UNIVERSITY OF PALERMO

Department of Economics, Business and Statistics

MicroRNA Interaction Network

Ph.D candidate:
Giorgio Bertolazzi

Ph.D. coordinator:
Prof. Andrea Consiglio
DSEAS, UNIPA

Supervisor:
Prof. Michele Tumminello
DSEAS, UNIPA

Company supervisor:
Dr. Claudia Coronnello
Ri.MED Foundation, Palermo

Internship supervisor:
Prof. Panayiotis V. Benos
CSB department, University of Pittsburgh

Keywords: **Bioinformatics; computational methods; sequence analysis; networks; multivariate statistics; machine learning**

XXXIII Ph.D. cycle - 2021


PANORMITANÆ STVDIORVM VNIVERSITATIS SIGILLVM


Fondazione
Ri.MED


UPMC
University of Pittsburgh
Medical Center


MIUR
pon
2014-2020
FONDI
STRUTTURALI
EUROPEI

UNIVERSITY OF PALERMO
Department of Economics, Business and Statistics

MicroRNA Interaction Network

Ph.D candidate:
Giorgio Bertolazzi

Ph.D. coordinator:
Prof. Andrea Consiglio
DSEAS, UNIPA

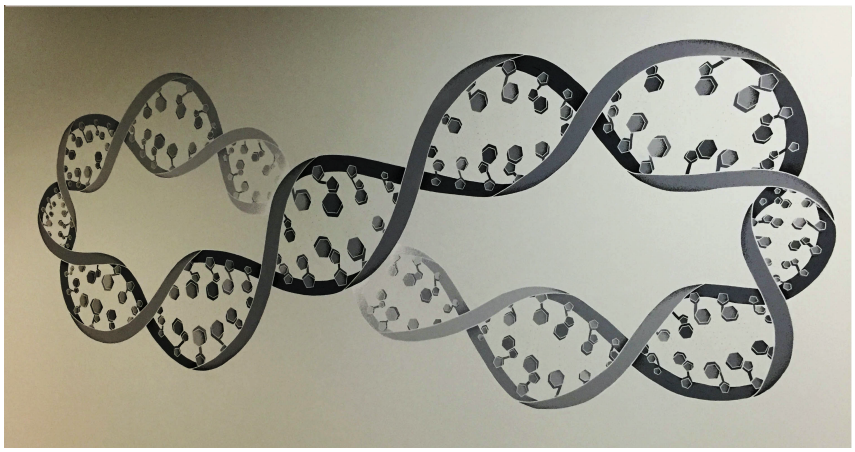
Supervisor:
Prof. Michele Tumminello
DSEAS, UNIPA

Company supervisor:
Dr. Claudia Coronnello
Ri.MED Foundation, Palermo

Internship supervisor:
Prof. Panayiotis V. Benos
CSB department, University of Pittsburgh

Keywords: **Bioinformatics; computational methods; sequence analysis; networks; multivariate statistics; machine learning**

XXXIII Ph.D. cycle - 2021



Achilles: What is that strange flag down at the other end of the track? It reminds me somehow of a print by my favorite artist, M.C. Escher.
Tortoise: That is Zeno's flag.
Achilles: Could it be that the hole in it resembles the holes in a Möbius strip Escher once drew? Something is wrong about that flag, I can tell.
Tortoise: The ring which has been cut from it has the shape of the numeral for zero, which is Zeno's favorite number.
Achilles: But zero hasn't been invented yet! It will only be invented by a Hindu mathematician some millennia hence. And thus, Mr. T., my argument proves that such a flag is impossible.
Tortoise: Your argument is persuasive, Achilles, and I must agree that such a flag is indeed impossible. But it is beautiful anyway, is it not?

Gödel, Escher, Bach: an Eternal Golden Braid
Douglas Hofstadter

Contents

Introduction

MicroRNAs (miRNAs) are small molecules that regulate gene expression through the binding of the target messenger RNA molecules. MicroRNA activity is fundamental in development, differentiation, and other cell functions [1]. Alterations in miRNA activity have been associated with many human diseases [2], such as cancer, diabetes, neurological disorders and dysfunctions of the immune response.
The biological processes related to miRNA activity are extraordinarily complex. For simplicity, we can represent miRNA action as the interaction of two kinds of molecules:

- MicroRNAs
- Messenger RNAs

where messenger RNAs (mRNAs) represent the targets of miRNA binding.
Those interactions can be represented as a bipartite network in which a miRNA can bind a vast number of mRNAs, and, vice versa, an mRNA can be the target of many miRNAs.
Network approach is widely used in molecular biology to study molecular interactions. A network, also called *graph* in topology, is a mathematical object that describes a real system's whole connectivity. It is composed of *nodes* and *links* (also called *vertices* and *edges*); miRNAs and mRNAs correspond to the nodes of a bipartite network, and links represent their interactions.
Network connectivity analysis is closely related to big data and computational issues. Indeed, the number of interactions is generally much bigger than the number of elements. For this reason, multivariate statistics and large-scale inference are helpful in this context.

MicroRNA network construction is strictly related to molecular binding prediction; the whole thesis focuses on developing and applying computa-

tional methods for miRNA target prediction. In particular, a machine learning approach is used to upgrade an existing target prediction algorithm named ComiR.
The thesis is divided into five chapters. Excluding the first one, each chapter corresponds to a paper related to miRNA binding prediction. The order of the chapters is based on a common thread that conceptually goes through algorithm development and usage:

1. Design; problem identification and thoroughly understanding it
2. Analysis; selection of the information on which the algorithm will be based
3. Implementation; algorithm development and performance testing
4. Usage; algorithm running in an empirical application
5. Research for a future upgrade

Following this workflow, we report a brief description of the thesis chapters;

1. **Bioinformatics Databases and Techniques for the Analysis of MicroRNA Interactions**
   The first chapter is an overview of bioinformatics and statistical issues related to the study of miRNA and the prediction of their targets.
2. **Analysis of miRNA Interactions**
   The second chapter reports an analysis of two proteins involved in miRNA activity [3]. This analysis highlights important information that explain miRNA binding behavior. In particular, mRNA coding region information significantly improves the prediction capacity of miRNA target prediction algorithms.
3. **MicroRNA Target Prediction**
   The third chapter regards the development of an algorithm for miRNA target prediction [4]. It represents an upgrade of ComiR algorithm. The results show that the new ComiR version has a higher prediction capacity than the previous ComiR version.
4. **SARS-Cov-2 Sequence Analysis**
   In the fourth chapter, the major miRNA target prediction algorithms

(including ComiR) have been used to identify a small group of miRNAs that potentially bind RNA-sequences of COVID-19 coronavirus [5]. The activity of those miRNAs has been studied to explain biological processes that could be involved with COVID-19 inflammatory state. In particular, miR-1207-5p may contribute to dysregulation of inflammatory responses in COVID-19 disease by targeting SARS-CoV-2 RNA and causing the over-expression of the gene CSF1.

5. **A Novel Statistical Test For Differential Expression Analysis**
   The fifth chapter aims to define a novel statistical test for gene differential expression analysis (DEA).
   DEA consists in the analysis of expression profiles over two groups of samples to identify over-expressed (*enriched*) and under-expressed (*under-represented*) genes. DEA is one of the most used approaches to compare different experimental conditions and highlight differences in gene functions over tissues [6]. We used DEA for identifying differentially expressed genes that compose the machine learning training set of ComiR. The analysis on gene expression data show that the novel statistical test can be integrated into the classical DEA to better identify differentially expressed genes. Moreover, the results suggest that our approach for DEA can be used to upgrade ComiR training set.

Excluding the first chapter, each chapter corresponds to a scientific paper:

2. *RIP-Chip analysis supports different roles for AGO2 and GW182 proteins in recruiting and processing microRNA targets*
   Perconti, Rubino, Contino, Bivona, Bertolazzi, Tumminello, Feo, Giallongo, Coronnello (2017) BMC bioinformatics, 20(Suppl 4):120

3. *An Improvement of ComiR Algorithm by Exploiting mRNA Coding Regions*
   Bertolazzi, Benos, Tumminello, Coronnello (2020) BMC Bioinformatics, 21(Suppl 8):201

4. *miR-1207-5p Can Contribute to Dysregulation of Inflammatory Response in COVID-19 via Targeting SARS-CoV-2 RNA*
   Bertolazzi, Cipollina, Benos, Tumminello, Coronnello (2020) Frontiers in Cellular and Infection Microbiology, Vol. 10, Article 586592

5. *A Novel Statistical Test For Differential Expression Analysis* (manuscript in preparation)

An overall reading of the thesis offers a view of miRNA binding prediction methodologies. Among them, we propose ComiR algorithm as a reliable tool for miRNA target prediction.
The development of the novel ComiR algorithm is accompanied by real data analysis. Specifically, ComiR algorithm has been successfully used to predict human miRNAs that potentially bind COVID-19 genome; we have analyzed endogenous miRNAs that target viral RNA strands to explain inflammatory processes probably involved in COVID-19 disease.
This thesis lays the foundations for the upgrade of ComiR webtool. The novel algorithm significantly improves the ComiR prediction capacity by including miRNA binding sites located on mRNA coding regions. ComiR facilitates the investigation of miRNA binding and can be used to build a large mRNA-miRNA network.

Chapter 1

Bioinformatics Databases and Techniques for the Analysis of MicroRNA Interactions

The study of microRNA (miRNAs) is an important branch of molecular biology. MicroRNAs are small RNA molecules (20-24 nucleotides) that regulate gene expression by binding messenger RNA molecules (mRNAs), whereas mRNAs represent miRNA binding targets.
The prediction of miRNA binding sites has a key role in the study of miRNA interactions. It is fundamental for the identification of miRNA targets and understanding miRNA regulation activity.
The main topic of this thesis is the prediction of miRNA binding sites using a computational approach. This approach requires a wide variety of multidisciplinary skills that involve different fields of scientific and technical knowledge, such as bioinformatics, molecular biology, chemistry, statistics, and topology.
The present chapter is an overview of multidisciplinary techniques and strategies for the analysis of miRNA interactions. We used these methodologies for developing miRNA target prediction algorithms and for data analyses. Therefore, the following sections outline the bioinformatics framework of the thesis:

1. The first section reports a biological introduction to miRNAs and describes the procedures for predicting miRNA targets; those procedures include the use of bioinformatics sequencing databases as input for miRNA target prediction algorithms.

2. The second section describes biological experiments used to measure gene expression and to validate miRNA interactions. Those experiments pro-

duce part of the data used in all the analyzes of the thesis.

3. The last section lingers on statistical techniques for the analysis of genetic big data. It introduces large scale inferences methodologies and clarifies our strategy for developing miRNA target prediction algorithms.

1.1 MicroRNA molecular interactions

The fist two miRNAs, named *lin-4* and *let-7*, were identified in 1993 as regulators of *Caenorhabditis elegans* development [7][8]. Initially, those molecules were considered an unusual worm specific gene expression regulation mechanism. Today, miRNAs are considered important regulators of gene expression in eukaryotes. MicroRNA-mediated gene regulation is part of a more sweeping mechanism known as *RNA interference* (RNAi) [9]. Disorders in RNAi mechanism are connected with a wide variety of human pathologies [2].
MicroRNA regulation activity depends on the recognition and binding of mRNA target molecules. Therefore the prediction of miRNA binding site located on mRNAs is fundamental in the study of miRNA interactions.
This section focuses on the description of miRNA target prediction procedures. Those procedures include the download and interpretation of databases containing miRNA and mRNA sequencing; nucleotide sequences represent the input for many algorithms used to predict miRNA binding sites.
A biological explanation of miRNA activity precedes the description of miRNA target prediction procedures. Before going into detail about miRNA mechanisms, a brief introduction of basic genetic concepts is reported. Readers with a basic knowledge of protein synthesis can skip the next paragraph.

1.1.1 Central dogma of molecular biology

The *genome* is the whole genetic material of an organism. It contains all the information needed for the growth and development of that organism. A genome consists of DNA; it incorporates both non-coding and coding DNA. *Genes* are generally defined as DNA coding sequences that contain the information used to synthesize proteins (non-coding sequences will be described in the next section).

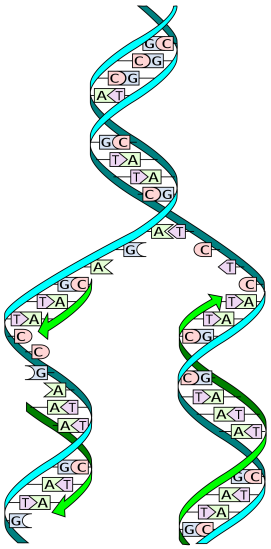


Figure 1.1: *DNA replication. The new strands are composed of complementary nucleotides* [11].

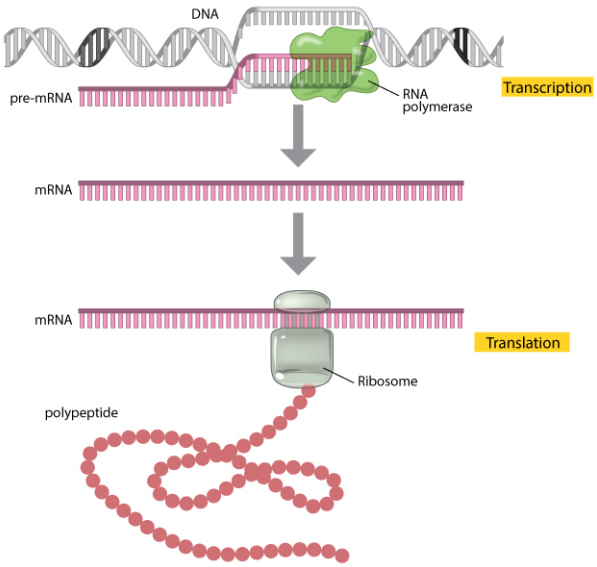


Figure 1.2: *Schematic description of the central dogma of molecular biology* [12].

The *central dogma of molecular biology*[1] describes how DNA information is transcribed into coding RNA to be used for protein synthesis. This process is the fundamental basis for life on earth. Below we provide a very brief description of the phases that characterize the flow of genetic information from DNA to protein synthesis:

1. **DNA replication**

   DNA replication is the process by which DNA makes a copy of itself during cell division. DNA is composed of a double helix of two complementary strands (Fig.1.1). The elements that compose DNA strands are called nucleotides; each nucleotide is characterized by a nitrogen-containing nucleobase (cytosine [C], guanine [G], adenine [A], or thymine [T]). Those bases determine the complementary pairing of DNA strands

[1] The wording *dogma* has a historical reason, and it is still used although is not a dogma at all. This word was proposed by Francis Crick in 1958. When Horace F. Judson asked Crick how and why he coined the expression "*the central dogma*" he said: "*A dogma was an idea for which there wasn't reasonable evidence. You see?... I just didn't know what dogma meant. And I could just as well have called it the "Central Hypothesis". Which is what I mean to say. Dogma was just a catchphrase...*[10]"

(T-A, G-C). During DNA replication, the two strands separate each other. The separated strands will act as templates for making the new strands of DNA, where the synthesis of the new strands is driven by an enzyme called DNA polymerase that allows the binding between complementary nucleotides.

2. **Transcription**

   Transcription is a biological process in which the genetic information contained within DNA is re-written into messenger RNA (mRNA).
   The formation of mRNA is mediated by RNA polymerase. This enzyme uses DNA strands as a template to create the mRNA. The new RNA strand is complementary to the DNA strand used as a template, except that adenine's complementary nucleobase is uracil [U] and not thymine anymore.

3. **Translation**

   The translation is the process in which macromolecular machines, called ribosomes, use mRNA information to synthesize proteins.
   After the transcription, mature mRNA molecules leave the nucleus and travel to the cytoplasm, where they find ribosomes. Ribosomes "read" the information contained in mRNAs and use this information as a template to assemble the chain of amino acids that compose a protein.

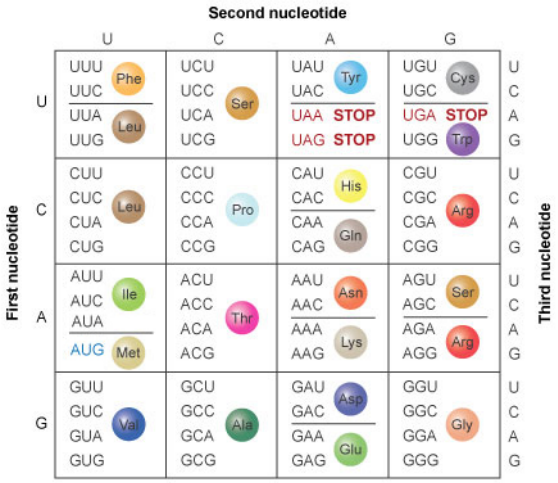


Figure 1.3: *Codons that codify specific amino acids. Multiple codons can code for the same amino acid* [12].

The language of nucleotides is based on triplets: Each group of three bases in mRNA constitutes a *codon*, and each codon is associated with a particular amino acid (Fig.1.3). The total number of triplets is $4^3 = 64$, which is bigger than the total number of amino acids (23); it implies that some triplets correspond to the same amino acid.
All the organisms share the same genetic language; it means that the association between codons and amino acids is identical all over the living beings. It is considered evidence that all the organisms of the planet have common ancestors.

1.1.2 Non-coding RNA

Non-coding DNA composes a large part of eukaryotic genomes. For several years the role of non-coding DNA was unknown. The scientific community considered those DNA strands without any biological function because they don't encode protein sequences. For this reason, this kind of DNA was named "*junk DNA*".
Today, many studies emphasize the importance of non-coding DNA for cell life and evolution [13]. Non-coding DNA strands (i.e., non-coding genes) are transcribed into functional non-coding RNAs. Exist different types of non-coding RNAs (e.g., transfer RNA, ribosomal RNA, regulatory RNA), whereas each RNA type has a specific function inside the cell.
MicroRNAs are a particular type of regulatory RNA. Their importance is related to the RNA interference regulatory process. In the next section, miRNA activity is described as part of a broader mechanism regulating gene expression.

1.1.3 RNA interference and gene expression regulation

A gene is expressed if its information is used in the synthesis of the correspondent RNA.
Gene expression level is estimated as the amount of mRNA molecules in the cell transcribed from that gene, and *gene expression profiling* is the expression measurement of thousands of genes simultaneously. It provides a global view of cellular activity and functions. The methods used to measure gene expression will be described in the next section.
The gene expression profile reflects cell functionality; each type of cell produces its own transcripts that depend on its role inside the organism. Moreover, the

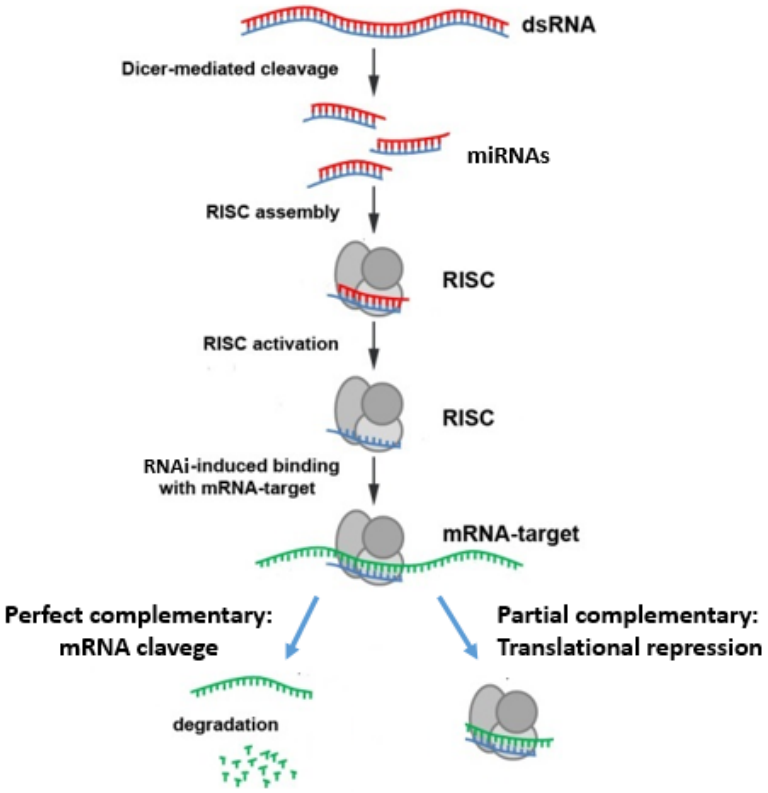


Figure 1.4: *Graphical description of RNA-interference process* [14].

transcripts of a single cell depend on organism's nutrients, environmental response, and life cycle; therefore, the transcripts vary over time and are never stable. For this reason, each cell uses many mechanisms that regulate the expression of its genes; in this way, a gene could be activated if required. On the other hand, gene activity can be suppressed if it is temporary not necessary.
The primary control point for gene expression is transcription initiation, but many other gene regulation mechanisms exist. Among them, we focus on the RNA-interference regulation process, and in particular on microRNA interference.

RNA interference (RNAi) is a cell process in which RNA molecules physically interact with mRNA molecules to suppress gene expression. RNAi represents a post-transcription regulation process widely used by eukaryotic organisms.
Below is a detailed description of the mechanisms that lead the formation, activation, and action of RNAi molecules (Fig.1.4):
In the first step of the RNAi process, a specific ribonuclease enzyme, called

Dicer, binds and cleaves long double-strand RNAs yielding short (21-23 nt) duplexes with 2-overhanged nucleotides at the 3'-ends [14]. RNA molecules produced by the Dicer are called *microRNAs* (miRNAs). Those molecules are the main protagonists during the interaction with mRNAs.
RNAi molecules' activity depends on the formation of a protein complex called *RNA-induced silencing complex* (RISC). The RISC incorporates one strand of RNAi molecules; that strand acts as a template for RISC to recognize complementary messenger RNA (mRNA) transcript. In different organisms, the RISC complex varies its composition, but one protein family, called *Argonaute*, always composes the RISC.
MicroRNA binding is based on nucleotide complementarity. Perfect complementarity between miRNA-mRNA pairs is quite rare, but also a six base-pair match could be sufficient for the binding.
MicroRNA binding is often not specific; indeed, a single miRNA molecule can potentially bind a vast number of mRNAs. On the other hand, a single mRNA could be the target of many miRNAs. In this context, an mRNA bounded by a miRNA represents the *target molecule* of the binding. Using similar terminology, a gene whose expression is regulated by a miRNA is called *target gene*.
The whole interactions between miRNAs and mRNAs compose a huge molecular network that is not entirely known. This thesis focuses on developing novel algorithms that allow the exploration of miRNA molecular connectivity; the following sections introduce the main computational methods for miRNA target prediction and describe the bioinformatics datasets used for the analyzes. Before going through those computational aspects (that are the thesis central topics), the next section lingers on the importance of miRNA activity in gene regulation processes.

1.1.4 Role of microRNAs in biological processes

MicroRNA activity is involved in several biological processes such as cell differentiation, apoptosis, development, and immune response through target gene regulation [1]. The discovery of miRNAs was a crucial point in molecular biology; it gives the possibility to study gene expression from a new perspective. RNAi is used in many experiments by introducing synthetic double-strand RNAs into cells to suppress specific genes selectively. It can help to identify the components necessary for a particular cellular process.
It has been proposed that approximately 30% of the human protein coding

genes are controlled by miRNAs [15]. For this reason, the processes in which miRNAs are involved are very heterogeneous.
Recent evidence supports the idea that miRNAs are fundamental during development; a development arrest has been observed in animal embryos characterized by dysregulations in RNAi process, e.g., Dicer deficiency caused death in invertebrates and zebrafish embryos. MicroRNA are also important in nervous system development, and miRNAs have been found in dendrites and axons of neurons [16]. During cardiac muscle development, miR-1 and miR-133 have been recognized as important regulators [17]. MicroRNAs have been found to regulate immune responses [18] and regulate genes involved in inflammatory states [19][20]. Host miRNAs also interact with viruses; many complex virus-specific mechanisms are not yet fully understood [21]. Finally miRNAs are clearly involved in angiogenesis [23] and apoptosis [22].
Alterations and mutations in miRNAs have been associated with several human diseases such as cancer, immune disorders, and neurological and cardiovascular diseases. Downregulation of miRNAs might unblock the suppression of oncogenic genes. On the other hand, overexpressed miRNAs might inhibit tumor suppressor genes; in both situations, miRNAs are strongly involved in tumors [24]. Altered miRNAs have been observed in autoimmune diseases such as arthritis, Systemic Lupus Erythematosus, and multiple sclerosis [25].
Recently, microRNAs have been detected in serum and plasma, and circulating microRNA profiles have been associated with some different tumor types, diseases such as stroke and heart disease, and altered physiological states such as pregnancy [26]. Therefore, the use of miRNAs as diagnostic biomarkers represents a new clinical application of miRNAs.
Finally, in the future, miRNAs could become used as drugs to treat gene expression disorders. This promise doesn't look so far; therapeutic application of miRNA are beginning to become reality [27] [28]. For example, miRNA activity could be used to combat viral infections [29]. This perspective leads the research to the study of miRNA-virus interactions. This topic will be addressed in Chapter 4.

1.1.5 Genome sequence databases

The prediction of miRNA binding sites is based on matching complementary sequences located on miRNA and mRNA strands. For this reason, miRNA target prediction algorithms generally require nucleotide sequences as input.

Public genome databases are available online, and the complete genome of several species is easily downloadable. Below we report a brief description of three sequencing databases that we have used for miRNA binding research on *Human* and *Drosophila* genomes (Chapters 2 and 3) and on SARS-CoV-2 virus (Chapter 4).

BioMart Ensembl genome database

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation, and transcriptional regulation [30]. Ensembl tools compute multiple alignments, predict regulatory functions, and evaluate the effect of rare genetic variants.
BioMart tool reports the whole genome annotation of more than 100 species. We have downloaded the entire transcriptome of two species; i.e., *Human* (GRCh38.p13) and *Drosophila melanogaster* (BDGP6.28). MicroRNA binding sites have been identified by searching complementary matches on those sequences. BioMart datasets are also downloadable from R console using the Bioconductor package *biomaRt* [31].

NCBI database

The NCBI web platform includes a series of databases relevant to biotechnology and biomedicine [32]. The Nucleotide database collects sequences from several sources, including GenBank, RefSeq, TPA, and PDB. Genome, gene, and transcript sequence data provide essential information for biomedical research and discovery.
A total of 15881 worldwide viral complete genomes was downloaded —updated to September 7th, 2020— from the Severe acute respiratory syndrome coronavirus 2 data hub of NCBI Virus database, by filtering for taxid = "2697049" and Nucleotide Completeness = "complete". The RefSeq sequence NC_045512 was used as reference to predict the binding sites of human miRNAs on the viral RNA.

miRBase: The microRNA database

The miRBase database is a database of published miRNA sequences and annotations. It represents the primary online repository for miRNA sequence data [33]. miRBase reports all known miRNA mature sequences and their locations.

We have downloaded all miRNA sequences of *Human* and *D. Melanogaster*. Those sequences have been used for miRNA binding prediction in Chapters 2, 3, and 4.

1.1.6 *In silico* predictions of microRNA binding sites

One of the most common experimental strategies used to investigate miRNA targets consists of the immunoprecipitation of RISC proteins [35][36]. The high costs of experiments oriented the miRNA target identification towards a computational approach; it consists of the reproduction of biological experiments using computational simulations. *In silico* predictions of molecular binding generally guide the decision on which *in vitro* experiments could be important to carry out.
MicroRNA binding prediction algorithms are generally based on Watson-Crick base-pair matching [37][38][39]. Perfect complementarity between miRNA-mRNA pairs is quite rare, but also a six base-pair match could be sufficient to suppress gene expression. Few other methods use the miRNA expression profile as additional information to predict miRNA targets: GenMir++[40], PicTar [41], Talasso [42], and ComiR [43][44].
Most of those algorithms consider only the binding sites located on 3'UTR region of mRNAs. In Chapter 3, we propose an upgrade of ComiR algorithm by considering the binding sites located on the coding region.
This section focuses on three tools for predicting miRNA binding sites: TargetScan, PITA, and miRanda. Those tools run a deep research on RNA sequences looking for nucleotide complementarity. The binding scores calculated by these tools represent the basic information used by ComiR algorithm for target prediction. For this reason, we run those algorithms on the entire *D. Melanogaster* and *Human* transcriptomes as show in Chapters 2 and 3.

TargetScan

MicroRNAs recognize their mRNA targets by base-pairing interactions that involve nucleotides 2-8 of miRNA (*seed region*).
TargetScan predicts miRNA targets by finding perfect Watson-Crick (WC) seed complementarity [45], whereas a perfect matching is called *canonical*.
TargetScan algorithm distinguishes four different types of canonical binding sites (Fig.1.9); 6mer indicates the weakest matching (only six nucleotides),

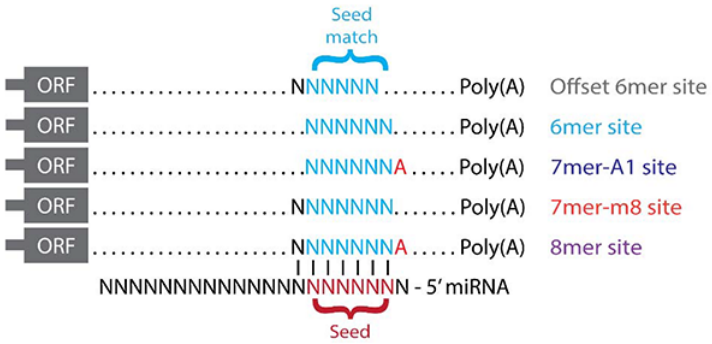


Figure 1.5: *Canonical site types classified by TargetScan algorithm* [47].

while 8mer indicates the strongest matching. In our analysis, we have considered matches that involve at least seven nucleotides; therefore, 6mer sites have been excluded.
The difference between binding site types depends not only on the number of binding sites in the WC matching but also on the presence of the base A at the first position downstream of the seed match. This nucleotide could pair with the first nucleotide of a miRNA whose first nucleotide is U. However, a conserved A was also observed next to seed matches for miRNAs that do not begin with an U base. This "*A anchor*" would increase the specificity of target prediction; therefore, the conservation of A characterize stronger matching [45]. Non-canonical sites are excluded during the binding research. Indeed the authors have demonstrated that non-canonical sites do not mediate repression despite binding the miRNA, which indicates that the vast majority of functional sites are canonical [46].

miRanda

miRanda algorithm detects potential miRNA binding sites in genomic sequences [48]. The main difference with Targetscan is that miRanda also researches non-canonical matches (Fig.1.6); therefore, sequence complementarity

```
Query:    3' gtCGAAAGTTTTACTAGAGTg 5' (eg. miRNA)
               |:|||||  |||||||||:
Ref:      5' taGTTTTCACAATGATCTCGg 3' (eg. 3'UTR)
```

Figure 1.6: *Example of non-canonical sequence matching obtained through miRanda output.*

is not the only criteria during the target research.
miRanda algorithm is divided into two phases:

1. Sequence matching to assess whether two sequences are complementary and computation of a matching score.

2. Each significant match is associated with an energy of physical interaction.

PITA

During miRNA binding, there is an energy cost of base-pairing interaction to make the target accessible for the binding. PITA algorithm [49] calculates miRNA binding scores taking into account the site accessibility of the targets. PITA interaction scores are computed as the difference between the free energy gained from the formation of the microRNA-target duplex and the energetic cost of unpairing the target to make it accessible to the microRNA. The lower is the score the higher is the probability of interaction.

1.1.7 Molecular Interaction Networks

Over the last decade, molecular biology methodologies have been strongly influenced by network theory [50]. System biology perspective focuses on physical interactions between biological elements, whereas the whole set of interactions between cell components is called *Interactome* [51].
The construction of the *Interactome* is an ambitious challenge in molecular biology. In this context, network models are perfectly suitable to represent and analyze molecular interactions. Molecular networks describe interactions and pathways which characterize cell processes. Fig.1.7 shows interactions between heterogeneous elements involved in cell processes, each of those physical and functional interactions can be represented using a network.
A wide variety of molecular networks have been defined; e.g., *metabolic networks* [54] represent relation between chemical reactions and their substrates using a bipartite representation; *gene regulatory networks* [55] describe the mechanism of gene expression regulation using graphical models; *protein-protein interaction networks* [56] represent protein binding. All those networks typically have a lot of properties in common with macro-scale systems. For example, they have a scale-free degree distribution and a preferential attachment

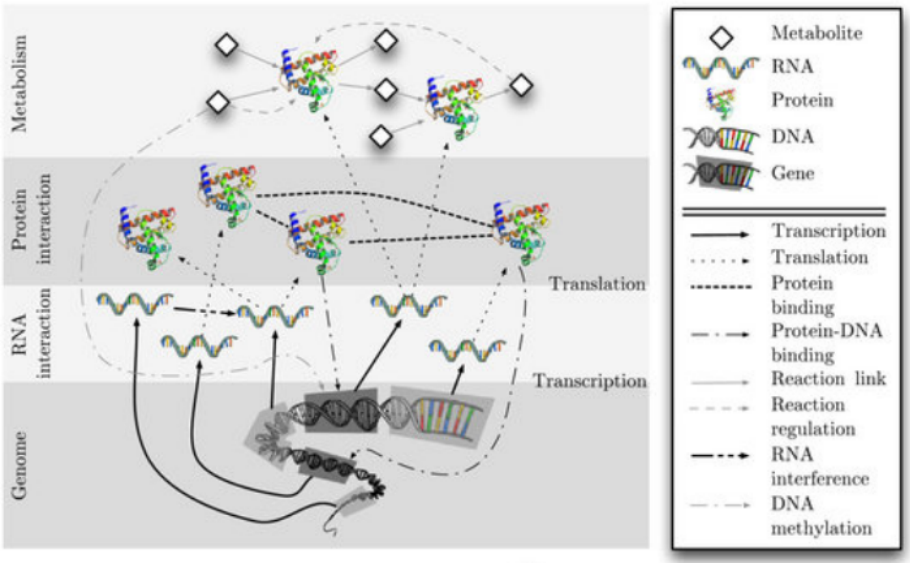


Figure 1.7: *Schematic overview on molecular interactions in the cell* [50].

behavior typical of scale-freeness.

MicroRNAs and their targets compose a bipartite network. But most of the miRNA connectivity is still unknown because the high cost of experiments limits the knowledge of specific binding. *In silico* predictions of miRNA interactions permit to highlight molecular interactions that are still unknown. In the following sections, we introduce the methods to be integrated with *in Silico* predictions for constructing miRNA bipartite networks.

1.2 Gene expression analysis

Most modern genetic studies are related to gene expression analysis; indeed, gene expression profiles represent the primary information to investigate cell activity and functions. In this thesis, the analysis of gene expression is used for many different purposes. For example, miRNA expression is used to improve the performance of miRNA target prediction algorithms (Chapters 2 and 3). Moreover, empirical validation of miRNA interactions requires a significant analysis of differentially expressed profiles (as described in section 1.3.2). In general, gene expression analysis is an important part of all the analyzes presented in the thesis.

The expression level of a gene corresponds to the amount of RNA transcript of that gene. This section reports a brief description of throughput techniques for quantifying gene expression levels.

1.2.1 Throughput biological experiments

There are three main approaches for recording gene expression: hybridization DNA microarrays, next-generation sequencing, and real-time quantitative PCR.

DNA microarray

DNA microarray is a technique used for to identify and quantify mRNA transcript presents in the cell [59]. In this thesis, microarray data are the most used data (Chapters 2, 3, and 5).
DNA microarray is physically composed by is a collection of microscopic DNA spots attached to a solid surface. Each DNA spot contains little probes of a specific DNA sequence. Those probes links with complementary DNA (cDNA) previously synthesized from RNA molecules.
The transcript from which the cDNA comes is typically collected from two different samples. For example, the baseline sample from a healthy individual and the experimental sample from a diseased individual (e.g., cancer sample). The two mRNA samples are then converted into cDNA using a reaction catalyzed by the enzyme reverse transcriptase[2], and each sample is marked with a fluorescent probe having a specific color. For example, the experimental cDNA sample may be marked with a red fluorescent dye, whereas the reference cDNA may be marked with a green fluorescent dye (Fig.1.8). The two samples are then mixed and allowed to bind the microarray slide. Microarray is considered a hybridization procedure because cDNA molecules from two different samples bind to the DNA probes.
Microarray data are obtained by quantifying fluorescent pixel intensities. The pre-processing of raw data can be divided in three main steps:

1. *Background correction* purifies data from background noise. A common method is the *normexp*; it models the observed pixel intensities as

[2] Reverse transcription (RT) is a process in which complementary DNA (cDNA) is created from an RNA template using an enzyme named reverse transcriptase. It occurs as part of specific mechanisms; for example, retroviruses use RT to replicate their genomes (e.g., HIV). RT represents an exception to the central dogma of molecular biology. The use of reverse transcriptase to measure gene expression was a revolution in molecular biology; indeed, it permitted the development of high throughput experiments.

the sum of two random variables; one normally distributed (background noise) and the other exponentially distributed (signal) [60].
Once background component has been quantified, gene expression can be calculated as follows:

$$\text{Density of Red} = R_{fg} - R_{bg}$$

$$\text{Density of Green} = G_{fg} - G_{bg}$$

$$FC - Expression = log_2\left(\frac{Density\,of\,Red}{Density\,of\,Green}\right) \tag{1.1}$$

where, fg = foreground, and bg = background.

2. *Normalization* consists in data scaling and transforming. It is used to make expression profiles from different conditions comparable (see Section 5.2).

3. *Summarization.* In microarray, we may have several values for the same spot or gene. Therefore, the information has to be synthesized in a single measure for each gene.

Next generation sequencing

Next-generation sequencing (NGS) [61] is a group of techniques based on a massive approach for identifying genome sequences and quantifying their expression levels. These techniques are also used for genome comparisons, identification of gene variants, and the research of new miRNAs.
NGS profiling is more accurate than hybridization-based technologies, and it will probably replace microarray technique in the future. Moreover, NGS expression levels are highly correlated with real-time PCR experiments [62]. Tab.1.1 reports a schematic comparison of high throughput techniques. As reported in the table, NGS is used not only for quantifying transcript amounts but also to determinate the primary structure of genome sequences.
Several platforms have been developed to run NGS (e.g., Illumina Genome Analyzer, SOLID, Roche), and they are widely used in biological research.

	Real-time PCR	**Microarray**	**NGS**
Throughput	Medium	High	Ultra high
Principle	PCR amplification	Hybridization	Sequencing
Time	<6h	2 days	1-2 weeks
Sample input	10 ng - 500 ng	100 ng - 1 μg	500 ng - 10 μg
Applications	Counting	Counting	Reading and Counting

Table 1.1: *Comparison of throughput techniques* [63]. *Throughput applications are divided into two main categories: "reading" and "counting". Reading consists of the sequencing itself; it is used to study an unknown genome or to research genomic variants. Counting consists of the profiling of gene expression; it is the quantification of the amounts of transcripts in the cell.*

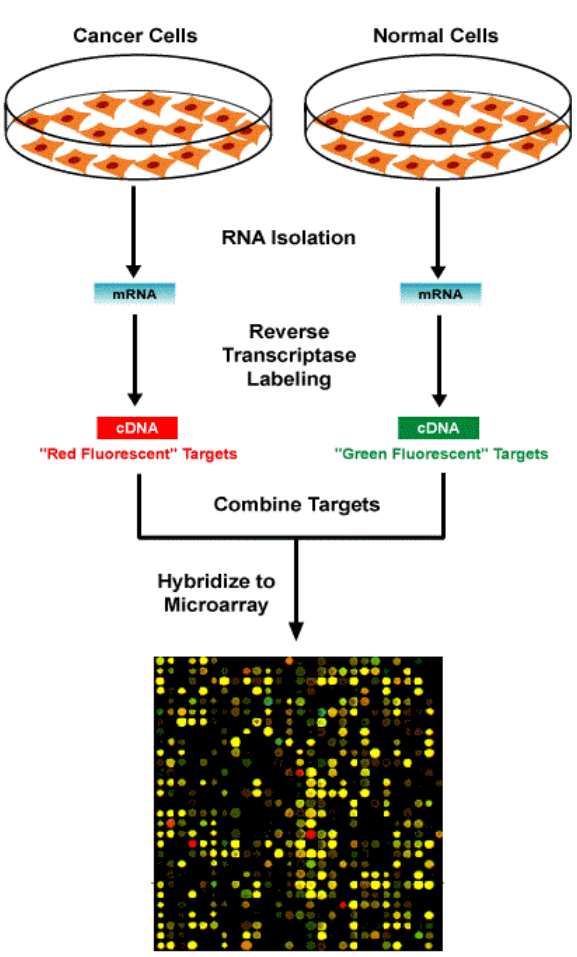


Figure 1.8: *Steps in microarray data collection* [64].

RT-PCR

Reverse transcript polymerase chain reaction (RT-PCR) is a technique for gene expression measurements [58]. It is generally used as a gold standard for the validation of other profiling approaches. PCR is also used for disease diagnosis; for example, during the COVID-19 pandemic, PCR has been the most used diagnostic test.
RNA sample is first reverse-transcribed to complementary DNA (cDNA) using the reverse transcriptase enzyme. The cDNA sample is amplified in billions of copies using the polymerase chain reaction (PCR), and the amount of cDNA is quantified. The quantification of amplified DNA molecules is carried out by marking cDNA using a fluorescent dye. The fluorescence signal increases proportionally to the amount of replicated DNA, and the DNA is quantified in "*real time*".

1.2.2 Experimental validation of miRNA targets

In silico algorithms predict a large number of miRNA interactions, but many of those interactions have to be discarded because they don't take place inside the cell. Therefore, a filtering criterion is important to select functional miRNA interactions, and empirical validation of computational results is often required.
The experimental approach permits more reliable identification of miRNA targets. Unfortunately, the high cost of experiments doesn't allow large scale inference. For this reason, the computational approach is used to investigate a large number of interactions and could orient biological experiments.
Empirically validated interactions are often used to build the training set of machine learning algorithms for miRNA target prediction. This is the case of ComiR algorithm [43][4] (Chapter 3).
MicroRNA target validation often requires the analysis of transcript changes over experiments (e.g., IP samples *vs* input samples). Indeed, differential expressed genes over experiments represent the most reliable validated miRNA targets. Section 1.3.2 deeply describe statistical methods for differential expression analysis. Below we briefly introduce the main biological experiments used for empirical validation of miRNA targets.

MicroRNA transfection

MicroRNA transfection consists of the introduction of exogenous miRNAs into eukaryotic cells. Exogenous miRNAs cause decreases in the abundance of mRNA transcript; therefore, mRNA targets can be identified by studying changes in transcript abundance [66]; therefore, differentially expressed genes are the probable miRNA targets.

Depletion of RISC proteins

MicroRNA activity depends on the formation of a ribonucleoprotein complex named the RNA-induced silencing complex (RISC). The depletion of a RISC protein causes the knockout of RISC activity and the consequent inhibition of miRNA activity [145]. It causes a variation in the abundance of mRNA targets; therefore, most over-expressed genes correspond to miRNA targets, and the analysis of differentially expressed genes can be used to identify them.

Immunoprecipitation of RISC proteins

One limitation of the previous approaches is that targets are inferred by considering only the changes in mRNA abundance. However, a miRNA could indirectly influence gene expression without direct binding. Moreover, analysis of transcript changes doesn't return information on which targets have an important role in carrying out the actual biological processes [36].
Immunoprecipitation (IP) of RISC proteins is a direct experimental method to identify miRNA targets. This technique uses an antibody that specifically binds to a particular RISC protein, and then the whole protein complex is isolated by precipitating the bounded proteins out of solution.
The immunoprecipitated is analyzed using high throughput methods for expression profiling, such as gene array (RIP-Chip) or sequencing (RIP-Seq), which allow the systematic identification of RISC-bound miRNAs and their target mRNA sequences. Finally, genes enriched and underrepresented in IP samples can be identified through a differentially expressed analysis.
IP experiments, like the previous experimental approaches, can't recognize miRNA binding sites. IP analysis identifies groups of molecules (miRNAs and mRNAs in our case) that bind each other. For this reason, a computational approach is generally used to search miRNA-mRNA couples that can potentially bind.

IP approach has been widely applied to the AGO protein family [67][68][69] [70]. Moreover, also the immunoprecipitation of GW182 RISC protein has been recently realized [71][72][73]. Chapter 2 reports a RIP-ChIP analysis of the role of AGO2 and GW182 in recruiting and processing miRNA targets. Using a differential expression analysis on IP experiments, we identified groups of miRNA target genes that represent the gold standard predictions for evaluating algorithm prediction capacities. Moreover, in Chapter 3, IP experiments are considered to build the training set of ComiR algorithm. The next section introduces the statistical methods to carry out differential expression analysis for identifying validated miRNA targets.

1.3 Large scale inference at time of genetic big data

The present section describes the statistical methodologies used for the data analyzes presented in the thesis. Specifically, we will outline the main steps for developing miRNA target prediction algorithms by connecting all the concepts introduced until now. Before that, few other notions have to be introduced.
In genetics, classical statistical methods cannot be used because of the high-dimensionality of high throughput data. Indeed, the number of dimensions is generally higher than the number of observations. This multivariate structure of the data causes computational problems; this situation is referred to as the *curse of dimensionality* [75]. For this reason, specific approaches have been developed for the analysis of genetic big data. In this section, we will describe the statistical methods used in the next chapters.

1.3.1 Pre-processing of gene expression data

The first step in statistical analysis is the pre-processing of raw data. It is an important part of the analysis because data transformations influence final results.
In this thesis, microarray data are frequently used. Microarray expression profiles generally have a *log-normal* distribution. Moreover, miRNA gene expression generally follows a *power law* distribution [76], therefore miRNA expression profiles don't have a proper scale, and miRNA variance could diverge. Log2-transformation is the most popular way to reduce expression variability

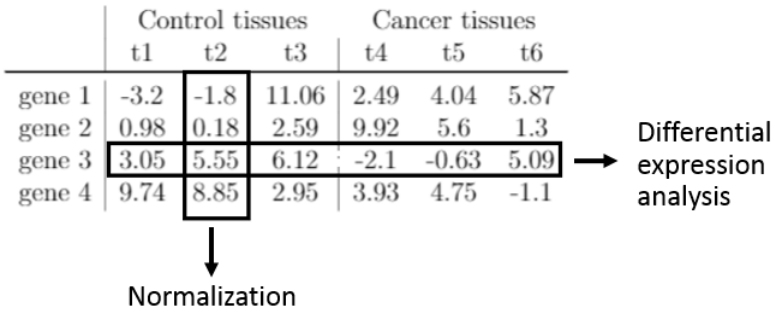


Figure 1.9: *Example of genetic dataset structure. Each column corresponds to a sample from a profiling experiment, while each rows reports the expression of a single gene over a group of tissues. Normalization is performed by column and DEA is performed by row.*

and to make the profile distribution symmetric.
In transcriptomics studies, different tissues are often compared to each other. It requires that the total amount of transcript is the same over different tissues, but this situation is generally not observed in raw data; therefore, data normalization always precedes this analysis. A simple solution is to divide expression values by the sum of the expressions. Another solution is the *quantile normalization* [77], it makes profile distributions identical in statistical properties.

1.3.2 Differential expression analysis

Differential expression analysis (DEA) is one of the most used approaches to compare different experimental conditions and highlight differences in gene functions over tissues [6]. In the present thesis, differentially expressed (DE) genes are analyzed in all chapters; In Chapters 2 and 3, DE genes from IP experiments compose the validated targets used for training and testing miRNA target prediction algorithms. Instead, Chapter 5 proposes a novel statistical test for DEA. Moreover, in Chapter 4, healthy and COVID-19 tissues are compared, and DE genes are identified as possible protagonists of the COVID-19 inflammatory state.
DEA consists in the analysis of expression profiles over two groups of samples to identify over-expressed (*enriched*) and under-expressed (*under-represented*) genes, where each sample group corresponds to a different experimental condition. Gene expression behavior over different conditions is related to specific

biological processes and functions. Therefore, DEA is widely used together with gene classification and enrichment analysis of Gene Ontology (GO) categories (to pursue this goal, GO-analysis on DE genes is a common procedure; it is described in section 1.3.6). For example, genes frequently over-expressed in a specific tissue are probably involved in cell activities that characterize that tissue. DEA is also carried out to investigate genetic diseases; a gene that is over-expressed in cancer tissues could be directly involved in cancer.
Let be $\overline{x}_i$ and $\overline{y}_i$ the average expression values of the i^{th} gene in two groups of tissues that we want to compare. The two most popular approaches for DEA are the fold-change analysis and the use of a large family of t-tests. Those two approaches are generally applied together.

1. Fold-change (FC) indicates the expression variation of a gene over two different conditions [78]. It can be calculated as a ratio generally expressed in log-scale:
$$log(FC_i) = log_2 \frac{\overline{x}_i}{\overline{y}_i} \tag{1.2}$$
The genes whose $log(FC)$ is greater (or lower) than a threshold (e.g., ± 2) are identified as differentially expressed, but threshold selection is arbitrary.

2. A large family of t-tests is the most widely used procedure for DEA [203] [204]. This procedure is strictly related to large scale inference, and p-values are corrected using multiple comparison procedures.

The t-test is based on parametric assumptions rarely satisfied. However, large samples allow an assumption relaxation, but the high cost of experiments makes it difficult to find. For this reason, in small skewed samples, t-test p-values are often not reliable [205]. Moreover, the small variance of low expressed genes makes the denominator of t-test statistics unnaturally smaller. It increases the total I type error and the number of significant genes. Alternative definitions of the t-test have been proposed to reduce the impact of small samples and low expression variability, e.g., *Significance Analysis of Microarray* (SAM) [207] and *moderated t-test* [206].
In SAM, a small constant s_0 is added in the denominator of the t-statistic to correct small variance values:
$$t_i = \frac{\overline{x}_i - \overline{y}_i}{s_i + s_0} \tag{1.3}$$

Instead, the *moderated-t-test* is based on an empirical Bayes approach for estimating sample variances towards a pooled estimate, resulting more stable inference when the number of arrays is small [206].
On the other hand, large sample t-tests produce too many significant genes; it depends on average expression differences truly different from zero but not large enough to be biologically meaningful. A common strategy to reduce the number of selected differentially expressed genes is to set an arbitrary threshold on the fold change (e.g., 1, 1.5, or 2) [208].
In Chapter 5, we proposed a novel statistical test for DEA. The results suggest its application together with the t-test approach to better understand the biological questions related to DEA.

1.3.3 Machine learning approach

Machine learning is a class of algorithms that carry out statistical predictions by recognizing empirical patterns in the data [86]. Those algorithms are connected to artificial intelligence because they improve their prediction capacity with empirical experience.
The machine learning approach was specifically developed for the analysis of big data. For this reason, it has been widely used in genetics. In Chapter 3, a *support vector machine* (SVM) model is used as a structural part of ComiR algorithm [4][43]; the scores of three miRNA target prediction algorithms (i.e., PITA, miRanda, and TargetScan) are combined in an SVM model to predict miRNA target genes. SVM is a supervised method; the model is trained on a set of elements already classified into two groups. Those classified elements are used as an example to identify empirical patterns that will permit the classification of new elements. In our case, the training set is composed of target and non-target genes of a group of miRNAs, where those genes have been empirically validated through biological experiments (i.e., DE genes from IP experiments).
The idea behind an SVM is to maximize the margin between two groups and minimize the total classification errors by placing a hyperplane in high dimensional space, where the number of dimensions corresponds to the number of variables given in input for the classification [85]. Therefore, the SVM classifies new elements on the base of the hyperplane previously adapted. The shape of the hyperplane depends on a kernel function, Fig.1.10 shows a classification based on a linear kernel, whereas many non-linear kernels have been proposed.

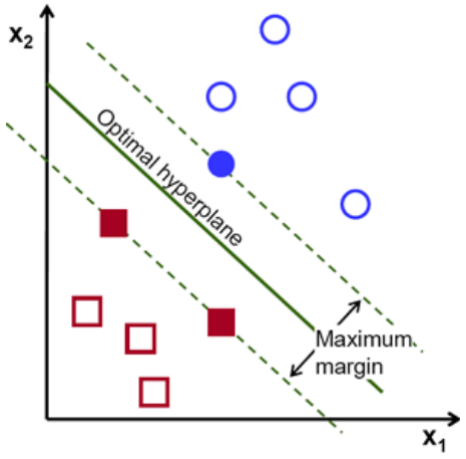


Figure 1.10: *Example of linear SVM classification using a two dimensional information.*

At this point, we have introduced the whole theory necessary to outline a schematic concept map for developing an algorithm for miRNA target prediction. Indeed, ComiR algorithm is based on the following steps:

1. Realize IP experiments (as described in Section 1.2.2)

2. Identify differentially expressed genes that compose the training set (as described in Section 1.3.2)

3. Build a machine learning model using binding scores previously calculated (score calculation has been described in Section 1.1.6)

Following this thread, the development of SVM algorithms will be presented in Chapters 2 and 3. In particular, we propose ComiR algorithm as a reliable tool for miRNA target prediction.
The last part of the present chapter completes the description of the most widespread statistical methods in genomics. Most of them are used for data analysis in the following chapters.

1.3.4 Analysis of complex systems

Many real systems are composed of a vast number of elements that interact among them. For example, human society is a complex system in which individuals experience many types of social interactions. Similarly, a single cell of an organism is a biological system based on the interaction of biomolecules.

Connectivity has a direct effect on the structure and the nature of the system. For this reason, the study of complex systems focuses on how elements interact among them. The analysis of element interactions is closely related to large-scale inference and computational issues; the number of interactions is generally much bigger than the number of elements. For this reason, the study of complex systems requires a flexible approach that allows to manage and analyze a huge amount of data and gives a simple representation of the system.

The network approach is universally used to study complex systems. A *network*, also called *graph* in topology, is a mathematical object that describes the whole connectivity of a system. It is composed of *nodes* and *links* (also called *vertices* and *edges*), whereas nodes correspond to the elements of the system and links represent their interactions.

Albert-László Barabási wrote that networks introduce a new way of thinking about the real-world. Indeed, networks are used in all research fields, including molecular biology. For example, metabolic networks describe chemical reaction among biomolecules [54], neuronal networks describe brain cell connectivity, regulatory networks study gene regulatory relations [55], and protein-protein networks show molecule binding [56].

The considerable success of network methodology can not be justified only by a new perspective in the study of complex systems. The end of the XX century has been characterized by a revolution in data collection, sharing, and storage; the high throughput technologies make big data available to everyone. The new data dimension orientated the scientific research towards innovative methodologies which carry out large scale inference. This framework produced the conditions for the development and popularity of the network approach in data analysis.

Gene network models

Many different mathematical models related to networks have been developed to describe properties and evolution of complex systems over time [87] (e.g., random graph, preferential attachment graph, small-world network). On the other hand, a wide range of probabilistic models has been defined to investigate gene network dependence relations. Those models are considered *probabilistic* because nodes are random variables that can take different values or states, while links connect statistically related variables.

Statistically Validated Networks (SVNs) [53][97] investigate network connectivity through large scale hypothesis testing. This approach has a fundamental advantage; interactions that occur by chance are excluded from the network. Correlation-based networks are a typical example of SVNs. They connect elements whose behavior is correlated. A famous correlated-based network is the *gene co-expression network* [52][53]; in this network, nodes correspond to genes, and links connect genes that have a significant co-expression over a group of conditions.
The construction of SVNs based on correlation requires a statistical test on a correlation coefficient for each couple of nodes. Let consider the nodes v_i and v_j, and the correlation coefficient $\hat{p}_{ij}$ estimated on their records. We are interested in the following hypothesis testing:

$$\begin{cases} H_{0ij} : \rho_{ij} = 0 \\ H_{1ij} : \rho_{ij} \neq 0 \end{cases} \tag{1.4}$$

A non-parametric approach based on the Spearman Correlation coefficient is often used to avoid normal assumptions during hypothesis testing (Quatto *et al.* [93]). All the links in the network come from a significant result of a statistical test. For this reason, we speak about networks that are *statistically validated*. The presence of spurious correlations is the main limitation of gene co-expression networks. Some authors [94] [95] propose a partial correlation-based approach to reduce the spurious correlation in the network.
Gaussian Graphical Models (GGMs)[88] represent an alternative approach conceptually close to SVNs; those models consider the conditional dependence between nodes. Therefore, spurious connections are removed from the network. The main limitation of GGMs is that they require the assumption of multivariate normal distribution, rarely satisfied in gene expression data. In general, graphical models (e.g., gaussian models [88], bayesian models [89], Boolean networks [90][91]) allow to study the relations among genes. For example, *Boolean networks* focus on network dynamics underlying how nodes regulate each other in terms of activation and suppression. In our recent work, we used Boolean gene networks to evaluate the effects of *in silico* mutations in Autism Spectrum Disorders [92]. In our current research, we use a new type of SVN [97] to study miRNA-mRNA bipartite networks. This topic is introduced in the *future research* at the end of the thesis.

1.3.5 Cluster analysis

A common genetic issue is the study of gene expression similarity over different experimental conditions. Clustering methods are used for identifying gene groups with a similar gene expression pattern within them [100].
The clustering problem doesn't have a univocal solution, and many clustering algorithms have been proposed. *K-means* method is the most famous although it depends on an arbitrary choice of the number of clusters. It identifies clusters minimizing the within-cluster sum of squares (i.e., intra-cluster variance).
Another widely used algorithm is the *complete-linkage* [102]; in each step, the algorithm includes an element in its nearest cluster, where the distance from a cluster is equaled to the distance from the farthest cluster's element.
In network theory, a cluster is a group of elements strongly connected among them and sparser connected with other elements of the system. Complex systems naturally organize in clusters; clustering is a common task in network theory. Two common clustering methods are:

- *Modularity Statistic* [103] searches the optimum partition that maximizes the number of links inside clusters and minimizes the number of links between different clusters.

- *Infomap algorithm* [104] operates cluster detection considering the flow induced by the links of a network; clusters consist of nodes among which the flow persists for a long time once entered.

Once gene clusters have been identified, the aspects that characterize those clusters have to be explored. *Gene enrichment analysis* is a common strategy for cluster characterization, as described in the next section.

1.3.6 Gene Ontology Enrichment Analysis

Gene Enrichment Analysis is widely used to identify biological processes, molecular functions, pathways, and cellular components associated with a group of genes. This methodology can be applied to gene groups from any type of analysis, such as network clustering and differential expression analysis. For example, in Chapter 5, we study differentially expressed genes through a Gene Ontology enrichment analysis [216].
The *Gene Ontology* (GO) [106] is a bioinformatics project that classifies genes

in *terms*. Each term represents a gene product property; therefore, genes that belong to a specific term are all involved in a common biological process. Gene Ontology is not the only gene classification proposed in the literature; indeed, in the last decade a large number of gene annotations have been used to perform gene enrichment analysis, e.g., Reactome pathway analysis [107], KEGG Pathway enrichment [108]. GO-Analysis, as many other gene enrichment methods, searches the terms with a significant number of genes in common with the gene group of interest; the significant terms characterize the gene group under exam. Gene Ontology is constantly evolving; indeed, the gene group under exam could enrich the significant terms from GO-analysis.

Significant terms are identified using a hypergeometric test (i.e., Fisher exact test). Consider N_G genes that belong to group G, and N_Q genes that experience the attribute Q, where Q represent a GO-term in our case, or any other gene annotated group. Under the null hypothesis that the attribute Q is uniformly distributed across a large pool of N genes, the probability that $N_{G,Q} = n_{G,Q}$ genes in group G have the attribute Q is:

$$Pr(N_{G,Q} = n_{G,Q}|N, N_G, N_Q) = \frac{\binom{N_G}{n_{G,Q}}\binom{N-N_G}{N_Q-n_{G,Q}}}{\binom{N}{N_Q}} \tag{1.5}$$

So we can associate a p-value with the observed number of $n_{G,Q}$ genes that belong to group G and have the attribute Q:

$$Pr(N_{G,Q} \geq n_{G,Q}) = 1 - \sum_{X=0}^{n_{G,Q}-1} Pr(X|N, N_G, N_Q) \tag{1.6}$$

A significant result of the test means that Q is over-expressed in group G; therefore, Q characterizes this cluster. Whereas the attribute Q is not necessarily the most common feature in the group; the significance means that the relative frequency of Q is significantly higher in group G than the relative frequency observed in the whole pool of genes. Tumminello *et al.* [109] uses the idea behind gene enrichment analysis to characterize clusters of any type of network. Using this approach, GO-terms conceptually correspond to any attribute of interest.

Recently, a new method of *gene network enrichment analysis* has been introduced in the literature; it integrates the information on interactions between genes provided by gene networks into enrichment analyses. This approach tests enrichment between sets of genes in a network [110] [111] [112].

GO-analysis and other enrichment methods allows the simultaneous investiga-

tion of several terms; therefore, it requires a multiple comparison correction, as shown in the next section.

1.3.7 Multiple Comparison Procedures

Most of the methods presented until now are closely related to large scale inference; indeed, a large family of statistical tests is a fundamental part of many bioinformatics methods. For example:

- Differential expression analysis calculates a t-test on each gene expression profile $\longrightarrow$ there are almost 25.000 coding genes.
- GO-ontology is generally applied on all terms containing at least three genes in common with the gene group of interest $\longrightarrow$ the total number of terms is almost 45.000.
- For building statistically validated networks (SVNs), the statistical test used for link validation has to be performed for each couple of nodes; therefore, considering N nodes, the total number of tests is

$$m = \binom{N}{2} = \frac{N(N-1)}{2} \tag{1.7}$$

  In this context, the first type error γ corresponds to the probability of wrongly draw a link in the network. On the other hand, the second type error β corresponds to the probability of miss a link that should be drowned.

The high number of statistical hypotheses is connected with a multiple testing problem. There are plenty of methods that allow controlling the whole error of the entire procedure [113]. In this section, the most popular multiple comparison procedures (MCPs) are introduced.
Bonferroni method is the most conservative correction for multiple comparisons. It controls the probability of rejecting at least one null hypothesis true; this probability is called *Family Wise Error Rate* (FWER).
Consider a group of null hypothesis $H_{01}, H_{02}, ..., H_{0m}$, and a given level α, such as:

$$FWER \leq \alpha \tag{1.8}$$

The constrain 1.8 is guaranteed fixing the first type error γ of every single test equals to:

$$\gamma = \frac{\alpha}{m} \tag{1.9}$$

hypothesis	not-rejected	rejected	total
true	U	F	m_0
false	Z	S	m_1
	W	R	m

Table 1.2: *Possible result of a multiple testing procedure.*

Therefore, the hypothesis H_{0i} is rejected if its p-value p_i is smaller than γ.
The Boole inequality demonstrates the validity of the constrain 1.8 for any number m_0 of true hypothesis:

$$FWER = Pr\left\{\bigcup_{I_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{I_0} Pr\left\{p_i \leq \frac{\alpha}{m}\right\} = m_0\frac{\alpha}{m} \leq \alpha \tag{1.10}$$

where, I_0 is the set of m_0 true hypothesis.
Moreover, at the price of an increase of conservativeness, the constrain is also guaranteed in case of dependency of hypotheses. In general, Bonferroni procedure is very conservative and could produce too few rejections.
Benjamini and Hochberg introduced a more powerful approach to multiple testing; it changes the error definition introducing the *False Discovery Rate* (FDR)[114] as a measurement of the whole error of the procedure.
Consider the results of a multiple testing procedure as shown in Tab. 1.2, the FDR is defined as the expected proportion of wrong rejections

$$FDR = \mathbb{E}\left[\frac{F}{R}\right] \tag{1.11}$$

1.11 is not defined for $R = 0$. Moreover, is not possible control the FDR if $m = m_0$. For these reasons, Benjamini and Hochberg proposed an alternative definition of the FDR:

$$FDR = \mathbb{E}\left[\frac{F}{R}|R > 0\right] Pr\{R > 0\} \tag{1.12}$$

Given a family of hypothesis $H_{01}, H_{02}, ..., H_{0m}$, the control of the FDR at the level q is obtained through the following steps:

1. Order the observed p-values in the vector $\{p_{(1)}, p_{(2)}, .., p_{(m)}\}$.

2. Find the index i_{max} such as

$$i_{max} = \max\left\{i : \ p_{(i)} \leq \frac{i}{m}q\right\}$$

3. the i^{th} p-value $p_{(i)}$ is considered significant if

$$i \leq i_{max}$$

Therefore, this procedure identifies the p-values $p_{(1)}, p_{(2)}, .., p_{(i_{max})}$ as significant guaranteeing the constraint $FDR \leq q$.
The presence of hypothesis dependencies requires an alternative calculation of the i_{max} index [115]. But it makes the procedure more conservative:

$$i_{max} = \max \left\{ i : \; p_{(i)} \leq \frac{i}{m} \frac{q}{h_i} \right\} \tag{1.13}$$

$$\text{with,} \qquad h_i = \sum_{j=1}^{i} \frac{1}{j}$$

Bradley Efron proposes an empirical Bayes approach to multiple comparisons [116]; it increases the statistical power of the testing procedure through a Bayes estimation of the FDR. This approach to multiple comparisons is used in the significant analysis of microarray implemented by the R package *samr*. Moreover, Quatto *et al.* [93] propose a new method for SVN construction based on the Efron approach that can be used for very sparse networks. In practice, we recommend the comparison of different correction approaches; the Bonferroni network identifies the strongest connections Instead FDR-network gives an overall view of the whole connectivity of the system.

Chapter 2

Analysis of miRNA Interactions:

RIP-Chip analysis supports different roles for AGO2 and GW182 proteins in recruiting and processing microRNA targets

The analysis presented in this chapter has been published on BMC Bioinformatics by the name *RIP-Chip analysis supports different roles for AGO2 and GW182 proteins in recruiting and processing microRNA targets* [3].
We analyzed the activities of two RISC proteins, AGO2 and GW182. The results highlight important information that explain miRNA binding behavior. In particular, mRNA coding region information significantly improve the performance of miRNA target prediction algorithms. Those results outline the theoretical framework for improving ComiR algorithm (as presented in Chapter 3).

Abstract

We performed three RIP-Chip experiments using either anti-AGO2 or anti-GW182 antibodies and compiled a data set made up of the miRNA and mRNA expression profiles of three samples for each experiment. Specifically, we analyzed the input sample, the immunoprecipitated fraction and the unbound sample resulting from the RIP experiment. We used the expression profile of the input sample to compute several variables, using formulae capable of integrating the information on miRNA binding sites, both in the 3'UTR and coding regions, with miRNA and mRNA expression level profiles. We com-

pared immunoprecipitated vs unbound samples to determine the enriched or underrepresented genes in the immunoprecipitated fractions, independently for AGO2 and GW182 related samples.
For each of the two proteins, we trained and tested several support vector machine algorithms capable of distinguishing the enriched from the underrepresented genes that were experimentally detected. The most efficient algorithm for distinguishing the enriched genes in AGO2 immunoprecipitated samples was trained by using variables involving the number of binding sites in both the 3'UTR and coding region, integrated with the miRNA expression profile, as expected for miRNA targets. On the other hand, we found that the best variable for distinguishing the enriched genes in the GW182 immunoprecipitated samples was the length of the coding region.
Due to the major role of GW182 in GW/P-bodies, our data suggest that the AGO2-GW182 RISC recruits genes based on miRNA binding sites in the 3'UTR and coding region, but only the longer mRNAs probably remain sequestered in GW/P-bodies, functioning as a repository for translationally silenced RNAs.

2.1 Role of RISC in microRNA binding

Argonaute (AGO) proteins and the GW182 protein family (also known as TNRC6 proteins) are involved in the cellular process which leads to gene silencing mediated by miRNAs, small endogenous non-coding RNAs that act as post-transcriptional regulators by base pairing to target mRNAs [117][118]. While miRNAs guide AGOs to target mRNAs, a direct interaction between AGO and GW182 proteins is required for the assembly of ribonucleoprotein complexes, named RISCs, and the recruitment of additional factors involved in gene silencing, which is ultimately achieved through the degradation of target mRNAs or translational repression [119][120]. Several studies of higher eukaryotes have indicated that, among the AGO proteins, AGO2 is catalytically active and involved in the mRNA cleavage process, whereas AGO1, 3 and 4 are catalytically inactive and mainly involved in translational repression [120][121]. In the cell cytoplasm, AGOs, together with GW182/TNRC6A and its mammalian paralogs, TNRC6B and TNRC6C, have a role in executing miRNA-mediated repression, either by silencing or decay, but the proteins also contribute to other functions in the nucleus, such as transcription and splicing

control [122][123]. On the other hand, GW182 is a marker of GW/P-bodies, dynamic cytoplasmic structures containing non-translating mRNAs, that have been associated with the cellular response to stress [124] and were first identified because human autoimmune sera recognized them [125][67]. Work over the past few years has significantly increased our understanding of the biology of GW/P-bodies in higher and lower eukaryotes. It has been shown that these bodies contain proteins involved in diverse post-transcriptional processes, such as mRNA degradation, nonsense-mediated mRNA decay, translational repression, RNA-mediated gene silencing, and may also function as a cytoplasmic domain for RNA storage.

Furthermore, RNA-binding protein immunoprecipitation, coupled with high throughput methods for expression profiling, such as gene array (RIP-Chip) or sequencing (RIP-Seq), has allowed the systematic identification of RISC-bound miRNAs and their target mRNA sequences in mammalian cells and the dissection of miRNA-mediated post-transcriptional regulatory networks. This approach has been widely applied to the AGO protein family, through the immunoprecipitation of either exogenously introduced tagged-proteins or endogenous proteins and the subsequent analysis of the associated RNAs [67][68][69][70]. So far, few reports have described a similar approach for GW182 and its paralogs using specific antibodies [71][72], and recently, Meister and co-workers reported a novel method, based on affinity purification, for the simultaneous isolation of all AGO-containing complexes [73].

The RIP-based high throughput method for expression profiling has been widely used to predict miRNA-target interactions in order to develop algorithms useful for identifying potential miRNA targets.

In order to get additional insight into the diverse cellular functions of RISCs, we performed RIP-Chip experiments using antibodies specific for AGO2 and GW182/TNRC6A. Data collection methods are described in the appendix at the end of this chapter.

Data from miRNA and mRNA expression profiles were combined, using existing target prediction results, to compute several variables that served to train and test various support vector machine (SVM) algorithms, searching for the more efficient variables for distinguishing enriched genes in the immunoprecipitated samples.

2.2 *In silico* prediction of microRNA-mRNA interactions

All the 3'UTR and coding sequences used to predict miRNA binding sites were selected from Ensembl.org. If the database contained more than one sequence for the same Ensembl ID, the longest sequence was selected. We only considered sequences at least 50 bases long. From Ensembl.org we selected 18,552 3'UTR and 19,420 coding sequences, of which 16,363 mRNAs were included in both sets and in the microarray platform used. MicroRNA binding sites were predicted using TargetScan [45], PITA [49] and miRanda [48] scripts. We computed two miRNA-mRNA interaction matrices (BS), one for 3'UTR and one for the coding regions, which contained the number of binding sites predicted for each miRNA seed on the selected sequences. For both BS matrices, we computed the respective density matrices (dBS) by dividing the number of predicted binding sites by the length of the considered sequence.
The expression profile of endogenous miRNAs has been shown to be determinant in predicting RISC machinery functional targets, and it is used by ComiR [43] to predict targets of a set of miRNAs. In addition to such collaborative effects, competition effects have a crucial role in miRNA regulatory function, as shown by the evidence of competing exogenous [127] and endogenous [128] effects. In summary, both miRNA and mRNA expression profiles have a crucial role in determining miRNA binding activity. For this reason we have used miRNA expression profiles to transform the primary scores of PITA, miRanda and TargetScan.

2.3 RIP-Chip Analysis

Microarray data pre-processing consisted of the following pipeline. The Feature Extraction Software already provided background subtracted, dye normalized and spatially detrended processed signal intensities. Intensities were normalized using the quantile normalization technique. First of all, an average linkage cluster analysis was performed in order to check instrumental replicate consistency, and then the average expression profile of instrumental replicates was computed. The obtained expression profiles were used to perform a post-hoc power analysis specific for microarray studies [131], and we obtained an observed power of 0.7, which implied that 70% of truly enriched genes were

expected to be discovered.
The pre-processed expression profiles were compared through hierarchical cluster analysis (average linkage), where distance was computed as dist = 1 – correlation. Genes enriched and underrepresented in immunoprecipitated (IP) samples were identified using the Significance Analysis of Microarrays (SAM) algorithm [65], implemented by the samr library in BioconductoR. The samr library associates a q-value with each gene, i.e., the lowest False Discovery Rate at which that gene is called significant. It is like the well-known p-value, but adapted to multiple-testing situations. A q-value of 5% was set as the threshold for significance in detecting enriched and underrepresented genes. Enriched genes detected by the SAM algorithm were compared with the enriched genes detected by REA [132], an algorithm developed specifically for RIP-Chip enrichment analysis.

2.3.1 AGO2 and GW182 proteins complexes handle different mRNA content

To gain new insight into the regulatory networks of gene expression involving functionally diverse RISCs in the cell cytoplasm, we used RIP-Chip to identify mRNAs and miRNAs selectively bound to these complexes in the MCF-7 cell line, which is widely used and representative of luminal breast cancer. We selected AGO2 and GW182 antibodies against core RISC proteins since AGO2 is the most abundantly expressed AGO protein in many cell types, including MCF-7 cells [134], and GW182/ TNRC6A has been shown to be the major binding partner for AGO2 [135]. We performed three independent RIP experiment, collecting the input (IN), immunoprecipitated (IP) and flow throught (FT) samples.
The efficiency of the AGO2 and GW182 antibodies in IPs was confirmed by the enrichment of both proteins in the IP fractions and their depletion in the FT fractions, while the lack of precipitation of either AGO2 or GW182 protein by control IgG confirmed the specificity of antibodies (Fig. 2.1a). We also examined, in AGO2-IP and GW182-IP, the enrichment of seven miRNAs highly expressed in the MCF7 cell line [70]. As shown in Fig.2.2a, all the analyzed miRNAs were significantly enriched by AGO2 and GW182-IP compared to controls (p-value$<$ 0.05, AGO2 or GW182-IP vs IgG-IP). As expected for

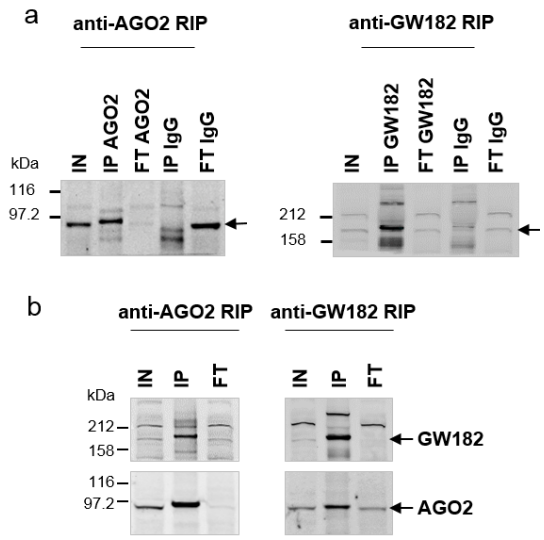


Figure 2.1: *Western Blot analysis of proteins immunoprecipitated and co-immunoprecipitated with anti-AGO2 or anti-GW182 antibody (IP). IgGs in* ***a****) are the negative controls. IN and FT made up 1% of the cytoplasmic lysate used for each IP sample. GW182 was specifically co-immunoprecipitated with AGO2 (****b****, left panel), and AGO2 was specifically co-immunoprecipitated with GW182 (****b****, right panel)*

proteins present in the same complex, Western Blot analysis confirmed the reciprocal co-immunoprecipitation of AGO2 and GW182 (Fig.2.1b). Whole genome and miRNA expression profiles, as determined by microarray analysis, gave rise to a novel dataset that is available through the NCBI GEO database (accession IDGSE109667). As shown in Fig.2.2b, the cluster analysis performed on whole genome expression profiles revealed that the mRNA expression profiles of the AGO2-IP samples (blue cluster) were homogeneous and different from the GW182-IP mRNA expression profiles (red cluster). The miRNA expression profile clustering showed only one homogenous cluster, the AGO2-IP sample cluster (Fig.2.2b, blue cluster). The comparison of AGO2-IP vs IN expression profiles revealed the underrepresentation, in the IP sample, of several miRNAs highly expressed in IN samples, a fact that implies a lower correlation between IP and IN expression profiles (see Additional file 1). On the other hand, GW182-IP and IN miRNA expression profiles were more similar to each other, and such behavior explains the absence of a GW-IP cluster in miRNA expression profile clustering.

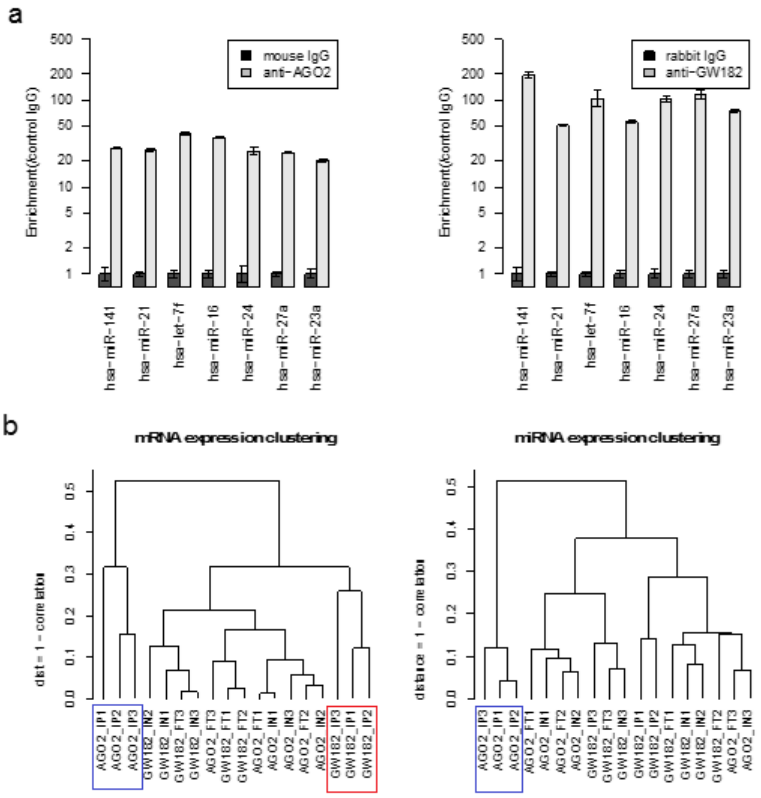


Figure 2.2: ***a)*** *Enrichment analysis of seven highly expressed miRNAs in anti-AGO2 and anti-GW182 IP compared to IgG-IP controls.* ***b)*** *Average Linkage Cluster analysis of mRNA and miRNA expression profiles of IP, IN and FT samples from three independent experiments; distance is computed as 1- Correlation (Pearson). AGO2-IP and GW182-IP mRNA expression profiles are highlighted in blue and green, respectively. In mRNA expression clustering, we considered all the 16,323 genes with a detected expression level in the samples considered. In miRNA expression clustering, we considered 508 miRNAs with a detected expression level in at least one sample.*

We also characterized the two proteins' behavior by detecting the enriched genes in AGO2-IP and GW182-IP. We observed that the most efficient comparison in retrieving miRNA targets was the one between IP vs FT, with respect to IP vs IN samples. Indeed, GSEA analysis showed more miRNA predicted targets in IP vs FT enriched genes than in the IP vs IN comparison. We first noticed that the intersection between the two sets of enriched genes in AGO2 and GW182-IP showed a poor, yet significant, overlap.
Our list of enriched genes in the AGO2 IP vs FT comparison showed a sta-

tistically significant overlap with the published list of 616 enriched genes for AGO2-IP in MCF-7 cells [70]. Unfortunately, no high throughput analysis results are yet publicly available for any anti-GW182 antibody, which makes it impossible to perform a similar comparison for enriched genes in GW182-IP. The two sets of enriched/underrepresented genes, named UP/LOW_AGO2 and UP/LOW_GW182, were used, in the analysis described below, to select the features capable of distinguishing the mRNA associated with the AGO2 and GW182 proteins, respectively.

Variable name	Formula	BS
F1	$\Sigma_i expr(miRNA_i) \times BS_{ij} \times expr(mRNA_j)$	number in 3'UTR
F2	$\Sigma_i expr(miRNA_i) \times BS_{ij}$	number in 3'UTR
F3	$\Sigma_i BS_{ij} \times expr(mRNA_j)$	number in 3'UTR
F4	$\Sigma_i BS_{ij}$	number in 3'UTR
F1d	$\Sigma_i expr(miRNA_i) \times dBS_{ij} \times expr(mRNA_j)$	density in 3'UTR
F2d	$\Sigma_i expr(miRNA_i) \times dBS_{ij}$	density in 3'UTR
F3d	$\Sigma_i dBS_{ij} \times expr(mRNA_j)$	density in 3'UTR
F4d	$\Sigma_i dBS_{ij}$	density in 3'UTR
F5	$\Sigma_i expr(miRNA_i) \times BS_{ij} \times expr(mRNA_j)$	number in CDS
F6	$\Sigma_i expr(miRNA_i) \times BS_{ij}$	number in CDS
F7	$\Sigma_i BS_{ij} \times expr(mRNA_j)$	number in CDS
F8	$\Sigma_i BS_{ij}$	number in CDS
F5d	$\Sigma_i expr(miRNA_i) \times dBS_{ij} \times expr(mRNA_j)$	density in CDS
F6d	$\Sigma_i expr(miRNA_i) \times dBS_{ij}$	density in CDS
F7d	$\Sigma_i dBS_{ij} \times expr(mRNA_j)$	density in CDS
F8d	$\Sigma_i dBS_{ij}$	density in CDS
F9	$expr(mRNA_j)$	Not applicable
L1	length of 3'UTR	Not applicable
L2	length of coding region	Not applicable

Table 2.1: *Definition of variables used to model miRNA activity. The column BS provides details about the miRNA predicted binding sites used to compute BS_{ij} (the binding sites matrix). For each variable, the Formula defines the values associated to each $mRNA_j$. CDS=coding region sequence*

2.3.2 Expression-based variables used for characterizing enriched genes in IP samples

To have better insight into the roles of the GW182 and AGO2 proteins in miRNA regulatory activity, and with the aim of selecting the most useful variables for distinguishing between enriched and underrepresented genes in IP samples, we tested formulas including mRNA and miRNA expression levels in IN samples and miRNA predicted binding sites on 3'UTR and coding regions of mRNAs. Specifically, we considered 19 variables, all computed by using features characterizing the mRNA sequences and IN sample gene expression. Table 2.1 describes all the considered variables. The defined variables display high correlations among each other, as shown in the correlation matrix reported in Fig.2.3, where variables are specifically computed for the AGO2_IN1 sample. Analogous results were obtained when using the expression profile information of other IN samples. Three main clusters of highly correlated variables were clearly visible, one that contains all the variables included in the formula for the mRNA expression profile, and the other two that relate to the presence of miRNA binding sites in the coding region and 3'UTR.

2.3.3 Enriched and underrepresented genes in anti-AGO2 RIP are efficiently distinguished by miRNA binding sites in mRNA coding regions weighted by miRNA expression

We first tested the performance of each of the 19 variables to distinguish the enriched genes (UP) in AGO2-IP vs FT from the underrepresented (LOW) genes. We computed the variables by using the expression profiles from each individual anti-AGO2 RIP experiment and performed a ROC analysis and a Wilcoxon test, using the UP/LOW genes detected comparing AGO2-IP vs FT as a reference set. Figure 2.4a and b show the obtained AUC values and the Wilcoxon-test p-values, both used as an estimation of performance in distinguishing UP genes from LOW genes.
We have also verified that Targetscan predictions have the best performance in distinguishing the enriched genes. Thus, we decided to use it in any further analyses to compute BS matrices. It was evident that the features belonging to the cluster related to the coding region length were the most efficient.

Indeed, F6 and F8 variables were the best variables for distinguishing between enriched and underrepresented genes in anti-AGO2 RIP samples. F8 counts the number of binding sites in the coding region of the mRNA, while the number of binding sites is weighted by the miRNA expression values in F6. Both F6 and F8 variables are highly correlated with the L2 variable, which could have been anticipated, since the longer the coding region is, the higher the number of binding sites detected in the region by any binding site prediction algorithm. Fig.3.4 clearly shows that F6, F8, and L2 variables assume lower values for LOW_AGO2 genes with respect to all genes.
On the other hand, the variable with the next highest performance, not belonging to the L2 cluster, was the F4d variable. Fig.3.4 shows that F4d assumes higher values for UP_AGO2 with respect to all genes. The behavior of F4d promised to be synergistic with F6 in distinguishing UP and LOW genes, and, therefore, we further discuss it in a separate section.

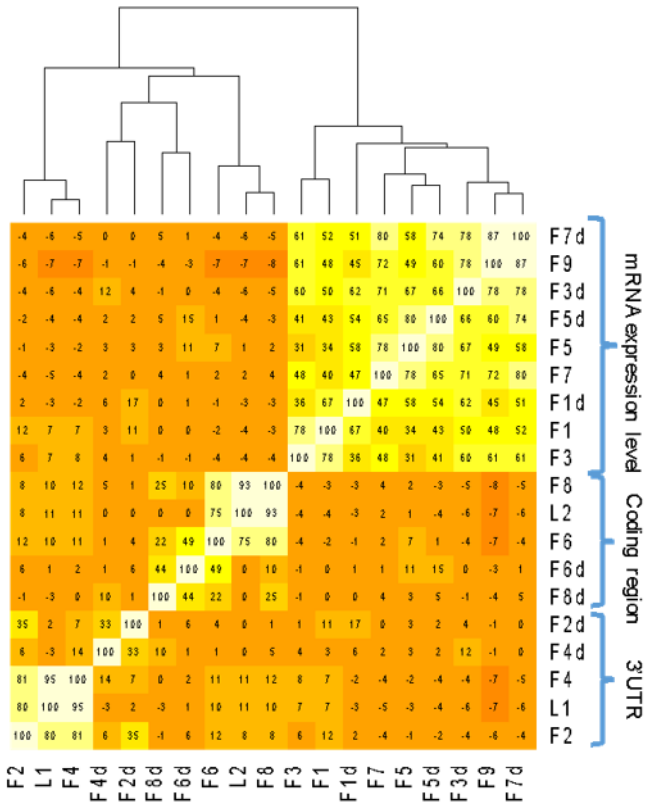


Figure 2.3: *Correlation matrix of variables listed in Table 2.1. Heatmap representation of the correlation block matrix of the variables computed with AGO2_IN1 miRNA and mRNA expression profiles. The reported numbers are the correlation values, expressed in the range [−100:100].*

Next, we verified that the high performance of variables F6 and F8 was specifically due to the effects of the miRNA expression profile in the formula. Specifically, we considered 1000 simulated miRNA expression profiles, as obtained by assigning the original expression profile to 50 random miRNAs, chosen from among all the miRNAs expressed in the sample, and 1000 simulated miRNA expression profiles, as obtained by shuffling the original 50 miRNAs found to be highly expressed (top 50 expressed). The first block of simulations was less conservative, and its aim was to test whether the identity of the top 50 expressed miRNAs was determinant for reaching the original performance; it was the only block of simulations meaningful for testing the performance of

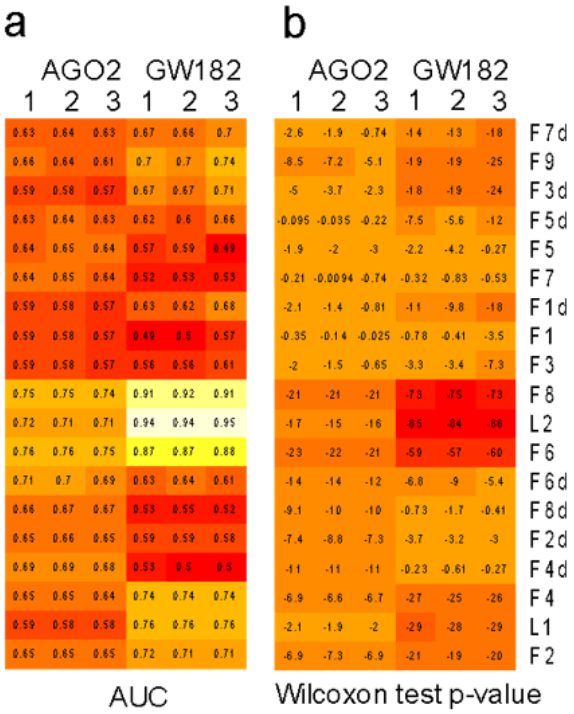


Figure 2.4: *Prediction capacity of variables listed in Table 2.1.* ***a****) ROC-AUC values obtained by classifying enriched/underrepresented genes associated with the variables computed with each IN expression profile.* ***b****) Wilcoxon test p-values (log10) obtained by comparing the variable values associated with the enriched/underrepresented gene sets. In both* ***a****) and* ***b****), the variables computed with the three AGO2 IN profiles were used to distinguish enriched and underrepresented genes in AGO2-IP vs FT. The variables computed with the three GW182 IN profiles were used to distinguish enriched and underrepresented genes in GW182-IP vs FT.*

the F8 variable. The second block of simulations was more conservative, and its aim was to assess whether the specific expression profile associated with the top 50 miRNAs was determinant. In both cases, the performance of the simulated F6 and F8 variables was significantly lower than the F6 and F8 variables obtained by including the original miRNA expression profile (see Fig.2.5a). We also tested simulations that were more conservative by holding the expression profile of the highly expressed miRNAs fixed while shuffling the expression of the remaining ones. Fig.2.5a shows the results of these simulations obtained by fixing up to five top expressed miRNAs. As the number of the top expressed miRNAs increased, the F6 variable performance became closer to that obtained with the original miRNA expression profile; in addition, the higher the number of miRNAs fixed, the closer it got to the original performance level. As a result, we concluded that the miRNA expression profile is crucial for distinguishing AGO2-associated miRNA targets, especially the expression profile of the first top expressed miRNAs, and that the most relevant miRNA binding sites are the ones found in the coding region.

2.3.4 Enriched and underrepresented genes in anti - GW182 RIP are efficiently distinguished by coding region length

The performance of each of the 19 variables was tested to distinguish between the enriched genes in GW182-IP vs FT and the underrepresented ones. Fig.2.4 a and b show that the features belonging to the cluster related to the coding region length are the most efficient at distinguishing between enriched genes in anti-GW182 RIP samples. In this case, the best feature for distinguishing the enriched genes in GW182-IP samples was the coding region length of the mRNA, i.e., the L2 variable, with a surprisingly very high performance (average AUC > 0.9). The average AUC associated with the F6 variable was also very high (average AUC $= 0.87$); however, the miRNA expression profile was not crucial for reaching such high performance since a shuffled expression profile was not significantly deficient in distinguishing the enriched genes (Fig.2.5b). In Fig. 2.6, we compare the ECDF of the coding region length of the UP and LOW genes in the anti-GW182 RIP experiments. The separation between UP and LOW genes in anti-GW182 RIP samples is evident in the coding region length values, though less in the 3'UTR length values.

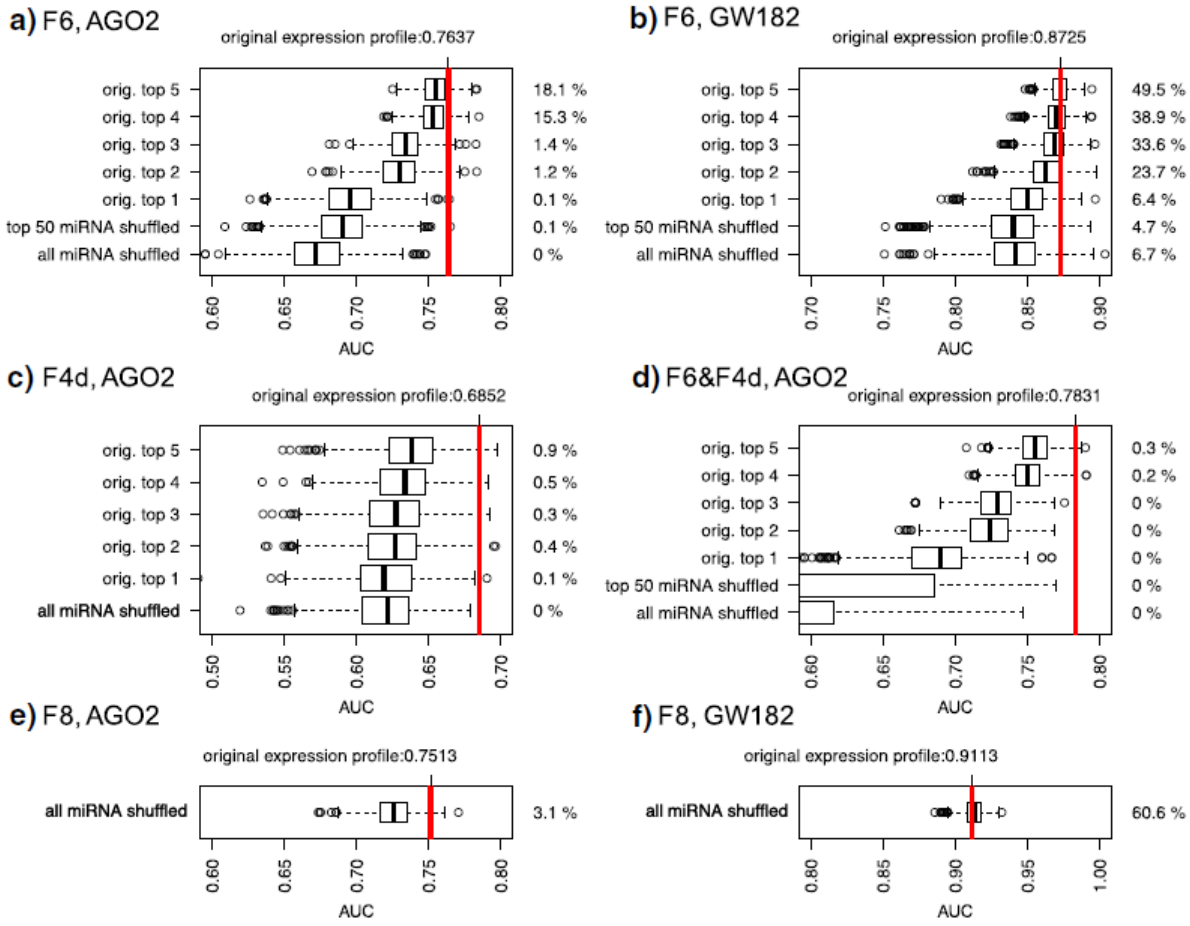


Figure 2.5: *Graphic representation of the effect of miRNA expression profile shuffling. Each boxplot represents the AUC values obtained with 1000 simulated miRNA expression profiles. The percentage on the right of each boxplot refers to the number of times an AUC value was greater than the AUC obtained with the original miRNA expression profile (red vertical line).* ***a****) Performance of simulated F6 variables in distinguishing AGO2 enriched/underrepresented genes.* ***b****) Performance of simulated F6 variables in distinguishing GW182 enriched/underrepresented genes.* ***c****) Performance of simulated F4d variables in distinguishing AGO2 enriched/underrepresented genes.* ***d****) Performance of simulated SVM models (F6 & F4d variables) in distinguishing AGO2 enriched/underrepresented genes.* ***e****) Performance of simulated F8 variables in distinguishing AGO2 enriched/underrepresented genes.* ***f****) Performance of simulated F8 variables in distinguishing GW182 enriched/underrepresented genes.*

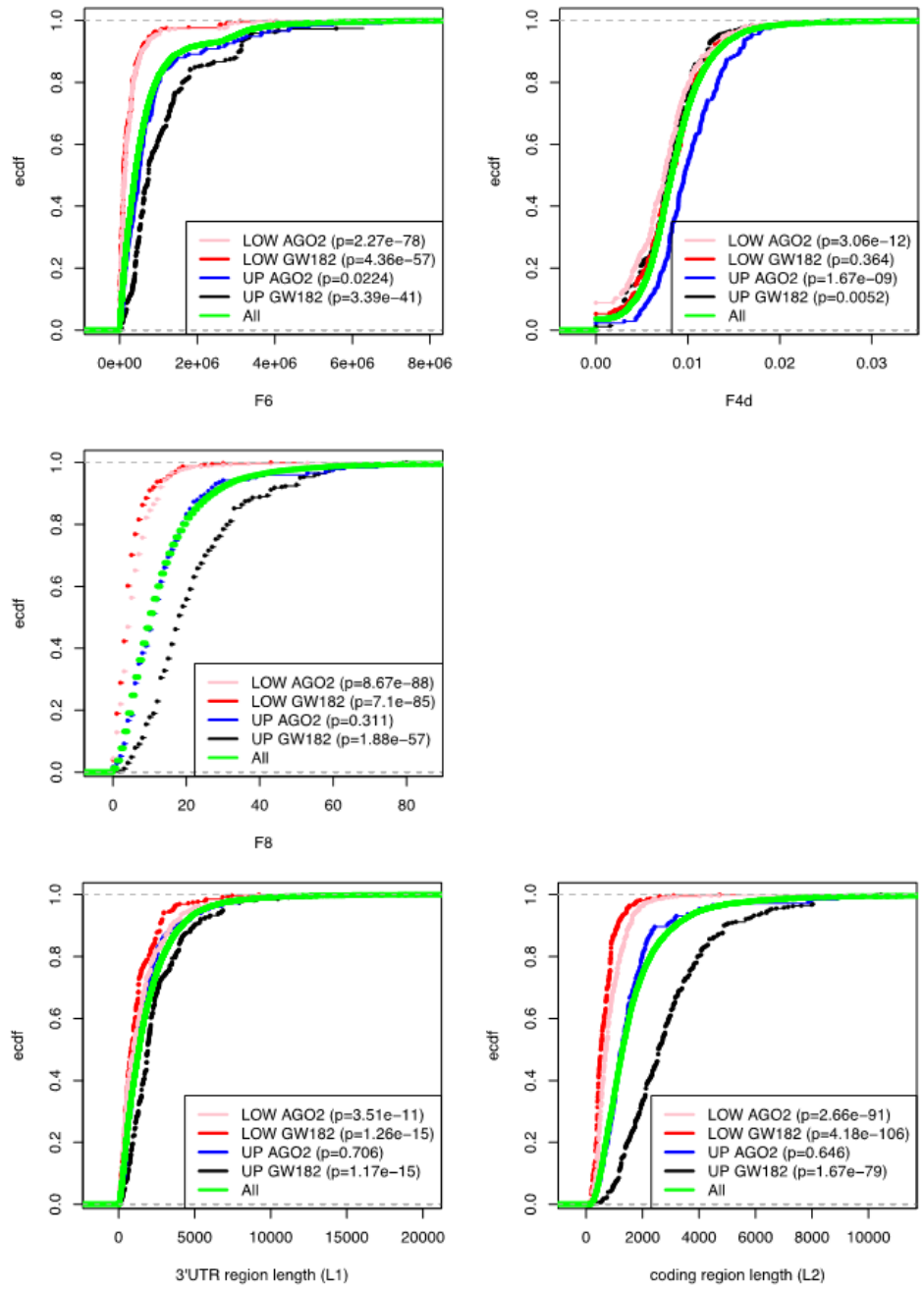


Figure 2.6: *Graphic representation of selected features values associated to enriched and underrepresented genes. Empirical cumulative distribution function (ECDF) of F6, F4d, F8, L1 and L2 variables computed for enriched (UP) and underrepresented (LOW) genes in AGO2 IP vs FT and GW182 IP vs FT analyses. The reported p-values were obtained by performing a Wilcoxon-test comparing the values assumed by the selected set of genes with the values assumed by all the genes (16,363, green lines).*

Wilcoxon tests were performed to compare the 3'UTR and coding region length of GW182_UP and DOWN genes with all gene lengths, and gave highly significant p-values. Anti-GW182 RIP gene expression profiles, which could be used to support our hypothesis that the mRNA coding region length is a relevant feature for GW182 activity, are not available, and none of the enriched group of genes reported in the literature regards breast cancer cells. Nevertheless, we considered the IP-enrichment results of 7820 genes published by Landthaler and collaborators [71], where the authors generated HEK293 cell lines stably expressing epitopetagged human AGO and GW proteins and used such cells to detect enriched mRNA in miRNA-containing ribonucleoprotein particles through a microarray analysis. They found a high overlap among the enriched targets of the AGO and GW182 family proteins by analyzing the top immunoprecipitated transcripts associated with the four AGO proteins vs the ones associated with the three GW182 proteins. Differently from [71], we considered the non-overlapping enriched genes, and we found that the mRNAs enriched only in GW182-IP had significantly longer 3'UTR and coding regions.

2.3.5 SVM models improve performance in distinguishing enriched genes

We tested whether a combination of two variables could significantly improve the classification of the performance of enriched/underrepresented genes. An SVM algorithm model[1] was trained with each pair of features, and the cross-validation AUC results for each pair are reported in Fig.2.7. The best performance in predicting AGO2-bound mRNAs was associated with the F6-F4d variable pair, with an AUC significantly higher than the one obtained with F6 only (AUC = 0.78; DeLong's test p-value < 0.05). The F4d variable takes into account the density of the binding sites in the 3'UTR, as predicted for the top 50 expressed miRNAs. The F4d variable performance by itself (AUC = 0.68) is the highest among the features not highly correlated with the F6 variable.

[1] SVM models were trained with linear kernel using the e1071 R library. The R library caret was used to test the SVM trained models with the Leave One Out Cross Validation (train-Control method = "LOOCV") testing procedure method "svmLinear2"). The next Chapter presents a deeper analysis of the prediction capacity of the SVM model; in addition to the LOOCV procedure, we have evaluated the model prediction capacity on two external test sets free from possible over-fitting problems.

We checked whether the identity of the top 50 expressed miRNAs was crucial for reaching such a performance by randomly changing the identity of the 50 miRNAs in the F4d formula, and holding the expression of an increasing number of top miRNAs fixed. The results are plotted in Fig.2.5c, they show that, when using randomly chosen miRNAs, the performance is significantly lower than the one obtained with the true top 50 expressed miRNAs. Differently from what was obtained for the F6 variable, to reach the performance obtained with the original miRNA expression profile, the expression of almost

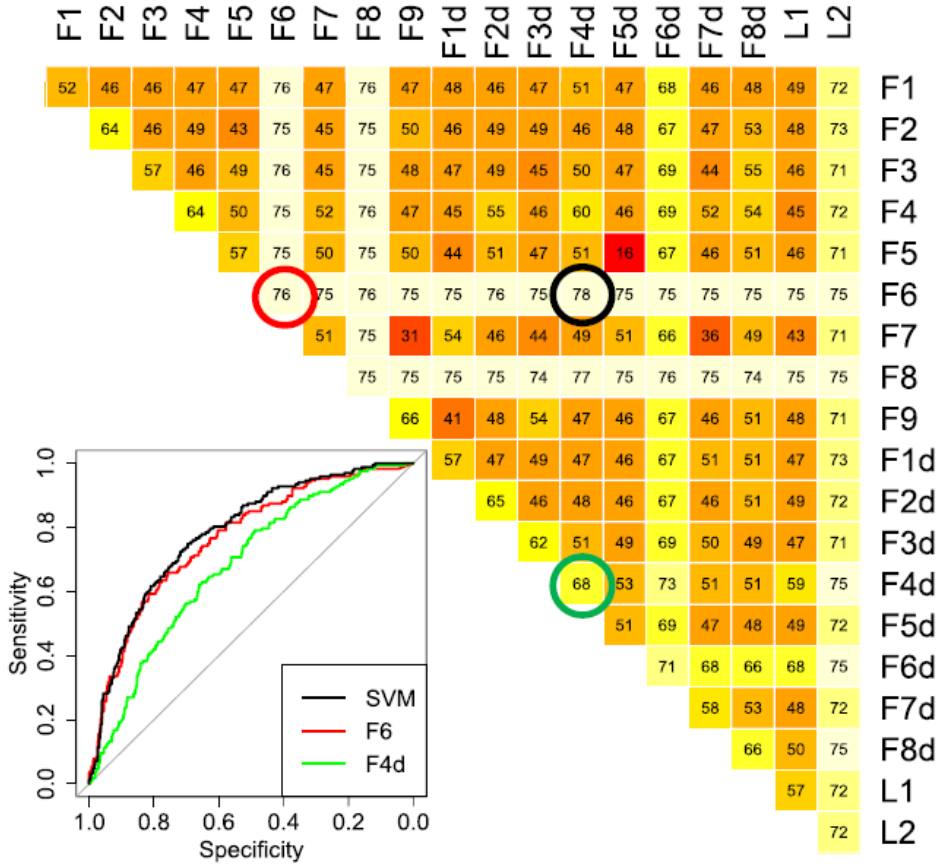


Figure 2.7: *Support Vector Machine models performance summary. AUC values of SVM models trained with any pair of variables defined in Table 1, used to classify enriched/underrepresented genes in AGO2-IP vs FT comparison. Variables were computed by using the AGO2_IN1 expression profiles. Values are in the range [0:100]. Values in the diagonal refer to single variable performance. The ROC plot at bottom left represents the results obtained with the best-performing SVM model (F6 and F4d, black line) and with the two single variables, F6 (red line) and F4d (green line).*

all the miRNAs had to be held, meaning that the identity of the top 50 miRNAs is substantially important to the F4d variable's performance.
Analogous simulations were done for the predictions obtained with the SVM model trained with the F6 and F4d variables (Fig.2.5d).
The results show that several miRNAs had to be fixed in order to reach a performance similar to that obtained with the original miRNA expression profile.
Finally, we tested how slightly different expression profiles, such as the ones obtained by experiment replica, may affect enriched/underrepresented gene classification. Specifically, we used an SVM model trained with features computed with miRNA expression profiles from one IN sample to classify genes with higher vs lower IP/FT ratio, computed in each of the three experiments. Our results show that higher performance was always obtained when predictions of IP/FT ratio values in one experiment were obtained with the miRNA expression profile belonging to the IN sample expression profile of the same experiment.
The pair of variables that best predicted the GW182-bound mRNAs was the L1 and L2 pair, i.e., the length of the 3'UTR and the coding region, respectively, but the improvement in the AUC value was not statistically significant (DeLong's test p-value > 0.05).

2.4 Discussion

We analyzed the activity of two endogenous interacting proteins, AGO2 and GW182, in MCF-7 cell cytoplasm. Both are involved in RISCs, and we analyzed the RNA co-immunoprecipitated with the selected proteins, which was expected to be enriched in genes involved in endogenous miRNA regulatory activity. Data from RIP-Chip experiments served to model miRNA activity by assigning variables based on miRNA expression profiles to each mRNA target, searching for the ones that would better distinguish the enriched genes in RIP samples. We expected that the detected variables could reveal which information was relevant for modeling miRNA activity and the RISC proteins' roles.
Our results show that mRNAs co-immunoprecipitated with the two proteins have different characteristics. Such a finding might appear in contrast with a previous analysis performed in HEK293 cell lines, in which tagged-AGO2 or tagged-GW182/TNRC6A proteins we re stably overexpressed and the AGO

protein family and the GW182 protein family were found to be associated with highly similar sets of transcripts [71]. The low consistency with this previous study might indicate a different composition of RISCs in MCF-7 cells than HEK293 cells. Moreover, analysis under physiological conditions vs overexpressed AGO or GW182 might also explain the differences, and the fact that the authors analyzed the top immunoprecipitated transcripts for the whole AGO family (AGO1–4) vs the GW182 family (TNRC6A-C) might have mitigated RNA enrichment differences with respect to what we obtained through the comparison of two specific proteins, i.e., AGO2 and TNRC6A. Indeed, it has been reported that AGO1 and AGO2 proteins interact with a distinct set of miRNAs [129] and, as a consequence, with different mRNA targets, whereas the GW182/TNRC6A protein interacts with the whole AGO protein family [118]. This evidence also justifies the high similarity we found between the miRNA expression profiles of GW182-IP and FT, in contrast with more specific miRNA expression profiles associated with the AGO2-IP and FT samples (Fig.2.2b). Furthermore, although a high degree of redundancy among the members of each protein family has been reported, it cannot be excluded that the use of different GW182 antibodies and/or slightly different experimental conditions, e.g., buffer stringency, might result in a different enrichment of RNAs in the immunoprecipitated samples. To this end, a systematic analysis of the data obtained using the same antibody in the same cell background, or the use of methods based on biochemical approaches, like the one described by Hauptmann and coworkers [73], might definitively clear up this point.

We found that the mRNAs co-immunoprecipitated with the AGO2 protein can be distinguished from the underrepresented mRNAs by considering the number of miRNA binding sites in the coding region, weighted by miRNA expression level. In order to improve the classification performance, we also trained an SVM with two features at a time, and we found that the additional feature to be considered was the density of the binding sites predicted in the 3'UTR of mRNA. We then performed simulations by shuffling the miRNA expression profiles in order to detect which miRNAs are relevant to composing the features used to distinguish enriched and underrepresented genes. When the performance obtained by randomly shuffling a set of miRNAs is significantly lower than the performance obtained with the original miRNA expression profile, we can assess that the set of miRNAs replaced is relevant in the classification. Results show that the only relevant miRNAs, when considering

binding sites in the mRNA coding regions, are the top two to three of those expressed. On the contrary, almost all of the top 50 expressed miRNAs are relevant when considering the binding sites in the 3'UTR of mRNA, with a prominent exception being the top expressed one, i.e., hsa-miR-21-5p. The expression level detected for hsamiR- 21-5p is very high, by itself covering 60% of the total miRNA expression profile, and we suppose that its distinctive behavior is related to saturation effects in miRNA activity, which we plan to investigate in further studies.

In addition to simulated miRNA expression profiles, we tested how switching miRNA expression profiles across our experimental replicates affects the performance of the classification algorithm. We found that even slight differences in the expression profiles of the single replicate IN samples gave rise to differences in enriched vs underrepresented gene classification, leading to the conclusion that the combination of mRNA and miRNA expression profiles from the same experiment gives the best performance.

On the other hand, we clearly observed that the mRNA co-immunoprecipitated with the GW182 protein was highly enriched with genes with longer coding regions. In this case, enriched/underrepresented gene classification does not depend on the miRNA expression profile, but only on 3'UTR and coding region lengths. We confirmed this result by analyzing the data from Landthaler and coworkers [71]. Our interpretation is that GW182 complexes preferentially sequester the longer mRNAs in the process of populating GW/P-bodies.

While functionally diverse RISCs lacking GW182 have been described [136], the interaction between mRNAs and GW182 is reported to be mediated by the miRNA and AGO proteins and, so far, no direct interaction has been demonstrated between GW182 and mRNA. Recently, Elkayam and coauthors [137] showed that, differently from AGO proteins, which have a single GW182-binding site, GW182 can recruit up to three copies of AGO proteins via its three distinct GW motifs. We believe that such a feature supports our results, since the longer the mRNA is, the higher the number of miRNA binding sites and the probability that RNA -loaded AGO proteins would find cooperative binding sites within the right distance to interact with the same GW182 protein. In this case, the model of single binding sites weighted by miRNA expression profile is probably oversimplified, and further analysis is required to include collaboration effects. To our knowledge, the involvement of mRNA length in GW182 recruitment is a novel observation that may contribute to

shedding light on the different activities of the AGO2 and GW182 proteins in various RISCs and/or in diverse cellular districts such as GW/P-bodies.

2.5 Conclusions

In this work, we aimed to unravel RISC activity by analyzing a novel RIP-Chip data set obtained by the immunoprecipitation of two RISC proteins, AGO2 and GW 182. We analyzed the overexpressed genes in the anti-AGO2 and anti-GW182 RIP samples vs the respective FT samples, and we revealed different features characterizing the enriched genes in the two data sets. AGO2-associated mRNAs are characterized by a high number of binding sites in the coding region for top expressed miRNAs and by a high density of binding sites in the 3'UTR region. On the other hand, GW182-asso ciated mRNAs are characterized by long coding regions. These different characteristics may underline the different roles played by the selected proteins in the RISC machinery activity. Our data confirm that the anti-AGO2 RIP gives an accurate picture of which RNA is involved in miRNA regulatory activity. Regarding the anti-GW182 RIP, data show no significant involvement of miRNA expression profiles in GW182-associated mRNA selection, at least within a simplified model of single binding sites weighted by miRNA expression profile. Our results support the hypothesis that, after being recruited by the miRNA machinery, only the mRNAs with longer coding regions are destined to be stored in GW/P bodies, while shorter mRNAs are most likely processed in different ways that lead to degradation rather than storage.

Chapter 3

MicroRNA Target Prediction:

An improvement of ComiR algorithm by exploiting coding region sequences of mRNAs

This chapter focuses on the prediction of miRNA targets. In particular, ComiR algorithm [44] is presented as a reliable tool for miRNA target prediction, and an improvement of ComiR algorithm is proposed by considering miRNA binding sites located on messenger RNA coding regions.
The importance of the coding region during miRNA binding activity has been demonstrated in our previous work [3] (described in Chapter 2). Starting from those results, we include coding region information in ComiR algorithm.
We find that ComiR algorithm trained with coding region information is more efficient in predicting the microRNA targets with respect to the algorithm trained with 3'UTR information. On the other hand, we show that 3'UTR based predictions can be seen as complementary to the coding region based predictions, which suggests that both predictions, from 3'UTR and coding regions, should be considered in a comprehensive analysis. The following results have been published on BMC Bioinformatics as "*An improvement of ComiR algorithm for microRNA target prediction by exploiting coding region sequences of mRNAs*" [4].

3.1 Background

MicroRNA genes (miRNAs) are small non-coding RNAs that post - transcriptionally regulate the expression level of messenger RNAs (mRNAs). MicroRNAs are critical in many important biological processes, and are important

markers for many diseases. miRNA regulation activity depends on the recognition of binding sites located on messenger RNA molecules (mRNAs), in this context mRNAs represent the *targets* of miRNA binding.
The ability of predicting miRNA targets is crucial to understand the processes they are involved in. MicroRNA regulation activity depends on the recognition of binding sites located on mRNA molecules. MicroRNA-mRNA interaction is mediated by a family of ribonucleoprotein complexes called *RNA-induced silencing complexes* (RISCs)[34]. The immunoprecipitation of RISC proteins is an experimental strategy used to investigate on miRNA targets [35][36].
The high costs of experiments oriented the miRNA target identification towards a computational approach; miRNA target prediction algorithms are generally based on Watson-Crick base-pair matching [37] [38] [39]. Perfect complementarity between miRNA-mRNA pairs is quite rare, but also a six base-pair match could be sufficient to suppress gene expression.
Few other methods use the miRNA expression profile as additional information, namely, GenMir++[40], PicTar [41], Talasso [42].
ComiR (Combinatorial miRNA targeting) [43][44] is a user friendly web tool realized to predict the targets of a set of microRNAs, starting from their expression profile. ComiR algorithm incorporates miRNA expression in a thermodynamic binding model, and it associates each gene with the score of being a target of a set of miRNAs.
ComiR was trained with the information regarding binding sites in the 3'UTR region. The miRNA targets identification has been mainly based on the search of mRNA binding sites contained in the 3'UTR region [138]. It is also known that miRNAs bind the coding region [139], in a previous work [3] we have showed that the coding region plays a role in distinguishing RISC machinery targets. Therefore, the information contained in the coding region can't be ignored for the miRNA target prediction.
Fig. 3.1 reports the number of outcomes of four queries to PUBCHEM and ISI Web of Science repositories, namely, the number of papers associated with the joint queries 1) "miRNA target prediction" "3'UTR"; 2) "miRNA target prediction" "coding region"; 3) "miRNA binding" "3'UTR"; and 4) "miRNA binding" "coding region". It's worth noting that, despite continuous evidences of the presence of binding sites in the mRNA coding region, the incidence of the word "3'UTR" is steadily one order of magnitude higher than the one of "coding region". Indeed, we hypothesize that words "miRNA" and "3'UTR" have

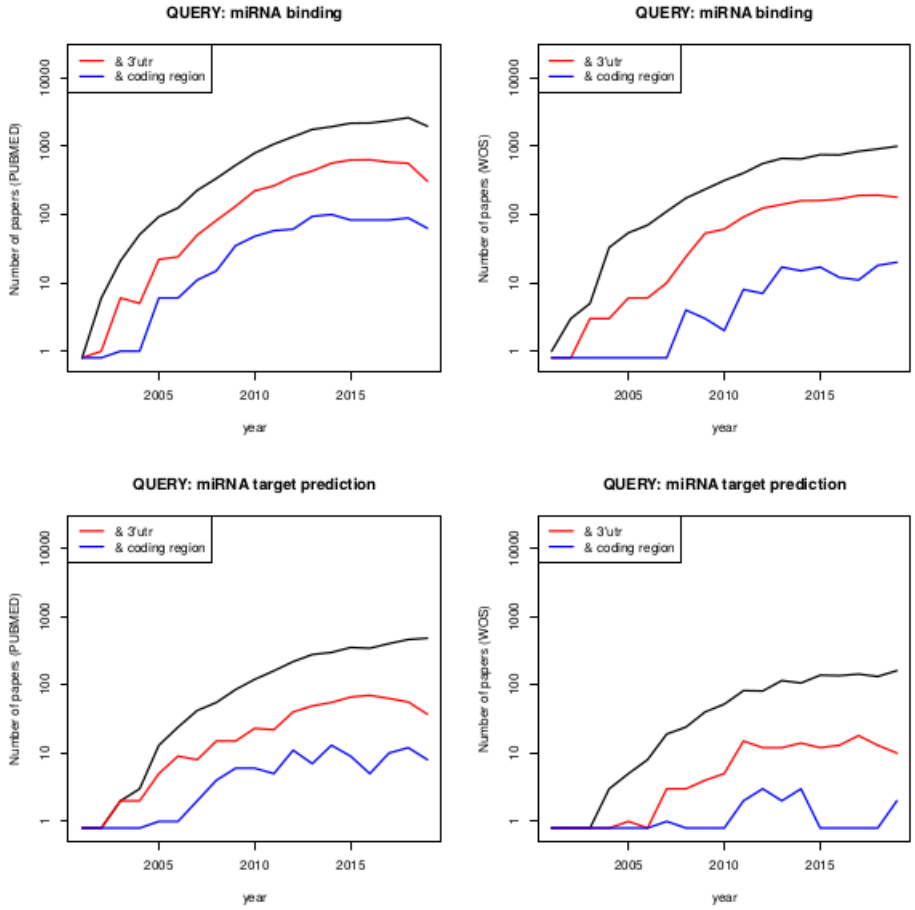


Figure 3.1: *Quantification of scientific production regarding miRNA topics, updated to Jan 2020. Left panels concern queries to the PUBMED repository, while right panels concern queries to the ISI Web of Science repository. Black lines indicate the temporal evolution of the number of papers found through the main query, which is indicated in the title of each panel. Red and blue lines indicate the temporal evolution of the number of papers found by combining the main query with the words "3'UTR" and "coding region", respectively.*

been linked together since the discovery of microRNAs [141] [142], whereas the association between "miRNA" and "coding region" is less explored, the focus of the actual version of ComiR on binding sites in the 3'UTR only is a typical example.

In this study, we propose to upgrade the ComiR algorithm, by introducing information about the binding sites contained in the coding region of the genes. We show that the information contained in the coding region significantly improves the accuracy of ComiR predictions.

3.2 ComiR algorithm

ComiR is a user friendly web tool described in [44]. The user has to provide a list of miRNAs and their expression levels. The output is a ranked vector of scores; therefore, each gene is associated with a reliability of being a target of the set of miRNAs given in input.
The original version of ComiR contains a Support Vector Machine (SVM) based algorithm that incorporates the miRNA target prediction results of four individual tools (i.e., PITA [49], miRanda [48] and TargetScan [45] and miRSVR [143]) in 3'UTR. Due to a break in maintenance of mirSVR scores, in this work, we will only consider the PITA, miRanda and TargetScan predictions.
In this work we include the coding region binding sites in ComiR algorithm, therefore we downloaded the 3'UTR and coding region sequences of genes in D. melanogaster species from Ensembl/bioMart (release BDGP6.22). The whole set of 469 mature miRNA sequences was downloaded from miRBase (release 22). For each miRNA, we applied PITA, miRanda and TargetScan algorithms, in order to detect the binding sites in both the 3'UTR and the coding region of each gene. These algorithms associate a score for each couple miRNA-mRNA, these scores have been integrated with the opportune miRNA expression profiles before running the SVM.

3.2.1 Incorporation of miRNA expression levels

The primary scores of PITA and miRanda are transformed using a thermodynamic binding model based on Fermi-Dirac equation, this allows to take into account the miRNA expression in the score calculation.
This model has been introduced by Coronnello *et al.* [43], it takes into account binding affinity and miRNA expression. Suppose that miRNA i (miR_i) has n_{ik} binding sites BS_{ijk} ($j = 1, ..., n_{ik}$ on mRNA k. The reversible reaction of binding between miRNA$_i$ and BS_{ijk} is:

$$miR_i + BS_{ijk} \rightleftharpoons miR_i : BS_{ijk}$$

The equilibrium binding constant of this reaction is:

$$K_i = \frac{[miR_i : BS_{ijk}]}{[miR_i][BS_{ijk}]} \tag{3.1}$$

Considering the Fermi-Dirac equation, the probability of binding is:

$$\begin{aligned} Pr(miR_i : BS_{ijk}) &= \frac{[miR_i : BS_{ijk}]}{[miR_i : BS_{ijk}]+[BS_{ijk}]} \\ &= \frac{1}{1+\frac{1}{K_i[miR_i]}} = \frac{1}{1+e^{(E_{ijk}-\mu)/RT}} \end{aligned} \tag{3.2}$$

where $E_{ijk} = -RT\,ln(K_i)$ is the standard free energy of binding and $\mu_i = RT\,log([miRNA_i])$.
The concentration $[miRNA_i]$ can be imputed using $miRNA_i$ expression, and the energy E_{ijk} corresponds to the score given by miRanda or PITA.
The total score of binding between $miRNA_i$ and $mRNA_k$ considers all the n_{ik} binding sites on $mRNA_k$:

$$S_{ik} = \sum_{j}^{n_{ij}} Pr(miR_i : BS_{ijk}) \tag{3.3}$$

We are interested in the score of a single gene respect a set of miRNAs, therefore the Fermi-Dirac (FD) score of $mRNA_k$ respect a group of N miRNAs is:

$$\begin{aligned} \text{FD score:}\quad S_k &= \textstyle\sum_i^N \sum_j^{n_{ij}} Pr(miR_i : BS_{ijk}) \\ &= \textstyle\sum_i^N \sum_j^{n_{ij}} \left\{1 + e^{(E_{ijk}-\mu_i)/RT}\right\}^{-1} \end{aligned} \tag{3.4}$$

In the case of TargetScan, a simple weighted sum has been calculated:

$$\text{WS score:}\quad S_k = \sum_{i}^{N} [miR_i]\, T_{ik} \tag{3.5}$$

where T_{ik} is the number of binding sites predicted by TargetScan for the miRNA$_i$:mRNA$_k$ pair.
This work is focused on *Drosophila melanogaster* (Dme) miRNA target prediction. We have considered 28 miRNAs that have at least 50 reads in the S2 cells [144]. The expression of those miRNAs have been used to compute Fermi-Dirac and Weighed-Sum scores, so the SVM has been run on these scores.

3.2.2 SVM Training Dataset

The training set and the testing set were obtained by comparing the results from two different experiments that regard the RISC protein Ago1: depletion of Ago1 [145] and Ago1 immunoprecipitation (IP) [144]. The comparison of these two experiments give four sets of genes:

- set I - 152 genes enriched in AGO1 IP and upregulated after AGO1 depletion
- set II - 1039 genes enriched in AGO1 IP and not upregulated after AGO1 depletion
- set III - 300 genes not enriched in AGO1 IP and upregulated in AGO1 depletion
- set IV - 5509 genes not enriched in AGO1 IP and not upregulated in AGO1 depletion.

We only considered Dme genes with annotated both the 3'UTR and coding region. Consequently, our final dataset was composed by 139 genes in set I, 929 genes in set II, 253 genes in set III and 4738 genes in set IV.
Similarly to the original version of ComiR, the SVM has been trained with set I (139 genes) as positive set and the 139 most highly expressed genes from set IV as negative set (named top-set-IV). While set III has been used for testing.

		Immunoprecipitation of Ago1	
		IP enriched	Not enriched
Ago1 depletion	Up-regulated	**Positive**	Set III
	Not up-regulated	Set II	**Negative**

Table 3.1: *Experiments that identifies positive and negative validated genes of training and testing set.*

3.3 Statistical Analysis

To evaluate which part of the gene sequence produce the best prediction accuracy, we have implemented a SVM on three combinations of different subsets of relevant variables: 1) PITA, miRanda and TargetScan scores on 3'UTR region; 2) PITA, miRanda and TargetScan scores on coding region; 3) all of the variables considered in points 1 and 2.
The performance of each SVM combination has been evaluated by implementing leave-one-out Cross-Validation (LOOCV) procedure (one by one, each gene is left out from the training set at each step of the procedure) Fig.3.2A compares the ROC curves obtained from a LOOCV analysis. The SVM trained on the coding region features has a higher predictions capacity than the SVM on 3'UTR region features (coding vs 3'UTR, De long test [146] p-value = 0.0005). On the other hand, the joint use of 3'UTR and coding region information doesn't significantly improve the performance (coding vs coding+3'UTR; p-value = 0.31).

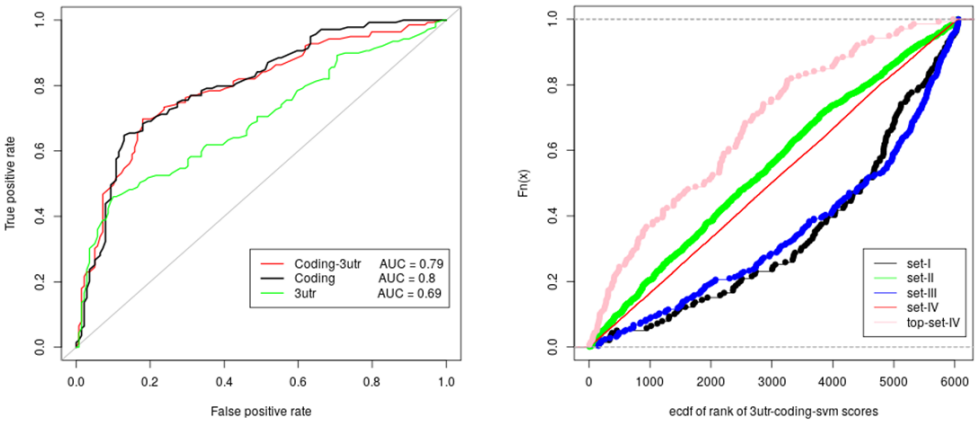
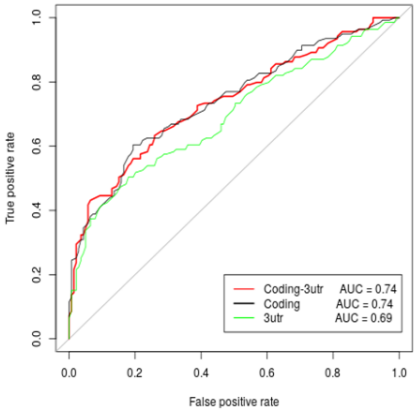
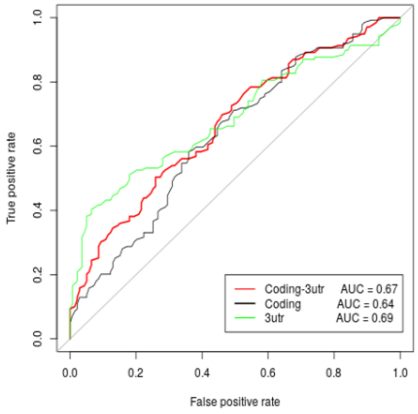


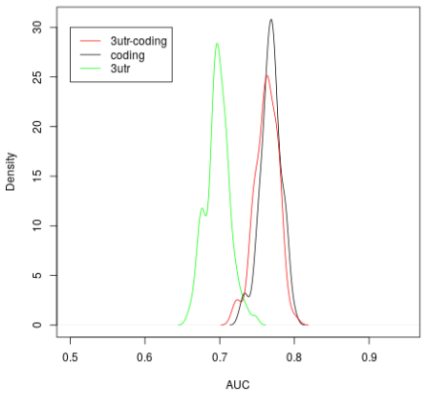
Figure 3.2: *Overview of SVM prediction outcome. The SVM is trained with set I as positive set and top-set-IV as negative set. A) shown ROC curves are the result of a LOOCV analysis. PITA, miRanda and Targetscan scores related to 3'UTR (green line), coding region (black line) and both (red line) are user to train the SVM. B) ECDF of the rank of ComiR scores obtained for the genes of set I (black), set II (green), set III (blue), set IV (red) and top-set-IV (pink).*



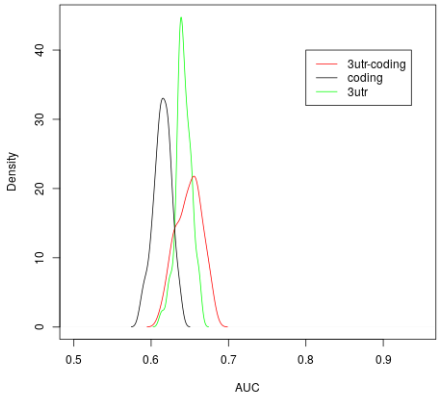
A



B



C



D

Figure 3.3: *Overview SVM performance when using set I as training positive set and set III as positive testing set and vice versa. A) ROC analysis results obtained by using set I and top-set-IV as training set and one example of ran-set-III and top2-set-IV as test set; B) ROC analysis results obtained by using one example of ran-set-III and top2-set-IV as training set and set I and top-set-IV as test set; C) AUC values distribution of 100 ROC analysis as described in 3.3A associated with different ran-set-III sampling. D) AUC values distribution of 100 ROC analysis as described in 3.3B associated with different ran-set-III sampling.*

In Fig.3.2B we compare the empirical cumulative distribution functions (ECDF) of the rank of ComiR scores obtained for the genes in the four sets of the dataset with the coding+3'UTR model.
Similar results are obtained for the predictions obtained with the 3'UTR only model and the coding region only model. We observe that both set I and set III show significantly higher ComiR scores than the whole dataset scores (Wilcoxon test p-value = 10e-12 and 10e-18 respectively). On the contrary, set II doesn't show significantly higher ComiR scores than the whole dataset. To further explore the behaviour of the sets I and III, and the performance of the SVM, we performed ROC analyses by alternatively using the two sets as training and testing set. Specifically, to obtain comparable AUC values, we randomly selected 139 genes from set III (ran-set-III) and the 139 most highly expressed genes (named top2-set-IV) after the first 139 included in the top-set-IV set.
Fig.3.3A shows the ROC analysis results obtained by using set I and top-set-IV as training set and one of the ran-set-III and top2-set-IV as test set. The described training and testing set were then switched and the ROC analysis results are shown in Fig.3.3B. We performed 100 of such tests, each time by randomly selecting a different ran-set-III set, and the distribution of the obtained AUC values is reported in Fig.3.3C-D. In this case, we keep obtaining

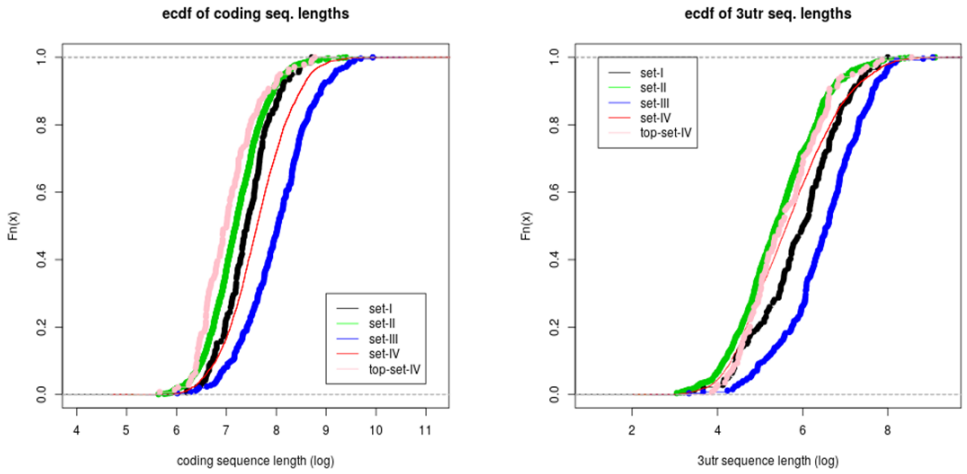
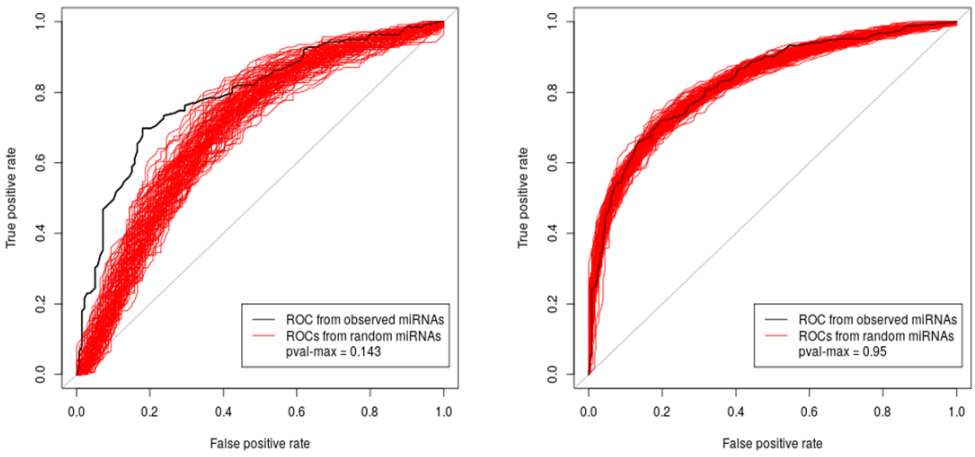


Figure 3.4: *Overview of gene sequences lengths. A) ECDF of 3'UTR sequence lengths, B) ECDF of coding region sequence lengths in the analyzed sets of genes.*

acceptable AUC values, in the range [0.6-0.8].

The lower AUC values, as compared to Fig3.2A, are due to the fact that the training and testing sets are selected from two different pools of genes and it is evident that a better efficiency is obtained when set I, instead than set III, is used as positive training set.
Fig.3.4 shows the ECDF of the sequence lengths of the analyzed sets. Genes of set III have significantly higher 3'UTR and coding region lengths. Due to the additive calculation of the scores used to feed the SVM, it is expected that the length of the sequence plays a role in distinguishing the targets.
To detect whether the SVM predictions are significantly dependent by the used miRNA expression profile, we performed a set of 100 LOOCV tests, each one performed by using a simulated miRNA expression profile to compute the training dataset. Specifically, each simulated miRNA expression profile was obtained by associating the original 28 expression values with a set of 28 randomly selected miRNAs (among the 469 Dme miRNAs).
Fig.3.5A shows the ROC analysis results obtained with the simulated profiles (red lines) in comparison with the original profile (black line). It is evident that the performance in predicting the targets is significantly higher when the



A **B**

Figure 3.5: *Overview of SVM performance with simulated miRNA expression profiles. The black line is associated to LOOCV test result obtained with the original miRNA expression profile, the red lined with the simulated profiles. In A) we used set I as positive training set, in B) we used set III.*

scores used to train the SVM are computed with the original miRNA expression profile. This effect is less evident when the set III is used as positive set (Fig.3.5B), probably due to the fact that set III is strongly characterized by long RNA sequences and this feature is predominant in the training.
Training the SVM with both the 3'UTR and coding region information doesn't produce an improvement in the prediction efficiency. Fig.3.6 shows the scatter plot of the 3'UTR-based predictions rank vs the coding region-based predictions rank of the positive and negative sets. It seems that the two SVM models trained with 3'UTR or coding region information separately, prioritize differently the genes. Moreover the importance of 3'UTR has been proved in literature, therefore we decide to include 3'UTR information in ComiR algorithm.

3.4 Discussion

The presence of miRNA binding sites in the coding region of the genes has been already described in the scientific literature [147], although it is less explored than the association of miRNAs with the 3'UTR. As mentioned already, the current version of ComiR only considers the binding sites predicted within the 3'UTR untranslated region. To fill such a gap of information, we decided to use the binding sites predicted in the coding region. Coding regions are significantly longer than the 3'UTR, and the computational effort needed to predict their binding sites is probably one of the reasons why target-prediction tools are not extensively applied to them. The main objective of the paper is therefore to test whether adding the binding sites on the coding regions improves the miRNA target prediction.
If we compare the old version of ComiR results [43] and the results obtained here by using the 3'UTR region only model, we noticed a significant drop in the performance of the upgraded version with respect to the first version of ComiR. We attributed this drop to the missing use of the mirSVR predictions and to the whole upgrade to the current release of the 3'UTR sequences used to run the used miRNA target prediction tools, which changed significantly the efficiency in predictions of each single tool. Nevertheless, it is desirable to ensure the maintenance of the algorithm by upgrading the predictions database with the most recent sequences releases. Our results show that, focusing on the results obtained with the current sequences releases, the ComiR algorithm

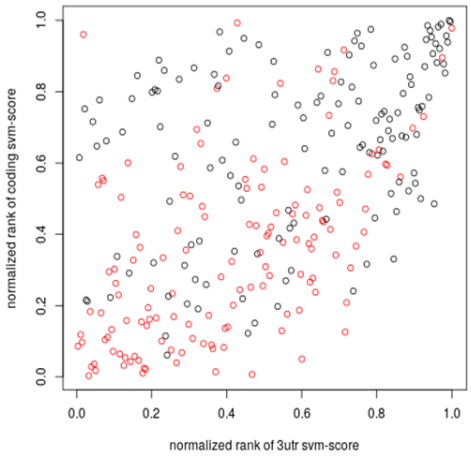


Figure 3.6: *Scatter plot of SVM scores obtained with coding region based model vs 3'UTR based model. The SVM is trained with set I as positive set and top-set-IV as negative set. Black points refers to the positive set, red points to the negative set.*

is significantly improved by considering the binding sites predicted in the coding region, outperforming the efficiency obtained by the algorithm when using only the 3'UTR binding sites. We observe that combining the information of both 3'UTR and coding region binding sites in the SVM model doesn't improve the performance of the prediction algorithm. This result is not due to a redundancy in 3'UTR and coding region information. In fact, using the information carried by the binding sites presence in 3'UTR and coding region separately leads to the prediction of different sets of genes, both showing a significant enrichment of the positive training set. Our conclusion is that both the trained SVMs should be utilized to obtain a complete vision of the target prediction, and further analysis will be conducted to unravel the peculiarities of the two different predicted sets. Our results suggest that ComiR scores prioritize the targets that are functionally degraded (set I and set III), while genes that are co-immunoprecipitated with the RISC protein AGO1 are not significantly predicted (set II). In addition, training the SVM with set II as positive set, generates a SVM model that doesn't predict efficiently the set I training set (data not shown). On the other hand, set III genes show significantly longer 3'UTR and coding region lengths, and this peculiarity could be the main reason for its good performance as positive set. We confirm to consider the set I as the most trustable positive set, because these genes are

confirmed by two independent experimental approaches, whereas set II and III contain genes that have been detected by only one experimental approach each. The asymmetry in the response and the characteristics of these two sets of genes lead to the observation that both the experimental approaches, i.e., the RISC machinery proteins inhibition and immunoprecipitation, should be applied to detect a valid miRNA target set.

3.5 Conclusion

Our results indicate that binding sites predicted in the genes coding region are valuable information in order to efficiently predict the functional targets of a set of miRNAs by their integration in the ComiR algorithm framework. We currently aim at finding the best way to combine the two scores obtained by training the SVM with the 3'UTR and the coding region separately. Further analysis will be conducted to analyze data from other species, by using positive and negative set of miRNA targets obtained through the comparison of results from both RISC proteins inhibition and immunoprecipitation.

Chapter 4

SARS-Cov-2 Sequence Analysis:

miR-1207-5p can contribute to dysregulation of inflammatory response in COVID-19 via targeting SARS-CoV-2 RNA

COVID-19 represents the first worldwide pandemic in a globalized world characterized by a considerable impact on healthcare systems and mortality [200]. The short time since the outbreak began is the reason why many aspects of the molecular interactions of SARS-CoV-2 in the human host are still unknown, especially its mechanisms on a transcriptional level. The role of host endogenous miRNAs in the propagation of viruses is a discussed theme, which leads to the uncovering of many complex virus-specific mechanisms, not yet fully understood.

The analysis described in this chapter is an enrichment of the paper "*miR-1207-5p can contribute to dysregulation of inflammatory response in COVID-19 via targeting SARS-CoV-2 RNA*" published in *frontiers in Cellular and Infection Microbiology* [5]. We have carried out an extensive analysis of human miRNA binding sites on the viral genome highlighting the role of human miRNAs in SARS-CoV-2 infection. It is known that exogenous RNA can compete for miRNA targets of endogenous mRNAs leading to their overexpression [201] . MicroRNA binding sites have been identified by comparing predictions of five algorithms. In particular, the novel ComiR version (presented in the previous chapter) produces more specific miRNA predictions than other algorithms.

Our results suggest that the SARS-CoV-2 virus can act as an exogenous competing RNA, facilitating the over-expression of its endogenous targets. Transcriptomic analysis of human alveolar and bronchial epithelial cells confirmed that the CSF1 gene, a known target of miR-1207-5p, is over-expressed following SARS-CoV-2 infection. CSF1 enhances macrophage recruitment and activation, and its overexpression may contribute to the acute inflammatory response observed in severe COVID-19. In summary, our results indicate that dysregulation of miR-1207-5p-target genes during SARS-CoV-2 infection may contribute to uncontrolled inflammation in most severe COVID-19 cases.
An introductory section on SARS-CoV-2 mechanisms and pathogenesis precedes the analysis.

4.1 Mechanism of SARS-CoV-2 infection

The SARS-CoV-2 virus is composed of a single-positive RNA strand contained in a glycoprotein membrane. Its "life" cycle consists of the following five steps: attachment, penetration, biosynthesis, maturation, and release (Fig.4.1).
At first the virus binds to host receptors (*attachment*), it enters host cells through endocytosis or membrane fusion (*penetration*).
Ones inside the host cell, the RNA(+) strand is used to make the enzyme RNA polymerase and is replicated to RNA(-), whereas the replication process is activated by TMPRSS2 protease.
The RNA(-) is used to make subgenetic mRNAs by transcribing and to make more RNA(+) by replication. Subgenetic mRNA is used to make viral proteins (*biosynthesis*). Then, new viral particles are moved to the endoplasmic reticulum-Golgi intermediate compartment to be assembled to compose a new virus (*maturation*). Finally, the new viruses are released through esocytosis.
Four structural proteins form coronaviruses; Spike (S), membrane (M), envelop (E), and nucleocapsid (N) [148]. Spike is a glycoprotein located on the capsid surface; this protein characterizes the virus's external appearance. Several analyses showed that the spike proteins bind the *angiotensin converting enzyme-2* (ACE2), so ACE2 is identified as a functional receptor for SARS-CoV [149]. Therefore, the spike protein has a crucial role in recognizing and binding the host [167]. Spike is composed of two functional subunits; the S1 subunit is responsible for binding to the host cell receptor, and the S2 subunit is for the fusion of the viral and cellular membranes [150].

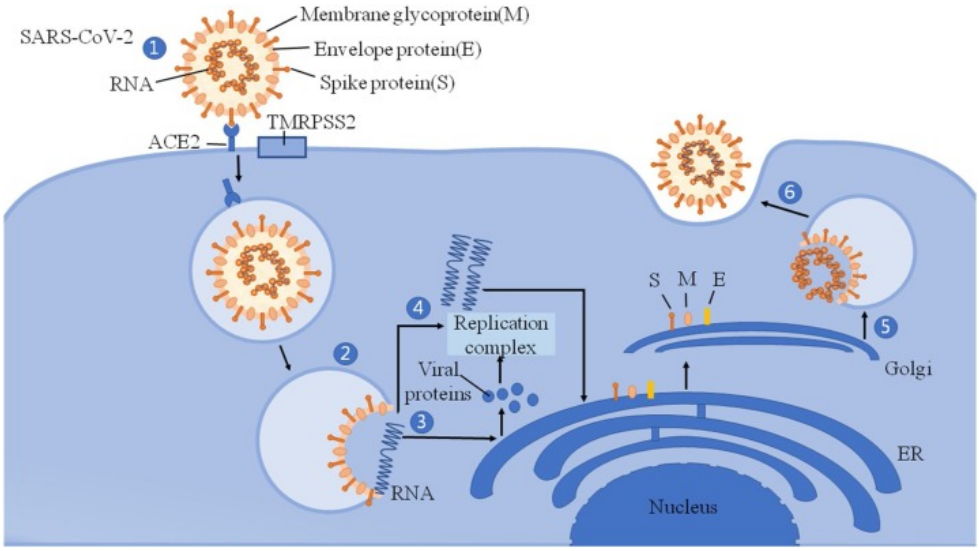


Figure 4.1: *SARS-CoV-2 invasion cellular mechanisms.* ***1.*** *SARS-CoV-2 entry.* ***2.*** *Membrane fusion and viral RNA release.* ***3.*** *Translation.* ***4.*** *Replication and protein synthesis.* ***5.*** *SARS-CoV-2 packaging in golgi.* ***6.*** *SARS-CoV-2 release* [151]

SARS-CoV-2 infection can be divided into three phases: phase I, an asymptomatic incubation period; phase II, non-severe symptomatic period; phase III, a severe respiratory symptomatic stage with high viral infectivity [152]. The damaged cells induce an inflammation state in lung tissues. Pro-inflammatory macrophages and granulocytes primarily mediate it. Lung inflammation is the main cause of respiratory disorders at the severe stage [153].

4.2 Coronavirus sequencing over host species

Coronavirus have been classified in four classes; α, β, γ, and δ [154]. They infect a wide variety of host species, whereas α and β coronaviruses infect only mammals. Moreover, β class includes the three main coronavirus acute respiratory diseases that affect humans; Severe acute respiratory syndrome (SARS-CoV), Middle East respiratory syndrome (MERS-CoV), and COVID19 disease (SARS-CoV-2).
The origin of those coronaviruses has been brought back to a spillover zoonotic phenomena. *Spillover event* occurs when a pathogen found in a reservoir population infects a novel host population belonging to another species. We have

performed an alignment search to compare the SARS-CoV-2 strand (NC_045512) with 2500 coronavirus genomes from different host species. BLAST algorithm [155] has been used to find regions of similarity between genetic sequences. Tab.4.1 shows the alignment scores obtained by BLAST. Bit-score is a numerical value that expresses the size of the search space that we would have to find a score as good as or better than the observed one under the hypothesis of random matching; higher score values correspond to a high similarity. We have found the highest similarity with Bat coronavirus RaTG13 and with pangolin coronavirus PCoV_GX. The similarity of the SARS-CoV-2 with bat and pangolin coronaviruses is higher than the similarity with the SARS-Cov virus. For us, it is further evidence of zoonosis.

The scientific literature supports our results; Zhou et al. [156] have shown that SARS-CoV-2 sequences are 96% identical at the whole-genome level to a bat coronavirus RaTG13, while SARS-CoV-2 shares 79.6% sequence identity to SARS-CoV. Moreover, Malayan pangolins have been identified as intermediate hosts that may have facilitated transfer to humans [157]. Although pangolin is fully protected by Chinese law, it is frequently illegally sold in oriental wet markets. Pangolin meat is eaten, and its scales have a high value and are used to cure various ills in oriental medicine. For this reason, some 4.000 or 5.000 pangolins are illegally imported from Java every year [158].

Genome code	**Bit-score**	**Virus name**
MN996532	48724	Bat coronavirus RaTG13
MT040335	28301	Pangolin coronavirus isolate PCoV_GX-P5L
MT040333	28293	Pangolin coronavirus isolate PCoV_GX-P4L
MT072864	28262	Pangolin coronavirus isolate PCoV_GX-P2V
MT040334	28256	Pangolin coronavirus isolate PCoV_GX-P1E
MT040336	28247	Pangolin coronavirus isolate PCoV_GX-P5E
MG772933	26943	Bat SARS-like coronavirus isolate bat-SL-CoVZC45
MG772934	22223	Bat SARS-like coronavirus isolate bat-SL-CoVZXC21
AY394996	15213	SARS coronavirus ZS-B
AY394997	15213	SARS coronavirus ZS-A

Table 4.1: *Top-10 BLAST bit-score obtained from the comparison between SARS-CoV-2 strand (NC_045512) and other coronavirus genomes from different species.*

Figure 4.2: *Rhinolophus pusillus is the most likely reservoir species of SARS-CoV-2.*

Figure 4.3: *Malayan pangolin could be the intermediate host of SARS-CoV-2.*

4.3 Analysis of interactions between SARS-CoV-2 stains and host miRNAs

The short time since the COVID-19 outbreak is why many aspects of the molecular interactions of SARS-CoV-2 in the human host are still unknown, especially its mechanisms at the transcriptional level. The present study aims to unravel the role of human miRNAs in SARS-CoV-2 infection. miRNAs are short non-coding RNA molecules with a post-transcriptional regulatory function [159]. They bind complementary sequences in mRNA molecules to inhibit the translation of their mRNA targets into proteins [160]. Host endogenous miRNA activity in viral propagation has been previously studied, and many complex virus-specific mechanisms have been identified. However, the precise role of miRNAs in viral infections is not yet fully understood [161]. This section shows the results of an extensive predictive analysis to identify human lung-specific miRNAs that may bind the SARS-CoV-2 RNA. Then, we considered the already experimentally validated miRNA interactions with endogenous genes to identify the host's miRNA regulatory sub-network affected by SARS-CoV-2 infection, looking at the virus as a competing RNA [162]. We finally evaluated the impact of such interactions on the expression profile of genes targeted by the identified miRNAs in human airway epithelial cells infected with SARS- CoV-2. Specifically, we identified miR-1207-5p as a

possible regulator of the S protein in SARS-CoV-2 RNA. As so, we suggest that the viral RNA competes with the CSF1 mRNA, a known target of miR-1207-5p [163], leading to CSF1 overexpression. To support our hypothesis, we have evaluated several published transcriptional datasets. The finding that the CSF1 gene is over-expressed in lung epithelial cells infected with SARS-CoV-2 supported our hypothesis. CSF1 controls the production, differentiation, and function of macrophages, and its overexpression may contribute to the acute inflammatory response observed in severe COVID-19. The results are preceded by three sections that introduces the methodologies used for the analysis.

4.3.1 Transcriptomics datasets and expression Analysis

Normal lung tissue expression profiles have been downloaded from TissueAtlas [164]. Raw miRNA expression data from 18 lung control tissues were normalized with quantile normalization and the average expression level for each miRNA was computed. We used the average expression profile computed from all the 18 control tissues to identify the top 100 expressed miRNAs in normal lung tissue. Tab. 4.2 summarizes the list of selected miRNAs and their average expression level in lung control tissues.
A wide collection of already available transcriptomics datasets with gene expression profiles after SARS-CoV-2 infection has been assembled from literature. When available, we considered the differential expression analysis results obtained by the authors. Otherwise, we preprocessed and analyzed the gene expression profiles to identify differentially expressed genes. When raw count RNAseq data was available, we used the DESeq2 [165] R pipeline to compare infected vs. not infected samples, and the Benjamini-Hochberg procedure [166] to compute adjusted p-values. The univariate threshold of statistical significance was set at 5%.

4.3.2 Analysis of SARS-CoV-2 sequence stability

The RefSeq sequence NC_045512 (recorded in Wuhan, January 2020) was used as reference to predict the binding sites of human miRNAs on the viral RNA. A total of 15881 worldwide viral complete genomes was downloaded —updated to September 7th, 2020— from the Severe acute respiratory syndrome coronavirus 2 data hub of NCBI Virus database, by filtering for taxid = “2697049” and

Nucleotide Completeness = "complete". Stability of particular viral genome regions was assessed by searching the exact match of the region in all the viral available genomes. To assess the statistical significance of the stability of each binding site, we associated a p-value with the number (m_{bs}) of viral sequences that showed a mutation in the region of the binding site. Such a p-value was calculated as the frequency with which a number of mutations larger or equal to m_{bs} was observed in all of the other regions with the same length of the binding site in the involved mRNA.

4.3.3 Prediction of miRNA binding sites on SARS-CoV-2 strands

Mature miRNA sequences were downloaded from miRbase, version 22. We used four miRNA target prediction tools to assess whether an RNA sequence is predicted to be a target of a miRNA: miRanda [48], PITA [49], Targetscan [46], and ComiR [43][44]. miRanda script was used with -score 0 and -energy 0 settings. PITA and Targetscan scripts were used with default settings. ComiR was used to compute the ComiR score associated with the targets of each single miRNAs. For each miRNA we identified as highly predicted targets the genes that passed all the following conditions:

- miRanda binding energy, lower than -20;
- PITA $\Delta\Delta$E, lower than -15
- TargetScan Binding Site, 8mer or 7mer
- ComiR score, greater than 0.85

We used the localization of the binding sites predicted by PITA, miRanda and Targetscan to further restrict the set of targets by considering only the binding sites predicted by all the three algorithms. The resulting targets are named as highly predicted targets.
Experimentally validated miRNA targets were downloaded from miRTarBase, where only the validation methods with strong evidence (i.e., Reporter assays, RT-qPCR, and Western- blot based experiments) have been considered.

miRNA ID	AEL	miRNA ID	AEL	miRNA ID	AEL	miRNA ID	AEL
miR-7975	50459	miR-30b-5p	1306	miR-30a-5p	724	miR-181a-5p	426
miR-7977	26791	miR-22-3p	1271	miR-34a-5p	717	miR-30d-5p	425
miR-8069	17366	miR-24-3p	1243	miR-23b-3p	705	miR-100-5p	419
miR-4516	11347	miR-6125	1227	miR-6085	705	miR-1260a	402
miR-451a	11344	miR-223-3p	1222	miR-199a-3p	667	miR-4466	399
miR-21-5p	11211	miR-26a-5p	1210	miR-7704	660	miR-3162-5p	399
miR-6089	10752	miR-4286	1164	miR-19b-3p	651	miR-6068	383
miR-3960	6238	miR-6800-5p	1097	**miR-103a-3p**	640	miR-4270	381
miR-6090	5613	miR-6088	1070	miR-5739	614	miR-342-3p	377
let-7b-5p	5106	let-7g-5p	1069	miR-7641	593	miR-4739	374
let-7a-5p	4879	miR-4687-3p	991	miR-20a-5p	593	miR-150-5p	363
miR-5100	4178	miR-1202	981	miR-130a-3p	589	miR-200c-3p	353
miR-6869-5p	3325	let-7i-5p	962	miR-26b-5p	588	miR-107	337
miR-16-5p	2943	let-7c-5p	910	let-7d-5p	572	miR-99a-5p	300
miR-4459	2598	miR-29b-3p	895	miR-27b-3p	552	miR-3656	299
miR-126-3p	2573	miR-1915-3p	853	miR-4530	548	miR-30c-5p	297
let-7f-5p	2376	**miR-4763-3p**	842	miR-3665	505	miR-106b-5p	285
miR-6749-5p	1955	miR-125b-5p	832	miR-5787	499	miR-4741	275
miR-4281	1918	miR-2861	824	miR-7107-5p	492	miR-642a-3p	268
miR-29a-3p	1898	miR-1225-5p	802	miR-4284	492	miR-1260b	252
miR-6087	1701	miR-15b-5p	772	miR-142-3p	487	miR-1246	250
miR-29c-3p	1609	miR-638	770	miR-145-5p	483	miR-200b-3p	247
miR-6821-5p	1406	**miR-1207-5p**	766	miR-1273g-3p	473	miR-497-5p	240
miR-27a-3p	1378	miR-195-5p	758	let-7e-5p	468	miR-6826-5p	237
miR-23a-3p	1318	miR-141-3p	733	miR-15a-5p	458	miR-6165	231

Table 4.2: *Summary of the most expressed miRNAs in normal lung tissues. In red, the miRNAs predicted by four algorithms (i.e., PITA, miRanda, TargetScan, and new ComiR); the original ComiR version predicts both red and blue miRNAs. In red bold and underlined, the 5 selected miRNAs.*
AEL = Average Expression Level.

4.3.4 Role of endogenous miRNAs in COVID-19 disease

The methodologies presented in the previous sections are used to investigate on SARS-CoV-2 genome in terms of mutations and interaction with endogenous miRNAs. The present section reports the main results obtained from data analysis.

Five human lung-specific miRNAs are predicted to target SARS-CoV-2 viral genome

Aiming to unravel the role of endogenous miRNA expressed in the human lung with respect to SARS-CoV-2 virus, we focused our analysis on the 100 most expressed miRNAs in normal lung [164], identified as described in section 4.3.1. We identified potential targets of these 100 miRNAs on SARS-CoV-2 RNA sequence (NCBI reference viral sequence NC_045512), using four miRNA target prediction tools [49][48][46][43]. Only 15 miRNAs were predicted to target the viral RNA by all the four algorithms (Fig. 4.4). Among the predicted miRNA:viral RNA interacting pairs, six specific binding sites (specific target locations) were identified by all four algorithms (Fig. 4.6 and 4.7).
The six sites were targeted by 5 miRNAs: miR-6089, miR- 6821-5p, miR-103a-3p, miR-4763-3p, and miR-1207-5p (we consider those miRNAs as strong predictions because the four algorithms have consistently predicted their binding sites, whereas different algorithms could predict distinct binding sites on the same miRNA). miR-4763-3p and miR-1207-5p miRNAs belong to the same miRNA family, sharing the same seed sequence (ggcaggg). In our analysis, we predict that they have a common binding site in the viral sequence, located in the region coding for the Spike (S) glycoprotein. Spike is a structural protein that allows Sars- Cov-2 to enter host cells by interacting with membrane receptors [167]. Human miRNAs miR-6089, miR- 6821-5p, and miR-4763-3p have their binding sites in the ORF1ab gene, specifically hitting the regions coding for Nsp10, formerly known as growth-factor-like protein (GFL), Nsp12, an RNA-dependent RNA polymerase, and Nsp13_ZBD gene, a helicase (Fig. 4.7). The three mentioned non- structural proteins are crucial in coronavirus replication, being part of a complex of 16 non-structural proteins entailed for viral RNA replication and transcription [167] [168]. miR-103a-3p binding site is located in the Nucleocapsid (N) protein coding region. N proteins are

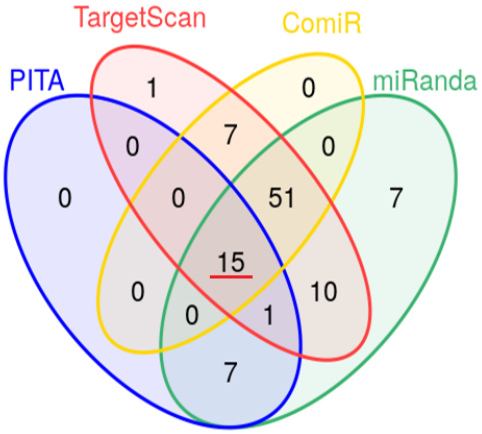


Figure 4.4: *miRNA-target prediction results of 100 top expressed miRNA in normal lung on COVID19 (NCBI Reference sequence NC_045512.2). Each group in the Venn diagram represents the set of miRNAs predicting as target the COVID19 sequence by applying one of the considered algorithms (PITA, Targetscan, miRanda and ComiR).*

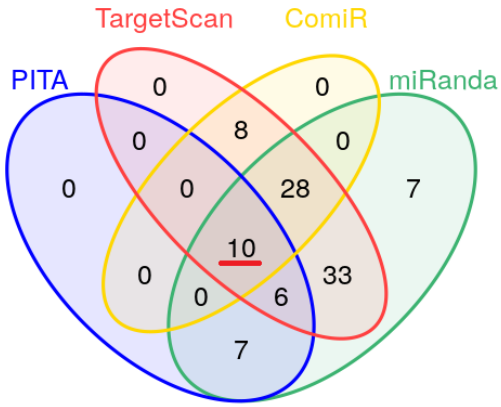


Figure 4.5: *miRNA-target prediction results of 100 top expressed miRNA in normal lung on COVID19. miRNA ComiR predictions from the original ComiR have been replaced with the predictions from the new version.*

structural proteins, that play key roles during the packaging of the viral RNA genome [167]. Whether the enhancement of the host's miRNAs regulatory machinery could inhibit the replication process or the production of the structural viral proteins, and as a consequence the virus diffusion through the host, is a hypothesis that needs to be experimentally validated and requires further investigation.

The new version of ComiR produces more specific predictions

ComiR algorithm is based on a support vector machine (SVM) trained on messenger RNA information from *Drosophila Melanogaster* experiments. The previous studies showed that ComiR algorithm can predict miRNA targets not only on *Drosophila* but also on other species such as *Human* and *C. Elegans*

ComiR	SVM training	predicted miRNAs	predictions by all the 4 algorithms	strong predictions by all the 4 algorithms
original	3'UTR	75	15	5
novel	coding	6	6	4
novel	3'UTR	44	8	3
novel	3'UTR/coding	46	10	5

Table 4.3: *Comparison of miRNA groups predicted to target SARS-CoV-2 genome using different ComiR algorithms. The new ComiR version predicts fewer miRNAs then the original version. But the new ComiR predictions are more specific than the original ComiR predictions.*

[43][44]. The ComiR predictions reported in the previous section have been obtained using the original version of ComiR [43][44]. We have recently developed a new ComiR version [4] as described in Chapter 3. The new ComiR algorithm uses the information contained in the coding region to improve its prediction capacity. Whereas the original ComiR only considers 3'UTR information for the SVM training (see Chapter 3).
In the present analysis, we have compared ComiR predictions from both versions. The new ComiR predictions have been obtained using both 3'UTR and coding region information for training; Tab.4.3 shows the comparison of miRNA predicted groups using different ComiR algorithms.

The new ComiR version predicts fewer miRNAs than the original version, but those novel predictions are more specific than the original ComiR predictions; indeed, the new algorithm has identified all the strong miRNA predictions (i.e., miRNAs whose binding sites have been consistently predicted by PITA, miRanda and TargetScan). In particular, the original ComiR predicted 75 miRNAs; five of them were identified as strong binders (i.e., miR-6089, miR-6821-5p, miR-103a-3p, miR-4763-3p, and miR-1207-5p). The new ComiR version trained using coding region information predicted only 6 miRNAs. Four of them are identified as strong binders (i.e., miR-6089, miR-6821-5p, miR-4763-3p, and miR-1207-5p). Whereas the new ComiR version trained using 3'UTR information predicted 44 miRNAs. Three of them are identified as strong binders (i.e., miR-6089, miR- 6821-5p, and miR-103a-3p).
Unifying the results from both ComiR training sets (based on 3'UTR and coding region), we predict all the strongest miRNA binding sites. Fig.4.5 shows

the intersection of miRNAs predicted by miRanda, Targetscan, PITA, and the new ComiR version.

Stability of predicted miRNA binding sites on SARS-CoV-2 RNA

The worldwide spread of COVID-19 infection exposes the viral genome to a high risk of mutation. For this reason, we checked the binding sites' sequence stability across the 15881 SARS-CoV-2 genomes annotated from all over the world in the NCBI virus database.
An example of mutation identification is shown in Tab. 4.4, it reports mutations in miRNA binding sites located on spike (the binding sites of other miRNAs are not shown).
To analyze such a stability, for each one of the six selected binding sites, we counted the number of viral sequences that presented a mutation. Results are reported in the third column of Fig. 4.6. We found that the binding regions are highly stable, which implies the consequent stability of binding site predictions across the currently circulating viruses. In addition, we compared the occurrences of mutations in each binding site (m_{bs}) with the occurrences in any other region of the same length in the involved viral coding RNA, as described in the section 4.3.2.
The obtained p-values (see Fig. 4.6) indicate that the stability of all the six binding sites does not show a significant deviation from the one of the whole mRNA in which they are located, respectively. Where the p-values have been calculated as the frequency with which a number of mutations larger or equal to m_{bs} was observed in all of the other regions with the same length of the binding site in the involved mRNA.

Host mRNAs competing with SARS-CoV-2 RNA are overexpressed in Lung Epithelial Cells

The viral sequence, once expressed, can interact with the host's miRNA regulatory machine by sequestering the selected miRNAs. Therefore, viral RNA may act as a miRNA sponge, with the same mechanism of competing endogenous RNA [162]. Among the five selected miRNAs, two have been previously studied in detail. miR-103a-3p activity has been widely studied in different tissues, i.e., gastric and colorectal cancer or liver [169][170][171] [172][173][174][175][176][177] [178], and a number of its targets has been validated. miR-1207-5p expression is high in the cytoplasmic fraction of human normal lung tissue while being

SARS-CoV-2 code	miRNA binding site in spike
NC_045512	gaacttcacaactgctcctgcca
MT958259	gaacttcacaactactcctgcca
MT451181	gaacttcacaactgcttctgcca
MT706284	gaacttcacaactgttcctgcca
MT706442	gaacttcacaactgttcctgcca
MT263443	gaacttcacaactgttcctgcca
MT873480	gaacttcacaacttctcctgcca
MT811251	gaacttcacaacttctcctgcca
MT627751	gaacttcacaacttctcctgcca
MT628092	gaacttcacaacttctcctgcca
MT334539	gaacttcacaacttctcctgcca
MT334540	gaacttcacaacttctcctgcca
MT800995	gaacttcacaattgctcctgcca
MT345855	gaatttcacaactgctcctgcca
MT873372	taacttcacaactgctcctgcca
MT831540	taacttcacaactgctcctgcca
MT745640	taacttcacaactgctcctgcca
MT745652	taacttcacaactgctcctgcca
MT795896	gaacttcacaactgctccngcca
MT252730	gaacttcacaacnnnnnnnnn

Table 4.4: *Mutations in miRNA binding sites located on spike (miRNA-4763-3p and miRNA-1207-5p are involved in the binding). Red letters indicate the presence of mutation with respect to NC_045512. Almost all mutations are of type $g \longrightarrow t$ or $c \longrightarrow t$. The last two rows contain sequences with unknown bases; we have replicated the analysis by excluding RNA regions with unknown bases, we found that p-values of Tab. 4.6 don't have significant variations.*

reduced in cancer [163]. miR-1207-5p has been first characterized as negative regulator of epithelial-to-mesenchymal transition (EMT) as it inhibits the expression of a number of genes involved in this process, including Snail, Smad2, Smad3 and Vimentin [163][179]. In addition to its role in EMT, miR-1207-5p plays an important role in shaping the inflammatory milieu. In this respect, CSF1 (colony-stimulating factor 1, also known as macrophage colony-stimulating factor, M- CSF) has been reported as one of its direct targets [163].

miRNA Binding site Start..Stop	Alignment	# of COVID19 with partial matching	p-value
miR-4763-3p: 13029..13052:	3' gggcggGUCGUGGUCGGGGACGGa 5' || || ||: |||||| 5' gtaatgCAACAGAAGTGCCTGCCa 3'	10	0.72
miR-6089: 15322..15345:	3' ggcggggcggggUGGGGCCGGAGg 5' |:: :|||||| 5' aacatgcttagaATTATGGCCTCa 3'	127	0.06
miR-6821-5p: 17443..17465:	3' ggggcggagcucgGUGGUGCGUg 5' ||||||||| 5' gctcaattacctgCACCACGCAc 3'	7	0.92
miR-4763-3p: 24781..24803:	3' gggcGGGUCGUGGUCGGGGACGGa 5' ::|| ||	||:|||||| 5' gaacTTCA-CAACTGCTCCTGCCa 3'	19	0.51
miR-1207-5p: 24782..24803:	3' ggGGAGGGUCGG-AGGGACGGu 5' |:|| || |: :|||||| 5' aaCTTCACAACTGCTCCTGCCa 3'	15	0.57
miR-103a-3p: 28717..28741:	3' aguaUCGGGAC-A-UGUUACGACGa 5' | :||||	|||||||||| 5' ccgcAATCCTGCTAACAATGCTGCa 3'	58	0.37

Figure 4.6: *Six miRNA:viral-RNA targets predicted by all methods ("high confidence targets"). Column-1: miRNA name, start/stop bases in the NC_045512 sequence; column-2: base alignment; column-3: number of SARS-CoV-2 sequences not containing an exact match for the binding site region; column-4: p-value*

In order to test whether infection of lung epithelial cells with SARS-CoV-2 cell infection affects the gene expression levels of the endogenous miRNA target genes, we used a recent dataset of gene expression profiles of human lung-derived cells infected with SARS-CoV-2 [180]. Authors examined the behavior of wild type adenocarcinomic human alveolar basal epithelial (A549) and airway epithelial (Calu3) cell lines. A549 cells show a low expression of ACE2 receptor, hence a limited coronavirus infection rate. Thus, the authors also analyzed A549 cells transfected with a vector expressing ACE2 (A549+ACE2). We used this dataset to analyze the transcriptional profiling of the experimentally validated targets of miR-1207-5p and miR- 103a-3p.

Figure 4.8.a presents the effect of viral infection on miR-1207- 5p and miR-103a-3p endogenous target gene expression. Log-fold change ($log_2(FC)$) values are calculated by comparing SARS- CoV-2 infected vs. mock treated cell lines

(as described in section 4.3.1).

We expect that endogenous direct targets will increase their expression level following SARS-CoV-2 infection since viral RNA will compete with the endogenous RNA. Some of the analyzed targets behave as expected, especially the ones that are in the range of 1,000–2,000 reads per million (rpm), including CREB1, CSF1, PTEN, and DICER1. Consistent with the known A549 limited infection rate, the expression of these genes is enhanced in ACE2-expressing A549 cells, and even more in Calu-3 cells that are highly permissive to SARS-CoV-2 replication. These findings support our hypothesis that the viral RNA may act as a competing RNA for a selection of host miRNAs leading to the increase of the expression level of their endogenous targets.

Highly expressed targets, for instance ADAM10, are not up- regulated as expected. This is probably due to the fact that these genes might be modulated by other highly expressed miRNAs not sequestered by the virus. Alternatively, the sponge effect that we are hypothesizing is not effective when the mRNA is highly expressed.

Figure 4.8.b presents the complexity of the miRNA-target network known up to now. Here we map all the experimentally validated interactions among the list of direct targets of miR-1207- 5p and miR-103a-3p, and 45 of the 100 most highly expressed miRNAs in healthy lungs, that show at least one interaction. For instance, we observe that ADAM10, one of the targets of miR- 103a-3p, is also regulated by miR-451a, the most highly expressed miRNA in lung. The presence of this regulator might be the reason why the expression of ADAM10 is not affected by the presence of the virus.

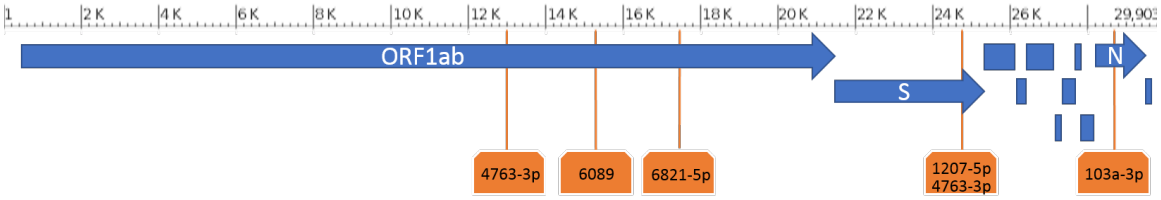


Figure 4.7: *Location of the five high confidence targets on the SARS-CoV-2 genome.*

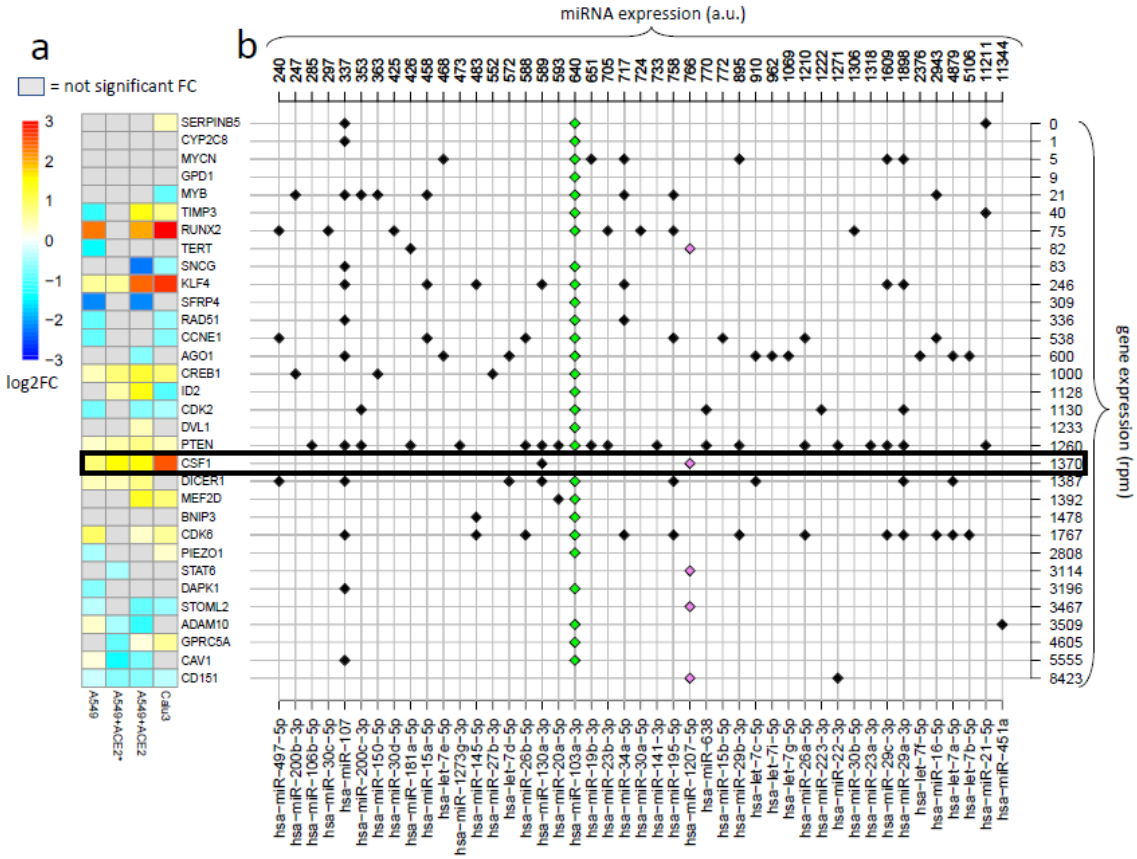


Figure 4.8: *Overview of validated targets of hsa-miR-1207-5p and hsa-miR-103a-3p (GEO dataset GSE147507).* ***A****) Heatmap of the log2FC in gene expression between SARS-CoV-2 infection vs. mock treatment in cells with different multiplicities of infection (MOI). Cells: A549 (low ACE2 expression), A549+ACE2 (ACE2-expressing A549 cells, low MOI = 0.2), A549+ACE2 (ACE2-expressing A549 cells, MOI = 2–5), Calu3 cells (MOI = 2–5) (GEO dataset GSE147507). Only the log2FC that are associated with adjusted p-value <0.05 are displayed. Target genes are ordered according to their average expression level in A549 cells.* ***B****) Map of annotated interactions among the targets of hsa-miR-1207-5p (pink) and hsa-miR-103a-3p (green) and other highly expressed in normal lung tissue miRNAs. miRNAs are ordered according to their expression level. Genes are in the same order as in panel A) and their expression levels are shown on the right of the grid.*

Binding of miR-1207-5p to SARS-CoV-2 RNA may lead to overexpression of EMT-related genes and CSF1

miR-1207-5p has been first characterized as negative regulator of EMT by controlling the expression of several genes including SMAD2, SMAD3, SMAD7,

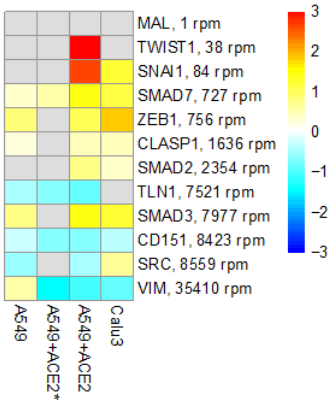


Figure 4.9: *Overview of genes involved in EMT process and reported to be regulated by miR-1207-5p. The heatmap shows the log2FC in gene expression between SARS-CoV-2 infection vs. mock treatment in the same cells as in Fig. 4.8. Targets are ordered according to their average expression level in A549 cells.*

CLASP1, ZEB1, and SNAIL1 [163][179]. EMT processes favor fibrotic events. Of interest, current data suggest that pulmonary fibrosis after COVID-19 recovery could be substantial [181][182][183]. Therefore, we tested the hypothesis that SARS-CoV-2 infection in bronchial epithelial cells may have an impact on the expression of these genes by reducing the availability of miR-1207-5p. Fig. 4.9 shows the results of the differential expression analysis for the genes involved in EMT that have been reported to be regulated by miR-1207-5p. The increase in their expression levels appears evident when cells are infected with SARS-CoV-2 virus, therefore supporting our hypothesis.

We further expanded our analysis by evaluating the impact of SARS-CoV-2 infection on the expression of CSF1. As reported in Fig. 4.8, CSF1 is one of the host gene targets most upregulated following viral infection. CSF1 is a predicted target of 3 out of 5 of the miRNAs targeting the virus sequence: miR-4763-3p, miR- 1207-5p and miR-6089. It is also an experimentally validated target of miR-1207-5p [163]. The only other known miRNA CSF1 regulator, among the 100 highly expressed miRNA in the lung, is miR-130a-3p, which is expressed at lower level than miR-1207-5p. CSF1 regulates the survival, proliferation, differentiation, and chemotaxis of tissue macrophages and dendritic cells (DC) that play a key role in innate immune responses. In the human lung, CSF1 can be released by airway epithelial cells in the airspace and its

local concentration contributes to control the recruitment and activation of DC and macrophages [184][185][186].

To further validate our hypothesis that the CSF1 mRNA is over-expressed after SARS-CoV-2 infection, we analyzed several recently published datasets. To this purpose, different types of experimental designs and platforms were taken into consideration. When available, we referred to the differential expression analysis performed by the authors. Specifically, we considered transcriptomics data analysis of infected vs. healthy samples from human lung biopsies as reported in [187], bronchoalveolar lavage fluid (BALF) in [188], peripheral blood mononuclear cells (PBMC) and BALF in [189], and whole blood in [190]. We also analyzed the single cell RNAseq data from whole blood reported in [191], infected NHBE cells in [193], and infected Calu3 cells in [192]. Data sets obtained by analyzing human samples were not useful to confirm our hypothesis. This can be due to several reasons. More specifically, the high-variability among patients, the cell heterogeneity of reported biological samples (such as bronchioalveolar lavage fluids and lung biopsies) with different efficiency of viral transfection and the low sample size make it really difficult to unravel fine regulatory mechanisms of virus-host interaction. On the contrary, when dataset derived from bronchial epithelial cells (both primary cells and cell lines) were analyzed, significant upregulation of CSF1 was observed therefore confirming our hypothesis. For example, [192] performed gene expression profiles of SARS-CoV-2 infected Calu3 cell line. Overexpression of CSF1 in Sars-CoV-2 infected versus mock treated cells confirmed our hypothesis. Furthermore, in [193] the authors performed single-cell RNA sequencing of human bronchial epithelial cells grown in air-liquid interface and infected with SARS-CoV-2. When looking at ciliated cells, the expression of CSF1 significantly increased in infected compared to mock cells. Of note, the expression of CSF1 was significantly higher in ciliated infected cells compared to bystander cells that remained uninfected in samples challenged with SARS-CoV-2. These findings suggest that viral replication inside the cells is required in order for CSF-1 to be over-expressed therefore supporting that a direct interaction between viral RNA and host miRNAs is required to alter the expression of CSF1 during infection.

4.4 Discussion

In 10–20% of the cases, SARS-CoV-2 infections may progress to interstitial pneumonia and acute respiratory distress syndrome (ARDS) especially in patients with older age and comorbidities. Clinical features of severe COVID-19 as well as their systemic cytokine profile suggest the occurrence of macrophage activation syndrome (MAS) [194][182]. High rates of viral replication have been listed among the factors that may drive severe lung pathology during infection by contributing to enhanced host cell cytolysis and production of inflammatory cytokines and chemokines by infected epithelial cells [182][195][196]. We propose that the high concentration of viral RNA in the cell may sequestrate miR-1207-5p therefore contributing to CSF1 release leading to enhanced macrophage recruitment and activation. In fact, increased release of CSF1 may represent a predisposing factor for MAS and cytokine storm secondary to viral infection [197][198]. Consistently, it has been recently reported that T-cell derived CSF-1, acting via intercellular crosstalk, may be associated with cytokine storm in COVID-19 [195]. In our proposed model, infected bronchial epithelial cells may be a source of CSF-1 contributing to local and systemic inflammatory profiles. In addition, reduced availability of miR-1207-5p may also promote EMT events therefore favoring fibrosis [181][182][183]. Although further experimental validation will be required to confirm direct interaction between miR-1207-5p and the SARS-CoV-2 genome, our proposed model has been confirmed using several published datasets. Results herein reported strongly suggest that upregulation of CSF1 due to interaction of miR-1207-5p with viral genome may occur when lung epithelial cells are infected with a high viral load. A limitation of the current study is the lack of data regarding protein levels and release. To address this issue, we carefully looked for published proteomics data in COVID19 literature, but, so far, no information about CSF1 protein levels has been published and therefore further studies will be carried out to address this point.

Nevertheless, transcriptional and post-transcriptional control of mRNA levels represent a key regulatory step for most inflammatory mediators during infection. In this respect, the discovery of novel potential mechanisms that contribute to modulate the mRNA levels of a specific inflammatory mediator in the context of SARS-Cov-2 infection may represent a step forward toward a better understanding of virus-host interaction molecular mechanisms.

A wide analysis of the SARS-CoV-2 transcriptome [199] revealed the presence

of several non-canonical sub- genomic RNAs. They consist in discontinuous transcriptions of the viral sequence, where the 5' leader region is fused to a non- conventional part of the genome. As a result, the obtained RNA contains only a portion of the viral mRNAs. It is tempting to speculate that they may play a role as competing RNA. Specifically, miR-1207-5p related binding site is located in the far downstream region of the viral gene Spike. As a consequence, almost all of the sub-genomic RNA sequences with the fusion occurring in the region of the Spike gene contain the miR-1207- 5p binding site. Although, these sub-genomic RNA sequences do not have the coding potential to yield the S protein, they could still act as miRNA sponges.
To conclude, our results suggest that the miR-1207-5p family may interact with SARS-CoV-2 viral genome leading to deregulation of CSF-1, which may enhance inflammatory responses in COVID-19 patients, and promoting EMT, which can contribute to pulmonary fibrosis, a possible sequela of COVID-19. Further experimental validation will be conducted to confirm molecular mechanisms of host-virus interaction and to investigate their involvement in disease progression.

4.5 Conclusion

We have identified a small group of five miRNAs whose binding sites in SARS-CoV-2 genome have been predicted by four different algorithms. We have also verified the stability of SARS-CoV-2 strands at the binding site of interest.
Our results support the idea that predicted miRNAs interact with SARS-CoV-2 strands. We propose the involvement of miR-1207-5p family and CSF1 in the progression of COVID-19 infection, and as possible targets for COVID-19 treatment. Further experimental validation is still due to confirm the binding of miR-1207-5p into the viral genome, and the role of SARS-CoV-2 in the regulation of CSF1 expression.

Chapter 5

A novel statistical test for differential expression analysis

As we have shown in the previous chapters, differentially expressed genes (DE genes) from IP experiments have been considered for training (and also testing) miRNA target prediction algorithms such as ComiR [4]. An improvement of the procedure to identify DE genes that compose ComiR training set can improve the algorithm prediction capacity. For this reason, differential expression analysis has a central role in this thesis; the present chapter introduces a novel procedure for DE gene identification.

Differential expression analysis (DEA) is widely used in transcriptomic studies. Some t-test variants have been proposed in the literature to identify enriched and under-represented transcripts; however, the deviation from the normal assumption in small samples makes t-test p-values not reliable.

We propose a novel exact statistical test, obtained from the hypergeometric distribution, able to identify genes missed by the t-test. The analysis of real gene expression datasets supports the efficiency of our method and suggests its application together with the t-test approach to better understand the biological questions related to differentially expressed genes.

5.1 Background

Differential expression analysis (DEA) is a large-scale inference procedure used to identify genes whose expression differ under different biological conditions. A large family of t-tests is the most widely used procedure for DEA [203] [204]. But this methodology depends on parametric assumptions rarely sat-

isfied. However, large samples allow an assumption relaxation, but the high cost of experiments makes it difficult to find. For this reason, in small skewed samples, t-test p-values are often not reliable [205].
Moreover, the small variance of low expressed genes makes the denominator of t-test statistics unnaturally smaller. It increases the total I type error and the number of significant genes. Alternative definitions of the t-test have been proposed to reduce the impact of small samples and low expression variability, e.g., *moderated t-test* [206] and *Significance Analysis of Microarray* (SAM) [207]. On the other hand, large sample t-tests produce too many significant genes; it depends on average expression differences truly different from zero but not large enough to be biologically meaningful.
A common strategy to reduce the number of selected differentially expressed genes is to set a threshold on the fold change (e.g. 1, 1.5, or 2) [208]. But this solution depends on an arbitrary parameter.
In this chapter, we proposed a novel statistical test for DEA obtained from multivariate hypergeometric distributions. The novel test is indicated as Hy-test. At the price of a slight loss of information, Hy-test presents several advantages;

- free from parametric assumptions
- allows implicit discretization of the expression profiles.
- provides more reliable p-values than the t-test p-values

The analysis presented here shows that the Hy-test is able to identify genes missed by the t-test. The results suggest that Hy-test and t-test can be used together to better understand the biological questions that are investigated by a DEA approach.

5.2 Preprocessing procedure for microarray data

As a preliminary step for p-value calculation, gene expression profiles have been normalized and discretized in three levels: $\{-1, 1, 0\}$, meaning "down-regulated", "upregulated", and "no-changed" respectively. Our discretization approach maximize the disagreement between the discretized levels of the two different experimental conditions. After that, exact p-values are calculated on the discretized expression profiles. The following sections give a detailed

description of our procedure for DEA. Before introducing the methodology, we linger on the data preprocessing:
As test bed, we consider gene expression profiles of breast cells in a pattern of paired tissues; 17.632 genes have been recorded in 75 tumor tissues and 75 normal tissues (than the analysis have been replicated by considering 67 kidney renal clear cell carcinoma - KIRC - paired with 67 normal tissues). Data has been downloaded from the TCGA website. The expression profiles of duplicated genes have been replaced with their mean expression. Moreover, The expression of each gene has been normalized using a quantile normalization [209] and then log-transformed.
However, some genes aren't expressed in almost all the tissues of interest. For this reason, we propose a method of molecule selection; for each tissue, we select the highest expressed genes that explain at least 50% of the whole expression. Considering the vector $\left(x_{(1)}, x_{(2)}, ..., x_{(N)}\right)$ of ordered gene expression values in one tissue. We selected the first-l genes such that:

$$l = \text{argmin}_k \left\{ k \left| \frac{\sum_{i=1}^{k} x_{(i)}}{\sum_{i=1}^{N} x_i} \geq 0.5 \right. \right\}$$

where, $x_{(i)}$ is the i^{th} expression value in the ordered vector, and N is the total number of genes.
Using this criterion for each tissue, we included genes that have been selected in at least one tissue; i.e., 14.569 genes in the breast tissue analysis, and 14.482 in the kidney tissue analysis. Considering those selected genes, two different expression analyzes have been carried out separately; one on breast tissues and another on kidney tissues.

5.3 Recording the expression profiles

The discretization of gene expression data (GED) is widely used in genomics analysis. Despite a certain loss of information, GED discretization is often used as a preprocessing step to reduce the noise of raw data and to obtain a more straightforward interpretation of the data [213].
Several algorithms require data discretization during the preprocessing (e.g. biclustering method [210]). Moreover, many network models require discrete data as input (e.g., *Bayesian Networks* and *logical networks* [212] [211]).
Despite the importance of discretization in transcriptomics, the criteria under

discretization methods are always arbitrary (e.g., the *FC-discretization* depends on a threshold arbitrary fixed, generally equal to 1, 1.5 or 2; the *equal width discretization* depends on a tuning parameter; the *ranking discretization* depends on the X^{th} percentile that identifies the top-X% genes).
Here, we introduce a novel approach to gene expression discretization based on reasonable criteria free from arbitrary parameters. Let's consider a gene expression profile recorded on two experimental conditions, e.g. normal and cancer tissues, for a total of n pairs of tissues. We estimate a threshold couple able to discretize gene expression as "downregulated", "upregulated", and "no-changed". The optimum thresholds are obtained by maximizing the disagreement between the discretized levels of the two different experimental conditions.
Applying the thresholds $\{k_1, k_2\}$ on the whole expression of a single gene, we obtain two discretized vectors, one for healthy tissues, say $\vec{v}_H$, and one for diseased tissues, say $\vec{v}_D$, with entries that take values $\{-1, 0, 1\}$ that means "downregulated", "no-changed", and "upregulated" respectively. The thresholds $\{k_1, k_2\}$ are estimated by maximizing the quantity

$$H(\vec{v}_H, \vec{v}_D) = n_{+,-} + n_{-,+}$$

where $n_{+,-}$ ($n_{-,+}$) is the number of tissue couples that present upregulated normal (cancer) tissues paired with downregulated cancer (normal) tissues. Optimization research has been carried out by using a genetic algorithm [215]. We have estimated a threshold for each gene of the dataset. This method can be easily adapted to extract a single cutoff couple for all genes.

5.4 Analytical derivation of an exact test for DEA

Let's consider a gene expression profile recorded on two experimental conditions, e.g. normal and cancer tissues, for a total of n pairs of tissues. We aim to calculate a p-value to evaluate if gene expression is significantly different over cancer and normal tissues. Consider two threshold couples, one for normal tissues and another for cancer tissues, able to discretize gene expression as "downregulated", "upregulated", and "no-changed" (later, we will extend the p-value calculation to the more realistic situation of a single threshold couple).

In this way, two vectors with n components are obtained, one for healthy tissues, say $\vec{v}_H$, and one for diseased tissues, say $\vec{v}_D$, with entries that take values $\{-1, 1, 0\}$. To investigate the differential expression we are interested in the quantity

$$H(\vec{v}_H, \vec{v}_D) = n_{+,-} + n_{-,+}, \tag{5.1}$$

where $n_{+,-}$ ($n_{-,+}$) is the number of tissue couples that present upregulated normal (cancer) tissues paired with downregulated cancer (normal) tissues. To associate a p-value with $H(\vec{v}_H, \vec{v}_D)$ it's necessary, as a preliminary step, to evaluate the probability that $n_{match} = n_{+,-} + n_{-,+}$ occurs by chance. Constraints on the total number of positive, negative, and null signs are set on both vectors in the null hypothesis. Specifically, the null model is based on external parameters $\vec{K}_H = (K_H^+, K_H^-, K_H^0)$ and $\vec{K}_D = (K_D^+, K_D^-, K_D^0)$, where K_H^i (K_D^i) is the total number of tissues with sign i in vector $\vec{v}_H$ (vector $\vec{v}_D$), with, $i \in \{-1.1, 0\}$. Such parameters are not independent. Indeed, $K_H^+ + K_H^- + K_H^0 = K_D^+ + K_D^- + K_D^0 = n$, where n is the total number of tissue couples in the dataset. We are interested in calculating the probability that matrix

$$C = \begin{pmatrix} n_{+,+} & n_{+,-} & n_{+,0} \\ n_{-,+} & n_{-,-} & n_{-,0} \\ n_{0,+} & n_{0,-} & n_{0,0} \end{pmatrix} \tag{5.2}$$

occurs by chance, subject to the aforementioned constraints. An entry $n_{i,j}$ of C represents the number of tissues that display sign i in vector $\vec{v}_H$ and sign j in $\vec{v}_D$. Notation C is used here because sometimes matrices such as the one above are indicated as “confusion" matrices. Entries of matrix C are not independent due to the constraints on the number of positive, negative, and null signs described above. Specifically, they are linearly dependent according to the following six equations:

$$\begin{cases} n_{+,+} + n_{+,-} + n_{+,0} = K_H^+ \\ n_{-,+} + n_{-,-} + n_{-,0} = K_H^- \\ n_{0,+} + n_{0,-} + n_{0,0} = K_H^0 \\ n_{+,+} + n_{-,+} + n_{0,+} = K_D^+ \\ n_{+,-} + n_{-,-} + n_{0,-} = K_D^- \\ n_{+,0} + n_{-,0} + n_{0,0} = K_D^0 \end{cases} \tag{5.3}$$

This linear system has rank equal to 5, because of the linear relationship between parameters: $K_H^+ + K_H^- + K_H^0 = K_D^+ + K_D^- + K_D^0 = n$. Therefore, it can

be solved as

$$\begin{cases} n_{+,0} = K_H^+ - n_{+,-} - n_{+,+} \\ n_{-,0} = K_H^- - n_{-,-} - n_{-,+} \\ n_{0,+} = K_D^+ - n_{-,+} - n_{+,+} \\ n_{0,-} = K_D^- - n_{-,-} - n_{+,-} \\ n_{0,0} = K_H^0 + K_D^0 - n + n_{-,-} + n_{-,+} + n_{+,-} + n_{+,+}. \end{cases} \tag{5.4}$$

This result indicates that matrix C is determined if so are $n_{-,-}, n_{-,+}, n_{+,-}, n_{+,+}$. Therefore the probability

$$\begin{aligned} P(C) &= P(n_{-,-}, n_{-,+}, n_{+,-}, n_{+,+} \mid \vec{K}_H, \vec{K}_D) = \\ &= P(n_{-,-}, n_{-,+} \mid n_{+,-}, n_{+,+}, \vec{K}_H, \vec{K}_D)\, P(n_{+,-}, n_{+,+} \mid \vec{K}_H, \vec{K}_D), \end{aligned} \tag{5.5}$$

where, according to a simple combinatorial analysis of the problem,

$$P(n_{+,-}, n_{+,+} | \vec{K}_H, \vec{K}_D) = \frac{\binom{K_D^+}{n_{+,+}}\binom{K_D^-}{n_{+,-}}\binom{K_D^0}{n_{+,0}}}{\binom{n}{K_H^+}} \tag{5.6}$$

and

$$P(n_{-,-}, n_{-,+} \mid n_{+,-}, n_{+,+}, \vec{K}_H, \vec{K}_D) = \frac{\binom{K_D^+ - n_{+,+}}{n_{-,+}}\binom{K_D^- - n_{+,-}}{n_{-,-}}\binom{K_D^0 - n_{+,0}}{n_{-,0}}}{\binom{n-K_H^+}{K_H^-}}. \tag{5.7}$$

The distribution of C allows to calculate the probability

$$P\left[H(\vec{v}_H, \vec{v}_D) = x\right] = P\left(n_{+,-} + n_{-,+} = x\right) = P(x) \tag{5.8}$$

as

$$\begin{aligned} P(x) &= \sum_{\{n_{+,+}, n_{-,-}, n_{-,+}\}} P(n_{-,-}, n_{-,+} \mid x - n_{-,+}, n_{+,+}, \vec{K}_H, \vec{K}_D) P(x - n_{-,+}, n_{+,+} \mid \vec{K}_H, \vec{K}_D) \\ &= \sum_{\{n_{+,+}, n_{-,-}, n_{-,+}\}} \frac{\binom{K_D^+}{n_{+,+}}\binom{K_D^-}{x-n_{-,+}}\binom{K_D^0}{n_{+,0}}}{\binom{n}{K_H^+}} \frac{\binom{K_D^+ - n_{+,+}}{n_{-,+}}\binom{K_D^- - x + n_{-,+}}{n_{-,-}}\binom{K_D^0 - n_{+,0}}{n_{-,0}}}{\binom{n-K_H^+}{K_H^-}}. \end{aligned}$$

According to this distribution, the p-value associated with an observation $\hat{x} = \hat{n}_{-,+} + \hat{n}_{+,-}$ is:

$$P\left(x \geq \hat{x}\right) = \sum_{\{n_{+,+}, n_{-,-}, n_{-,+}, x \geq \hat{x}\}} \frac{\binom{K_D^+}{n_{+,+}}\binom{K_D^-}{x-n_{-,+}}\binom{K_D^0}{n_{+,0}}}{\binom{n}{K_H^+}} \frac{\binom{K_D^+ - n_{+,+}}{n_{-,+}}\binom{K_D^- - x + n_{-,+}}{n_{-,-}}\binom{K_D^0 - n_{+,0}}{n_{-,0}}}{\binom{n-K_H^+}{K_H^-}} \tag{5.9}$$

If thresholds for upregulation and downregulation are set to be the same for both normal and cancer tissues, we have to modify the previous formula. Let's consider the following quantities:

$$\begin{aligned} K^+ &= K_D^+ + K_H^+ \\ K^- &= K_D^- + K_H^- \\ K^0 &= K_D^0 + K_H^0 \\ 2n &= K^+ + K^- + K^0 \end{aligned}$$

where, $2n$ is the total number of tissues, which are paired in n couples. In this case, the null hypothesis is attained by assuming that n tissues are randomly selected to be pathological, and paired with the others, which are supposed to be the healthy ones. Therefore:

$$P\left(x \geq \hat{x}\right) = \sum_{Q} \frac{\binom{K^+}{K_D^+}\binom{K^-}{K_D^-}\binom{K^0}{K_D^0}}{\binom{2n}{n}} \frac{\binom{K_D^+}{n_{+,+}}\binom{K_D^-}{x-n_{-,+}}\binom{K_D^0}{n_{+,0}}}{\binom{n}{K_H^+}} \frac{\binom{K_D^+ - n_{+,+}}{n_{-,+}}\binom{K_D^- - x + n_{-,+}}{n_{-,-}}\binom{K_D^0 - n_{+,0}}{n_{-,0}}}{\binom{n-K_H^+}{K_H^-}} \tag{5.10}$$

where $Q = K_D^+, K_D^-, n_{+,+}, n_{-,-}, n_{-,+}$, such that $x \geq \hat{x}$. Therefore, in contrast with equation 5.9, the quantities K_D^+ and K_D^- are not fixed, and $\sum_Q$ explores all possible values under the constrain $K_D^+ + K_D^- + K_D^0 = n$.
In this manuscript, the Hy-test refers to equation 5.10. We use this test on a large set of genes, therefore a multiple comparison correction is required. Significant p-values are associated with differentially expressed genes.
To investigate p-value distribution, we have calculated Hy-test p-values on a randomized dataset obtained shuffling the expression profile of each gene in the breast cancer dataset. We found that p-values are not uniformly distributed, but none of those p-values is significant considering a Benjamini-Hochberg correction at 20% level. On the other hand, also the t-test p-values are not uniformly distributed because the gene expression distribution is not normal.

5.5 Quantitative analysis of GO-terms

The comparison between the Hy-test and the classical t-test has been done studying significant terms from a Gene Ontology (GO) enrichment analysis [216]; GO-Enrichment analysis has been carried out on two sets of significant

genes; one from the Hy-test and another from the t-test. From those two sets we have obtained two separated lists of significant GO-terms. GO-analysis has been done using topGO package from Bioconductor and focusing on biological process terms. Fisher exact p-values have been associated with GO-terms. Their significance have been evaluated considering the Bonferroni correction at level 0.05.

To identify GO-terms (e.g., *cell cycle*) conceptually associated with a specific cell line (b*reast cancer* in this analysis), we have defined a novel procedure that searches PubMed articles related to the biological concepts under exams; i.e., *breast cancer* and *cell cycle* in this example. A significant number of articles related to both those concepts indicates a conceptual association between them. The PubMed research has been carried out using the R package RISmed. Articles published between January 2000 and October 2020 have been considered.

The probability of observing $n_{C,T}$ PubMed articles with both keywords "*breast cancer*" and "*cell cycle*" is

$$Pr(N_{C,T} = n_{C,T}|N, N_C, N_T) = \frac{\binom{N_C}{n_{C,T}}\binom{N-N_C}{N_T-n_{C,T}}}{\binom{N}{N_T}} \tag{5.11}$$

where N is the number of articles with the keyword *breast*, N_C is the number of articles with both the keywords *breast* and *cancer*, N_T is the number of articles with both "*breast cancer*" and "*cell cycle*" as keywords.

Using an hypergeometric test we have associated to each term a p-value of conceptual association;

$$Pr(N_{C,T} \geq n_{C,T}) = 1 - \sum_{X=0}^{n_{C,T}-1} Pr(X|N, N_C, N_T) \tag{5.12}$$

Statistical significance has been evaluated by considering a Bonferroni correction at level 0.05.

5.6 Data analysis results

Data has been downloaded from the TCGA website. We have selected 14.569 genes recorded on 75 normal tissues and 75 in breast cancer tissues, as described in section 5.2. Log-fold change ($log(FC)$) has been calculated over cancer and normal tissues, and genes with $|log(FC)| \geq 1$ have been selected.

To identify differentially expressed genes, we used both the t-test and Hy-test.

We obtained 1.307 significant genes from Hy-test and 3.269 from the t-test (Fig.5.1). A GO-enrichment analysis has been carried out on those two lists of significant differentially expressed genes (biological process terms have been considered). 104 significant terms have been obtained from Hy-test significant genes and 213 from t-test significant genes. The associations of significant terms with *breast cancer* have been evaluated by researching PubMed papers as described in section 5.5. Terms that are significantly associated with *breast cancer* are reported in Tab.5.2 and Tab.5.3. Eight of those terms have been found by both procedures.
Finally, to validate our proposal, we carried on the whole analysis on 67 kidney renal carcinoma tissues paired with healthy tissues. The results are shown in the bottom part of Tab.5.1 and in Tab.5.4.

		DEA	**Enrichment analysis**	**PubMed research**
		Sign. genes	Enriched terms	Sign. associated terms
breast	Hy-test	1.307	104	16
	t-test	3.269	213	43
	Intersection	1.113	38	8
kidney	Hy-test	2.702	163	13
	t-test	3.988	366	29
	Intersection	2.165	149	12

Table 5.1: *Numbers of significant DE genes and terms found in each step of the analysis on breast and kidney tissues*

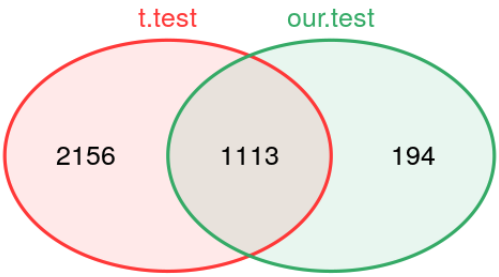


Figure 5.1: *Significant DE genes from the analysis on breast*

Figure 5.2: *Significant DE genes from the analysis on kidney*

Sign. GO-term	**Analysis**	**term size**	**BR term size**	**p-value**
angiogenesis	both	5765	5213	0
cell proliferation	both	21217	19745	0
tissue development	both	5586	4508	4.52e-09
cell migration	both	9799	9355	0
growth	both	49004	41783	0
cell motility	both	2235	2083	0
cell division	both	819	749	0
localization of cell	both	2319	2030	0
cell cycle checkpoint	Hy-test	503	470	0
mitotic cell cycle	Hy-test	453	403	2.58e-10
DNA replication	Hy-test	853	783	0
cell cycle	Hy-test	10893	10190	0
cell cycle process	Hy-test	482	446	0
cell cycle phase transition	Hy-test	296	278	1.55e-14
negative regulation of cell cycle	Hy-test	447	430	0
regulation of cell cycle	Hy-test	2479	2308	0

Table 5.2: *GO-terms significantly associated with "breast cancer" among significant GO-terms found using our procedure. GO-terms at the top of the table have been found also by the t-test procedure (found by both procedures). "term size" is the number of genes that compose a GO-term. "BR term size" is the number of GO-term genes associated with "breast cancer".*

Sign. GO-term	Analysis	term size	BR term size	p-value
anion transport	*t*-test	156	150	7.49e-11
chemotaxis	*t*-test	344	304	1.87e-07
cell communication	*t*-test	538	475	1.12e-10
cell adhesion	*t*-test	4248	3779	0
signal transduction	*t*-test	1881	1683	0
positive regulation of cell proliferation	*t*-test	654	619	0
response to hormone	*t*-test	3230	3049	0
regulation of signal transduction	*t*-test	416	376	5.58e-12
regulation of signaling receptor activity	*t*-test	636	596	0
regulation of hormone levels	*t*-test	446	393	7.61e-09
drug transport	*t*-test	986	927	0
negative regulation of angiogenesis	*t*-test	117	110	1.31e-06
biological adhesion	*t*-test	456	394	1.20e-06
regulation of signaling	*t*-test	4129	3850	0
signaling	*t*-test	19565	18204	0
cell differentiation	*t*-test	4969	4138	0
regulation of cell migration	*t*-test	1970	1889	0
positive regulation of cell migration	*t*-test	216	212	0
regulation of transporter activity	*t*-test	109	105	4.18e-08
regulation of localization	*t*-test	600	529	1.76e-11
regulation of cell proliferation	*t*-test	3985	3753	0
regulation of membrane potential	*t*-test	365	343	0
response to drug	*t*-test	4022	3804	0
protein kinase B signaling	*t*-test	501	466	0
regulation of cell differentiation	*t*-test	983	867	0
negative regulation of cell differentiation	*t*-test	159	150	7.17e-09
regulation of angiogenesis	*t*-test	795	728	0
cell development	*t*-test	14620	13316	0
regulation of inflammatory response	*t*-test	169	155	9.42e-07
regulation of biological process	*t*-test	115	108	1.93e-06
leukocyte migration	*t*-test	54	54	1.10e-06
regulation of transport	*t*-test	343	300	2.06e-06
biological regulation	*t*-test	1454	1349	0
regulation of molecular function	*t*-test	595	556	0
regulation of cell motility	*t*-test	477	450	0

Table 5.3: *GO-terms significantly associated with "breast cancer" among significant GO-terms found only by the t-test procedure.*

Sign. GO-term	**Analysis**	**term size**	**KIRC term size**	**p-value**
programmed cell death	*Hy*-test	423	127	0
angiogenesis	*t*-test	1515	470	0
kidney development	*t*-test	33777	2984	0
behavior	*t*-test	2286	290	0
negative regulation of cell proliferation	*t*-test	72	18	1.93e-06
tissue development	*t*-test	5028	493	2.17e-13
cell differentiation	*t*-test	3350	413	0
regulation of cell migration	*t*-test	316	94	0
B cell proliferation	*t*-test	577	77	7.66e-08
regulation of cell differentiation	*t*-test	586	88	2.55e-11
regulation of angiogenesis	*t*-test	230	69	0
cell development	*t*-test	9412	1641	0
gland development	*t*-test	475	77	1.04e-11
cell motility	*t*-test	423	95	0
epithelial cell proliferation	*t*-test	1374	180	1.88e-15
localization of cell	*t*-test	2031	211	2.07e-08
T cell migration	*t*-test	164	32	1.48e-07
regulation of cell motility	*t*-test	85	24	2.98e-09
cell activation	both	7629	768	0
cell killing	both	222	74	0
immune system development	both	646	87	7.58e-09
cell adhesion	both	2107	239	6.59e-13
cell proliferation	both	6239	1226	0
cell migration	both	1848	501	0
biological adhesion	both	183	40	1.33e-10
regulation of signaling	both	2541	241	2.87e-06
signaling	both	12173	1129	0
T cell proliferation	both	762	101	1.21e-09
regulation of cell proliferation	both	974	205	0
biological regulation	both	906	148	0

Table 5.4: *GO-terms significantly associated with "kidney cancer" among significant GO-terms found using our procedure and the t-test procedure.*

5.7 Discussion

With the advent of next generation sequencing (NGS) technologies, transcriptomic studies have gained a central role in almost all fields of biology and medicine. However, this approach still possesses several issues and biases, mainly derived from the diversity of biological samples, library preparation, sequencing platforms and bioinformatic analyses [217, 218, 219, 220, 221, 222, 223, 224, 225, 226], whose detailed description is beyond the aim of this paper. DEA plays a central role in comparative transcriptomic studies, that indeed represent the vast majority of gene expression analyses (with the notable exception of de novo assembly [235, 236, 237, 238], and few others [239, 240, 241]). The definition and thus the retrieval of genes that are differentially expressed in different conditions is the core action that define a transcriptomic comparative study. Working with data that are generated by a plethora of procedures in a very noisy and variable system such as a biological one is, pave the necessity to adopt different approaches to analyze the phenomena under investigation. Indeed, Hy-test can be adopted together with the canonical t-test to retrieve information that would be otherwise missed, as confirmed by the analyses on real data on breast and kidney cancers we present here.

Accordingly with literature data, the t-test, in comparison with our Hy-test, increases the number of significant genes retrieved from DEA [208] broadening the differential gene ontology enrichment. Hy-test is more selective both on retrieving DE genes and terms of GO, but Hy-test is not only able to narrow the window of selected genes, focusing the analysis, it is also able to retrieve specific terms of GO that would be otherwise missing from the subsequent analyses. This is particular evident in the breast cancer dataset where the vast majority of DE genes retrieved by Hy-test (85%) are also collected by t-test, but the enrichment analysis shows only a moderate overlapping (36%), Tab.5.1 strongly suggesting that Hy-test is indeed able to collect the same core DE genes, but it is also able to retrieve a different set of genes that points to functions of biological relevance that would be otherwise missed, as discussed in detail below. This is also true for the kidney dataset but with less evident differences. In that, about 80% of DE genes and 91% of enriched terms overlap between the two tests. However, even in kidney dataset, the Hy-test was able to pinpoint the "programmed cell death" GO term that would be otherwise missed, and indeed "programmed cell death" plays a central role in kidney cancer, as described in detail below.

5.7.1 Breast cancer

In the case of the real breast cancer profiles analyzed, both t-test and Hy-test reveal that DE genes are enriched in functions involved in tissue development, as expected [242, 243, 244, 245, 246, 247, 248, 249, 250, 242], and while only the t-test approach focuses in signal transduction [251, 252, 253], the Hy-test only highlights a central role of the regulation of cell cycle in breast cancer, as strongly supported by literature [254, 255, 256, 257, 258].
In details, the mammary gland is a tissue characterized by a high proliferation rate, and the developmental programs are prompt to be subverted to promote cancer progression. In the gland, many cells are extremely polarized, and when the maintenance of this organization is disrupted by extrinsic or intrinsic factors, this disruption may act as a promoter of hyperplasia and transformation [243]. Several studies suggest also that the disruption of the typical apical-basal polarity may even contribute to the metastatic event [248]. The deregulation of extracellular matrix proteins and signaling is sufficient to promote breast cancer development and progression [247]. Signal transduction has a central role in breast cancer; indeed, breast cancer molecular classification usually follows the presence or absence of specific hormone and growth factor receptors [259, 260] with direct implications in diagnosis, prognosis and therapy. Signal transduction pathways are of course cardinal in the maintenance of cancer clones and in the progression of the disease, such as the PI3K/Akt/mTOR pathway [261] and their inhibition has been evaluated for long as a potential therapeutic approach [262, 263, 264].
Both tissue development and signal transduction have a central role in breast cancer, but the t-test lacked to retrieve the cell intrinsic cell cycle deregulation GO terms that has been pinpointed by the Hy-test only. Indeed, cell cycle deregulation is crucial in breast cancer development and e.g. cell cycle control machinery is a target of novel therapeutic strategies such as CDK4/6 inhibitors [254, 255, 256, 257].

5.7.2 Kidney renal clear cell carcinoma

In the case of kidney cancers, the differences between the two approaches are even more straightforward. Both approaches retrieve an enrichment in cell signaling in particular in the contest of the immunological microenvironment [265, 266, 267, 268] (and the t-test only add a level in the involvement

of functions related to kidney development [269]), but Hy-test only focus on "programmed cell death" which is central in kindey cancer and even in therapeutic approaches to the diseases [266].
In details, it is known that the reshape of the metabolism is one of the key steps that kidney tumor cells must undergo during cancer progression and this event strongly relies on the cross-talk between the cancer cells and the tumor microenvironment [265]. In particular, the inflammatory microenvironment is involved in the development of pre-neoplastic alterations and the development of kidney cancer [270]. In the growing tumor, it has been reported a role of tumor associated macrophages (TAM) and tumor infiltrating neutrophils [271]. TAMs have a role in tumor progression, but are also an attractive therapeutic target in kidney cancer [272]. For patients with renal clear cell carcinoma has been even proposed a model based on few immune-related genes that can predicted the prognosis based on tumor immune microenvironments [273].
Interestingly, FDA recently approved therapies targeting the immunological checkpoint protein "programmed death 1 (PD-1)" for metastatic kidney cancer [266, 274]. Indeed, immunotherapies are expected to become the first line treatment option in kidney cancer in the near future [275, 276].
Kidney cancer cells must possess a way to work around programmed cell death, one of the main anticancer mechanism, that lead a precancerous cell to sense cellular and /or genomic damage and prompt it to commit suicide before the precancerous injuries can became a functional cancer commitment. Adopting an oversimplification, the vast majority of programmed cell death programs are orchestrated by p53 network [277, 278, 279].
The von Hippel-Lindau protein (pVHL) is mutated in the vast majority of clear cell renal carcinoma, the most common kidney cancer, and it has been implicated in the control of tumor suppression via the hypoxia inducible factor (HIF) pathway [280, 281]. the VHL-HIF axis is central in regulating apoptosis (the main form of programmed cell death) via several pathways, such as those mediated by BNIP3 [281]. The lack of pVHL activity protects renal cancer cells against mitochondria activated apoptosis [282]. pVHL even directly transactivate p53 [283] and is directly involved in the control of mitotic fidelity and in avoiding aneuploidy [284]; it has even reported that p53 and pVHL act synergistically in the regulation of cell proliferation and apoptosis in cell renal cell carcinoma [285].
Considering that the programmed cell death subversion plays a central role in

kidney cancer development, it is intriguing to ascertain that only the Hy-test lead to retrieve this GO term from the enrichment analysis, strongly suggesting that a dual approach that use both Hy-test and t-test can better describe the true meaning of a DEA on real data.

5.8 Conclusions

A novel exact statistical test for DEA has been defined and applied on real gene expression data. The study of differentially expressed genes through GO-Analysis is a standard procedure in molecular biology. Therefore, the enriched GO-terms have been identified. Moreover, we have presented a novel methodology for a more in-depth GO-analysis interpretation by calculating an association measurement between biological concepts (i.e., GO-terms, "*breast cancer*", and "kidney cancer").
Overall the results here presented strongly suggest that the application of Hy-test together with t-test can be useful to better understand the biological questions investigated by a DEA approach.

Future Researches

This section focuses on current and future researches strictly related to the analyzes presented in the present thesis.
The upcoming analyzes depend on empirical experiments that will be carried out soon (e.g., IP experiments on *Human* and other species). Meanwhile, we have already developed a theoretical framework for future analyzes. The following sections report a brief description of novel statistical models and algorithms.

Upgrade of ComiR web tool

In Chapter 3, we tested whether including coding region binding sites in ComiR algorithm improves its performance for predicting microRNA targets. The analysis focused on the *D. melanogaster* genome; databases with the currently available releases of mRNA and microRNA sequences have been used for ComiR upgrade. As a result, we find that ComiR algorithm trained with the information related to the coding regions is more efficient in predicting the microRNA targets, with respect to the algorithm trained with 3'UTR information.
On the other hand, we show that 3'UTR based predictions can be seen as complementary to the coding region based predictions, which suggests that both predictions, from 3'UTR and coding regions, should be considered in a comprehensive analysis. Furthermore, we observed that the lists of targets obtained by analyzing data from one experimental approach only, that is, inhibition or immunoprecipitation of AGO1, are not reliable enough to test the performance of our microRNA target prediction algorithm. Further analysis will be conducted to investigate the effectiveness of the tool with data from other species, provided that validated datasets, as obtained from the comparison of RISC proteins inhibition and immunoprecipitation experiments.

Moreover, in Chapter 4, we have compared miRNA binding sites predicted by four algorithms (indeed, the comparison of predicted targets is not enough because different algorithms could predict distinct miRNA binding sites located on the same mRNA strand). The analysis of SARS-CoV-2 genome shows that the strongest predictions refer to binding sites consistently predicted by different algorithms. For this reason, we believe that the identification of consistent predicted binding sites could improve ComiR prediction capacity. Therefore, this information will be included in the ComiR machine learning model, and its prediction capacity will be tested.
ComiR training set is currently composed of differentially expressed genes from empirical experiments (as described in Chapter 3). In Chapter 5, we presented a novel method for differential expression analysis; therefore, the classical approach for building ComiR training set will be compared with an approach that integrates the novel statistical test presented in Chapter 5. ComiR predictions will be compared to evaluate the best training and features over different species. Finally, the best ComiR model will be used for upgrading the ComiR web tool.

Analysis of miRNA-mRNA bipartite Networks

A bipartite system is a particular type of complex system composed of two sets of elements, where the elements that belong to different sets are qualitatively different from each other. Bipartite systems can be represented by constructing a bipartite network whose elements of one set only interact with the other set elements. In the present analysis, nodes correspond to miRNAs and mRNAs, and links represent their interactions.
We aim to use target prediction algorithms for building a large miRNA-mRNA bipartite Network. A common strategy for analyzing this kind of networks is the construction of a projected network [96]; it connects nodes with at least one neighbor in common. A cluster analysis on the projected network permits identifying gene groups targeted by the same miRNAs (clustering on gene projected set) and miRNA groups with similar binding behavior (clustering on miRNA projected set). Finally, a GO-analysis can characterize miRNA and gene groups providing a better understanding of miRNA regulatory process.
Projected networks contain several links that come from a random co-occurrence of neighbors in the original bipartite network. Tumminello *et al.* [97][98] in-

troduced a statistical approach to purify the projected network from links that occur by chance. A family of statistical tests is performed, and a p-value is associated with each link of the projected network. The significant links can not be explained in terms of random connectivity, and they compose the statistically validated projected network (projected SVN).
Let focus on the methodology to construct a projected SVN. Consider a Bipartite System **S** composed by set A and set B, we want to build the projected SVN on set A (this procedure is symmetric on set B). Assume that the elements i and j of set A have $N_{ij} = n_{ij}$ neighbors in common. Under the hypothesis of random connectivity, the probability that i and j have n_{ij} neighbors in common is

$$Pr(N_{ij} = n_{ij}|N_B, N_i, N_j) = \frac{\binom{N_i}{n_{ij}}\binom{N_B - N_i}{N_j - n_{ij}}}{\binom{N_B}{N_j}} \tag{5.13}$$

where N_B is the number of nodes in the set B, N_i and N_j are the degrees of i and j respectively.
Frequently, node degrees are very heterogeneous and follow a *scale-free* distribution; e.g., the genome of a unicellular organism contains the information about a small group of protein, whereas a complex organism has a genome that includes a vast number of proteins. In heterogeneity presence, the calculation of probability 5.13 is wrongly influenced by the degree of set B nodes. For this reason, Tumminello proposed to stratify the bipartite system according to the degree of set B. Each subsystem S_k consists of all the N_B^k set B elements with a given degree k and the set A elements linked with them. So, the probability that i and j have $N_{ij}^k = n_{ij}^k$ neighbors in common in set B^k is

$$Pr(N_{ij}^k = n_{ij}^k|N_B^k, N_i^k, N_j^k) = \frac{\binom{N_i^k}{n_{ij}^k}\binom{N_B^k - N_i^k}{N_j - n_{ij}^k}}{\binom{N_B^k}{N_j^k}} \tag{5.14}$$

where, N_B^k is the number of nodes in the set B^k, N_i^k and N_j^k are the degrees of i and j in the subsystem S^k respectively.
So we can associate a p-value to the number of neighbors in common $N_{ij}^k = n_{ij}^k$ for each node couple i-j:

$$
\begin{aligned}
Pr(N_{ij}^k \geq n_{ij}^k) &= \sum_{X=n_{ij}^k}^{min(N_j^k, N_i^k)} Pr(X|N_B^k, N_i^k, N_j^k) \\
&= 1 - \sum_{X=0}^{n_{ij}^k - 1} Pr(X|N_B^k, N_i^k, N_j^k)
\end{aligned}
\tag{5.15}
$$

Significant p-values identify the validated links of the projected network related to the subsystem S_k. The projected SVN of the whole system $\mathbf{S}$ is obtained linking node couples with at least one validated link over all the subsystems. The unified projected SVN is weighted by associating each link with the number of subsystems in which the link of interest has been validated.

Of course, this procedure is affected by a problem of multiple comparisons. Therefore, the p-values have to be corrected using an appropriate threshold. The number of multiple comparisons could be much higher than $N_A(N_A-1)/2$; indeed, in each subsystem, we have to perform a statistical test for each couple of nodes of set A that have at least one neighbor in common. So if the degree of elements if set B is between k_{max}^B and k_{min}^B, the total number of tests is $m \leq (k_{max}^B - k_{min}^B)N_A(N_A-1)/2$.

An alternative approach to solving the heterogeneity problem considers a biased urn model based on Wallenius distribution [99]. The main difference in respect to the hypergeometric model is that urn balls have a different probability of being picked. Wallenius distribution allows to calculate exact p-values without the expedient of the system stratification, but this approach is costly from a computational point of view.

MicroRNA-mRNA rewiring network

MicroRNA binding behavior is extremely complex; indeed, many elements influence molecule binding. Binding score calculation provides an idea of miRNA binding propensity, but those scores don't consider how molecules influence each other. We propose a novel network model for simulating miRNA binding, taking into account the whole connectivity of the system.

MicroRNA binding depends not only on the accessibility of mRNA binding sites but also on the total amount of miRNAs and mRNA in the cell. Indeed, the amount of a single miRNA can influence the binding of the other miRNAs by acting as a competitor [201] (the transcript amount is quantified through

miRNA and mRNA expression levels).
Therefore, in our model, molecule amounts represent the initial system condition. Starting with this information, each simulation explores the binding behavior by considering miRNA binding sites located on mRNAs. Each bound that occurs during a simulation reduces the total molecule amount. The simulation ends when the miRNA (or mRNA) amount is all involved in the simulated bonds.
Our simulation algorithm is based on a network approach conceptually closed to *degree-preserving rewiring* [202]; miRNA-mRNA binding can be represented through a bipartite network in which miRNAs and mRNAs correspond to nodes, and links represent their interactions. In the network, node degrees are equal to miRNA and mRNA expression levels. In this way, node degrees reflect the molecule amounts.
The algorithm is based on the following steps:

1. B miRNAs and B mRNAs are extracted from the transcript amount

2. The two groups extracted in step 1. are randomly matched to form the candidate links that could be included in the network. Each link is associated with its binding score.

3. The links selected in step 2. are randomly extracted with repetition (the selection probability is proportional to the binding score). The links that have been extracted in this step are added to the network.

4. Node degrees are updated by subtracting miRNA and mRNA amounts involved in the links selected in step 3.

5. The algorithm stops if miRNA (or mRNA) amounts are all allocated in the weighted links. Otherwise, go to step 1.

We have already implemented the rewiring network algorithm on the R software environment. Preliminary results show a good computational efficiency on large networks. Moreover, the rewiring network model allows to evaluate the effects of *in silico* mutations on miRNA binding.

Conclusions

The present thesis offers a view of miRNA binding prediction methodologies. Among them, we propose ComiR algorithm as a reliable tool for miRNA target prediction, and an improvement of its prediction capacity is deeply discussed. The ComiR upgrade has been strongly supported by the results presented in Chapter 2. In this work, we analyzed the overexpressed genes in the anti-AGO2 and anti-GW182 RIP samples vs the respective FT samples, and we revealed different features characterizing the enriched genes in the two data sets. In particular, both AGO2/GW182-associated mRNAs are characterized by miRNA binding sites located on the coding regions. Indeed, we found that coding region information significantly improves the prediction capacity of mRNA targets. AGO2-associated mRNAs are characterized by a high number of binding sites in the coding region for top expressed miRNAs and by a high density of binding sites in the 3'UTR region. On the other hand, GW182-associated mRNAs are characterized by long coding regions. Therefore, those proteins play different roles in the RISC machinery activity.

Starting from those results, we have upgraded the ComiR algorithm by considering coding region information as described in Chapter 3. Our results indicate that binding sites predicted in coding regions are valuable information to efficiently predict the functional targets of a set of miRNAs by their integration in the ComiR algorithm framework.

In Chapter 4, the novel ComiR version has been successfully used to predict human miRNAs that potentially bind the SARS-CoV-2 genome. Our results support the idea that five predicted miRNAs interact with SARS-CoV-2 strands. Moreover, a sequencing analysis shows that SARS-CoV-2 strands are strongly stable in the regions where miRNA binding sites are located.

We propose the involvement of miR-1207-5p family and CSF1 gene in the progression of COVID-19 disease, and as possible targets for COVID-19 treatment. Further experimental validation is still due to confirm the binding of

miR-1207-5p into the viral genome and the role of SARS-CoV-2 in the regulation of CSF1 expression.

Comparing different miRNA target prediction algorithms on SARS-CoV-2 highlights the high capacity of the novel ComiR algorithm to find specific miRNA binding sites. Although those results strongly support the novel ComiR version as a reliable tool, we currently aim at finding the best way to combine the two scores obtained by training the SVM with the 3'UTR and the coding region separately. Further analysis will be conducted to analyze data from other species by using positive and negative sets of miRNA targets obtained by comparing results from both RISC proteins inhibition and immunoprecipitation. Therefore, upcoming experiments will provide novel features for developing ComiR machine learning models.

Another aspect of ComiR algorithm that will be investigated in our future research is the training set construction; the novel statistical test presented in Chapter 5, named Hy-test, could be used to identify DE genes that compose the ComiR training set. Hy-test can be adopted together with the canonical t-test to retrieve information that would be otherwise missed, as confirmed by the analyses on real data on breast and kidney cancer tissues presented in Chapter 5. Hy-test is more selective on retrieving DE genes, but Hy-test is not only able to narrow the whole window of DE genes. The results suggest that the application of Hy-test together with t-test can be useful for clearer identification of DE genes.

An overall view of the results presented in this thesis strongly supports an improvement of ComiR web tool. In particular, coding region information has a high miRNA target prediction capacity and significantly improves the performance of ComiR algorithm. Moreover, binding sites coherently predicted by different algorithms appear highly specific and could produce an additional improvement of ComiR prediction capacity. Finally, we currently aim to extend those results to other species by performing new experiments.

List of Figures

List of Tables

Acronyms

3'UTR: Three prime untranslated region

AGO: Argonaute protein family

AUC: Area under the (ROC) curve

Anti-AGO2/Anti-GW182: Antibodies of AGO2/GW182

BS: Binding site

CDS: Coding region sequence

ComiR: Combinatorial miRNA targeting

DE gene: Differentially expressed gene

DEA: Differential expression analysis

Dme: Drosophila melanogaster

ECDF: Empirical Cumulative Function Distribution

FC: Fold-change

FD: Fermi-Dirac

FDR: False discovery rate

FT sample: Flow-through sample

FWER: Family wise error rate

GO-Analysis: Gene Ontology enrichment analysis

GO: Gene Ontology

IN sample: Input sample

IP: Immunoprecipitation

LOOCV: Leave One Out Cross Validation

MCP: Multiple comparison procedure

NGS: Next generation sequencing

RIP − Chip: RIP coupled to microarray

RIP: RNA-binding protein immunoprecipitation

RISC: RNA-induced silencing complex

RNAi: RNA interference

RT-PCR: Reverse transcript polimerase chain reaction

RT: Reverse transcription

SARS-Cov-2: Severe acute respiratory syndrome 2

SVM: Support vector machine

SVN: Statistically validated network

WS: Weighed-sum

cDNA: Complementary DNA

mRNA: Messenger RNA

miRNA: microRNA

Glossary

AGO (Argonaute) Protein family that is an essential component of the RISC. Argonaute family includes AGO1 and AGO2 proteins.

AUC (Area under the ROC curve) Index used to compare ROC curves.

Antibody - *about IP experiments* - molecule that specifically binds to a particular protein to be immunoprecipitated.

Binding site (BS) Region on a molecule that specifically binds to another molecule.

CSF1 Gene that encodes a cytokine differentiation, and function of macrophages.

Coding region (CDS) Portion of a gene's DNA or RNA that codes for protein synthesis.

Cross validation (CV) Techniques for assessing how the results of a machine learning model depends on the empirical data used for training.

DeLong's test Statistical test used for comparing AUC values.

Differential expression analysis (DEA) Analysis of expression profiles over two groups of samples to identify over-expressed (*enriched*) and under-expressed (*underrepresented*) genes.

Enrichment 1. In differential expression analysis; a significant increased presence or expression of a given RNA. *Enriched* is often used as a synonym for *over-expressed*. **2.** In GO-analysis: Identification of a group of genes that enriches a previous GO classification.

Flow-through (FT) RNA extracted from the residual fraction of cells lysate, after IP experiment.

GW182 (TNRC6) Protein that

is part of the RISC.

Gene ontology (GO) Bioinformatics project that classify genes in *terms* that represent gene products.

High throughput Techniques that include genome sequencing, transcriptomics, and other genome-related microarray measurements such as chip-on-chip. The term *high throughput* indicates the velocity of data processing and the high quality of this data.

Immunoprecipitation (IP) Technique of precipitating a protein antigen out of solution using an antibody that specifically binds to that particular protein. This process can be used to isolate a particular protein and its binding molecules (e.g., AGO1 and binding mRNAs) from a sample.

In Silico Biological experiment performed on computer or via computer simulation.

Input sample (IN) RNA extracted from whole cells lysate.

Macrophages A large cell, part of the body's immune system, that can ingest pathogenic microorganisms such as bacteria and virus.

Messenger RNA (mRNA) Molecule responsible for carrying the genetic code transcribed from DNA to specialized sites within the cell (known as ribosomes), where the information is translated into protein composition.

MicroRNA (miRNA) A small RNA molecule that is encoded by a cell and can 'silence' the expression of a particular target gene within the cell. miRNAs bind to target messenger RNA (mRNA) molecules and suppress translation of the mRNA into protein.

Microarrays Technique used to quantify gene expression by determining the total output of messenger RNAs.

Nucleotide Primary component of genetic material. DNA and RNA are made up of long chains of nucleotides. Specifically, a nucleotide is an organic compound consisting of a nitrogen-containing purine or pyrimidine base linked to a sugar (ribose or deoxyribose) and a phosphate group.

P-bodies (Processing bodies) Protein granules primarily composed of translationally re-

pressed mRNAs and proteins related to mRNA decay. Their formation takes place during the post-transcriptional regulation related to miRNAs.

RIP-chip (RNA immunoprecipitation chip) Technique which combines RNA immunoprecipitation with a microarray expression measurement.

RISC (RNA-induced silencing complex) Protein complex that recognize complementary messenger RNAs by incorporating a miRNA strand. It has a central role in miRNA regulation activity.

RNA (ribonucleic acid) Complex organic compound in living cells that is concerned with protein synthesis. In some viruses, RNA is also the hereditary material. Most RNA is synthesized in the nucleus and then distributed to various parts of the cytoplasm. An RNA molecule consists of a long chain of nucleotides.

ROC curve Graphical plot that illustrates the diagnostic ability of a classifier. Classification ability is evaluated by varying a threshold on the classification score.

Spillover Event that occurs when a pathogen (e.g., virus) comes into contact with a novel host population.

Strand A single long chain of nucleotides.

Support vector machine (SVM) Supervised learning model used for classification (in this thesis, SVM is used to classify miRNA targets and non-targets).

Three prime untranslated region (3′-UTR) Section of messenger RNA (mRNA) that immediately follows the translation termination codon. In this region is located most of the miRNA binding sites.

Training set Empirical information used for training a machine learning model.

Transcription The process in living cells in which the genetic information of DNA is transferred to a molecule of messenger RNA (mRNA) as the first step in protein synthesis.

Transcriptome The full complement of RNA transcripts of the genes of a cell or organism. It includes messenger RNAs and non-coding RNA (e.g., miRNAs).

Underrepresented In differential expression analysis; synonym for *under-expressed*.

Wilcoxon test Statistical test used for comparing Empirical cumulative function distributions (ECFD).

Author contributions

Excluding the first chapter, each chapter of the thesis have been obtained starting from a scientific paper. Therefore, different authors gave important contributions to write the papers on which the thesis is based. The following items highlight the main contributions:

Chapter 1 Giorgio Bertolazzi[1] reviewed the methodologies related to miRNA target prediction and wrote the chapter supervised by Claudia Coronnello[2] and Michele Tumminello[1].

Chapter 2 Giorgio Bertolazzi and Michele Tumminello performed statistical data analysis. Giovanni Perconti[3] defined and optimized the RIP experimental protocols and performed the RIP experiments. Patrizia Rubino[3] performed the RNA extractions, RT and PCR. Flavia Contino[4] and Serena Bivona[4,5] performed the microarray experiments. Salvatore Feo[4,5] provided intellectual insight into the microarray experimental design. Agata Giallongo[3] provided oversight and guidance throughout the project. Claudia Coronnello conceived the statistical methods, performed the analyses and wrote the manuscript.

Chapter 3 Giorgio Bertolazzi collected and analyzed the data, supervised by Michele Tumminello and Panayiotis V. Benos[6]. Claudia Coronnello provided

[1] Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy

[2] Fondazione Ri.MED, Palermo, Italy

[3] Istituto di Biomedicina ed Immunologia Molecolare (IBIM) CNR, Palermo, Italy

[4] Dipartimento di Scienze e Tecnologie Biologiche Chimiche e Farmaceutiche, Università degli Studi di Palermo, Palermo, Italy

[5] ATEN Center, Università degli Studi di Palermo, Palermo, Italy

[6] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, USA

oversight and guidance throughout the project and wrote the manuscript.

Chapter 4 Giorgio Bertolazzi performed statistical data analysis and sequencing analysis. Chiara Cipollina[2,7] and Claudia Coronnello conceptualized the biological aspects and methods. Claudia Coronnello provided the sequencing data. Michele Tumminello and Claudia Coronnello conceived the statistical methodology. All authors made their contribution in writing the manuscript.

Chapter 5 Giorgio Bertolazzi and Gianluca Sottile[1] performed statistical data analysis. Michele Tumminello formulated the novel statistical test and conceived the methodology. Walter Arancio[2] elaborated on biological aspects. Claudia Coronnello provided the gene expression data and supervised the analysis.

The whole thesis represent the synthesis of Giorgio Bertolazzi's Ph.D. training and research supervised by Michele Tumminello, Claudia Coronnello and Panayiotis V. Benos.

[7] Institute for Biomedical Research and Innovation, National Research Council, Palermo, Italy

Acknowledgements

I would like to thank my supervisor Michele for his consistent support, encouragement, and contagious enthusiasm. I'd also like to acknowledge my supervisors Claudia for her guidance in bioinformatics issues, and Prof. Takis Benos for his hospitality and guidance during my period abroad.
Finally, I'd like to thank my Ph.D. colleagues and professors with whom I have shared a very good time in the lab; in particular, I am grateful to Prof. Consiglio for his advice.

Bibliography

[1] Tufekci, Meuwissen, Genç (2014) *The Role of microRNAs in Biological Processes*, Methods Mol Biol. 1107:15-31. doi: 10.1007/978-1-62703-748-8_2

[2] Erson, Petty (2008) *MicroRNAs in development and disease.* Clin Genet 74:296–306

[3] Perconti, Rubino, Flavia Contino, Bivona, Bertolazzi, Tumminello, Feo, Giallongo, Coronnello (2017) *RIP-Chip analysis supports different roles for AGO2 and GW182 proteins in recruiting and processing microRNA targets*, BMC bioinformatics 20(Suppl 4):120

[4] Bertolazzi, Benos, Tumminello, Coronnello (2020) *An Improvement of ComiR Algorithm by Exploiting mRNA Coding Regions*, BMC Bioinformatics, 21(Suppl 8):201

[5] Giorgio Bertolazzi, Chiara Cipollina, Panayiotis V. Benos, Michele Tumminello, Claudia Coronnello (2020) *miR-1207-5p Can Contribute to Dysregulation of Inflammatory Response in COVID-19 via Targeting SARS-CoV-2 RNA*, Frontiers in Cellular and Infection Microbiology, Vol. 10, Article 586592

[6] Kvam VM, Liu P, Si Y (2011) *A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data.* Am J Bot 99(2):248–256

[7] Wightman B, Ha I, Ruvkun G (1993) *Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans*, Cell 75: 855–862

[8] Shabalina SA, Koonin EV (2008) *Origins and evolution of eukaryotic RNA interference*, Trends Ecol Evol 23:578–587

[9] Erson-Bensan (2013) *Introduction to MicroRNAs in Biological Systems*, part of the Methods in Molecular Biology book series (MIMB, volume 1107)

[10] Thierry Bardini (2010) *Junkware*, University of Minnesota press

[11] Wikipedia, DNA replication, $https://en.wikipedia.org/wiki/DNA_replication$

[12] Scitable nature education, $https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/$

[13] Palazzo, Gregory (2014) *The Case for Junk DNA*, PLOS Genetics, $https://doi.org/10.1371/journal.pgen.1004351$

[14] Petrova, Zenkova, Chernolovskaya (2012) Structure - Functions Relations in Small Interfering RNAs, DOI: 10.5772/53945

[15] Ha TY (2011) *MicroRNAs in human diseases: from cancer to cardiovascular disease.* Immune Netw 11:135–154

[16] Brittis PA, Lu Q, Flanagan JG (2002) *Axonal protein synthesis provides a mechanism for localized regulation at an intermediate target.* Cell 110:223–235

[17] Lee CT, Risom T, Strauss WM (2006) *MicroRNAs in mammalian development.* Birth Defects Res C Embryo Today 78:129–139

[18] O'Connell RM, Taganov KD, Boldin MP *et al* (2007) *MicroRNA-155 is induced during the macrophage infl ammatory response.* Proc Natl Acad Sci U S A 104:1604–1609

[19] Xu *et al.* (2018) *Exosome[U+2010]encapsulated miR[U+2010]6089 regulates inflammatory response via targeting TLR4*, ORIGINAL RESEARCH ARTICLE, DOI: 10.1002/jcp.27014

[20] Yan *et al.* (2017) *MicroRNA-6869-5p acts as a tumor suppressor via targeting TLR4/NF-B signaling pathway in colorectal cancer*, ORIGINAL RESEARCH ARTICLE DOI: 10.1002/jcp.26316

[21] Bruscella P, Bottini S, Baudesson C, Pawlotsky JM, Feray C, Trabucchi M. Viruses and miRNAs: More Friends than Foes. Front Microbiol. 2017;8:824. Published 2017 May 15. doi:10.3389/fmicb.2017.00824

[22] Lynam[U+2010]Lennon, Maher, Reynolds (2009) *The roles of microRNA in cancer and apoptosis*, Biological Reviews, Volume84, Issue1.

[23] Wang *et al.* (2008) *The Endothelial-Specific MicroRNA miR-126 Governs Vascular Integrity and Angiogenesis*, Developmental Cell Volume 15, Issue 2, 12 August 2008, Pages 261-271

[24] Munker R, Calin GA (2011) *MicroRNA profiling in cancer.* Clin Sci (Lond) 121:141–158

[25] Furer V, Greenberg JD, Attur M *et al* (2010) *The role of microRNA in rheumatoid arthritis and other autoimmune diseases.* Clin Immunol 136:1–15

[26] Reid G, Kirschner MB, van Zandwijk N (2011) *Circulating microRNAs: association with disease and potential use as biomarkers.* Crit Rev Oncol Hematol 80:193–208

[27] Bernardo, Ooi, Lin McMullen (2015) *miRNA therapeutics: a new class of drugs with potential therapeutic applications in the heart*, FUTURE MEDICINAL CHEMISTRYVOL. 7, NO. 13.

[28] Soifer, Rossi, PålSaetrom (2007) *MicroRNAs in Disease and Potential Therapeutic Applications*, Molecular Therapy Volume 15, Issue 12

[29] Pedersen et al. (2007) *Interferon modulation of cellular microRNAs as an antiviral mechanism*, Nature volume 449, pages 919–922

[30] Yates *et al.* (2020) *Ensembl 2020.* PubMed, doi:10.1093/nar/gkz966

[31] Durinck *et al.* (2005) *BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis*, Bioinformatics, Volume 21, Issue 16

[32] Geer *et al.* (2010) *The NCBI BioSystems database.* Nucleic Acids Res. PubMed PMID: 19854944

[33] Griffiths-Jones, Grocock, Dongen, Bateman, Enright (2006) *miRBase: microRNA sequences, targets and gene nomenclature*, Nucleic Acids Research, Volume 34, Issue suppl1

[34] Pratt, MacRae (2009) *The RNA-induced Silencing Complex: A Versatile Gene-silencing* JOURNAL OF BIOLOGICAL CHEMISTRY Vol 284,numb 27

[35] Wook Chi, Zang, Mele, Darnell (2009) *Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps*, NATURE | Vol 460

[36] Hendrickson, Hogan, Herschlag, Ferrell, Brown (2008) *Systematic Identification of mRNAs Recruited to Argonaute 2 by Specific microRNAs and Corresponding Changes in Transcript Abundance*, PLOS one

[37] Wang, El Naqa (2008) *Prediction of both conserved and nonconserved microRNA targets in animals.* Bioinformatics 24: 325–332.

[38] Yousef, Jung, Kossenkov, Showe, Showe (2007) *Nave Bayes for microRNA target predictionsmachine learning for microRNA targets*, Bioinformatics 23: 2987–2992.

[39] Friedman, Farh, Burge, Bartel (2009) *Most mammalian mRNAs are conserved targets of microRNAs*, Genome Res 2009, 19: 92–105.

[40] Huang, Babak, Corson, Chua, Khan, et al (2007) *Using expression profiling data to identify human microRNA targets*, Nat Methods 4:1045–1049.

[41] Krek, Grun, Poy, Wolf, Rosenberg, et al. (2005) *Combinatorial microRNA target predictions.* Nature Genetics 37: 495–500.

[42] Muniategui Nogales-Cadenas, Vazquez, Aranguren, Agirre, *et al.* (2012) *Quantification of miRNA-mRNA interactions*, PLoS ONE 7: e30766.

[43] Coronnello, Hartmaier, Arora, Huleihel, Pandit, Bais, Butterworth, Kaminski, Stormo, Oesterreich, Benos (2012) *Novel Modeling of Combinatorial miRNA Targeting Identifies SNP with Potential Role in Bone Density*, PLoS Comp Bio 8:12-e1002830

[44] Coronnello, Benos (2013) *ComiR: combinatorial microRNA target prediction tool*, Nucleic Acids Research 41:W159-W164

[45] Lewis, Burge, Bartel (2005) *Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets*, Cell, Vol. 120, 15–20

[46] Agarwal, Bell, Nam, Bartel (2014) *Predicting effective microRNA target sites in mammalian mRNAs*, eLife, DOI: 10.7554/eLife.05005

[47] TargetScan web site: *http://www.targetscan.org/vert_72/*

[48] Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. (2003) *MicroRNA targets in drosophila.* Genome Biol. 5(1):R1.

[49] Kertesz, Iovino2, Unnerstall, Gaul, Segal (2007) *The role of site accessibility in microRNA target recognition*,nature genetics, doi:10.1038/ng2135

[50] Winterbach, Mieghem, Reinders, Wang, de Ridder (2013) *Topology of molecular interaction networks* BMC System Biology

[51] Cusick, Klitgard, Vidal, Hill (2005) *Interactome: Gateway into systems biology*, Hum Mol Genet 14:R171–181.

[52] Carter, Brechbühler, Griffin, Bond (2004) *Gene co-expression network topology provides a framework for molecular characterization of cellular state*, Bioinformatics

[53] Batushansky, Toubiana, Fait (2016),

[54] Jeong, Tombor, Albert, Oltvai, Barabási (2000) *The large-scale organization of metabolic networks.* Nature 407(6804):651–654.

[55] Karlebach, Shamir (2008) *Modelling and analysis of gene regulatory networks*, Nature Reviews Molecular Cell Biology volume 9, pages770–780(2008)

[56] Schwikowski, Uetz, Fields (2000) *A network of protein–protein interactions in yeast*, Nature Biotechnology volume 18, pages1257–1261

[57] Normand and Yanai *An Introduction to High-Throughput Sequencing Experiments: Design and Bioinformatics Analysis*, Part of the Methods in Molecular Biology book series (MIMB, volume 1038)

[58] Chen C, Ridzon DA, Broomer AJ *et al* (2005) *Real-time quantifi cation of microRNAs by stemloop RT-PCR.* Nucleic Acids Res 33(20):e179

[59] Liu CG, Calin GA, Volinia S, Croce CM *et al* (2008) *MicroRNA expression profi ling using microarrays.* Nat Protoc 3(4):563–578

[60] Silver, Ritchie, Smyth (2009) *Microarray background correction: maximum likelihood estimation for the normal–exponential convolution*, Biostatistics, Volume 10, Issue 2

[61] Hafner M, Landgraf P, Ludwig J *et al* (2008) *Identifi cation of microRNAs and other small regulatory RNAs using cDNA library sequencing.* Methods 44(1):3–12

[62] Nagalakshmi U, Wang Z, Waern K *et al* (2008) *The transcriptional landscape of the yeast genome defi ned by RNA sequencing.* Science 320(5881):1344–1349

[63] Bala Gür Dedeoğlu (2014) *High-throughput Approaches for microRNA Expression Analysis*, Methods Mol Biol. 2014;1107:91-103. doi: 10.1007/978-1-62703-748-8_6.

[64] Immage taken from: *https : //commons.wikimedia.org/wiki/File : Microarray – schema.gif*

[65] Tusher VG, Tibshirani R, Chu G. (2011) *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A. 98(9):5116–21.

[66] Khan, Betel, Miller, Sander, Leslie, Marks (2009) *Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs*, Nature Biotechnology volume 27, pages549–555

[67] Tan LP, Seinen E, Duns G, de Jong D, Sibon OC, Poppema S, *et al.* (2009) *A high throughput experimental approach to identify miRNA targets in human cells.* Nucleic Acids Res.37(20):e137.

[68] Karginov FV, Conaco C, Xuan Z, Schmidt BH, Parker JS, Mandel G, *et al.* (2007) *A biochemical approach to identifying microRNA targets.* Proc Natl Acad Sci USA, 104(49):19291–6.

[69] Burroughs AM, Ando Y, de Hoon MJ, Tomaru Y, Suzuki H, Hayashizaki Y, *et al.* (2011) *Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin.* RNA Biol. 8(1):158–77.

[70] Fan M, Krutilina R, Sun J, Sethuraman A, Yang CH, Wu ZH, *et al.* (2013) *Comprehensive analysis of miRNA) targets in breast cancer cells.* J Biol Chem. 288(38):27480–93.

[71] Landthaler M, Gaidatzis D, Rothballer A, Chen PY, Soll SJ, Dinic L, et al. (2008) *Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs.* RNA. 14(12):2580–96.

[72] Yamagishi M, Katano H, Hishima T, Shimoyama T, Ota Y, Nakano K, et al. (2015) *Coordinated loss of microRNA group causes defenseless signaling in malignant lymphoma.* Sci Rep. 5:17868.

[73] Hauptmann J, Schraivogel D, Bruckmann A, Manickavel S, Jakob L, Eichner N, et al. (2015) *Biochemical isolation of Argonaute protein complexes by ago-APP.* Proc Natl Acad Sci U S A. 2015;112(38):11841–5

[74] Eulalio A, Rehwinkel J, Stricker M, Huntzinger E, Yang SF, et al. (2007) *Targetspecific requirements for enhancers of decapping in miRNA-mediated gene silencing.* Genes Development 21: 2558–2570.

[75] Bellman RE (1961) *Adaptive control processes: a guided tour*. Princeton University Press, Princeton, NJ

[76] Goldstein, Morris, Yena (2004) *Problems with fitting to the power-law distribution*, Eur. Phys. J. B 41, 255–258

[77] Kasper D. Hansen, Rafael A. Irizarry, Zhijin WU (2012) *Removing technical variability in RNA-seq data using conditional quantile normalization*, Biostatistics, Volume 13, Issue 2, 204–216

[78] Cui X, Churchill GA (2003) *Statistical tests for differential expression in cDNA microarray experiments*. Genome Biol 4(210):1–10

[79] Cui and Churchill (2003) *Statistical tests for differential expression in cDNA microarray experiments*, Genome Biology volume 4, Article number: 210 (2003)

[80] Pan (2002) *A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments*, BIOINFORMATICS Vol. 18 no. 4, Pages 546–554

[81] Morten W.Fagerland and LeivSandvik (2009) *Performance of five two-sample location tests for skewed distributions with unequal variances*, Contemporary Clinical Trials Volume 30, Issue 5, September 2009, Pages 490-496

[82] Gordon K. Smyth (2004) *Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments*, Statistical Applications in Genetics and Molecular Biology, Volume 3, Issue 1

[83] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu (2001) *Significance analysis of microarrays applied to the ionizing radiation response*. PNAS,vol. 98,no. 9,5116–5121

[84] Davis J. McCarthy, Gordon K. Smyth (2009) *Testing significance relative to a fold-change threshold is a TREAT*, Bioinformatics, Volume 25, Issue 6, Pages 765–771

[85] Zararsiz G, Elmali F, Ozturk A (2012) *Bagging support vector machines for leukemia classifications*. Int J Comput Sci 9(6):355–358

[86] Maxwell W. Libbrecht William Stafford Noble (2015) *Machine learning applications in genetics and genomics*, Nature Reviews Genetics volume 16, pages321–332

[87] Newman (2010) *Networks: an introduction*, Oxford University Press

[88] Yuan, Lin (2007) *Model selection and estimation in the Gaussian graphical model*, Biometrika pp. 1–17

[89] Perrin, Ralaivola, Mazurie, Bottani, Mallet, d'Alché–Buc (2003) *Gene networks inference using dynamic Bayesian networks*, Bioinformatics, Volume 19, Issue suppl 2, 27

[90] Akutsu, Miyano (1998) *Identification of genetic network from a small number of gene expression patterna under the boolean network model*, biocomputing'99, https://doi.org/10.1142/9789814447300_0003

[91] Faure, Naldi, Chaouiya, Thieffry (2006) *Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle*, Bioinformatics Vol. 22

[92] Contino, Bertolazzi, Calì, Cantone, Vera-González, Romano (2020) *Boolean Networks: A Primer*, Systems Medicine; Integrative, Qualitative and Computational Approaches Volume 2, Pages 41-53

[93] Quatto, Margaritella, Costantini, Baglio, Garegnani, Nemni, Pugnetti (2019) *Brain networks construction using Bayes FDR and average power function*, Statistical Methods in Medical Research.

[94] Reverter, Chan (2008) *Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks*, BIOINFORMATICS ORIGINAL PAPER Vol. 24

[95] Kenett, Tumminello, Madi, Gur-Gershgoren, Mantegna, Ben-Jacob (2010) *Dominating Clasp of the Financial Sector Revealed by Partial Correlation Analysis of the Stock Market*, PLOS one

[96] Zhou, Ren, Medo, Zhang (2007) *Bipartite network projection and personal recommendation*, PHYSICAL REVIEW E 76

[97] Tumminello, Micciché, Lillo, Piilo, Mantegna (2011) *Statistically Validated Networks in Bipartite Complex Systems*, PLOS one

[98] Tumminello, Edling, Liljeros, Mantegna , Sarnecki (2013) *The Phenomenology of Specialization of Criminal Suspects*, PLOS one

[99] Puccio, Vassallo, Piilo, Tumminello (2018) *Covariance and correlation estimators in bipartite complex systems with a double heterogeneity* Journal of Statistical Mechanics: Theory and Experiment

[100] Eisen, Spellman, Brown, Botstein (1998) *Cluster analysis and display of genome-wide expression patterns*, Proceedings of the National Academy of Sciences of the United States of America, Volume 95, Issue 25, Pages 14863-14868

[101] Kodinariya, Makwana (2013) *Review on determining number of Cluster in K-Means Clustering*, International Journal of Advance Research in Computer Science and Management Studies

[102] Defays (1997) *An efficient algorithm for a complete link method*, The Computer Journal, Volume 20, Issue 4, Pages 364–366

[103] Newman (2006) *Modularity and community structure in networks*, PNAS

[104] Edler, Lancichinetti, Rosvall *Community detection and visualization of networkswith the map equation frameworkLudvig Bohlin*, Measuring scholarly impact: theory and practice, Springer.

[105] Zheng, Wang (2008) *GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis*, Nucleic Acids Research, Volume 36

[106] Ashburner, Ball, Blake, Botstein, Butler, Cherry, Davis, Dolinski, Dwight, Eppig, Harris, Hill, Issel-Tarver, Kasarskis, Lewis, Matese, Richardson, Ringwald, Rubin, Sherlock (2000) *Gene ontology: tool for the unification of biology*, Nat. Genet. 25 25

[107] Fabregat et al. (2017) *Reactome pathway analysis: a high-performance in-memory approach*, BMC Bioinformatics volume 18:142

[108] Kanehisa et al. (2010) *KEGG for representation and analysis of molecular networks involving diseases and drugs*, Nucleic Acids Research, Volume 38.

[109] Tumminello, Micciché, Lillo, Varho, Piilo, Mantegna (2011) *Community characterization of heterogeneous complex systems*, Journal of Statistical Mechanics: Theory and Experiment

[110] Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, Lehtiö J, Pawitan Y. (2012) *Network enrichment analysis: extension of gene-set enrichment analysis to gene networks.* BMC Bioinf. 13(1):226.

[111] McCormack T, Frings O, Alexeyenko A, Sonnhammer E. (2013) *Statistical assessment of crosstalk enrichment between gene groups in biological networks.* PLoS One. 8(1):54945.

[112] Signorelli, Vinciotti, Wit (2016) *NEAT: an efficient network enrichment analysis test*, BMC Bioinformatics volume 17, Article number: 352

[113] Chen, Feng, Hi (2017) *A general introduction to adjustment for multiple comparisons*, Journal Thoracic Diseases. 9(6): 1725–1729

[114] Benjamini, Hochbergm (1995) *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of royal statistical society.

[115] Benjamini, Yekutieli (2001) *The control of the false discovery rate in multiple testing under dependency.* Ann Stat 2001;29:1165-88.

[116] Efron (2010) *Large-scale inference. Empirical Bayes methods for estimation, testing and prediction.* Cambridge: Cambridge University Press.

[117] . Eulalio A, Tritschler F, Izaurralde E. The GW182 protein family in animal cells: new insights into domains required for miRNA-mediated gene silencing. RNA. 2009;15(8):1433–42.

[118] Pfaff J, Meister G. (2013) *Argonaute and GW182 proteins: an effective alliance in gene silencing.* Biochem Soc Trans. 41(4):855–60.

[119] Pfaff J, Hennig J, Herzog F, Aebersold R, Sattler M, Niessing D, et al. (2013) *Structural features of Argonaute-GW182 protein interactions.* Proc Natl Acad Sci U S A. 2013;110(40):E3770–9.

[120] Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, et al. (2004) *Argonaute2 is the catalytic engine of mammalian RNAi.* Science. 305(5689):1437–41.

[121] Ender C, Meister G. (2010) *Argonaute proteins at a glance.* J Cell Sci. 2010;123(Pt 11):1819–23.

[122] Huang V, Li LC. (2014) *Demystifying the nuclear function of Argonaute proteins.* RNA Biol. 2014;11(1):18–24.

[123] Hicks JA, Li L, Matsui M, Chu Y, Volkov O, Johnson KC, et al. (2017) *Human GW182 paralogs are the central organizers for RNA-mediated control of transcription.* Cell Rep. 2017;20(7):1543–52.

[124] Souquere S, Mollet S, Kress M, Dautry F, Pierron G, Weil D. (2009) *Unravelling the ultrastructure of stress granules and associated P-bodies in human cells.* J Cell Sci. 2009;122(Pt 20):3619–26.

[125] Eystathioy T, Chan EK, Takeuchi K, Mahler M, Luft LM, Zochodne DW, et al. (2003) *Clinical and serological associations of autoantibodies to GW bodies and a novel cytoplasmic autoantigen GW182.* J Mol Med (Berl). 2003;81(12):811–8.

[126] Sales G, Coppe A, Bisognin A, Biasiolo M, Bortoluzzi S, Romualdi C. (2020) *MAGIA, a web-based tool for miRNA and genes integrated analysis.* Nucleic Acids Res. 2010;38(Web Server):W352–9.

[127] Ebert MS, Neilson JR, Sharp PA. (2007) *MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells.* Nat Methods. 2007;4(9):721–6.

[128] Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. (2007) *A ceRNA hypothesis: the Rosetta stone of a hidden RNA language?* Cell. 2011;146(3):353–8.

[129] Turchinovich A, Burwinkel B. (2012) *Distinct AGO1 and AGO2 associated miRNA profiles in human cells and blood plasma.* RNA Biol. 2012;9(8):1066–75.

[130] Wang D, Zhang Z, O'Loughlin E, Lee T, Houel S, O'Carroll D, et al. (2012) *Quantitative functions of Argonaute proteins in mammalian development.* Genes Dev. 2012;26(7):693–704.

[131] Power Calculations for Matched-pairs designs, available at: https://sph.umd.edu/department/epib/sample-size-and-power-calculations-microarraystudies. Last accessed on November 28, 2018.

[132] Erhard F, Dolken L, Zimmer R. (2013) *RIP-chip enrichment analysis. Bioinformatics.* 2013;29(1):77–83.

[133] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. (2011) *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinf. 2011;12:77.

[134] Voller D, Linck L, Bruckmann A, Hauptmann J, Deutzmann R, Meister G, et al. (2016) *Argonaute family protein expression in Normal tissue and Cancer entities.* PLoS One. 11(8):e0161165.

[135] Kalantari R, Hicks JA, Li L, Gagnon KT, Sridhara V, Lemoff A, et al. (2016) *Stable association of RNAi machinery is conserved between the cytoplasm and nucleus of human cells.* RNA. 2016;22(7):1085–98.

[136] Wu PH, Isaji M, Carthew RW. (2013) *Functionally diverse microRNA effector complexes are regulated by extracellular signaling.* Mol Cell. 2013;52(1):113–23.

[137] Elkayam E, Faehnle CR, Morales M, Sun J, Li H, Joshua-Tor L. (2017) *Multivalent recruitment of human Argonaute by GW182.* Mol Cell. 2017;67(4):646–58 e3.

[138] Bartel (2004) *MicroRNAs: genomics, biogenesis, mechanism, and function*, Cell 116: 281–297.

[139] Brümmer, Hausser, (2014) *MicroRNA binding sites in the coding region of mRNAs: extending the repertoire of post-transcriptional gene regulation*, Bioessays 36(6):617-26.

[140] Perconti, Rubino, Contino, Bivona, Bertolazzi, Tumminello, Feo, Giallongo, Coronnello (2019) *RIP-Chip analysis supports different roles for AGO2 and GW182 proteins in recruiting and processing microRNA targets.* BMC Bioinformatics 20, 120

[141] Stark et al. (2003) *Identification of Drosophila MicroRNA Targets*, Plos Biology *doi.org*/10.1371/*journal.pbio*.0000060

[142] Han et al. (2004) *The Drosha–DGCR8 complex in primary microRNA processing*, genes and development *doi*/10.1101/*gad*.1262504

[143] Betel D, Koppal A, Agius P, Sander C, Leslie C. (2010) *Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites*, Genome Biology , 11:R90

[144] Hong, Hammell, Ambros, Cohen (2009) *Immunopurification of Ago1 miRNPs selects for a distinct class of microRNA targets*, Proceedings of the National Academy of Sciences of the United States of America 106: 15085–15090

[145] Eulalio, Rehwinkel, Stricker, Huntzinger, Yang, et al. (2007) *Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing*, Genes Development 21: 2558–2570.

[146] DeLong, DeLong, Clarke-Pearson (1988) *Comparison the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach*, Biometrica 44, 837-854.

[147] Liu C, Mallick B, Long D, Rennie WA, Wolenc A, Carmack CS, Ding Y (2013) *CLIP-based prediction of mammalian microRNA binding sites.* Nucleic Acids Research. 41:14 Page e138

[148] B.J. Bosch, R. van der Zee, C.A. de Haan, P.J. Rottier (2003) *The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex*, Journal 77, 8801–8811.

[149] A.C. Walls, Y.J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, D. Veesler (2020) *Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein*, Journal. , https://doi.org/10.1016/j.cell.2020.02.058.

[150] Yuki, Fujiogi, Koutsogiannaki (2020) *COVID-19 pathophysiology: A review*, Clinical Immunology

[151] Yuhao at al. (2020) *New understanding of the damage of SARS-CoV-2 infection outside the respiratory system*, Biomedicine Pharmacotherapy 127, 110195

[152] Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. (2020) *Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China.* Jama. https://doi.org/10.1001/jama.2020.1585.

[153] Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, et al. (2020) *Pathological findings of COVID-19 associated with acute respiratory distress syndrome*, The Lancet Respiratory medicine. 2213-2600(20)30076

[154] Lic, Yang, Ren (2020) *Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species*, Infection, Genetics and Evolution Volume 82.

[155] Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden T.L. (2019) *Magic-BLAST, an accurate RNA-seq aligner for long and short reads.* BMC Bioinformatics. 2019 Jul 25;20(1):405.

[156] Zhou, Yang, Shi (2020) *A pneumonia outbreak associated with a new coronavirus of probable bat origin*, Nature volume 579, pages270–273

[157] Tsan-Yuk Lam, Na Jia, Cao et al. *Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins*, Nature

[158] *Chinese Medicine and the Pangolin.* Nature 141, 72 (1938). https://doi.org/10.1038/141072b0

[159] Bartel, D. P. (2004). *MicroRNAs: Genomics, Biogenesis, Mechanism, and Function.* Cell 116 (2), 281–297. doi: 10.1016/S0092-8674(04)00045-5

[160] Bartel, D. P. (2009). *MicroRNAs: Target Recognition and Regulatory Functions.* Cell 136 (2), 215–233. doi: 10.1016/j.cell.2009.01.002

[161] Bruscella, P., Bottini, S., Baudesson, C., Pawlotsky, J. M., Feray, C., and Trabucchi, M. (2017). *Viruses and miRNAs: More friends than foes.* Front. Microbiol 8, 824. doi: 10.3389/fmicb.2017.00824

[162] Sumazin, P., Yang, X., Chiu, H. S., Chung, W. J., Iyer, A., Llobet-Navas, D., et al. (2011). *An extensive MicroRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma.* Cell 147 (2), 370– 381. doi: 10.1016/j.cell.2011.09.041

[163] Dang, W., Qin, Z., Fan, S., Wen, Q., Lu, Y., Wang, J., et al. (2016). *miR-1207-5p suppresses lung cancer growth and metastasis by targeting CSF1.* Oncotarget 7, 32421–32432. doi: 10.18632/oncotarget.8718

[164] Ludwig, N., Leidinger, P., Becker, K., Backes, C., Fehlmann, T., Pallasch, C., et al. (2016). *Distribution of miRNA expression across human tissues.* Nucleic Acids, Res 44 (8), 3865–3877. doi: 10.1093/nar/gkw116

[165] Love, M. II, Huber, W., and Anders, S. (2014). *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol 15, 550. doi: 10.1186/s13059-014-0550-8

[166] Benjamini, Y., and Hochberg, Y. (1995). *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological. J. R. Stat. Soc. Ser. B (Methological) 57 (1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

[167] Masters, P. S. (2006). *The Molecular Biology of Coronaviruses.* Adv. Virus Res 66, 193–292. doi: 10.1016/S0065-3527(06)66005-3

[168] Bouvet, M., Lugari, A., Posthuma, C. C., Zevenhoven, J. C., Bernard, S., Betzi, S., et al. (2014). *Coronavirus Nsp10, a critical co-factor for activation of multiple replicative enzymes.* J. Biol. Chem 289, 25783–25796. doi: 10.1074/ jbc.M114.577353

[169] Liao, Y., and Lönnerdal, B. (2010). *Global MicroRNA characterization reveals that miR-103 is involved in IGF-1 stimulated mouse intestinal cell proliferation.* PLoS One 5 (9), e12976. doi: 10.1371/journal.pone.0012976

[170] Martello, G., Rosato, A., Ferrari, F., Manfrin, A., Cordenonsi, M., Dupont, S., et al. (2010). *A microRNA targeting dicer for metastasis control.* Cell 141 (7), 1195– 1207. doi: 10.1016/j.cell.2010.05.017

[171] Annibali, D., Gioia, U., Savino, M., Laneve, P., Caffarelli, E., and Nasi, S. (2012). *A new module in neural differentiation control: Two microRNAs upregulated by retinoic acid, miR-9 and -103, target the differentiation inhibitor ID2.* PLoS One 7 (7), e40269. doi: 10.1371/journal.pone.0040269

[172] Chen, H. Y., Lin, Y. M., Chung, H. C., Lang, Y.D., Lin, C. J., Huang, J., et al. (2012). *MiR-103/107 promote metastasis of colorectal cancer by targeting the metastasis suppressors DAPK and KLF4*. Cancer Res. doi: 10.1158/0008- 5472.CAN-12-0667

[173] Yu, D., Zhou, H., Xun, Q., Xu, X., Ling, J., and Hu, Y. (2012). *microRNA-103 regulates the growth and invasion of endometrial cancer cells through the downregulation of tissue inhibitor of metalloproteinase 3*. Oncol. Lett. 3 (6), 1221–1226. doi: 10.3892/ol.2012.638

[174] Zhang, S. Y., Surapureddi, S., Coulter, S., Ferguson, S. S., and Goldstein, J. A. (2012). *Human CYP2C8 is post-transcriptionally regulated by microRNAs 103 and 107 in human liver*. Mol. Pharmacol 82 (3), 529–540. doi: 10.1124/ mol.112.078386

[175] Geng, L., Sun, B., Gao, B., Wang, Z., Quan, C., Wei, F., et al. (2014). *MicroRNA-103 promotes colorectal cancer by targeting tumor suppressor DICER and PTEN*. Int. J. Mol. Sci 15 (5), 8458–8472. doi: 10.3390/ijms15058458

[176] Liang, J., Liu, X., Xue, H., Qiu, B., Wei, B., and Sun, K. (2015). *MicroRNA-103a inhibits gastric cancer cell proliferation, migration and invasion by targeting c-Myb*. Cell Prolif 48 (1), 78–85. doi: 10.1111/cpr.12159

[177] Zhang, Y., Qu, X., Li, C., Fan, Y., Che, X., Wang, X., et al. (2015). *miR-103/107 modulates multidrug resistance in human gastric carcinoma by downregulating Cav-1*. Tumor Biol 36, 2277–2285. doi: 10.1007/s13277-014- 2835-7

[178] Asiaee, A., Abrams, Z. B., Nakayiza, S., Sampath, D., and Coombes, K. R. (2019). *Explaining Gene Expression Using Twenty-One MicroRNAs*. J. Comput. Biol 27 (7), 1157–1170. doi: 10.1089/cmb.2019.0321

[179] Qin, Z., He, W., Tang, J., Ye, Q., Dang, W., Lu, Y., et al. (2016). *MicroRNAs Provide Feedback Regulation of Epithelial-Mesenchymal Transition Induced by Growth Factors*. J. Cell. Physiol 231 (1), 120–129. doi: 10.1002/jcp.25060

[180] Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W. C., Uhl, S., Hoagland, D., Moller, R., et al. (2020). *Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19*. Cell 181 (5), 1036–1045. doi: 10.1016/ j.cell.2020.04.026

[181] Cabrera-Benitez, N. E., Laffey, J. G., Parotto, M., Spieth, P. M., Villar, J., Zhang, H., et al. (2014). *Mechanical ventilation-associated lung fibrosis in acute respiratory distress syndrome: A significant contributor to poor outcome*. Anesthesiology 121, 189–198. doi: 10.1097/ALN.0000000000000264

[182] Merad, M.,andMartin,J.C.(2020). *Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages*. Nat. Rev. Immunol 20, 355–362. doi: 10.1038/s41577-020-0331-4

[183] Spagnolo, P., Balestro, E., Aliberti, S., Cocconcelli, E., Biondini, D., Della Casa, G., et al. (2020). *Pulmonary fibrosis secondary to COVID-19: a call to arms?* Lancet Respir. Med 8 (8), 750–752. doi: 10.1016/s2213-2600(20)30222-8

[184] Louis, C., Cook, A. D., Lacey, D., Fleetwood, A. J., Vlahos, R., Anderson, G. P., et al (2015). Specific Contributions of CSF-1 and GM-CSF to the Dynamics of the Mononuclear Phagocyte System. J. Immunol 195 (1), 134–144. doi: 10.4049/ jimmunol.1500369

[185] Moon, H. G., Kim, S., Jeong, J. J., Han, S. S., Jarjour, N. N., Lee, H., et al. (2018). *Airway Epithelial Cell-Derived Colony Stimulating Factor-1 Promotes Allergen Sensitization.* Immunity 49 (2), 275–287. doi: 10.1016/ j.immuni.2018.06.009

[186] Turianová,L.,Lachová,V.,Svetlıḱova,D.,Kostrábová,A.,andBetáková,T.(2019). *Comparison of cytokine profiles induced by nonlethal and lethal doses of influenza A virus in mice.* Exp. Ther. Med 18 (6), 4397–4405. doi: 10.3892/ etm.2019.8096

[187] Vishnubalaji, R., Shaath, H., and Alajez, N. M. (2020). *Protein coding and long noncoding RNA (lncRNA)) transcriptional landscape in SARS-CoV-2 infected bronchial epithelial cells highlight a role for interferon and inflammatory response.* Genes (Basel) 11 (7), 760. doi: 10.3390/genes11070760

[188] Zhou, Z., Ren, L., Zhang, L., Zhong, J., Xiao, Y., Jia, Z., et al. (2020). *Heightened Innate Immune Responses in the Respiratory Tract of COVID-19 Patients.* Cell Host Microbe 27 (6), 883–890. doi: 10.1016/j.chom.2020.04.017

[189] Xiong, Y., Liu, Y., Cao, L., Wang, D., Guo, M., Jiang, A., et al. (2020). *Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients.* Emerg. Microbes Infect 9 (1), 761–770. doi: 10.1080/22221751.2020.1747363

[190] Ong, E. Z., Chan, Y. F. Z., Leong, W. Y., Lee, N. M. Y., Kalimuddin, S., Mohideen, S. M. H., et al. (2020). *A Dynamic Immune Response Shapes COVID-19 Progression.* Cell Host Microbe 27 (6), 879–882. doi: 10.1016/ j.chom.2020.03.021

[191] Wilk,A.J.,Rustagi,A.,Zhao,N.Q.,Roque,J.,Martıńez-Colón,G.J.,McKechnie, J. L., et al. (2020). *A single-cell atlas of the peripheral immune response in patients with severe COVID-19.* Nat. Med 26, 1070–1076. doi: 10.1038/s41591- 020-0944-y

[192] Emanuel, W., Mosbauer, K. , Franke, V., Diag, A., Gottula, L. T., Arsie, R., et al. (2020). *Bulk and single-cell gene expression profiling of SARS-CoV-2 infected human cell lines identifies molecular targets for therapeutic intervention.* bioRxiv. doi: 10.1101/2020.05.05.079194

[193] Ravindra, N., Alfajaro, M. M., Gasque, V., Habet, V., Wei, J., Filler, R. B., et al. (2020). *Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium.* BioRxiv Prepr Serv Biol. doi: 10.1101/2020.05.06.081695

[194] McGonagle, D., O'Donnell, J. S., Sharif, K., Emery, P., and Bridgewood, C. (2020). *Immune mechanisms of pulmonary intravascular coagulopathy in COVID-19 pneumonia.* Lancet Rheumatol 2 (7), E437–E445. doi: 10.1016/S2665-9913(20) 30121-1

[195] Wen, W., Su, W., Tang, H., Le, X., Zhang, X., Zheng, Y., et al. (2020). *Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing.* Cell Discovery 6, 31. doi: 10.1038/s41421-020-0168-9

[196] Gardinassi, L. G., Souza, C. O. S., Sales-Campos, H., and Fonseca, S. G. (2020). *Immune and Metabolic Signatures of COVID-19 Revealed by Transcriptomics Data Reuse.* Front. Immunol 11, 1636. doi: 10.3389/fimmu.2020.01636

[197] Akashi, K., Hayashi, S., Gondo, H., Mizuno, S., Harada, M., Tamura, K., et al. (1994). *Involvement of interferon-g and macrophage colony-stimulating factor in pathogenesis of haemophagocytic lymphohistiocytosis in adults.* Br. J. Haematol 87 (2), 243–250. doi: 10.1111/j.1365-2141.1994.tb04905.x

[198] Maruyama, J., and Inokuma, S. (2010). *Cytokine profiles of macrophage activation syndrome associated with rheumatic diseases.* J. Rheumatol 37 (5), 967–973. doi: 10.3899/jrheum.090662

[199] Kim, D., Lee, J. Y., Yang, J. S., Kim, J. W., Kim, V. N., and Chang, H. (2020). *The Architecture of SARS-CoV-2 Transcriptome.* Cell 181 (4), 914–921. doi: 10.1016/j.cell.2020.04.011

[200] Piccininni et al. (2020) *Use of all cause mortality to quantify the consequences of covid-19 in Nembro, Lombardy: descriptive study*, BMJ 2020;369:m1835

[201] Marvin Jens, Nikolaus Rajewsky (2015) *Competition between target sites of regulators shapes post-transcriptional gene regulation*, Nature Reviews Genetics volume 16, pages113–126

[202] Mieghem, Wang, Ge,Tang, Kuipers (2010) *Influence of assortativity and degree-preserving rewiring on the spectra of networks*, The European Physical Journal B volume 76, pages643–652

[203] Cui and Churchill (2003) *Statistical tests for differential expression in cDNA microarray experiments*, Genome Biology volume 4, Article number: 210 (2003)

[204] Pan (2002) *A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments*, BIOINFORMATICS Vol. 18 no. 4, Pages 546–554

[205] Morten W.Fagerland and LeivSandvik (2009) *Performance of five two-sample location tests for skewed distributions with unequal variances*, Contemporary Clinical Trials Volume 30, Issue 5, September 2009, Pages 490-496

[206] Gordon K. Smyth (2004) *Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments*, Statistical Applications in Genetics and Molecular Biology, Volume 3, Issue 1

[207] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu (2001) *Significance analysis of microarrays applied to the ionizing radiation response*. PNAS,vol. 98,no. 9,5116–5121

[208] Davis J. McCarthy, Gordon K. Smyth (2009) *Testing significance relative to a fold-change threshold is a TREAT*, Bioinformatics, Volume 25, Issue 6, Pages 765–771

[209] Bolstad B. M., Irizarry R. A., Astrand, M, and Speed, T. P. (2003) A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance, Bioinformatics 19(2) ,pp 185-193.

[210] Dussant, Gallo, Carballido, Ponzoni (2017) *Analysis of Gene Expression Discretization Techniques in Microarray Biclustering*, Springer International Publishing.

[211] Dimitrova, Vera Licona, McGee, et al. (2010) *Discretization of time series data*, Journal of Computational Biology 17(6):853–69.

[212] Karlebach and Shamir (2008) *Modelling and analysis of gene regulatory networks*, moLecuLar ceLL bioLogy Vol.8

[213] Gallo, Cecchini, Carballido, Micheletto and Ponzoni (2015) *Discretization of gene expression data revised*, Briefings in Bioinformatics, 17(5), 2016, 758-770.

[214] Li, Liu, Bai, Cai, Ji, Guo, Zhu (2010) *Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks*, BMC bioinformatics.

[215] Darrell Whitley (1994) *A genetic algorithm tutorial*, Statistics and Computing volume 4, pages65–85

[216] Zheng, Wang (2008) *GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis*, Nucleic Acids Research, Volume 36

[217] Marguerat and Bähler (2010) *RNA-seq: from technology to biology*, Cellular and Molecular Life Sciences volume 67, pages569–579(2010)

[218] Costa et al. (2010) *Uncovering the complexity of transcriptomes with RNA-Seq*,BioMed Research International, Article ID 853916

[219] Roy et al. (2011) *A comparison of analog and Next-Generation transcriptomic tools for mammalian studies*, Briefings in Functional Genomics, Volume 10, Issue 3, May 2011, Pages 135–150

[220] Fang and Cui (2011) *Design and validation issues in RNA-seq experiments*, Briefings in Bioinformatics, Volume 12, Issue 3, May 2011, Pages 280–287

[221] Xuan et al. (2012) *Next-generation sequencing in the clinic: promises and challenges*, Cancer Letters Volume 340, Issue 2, 1 November 2013, Pages 284-295

[222] Thomas Milan and Brian T. Wilhelm (2017) *Mining Cancer Transcriptomes: Bioinformatic Tools and the Remaining Challenges*, Molecular Diagnosis Therapy volume 21, pages249–258(2017)

[223] Marinov (2017) *On the design and prospects of direct RNA sequencing*, Briefings in Functional Genomics, Volume 16, Issue 6, November 2017, Pages 326–335

[224] Brothers II et al. (2018) *Integrity, standards, and QC-related issues with big data in pre-clinical drug discovery*, Biochemical Pharmacology Volume 152, June 2018, Pages 84-93 Biochemical Pharmacology

[225] Pacini and Koziol (2018) *Bioinformatics challenges and perspectives when studying the effect of epigenetic modifications on alternative splicing*, *https* : *//doi.org*/10.1098/*rstb*.2017.0073

[226] Hussain Ahmed Chowdhury et al. (2018) *Differential Expression Analysis of RNA-seq Reads: Overview, Taxonomy, and Tools*, IEEE/ACM Transactions on Computational Biology and Bioinformatics (Volume: 17, Issue: 2, March-April 1 2020)

[227] van der Lee R, Correard S, Wasserman WW. *Deregulated Regulators: Disease-Causing cis Variants in Transcription Factor Genes*, Trends in Genetics Volume 36, Issue 7, July 2020, Pages 523-539

[228] Culyba (2019) *Ordering up gene expression by slowing down transcription factor binding kinetics*, Current Genetics volume 65, pages401–406(2019)

[229] Dubois-Chevalier et al. *Organizing combinatorial transcription factor recruitment at cis-regulatory modules*, *https* : *//doi.org*/10.1080/21541264.2017.1394424

[230] Hacker and Wagner (2017) *Transcription factor decoy technology: A therapeutic update*, Biochemical Pharmacology Volume 144, 15 November 2017, Pages 29-34

[231] Kostas A. Papavassiliou, Athanasios G. Papavassiliou (2016) *Transcription Factor Drug Targets*, *ttps* : *//doi.org*/10.1002/*jcb*.25605

[232] Castellano et al. (2009) *The involvement of the transcription factor Yin Yang 1 in cancer development and progression*, *https* : *//doi.org*/10.4161/*cc*.8.9.8314

[233] Wagner and Lynch (2008) *The gene regulatory logic of transcription factor evolution*, Trends in Ecology Evolution Volume 23, Issue 7, July 2008, Pages 377-385

[234] Lee PharmD et al. (2013) *Nuclear factor kappa B: important transcription factor and therapeutic target*, *https* : *//doi.org*/10.1177/009127009803801101

[235] Martin and Wang (2011) *Next-generation transcriptome assembly*, Nature Reviews Genetics volume 12, pages671–682

[236] Castillo and Buell (2013) *Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence*, Natural Product Reports

[237] Moreton et al. (2016) *Assembly, Assessment, and Availability of De novo Generated Eukaryotic Transcriptomes*, Front. Genet., 11 January 2016, *https* : *//doi.org*/10.3389/*fgene*.2015.00361

[238] Spyros Oikonomopoulos et al. (2020) *Methodologies for Transcript Profiling Using Long-Read Technologies*, Front. Genet., 07 July 2020 | *https* : *//doi.org*/10.3389/*fgene*.2020.00606

[239] Prange et al. (2014) *The genome-wide molecular signature of transcription factors in leukemia*, Experimental Hematology Volume 42, Issue 8, August 2014, Pages 637-650

[240] Wang et al. (2015) *Pathway and network approaches for identification of cancer signature markers from omics data*, J Cancer. 2015; 6(1): 54–65.

[241] Duran-Pinedo (2020) *Metatranscriptomic analyses of the oral microbiome*, *https* : *//doi.org*/10.1111/*prd*.12350

[242] Corso G, Figueiredo J, De Angelis SP, et al. (2020) *E-cadherin deregulation in breast cancer*, S.J Cell Mol Med. 2020 Jun;24(11):5930-5936.

[243] Rejon C, Al-Masri M, McCaffrey L. (2016) *Cell Polarity Proteins in Breast Cancer Progression*, J Cell Biochem. 2016 Oct;117(10):2215-23. Epub 2016 Jun 30. PMID: 27362918 Review.

[244] Yin et al. (2018) *Wnt signaling in human and mouse breast cancer: Focusing on Wnt ligands, receptors and antagonists*, *https* : *//doi.org*/10.1111/*cas*.13771

[245] Rausch et al. (2017) *The Linkage between Breast Cancer, Hypoxia, and Adipose Tissue*, *https* : *//doi.org*/10.3389/*fonc*.2017.00211

[246] Zhou et al. (2015) *Claudin 1 in Breast Cancer: New Insights*, J. Clin. Med. 2015, 4(12), 1960-1976

[247] Zhu et al. (2014) *Integrated extracellular matrix signaling in mammary gland development and breast cancer progression*, Histol Histopathol. 2014 Sep; 29(9): 1083–1092.

[248] Chatterjee and McCaffrey (2014) *Emerging role of cell polarity proteins in breast cancer progression and metastasis*, Breast Cancer (Dove Med Press). 2014; 6: 15–27.

[249] Yiafan et al. (2013) *Epithelial-mesenchymal Transition: A Hallmark of Breast Cancer Metastasis*, American Scientific Publishers, Volume 1, Number 1, March 2013, pp. 38-49(12)

[250] Bazzoun et al. (2013) *Polarity proteins as regulators of cell junction complexes: implications for breast cancer*, Pharmacology Therapeutics Volume 138, Issue 3, June 2013, Pages 418-427

[251] Rajan A, Nadhan R, Latha NR, Krishnan N, Warrier AV, Srinivas P. (2020) *Deregulated estrogen receptor signaling and DNA damage response in breast tumorigenesis*, Biochim Biophys Acta Rev Cancer. 2020 Nov 28;1875(1):188482.

[252] Akram M, Iqbal M, Daniyal M, Khan AU. (2017) *Awareness and current knowledge of breast cancer*, Biological Research volume 50, Article number: 33

[253] Tan PH, Ellis I, Allison K et al. (2020) *The 2019 World Health Organization classification of tumours of the breast*. Histopathology 77(2), 181–185

[254] Piezzo et al. (2020) *Targeting Cell Cycle in Breast Cancer: CDK4/6 Inhibitors*, Int. J. Mol. Sci. 2020, 21(18), 6479

[255] Ding et al. (2020) *The Roles of Cyclin-Dependent Kinases in Cell-Cycle Progression and Therapeutic Strategies in Human Breast Cancer*, Int. J. Mol. Sci. 2020, 21(6)

[256] Fedele et al. (2019) *A clinical evaluation of treatments that target cell cycle machinery in breast cancer*, *https* : *//doi.org*/10.1080/14656566.2019.1672659

[257] Thu et al. (2018) *Targeting the cell cycle in breast cancer: towards the next phase*, *https* : *//doi.org*/10.1080/15384101.2018.1502567

[258] Butt et al. (2008) *Cell cycle machinery: links with genesis and treatment of breast cancer*, Innovative Endocrinology of Cancer pp 189-205

[259] Akram M, Iqbal M, Daniyal M, Khan AU. (2017) *Awareness and current knowledge of breast cancer*, Biological Research volume 50, Article number: 33 (2017)

[260] Tan PH, Ellis I, Allison K et al. (2020) *The 2019 World Health Organization classification of tumours of the breast*. Histopathology 77(2), 181–185

[261] Ortega et al. (2020) *Signal Transduction Pathways in Breast Cancer: The Important Role of PI3K/Akt/mTOR*, Journal of Oncology, Article ID 9258396

[262] Bedard et al. (2008) *Overcoming endocrine resistance in breast cancer: are signal transduction inhibitors the answer?*, Breast Cancer Research and Treatment volume 108, pages307–317

[263] L. Gharaibeh et al. (2020) *Notch1 in Cancer Therapy: Possible Clinical Implications and Challenges*, Molecular Pharmacology November 2020, 98 (5) 559-576

[264] STRAVODIMOU1 and VOUTSADAKIS (2020) *The Future of ER+/HER2- Metastatic Breast Cancer Therapy: Beyond PI3K Inhibitors*, Anticancer Research September 2020 vol. 40 no. 9 4829-4841

[265] Wettersten et al. (2020) *Reprogramming of Metabolism in Kidney Cancer*, Seminars in Nephrology, 2020 - Elsevier

[266] Aggen (2020) *Targeting PD-1 or PD-L1 in Metastatic Kidney Cancer: Combination Therapy in the First-Line Setting*, DOI: 10.1158/1078 – 0432.CCR – 19 – 3323

[267] Drake et al. (2018) *The Immunobiology of Kidney Cancer*, Journal of Clinical Oncology, Volume 36, Issue 36

[268] Nagy et al. (2016) *High risk of development of renal cell tumor in end-stage kidney disease: the role of microenvironment*, Tumor Biology volume 37, pages9511–9519(2016)

[269] Drake et al. (2020) *Stromal β-catenin activation impacts nephron progenitor differentiation in the developing kidney and may contribute to Wilms tumor*, doi: 10.1242/*dev*.189597

[270] Peterfi et al. (2019) *IL6 Shapes an Inflammatory Microenvironment and Triggers the Development of Unique Types of Cancer in End-stage Kidney*, Anticancer Research April 2019 vol. 39 no. 4

[271] Margaroli et al. (2020) *The immunosuppressive phenotype of tumor-infiltrating neutrophils is associated with obesity in kidney cancer patients*, OncoImmunology, Volume 9, 2020 - Issue 1

[272] Kovaleva et al. (2016) *Tumor Associated Macrophages in Kidney Cancer*, *https* : //*doi.org*/10.1155/2016/9307549

[273] Zou et al. (2020) *A 14 immune-related gene signature predicts clinical outcomes of kidney renal clear cell carcinoma*, PubMed 33194402

[274] Xu et al. (2020) *Checkpoint inhibitor immunotherapy in kidney cancer*, Nature Reviews Urology volume 17, pages137–150

[275] Hammers et al. (2016) *Immunotherapy in kidney cancer: the past, present, and future*, Current Opinion in Urology: November 2016 - Volume 26 - Issue 6 - p 543-547

[276] Kamli et al. (2019) *Limitations to the Therapeutic Potential of Tyrosine Kinase Inhibitors and Alternative Therapies for Kidney Cancer*, Ochsner Journal June 2019, 19 (2) 138-151

[277] Zhu G, Pan C, Bei JX, Li B, Liang C, Xu Y, Fu X. (2020) *Mutant p53 in Cancer Progression and Targeted Therapies*, Front Oncol. 2020 Nov 6;10:595187. PMID: 33240819 Free PMC article. Review.

[278] Liu Y, Leslie PL, Zhang Y. (2020) *Life and Death Decision-Making by p53 and Implications for Cancer Immunotherapy*m Trends Cancer. 2020 Nov 13:S2405-8033(20)30279-X.

[279] Levine AJ. (2020) *p53: 800 million years of evolution and 40 years of discovery*, Nat Rev Cancer. 2020 Aug;20(8):471-480. doi: 10.1038/s41568-020-0262-1. Epub 2020 May 13.PMID: 32404993 Review.

[280] Schmidt and Linehana (2016) *Genetic predisposition to kidney cancer*, Seminars in Oncology, Volume 43, Issue 5, October 2016, Pages 566-574

[281] Abhishek A.Chakraborty (2020) *Coalescing lessons from oxygen sensing, tumor metabolism, and epigenetics to target VHL loss in kidney cancer*, Seminars in Cancer Biology Volume 67, Part 2, December 2020, Pages 34-42

[282] Yamaguchi, Harada, Hirota (2016) *VHL-deficient renal cancer cells gain resistance to mitochondria-activating apoptosis inducers by activating AKT through the IGF1R-PI3K pathway*, Tumor Biology volume 37, pages13295–13306

[283] Roe et al. (2011) *Phosphorylation of von Hippel-Lindau protein by checkpoint kinase 2 regulates p53 transactivation*, $https://doi.org/10.4161/cc.10.22.18096$

[284] Hell et al. (2014) *Tumor suppressor VHL functions in the control of mitotic fidelity*, DOI: 10.1158/0008-5472.CAN-13-2040

[285] Zhao et al. (2016) *Synergy between von Hippel-Lindau and P53 contributes to chemosensitivity of clear cell renal cell carcinoma*, $https://doi.org/10.3892/mmr.2016.5561$