PHGR1, LOC389033, C10orf79) appeared as a favorable prognostic gene since their high expression correlated with a better clinical outcome. Hereafter, we focused only on the 11 basal-like specific switch genes whose activation appeared to be associated with the worst prognosis. Their clinical relevance was also confirmed using other BC datasets collected in the Kaplan-Meier plotter website [98] (Table 4.1, RNA level).

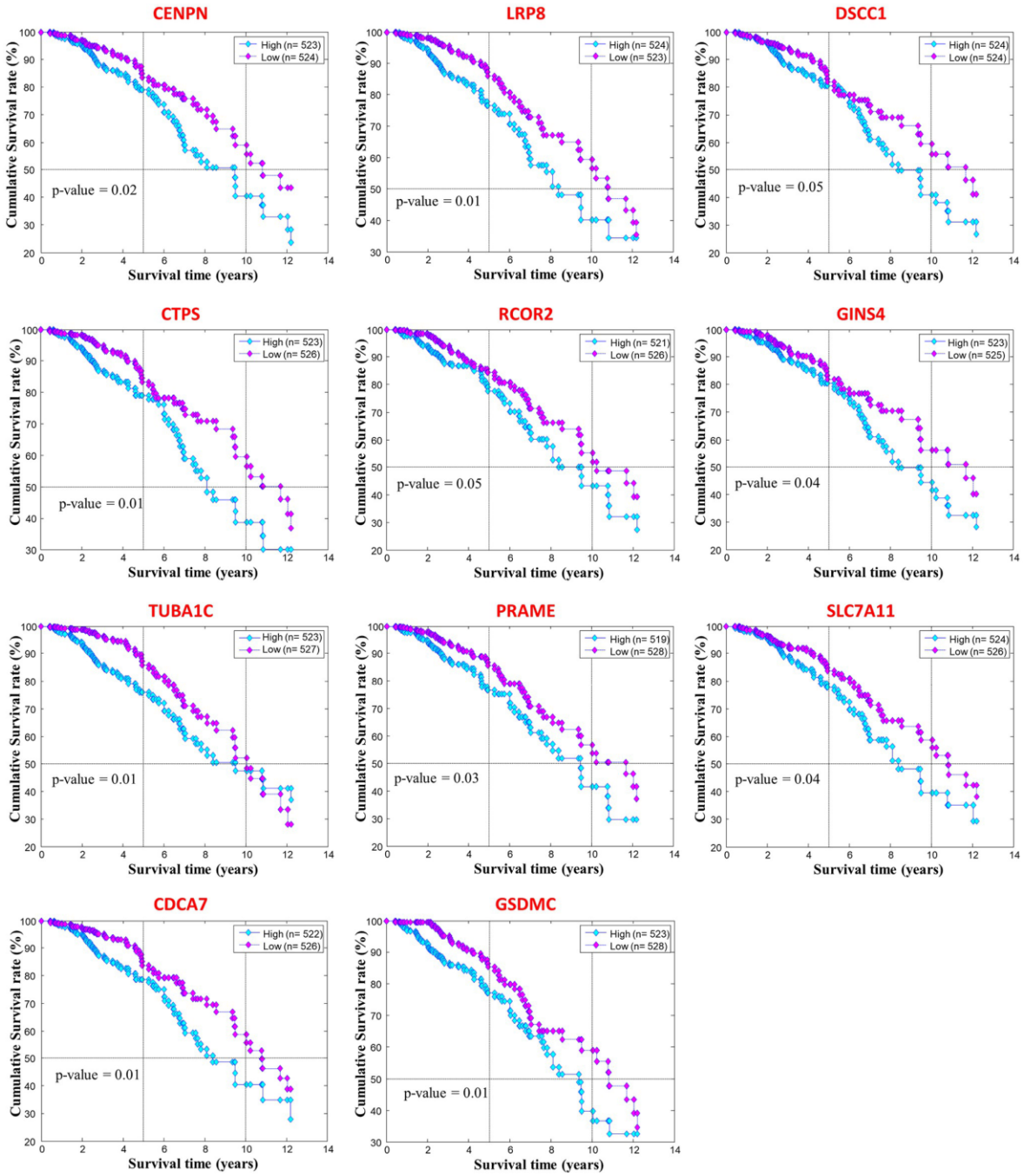

Figure 4.2. Switch genes with an unfavorable prognostic value from the survival analysis on TCGA data. Kaplan-Meier analyzes to evaluate the correlations between the expression of the basal-like specific switch genes and the OS in TCGA BC patients. Low- and high expression groups refer to patients with expression levels lower and greater than the 50th percentile, respectively.

Table 4.1. Summary of the properties of the basal-like prognostic biomarkers. Abbreviations: TFs, Transcription Factors; CNVs, Copy Number Variations; KM, Kaplan-Meier; IHC, Immunohistochemistry; PPI, protein-protein interactions; TCGA, The Cancer Genome Atlas; HPA, Human Protein Atlas; BC, Breast Cancer; BL, Basal-like; LumA, Luminal A; amp, amplified; del, deleted; hypo, hypomethylated. Asterisk (*) was used to highlight values not satisfying the chosen thresholds as well as not available data.

	DNA			**RNA**					**Protein**
	TFs	**CNVs**	**Methylation**	**SWIM**	**KM analysis (log rank p-value)**		**model fitting (index R^2)**		**IHC staining**
	TRRUST/Pscan/PPI	**TCGA**	**TCGA**	**TCGA**	**TCGA**	**other datasets**	**subtype**	**stage**	**HPA**
CENPN	*NRF1*	amp in BL/del in LumA	hypo in BL	switch genes	0.02	4.9E-6	0.99	0.96	not available*
LRP8	*HIC1*	amp in BL/del in LumA	-	switch genes	0.01	2.4E-4	0.98	0.63*	more expressed in BC
DSCC1	*HMBOX1*	amp in BL	-	switch genes	0.05	3.5E-8	0.95	0.78	more expressed in BC
CTPS	*MYC, TWIST1-2, NRF1*	amp in BL/del in LumA	hypo in BL	switch genes	0.01	8.2E-5	0.94	0.72	more expressed in BC
RCOR2	-	-	-	switch genes	0.05	4.3E-3	0.93	0.47*	more expressed in BC
GINS4	-	-	-	switch genes	0.04	6.4E-3	0.90	0.68	more expressed in BC
TUBA1C	*TP53, NFKB1*	del in BL	-	switch genes	0.01	1.3E-6	0.89	0.76	more expressed in BC
PRAME	*NRF1, SOX9, RARA*	amp in BL/del in LumA	hypo in BL	switch genes	0.03	9.9E-6	0.83	0.76	not available*
SLC7A11	-	-	-	switch genes	0.04	0.03	0.80	0.46*	not available*
CDCA7	*MYC, E2F1*	amp in BL	-	switch genes	0.01	1.3E-4	0.73	0.32*	not available*
GSDMC	-	amp in BL	hypo in BL	switch genes	0.01	4.9E-4	0.64*	0.05*	not available*

Overexpression of the basal-like prognostic biomarkers

A differential expression analysis showed that the 11 basal-like specific switch genes, whose unfavorable prognostic value was statistically significant from the previous survival analysis, were all up-regulated in the basal-like cancer condition compared to the normal condition. Yet, by performing an ANOVA test and multiple pairwise-comparisons among all the BC subtypes, we found that each comparison was statistically significant and the expression value of the 11 basal-like specific switch genes was greater in the basal-like versus the others BC subtypes and always greater than the median used in the KM survival analysis, leading to an association between worst prognosis patients (high-expression groups in the KM plots) and basal-like

affected subjects (Fig. 4.3). Taken all together, these findings prompted us to identify these 11 switch genes as potential prognostic biomarkers for basal-like subtype. To statistically quantify the increasing trend of the median expression values of these 11 switch genes as the phenotype varies from physiological to pathological condition passing across the different BC subtypes, we exploited a linear regression model, where the index R squared estimates the goodness-of-fit. We found that all but one showed a very strong straight-line relationship (R-squared ≥ 0.7) between their median expression and the tumor subtypes (Table 4.1, RNA level), with the CENPN as the first on the list (R-squared = 0.99). These results were mostly confirmed by performing the same analysis using the pathological staging of the BC patients affected by PAM50 subtypes (Table 4.1, RNA level). Indeed, we observed that 6basal-like specific switch genes (i.e., CENPN, DSCC1, CTPS, GINS4, TUBA1C, PRAME) reached an R-squared (rounded to one decimal place) ≥ 0.7 also with respect to the staging (Table 4.1, RNA level). The increasing trend of the top-ranked switch genes (highest R-squared) both with respect to the subtypes and the staging is depicted in Fig. 4.4a and 11b, respectively. To explore the expression patterns of the proteins encoded by the 11 prognostic switch genes, we queried the Human Protein Atlas (HPA) that provided representative immunohistochemistry images in BC tissues and normal breast tissues. As expected, we found that six of these proteins were overexpressed in BC tissues compared to normal breast tissues (Fig. 4.5 and Table 4.1, Protein level). For the remaining ones, there are pending cancer and normal tissue analysis on the HPA and the immunohistochemistry images are not currently available (Table 4.1, Protein level).

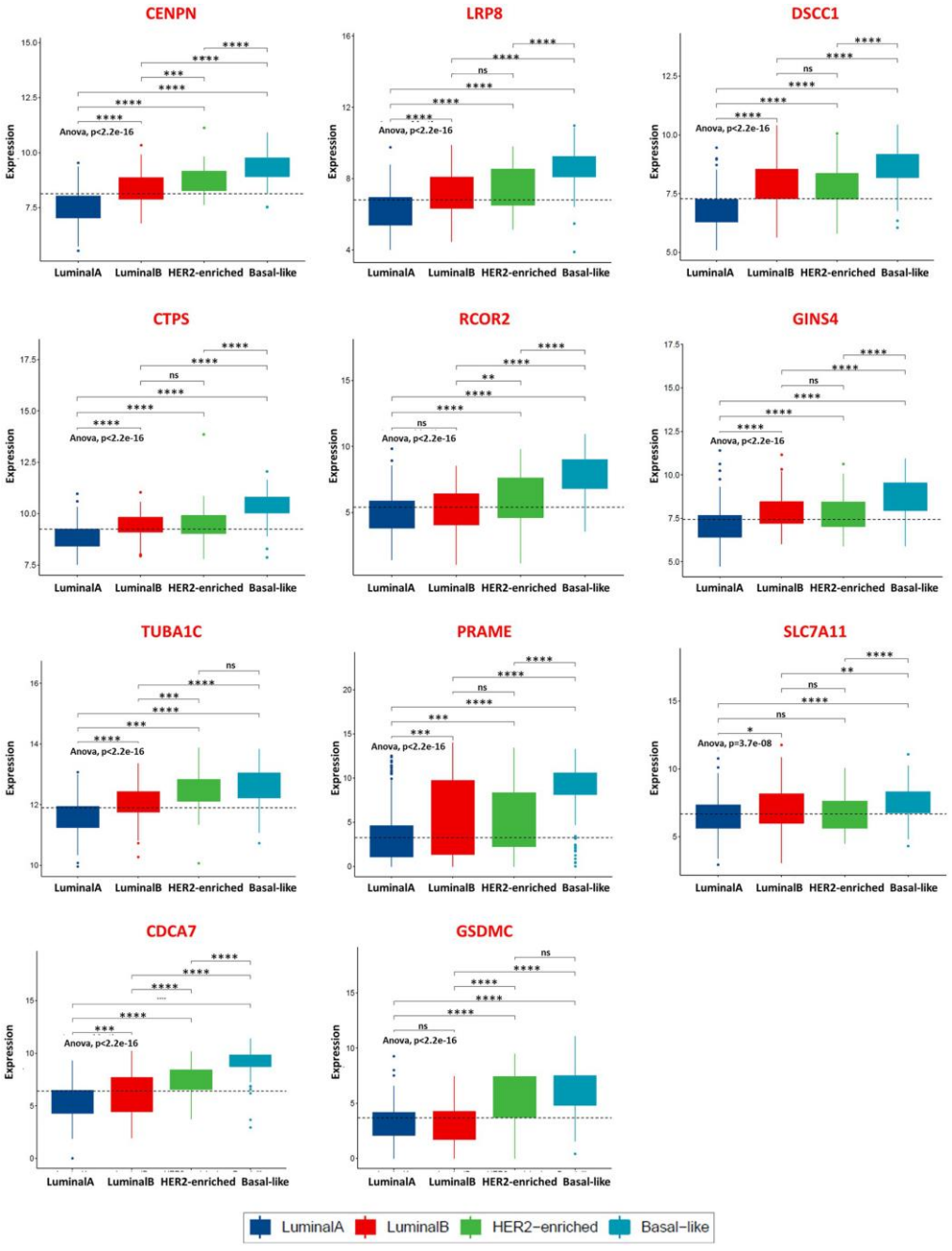

Figure 4.3. Switch genes with an unfavorable prognostic value in PAM50 BC subtypes. Gene expression levels of the 11 basal-like specific switch genes point out from the Kaplan-Meier survival analysis. The black dashed line reported in each plot indicates the median value used in the Kaplan-Meier survival analysis to split the low-expression and high-expression group. One-way ANOVA test was used to compare the means of the selected genes among the patients' groups. T-test was used to perform multiple pairwise comparisons and statistical significance was indicated by the star symbols (i.e., ns: $p > 0.05$, *:$p \leq 0.05$, ** :p ** 0.01, *** :$p \leq 0.001$, *** : $p \leq 0.0001$).

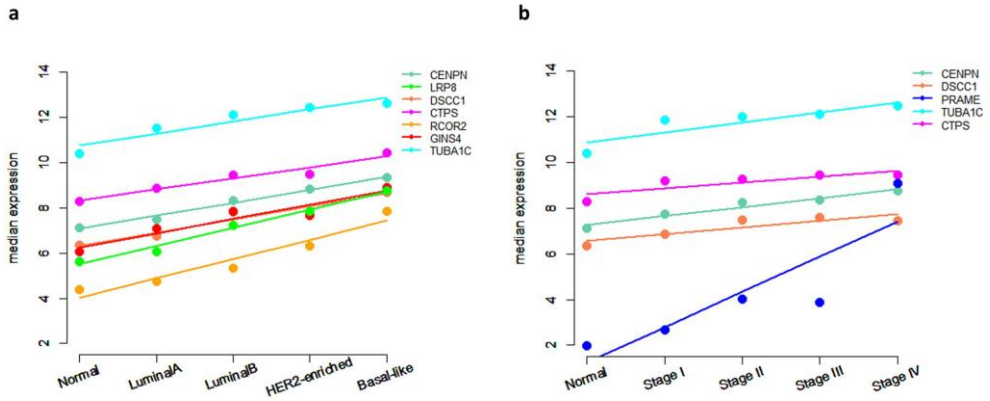

Figure 4.4. Linear regression model fitting. The median expression of the basal-like prognostic biomarkers is plotted against the phenotype varying from physiological to pathological condition (a) and against the pathological staging (b). Solid lines represent how the linear model fits the data. The results corresponding to the highest values of the model fitting index R-squared ≥ 0.9 for the subtype (a) and ≥ 0.7 for the staging (b) are shown**.**

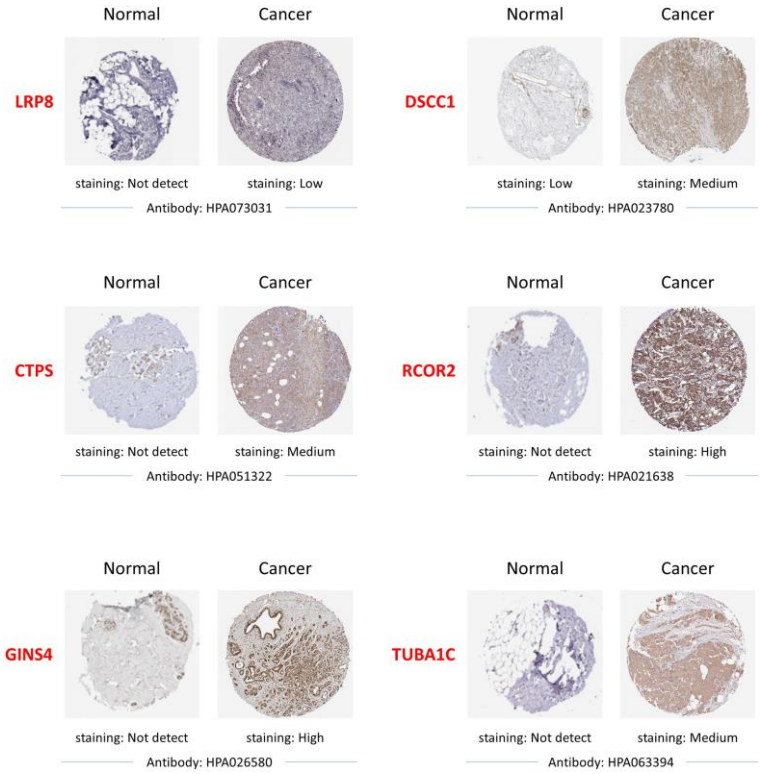

Figure 4.5. Immunohistochemistry results from the Human Protein Atlas. Representative immunohistochemistry images of the indicated switch genes in BC tissues and normal breast tissues obtained from the Human Protein Atlas.

Gene regulatory network of the basal-like prognostic biomarkers

To provide some hints on which transcription factors (TFs) could regulate the expression of the 11 switch genes proposed as prognostic biomarkers for basal-like subtype, we built a gene regulatory network by combining information on both computationally predicted and experimentally validated TF-target relationships. We firstly exploited Pscan web tool [94] to predict TFs putatively able to bind the promoter regions of the selected switch genes. Then, we filtered the Pscan predictions keeping only the TFs known to physically interact with at least one switch genes in the human interactome [8]. These TF-target relationships were finally complemented with those experimentally validated from TRRUST database [95]. The final gene regulatory network was composed of seven switch genes and twelve TFs, including well-known TFs that, if deregulated, contribute to neoplastic transformation as MYC, TP53 and NFKB1 (Fig. 4.6a and Table 4.1, DNA level). Interestingly, among the detected TFs, we also found four TFs (i.e., TP63, TWIST2, HIC1 and RARA) whose high expression appeared to be associated with the best prognosis for BC patients (Fig. 4.6b). In accordance with this result, we observed that these four favorable TFs reached their highest value in the patients affected by the less aggressive BC subtype, i.e., luminal A (Fig. 4.6c). It is worth noting that the other TFs of the gene regulatory network, in general, did not show a relevant increasing/decreasing trend across the different BC subtypes (Fig. 4.6c), indicating that the overexpression of their target basal-like specific switch genes maybe not ascribed to their transcriptomic variations but rather to other genetic and/or epigenetic alterations.

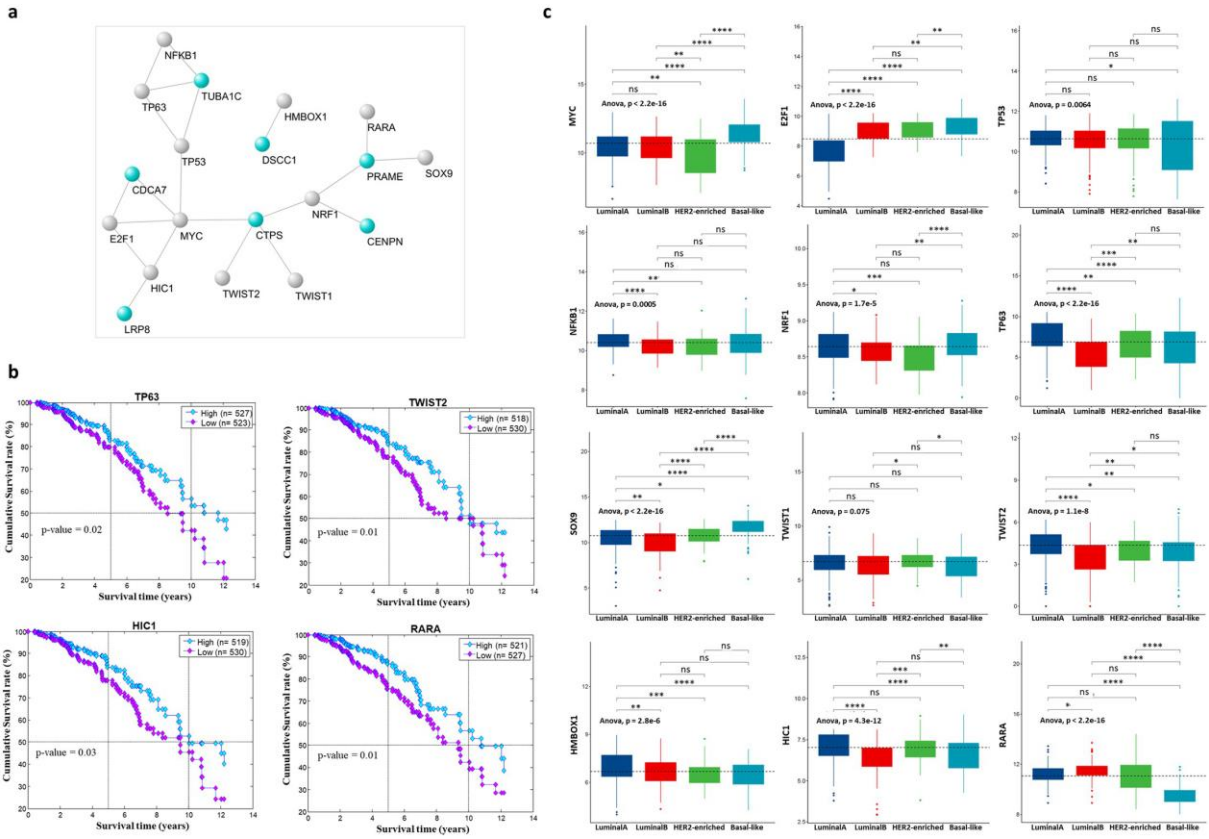

Figure 4.6. Gene regulatory network of the basal-like prognostic biomarkers. a) Network of the regulatory interactions among the identified switch genes and the known transcription factors (TFs). Light blue nodes represent switch genes; grey nodes represent transcription factors. b) TFs with a statistically significant prognostic value according to the Kaplan-Meier survival analysis. Kaplan-Meier analyzes to evaluate the correlations between the expression of the TFs and the OS in TCGA BC patients. Low- and high-expression groups refer to patients with expression levels lower and greater than the 50th percentile, respectively. c) Expression of the TFs in the gene regulatory network across the PAM50 BC subtypes. The black dashed line reported in each plot indicates the median value used in the Kaplan-Meier survival analysis to split the low-expression and high-expression group. One-way ANOVA test was used to compare the means of the selected genes among the patients' groups. T-test was used to perform multiple pairwise-comparisons and statistical significance was indicated by the star symbols (i.e., ns: $p > 0.05$, *:$p \leq 0.05$, ** :p ** 0.01, *** :$p \leq 0.001$, *** : $p \leq 0.0001$).

Genomic and epigenomic alterations of the basal-like prognostic biomarkers

Next, we investigated if the overexpression of the 11 basal-like prognostic biomarkers may depend on basal-like specific genomic alterations, such as Copy Number Variations (CNVs) and/or epigenomic alteration such as DNA methylation changes. We compared the CNVs and DNA methylation status of these 11 genes in basal-like subtype with respect to the less

aggressive BC subtype, i.e., luminal A. The CNVs analysis was performed on a total of 317 TCGA-BRCA patients (92 basal-like and 225 luminal A) for which CNVs data were available. Hierarchical clustering analysis on this data identified three main clusters and showed a different pattern of amplification and deletion in the selected genes between basal-like and luminal A patients (Fig. 4.7a). Interestingly, Cluster 1 appeared to be enriched in basal-like samples (64/67, 96%), whereas Cluster 2 (151/ 177, 85%) and Cluster 3 (71/73, 97%) were enriched in luminal A samples. Specifically, most of the basal-like patients belong to Cluster 1(64/92, 70%; highlighted in dark blue in Fig. 4.7a) and almost all luminal A belong to Cluster 2and Cluster 3(222/225, 99%; highlighted in green in Fig. 4.7a). Cluster 1 features were mostly related to DSCC1, GSDMC amplifications (> 1copy amplification per gene) along with TUBA1C deletion (>1 copy deletion per gene). Aberrant DNA methylation is another epigenetic alteration that plays a fundamental role in precipitating the development of a large and diverse number of human cancers [99]. For this reason, we investigated a potential correlation between DNA methylation patterns and mRNA expression profiles of the 11 basal-like prognostic biomarkers in basal-like and luminal A patients. The DNA methylation data analysis was performed on a total of 152 TCGA-BRCA patients (37 basal-like and 152 luminal A) for which DNA methylation data were available. Hierarchical clustering analysis on this data identified two main clusters and showed a different DNA methylation status of the selected genes between basal-like and luminal A patients (Fig. 4.7b). Cluster 1 was enriched in basal-like patients (25/37, 68%, highlighted in dark blue in Fig. 4.7b) and could be associated with a low methylation level especially for CENPN, PRAME, GSDMC and CTPS genes (Table 4.1, DNA level). On the other hand, Cluster 2isenriched in luminal A patients (98/115, 85%, highlighted in green in Fig. 4.7b). We compared the frequency of amplification and deletion events between basal-like and luminal A, using Fisher's exact test and we assessed the levels of methylation of the 11 basal-like prognostic biomarkers in the two groups (Fig. 4.6c). We observed different scenarios of CNV alteration along with DNA methylation status of the 11 basal-like prognostic biomarkers. CTPS, CENPN and PRAME had a higher frequency of amplification events (> 1 copy amplification per gene) in basal-like, higher frequency of deletion events in luminal A group ($p < 0.05$, Fisher exact test) and they hypomethylated in

basal-like patients (Fig. 4.7c). This first scenario showed the highest concordance between CNV alteration, DNA methylation levels and mRNA overexpression of these three genes in the basal-like group. Then, GSDMC was characterized by a higher frequency of amplification events in the basal-like group ($p < 0.05$, Fisher exact test) and was hypomethylated in basal-like patients (Fig. 4.7c), probably overlapping with its mRNA overexpression in the basal-like group. LRP8 was more amplified in the basal-like group and more deleted in luminal A patients ($p < 0.05$, Fisher exact test), supporting a putative correlation with its mRNA overexpression in the basal-like group. DSCC1 and CDCA7 had a higher frequency of amplification in basal-like patients ($p < 0.05$, Fisher exact test), which could be correlated with their mRNA overexpression in that group. Difficult to place was the result of TUBA1C, as we found that this gene has a higher frequency of deletion events in basal-like compared to luminal A group.

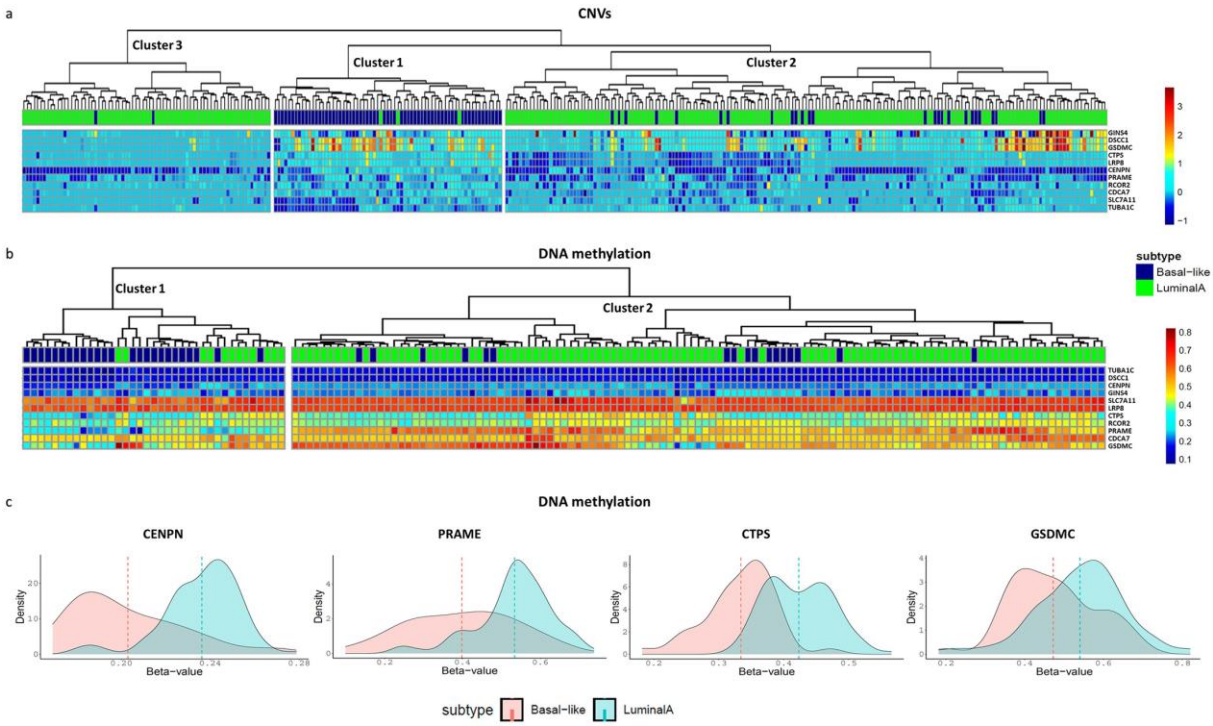

Figure 4.7. Genomic and epigenomic alterations of the basal-like prognostic biomarkers. a) Heatmap with dendrogram representing the unsupervised hierarchical clustering analysis based on CNVs data of TCGA-BRCA patients. The rows in the heatmap represent the 11 basal-like prognostic biomarkers. The columns correspond to basal-like and luminal ATCGA-B RCA patients: basal-like are indicated in dark blue and luminal Ain green. The cells of the heatmap represent the log2 segment mean value of CNVs (ranging from -1 up to 3.5), for which color code is indicated in the scale on the right-hand side of the figure. b) Heatmap with dendrogram representing the unsupervised hierarchical clustering analysis based on DNA methylation data of TCGA-BRCA patients. The rows in the heatmap represent the 11 basal-like prognostic biomarkers. The columns correspond to basal-like and luminal ATCGA-BRCA patients: basal-like are indicated in dark blue and luminal Ain green. The cells of the heatmap represent beta-value (ranging from 0to 1) extracted from Illumina 450k normalized data, for which color code is indicated in the scale on the right-hand side of the figure. c) Distribution plot of beta-value of CENPN, GSDMC, PRAME and CTPS genes in basal-like and luminal A patients. Dashed lines represent the mean of beta-values for each patients' group.

4.4. Discussion

In a recent study [5], the authors identified 108 switch genes specific for the basal-like subtype. The present analysis allowed to identify among them 11 basal-like specific switch genes with an unfavorable prognostic value (i.e., CTPS, CDCA7, GSDMC, LRP8, TUBA1C, CENPN, PRAME, SLC7A11, GINS4, DSCC1, RCOR2). We found that these 11 switch genes showed their highest mRNA overexpression in the basal-like compared to the other BC subtypes, and

this data further strengthens the hypothesis that these switch genes could be poor prognostic biomarkers in basal-like subtype affected patients (Fig. 4.3). After that, by a linear regression model, we found a straight-line relationship (from 0.7 up to 0.99) among CENPN, LRP8, DSCC1, CTPS, RCOR2, GINSS4, TUBA1C and PRAME with tumor subtypes and staging, while SLC7A11 and CDCA7 correlated only with subtypes. No correlation between GSDMC with subtypes and staging were found. The protein levels of these 11 switches in BC specimens were evaluated by querying the Human Protein Atlas. IHC results were examined confirming that 6(CTPS, LRP8, TUBA1C, DSCC1, GINS4, RCOR2) of the 11 proteins were overexpressed in BC tissues compared to normal ones. For the remaining proteins, IHC results were not yet available in the Human Protein Atlas (CDCA7, GSDMC, SLC7A11, PRAME and CENPN), nevertheless, the above citations confirmed us that all these switch proteins were overexpressed both in BC cell lines and tissues. These results led us to suspect their role in the neoplastic transformation. In fact, data from the literature, follow detailed, give to these molecules a tumorigenic characteristic being found deregulated in different human cancers including TNBC subtype, to make more robust our findings. CTPS1 (CTP synthase 1) gene, encodes an enzyme responsible for the catalytic conversion of UTP (uridine triphosphate) to CTP (cytidine triphosphate). This reaction is an important step in the biosynthesis of phospholipids and nucleic acids. Increased levels of the protein were linked to several mammalian cancer types such as sarcoma [100], hepatoma [101, 102] and leukemia [102], where the activity of this enzyme is both transformations- and progression linked, marking out this enzyme as an important target in the design of chemotherapy. More important, invitro experiments performed on BC cell lines demonstrated that CTP depletion results in a senescence-like growth arrest through activation of p53, whereas cells with mutated p53 undergo differentiation or apoptotic cell death [103]. LRP8 (LDL receptor-related protein 8) gene, encodes a member of the low-density lipoprotein receptor (LDLR) family. A recent study demonstrated that LRP8 was more strongly expressed in BC without hormone receptor expression (TNBC and HER2 positive) than in luminal tumors (Luminal A and Luminal B) [104]. Authors found that LRP8 depletion promoted apoptosis, impaired cell proliferation and colony formation suggesting that LRP8 has tumorigenic properties. These findings were further

confirmed by experiments showing that LRP8 depletion slowed tumor growth in an in vivo xenograft model. Moreover, inhibition of LRP8 was found to attenuate Wnt/β-catenin signaling to suppress BC stem cells (BCSCs) enriched in TNBC and responsible for chemoresistance and metastasis [104, 105]. Tubulin alpha-1C chain is a protein that in humans is encoded by the TUBA1C gene. TUBA1C is a member of the tubulin families and several studies demonstrated that its upregulation promotes oncogenesis and predicts poor prognosis in different tumor types [106, 107]. TUBA1C, TUBA1B and the β-tubulin isoform TUBB were found as isoforms with the highest expression levels compared to other isoforms in BC cell lines, and TUBA1C and TUBB were overexpressed in BC tumors compared to the normal breast tissues [108]. Also, the prognostic role of TUBA1C as a marker linked to the progression of BC was highlighted by [109], it was associated with lower OS in BC patients [110], and GTSE1 and TUBA1C combined predicted 100% probability of developing TNBC in whites [111]. Recently, overexpression of DSCC1 (DNA replication and sister chromatid cohesion 1) was found to increase proliferation, invasion and migration of BC cells, as well as its knockdown showed opposite outcomes [112, 113]. Besides, the authors found that DSCC1 could promote BC progression by activating the Wnt/β-catenin signaling and inhibiting p53 protein. PRAME nuclear receptor transcriptional regulator gene encodes an antigen that is preferentially expressed in human melanomas. The approved mutual link between BC and melanoma conditions emphasized the idea of utilizing this marker for targeting BC progression as well. Indeed, this protein was found to be involved in BC growth and metastasis and promote epithelial-to-mesenchymal transition in TNBC [114–116], suggesting that PRAME could serve as a prognostic biomarker and/or therapeutic target in TNBC. Cancer cell requires excess nutrients to meet their biosynthetic and bioenergetics needs and to maintain appropriate redox balance. Glucose and glutamine are important nutrients supporting cancer cell survival. SLC7A11 (solute carrier family 7member 11) gene encodes a member of a heteromeric, sodium-independent, anionic amino acid transport system that is highly specific for cysteine and glutamate; imports extracellular cystine coupled to the efflux of intracellular glutamate. SLC7A11 expression can be induced under various stress conditions, likely as an adaptive response to enable cells to restore redox homeostasis and maintain survival under stress

conditions [117]. The upregulation of SLC7A11 was found correlated with a poor response to treatment in different cancers including breast [118]. Recent evidence support that cancer cells upregulate SLC7A11 expression through diverse mechanisms to enhance their antioxidant defense and to suppress ferroptosis, a key tumor suppression mechanism [119]. The gasdermin (GSDM) superfamily consist of several molecules involved in cell pyroptosis. Recently, various studies have revealed the dysfunction and abnormal expression of the GSDM family in multiple human cancers, implying the potential roles in tumorigenesis. GSDMC (gasdermin C), a member of GSDM superfamily was found to promote cell proliferation in Colorectal Cancer [120], and high expression of GSDMC in BC [121] and lung adenocarcinoma [122] correlates with poor survival. CDCA7 (cell division cycle associated 7), was found to be elevated in various types of human cancer, including colon, lung, prostate and BCs [123], suggesting that this protein might play an important role in the development of cancer. Interestingly, CDCA7 isa DNA-binding protein and regulates the gene expression of the tumor-promoting effect of cMyc and E2F1. Recently the role of CDCA7 in TNBC subtype was partially clarified and authors found that high expression of CDCA7 was associated with metastatic relapse status and predicted poorer disease-free survival in patients with TNBC via transcriptionally upregulating the expression of EZH2 [124]. Centromere proteins (CENPs), which comprise 18 subtypes, are related dynamically to association and dissociation during mitosis with microtubule regulation. Among the CNPs, the protein encoded by CENPN (centromere protein N) gene, binds directly to the centromere-targeting domain of CENP-A. CENP-N depletion causes down-regulation of several CENPs and is considered essential for making anew centromere. Other functions of CENP-N, including its deregulation in BC are unclear, except the study that associated elevated expression of this protein with significantly increased mortality and risk of recurrence in BC smokers in contrast with non-smokers BC subjects [125]. RCOR2 (REST corepressor 2) is a protein-coding gene. Gene Ontology (GO) annotations related to this gene include DNA-binding transcription factor activity and transcription corepressor activity. To date, its involvement in the growth and progression of BC has not been investigated, yet. GINS4 is a subunit of the GINS complex (GINS1, GINS2, GINS3, and GINS4 subunits) involved in the initiation and progression of DNA replication

[126]. GINS4 was found highly expressed in lung, bladder and Colorectal Cancers, and its downregulation in the bladder and Colorectal Cancers inhibits growth and cell cycle and accelerate cell apoptosis progression in vitro as well as inhibits tumorigenesis in vivo [127, 128]. As for RCOR2 protein, GINS4 involvement in the growth and progression of BC has not been investigated, yet. Based on these findings, we felt compelled to understand which regulatory events might be responsible for their upregulation in basal-like subtype. So, we investigated whether the deregulated expression of the selected switch genes could be related to the activity of known transcription factors, copy number variation and DNA methylation. The construction of a gene regulatory network showed how these switch genes interact with several TFs known to be altered in cancer condition (MYC, TP53 and NFKB1), including in TNBC [129–131]. Nevertheless, we did not expect, but we were not surprised, that some of the identified TFs (TP63, TWIST2, HIC1 and RARA) were overexpressed in luminal A rather than in basal-like patients. So, being found also linked to a better prognosis, these results bring us to the hypothesis that these TFs could not be involved in the basal-like switch genes activation. Interestingly, we found that for most of the 11 switch genes their overexpression seems to be ascribed to genetic and/or epigenetic alterations. Indeed, we found that CTPS, CENPN, PRAME and GSDMC were found both hypomethylated and amplified in basal-like subtype as well as, except for GSDMC, also deleted in luminal A subtype; together these results are strongly in line with their expression data alterations found in the basal-like subtype. In the same way, also DSCC1 and CDCA7 were found amplified in basal-like, and CNVs profiles analysis demonstrated that the copy number amplification of two switch genes, DSCC1 and GSDMC, clustered for basal-like patients. Results on TUBA1C were somewhat controversial as this gene was found to be amplified in luminal A subtype and no genetic or epigenetic changes were found in basal-like subtype; for this switch gene seems that neither amplification nor methylation status is responsible for its overexpression in the basal-like subtype. Taken together these data can enrich the putative pathophysiological and prognostic role of these genes in BC basal-like subtype.

Chapter 5: Parallel multi-omics integration to identify repurposable drugs for COVID-19

In silico drug repurposing in COVID-19: A network-based analysis

5.1. Introduction

Drug repurposing consists of the use of an existing active pharmaceutical ingredient already on the market for a different indication [132]. This approach offers several advantages compared with the development of a new drug, including a faster and cheaper process due to consolidated knowledge regarding the drug 's safety and toxicity and higher success rates in introducing the drug to the market since it has already been tested in clinical trials [132]. The pharmacological base of drug repurposing relies on the fact that some diseases share common biological targets and that one drug may have several targets and thus may be able to treat different diseases [132]. In this framework, network-based approaches that leverage the Network Medicine principles could offer valuable help in identifying potential candidates for systematic drug repurposing [3, 133–136]. According to the Network Medicine construct, proteins associated to a specific disease tend to clusterize in the same network neighborhood of the human interactome and form disease modules that overlap for diseases showing significant molecular similarity, elevated co-expression, similar symptoms, and high comorbidity, whereas are well-separated for diseases that lack any detectable pathobiological relationships [3, 137]. As a consequence, if two disease modules overlap or are in the immediate vicinity within the interactome, local perturbations causing one disease can disrupt pathways of the other disease module as well, resulting in common clinical and pathobiological characteristics [137].

This hypothesis can be used to uncover new uses for existing drugs by identifying the disease modules located in the vicinity of drug therapeutic targets [33], or by identifying overlapping disease modules in the human interactome. [5, 24, 138–145].

Here, we performed a parallel multi-omics data integration of transcriptomics and interatomic data to study COVID-19 and other related inflammatory diseases. In the last three years, many computational tools for drug repurposing in COVID-19 patients were developed and most are based on three-dimensional analysis of the drug structure in relation to the viral and/or host targets and their binding affinities and interactions [146]. Among identified repurposable drugs, some target viral proteins, including antiviral drugs that inhibit viral RNA polymerase (e.g., favipiravir, remdesivir) or viral protease (e.g., lopinavir, prulifloxacin, tegobuvir, bictegravir, nelfinavir, and darunavir) [146]. Other drugs act on human cells and can block virus entry by several mechanisms, including inhibiting TMPRSS2 and other cell-surface proteases involved in SARS-CoV-2 activation (e.g., camostat mesylate and bromhexine), blocking clathrin-mediated endocytosis (e.g., chlorpromazine, baricitinib, and ruxolitinib), or preventing endocytosis by increasing endosomal pH (e.g., chloroquine and hydroxychloroquine). Although in vitro studies showed controversial results, these drugs have advanced to clinical trials either alone or in combination [147]. In parallel with targeting virus replication and cell entry, it is becoming evident that the host immune response plays a pivotal role in disease evolution. It was reported that patients with severe COVID19 disease present, in the early phases, hyperactivation of the innate immune response with cytokine storm resulting in a massive inflammatory response that later turns toward massive chronic basal inflammation characterized by a refractory immune state [148]. However, inappropriate adaptive immune response seems to play a crucial role in the late phase of the disease [149, 150], which is probably linked to immune checkpoint activation and immune system exhaustion [149]. This massive immune response has paved the way for testing several immunomodulatory agents in parallel with antiviral drugs [148, 151]. Several immunomodulatory and anti-inflammatory agents were tested to control cytokine storm. Tocilizumab, a monoclonal antibody against IL-6 receptors normally used for the treatment of diseases such as rheumatoid arthritis, was promising at first, though subsequent clinical trials did not provide unequivocal results on the benefit of tocilizumab in COVID-19 patients [152]. Corticosteroids appear to be effective in the treatment of COVID-19 patients, and many trials have confirmed that dexamethasone may be used for hospitalized subjects with severe SARS-

CoV-2 infection. Other drugs able to control inflammation, such as baricitinib, ruxolitinib, and eculizumab, are currently under clinical evaluation. Some anticoagulant and antiplatelet drugs have also been suggested to be effective in the treatment of COVID-19 patients. Heparin was found to limit hypercoagulability in COVID-19 patients, exert anti-inflammatory effects, and reduce mortality in hospitalized patients. Several clinical trials are currently evaluating heparin treatment efficacy in hospitalized patients with COVID-19 [153]. Despite these findings, robust clinical evidence is currently only available for a very limited number of drugs. Here, three different in-silico analyses were exploited. In fact, a single study might not be enough to cover the multiform clinical frame of the disease. We used transcriptomic data from whole blood cells, including all innate and adaptive immune system cells, of patients with COVID-19 and other inflammatory conditions, infections, or conditions with some clinical features in common with COVID-19. For each disease, we identified the genes that were most deregulated compared with healthy subjects. We then selected functionally related genes and verified that they were co-localized in the human interactome, thus generating a functional coherent disease module. This allowed us to identify drugs targeting proteins that were within or in proximity to the COVID19 module. Moreover, we also identified drugs that could be potentially repositioned for COVID-19 among those with an original medical indication for a disease whose module was in the COVID-19 neighborhood. Our in-silico analysis provided new pharmacological hypotheses to be explored and experimentally validated.

5.2. Material and Methods

Data collection and processing

Whole blood transcriptomic data for COVID-19 and 21 other diseases, including bacterial and viral infections, inflammatory diseases, immunodeficiency, primary lung, and coagulation disorders [7], were selected from the Gene Expression Omnibus (GEO) database. All datasets also included transcriptomic data of healthy controls.

Notably, the COVID-19 dataset was only recently deposited and is the first available concerning whole-genome gene expression data on whole blood cells. All patients were hospitalized for community- acquired lower respiratory tract infection with SARS-CoV-2 within the first 24 h of hospital admission. The human interactome was downloaded from Cheng et al. [8]. This version of the interactome is composed of 217, 160 protein–protein interactions connecting 15,970 unique proteins. Drug-target interactions were downloaded from DrugBank [154], which contains 13,563 drug entries, including 2627 approved small molecule drugs, 1373 approved biologics, 131 nutraceuticals, and over 6370 experimental drugs (released 22–04–2020) [154]. The target Uniprot IDs provided by DrugBank were mapped to Entrez gene IDs using the BioMart – Ensembl tool [155]. For our analysis, we selected a total of 1873 Food and Drug Administration (FDA)-approved drugs with at least one annotated target.

Disease-modulated genes and their localization in the human interactome

In order to identify genes that were most modulated by the disease, we computed differentially expressed genes between pathological and healthy conditions for every dataset using the following R packages: limma [56], to analyze microarray data, and Deseq2 [156], to analyze RNA-seq data. We mapped a list of the disease's modulated genes on the human interactome to identify the ones in the same connected subnetwork (i.e., the largest connected component), and thus functionally related. To test whether this subnetwork forms a statistically significant disease module, for each analyzed disease we randomly selected groups of proteins in the human interactome with the same size and degree distribution as the original list of disease deregulated genes and the following three metrics were computed: 1) the size of the largest connected component (LCC); 2) the number of interactions in the LCC; and 3) the total number of interactions. The three metrics were then z-score normalized by applying a degree-preserving randomization procedure, expecting a p value ≤ 0.05 for genes forming a statistically significant disease module [14]. Log2FC thresholds were chosen to guarantee the topological organization of disease deregulated genes in statistically significant modules.

SAveRUNNER

A detailed explanation of SAveRUNNER software is reported in Chapter 2 section 3 of this thesis

Network-based disease similarity

To measure the vicinity between the COVID-19 module and the other disease modules in the human interactome network, we used the non-Euclidean separation distance [137] defined in Eq. (5.1):

$$s(A, B) = p_{AB} - \frac{p_{AA} + p_{BB}}{2} \qquad (5.1)$$

where p(A,B) is the network proximity defined in Eq. (5.2):

$$p(A,B) = \frac{1}{|A| + |B|}\left[\sum_{a \epsilon A} d(a,b) + \sum_{b \epsilon B} d(b,a)\right] \qquad (5.2)$$

and d(a,b) is the shortest distance between the element a of module A and the element b of module B. A negative value for the separation measure indicates that two disease modules are in the same neighborhood of the human interactome, and thus they overlap; whereas a positive value for the separation measure indicates that two disease modules are topologically well separated. To evaluate the significance of module separation across two disease-specific modules (A, B), we built a reference distance distribution corresponding to the expected distance between two randomly selected groups of proteins with the same size and degree distribution as the original two disease-specific modules (A, B). The random selection was repeated 1000 times to build the reference distance distribution. The module separation measure was z-score normalized by using the mean and standard deviation of the reference distribution. Subsequently, the p value for the given z statistic was calculated. A p value ≤ 0.05 indicated that the separation between two disease-specific modules in the human interactome was more (or less) than that expected by chance.

Random Walk with Restart

The Random Walk with Restart (RWR) algorithm is another network- based approach to measure the closeness between the COVID-19 module and the other 19 disease modules in the human interactome network. RWR is an algorithm based on an intuitive concept that revolves around random walks. Given a random walker starting from a given node x, there are two different options at each iteration: either moving to one of its neighboring nodes or returning to x with a certain probability. Formally, the RWR algorithm can be described by Eq. (5.3):

$$R_t = \gamma W R_{t-1} + (1-\gamma)E \qquad (5.3)$$

where W is the network adjacency matrix, representing the matrix of transitions between nodes, whose element W[i,j] denotes the transition probability of going from node j to node i; E is the starting point vector, whose element E[i] is equal to 1 if i is a starting node, 0 otherwise; R_t is a probabilities vector, whose element R_t[i] denotes the probability of being at node i at iteration

t; γ is a number ranging in (0,1), and (1− γ) expresses the probability of "restarting" from the starting point node at each iteration. At iteration t = 0, the value of R_{t-1} is equal to E. The probabilities vector Rt will be iteratively calculated until the point of converge is reached (i.e., $R_t = R_{t-1}$, or the difference between the probability to stay and the probability to move is lower than a given threshold). Eventually, the RWR returns the vector R of the steady-state probabilities for each node in the network as output.

We ran RWR by considering the adjacency matrix W^{mxm}, built from the human interactome as a transition matrix, and the genes in the COVID-19 module as elements of the vector E. For each disease module, we averaged RWR steady-state probabilities corresponding with each module element and obtained a mean probability for each disease, i.e., the probability to reach it starting from COVID-19. This disease probability was then normalized by using the modified z-score defined in Eq. (5.4):

$$z_{mod} = c \cdot \frac{x - \hat{x}}{MAD} \qquad (5.4)$$

where x is the disease probability, $\hat{x}$ is the median value of distribution of all disease probabilities, MAD is the median absolute deviation defined as the median of the absolute difference of the observation from the sample median (i.e., median($|x - \hat{x}|$), and c is a scale factor equal to 0.6745, such that for normal distribution, zmod is equal to the standard z-score [157]. We termed this normalized disease probability COVID-19 *closeness*. Values of COVID-19 *closeness* that were outside the overall distribution pattern of the normalized disease probabilities were defined as outliers. A commonly used rule is to define a data point as an outlier if it is more than 1.5·IQR above the third quartile or below the first quartile. This means that low outliers are below $25^{th} - 1.5 \cdot IQR$ (i.e., the farthest diseases from COVID-19) and high outliers are above $75^{th} + 1.5 \cdot IQR$ (i.e., the closest diseases to COVID-19). Values of COVID-19 *closeness* that are outside the upper and lower quartiles are usually indicated as upper and lower whiskers, respectively. Diseases corresponding to high outliers as well as upper whiskers are more likely to be reached by the random walker starting from COVID-19.

5.3. Results

Functionally related and co-localized disease-related genes in the human interactome

In this study, we first compared the gene expression profile of COVID-19 and 21 diseases in which inflammatory and immune processes are involved with the profile of healthy controls to identify the highest modulated genes under pathological conditions. We mapped these genes on the human interactome, which is a network of proteins (nodes) in which the edges are the physical and functional interactions occurring between them, to evaluate whether they had the propensity to aggregate in local, disease specific neighborhoods of the human interactome, thus making them functionally related genes.

Table 5.1. Module search results for the analyzed datasets. LCC = largest connected component.

Disease name	Number of LCC nodes		Number of LCC edges		Number of total interactions	
	observation	*p-value*	*observation*	*p-value*	*observation*	*p-value*
Ankylosing spondylitis	216	0.6	825	0.005	827	0.008
Crohn's disease	399	2E-13	1120	3E-39	1147	2E-43
Chronic obstructive pulmonary disease (COPD)	163	0.0002	409	0.0002	418	0.0003
Chronic spontaneous urticaria	183	0.5	444	0.1	476	0.02
Community-acquired pneumonia	421	0.01	1347	1E-11	1359	2E-11
Common variable immunodeficiency	25	0.02	24	0.03	30	0.01
COVID-19	314	0.00001	441	3E-9	514	5E-17
Dermatomyositis	11	2E-11	23	2E-25	24	5E-17
H3N2 flu	170	0.03	270	0.001	282	6E-4
H1N1 flu	278	0.01	1309	2E-25	1331	5E-27
Inflammatory bowel disease	412	5E-16	1139	9E-38	1161	7E-42
Inclusion body myositis	3	0.01	4	0.01	5	0.007
Infective endocarditis	202	0.05	359	0.003	373	0.001
Primary lung cancer	333	0.0006	1185	2E-24	1192	2E-24
Polymyositis	12	0.004	14	0.00004	16	0.0002
Rheumatoid arthritis	178	0.03	321	1E-5	343	2E-7
Sarcoidosis	476	0.04	1731	4E-22	1745	3E-22
SARS	33	0.04	62	2E-6	65	3E-6
Septic shock	363	0.001	1202	1E-13	1223	6E-15
Tuberculosis	469	0.006	1829	8E-22	1843	7E-22
Ulcerative colitis	329	0.05	488	0.04	508	0.05
Venous thromboembolism	513	0.04	2975	0.00002	2977	0.00003

For each disease, we extracted the largest connected component from the subnetwork composed of genes that were modulated in that specific disease condition and verified whether these genes presented a statistically significant ability to generate a disease module. For subsequent analyses, we selected only the diseases that satisfied this module hypothesis, in

accordance with the organizing principles of network medicine [3, 14, 137]. Notably, we found that for all diseases analyzed (except for ankylosing spondylitis and chronic spontaneous urticaria), the deregulated genes formed statistically significant modules (Table 5.1). Thus, we considered 20 diseases, including COVID-19, for subsequent analyses.

Drug-disease module vicinity

To discover novel repurposable drugs and evaluate the magnitude to which a given drug can be repositioned for COVID-19, we exploited the recently developed SAveRUNNER algorithm [33]. The rationale behind SAveRUNNER builds on the hypothesis that for a drug to be effective against a specific disease, its associated targets (drug module) and the disease-specific associated genes (disease module) should be located nearby in the human interactome [8] (Fig. 5.1A). Using SAveRUNNER, we computed the similarity between each drug module and the COVID-19 module together with the corresponding statistical significance obtained through a degree-preserving randomization procedure. We obtained a weighted bipartite drug-disease network, where the link between a drug and a disease was appreciated if the corresponding drug targets and disease genes are located nearby in the interactome to a greater extent than what would be expected by chance (Fig. 5.1B). The weight of their interaction corresponds with the similarity measure between the corresponding drug and disease module. In our study, SAveRUNNER identified 399 repurposable drugs for COVID-19. Focusing on the top-ranked predicted drugs (similarity greater than 0.8), we observed molecules involved in the modulation of the coagulation system (e.g., heparin and tranexamic acid), antihistaminic drugs, mast cell stabilizers (e.g., chlorzoxazone and chlorpheniramine), anti-proliferative drugs including tyrosine kinase (TRK) inhibitors and antibiotics (e.g., larotrectinib and ciprofloxacin), alpha-adrenergic receptor agents (e.g., clonidine and prazosin); drugs affecting the central nervous system (e.g., perfenazine and droperidol), and inhibitors of the sodium voltage-gated channel alpha subunit 5 (SCN5A), which is involved in cardiac rhythm control (e.g., propafenone and prilocaine), among others (Table 5.2 and Fig. 5.1B).

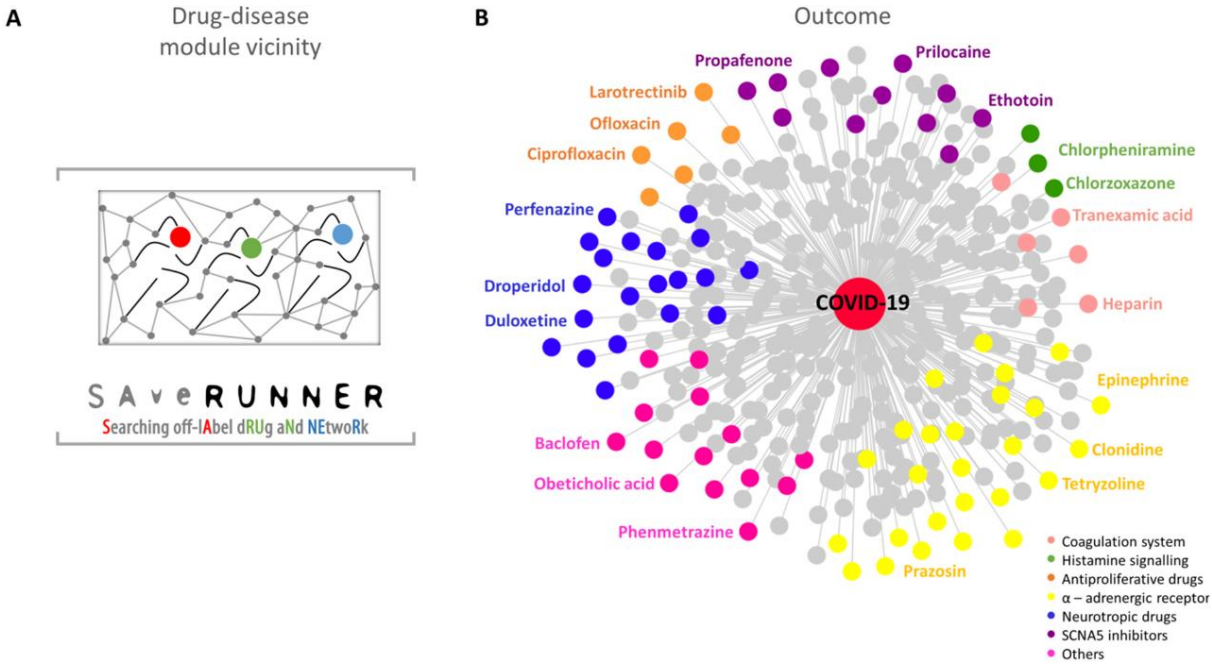

Figure 5.1. SAveRUNNER. A) The network-based algorithm used to identify off-label drug indications against COVID-19 [40]. B) The SAveRUNNER outcome network showing the high-confidence predicted drug-disease associations (p-value ≤ 0.05) connecting COVID-19 with 399 FDA-approved non-COVID-19 drugs. Drugs are colored according to the targeted pathways reported in the legend.

Table 5.2. Potential Drug Repurposing for COVID-19. SAveRUNNER-identified repurposable drugs for SARS-CoV-2 showing a similarity greater than 0.8. *Indicates drugs under investigation in clinical trials.

Approved Terapeutic Use	Drugs				Potential application in COVID-19
Antiplatelet/anticoagulant	Heparin*				Reduce the risk of thrombosis
Fibrinolytics	Streptokinase				
Pro-coagulants	Tranexamic acid*	Aminocaproic Acid			Limit viral entry
Polysulphates	Pentosan polysulphate	Chondroitin sulphate			
Histamine H1 - receptors antagonists	Diphenylpyraline	Chlorphenirami ne			Limit cytokine storm
Mast cells stabilizers	Chlorzoxazone				
Tropomyosin receptor kinase B inhibitors	Entrectinib	Larotrectinib			Reduce immune cells proliferation
Other Antiproliferatives	d-Serine	Podofillotoxin			
Fluoroquinolones	Ofloxacin	Ciprofloxacin			
α1-adrenergic receptors blockers	Nicergoline	Dapiprazole	Moxisylyte	Prazosin*	Limit cytokine storm
	Silodosin	Tamsulosin	Alfuzosin	Phenoxybenzamine	
	Phentolamine				
α2-adrenergic receptors agonists	Apraclonidine	Guanabenz	Guanfacine	Levonordefrin	
	Brimonidine	Clonidine*			
α1-agonists	Ergometrine	Metaraminol	Tetryzoline	Methoxamine	Sustain blood pressure in case of septic shock
	Midodrine	Oxymetazoline	Phenylephrine	Phenylpropanolamine	
	Xylometazoline	Epinephrine	Naphazoline		
Phenothiazines and Antipsychotics	Perphenazine	Thioridazine	Thiothixene	Periciazine	
	Pipotiazine	Prochlorperazin e	Flupentixol		
Serotonin-norepinephrine reuptake Inhibitors (SNRI)	Duloxetine	Sibutramine	Venlafaxine		Limit cytokine storm through modulation of Dopaminergic, Adrenergic and/or Serotoninergic receptors
Serotonin antagonist and reuptake inhibitors (SARI)	Nefazodone	Lorpiprazole			
Dopaminergics	Armodafinil	Diethylpropion	Modafinil	Solriamfetol	
	Benzphetamine	Fenoldopam			
Antidopaminergics	Droperidol	Methylergometr ine	Acetophenazi ne	Lumateperone	
Anti-epileptics	Fosphenytoin	Ethotoin	Mephenytoin		
Antiarrhythmic agents	Ajmaline	Encainide	Indecainide	Moricizine	Limit macrophages acrivation
	Tocainide	Propafenone	Vernakalant		
Local anaesthetics	Benzonatate	Prilocaine			

Disease-disease module vicinity

To find similarities between COVID-19 and the other 19 considered diseases, we implemented two network-based approaches: (1) network module separation, which quantified the topological distance between a disease module and the COVID-19 module in the human interactome network (Fig. 5.2A); (2) the RWR algorithm, which calculated the probability of a random walker reaching a disease module starting from the COVID-19 module (Fig. 5.3A). We observed that every separation value was statistically significant (see Materials and Methods), and we considered diseases whose separation values were less than the 15th percentile of the distribution of all separation values. We found H1N1 flu, Crohn's disease, inflammatory bowel disease, and septic shock to be the closest diseases to COVID-19 (Fig. 5.3B). From the RWR algorithm, we selected only those diseases that ranked within the outliers or upper whiskers, i.e., diseases that were more likely to be reached by the random walker starting from COVID-19 (see Materials and Methods). These diseases included dermatomyositis, polymyositis, and inclusion body myositis in the outlier category, whereas SARS, H1N1, and H3N2 flu appeared as upper whiskers (Fig. 5.3B). Interestingly, the RWR approach confirmed the results obtained with network module separation, in which H1N1 flu resulted as one of the closest diseases to COVID-19 in the interactome, with respect to the other analyzed diseases.

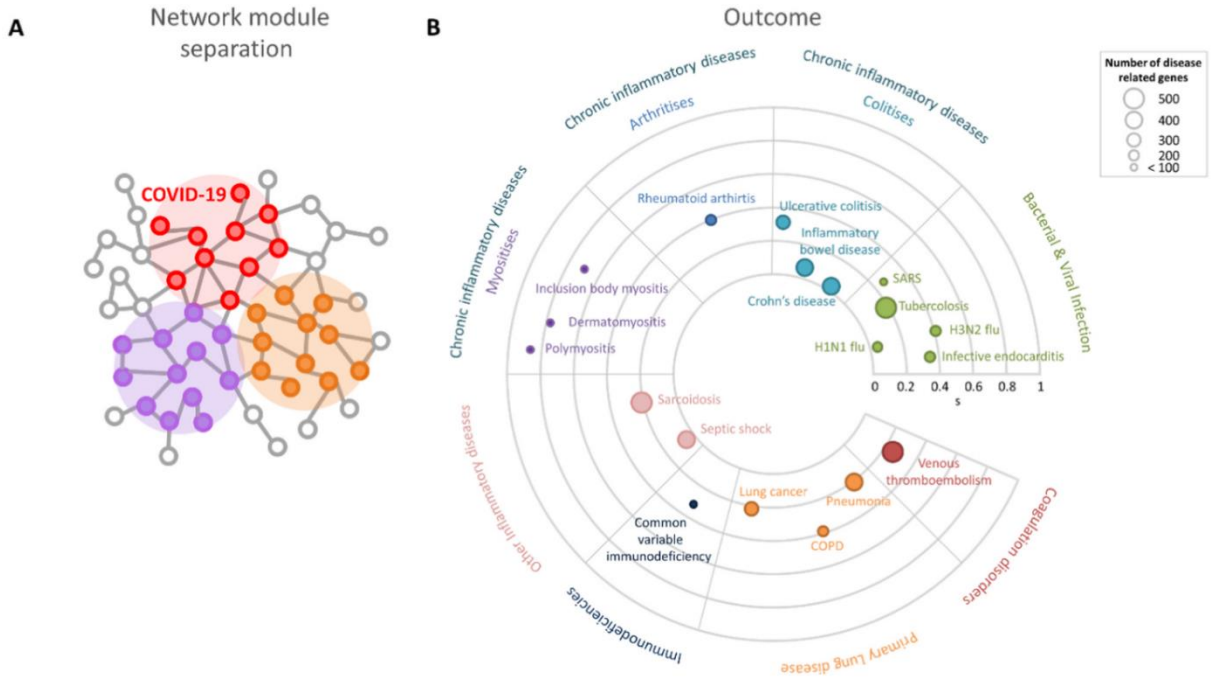

Figure 5.2. Network module separation. A) Schematic representation of disease-related gene modules and their topological distances in the human interactome. B) Radial plot reporting the network module separation measure(s) between COVID-19 and the other 19 diseases. Each disease is represented as a circle whose radius reflects the number of disease-related genes associated with it. The farther a disease is from the center, the more distant its module is from the COVID-19 module in the human interactome.

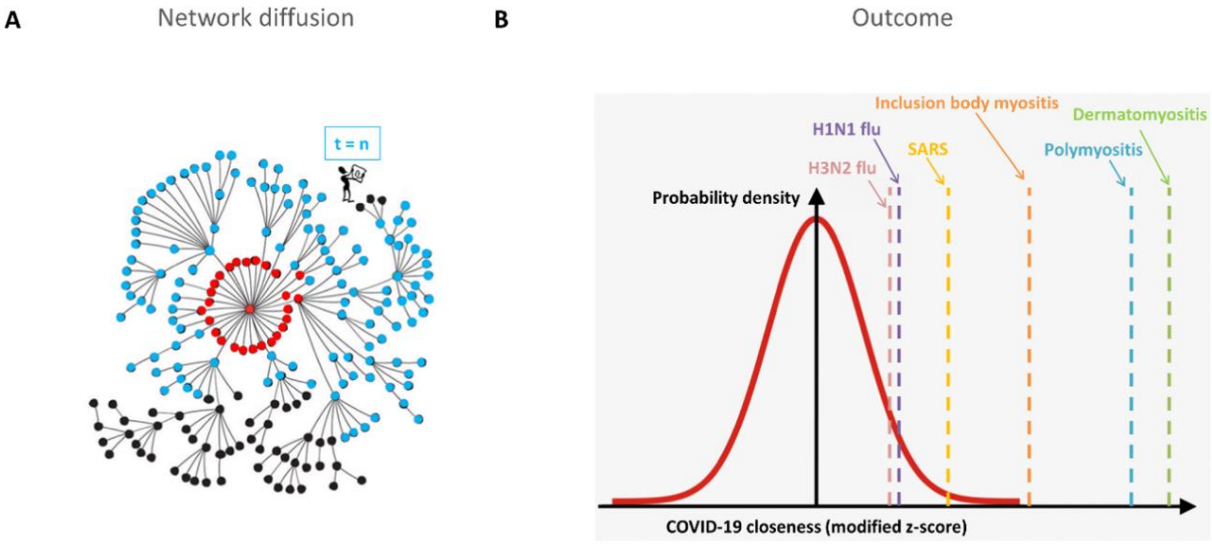

Figure 5.3. Random Walk with Restart (RWR). A) Sketch of the RWR algorithm applied on the human interactome. Red nodes represent the starting point nodes, light blue nodes represent all visited nodes at the end of the algorithm run, and black nodes represent the nodes of the human interactome that were not visited. B) Distribution of modified z-score-normalized probabilities (COVID-19 *closeness*) of nodes that were visited by the RWR algorithm starting from nodes belonging to the COVID-19 module. Diseases that are high outliers and upper whiskers are highlighted in the figure.

5.4. Discussion

To analyze the multiform clinical frame of SARS-CoV-2 from different scenarios, we used three different network medicine approaches to select drugs commonly used for the treatment of other conditions that could be repurposed for use in COVID-19 patients. It should be bear in mind that computational approaches are useful to generate new pharmacological hypotheses that need to be tested and validated experimentally. The potential use of the identified drugs for COVID-19 treatment will have to be carefully evaluated taking into account their possible side effects that can be found at DrugBank database website (https://go.drugbank.com/) [154]. Being aware of this important limitation we discussed the different identified drug classes considering the available information from literature focusing on their possible effects on the immune response modulation and infection natural history in COVID-19 patients.

SAveRUNNER drugs predictions

SAveRUNNER algorithm was used to identify repurposable drugs from DrugBank that could target the COVID-19 module or its neighborhood. Results show that drugs involved in the modulation of the coagulation system, histamine receptors, mast cell stability, immune cell proliferation, adrenergic receptors, serotonin receptors, or sodium channel SCN5A (sodium voltage-gated channel alpha subunit 5) function may have a great impact on immune system response in COVID-19 patients. SAveRUNNER identified drugs acting on COVID-19-related genes, regardless of their specific effects. Consequently, some drugs identified may be beneficial while others might be detrimental, and a critical and clinical evaluation that also considers the stage of SARS-CoV-2 infection is always essential [149]. Below we briefly discuss drug classes with a similarity greater than 0.8 (Table 5.2).

- *Drugs active on the coagulation system*

  SAveRUNNER analysis found different compounds with a mechanism of action involving the modulation of the coagulation system that could be repurposed for COVID-19. This observation positively correlates with SARS-CoV-2 infection, where a severe impairment in the coagulation system leading to thrombosis is frequently observed [158]. Indeed, heparin was tested as a prophylactic treatment and was demonstrated to improve disease-specific mortality [159]. Heparin clearly emerged in our analysis, supporting the potential and accuracy of SAveRUNNER software in identifying repurposable drugs. Several clinical trials testing the effect of different heparin formulations in COVID-19 are ongoing, including NEBUHEPA (NCT04530578), which is evaluating the effect of nebulized heparin in patients with COVID-19-related acute respiratory distress syndrome (ARDS). The SAveRUNNER analysis identified chondroitin sulphate and pentosan polysulphate, which showed a lower activity compared to heparin in inhibiting platelet aggregation inhibition and was also shown to interact with spike proteins, thus reducing virion internalization and blocking inflammation and the cytokine storm associated with antigenic epitope exposition [160]. Another drug that has emerged is streptokinase, a fibrinolytic drug used in severe acute thrombosis. A case series report showed that streptokinase is

effective in COVID-19 patients [161]. Intriguingly, the analysis also identified tranexamic acid, which is normally used as a pro-coagulant agent during bleeding due to its ability to inhibit circulating plasminogen and other proteases, leading to thrombus stabilization [162]. Since plasminogen is one of the proteases necessary for spike-protein cleavage, thus allowing virion interaction with angiotensin-converting enzyme 2 (ACE2)-expressing cells [162], the possibility of using tranexamic acid in COVID-19 patients is currently being tested in clinical trials (NCT04338126).

- *H1-inhibitors and mast cell stabilizers*

Histamine is a proinflammatory molecule produced by mast cells that mediates type I hypersensitivity reactions. Mast cell abundance in human airways supports the potential relevance of this mediator in SARS-CoV-2 infection. Short-term effects of mast cell degranulation and histamine release include increased vascular permeability, vasodilation, immune cell recruitment, and platelet activation [163, 164]. Moreover, histamine release induces interleukin 6 (IL-6), leukotrienes, and the production of other inflammatory prostaglandins, thus triggering the activation of innate response [164]. Mast cell stabilization and blocking histamine signaling might be fundamental in controlling the cytokine storm, which is typical of the early stages of SARS-CoV-2 infection [164]. Indeed, our analysis highlighted a potential repurposing of chlorzoxazone, a mast cell stabilizer that blocks calcium channels and inhibits degranulation as well as leukotriene and cytokine production [165]. Similarly, diphenylpyraline and chlorpheniramine, which are commonly used antihistaminic drugs, could potentially block the early phase of cytokine storm during SARS-CoV-2 infection [164]. Of note, emerging evidence currently supports a direct antiviral effect of targeting the histamine pathway in SARS-CoV-2 in vitro [166].

- *Antiproliferative drugs and antibiotics with antiproliferative activity*

This category includes both tyrosine kinase inhibitors (TRK inhibitors) [Entrectinib, Larotrectinib] and drugs inhibiting the activity of human topoisomerase II, such as fluoroquinolones (ofloxacin, ciprofloxacin) [167, 168]. We can reasonably expect that their

inhibition of cell proliferation might be useful to limit the immune cell proliferation and consequent cytokine storm during SARS-CoV-2 infection [149, 169, 170]. Of note, the potential relevance of TRK inhibitor repurposing was confirmed by other drug repurposing studies [171]. In addition, fluoroquinolones were found to have antiviral activity in vitro, thus supporting a potential benefit in COVID-19 patients in limiting bacterial superinfection [172].

- *α-adrenergic receptor agents*

SAveRUNNER analysis found that α1-antagonists and α2-agonists could be repurposed for COVID-19. Indeed, data suggest that α1 adrenergic receptor activation may induce pro-inflammatory cytokine secretion in innate cells, thus suggesting the possibility that blocking α1 adrenergic receptors might limit the cytokine storm that characterizes severe COVID-19 patients [173]. Rose et al. found that men with a confirmed or suspected COVID-19 diagnosis who were on treatment with α1 adrenergic receptor antagonists prior to hospitalization had reduced in-hospital mortality (OR: 36%) compared to those who were not taking α1 adrenergic receptor antagonist medications [174]. As such, blocking alpha-adrenergic signaling in the immune system might be successful, particularly in early-stage infection, and indeed prazosin (α1-antagonist) is now being tested in a clinical trial (NCT04365257). Similarly, α2-agonists such as clonidine could be repurposed during COVID-19 to limit ARDS and inflammatory response [175]. Intriguingly, α1-agonism could stimulate immune system response and could be considered in COVID-19 patients in case of septic shock [176].

- *Drugs active in the central nervous system*

Several drugs active in the central nervous system were identified as repurposable for COVID-19 following SAveRUNNER analysis, particularly tricyclic compounds, drugs active in serotonin signaling (i.e., SNRI and SARI), and dopaminergic and dopamine antagonists. Interestingly, Hoertel et al. [177] suggested a possible role of both SSRI and non-SSRI antidepressants in reducing the risk of death and intubation in patients

hospitalized for SARS-CoV-2 infection. Recent evidence highlighted that these drugs may influence both innate and adaptive immunity:

- Phenotyazine and antipsychotic drugs have known effects on α-adrenergic and histaminergic receptors and could therefore act as possible immune system modulators [164, 178]. Moreover, different drugs belonging to this class were shown to possess antiviral properties, suggesting potential repurposing for COVID-19 [179].

- SNRI and SARI might also have a modulatory effect on the immune system, particularly on lymphocytes, which express serotonin receptor 5-hydroxytriptamine 2 (5-HT2). In rat and mouse models, fluoxetine treatment produced a significant reduction in TNFα and IFNγ production. In SARS-CoV-2 infection, drugs modulating the serotonin signaling might be repurposed as cytokine storm regulators [180].

- Dopamine receptors are expressed in different immune cell subtypes and their effect on immune response modulation is still debated. Dopaminergic stimulation reduces TNF-α and ROS production in neutrophils, though it stimulates mast cell degranulation and monocyte chemotaxis. In addition, dopamine stimulation appeared to be protective in a mouse model of peritonitis [181]. In this scenario, targeting the dopaminergic pathway emerged as a potential strategy to limit cytokine storm during COVID-19. However, a clear role of dopaminergic system activation in the context of immune response is debated, and more research is necessary to better define the role of dopamine in immune system modulation [181, 182].

- *Drugs acting on SCN5A sodium channels*

SCN5A sodium channels are commonly expressed in excitable tissue, particularly neurons and myocytes. Most identified SCN5A inhibitors are anti-arrhythmic drugs,

local anesthetics, or anti-epileptics. Recent evidence highlighted that the SCN5Achannel is involved in macrophage activation and plays a pivotal role in host antiviral response by inducing the phosphorylation and nuclear translocation of the transcription factor ATF2 [183]. Moreover, in LPS-activated macrophages, SCN5A regulates endosomal acidification and stimulates phagocytosis. Although this process protects the host during acute infections, it may also promote tissue injury [184]. Interestingly, endosomal and lysosomal acidification allow viral cellular entrance [149]. In this context, SCN5A inhibitors may contribute to controlling both systemic inflammation and viral infection. Other studies reported that macrophages present an anti-inflammatory phenotype in mice expressing human SCN5A [185]. Consistent with these observations, propafenone, a SCN5A inhibitor used for its anti-arrhythmic properties, was suggested as a possible inhibitor of spike protein cleavage and SARS-CoV-2 cellular penetration [186]. However, SCN5A inhibitors present a series of limitations that need to be considered, including arrhythmia. Among other drugs, SAveRUNNER also identified baclofen, a GABA-B agonist commonly used in neurodegenerative diseases as an antispastic. A recent computational analysis identified it as a TNF α inhibitor [187]. Since TNFα is one of the main inflammatory signals of innate immunity, baclofen might be repurposed as a mitigator of cytokine storm in SARS-CoV-2 infection [149]. Studies in mouse models found that FXR activation reduces the levels of circulating NF-KB and other proinflammatory cytokines, such as MCP-1. Obeticholic acid, an FXR agonist mainly indicated for the treatment of biliostasis, was shown to exert anti-inflammatory activities observed in the reduction of liver inflammation [188] and was identified by a computational study and proven in vitro to inhibit SARS-CoV-2 ligation to human ACE2 [189].

Network module separation drugs predictions

Network module separation was used to find diseases with a module close to that of COVID-19. Our hypothesis is that drugs used to treat these diseases may also be beneficial in COVID-19 patients. We found that septic shock, Crohn's disease, inflammatory bowel disease (IBD),

and H1N1 flu modules were very close to the COVID-19 module. Consistently, it was observed that COVID-19 patients with increased immune system activation present an elevated incidence of sepsis [149]. Moreover, the significative proximity between the COVID-19, IBD, and Crohn's disease modules is not surprising since literature data support that COVID-19 and IBD immune system activation share several similarities and that some drugs used for IBD appear to also be effective for COVID-19 patients [190]. Of the drugs reported in the DrugBank database for septic shock treatment, both epinephrine and norepinephrine were also identified by the SAveRUNNER algorithm, suggesting that SARS-CoV-2 infection and septic shock share common epinephrine or norepinephrine targets. Naloxone is another drug used to treat septic shock and is currently under investigation in COVID-19 patients. Mesalazine and sulfasalazine, two anti-inflammatory drugs, are used for Crohn's disease treatment and have also been identified by the SAveRUNNER algorithm. Two monoclonal anti-tumor necrosis factor alpha antibodies (adalimumab and infliximab) and many corticosteroid drugs (budesonide, methotrexate, prednisolone, prednisone, and hydrocortisone) used for Crohn's disease are currently being tested in COVID-19 in several clinical trials. This is not surprising since the role of corticosteroids as anti-inflammatory drugs is well known. Of note, the RECOVERY randomized clinical trial showed that the use of dexamethasone resulted in lower 28day mortality in those receiving either invasive mechanical ventilation or oxygen alone [191]. In addition, an increase in TNF α, a strong pro-inflammatory cytokine, was observed in patients affected by COVID-19 [151]. Some evidence suggests that TNFα inhibition may downregulate ACE2 expression and shedding, thus reducing viral entry into cells [151]. Several drugs, including an antiviral drug (oseltamivir), anti-inflammatory drugs (naproxen and acetylsalicylic acid), a beta-2 adrenergic receptor agonist (salbutamol), and an analgesic/antipyretic drug (acetaminophen), are used for flu treatment and are currently in clinical trials. However, several antiviral drugs, including oseltamivir, do not seem to exert a robust effect against the SARS-CoV-2 virus [147]. The immune system and the sympathetic nervous system are highly connected through post-ganglionic sympathetic nerve fibers, which secrete norepinephrine that innervates both primary and secondary lymphoid tissues. Both innate and adaptive immune system cells express adrenergic receptors, mainly β2. There is

evidence that glucocorticoids and other β2-receptor agonists suppress macrophage secretion of TNFα and other inflammatory cytokines in response to lipopolysaccharide, reducing inflammatory damage. It was shown that norepinephrine drives alternative M2 macrophage development, characterized by an anti-inflammatory phenotype [173], and that β2 adrenergic receptors modulate the activation of several innate immune cells and consequently modulate T and B cell response. However, the role of β2 receptors on the immune system is still debated and some authors have reported a pro-inflammatory role [192].

RWR approach drugs predictions

The RWR algorithm was used to search for diseases whose drugs may also perturb the COVID-19 module. We identified H1N1 and H3N2 flu, SARS-CoV-1 infection, dermatomyositis, polymyositis, and inclusion body myositis. These findings support the documented similarity between SARS-CoV-2 and SARS-CoV-1 infections [149]. Interestingly, both the network module separation approach and the RWR highlighted the disease modules of H1N1 flu and SARS-CoV-2 infection. Several corticosteroids used for dermatomyositis and polymyositis (prednisolone, prednisone, hydrocortisone, methylprednisolone, betamethasone, and methylprednisolone hemisuccinate) and corticotropin are currently in clinical trials for COVID-19. Notably, triamcinolone is a corticosteroid drug that is used for diseases identified by both the network module separation approach and RWR (Table 5.3). In conclusion, we used a network medicine approach to generate new pharmacological hypotheses for the COVID-19 treatment. While some of the in-silico identified drugs are already under evaluation in clinical trials, others were proposed by expert opinion or other computational studies to be potentially effective in COVID-19 patients. SAveRUNNER analysis also identified novel drug categories, including drugs known to be active in the central nervous system and sodium channel blockers, that could be repurposed in COVID-19 patients. The in-silico methodology has many limitations, including the need to test and validate the identified drugs. Indeed, the potential benefits as well as the risks of possible adverse reactions, mainly due to the multitarget action of many compounds, must be carefully evaluated and proved. Moreover, an efficient

translation should also consider pharmacokinetic aspects that could impact the clinical applicability of repurposed drugs.

Table 5.3. Drugs used for the diseases identified by the network module separation approach and RWR. *indicates drugs identified by SAveRUNNER. Drugs currently under investigation in clinical trials, as reported in DrugBank, are highlighted in bold.

Network module separation			**Random Walk with Restart**	
Septic shock	**Crohn's disease**	**H1N1 flu**	**Dermatomyositis**	**Polymyositis**
Epinephrine*	Mesalazine*	**Acetaminophen***		**Corticotropin**
Norepinephrine*	Sulfasalazine*	Cetirizine*		**Methylprednisolone**
Naloxone	**Adalimumab**	Chlorpheniramine*	**Betamethasone**	
	Budesonide	Phenylephrine*	Bupivacaine	
			Methylprednisolone	
	Infliximab	Pseudoephedrine*	**hemisuccinate**	
		Naproxen		
		Oseltamivir		
		Salbutamol		
		Acetylsalicylic acid		
		Ascorbic acid		
	Methotrexate		**Methotrexate**	
	Triamcinolone		Triamcinolone	
	Prednisolone		**Prednisolone**	
	Prednisone		**Prednisone**	
	Hydrocortisone		**Hydrocortisone**	

Chapter 6: Parallel multi-omics data integration for studying COPD

Network-based integration of gene expression and methylation data to build a consensus network for COPD

6.1. Introduction

Chronic obstructive pulmonary disease (COPD) is a heterogeneous, chronic inflammatory process airways often involving destruction of adjacent alveoli and vasculature. Symptoms range from the chronic productive cough to debilitating dyspnea. COPD is determined by both genetic and environmental factors and is the third leading cause of death worldwide. Cigarette smoking is a major environmental risk factor for COPD, however only a minority of smokers develop COPD and there is significant variability in lung function across smokers with similar cigarette exposure histories [193]. Among the primary difficulties in treating COPD patients is the clinical heterogeneity of the disease that is likely a result of genetic variation, and the disease progression that can vary from stability to exacerbation. Various contributors to COPD pathogenesis were also suggested, including protease-antiprotease imbalance, oxidant-antioxidant imbalance, cellular senescence, autoimmunity, chronic inflammation, deficient lung growth and development, and ineffective lung repair. However, the pathobiological mechanisms for COPD remain incompletely understood [194].

Approaches aiming to gain key insights into the genes driving the underlying disease molecular machinery are mainly based on single-omics data analysis. Among them, SWIM was recently applied to transcriptomic data from lung tissues of two well-characterized COPD case-control populations to study the differences between lung samples from normal subjects (represented by smokers with normal spirometry) and COPD cases [142].

Due to the important regulatory function of the gene expression, the DNA methylation of CpG sites is another important source of variability between diseased and normal cells [195]. Indeed, methylation at the CpG sites in the gene promoter can resulted in silencing of the gene expression [196]. Environmental exposure such as cigarette smoke can influence the DNA methylation of the genes [197, 198]. Recently, DNA methylation array analysis were used to identify genes potentially involved in COPD [199].

Despite the successful progresses obtained by using single omics data analysis, approaches aimed to integrate the different multi-omics data are needed to study the flow of information between different biomolecular layers helping to gain insights into the molecular mechanisms of the disease. Here, we performed a parallel multi-omics data integration by applying SWIM on the transcriptomic and DNA methylation data of a cohort of 823 former-smoking COPD cases and controls derived from a lung tissue cohort of the Lung Tissue Research Consortium at the Channing division of Network Medicine Division at Brigham and Women's Hospital. We started by selecting from the entire COPD cohort genes that are in common between RNA-seq and DNA methylation data, and then we applied SWIM on the two datasets, separately. The resulting correlation networks were then integrated by using the differential gene correlation analysis (DGCA) [200] that allowed us to define a *consensus network*, where nodes are genes both differentially expressed and differentially methylated in COPD cases. A link occurs between two nodes if they were interacting in both RNA-seq- and DNA methylation-based correlation networks (Fig. 6.1).

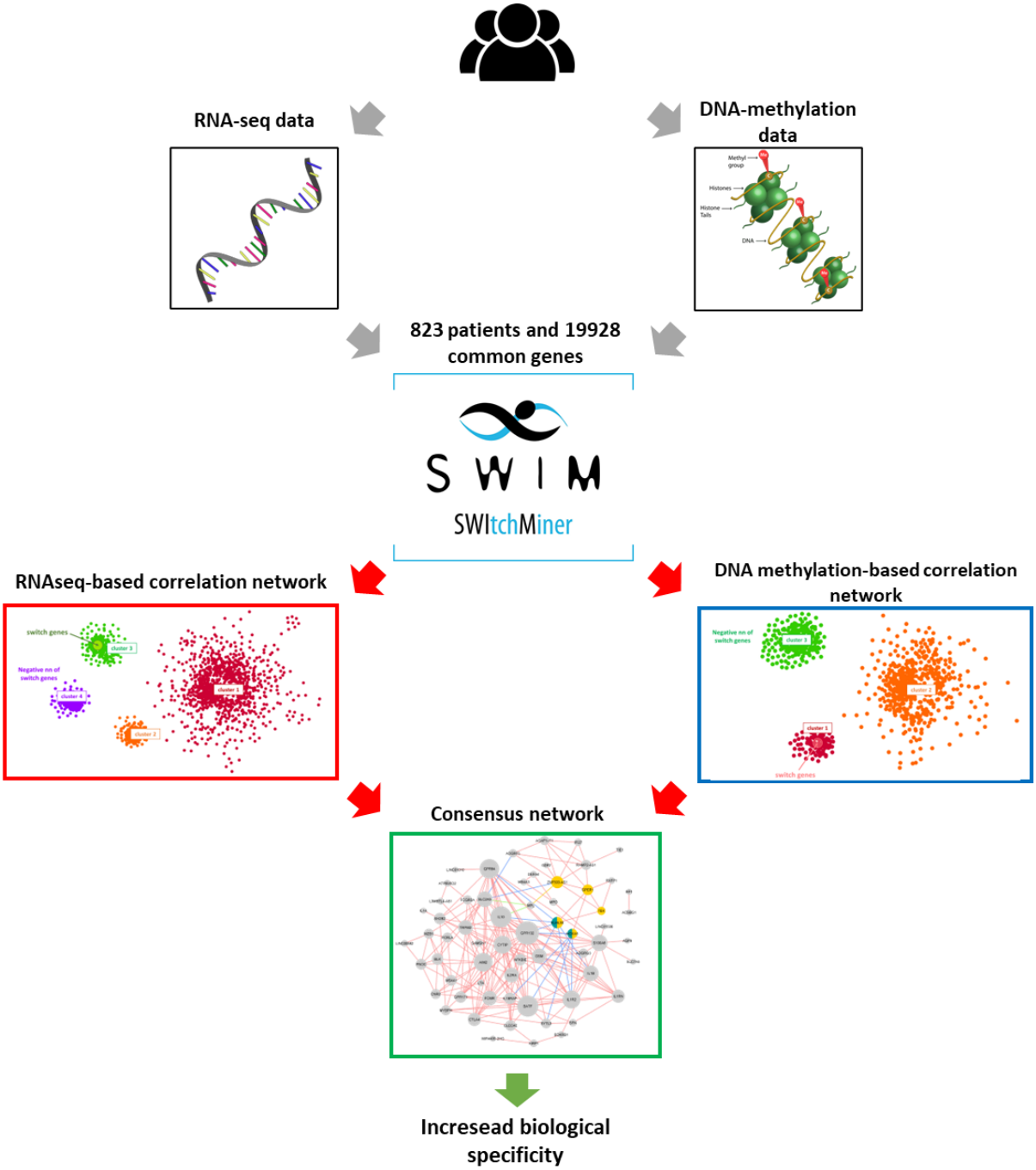

Figure 6.1. Study design. We selected from the entire COPD cohort those genes that are in common between RNA-seq and DNA methylation data and we applied SWIM tool on the two datasets, separately. Then, we integrated the two correlation networks through the differential gene correlation analysis. This analysis led to define a *consensus network*, where nodes are genes that are both differentially expressed and differentially methylated in COPD cases with respect to controls. A link occurs between two nodes if they were interacting in both RNA-seq- and DNA methylation-based correlation networks.

6.2. Materials and Methods

Data collection and processing

We considered 452 former-smoking COPD cases and 371 controls (823 in total) derived from a lung tissue cohort of the Lung Tissue Research Consortium from the Channing division of Network Medicine Division at Brigham and Women's Hospital, for which RNA-seq and DNA methylation data were available.

RNA-seq data preprocessing and normalization

The RNA-seq matrix was retrieved as raw count matrix for which Batch correction and normalization pre-processing were performed. The raw counts were corrected for Batch effects related to the batch plate used during the sequencing experiment using ComBat-seq algorithm [201]. Afterwards, a normalized matrix was produced by using EDASeq R package which provide within-lane normalization and between-lane normalization procedures that also consider the GC-content effect on the gene counts detected in the sequencing experiment.

DNA methylation data preprocessing and normalization

The DNA methylation matrix was retrieved as normalized matrix for which Batch plate correction and gene-level summarization processes were performed. The DNA methylation matrix were corrected for Batch plate effects related to the array experiment using Combat [202]. We were interested in studying a gene-level DNA methylation of the promoter region, which is related in many case to the repression of transcription [203]. To this aim, we identified the probes that fall 1500 bp upstream to the Transcription Starting Site (TSS1500) of each gene and averaged the values.

SWIM application on RNA-seq and DNA methylation data

An explanation of SWIM software is given in the Chapter 2 section 3, of this thesis

6.3. Results

We exploited the SWIM network-based analysis to build and analyze the RNA-seq and DNA methylation-based correlation network associated to COPD disease. Starting from the 19928 genes selected, SWIM proceeded calculating the differentially expressed genes (DEGs) and differentially methylated genes (DMGs). We obtained 2258 significantly differential expressed genes (DEGs) at 5% false discovery rate (FDR), of which 1180 (52.6%) upregulated in COPD cases and 1078 (47.4%) downregulated. Concerning DNA methylation data, we discovered 1002 DMGs at 5% false discovery rate (FDR), of which 283 DMGs (28,3%) hypermethylated and 718 DMGs (71,7%) hypomethylated in COPD cases.

SWIM used DEGs and DMGs data to build the correlation network based on the Pearson correlation coefficient. A correlation threshold equal to 0.45 (i.e., 95th percentile of the entire correlation distribution) and to 0.57 (i.e., 87th percentile of the entire correlation distribution) was set for the RNA-seq data and for the DNA methylation data, respectively.

The RNA-seq-based correlation network of COPD status contained 1907 nodes and 169358 edges, including 1182 date hubs, 415 party hubs, and 142 fight-club hubs. The DNA methylation-based correlation network specific for COPD contained 915 nodes, including 311 dates, 467 party, and 87 fight club hubs. SWIM found 131 switch genes in the RNA-seq specific correlation network (Fig. 6.2a), most of them resulted as downregulated in COPD cases (n 110/131, 84.0%). Meanwhile, SWIM found 66 switch genes in the DNA methylation-based correlation network and all of them were hypermethylated in COPD cases (Fig. 6.2b).

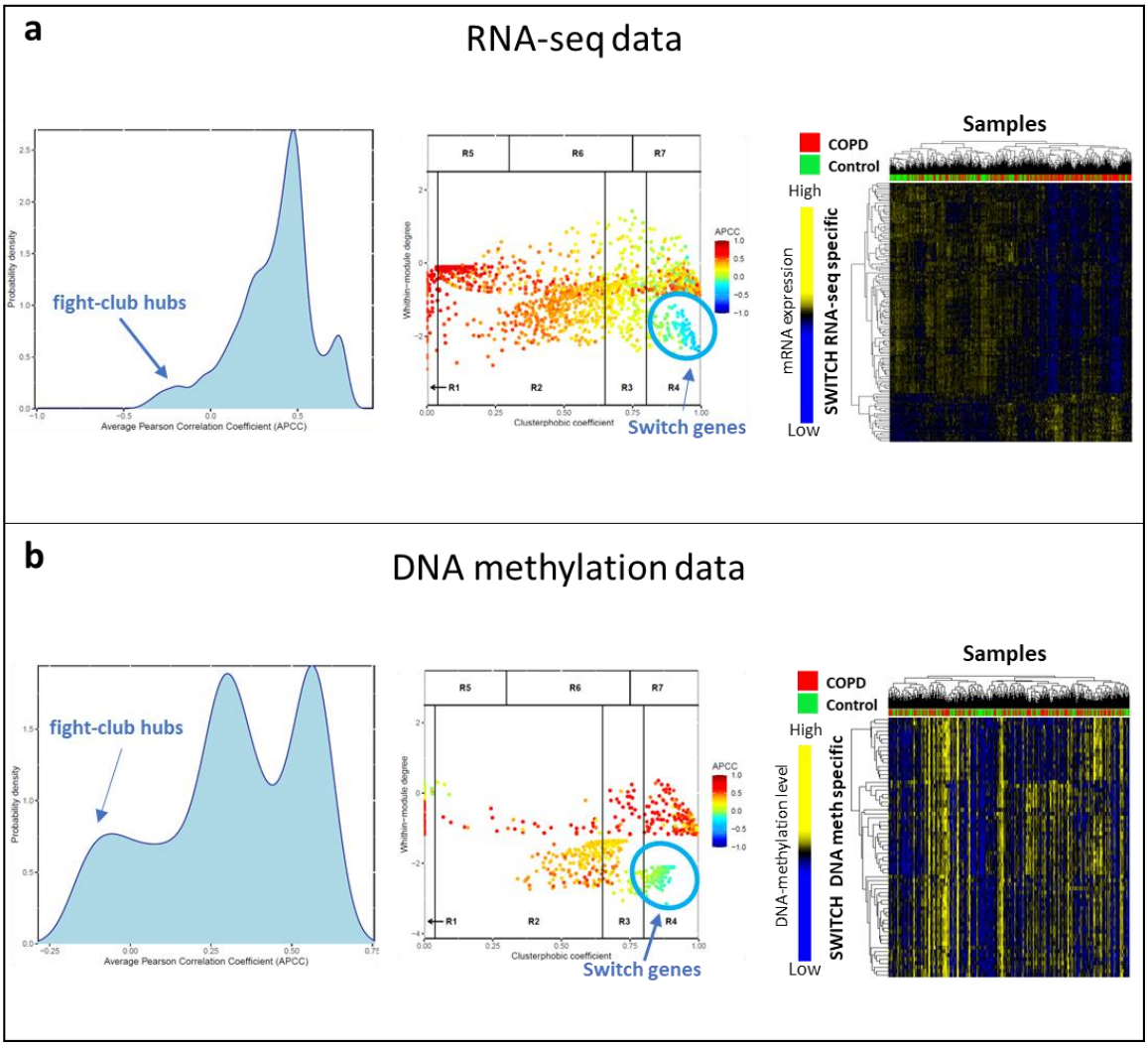

Figure 6.2. SWIM analysis on RNA-seq (a) and DNA methylation data (b). In each panel, from left to right, the following graphs are reported: i) the APCC distribution where the peak fight-club hubs are highlighted; ii) the heat cartography map where the switch genes are indicated; iii) the dendrogram and heat map of the RNA expression (a) or DNA methylation (b) levels of the identified switch genes.

The SWIM clustering analysis performed on the RNA-seq based correlation network end-up with 4 clusters with variable size (1040 nodes in cluster 1, 229 nodes in cluster 2, 298 nodes in cluster 3 and 340 nodes in cluster 4; Fig. 6.3a left). Most of switch genes fell in one cluster (cluster 3), resulting all downregulated in COPD cases; whereas their negative nearest neighbors mostly fell in another cluster (cluster 4), resulting all upregulated in COPD cases

(Fig. 6.3a right). In a recent paper [31], the authors demonstrated that the switch genes of a specific disease constituted a subnetwork of co-localized and functionally-related nodes with a coherent pattern of molecular co-abundance, thus satisfying all the hypotheses of the network medicine in the same way as disease genes themselves do. In view of these findings, we focused on the 84 downregulated switch genes (out of 111) falling in cluster 3 of the RNA-seq based correlation network (Fig. 6.3a right). By performing a functional enrichment analysis, we found that the 84 switch genes were enriched in transcription factor pathways, meanwhile their negative nearest neighbors were enriched in inflammatory signaling pathway such as TNF, IL17 NF-K B, JAK-STAT and MAPK signaling pathway. These results suggested a possible regulatory mechanism where the downregulation of switch genes could be correlated to an overexpression of inflammatory and immune components which is known to be crucial for COPD [204].

The SWIM clustering analysis performed on the DNA methylation-based correlation network defined 3 clusters (111 nodes in cluster 1, 500 nodes in cluster 2 and 304 nodes in cluster 3; Fig. 6.3b left). In this case, almost all of switch genes populated one cluster (cluster 1), resulting all hypermethylated in COPD patients; whereas their negative nearest neighbors fall in another cluster (cluster 3), resulting all hypomethylated in COPD patients (Fig. 6.3b right). As before, we focused on the switch genes falling in a same cluster and having a coherent pattern of co-abundance, i.e., the 64 hypermethylated switch genes (out of 66) falling in cluster 1 of the DNA methylation-based correlation network (Fig. 6.3b right). These 64 switch genes were found to be involved in interleukin-1 and interleukin 11 pathways, meanwhile their nearest negative neighbors were enriched in T-cell receptor, Cytokines and inflammatory response, Vitamin D, and STING pathway.

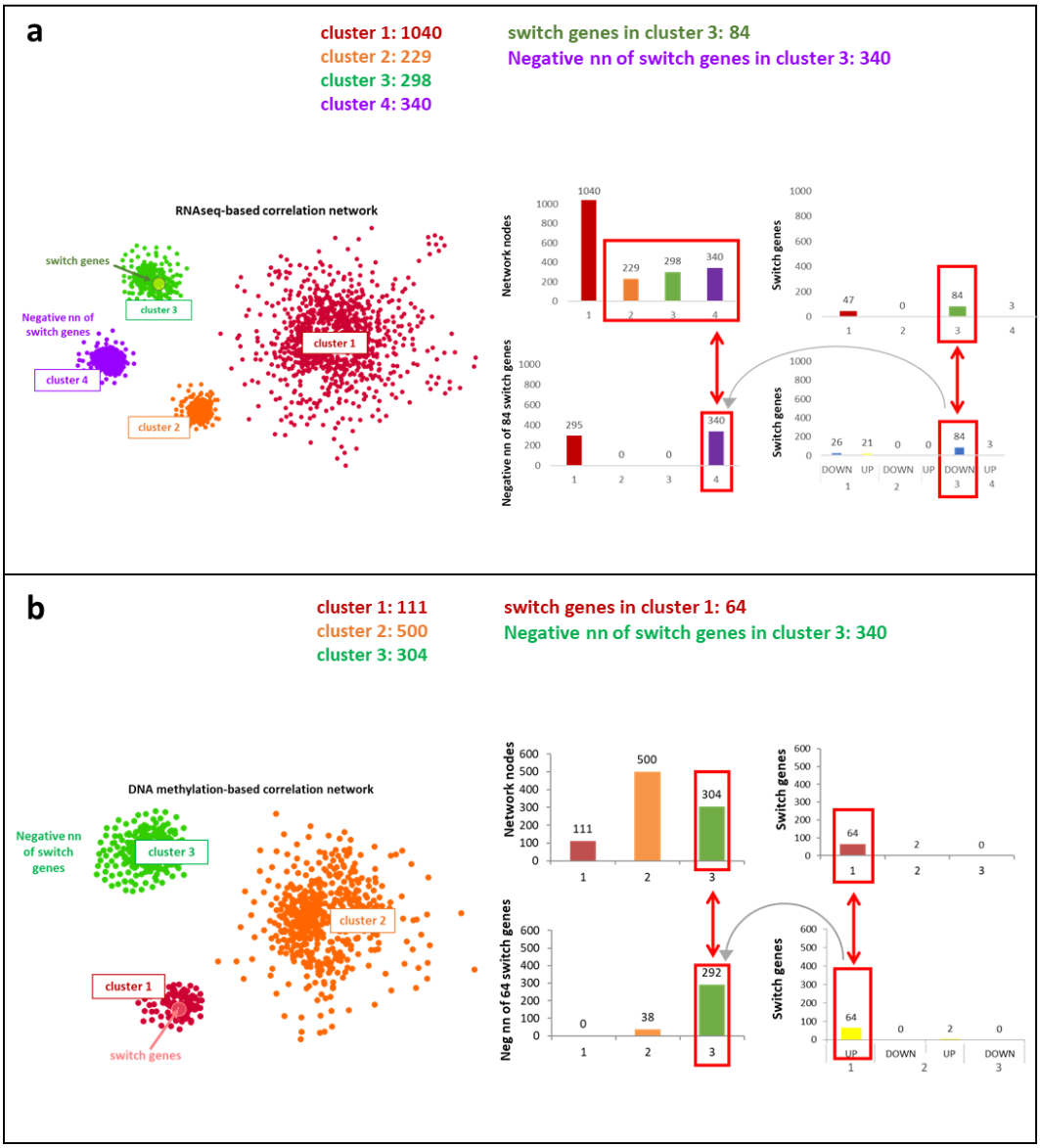

Figure 6.3. SWIM-based correlation network and cluster definition of RNA-seq (a) and DNA methylation (b) data. a) The left panel shows the structure of the RNA-seq-based correlation network, where the clusters that include most of the switch genes (cluster 3), and their negative nearest neighbors (cluster 4), are highlighted. The right panel shows the number of network nodes, switch genes, and negative nearest neighbors of the switch genes in each cluster. Red boxes indicate the switch genes in cluster 3 and their negative nearest neighbors considered in the analysis. b) The left panel shows the structure of the DNA methylation-based correlation network, where the clusters that include most of the switch genes (cluster 1), and their negative nearest neighbors (cluster 3), are highlighted. The right panel shows the number of network nodes, switch genes, and negative nearest neighbors of the switch genes in each cluster. Red boxes indicate the switch genes in cluster 1 and their negative nearest neighbors considered in the analysis.

To integrate the results of the RNA-seq and DNA-methylation based correlation networks, we exploited DGCA that allowed us to build a *consensus network*, in which we preserved the nodes and the edges in common between the two specific GCNs omics. In the *consensus network*, nodes are genes both differentially expressed and differentially methylated in COPD cases compared to controls, and a link occurs between two genes if a correlation exists between their expression and methylation profiles, simultaneously.

The obtained *consensus network* encompassed 63 nodes and 216 edges categorized into four classes (+/+, -/-, +/-, -/+) based on the changes in the gene pair correlation between the two single-omics networks. The first class (+/+) is characterized by a positive Pearson correlation value in both the RNA-seq and DNA methylation-based correlation networks (Fig. 6.4, red edges). The second class (-/-) is characterized by a negative Pearson correlation value in both the RNA-seq and DNA methylation-based correlation networks (Fig.5.4, blue edges). The third (+/-) and fourth (-/+) classes are characterized by a change of sign in the Pearson correlation coefficient between the RNA-seq and DNA methylation-based correlation networks (Fig. 6.4, yellow and green edges).

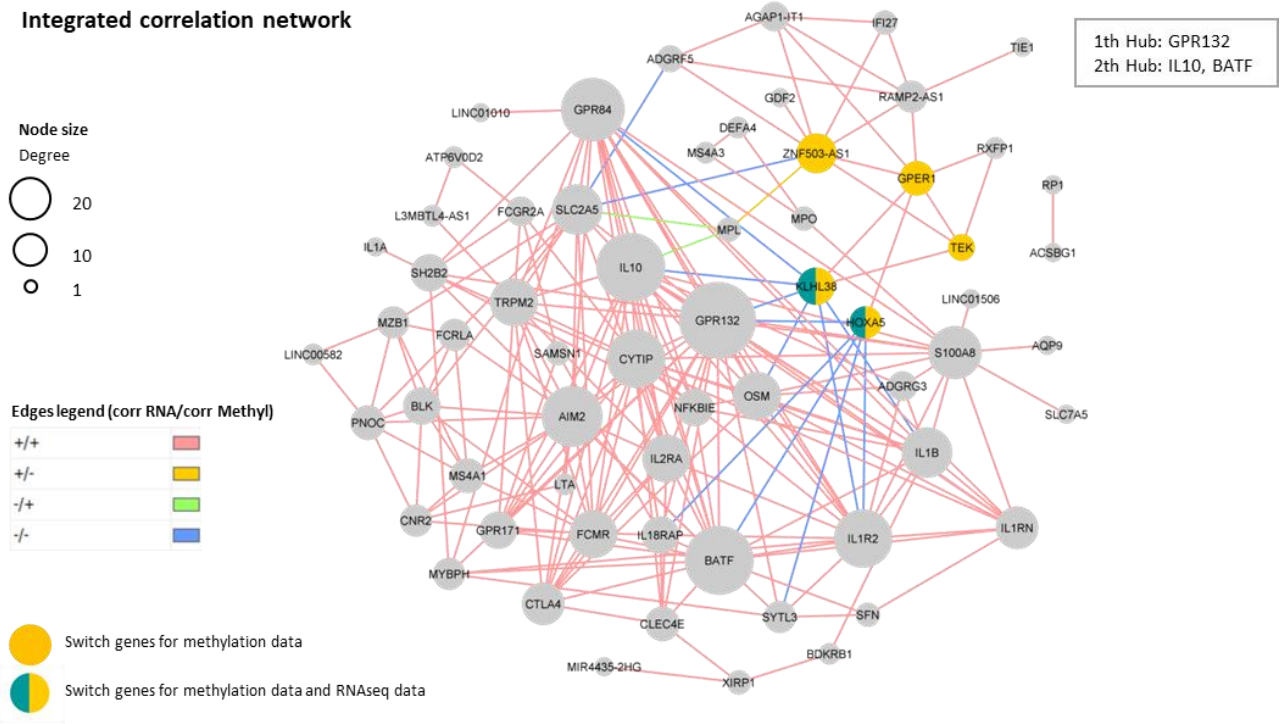

Figure 6.4. ***Consensus network*** **derived from the integration of the two omics data.** Nodes size is proportional to their degree. The edges are divided in 4 classes: The (+/+) class (red edges) is characterized by a positive Pearson correlation value in both the RNA-seq and DNA methylation-based correlation networks. The (-/-) class (blue edges) is characterized by a negative Pearson correlation value in both the RNA-seq and DNA methylation-based correlation networks. The (+/-) class (yellow edges) is characterized by a positive Pearson correlation value in the RNA-seq-based correlation network and a negative Pearson correlation value in the DNA methylation-based correlation networks. The (-/+) class (green edges) is characterized by a negative Pearson correlation value in the RNA-seq-based correlation network and a positive Pearson correlation value in the DNA methylation-based correlation network.

6.4. Discussion

By analyzing the *consensus network*, we found three switch genes (i.e., ZNF503-AS1, GPER1 and TEK) for the DNA methylation data and two switch genes (KLHL38 and HOXA5) for both RNA-seq and DNA methylation-based data, mainly connected through edges belonging to -/- class. We observed a statistically significant downregulation and a hypermethylation of GPER1 and TEK/TIE2 in COPD cases compared to controls. GPER1 was associated to an anti-inflammatory function in diverse studies [205, 206]. Moreover, angiopoietin receptor

TEK/TIE2 had an anti-inflammatory effects by preventing the leakage of pro-inflammatory plasma proteins and leukocytes from blood vessels [207, 208]. Therefore, reactivating the anti-inflammatory action of GPER1 and TEK/TIE2, that appears to be switched off in COPD cases, could be important to counteract the chronic inflammation in COPD context. No biologically insights could be drawn from literature for ZNF503-AS1 gene.

We observed a statistically significant downregulation and hypermethylation of HOXA5 in COPD cases with respect to controls. The authors of [209] discovered that the loss of function of HOXA5 in lung tissues led to an emphysema-like morphology because of impaired alveogenesis in murine model. The hereditability of a hypermethylated promoter of HOXA5, or environmental exposure that drives this epigenetic modification, is likely to be associated with a downregulation of HOXA5, which may result in a non-functional alveoli development.

We observed a statistically significant downregulation and hypermethylation of KLHL38 in COPD cases with respect to controls. The role of this gene seemed to be controversial, indeed the authors of [210] identified KLHL38 as downregulated in human lung tissue due to GPR126 overexpression. However, in another study [211] KLHL38 was observed as overexpressed in non-small cells lung cancer, suggesting an activator role of Akt-signaling pathway leading to cell proliferation, migration, and invasion.

We performed a functional enrichment analysis of the nodes of the *consensus network,* and we compared the results with respect to those ones obtained by executing the same analysis on nodes of RNA-seq and DNA methylation networks, separately (Fig. 6.5). We found 15 pathways that were specific for the *consensus network*, including NOD-like receptor, T cell receptors signaling pathways and Th17 cell differentiation, which were consistent with an inflammatory condition observed in COPD cases (Fig. 6.5a). Moreover, we observed 20 pathways that were specific for the RNA-seq-based-correlation network, including IL-17, TNF, PI3K-Akt, NF-kappa β, and MAPK signaling pathway, which were consistent with the chronic inflammatory state underlying the COPD disease (Fig. 6.5a). No enriched pathways were found for the DNA methylation-based correlation network (Fig. 6.5a). While being poorer

in terms of nodes with respect to the RNA-seq and DNA methylation-based correlation networks, the *consensus network* resulted richer in terms of functional characterization. In fact, it appears to be functionally related to the innate immune response mediated by the NOD-like receptors and adaptive immune response mediated by T-cells, lacking in the single-omics analysis.

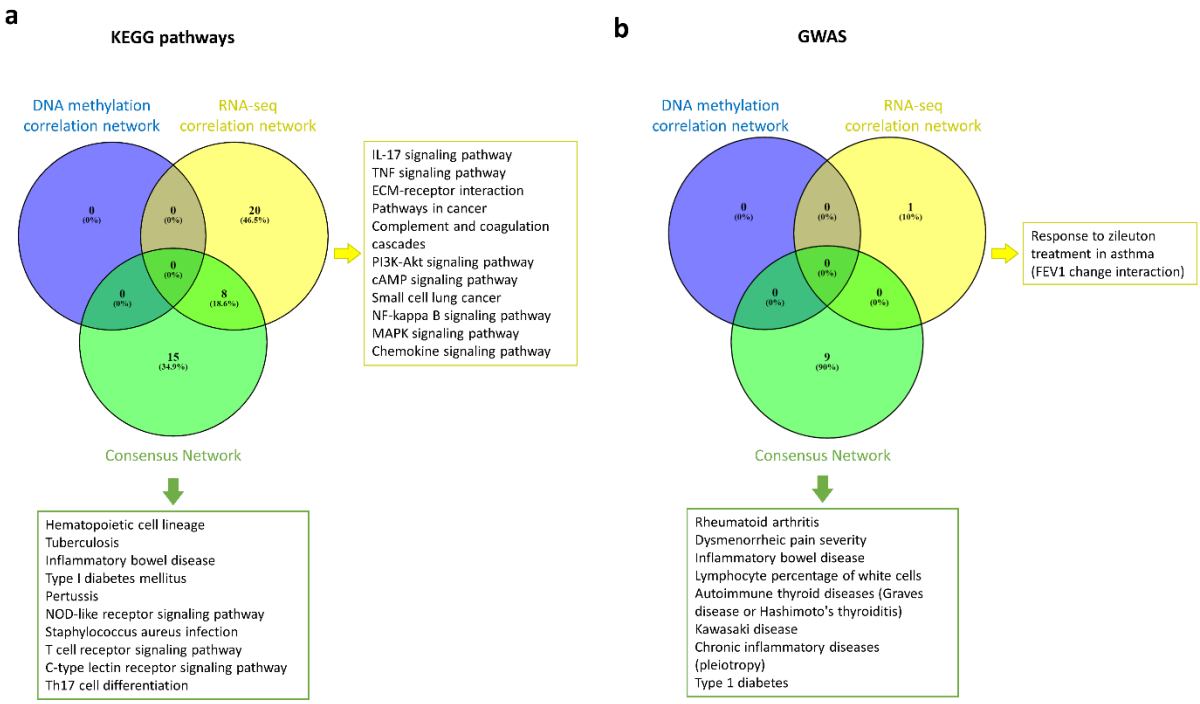

Figure 6.5. Venn Diagram. Enrichment analysis of the genes of the *consensus network*, RNA-seq and DNA-methylation-based correlation network in KEGG pathways (a) and GWAS genes (b).

By performing a GWAS enrichment analysis, we observed that the nodes of the *consensus network* were statistically significant associated to 9 diseases, including Rheumatoid arthritis, inflammatory bowel disease, autoimmune thyroid disease and chronic inflammatory disease (Fig. 6.5b). These diseases can share some common molecular mechanism with COPD, especially in relation with the prevalence of a chronic and autoimmune inflammatory condition

common to all these diseases (Fig. 6.5b). Meanwhile, no disease association was found by performing the same analysis on the single-omics-based correlation networks (Fig. 6.5b).

Altogether, these findings suggest that we have defined a group of differentially expressed and methylated genes that have a considerable biological specificity and could be related to the inflammatory pathological mechanism of COPD.

References

1. Jameson JL, Longo DL (2015) Precision Medicine — Personalized, Problematic, and Promising. N Engl J Med 372:2229–2234. https://doi.org/10.1056/NEJMsb1503104

2. Subramanian I, Verma S, Kumar S, et al (2020) Multi-omics Data Integration, Interpretation, and Its Application. Bioinforma Biol Insights 14:1177932219899051. https://doi.org/10.1177/1177932219899051

3. Barabási A-L, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12:56–68. https://doi.org/10.1038/nrg2918

4. Weinstein JN, Collisson EA, Mills GB, et al (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45:1113–1120. https://doi.org/10.1038/ng.2764

5. Grimaldi AM, Conte F, Pane K, et al (2020) The New Paradigm of Network Medicine to Analyze Breast Cancer Phenotypes. Int J Mol Sci 21:6690. https://doi.org/10.3390/ijms21186690

6. Sibilio P, Bini S, Fiscon G, et al (2021) In silico drug repurposing in COVID-19: A network-based analysis. Biomed Pharmacother 142:111954. https://doi.org/10.1016/j.biopha.2021.111954

7. Barrett T, Wilhite SE, Ledoux P, et al (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–D995

8. Cheng F, Desai RJ, Handy DE, et al (2018) Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun 9:2691. https://doi.org/10.1038/s41467-018-05116-5

9. Meng C, Kuster B, Culhane AC, Gholami AM (2014) A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics 15:162. https://doi.org/10.1186/1471-2105-15-162

10. Argelaguet R, Velten B, Arnol D, et al (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol 14:e8124. https://doi.org/10.15252/msb.20178124

11. Rich JT, Neely JG, Paniello RC, et al (2010) A practical guide to understanding Kaplan-Meier curves. Otolaryngol Neck Surg 143:331–336. https://doi.org/10.1016/j.otohns.2010.05.007

12. Louhimo R, Hautaniemi S (2011) CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics 27:887–888. https://doi.org/10.1093/bioinformatics/btr019

13. Caldera M, Buphamalai P, Müller F, Menche J (2017) Interactome-based approaches to human disease. Curr Opin Syst Biol 3:88–94. https://doi.org/10.1016/j.coisb.2017.04.015

14. Paci P, Fiscon G, Conte F, et al (2021) Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. Npj Syst Biol Appl 7:1–11. https://doi.org/10.1038/s41540-020-00168-0

15. Vidal M, Fields S (2014) The yeast two-hybrid assay: still finding connections after 25 years. Nat Methods 11:1203–1206. https://doi.org/10.1038/nmeth.3182

16. Lin J-S, Lai E-M (2017) Protein–Protein Interactions: Co-Immunoprecipitation. In: Journet L, Cascales E (eds) Bacterial Protein Secretion Systems: Methods and Protocols. Springer, New York, NY, pp 211–219

17. Papanikolaou N, Pavlopoulos GA, Theodosiou T, Iliopoulos I (2015) Protein–protein interaction predictions using text mining methods. Methods 74:47–53. https://doi.org/10.1016/j.ymeth.2014.10.026

18. Lu H-C, Fornili A, Fraternali F (2013) Protein–protein interaction networks studies and importance of 3D structure knowledge. Expert Rev Proteomics 10:511–520. https://doi.org/10.1586/14789450.2013.856764

19. Shen J, Zhang J, Luo X, et al (2007) Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci 104:4337–4341. https://doi.org/10.1073/pnas.0607879104

20. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4:Article17. https://doi.org/10.2202/1544-6115.1128

21. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559. https://doi.org/10.1186/1471-2105-9-559

22. Han J-DJ, Bertin N, Hao T, et al (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature 430:88–93

23. Carter SL, Brechbühler CM, Griffin M, Bond AT (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinforma Oxf Engl 20:2242–2250. https://doi.org/10.1093/bioinformatics/bth234

24. Paci P, Colombo T, Fiscon G, et al (2017) SWIM: a computational tool to unveiling crucial nodes in complex biological networks. Sci Rep 7:44797. https://doi.org/10.1038/srep44797

25. Paci P, Fiscon G (2022) SWIMmeR: an R-based software to unveiling crucial nodes in complex biological networks. Bioinformatics 38:586–588. https://doi.org/10.1093/bioinformatics/btab657

26. Falcone R, Conte F, Fiscon G, et al (2019) BRAFV600E-mutant cancers display a variety of networks by SWIM analysis: prediction of vemurafenib clinical response. Endocrine 64:406–413. https://doi.org/10.1007/s12020-019-01890-4

27. Fiscon G, Conte F, Licursi V, et al (2018) Computational identification of specific genes for glioblastoma stem-like cells identity. Sci Rep 8:7769. https://doi.org/10.1038/s41598-018-26081-5

28. Paci P, Colombo T, Fiscon G, et al (2017) SWIM: a computational tool to unveiling crucial nodes in complex biological networks. Sci Rep 7:44797

29. Fiscon G, Conte F, Paci P (2018) SWIM tool application to expression data of glioblastoma stem-like cell lines, corresponding primary tumors and conventional glioma cell lines. BMC Bioinformatics 19:436. https://doi.org/10.1186/s12859-018-2421-x

30. Paci P, Fiscon G, Conte F, et al (2020) Integrated transcriptomic correlation network analysis identifies COPD molecular determinants. Sci Rep 10:1–18. https://doi.org/10.1038/s41598-020-60228-7

31. Paci P, Fiscon G, Conte F, et al (2021) Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. Npj Syst Biol Appl 7:1–11. https://doi.org/10.1038/s41540-020-00168-0

32. Palumbo MC, Zenoni S, Fasoli M, et al (2014) Integrated Network Analysis Identifies Fight-Club Nodes as a Class of Hubs Encompassing Key Putative Switch Genes That Induce Major Transcriptome Reprogramming during Grapevine Development. Plant Cell Online tpc.114.133710. https://doi.org/10.1105/tpc.114.133710

33. Fiscon G, Conte F, Farina L, Paci P (2021) SAveRUNNER: A network-based algorithm for drug repurposing and its application to COVID-19. PLOS Comput Biol 17:e1008686. https://doi.org/10.1371/journal.pcbi.1008686

34. Fiscon G, Paci P (2021) SAveRUNNER: an R-based tool for drug repurposing. BMC Bioinformatics 22:150. https://doi.org/10.1186/s12859-021-04076-w

35. Fiscon G, Sibilio P, Funari A, et al (2022) Identification of Potential Repurposable Drugs in Alzheimer's Disease Exploiting a Bioinformatics Analysis. J Pers Med 12:1731. https://doi.org/10.3390/jpm12101731

36. Fiscon G, Conte F, Amadio S, et al (2021) Drug Repurposing: A Network-based Approach to Amyotrophic Lateral Sclerosis. Neurotherapeutics. https://doi.org/10.1007/s13311-021-01064-z

37. Amadio S, Conte F, Esposito G, et al (2022) Repurposing Histaminergic Drugs in Multiple Sclerosis. Int J Mol Sci 23:6347. https://doi.org/10.3390/ijms23116347

38. Conte F, Sibilio P, Fiscon G, Paci P (2022) A Transcriptome- and Interactome-Based Analysis Identifies Repurposable Drugs for Human Breast Cancer Subtypes. Symmetry 14:2230. https://doi.org/10.3390/sym14112230

39. Siegel RL, Miller KD, Fuchs HE, Jemal A (2021) Cancer Statistics, 2021. CA Cancer J Clin 71:7–33. https://doi.org/10.3322/caac.21654

40. Nguyen HT, Duong H-Q (2018) The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy (Review). Oncol Lett 16:9–18. https://doi.org/10.3892/ol.2018.8679

41. Ganesh K, Stadler ZK, Cercek A, et al (2019) Immunotherapy in colorectal cancer: rationale, challenges and potential. Nat Rev Gastroenterol Hepatol 16:361–375. https://doi.org/10.1038/s41575-019-0126-x

42. Chalmers ZR, Connelly CF, Fabrizio D, et al (2017) Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Med 9:34. https://doi.org/10.1186/s13073-017-0424-2

43. Goodman AM, Kato S, Bazhenova L, et al (2017) Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. Mol Cancer Ther 16:2598–2608. https://doi.org/10.1158/1535-7163.MCT-17-0386

44. Cheng Y-W, Pincas H, Bacolod MD, et al (2008) CpG Island Methylator Phenotype Associates with Low-Degree Chromosomal Abnormalities in Colorectal Cancer. Clin Cancer Res 14:6005–6013. https://doi.org/10.1158/1078-0432.CCR-08-0216

45. Belardinilli F, Capalbo C, Malapelle U, et al (2020) Clinical Multigene Panel Sequencing Identifies Distinct Mutational Association Patterns in Metastatic Colorectal Cancer. Front Oncol 10:

46. Capalbo C, Belardinilli F, Raimondo D, et al (2019) A Simplified Genomic Profiling Approach Predicts Outcome in Metastatic Colorectal Cancer. Cancers 11:147. https://doi.org/10.3390/cancers11020147

47. De Nicola F, Goeman F, Pallocca M, et al (2018) Deep sequencing and pathway-focused analysis revealed multigene oncodriver signatures predicting survival outcomes in advanced colorectal cancer. Oncogenesis 7:1–10. https://doi.org/10.1038/s41389-018-0066-2

48. Guinney J, Dienstmann R, Wang X, et al (2015) The consensus molecular subtypes of colorectal cancer. Nat Med 21:1350–1356. https://doi.org/10.1038/nm.3967

49. Muzny DM, Bainbridge MN, Chang K, et al (2012) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487:330–337. https://doi.org/10.1038/nature11252

50. André T, Shiu K-K, Kim TW, et al (2020) Pembrolizumab in Microsatellite-Instability–High Advanced Colorectal Cancer. N Engl J Med 383:2207–2218. https://doi.org/10.1056/NEJMoa2017699

51. Le DT, Durham JN, Smith KN, et al (2017) Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science 357:409–413. https://doi.org/10.1126/science.aan6733

52. Mermel CH, Schumacher SE, Hill B, et al (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 12:R41. https://doi.org/10.1186/gb-2011-12-4-r41

53. Nilsen G, Liestøl K, Van Loo P, et al (2012) Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. BMC Genomics 13:591. https://doi.org/10.1186/1471-2164-13-591

54. Mayakonda A, Lin D-C, Assenov Y, et al (2018) Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res 28:1747–1756. https://doi.org/10.1101/gr.239244.118

55. COSMIC | SBS - Mutational Signatures. https://cancer.sanger.ac.uk/signatures/sbs/. Accessed 27 Oct 2022

56. Ritchie ME, Phipson B, Wu D, et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47. https://doi.org/10.1093/nar/gkv007

57. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559. https://doi.org/10.1186/1471-2105-9-559

58. Zhang B, Horvath S (2005) A General Framework for Weighted Gene Co-Expression Network Analysis. Stat Appl Genet Mol Biol 4:. https://doi.org/10.2202/1544-6115.1128

59. Nirmal AJ, Regan T, Shih BB, et al (2018) Immune Cell Gene Signatures for Profiling the Microenvironment of Solid Tumors. Cancer Immunol Res 6:1388–1400. https://doi.org/10.1158/2326-6066.CIR-18-0342

60. Hu F-F, Liu C-J, Liu L-L, et al (2021) Expression profile of immune checkpoint genes and their roles in predicting immunotherapy response. Brief Bioinform 22:bbaa176. https://doi.org/10.1093/bib/bbaa176

61. Yaeger R, Chatila WK, Lipsyc MD, et al (2018) Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer. Cancer Cell 33:125-136.e3. https://doi.org/10.1016/j.ccell.2017.12.004

62. Pino MS, Chung DC (2010) The Chromosomal Instability Pathway in Colon Cancer. Gastroenterology 138:2059–2072. https://doi.org/10.1053/j.gastro.2009.12.065

63. Boland CR, Goel A (2010) Microsatellite Instability in Colorectal Cancer. Gastroenterology 138:2073-2087.e3. https://doi.org/10.1053/j.gastro.2009.12.064

64. Helleday T, Eshtad S, Nik-Zainal S (2014) Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet 15:585–598. https://doi.org/10.1038/nrg3729

65. Alexandrov LB, Kim J, Haradhvala NJ, et al (2020) The repertoire of mutational signatures in human cancer. Nature 578:94–101. https://doi.org/10.1038/s41586-020-1943-3

66. Drost J, van Boxtel R, Blokzijl F, et al (2017) Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science 358:234–238. https://doi.org/10.1126/science.aao3130

67. Strickler JH, Hanks BA, Khasraw M (2021) Tumor Mutational Burden as a Predictor of Immunotherapy Response: Is More Always Better? Clin Cancer Res 27:1236–1241. https://doi.org/10.1158/1078-0432.CCR-20-3054

68. Markowitz SD, Bertagnolli MM (2009) Molecular Basis of Colorectal Cancer. N Engl J Med 361:2449–2460. https://doi.org/10.1056/NEJMra0804588

69. Ogino S, Kawasaki T, Kirkner GJ, et al (2006) CpG Island Methylator Phenotype-Low (CIMP-Low) in Colorectal Cancer: Possible Associations with Male Sex and KRAS Mutations. J Mol Diagn 8:582–588. https://doi.org/10.2353/jmoldx.2006.060082

70. Jia M, Gao X, Zhang Y, et al (2016) Different definitions of CpG island methylator phenotype and outcomes of colorectal cancer: a systematic review. Clin Epigenetics 8:25. https://doi.org/10.1186/s13148-016-0191-8

71. Hinoue T, Weisenberger DJ, Lange CPE, et al (2012) Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res 22:271–282. https://doi.org/10.1101/gr.117523.110

72. Tao Y, Kang B, Petkovich DA, et al (2019) Aging-like Spontaneous Epigenetic Silencing Facilitates Wnt Activation, Stemness, and BrafV600E-Induced Tumorigenesis. Cancer Cell 35:315-328.e6. https://doi.org/10.1016/j.ccell.2019.01.005

73. Di Micco R, Fumagalli M, Cicalese A, et al (2006) Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. Nature 444:638–642. https://doi.org/10.1038/nature05327

74. Serrano M, Lin AW, McCurrach ME, et al (1997) Oncogenic ras Provokes Premature Cell Senescence Associated with Accumulation of p53 and p16INK4a. Cell 88:593–602. https://doi.org/10.1016/S0092-8674(00)81902-9

75. Javier BM, Yaeger R, Wang L, et al (2016) Recurrent, truncating SOX9 mutations are associated with SOX9 overexpression, KRAS mutation, and TP53 wild type status in colorectal carcinoma. Oncotarget 7:50875–50882. https://doi.org/10.18632/oncotarget.9682

76. Belhadj S, Quintana I, Mur P, et al (2019) NTHL1 biallelic mutations seldom cause colorectal cancer, serrated polyposis or a multi-tumor phenotype, in absence of colorectal adenomas. Sci Rep 9:9020. https://doi.org/10.1038/s41598-019-45281-1

77. Chalabi M, Fanchi LF, Dijkstra KK, et al (2020) Neoadjuvant immunotherapy leads to pathological responses in MMR-proficient and MMR-deficient early-stage colon cancers. Nat Med 26:566–576. https://doi.org/10.1038/s41591-020-0805-8

78. Williams DS, Mouradov D, Jorissen RN, et al (2019) Lymphocytic response to tumour and deficient DNA mismatch repair identify subtypes of stage II/III colorectal cancer associated with patient outcomes. Gut 68:465–474. https://doi.org/10.1136/gutjnl-2017-315664

79. Zaborowski AM, Winter DC, Lynch L (2021) The therapeutic and prognostic implications of immunobiology in colorectal cancer: a review. Br J Cancer 125:1341–1349. https://doi.org/10.1038/s41416-021-01475-x

80. Pagès F, Mlecnik B, Marliot F, et al (2018) International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. The Lancet 391:2128–2139. https://doi.org/10.1016/S0140-6736(18)30789-X

81. Kikuchi T, Mimura K, Okayama H, et al (2019) A subset of patients with MSS/MSI-low-colorectal cancer showed increased CD8(+) TILs together with up-regulated IFN-γ. Oncol Lett 18:5977–5985. https://doi.org/10.3892/ol.2019.10953

82. Egen JG, Kuhns MS, Allison JP (2002) CTLA-4: new insights into its biological function and use in tumor immunotherapy. Nat Immunol 3:611–618. https://doi.org/10.1038/ni0702-611

83. Wang J-Y, Wang W-P (2020) B7-H4, a promising target for immunotherapy. Cell Immunol 347:104008. https://doi.org/10.1016/j.cellimm.2019.104008

84. Chen Y-L, Lin H-W, Chien C-L, et al (2019) BTLA blockade enhances Cancer therapy by inhibiting IL-6/IL-10-induced CD19high B lymphocytes. J Immunother Cancer 7:313. https://doi.org/10.1186/s40425-019-0744-4

85. Ferlay J, Colombet M, Soerjomataram I, et al (2019) Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 144:1941–1953. https://doi.org/10.1002/ijc.31937

86. D Z, A I, C S, M P (2015) Clinical management of breast cancer heterogeneity. Nat Rev Clin Oncol 12:. https://doi.org/10.1038/nrclinonc.2015.73

87. Goldhirsch A, Winer EP, Coates AS, et al (2013) Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol 24:2206–2223. https://doi.org/10.1093/annonc/mdt303

88. Koboldt DC, Fulton RS, McLellan MD, et al (2012) Comprehensive molecular portraits of human breast tumours. Nature 490:61–70. https://doi.org/10.1038/nature11412

89. Parker JS, Mullins M, Cheang MCU, et al (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol Off J Am Soc Clin Oncol 27:1160–1167. https://doi.org/10.1200/JCO.2008.18.1370

90. Gazinska P, Grigoriadis A, Brown JP, et al (2013) Comparison of basal-like triple-negative breast cancer defined by morphology, immunohistochemistry and transcriptional profiles. Mod Pathol 26:955–966. https://doi.org/10.1038/modpathol.2012.244

91. Lin NU, Vanderplas A, Hughes ME, et al (2012) Clinicopathologic features, patterns of recurrence, and survival among women with triple-negative breast cancer in the National Comprehensive Cancer Network. Cancer 118:5463–5472. https://doi.org/10.1002/cncr.27581

92. Nagy Á, Lánczky A, Menyhárt O, Győrffy B (2018) Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. Sci Rep 8:9227. https://doi.org/10.1038/s41598-018-27521-y

93. Győrffy B (2021) Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. Comput Struct Biotechnol J 19:4101–4109. https://doi.org/10.1016/j.csbj.2021.07.014

94. Zambelli F, Pesole G, Pavesi G (2009) Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic Acids Res 37:W247–W252. https://doi.org/10.1093/nar/gkp464

95. Han H, Cho J-W, Lee S, et al (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res 46:D380–D386. https://doi.org/10.1093/nar/gkx1013

96. Khan A, Fornes O, Stigliani A, et al (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res 46:D260–D266. https://doi.org/10.1093/nar/gkx1126

97. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30:1363–1369. https://doi.org/10.1093/bioinformatics/btu049

98. Györffy B, Lanczky A, Eklund AC, et al (2010) An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Res Treat 123:725–731. https://doi.org/10.1007/s10549-009-0674-9

99. Klutstein M, Nejman D, Greenfield R, Cedar H (2016) DNA Methylation in Cancer and Aging. Cancer Res 76:3446–3450. https://doi.org/10.1158/0008-5472.CAN-15-3278

100. G W, Me B, Rc J, et al (1983) Purine and pyrimidine enzymic programs and nucleotide pattern in sarcoma. Cancer Res 43:

101. Kizaki H, Williams JC, Morris HP, Weber G (1980) Increased cytidine 5'-triphosphate synthetase activity in rat and human tumors. Cancer Res 40:3921–3927

102. Williams JC, Kizaki H, Weber G, Morris HP (1978) Increased CTP synthetase activity in cancer cells. Nature 271:71–73. https://doi.org/10.1038/271071a0

103. Huang M, Whang P, Lewicki P, Mitchell BS (2011) Cyclopentenyl Cytosine Induces Senescence in Breast Cancer Cells through the Nucleolar Stress Response and Activation of p53. Mol Pharmacol 80:40–48. https://doi.org/10.1124/mol.110.070284

104. Maire V, Mahmood F, Rigaill G, et al (2019) LRP8 is overexpressed in estrogen-negative breast cancers and a potential target for these tumors. Cancer Med 8:325–336. https://doi.org/10.1002/cam4.1923

105. Lin C-C, Lo M-C, Moody R, et al (2018) Targeting LRP8 inhibits breast cancer stem cells in triple-negative breast cancer. Cancer Lett 438:165–173. https://doi.org/10.1016/j.canlet.2018.09.022

106. Wang J, Chen W, Wei W, Lou J (2017) Oncogene TUBA1C promotes migration and proliferation in hepatocellular carcinoma and predicts a poor prognosis. Oncotarget 8:96215–96224. https://doi.org/10.18632/oncotarget.21894

107. Albahde MAH, Zhang P, Zhang Q, et al (2020) Upregulated Expression of TUBA1C Predicts Poor Prognosis and Promotes Oncogenesis in Pancreatic Ductal Adenocarcinoma via Regulating the Cell Cycle. Front Oncol 10:

108. Nami B, Wang Z (2018) Genetics and Expression Profile of the Tubulin Gene Superfamily in Breast Cancer Subtypes and Its Relation to Taxane Resistance. Cancers 10:274. https://doi.org/10.3390/cancers10080274

109. Wang CCN, Li CY, Cai J-H, et al (2019) Identification of Prognostic Candidate Genes in Breast Cancer by Integrated Bioinformatic Analysis. J Clin Med 8:1160. https://doi.org/10.3390/jcm8081160

110. Chen D, Li Y, Wang L, Jiao K (2015) SEMA6D Expression and Patient Survival in Breast Invasive Carcinoma. Int J Breast Cancer 2015:e539721. https://doi.org/10.1155/2015/539721

111. Ramos J, Yoo C, Felty Q, et al (2020) Sensitivity to differential NRF1 gene signatures contributes to breast cancer disparities. J Cancer Res Clin Oncol 146:2777–2815. https://doi.org/10.1007/s00432-020-03320-9

112. Jin G, Wang W, Cheng P, et al (2020) DNA replication and sister chromatid cohesion 1 promotes breast carcinoma progression by modulating the Wnt/β-catenin signaling and p53 protein. J Biosci 45:127

113. Kim J-T, Cho HJ, Park SY, et al (2019) DNA Replication and Sister Chromatid Cohesion 1 (DSCC1) of the Replication Factor Complex CTF18-RFC is Critical for Colon Cancer Cell Growth. J Cancer 10:6142–6153. https://doi.org/10.7150/jca.32339

114. Sun Z, Wu Z, Zhang F, et al (2016) PRAME is critical for breast cancer growth and metastasis. Gene 594:160–164. https://doi.org/10.1016/j.gene.2016.09.016

115. Al-Khadairi G, Naik A, Thomas R, et al (2019) PRAME promotes epithelial-to-mesenchymal transition in triple negative breast cancer. J Transl Med 17:9. https://doi.org/10.1186/s12967-018-1757-3

116. Epping MT, Hart A a. M, Glas AM, et al (2008) PRAME expression and clinical outcome of breast cancer. Br J Cancer 99:398–403. https://doi.org/10.1038/sj.bjc.6604494

117. Koppula P, Zhang Y, Zhuang L, Gan B (2018) Amino acid transporter SLC7A11/xCT at the crossroads of regulating redox homeostasis and nutrient dependency of cancer. Cancer Commun 38:12. https://doi.org/10.1186/s40880-018-0288-x

118. Yang Y, Yee D (2014) IGF-I Regulates Redox Status in Breast Cancer Cells by Activating the Amino Acid Transport Molecule xC−. Cancer Res 74:2295–2305. https://doi.org/10.1158/0008-5472.CAN-13-1803

119. Koppula P, Zhuang L, Gan B (2021) Cystine transporter SLC7A11/xCT in cancer: ferroptosis, nutrient dependency, and cancer therapy. Protein Cell 12:599–620. https://doi.org/10.1007/s13238-020-00789-5

120. Miguchi M, Hinoi T, Shimomura M, et al (2016) Gasdermin C Is Upregulated by Inactivation of Transforming Growth Factor β Receptor Type II in the Presence of Mutated Apc, Promoting Colorectal Cancer Proliferation. PLOS ONE 11:e0166422. https://doi.org/10.1371/journal.pone.0166422

121. Hou J, Zhao R, Xia W, et al (2020) PD-L1-mediated gasdermin C expression switches apoptosis to pyroptosis in cancer cells and facilitates tumour necrosis. Nat Cell Biol 22:1264–1275. https://doi.org/10.1038/s41556-020-0575-z

122. Wei J, Xu Z, Chen X, et al (2020) Overexpression of GSDMC is a prognostic factor for predicting a poor outcome in lung adenocarcinoma. Mol Med Rep 21:360–370. https://doi.org/10.3892/mmr.2019.10837

123. Osthus RC, Karim B, Prescott JE, et al (2005) The Myc Target Gene JPO1/CDCA7 Is Frequently Overexpressed in Human Tumors and Has Limited Transforming Activity In vivo. Cancer Res 65:5620–5627. https://doi.org/10.1158/0008-5472.CAN-05-0536

124. Ye L, Li F, Song Y, et al (2018) Overexpression of CDCA7 predicts poor prognosis and induces EZH2-mediated progression of triple-negative breast cancer. Int J Cancer 143:2602–2613. https://doi.org/10.1002/ijc.31766

125. Andres SA, Bickett KE, Alatoum MA, et al (2015) Interaction between smoking history and gene expression levels impacts survival of breast cancer patients. Breast Cancer Res Treat 152:545–556. https://doi.org/10.1007/s10549-015-3507-z

126. Chang YP, Wang G, Bermudez V, et al (2007) Crystal structure of the GINS complex and functional insights into its role in DNA replication. Proc Natl Acad Sci 104:12685–12690. https://doi.org/10.1073/pnas.0705558104

127. Rong Z, Luo Z, Zhang J, et al (2020) GINS complex subunit 4, a prognostic biomarker and reversely mediated by Krüppel-like factor 4, promotes the growth of colorectal cancer. Cancer Sci 111:1203–1217. https://doi.org/10.1111/cas.14341

128. Yamane K, Naito H, Wakabayashi T, et al (2016) Regulation of SLD5 gene expression by miR-370 during acute growth of cancer cells. Sci Rep 6:30941. https://doi.org/10.1038/srep30941

129. Camarda R, Zhou AY, Kohnz RA, et al (2016) Inhibition of fatty acid oxidation as a therapy for MYC-overexpressing triple-negative breast cancer. Nat Med 22:427–432. https://doi.org/10.1038/nm.4055

130. Shahbandi A, Nguyen HD, Jackson JG (2020) TP53 Mutations and Outcomes in Breast Cancer: Reading beyond the Headlines. Trends Cancer 6:98–110. https://doi.org/10.1016/j.trecan.2020.01.007

131. Kim J-Y, Jung HH, Ahn S, et al (2016) The relationship between nuclear factor (NF)-κB family gene expression and prognosis in triple-negative breast cancer (TNBC) patients receiving adjuvant doxorubicin treatment. Sci Rep 6:31804. https://doi.org/10.1038/srep31804

132. Jourdan J-P, Bureau R, Rochais C, Dallemagne P (2020) Drug repositioning: a brief overview. J Pharm Pharmacol 72:1145–1151. https://doi.org/10.1111/jphp.13273

133. Conte F, Fiscon G, Licursi V, et al (2020) A paradigm shift in medicine: A comprehensive review of network-based approaches. Biochim Biophys Acta BBA - Gene Regul Mech 1863:194416. https://doi.org/10.1016/j.bbagrm.2019.194416

134. Sonawane AR, Weiss ST, Glass K, Sharma A (2019) Network Medicine in the Age of Biomedical Big Data. Front Genet 10:

135. Silverman EK, Schmidt HHHW, Anastasiadou E, et al (2020) Molecular networks in Network Medicine: Development and applications. WIREs Syst Biol Med 12:e1489. https://doi.org/10.1002/wsbm.1489

136. Fiscon G, Conte F, Farina L, Paci P (2018) Network-Based Approaches to Explore Complex Biological Systems towards Network Medicine. Genes 9:. https://doi.org/10.3390/genes9090437

137. Menche J, Sharma A, Kitsak M, et al (2015) Uncovering disease-disease relationships through the incomplete interactome. Science 347:1257601. https://doi.org/10.1126/science.1257601

138. Fiscon G, Conte F, Licursi V, et al (2018) Computational identification of specific genes for glioblastoma stem-like cells identity. Sci Rep 8:7769. https://doi.org/10.1038/s41598-018-26081-5

139. Fiscon G, Conte F, Paci P (2018) SWIM tool application to expression data of glioblastoma stem-like cell lines, corresponding primary tumors and conventional glioma cell lines. BMC Bioinformatics 19:436. https://doi.org/10.1186/s12859-018-2421-x

140. Falcone R, Conte F, Fiscon G, et al (2019) BRAFV600E-mutant cancers display a variety of networks by SWIM analysis: prediction of vemurafenib clinical response. Endocrine 64:406–413. https://doi.org/10.1007/s12020-019-01890-4

141. Fiscon G, Pegoraro S, Conte F, et al (2021) Gene network analysis using SWIM reveals interplay between the transcription factor-encoding genes HMGA1, FOXM1, and MYBL2 in triple-negative breast cancer. FEBS Lett 595:1569–1586. https://doi.org/10.1002/1873-3468.14085

142. Paci P, Fiscon G, Conte F, et al (2020) Integrated transcriptomic correlation network analysis identifies COPD molecular determinants. Sci Rep 10:3361. https://doi.org/10.1038/s41598-020-60228-7

143. Pecce V, Verrienti A, Fiscon G, et al (2021) The role of FOSL1 in stem-like cell reprogramming processes. Sci Rep 11:14677. https://doi.org/10.1038/s41598-021-94072-0

144. Panebianco V, Pecoraro M, Fiscon G, et al (2020) Prostate cancer screening research can benefit from network medicine: an emerging awareness. Npj Syst Biol Appl 6:1–6. https://doi.org/10.1038/s41540-020-0133-0

145. Fiscon G, Conte F, Farina L, et al (2019) Identification of Disease–miRNA Networks Across Different Cancer Types Using SWIM. In: Laganà A (ed) MicroRNA Target Identification: Methods and Protocols. Springer, New York, NY, pp 169–181

146. Sahoo BM, Ravi Kumar BVV, Sruti J, et al (2021) Drug Repurposing Strategy (DRS): Emerging Approach to Identify Potential Therapeutics for Treatment of Novel Coronavirus Infection. Front Mol Biosci 8:628144. https://doi.org/10.3389/fmolb.2021.628144

147. Nitulescu GM, Paunescu H, Moschos SA, et al (2020) Comprehensive analysis of drugs to treat SARS-CoV-2 infection: Mechanistic insights into current COVID-19 therapies (Review). Int J Mol Med 46:467–488. https://doi.org/10.3892/ijmm.2020.4608

148. Kim JS, Lee JY, Yang JW, et al (2021) Immunopathogenesis and treatment of cytokine storm in COVID-19. Theranostics 11:316–329. https://doi.org/10.7150/thno.49713

149. López-Collazo E, Avendaño-Ortiz J, Martín-Quirós A, Aguirre LA (2020) Immune Response and COVID-19: A mirror image of Sepsis. Int J Biol Sci 16:2479–2489. https://doi.org/10.7150/ijbs.48400

150. Pum A, Ennemoser M, Adage T, Kungl AJ (2021) Cytokines and Chemokines in SARS-CoV-2 Infections—Therapeutic Strategies Targeting Cytokine Storm. Biomolecules 11:91. https://doi.org/10.3390/biom11010091

151. Rizk JG, Kalantar-Zadeh K, Mehra MR, et al (2020) Pharmaco-Immunomodulatory Therapy in COVID-19. Drugs 80:1267–1292. https://doi.org/10.1007/s40265-020-01367-z

152. Kulanthaivel S, Kaliberdenko VB, Balasundaram K, et al (2021) Tocilizumab in SARS-CoV-2 Patients with the Syndrome of Cytokine Storm: A Narrative Review. Rev Recent Clin Trials 16:138–145. https://doi.org/10.2174/1574887115666200917110954

153. Scavone C, Mascolo A, Rafaniello C, et al (2021) Therapeutic strategies to fight COVID-19: Which is the status artis? Br J Pharmacol n/a: https://doi.org/10.1111/bph.15452

154. Wishart DS, Feunang YD, Guo AC, et al (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46:D1074. https://doi.org/10.1093/nar/gkx1037

155. Kinsella RJ, Kähäri A, Haider S, et al (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database J Biol Databases Curation 2011:bar030. https://doi.org/10.1093/database/bar030

156. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8

157. Iglewicz B, Hoaglin D (1993) How to detect and handle outliers. ASQC Basic Ref Qual Control Stat Tech 16:

158. Iba T, Warkentin TE, Thachil J, et al (2021) Proposal of the Definition for COVID-19-Associated Coagulopathy. J Clin Med 10:191. https://doi.org/10.3390/jcm10020191

159. Price LC, Garfield B, Bleakley C, et al (2020) Rescue therapy with thrombolysis in patients with severe COVID-19-associated acute respiratory distress syndrome. Pulm Circ 10:2045894020973906. https://doi.org/10.1177/2045894020973906

160. Liu J, Li J, Arnold K, et al (2020) Using heparin molecules to manage COVID-2019. Res Pract Thromb Haemost. https://doi.org/10.1002/rth2.12353

161. Caballero López A, Herrera Cartaya C, Chávez González E, et al (2020) Pulmonary Thrombosis in COVID-19 Treated by Thrombolysis: A Small Case Series Using Streptokinase. Semin Thromb Hemost. https://doi.org/10.1055/s-0040-1716872

162. Ogawa H, Asakura H (2020) Consideration of Tranexamic Acid Administration to COVID-19 Patients. Physiol Rev 100:1595–1596. https://doi.org/10.1152/physrev.00023.2020

163. Demopoulos C, Antonopoulou S, Theoharides TC (2020) COVID-19, microthromboses, inflammation, and platelet activating factor. BioFactors Oxf Engl 46:927–933. https://doi.org/10.1002/biof.1696

164. Eldanasory OA, Eljaaly K, Memish ZA, Al-Tawfiq JA (2020) Histamine release theory and roles of antihistamine in the treatment of cytokines storm of COVID-19. Travel Med Infect Dis 37:101874. https://doi.org/10.1016/j.tmaid.2020.101874

165. Sestili P, Stocchi V (2020) Repositioning Chromones for Early Anti-inflammatory Treatment of COVID-19. Front Pharmacol 11:. https://doi.org/10.3389/fphar.2020.00854

166. Reznikov LR, Norris MH, Vashisht R, et al (2020) Identification of antiviral antihistamines for COVID-19 repurposing. Biochem Biophys Res Commun. https://doi.org/10.1016/j.bbrc.2020.11.095

167. Yadav V, Talwar P (2019) Repositioning of fluoroquinolones from antibiotic to anti-cancer agents: An underestimated truth. Biomed Pharmacother 111:934–946. https://doi.org/10.1016/j.biopha.2018.12.119

168. Abdel-Aal MAA, Abdel-Aziz SA, Shaykoon MSA, Abuo-Rahma GE-DA (2019) Towards anticancer fluoroquinolones: A review article. Arch Pharm (Weinheim) 352:1800376. https://doi.org/10.1002/ardp.201800376

169. Rasaeifar B, Gomez-Gutierrez P, Perez JJ (2020) Molecular Features of Non-Selective Small Molecule Antagonists of the Bradykinin Receptors. Pharmaceuticals 13:259. https://doi.org/10.3390/ph13090259

170. Assar S, Nosratabadi R, Azad HK, et al (2020) A Review of Immunomodulatory Effects of Fluoroquinolones. Immunol Invest 0:1–20. https://doi.org/10.1080/08820139.2020.1797778

171. J A, Francis D, C S S, et al (2020) Repurposing simeprevir, calpain inhibitor IV and a cathepsin F inhibitor against SARS-CoV-2 and insights into their interactions with Mpro. J Biomol Struct Dyn 1–12. https://doi.org/10.1080/07391102.2020.1813200

172. Scroggs SLP, Offerdahl DK, Flather DP, et al (2020) Fluoroquinolone Antibiotics Exhibit Low Antiviral Activity against SARS-CoV-2 and MERS-CoV. Viruses 13:. https://doi.org/10.3390/v13010008

173. Sharma D, Kunamneni A (2020) Recent progress in the repurposing of drugs/molecules for the management of COVID-19. Expert Rev Anti Infect Ther 0:1–9. https://doi.org/10.1080/14787210.2021.1860020

174. Rose L, Graham L, Koenecke A, et al (2020) The Association Between Alpha-1 Adrenergic Receptor Antagonists and In-Hospital Mortality from COVID-19. MedRxiv Prepr Serv Health Sci. https://doi.org/10.1101/2020.12.18.20248346

175. Hyoju SK, Zaborina O, van Goor H (2020) SARS-CoV-2 and the sympathetic immune response: Dampening inflammation with antihypertensive drugs (Clonidine and Propranolol). Med Hypotheses 144:110039. https://doi.org/10.1016/j.mehy.2020.110039

176. Luo P, Liu D, Li J (2021) Epinephrine use in COVID-19: friend or foe? Eur J Hosp Pharm 28:e1–e1. https://doi.org/10.1136/ejhpharm-2020-002295

177. Hoertel N, Sánchez-Rico M, Vernet R, et al (2021) Association between antidepressant use and reduced risk of intubation or death in hospitalized patients with COVID-19: results from an observational study. Mol Psychiatry 1–14. https://doi.org/10.1038/s41380-021-01021-4

178. Grisanti LA, Perez DM, Porter JE (2011) Modulation of immune cell function by α(1)-adrenergic receptor activation. Curr Top Membr 67:113–138. https://doi.org/10.1016/B978-0-12-384921-2.00006-9

179. Otręba M, Kośmider L, Rzepecka-Stojko A (2020) Antiviral activity of chlorpromazine, fluphenazine, perphenazine, prochlorperazine, and thioridazine towards RNA-viruses. A review. Eur J Pharmacol 887:173553. https://doi.org/10.1016/j.ejphar.2020.173553

180. Di Rosso ME, Palumbo ML, Genaro AM (2016) Immunomodulatory effects of fluoxetine: A new potential pharmacological action for a classic antidepressant drug? Pharmacol Res 109:101–107. https://doi.org/10.1016/j.phrs.2015.11.021

181. Pinoli M, Marino F, Cosentino M (2017) Dopaminergic Regulation of Innate Immunity: a Review. J Neuroimmune Pharmacol Off J Soc NeuroImmune Pharmacol 12:602–623. https://doi.org/10.1007/s11481-017-9749-2

182. Thomas Broome S, Louangaphay K, Keay KA, et al (2020) Dopamine: an immune transmitter. Neural Regen Res 15:2173–2185. https://doi.org/10.4103/1673-5374.284976

183. Jones A, Kainz D, Khan F, et al (2014) Human Macrophage SCN5A Activates an Innate Immune Signaling Pathway for Antiviral Host Defense *. J Biol Chem 289:35326–35340. https://doi.org/10.1074/jbc.M114.611962

184. Carrithers MD, Dib-Hajj S, Carrithers LM, et al (2007) Expression of the voltage-gated sodium channel NaV1.5 in the macrophage late endosome regulates endosomal acidification. J Immunol Baltim Md 1950 178:7822–7832. https://doi.org/10.4049/jimmunol.178.12.7822

185. Rahgozar K, Wright E, Carrithers LM, Carrithers MD (2013) Mediation of protection and recovery from experimental autoimmune encephalomyelitis by macrophages expressing the human voltage-gated sodium channel NaV1.5. J Neuropathol Exp Neurol 72:489–504. https://doi.org/10.1097/NEN.0b013e318293eb08

186. Singh S, Florez H (2020) Bioinformatic study to discover natural molecules with activity against COVID-19. F1000Research 9:1203. https://doi.org/10.12688/f1000research.26731.1

187. Li G, Ruan S, Zhao X, et al (2021) Transcriptomic signatures and repurposing drugs for COVID-19 patients: findings of bioinformatics analyses. Comput Struct Biotechnol J 19:1–15. https://doi.org/10.1016/j.csbj.2020.11.056

188. Verbeke L, Mannaerts I, Schierwagen R, et al (2016) FXR agonist obeticholic acid reduces hepatic inflammation and fibrosis in a rat model of toxic cirrhosis. Sci Rep 6:33453. https://doi.org/10.1038/srep33453

189. Carino A, Moraca F, Fiorillo B, et al (2020) Hijacking SARS-CoV-2/ACE2 Receptor Interaction by Natural and Semi-synthetic Steroidal Agents Acting on Functional Pockets on the Receptor Binding Domain. Front Chem 8:572885. https://doi.org/10.3389/fchem.2020.572885

190. Yang C, Xiao S-Y (2021) COVID-19 and inflammatory bowel disease: a pathophysiological assessment. Biomed Pharmacother 111233. https://doi.org/10.1016/j.biopha.2021.111233

191. RECOVERY Collaborative Group, Horby P, Lim WS, et al (2021) Dexamethasone in Hospitalized Patients with Covid-19. N Engl J Med 384:693–704. https://doi.org/10.1056/NEJMoa2021436

192. Barbieri A, Robinson N, Palma G, et al (2020) Can Beta-2-Adrenergic Pathway Be a New Target to Combat SARS-CoV-2 Hyperinflammatory Syndrome?—Lessons Learned From Cancer. Front Immunol 11:. https://doi.org/10.3389/fimmu.2020.588724

193. Burrows B, Knudson RJ, Cline MG, Lebowitz MD (1977) Quantitative Relationships between Cigarette Smoking and Ventilatory Function. Am Rev Respir Dis 115:195–205. https://doi.org/10.1164/arrd.1977.115.2.195

194. Silverman EK, Crapo JD, Make BJ (2018) Chronic Obstructive Pulmonary Disease. In: Jameson JL, Fauci AS, Kasper DL, et al (eds) Harrison's Principles of Internal Medicine, 20th ed. McGraw-Hill Education, New York, NY

195. Feinberg AP (2008) Epigenetics at the Epicenter of Modern Medicine. JAMA 299:1345–1350. https://doi.org/10.1001/jama.299.11.1345

196. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13:484–492. https://doi.org/10.1038/nrg3230

197. Wan ES, Qiu W, Baccarelli A, et al (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet 21:3073–3082. https://doi.org/10.1093/hmg/dds135

198. Breitling LP, Yang R, Korn B, et al (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet 88:450–457. https://doi.org/10.1016/j.ajhg.2011.03.003

199. Morrow JD, Cho MH, Hersh CP, et al (2016) DNA methylation profiling in human lung tissue identifies genes associated with COPD. Epigenetics 11:730–739. https://doi.org/10.1080/15592294.2016.1226451

200. McKenzie AT, Katsyv I, Song W-M, et al (2016) DGCA: A comprehensive R package for Differential Gene Correlation Analysis. BMC Syst Biol 10:106. https://doi.org/10.1186/s12918-016-0349-1

201. Zhang Y, Parmigiani G, Johnson WE (2020) ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics Bioinforma 2:lqaa078. https://doi.org/10.1093/nargab/lqaa078

202. Leek JT, Johnson WE, Parker HS, et al (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28:882–883. https://doi.org/10.1093/bioinformatics/bts034

203. Tate PH, Bird AP (1993) Effects of DNA methylation on DNA-binding proteins and gene expression. Curr Opin Genet Dev 3:226–231. https://doi.org/10.1016/0959-437X(93)90027-M

204. Celli BR, Locantore N, Yates J, et al (2012) Inflammatory Biomarkers Improve Clinical Prediction of Mortality in Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med 185:1065–1072. https://doi.org/10.1164/rccm.201110-1792OC

205. Pelekanou V, Kampa M, Kiagiadaki F, et al (2016) Estrogen anti-inflammatory activity on human monocytes is mediated through cross-talk between estrogen receptor ERα36 and GPR30/GPER1. J Leukoc Biol 99:333–347. https://doi.org/10.1189/jlb.3A0914-430RR

206. Harding AT, Goff MA, Froggatt HM, et al (2021) GPER1 is required to protect fetal health from maternal inflammation. Science 371:271–276. https://doi.org/10.1126/science.aba9001

207. Fiedler U, Augustin HG (2006) Angiopoietins: a link between angiogenesis and inflammation. Trends Immunol 27:552–558. https://doi.org/10.1016/j.it.2006.10.004

208. Parikh SM (2017) The Angiopoietin-Tie2 Signaling Axis in Systemic Inflammation. J Am Soc Nephrol 28:1973–1982. https://doi.org/10.1681/ASN.2017010069

209. Mandeville I, Aubin J, LeBlanc M, et al (2006) Impact of the Loss of Hoxa5 Function on Lung Alveogenesis. Am J Pathol 169:1312–1327. https://doi.org/10.2353/ajpath.2006.051333

210. Hall RJ, O'Loughlin J, Billington CK, et al (2021) Functional genomics of GPR126 in airway smooth muscle and bronchial epithelial cells. FASEB J 35:e21300. https://doi.org/10.1096/fj.202002073R

211. Xu Y, Wang C, Jiang X, et al (2021) KLHL38 involvement in non-small cell lung cancer progression via activation of the Akt signaling pathway. Cell Death Dis 12:1–14. https://doi.org/10.1038/s41419-021-03835-0

List of Figures

List of Tables