Inference and Analysis of Multilayered Mirna-Mediated Networks in Cancer

Duc Do
Marquette University

INFERENCE AND ANALYSIS OF MULTILAYERED MIRNA-MEDIATED NETWORKS IN CANCER

by

Duc Do, B.S.

A Dissertation Submitted to the Faculty of the
Graduate School, Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy

Milwaukee, Wisconsin

August 2018

ABSTRACT
INFERENCE AND ANALYSIS OF MULTILAYERED MIRNA-MEDIATED NETWORKS IN CANCER

Duc Do, B.S.

Marquette University, 2018

MicroRNAs (miRNAs) are small noncoding transcripts that can regulate gene expression, thereby controlling diverse biological processes. Aberrant disruptions of miRNA expression and their interactions with other biological agents (e.g., coding and noncoding transcripts) have been associated with several types of cancer. The goal of this dissertation is to use multidimensional genomic data to model two different gene regulation mechanisms by miRNAs in cancer. This dissertation results from two research projects. The first project investigates a miRNA-mediated gene regulation mechanism called competing endogenous RNA (ceRNA) interactions, which suggests that some transcripts can indirectly regulate one another's activity through their interactions with a common set of miRNAs. Identification of context-specific ceRNA interactions is a challenging task. To address that, we proposed a computational method called Cancerin to identify genome-wide cancer-associated ceRNA interactions. Cancerin incorporates DNA methylation (DM), copy number alteration (CNA), and gene and miRNA expression datasets to construct cancer-specific ceRNA networks. Cancerin was applied to three cancer datasets from the Cancer Genome Atlas (TCGA) project. We found that the RNAs involved in ceRNA interactions were enriched with cancer-related genes and have high prognostic power. Moreover, the ceRNA modules in the inferred ceRNA networks were involved in cancer-associated biological processes. The second project investigates what biological functions are regulated by both miRNAs and transcription factors (TFs). While it has been known that miRNAs and TFs can coregulate common target genes having similar biological functions, it is challenging to associate specific biological functions to specific miRNAs and TFs. In this project, we proposed a computational method called CanMod to identify gene regulatory modules. Each module consists of miRNAs, TFs and their coregulated target genes. CanMod was applied on the breast cancer dataset from TCGA. Many hub regulators (i.e., miRNAs and TFs) found in the inferred modules were known cancer genes, and CanMod was able to find experimentally validated regulator-target interactions. In addition, the modules were associated with distinguishable and cancer-related biological processes. Given the biological findings obtained from Cancerin and CanMod, we believe that the two computational methods are valuable tools to explore novel miRNA involvement in cancer.

ACKNOWLEDGEMENTS

Duc Do, B.S.

I want to send my utmost gratitude to my advisor, Dr. Serdar Bozdag, for his expertise, guidance, patience, and encouragement during my whole PhD career. It was a great pleasure working with him during a challenging and enriching process. I am also very much thankful to Dr. George Corliss for his constructive advice and criticism. I also want to thank my other committee members in no particular order, Dr. Allison Abbott, Dr. Naveen Bansal, and Dr. Mehdi Maadooliat. Thank you very much for your helpful comments and feedback over the years. They are integral for my professional growth. I am also grateful to the Department of Mathematics, Statistics and Computer Science and Marquette Graduate School for all of their support over the last five years.

I would like to thank my parents, Loc Do and Ha Pham, and my wife, Linh Pham, for providing me with unwavering support and always be there for me during the last challenging five years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This dissertation describes two computational methods that aim to model two gene regulatory mechanisms governed by microRNAs (miRNAs) in cancer biology. The dissertation is organized as follow. Chapters 1 introduces the biological background and the research motivations. Chapter 2 and Chapter 3 describe the two computational methods in detail and analyze the results obtained from applying the methods to cancer datasets. Chapter 4 summarizes the contributions of the dissertation and offers several future research directions.

This introductory chapter starts with a brief history of miRNA discovery and a discussion about miRNA involvement in complex diseases such as cancer. Next, it describes the overall research motivation of the dissertation. Then it presents the biological background of miRNAs, including miRNA biogenesis and basic functional roles of miRNAs. Next, it briefly discusses other types of important gene regulators besides miRNAs. The chapter ends with the specification of the research goals of the dissertation.

1.1 MicroRNA - a crucial gene regulator emerged from the "junk DNA" world

The discovery of the DNA structure in 1953 by James Watson and Francis Crick has guided molecular biology research to a new direction - understanding the intricacy of cellular functions via understanding how genetic information is stored and "decoded" in the cells [Pray, 2008]. In a seminal presentation in 1957, Crick presented the "central dogma" of molecular biology [Cobb, 2017]. The "central dogma" says that our genetic information is stored as a DNA code, which can be converted into messenger RNAs (mRNAs) (Fig. 1.1). The mRNAs are "decoded" to produce proteins, which are the agents that dictate different biological functions. This fundamental understanding of protein synthesis from the DNA level has played an essential role in many scientific breakthroughs and is undeniably one of the greatest scientific achievements in the 20^{th} century.

The "central dogma" implies that important instructions for protein synthesis are packed into the protein-coding genes that produce mRNAs. However, it is known that there are many sequences that are transcribed to RNAs but could not be translated into protein [Cheng et al., 2005, Van Bakel et al., 2010]. Those RNAs are known as non-coding RNAs (ncRNAs). While some ncRNAs are functional molecules (e.g., transfer RNAs, ribosomal RNAs), most ncRNAs have unknown functions [Hüttenhofer et al., 2005]. The DNA sequences that do not encode pro-

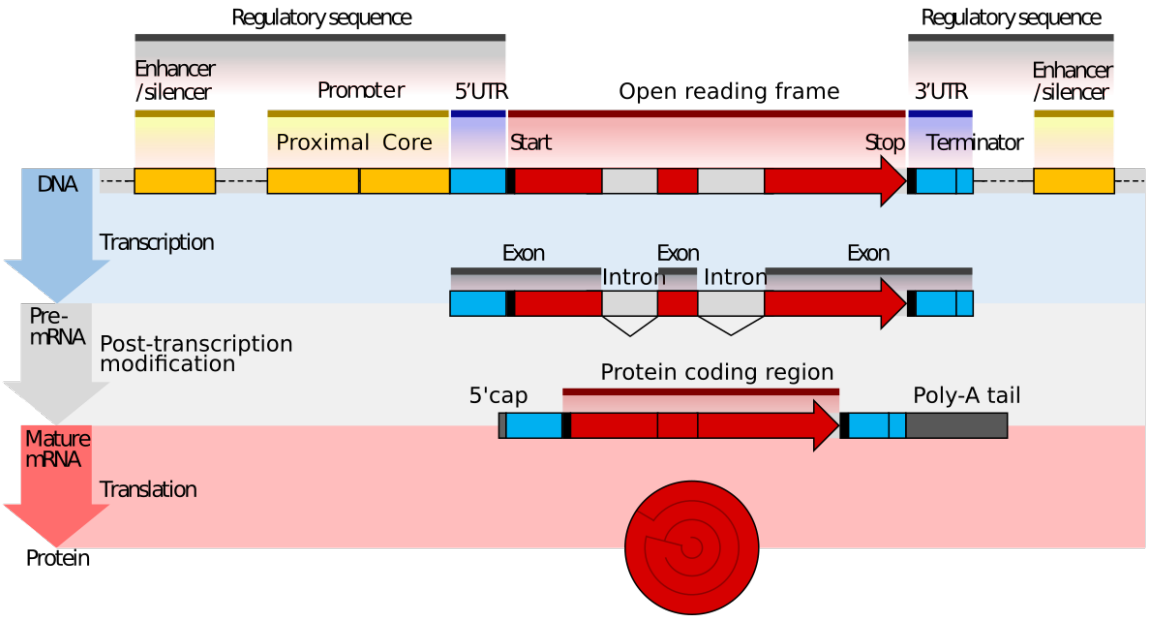

Figure 1.1: **Central dogma of molecular biology.** The central dogma describes the genetic information flow in cells from DNA to mRNA to protein. First, during the transcription process, precursor mRNAs (pre-mRNAs) are synthesized from DNA templates. Then, during the post-transcription modification process, the pre-mRNAs are converted into mature messenger RNAs (mRNAs). Finally, during the translation process, the sequences of nucleotides in mRNAs are translated into proteins. (https://commons.wikimedia.org/wiki/Category:Central_dogma_of_molecular_biology/media/File:Gene_structure_eukaryote_2_annotated.svg. Public Domain.)

teins were termed "junk DNAs" [Ehret and De Haller, 1963, Ohno, 1972, Palazzo and Lee, 2015]. Prior to the 21th century, the scientific community primarily focused on identifying and understanding the importance of protein-coding genes and overlooked the "junk DNA" world.

However, non-coding genes/RNAs have attracted much attention thanks to the completion of the Human Genome Project (HGP) in 2003 [International Human Genome Sequencing Consortium, 2004], which was followed by the ongoing

Encyclopedia of DNA Elements (ENCODE) project [ENCODE Project Consortium, 2004]. While the HGP sequenced and mapped the whole human genome, the ENCODE project aims to explore the biological significance of the genome sequence. The HGP found that there are around 20,000 protein-coding genes in human DNA, but surprisingly these genes make up only around 1% of the entire genome [ENCODE Project Consortium, 2007]. The remaining 99% are "junk DNA" regions that do not encode proteins. Thus, one major goal of the ENCODE project is to determine the biological functions of the non-coding genes [Djebali et al., 2012, Derrien et al., 2012]. A small portion of those genes, named miRNAs, emerged as one of the important molecular players. By binding to the protein-coding transcripts, miRNAs play an important role in regulating the production of proteins, thereby regulating many important biological functions.

1.1.1 Discovery of miRNAs

Discovery of the first miRNA was traced back to 1993 by joint efforts of different research groups while working on the nematode *Caenorhabditis elegans*. Ambros and his colleagues Rosalind Lee and and Rhonda Feinbaum discovered that the normal development of this organism required both the transcription of the gene linc-4 and the downregulation of the protein LIN-14 [Ambros, 1989, Lee et al., 1993]. They made a crucial observation that the lin-4 gene could not produce

protein. Instead, the lin-4 gene produced two small non-protein-coding RNAs that were approximately 21 and 61 nucleotides (nt) in length.

Soon after the lin-4 discovery by the Ambros group, Ruvkun and his colleagues Wightman and Ha made an important observation that there are repeated regions on LIN-14 containing complementary sequences to the 21-nt long lin-4 RNA [Arasu et al., 1991, Wightman et al., 1993]. Their follow up experiments showed that the binding of the small lin-4 RNA to the LIN-14 RNA decreased the LIN-14 protein expression. This was the first finding presenting an important regulatory mechanism that involved a binding between a small non-coding RNA with a protein-coding RNA.

It took seven years for the second miRNA to be discovered. In 2000, Reinhart in Ruvkun's laboratory discovered that let-7, another 21-nt long RNA, had a crucial role in the final larval developmental transition from larval stage to adulthood in *C. elegans* [Reinhart et al., 2000]. Quickly after Reinhart's discovery, let-7 was found also in many other organisms, including humans. The findings about the conservation of let-7 across different species sparked a research interest in miRNAs [Roush and Slack, 2008]. Since then, different important functional roles of miRNA have been found not only in normal biological processes but also in diseases such as cancer.

1.1.2 MiRNA involvement in cancer

Cancer is a disease in which abnormal cells divide without control and become invasive. Gene expression is dramatically deregulated in cancer [Cooper and Hausman, 2000]. In 2002, George Calin and colleagues presented the first report of miRNA involvement in cancer [Calin et al., 2002]. They observed that two miRNAs mir-15a and mir-16-1 were either consistently deleted or downregulated in B cell chronic lymphocytic leukemia (B-CLL), suggesting their potential role as tumor suppressors ("brakes" to inhibit tumorgenesis). Later it was confirmed that the deletion or downregulation of these two miRNAs led to the activation of their target oncogenes ("gas pedals" to accelerate tumorgenesis) CCND2 and CCND3 [Klein et al., 2010]. Since then, the involvement of miRNAs has been reported in many types of cancers such as breast, kidney, head and neck [Iorio et al., 2005, Catto et al., 2011, Tran et al., 2007].

1.1.3 Research motivation

There are many possible interactions between miRNAs and target transcripts. Moreover, miRNAs' and their target genes' activities can vary considerably in different contexts such as normal cells versus cancer cells. Thus, identifying the functional significance of miRNA-target interactions in different cellular con-

ditions is challenging [Kwan et al., 2016, Catalanotto et al., 2016]. In addition, since genes under miRNA regulation can also be regulated by many other types of regulators such as transcription factors (TFs), copy number alteration (CNA), and DNA methylation (DM), the interplay between miRNAs and those regulatory factors can have important biological functions but not well understood [Hayes et al., 2014, Jones, 2015]. There is still much unknown about different gene regulation mechanisms that miRNAs participate in. Fortunately, thanks to developments of technology, vast amounts of biological data have been generated, which enables us to gain important information of miRNA functions [Motameny et al., 2010, Aldridge and Hadfield, 2012].

By leveraging multiple types of biological data, this dissertation aims to model two important cancer-associated gene regulation mechanisms governed by miRNAs. The first mechanism involves the "competing endogenous RNA (ceRNA)", which suggested that two RNA transcripts can indirectly regulate each other through their interactions with their common miRNA regulators. The second mechanism involves the coordination of miRNAs and TFs in coregulating common target genes that are in charge of important biological processes. These two mechanisms will be discussed in more detail in Section 1.5 Research Goal and especially later in Chapters 2 and 3.

The next two sections specify the important biological background of miRNAs to facilitate the subsequent discussion and analysis in the dissertation.

1.2 MiRNA biogenensis

Typically, miRNA genes are transcribed by Polymerase II RNA to produce long primary transcripts (pri-miRNAs) that are approximately 400 nt long [Denli et al., 2004, Bartel, 2018]. The pri-miRNAs are bound by a microprocessor complex comprised of Drosha, an RNase III families of enzymes, and a gene DGCR8 [Gregory et al., 2004]. The microprocessor complex cleaves the 5'cap and poly-A tail out of the pri-miRNAs, leaving a hairpin pri-miRNAs structure [Ha and Kim, 2014].

After the pre-miRNAs are produced in the nucleus, they are exported into the cytoplasm by a nucleocytoplasmic shuttler Exportin-5 [Kim, 2004] (Fig. 1.2 point 1). In the cytoplasm, an RNA III enzyme named Dicer binds to the pri-miRNAs and cleaves the loop connecting the 3' and 5' arms of pri-miRNAs [Lund and Dahlberg, 2006] (Fig. 1.2 point 2). The cytoplasmic processing by Dicer yields double-stranded RNAs, named the miRNA-3p/miRNA-5p duplex. In most cases, only one of the miRNAs in the duplex is involved in gene regulation, and that miRNA is considered to be a mature miRNA. The remaining one is normally degraded and considered to be a passenger miRNA [Bartel, 2004, Lau et al., 2001]. Nevertheless, there

are some miRNA duplexes where both miRNAs are functionally active in regulating gene expression, and both are considered as mature miRNAs [Yekta et al., 2004].

1.3 Gene regulation by miRNA

The best-known function of miRNAs is their ability to repress translation of protein-coding genes. The process starts with the miRNA duplex binding to an Argonaute (AGO) protein to form RNA-induced silencing complex (RISC) (Fig. 1.2 point $2\frac{1}{2}$). The binding with AGO separates the duplex into the mature miRNA from the passenger miRNA. On the 5' end of the mature miRNAs resides a region, which is complementary or partially complementary with the miRNA-binding-sites ("seed" region or miRNA response elements - MREs) located on the 3'UTR of the target mRNAs [Pasquinelli, 2012, Thomson et al., 2011] (Fig. 1.2 point 4). It has been demonstrated that a transcript can contain multiple MREs for one or multiple miRNAs, implying a many-to-many relation between mRNAs and miRNAs [Pasquinelli, 2012]. Mature miRNAs recognize their target mRNAs based on the sequence complementary and guide their associated RISCs to bind to the MREs on their target mRNAs, resulting in the decrease of the protein output of the target mRNAs [Catalanotto et al., 2016].

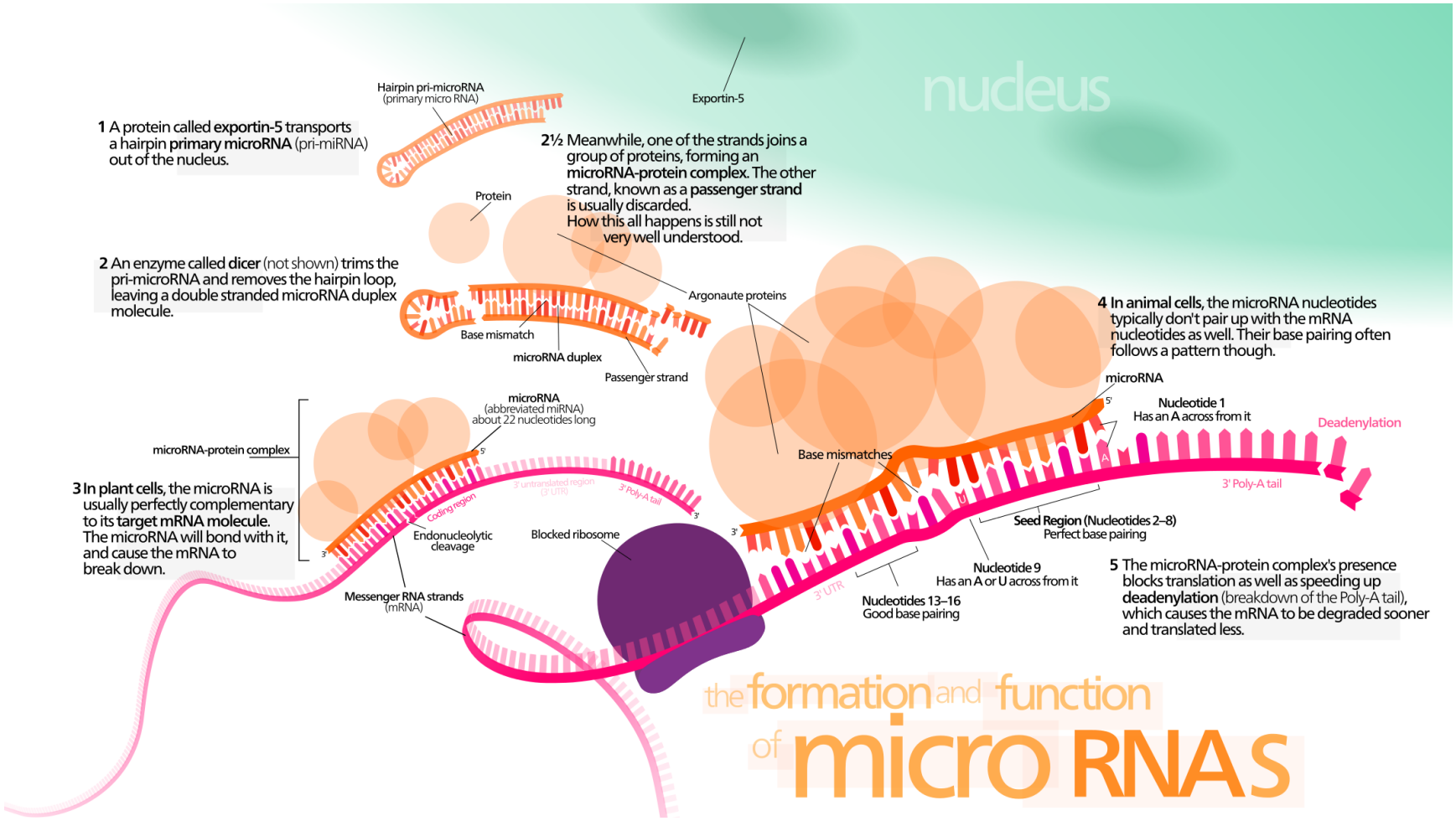

Figure 1.2: **Illustration of miRNA biogenesis and gene regulation by miRNA.** (https://en.wikipedia.org/wiki/ MicroRNA/media/File:MiRNA.svg. Public Domain).

1.4 Other types of gene regulators

This dissertation explores the regulatory relationship between miRNAs and their target genes. However, it is important to note that expression of a gene is regulated by other regulatory factors besides miRNAs such as transcription factor (TF), copy number alteration (CNA), and DNA methylation (DM).

1.4.1 Transcription factor (TF)

TFs are a type of protein that regulate their target genes at the transcriptional level by binding to the genes' promoter, enhancer, or repressor regions [Voss and Hager, 2014] (see Fig. 1.3). TFs can either stimulate or repress the transcription of their target genes, leading to either the increase or decrease of the target genes' expression [Spitz and Furlong, 2012]. Similar to the many-to-many relation between miRNAs and their target genes, a TF can simultaneously regulate many target genes, and a target gene can be regulated by multiple TFs.

1.4.2 Copy number alteration (CNA)

CNA is a type of structural alteration in the genome that result in gain or loss in copies of sections of DNA [Beroukhim et al., 2010]. In humans, each gene has two copies. Genes residing in the DNA regions that undergo copy number alter-

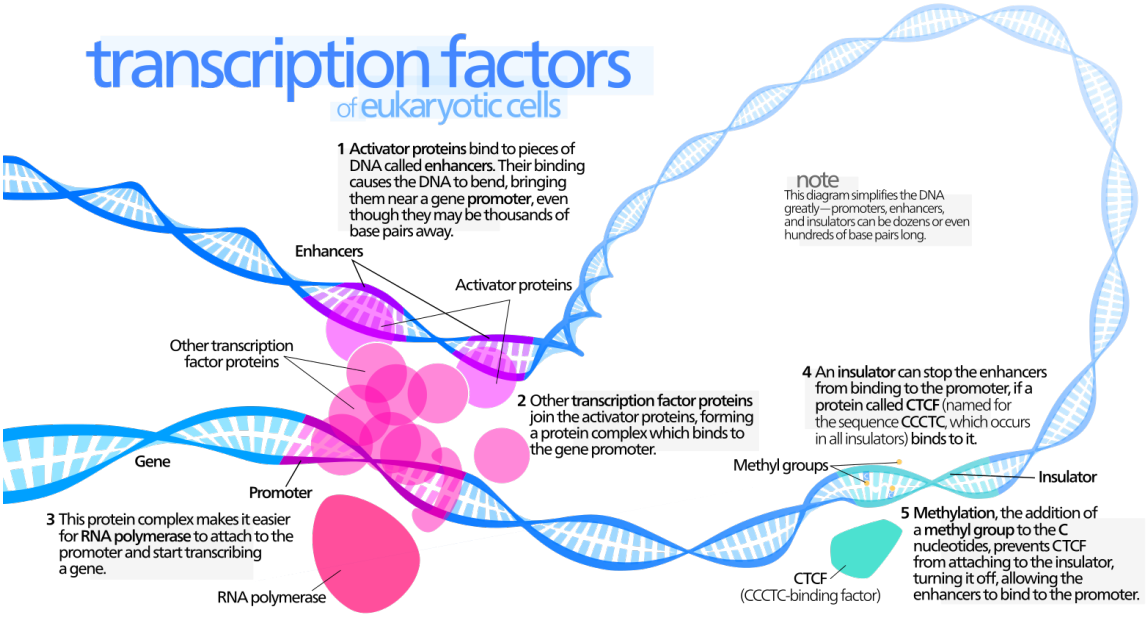

Figure 1.3: **Gene regulation at transcriptional level.** (https://commons.wikimedia.org/wiki/File:Transcription_Factors.svg. Public Domain).

ation have their copy numbers changed. Genes with copy number gain often have their expression increased; in contrast, genes with copy number loss often have their expression decreased [Taylor et al., 2008]. Fig. 1.4 presents the cases of CNA due to gene duplication and deletion of genes. CNA has been associated with complex traits in human and aberrant CNA has been implicated in many diseases including cancer [Taylor et al., 2008, Shlien and Malkin, 2009].

1.4.3 DNA methylation (DM)

DM is a chemical change to DNA, which occurs when methyl groups are

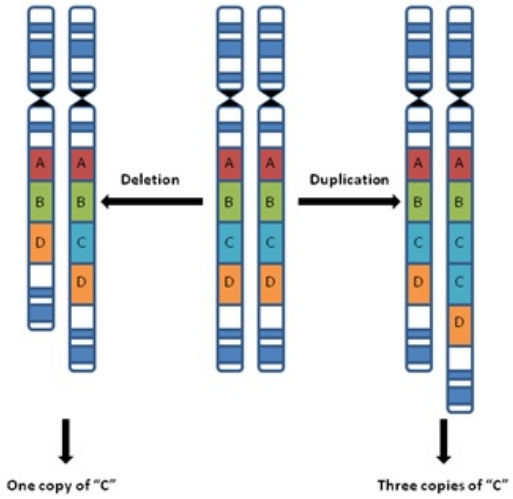

Figure 1.4: **Duplication or deletion of genes on a chromosome.** (http://readingroom.mindspec.org/wp-content/genetics_CNV.jpg. Public Domain).

added to the nucleotide cytosine [Phillips, 2008] (see Fig. 1.5). DM can alter the activity of genes residing on the sequence without changing the primary nucleotide sequence. Thus, DM is considered as an epigenetic gene regulation mechanism [Phillips, 2008]. When DM occurs on gene promoter regions, it often represses gene transcription and therefore decreases the gene expression [Jones, 2012]. DM can deactivate activities of important tumor-suppressor genes, which can trigger tumor formation and development processes [Baylin, 2005].

The different regulatory factors including TFs, CNA, DM, and miRNAs represent multilayers of gene expression regulation from transcriptional regulation (TFs and CNA) and epigenetic regulation (DM), to post-transcriptional regulation (miRNAs). Thus, studies that explore the functional importance of miRNAs based

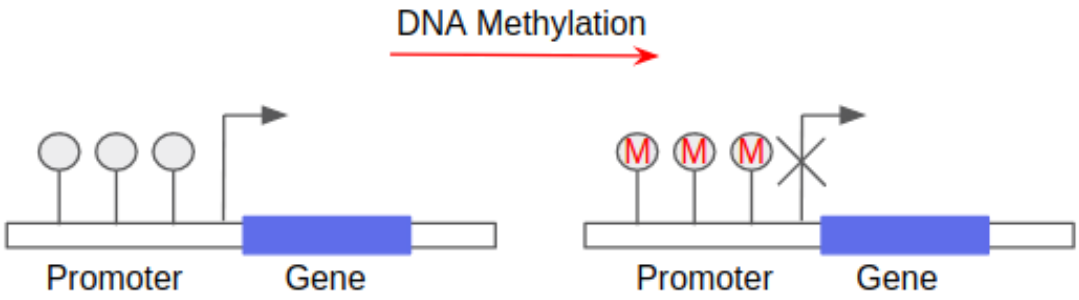

Figure 1.5: **Transcriptional repression via DNA methylation.**

on how they regulate expression of their target genes should be mindful about the other regulatory factors to avoid spurious conclusions about miRNA-target gene relation. The two computational methods proposed in this dissertation both incorporate those regulatory factors in their pipelines.

Given the related biological backgrounds, the following section will specify the research goals of this dissertation.

1.5 Research goals

There is still much unknown about different gene regulation mechanisms in which miRNAs participate and how they are associated with diseases such as cancer. This dissertation focuses on two specific gene regulation mechanisms involving miRNAs. The first gene regulation mechanism involves identifying indirect interaction between RNA transcripts that are coregulated by common miRNAs. The

second gene regulation mechanism involves identifying the gene regulatory modules consisting of miRNAs, TFs, and their coregulated genes. To that end, we developed two computational methods, namely Cancerin and CanMod, to identify ceRNA interactions and gene regulatory modules in cancer, respectively. We believe that the two methods can be used to provide meaningful insights of miRNA involvement in cancer biology. The next two sections describe research goals behind the two computational methods.

1.5.1 Identification of cancer-associated endogenous competing RNA interactions

Competing endogenous RNA (ceRNA) interactions involve indirect interactions between RNA molecules via their interactions with their common miRNA regulators [Tay et al., 2011, Cesana et al., 2011]. The ceRNA hypothesis [Salmena et al., 2011] posits that a change of expression level in one ceRNA would alter its miRNA regulators' abundance, which in turn alters the expression level of other target ceRNAs of these miRNAs. For example, a highly expressed ceRNA can sequester many miRNA molecules, reducing the total miRNA abundance and leading to the derepression of other target ceRNAs of these miRNAs. CeRNA interactions are not only among protein coding RNAs (i.e., mRNAs). Recent studies have found that non-coding RNAs such as long non-coding RNAs (lncRNAs) and pseudogenes

also involved in ceRNA interactions [Xia et al., 2014, Zhang et al., 2018, An et al., 2017].

CeRNA interactions have been shown to regulate important biological processes, and disruption of ceRNA interactions has been implicated in multiple types of diseases including cancer [Tay et al., 2014, Sanchez-Mejias and Tay, 2015, Li et al., 2017]. Identification and construction of genome-wide and condition-specific ceRNA interaction networks could facilitate better understanding of ceRNA regulatory mechanisms and their biological significance, especially in cancer biology. Under that motivation, in Chapter 2, we propose a computational pipeline called Cancerin, which infers ***Can***cer-associated ***ceR***NA ***i***teraction ***n***etworks. Cancerin was applied to three cancer datasets. In brief, the analysis results show that compared to existing methods, Cancerin is able to identify cancer-related ceRNA interactions with higher accuracy. The ceRNA interactions obtained by Cancerin could be used to help researchers acquire new insights on the roles of miRNAs in cancer formation and development.

1.5.2 Identification of cancer-associated gene regulatory modules

Besides miRNAs, a gene also can be regulated by many other factors including TFs [Voss and Hager, 2014, Jones, 2015]. Abnormal alternation of gene regulation by either miRNA or TF can disrupt important biological processes and lead

to tumor formation and development [Wang et al., 2014, Gabay et al., 2014, Patki et al., 2013]. While numerous studies have advanced our understanding of the roles of miRNAs and TFs in cancer pathology, not much is known on how miRNAs and TFs are coordinated in regulating cancer-related biological functions.

An oncogenic process often involves many genes, and those genes can be coregulated by multiple miRNAs and TFs [Martinez and Walhout, 2009, Hayes et al., 2014]. Being able to identify such functional gene regulatory modules, which consist of miRNAs, TFs, and their coregulated genes can further our understanding of gene regulation mechanisms in cancer biology. Because of the many-to-many relation between genes and their regulators (i.e., miRNAs and TFs), and how gene regulation can vary significantly in different cellular conditions, deciphering the coregulatory relationship between TFs, miRNAs to identify context-specific gene regulatory modules is a challenging problem. To that end, in Chapter 3, we propose a computational method called CanMod, which aims at identifying ***Can***cer-associated Gene Regulatory ***Mod***ules. CanMod was applied to the breast cancer dataset from TCGA. In brief, CanMod was able to infer gene modules that are significantly associated with cancer-related biological processes and pathways. CanMod is a valuable tool for researchers to gain understanding of the interplay between miRNAs and TFs in regulating cancer-related biological processes.

1.6 Datasets and Data Preprocessing

Before discussing the two computational methods in detail in Chapters 2 and 3, this section describes the what datasets were used and how the data were preprocessed to be used as input for the two computational methods.

1.6.1 Datasets

Both of the methods were applied to real cancer datasets from The Cancer Genome Atlas (TCGA) [Grossman et al., 2016]. TCGA is a public database that stores various biological and clinical data types of both normal samples and tumor samples of over 30 cancer types. TCGA uses different high-throughput techniques to analyze the cancer patient samples. Those techniques include DNA sequencing, gene expression (mRNA and miRNA) profiling, CNA profiling, and DM profiling. Since both Cancerin and CanMod leveraged those data types, TCGA provides valuable datasets to apply the two methods.

We used the R Bioconductor package TCGABiolinks [Colaprico et al., 2016] to download the genomic data of normal and solid tumor tissues for three types of cancer from TCGA [Grossman et al., 2016]. The cancer types included breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), and head and neck squamous cell carcinoma (HNSC). For those cancer types, we retrieved mRNA and miRNA gene expression data, CNA data, and DM data. Expression of lncR-

NAs was retrieved from the TANRIC database [Li et al., 2015]. We also retrieved the survival data (number of days until death) of the cancer patients from whom the genomic data were collected.

In addition, the 3'UTR sequences of 18,959 mRNAs and 13,870 lncRNAs were downloaded from the GENCODE Release 26 (GRCh38.p10) [Harrow et al., 2012], and the sequences of 2,588 mature miRNAs and were downloaded from miRBase release 21 [Kozomara and Griffiths-Jones, 2013].

In both Cancerin and CanMod, putative regulator-target gene interactions (i.e., miRNA-target and TF-target interactions) were used to reduce the search space of regulators for each target gene. Putative miRNA-mRNA interactions were retrieved from StarBase v2.0 [Li et al., 2013] and TargetScan 7.1 [Agarwal et al., 2015] databases. StarBase predicts miRNA-RNA interactions based on inferring the direct miRNA-RNA binding events from 108 CLIP-seq datasets [Li et al., 2013]. TargetScan predicts an mRNA to be a miRNA's target if the mRNA contains binding sites that are complementary to the seed regions of the miRNA [Agarwal et al., 2015]. Appendix A provides details about common methods to determine putative miRNA-target gene interactions.

Putative miRNA-lncRNA interactions were retrieved from Starbase v2.0 [Li et al., 2013], DIANA-LncBase v2 [Paraskevopoulou et al., 2015], and LnCeDb [Das et al., 2014]. The putative miRNA-lncRNA interactions in the DIANA-LncBase v2

are inferred using the DIANA-microT algorithm [Paraskevopoulou et al., 2013]. It is a machine-learning approach that estimates miRNA-RNA target binding score based on weighting multiple features such as sequence complementary, free binding energy, and conservation profile [Paraskevopoulou et al., 2013]. The putative miRNA-lncRNA interactions in the LnCeDb are aggregated from two sources. The first source includes interactions from the Mircode database [Jeggari et al., 2012], which uses seed complementarity and evolutionary source to infer interactions. The second source includes interactions inferred by its own sequence-based miRNA-RNA target prediction algorithm, which is a speed-up version of the Smith–Waterman sequence alignment algorithm [Smith and Waterman, 1981].

To assess the sequence complementarity between miRNAs and their mRNA/lncRNA targets, we retrieved the sequences of 2,588 mature miRNAs from the miRBase release 21 [Kozomara and Griffiths-Jones, 2013]. The 3'UTR sequences of 18,959 mRNAs and the sequences of 13,870 lncRNAs were retrieved from the GENCODE Release 26 (GRCh38.p10) [Harrow et al., 2012].

The putative TF-gene interactions are retrieved from the TRED [Zhao et al., 2005] and TTRUST (version 2) [Han et al., 2015] databases. TRED is an integrated repository for TF binding events to both enhancer and promoter regions in mammals [Zhao et al., 2005]. The binding events in TRED are curated through a literature text mining algorithm. TTRUST is also a repository of putative TF-gene

interactions, which is constructed using a sentence-based text mining algorithm, followed by manual curation.

1.6.2 Data preprocessing

This section describes how the data were preprocessed to be used as input for the Cancerin and CanMod.

1.6.2.1 Gene expression preprocessing and differential expression analysis

Expression of a gene is quantified as the overall abundance of the gene's transcripts. RNA-Seq is a high-throughput sequencing technology that is used to quantify genome-wide transcript abundance [Kukurba and Montgomery, 2015]. A transcript abundance is quantified by the number of sequenced reads, called raw counts, that are aligned and mapped to the transcript [Kukurba and Montgomery, 2015].

From TCGA, we retrieved the Illumina HiSeq 2000 sequencing data, which provided the raw count values of genome-wide mRNAs and miRNAs. The presence of low-count mRNAs and miRNAs can decrease the sensitivity of different statistical analyses used in Cancerin and CanMod. The low-count RNAs are defined as the RNAs that were not expressed in the majority of samples [Robinson et al., 2010]. We used the R Bioconductor package edgeR [Robinson et al., 2010] to filter

out low-count RNAs. First, the raw count values of each gene were converted to counts-per-million (CPM) values, which handles the library size bias between samples [Robinson et al., 2010]. Then, to remove the RNAs that were not expressed in the majority of samples, across all the samples for each cancer dataset, an RNA was filtered out if its CPM value was less than 1 in more than t samples, where t was set to the larger between the tumor and the normal group size.

We also used the edgeR package to identify differentially expressed (DE) mRNAs and DE miRNAs between the normal and the tumor samples. First, the package was used to model count data with a negative binomial (NB) distribution. After the data was fitted under NB models, we applied the Fisher's exact test to identify DE mRNAs and DE miRNAs [Robinson et al., 2010]. As expression of lncRNAs was in RPKM units and was normalized to follow a normal distribution, to find DE lncRNAs, we fitted a linear model for each lncRNA using the lmFit function in R package limma [Ritchie et al., 2015]. A miRNA, mRNA, or lncRNA was considered to be DE if its adjusted Bonferroni-Hochberg p-value [Benjamini and Hochberg, 1995] was smaller than 0.01.

To ensure the expression of the DE mRNAs, miRNAs, and lncRNAs was in the same units, we converted raw counts of DE mRNAs and DE miRNAs to reads-per kilobase-million (RPKM) values. We used log2(RPKM+0.001) to present the expression of all DE RNAs. Expression of those RNAs were z-normalized across all

the tumor samples since we only used the tumor samples after the preprocessing step.

1.6.2.2 Copy number alteration (CNA)

We downloaded the level 3 CNA data (Affymetrix SNP Array 6.0) from TCGA. The CNA data provides estimated mean copy numbers of chromosomal segments in the whole genome. Using the genomic location information of 22,310 protein coding genes provided by GENCODE Release 26 (GRCh38.p10), we applied the R Bioconductor package CNTools [Zhang, 2016] to convert the segmented CNA data into a gene-level data matrix, where each entry represented the copy number value of a gene in a specific sample.

1.6.2.3 DNA Methylation (DM)

We downloaded the level 3 DM data (Infinium HumanMethylation450 BeadChip) from TCGA. The DM data measures the methylation level of approximately 450,000 CpG sites genome-wide. The methylation level of each CpG site (i.e., β value) was estimated as the ratio of the methylated probe intensity to the overall intensity (sum of methylated and unmethylated probe intensities). Thus β ranges between 0 (hypomethylated) and 1 (hypermethylated). Previous studies indicated that the methylation of CpG sites in promoter regions were associated with gene

expression change [Nagae et al., 2011, Fernandez et al., 2012]. Therefore, we only considered β values of CpG sites in genes' promoter regions. To compute gene-centric methylation values, we used the Bioconductor annotation package IlluminaHumanMethylation450kanno.ilmn12.hg19 [Hansen, 2015] to identify the probes positioned at the genes' promoter regions (upstream 200 to 1500 base pairs away from of gene transcription start site). A gene's methylation level was estimated as the mean of its associated upstream probes' β values.

The next two chapters will provide detailed descriptions of the two computational methods and the biological findings obtained from applying the methods to the cancer datasets.

CHAPTER 2

CANCERIN: A COMPUTATIONAL METHOD TO IDENTIFY CANCER-ASSOCIATED COMPETING ENDOGENOUS RNA INTERACTIONS MEDIATED BY MIRNA REGULATION

(This chapter is adapted from the research article [Do and Bozdag, 2018]. The article was accepted for publication in PLoS Computational Biology in June, 2018. We have full permission to reuse the article's contents in this chapter.)

2.1 Abstract

MicroRNAs (miRNAs) inhibit expression of target genes by binding to their RNA transcripts. It has been recently shown that RNA transcripts targeted by the same miRNA could "compete" for the miRNA molecules and thereby indirectly regulate each other. Experimental evidence has suggested that the aberration of such miRNA-mediated interaction between RNAs – called competing endogenous RNA (ceRNA) interaction – can play important roles in tumorigenesis. Given the difficulty of deciphering context-specific miRNA binding and the existence of various gene regulatory factors such as DNA methylation and copy number alteration, inferring context-specific ceRNA interactions accurately is a computationally challenging task. Here we propose a computational method called Cancerin to identify cancer-associated ceRNA interactions. Cancerin incorporates DNA methylation (DM), copy number alteration (DM), gene and miRNA expression datasets to construct cancer-specific ceRNA networks. We applied Cancerin to three can-

cer datasets from the Cancer Genome Atlas (TCGA) project. Our results indicated that ceRNAs were enriched with cancer-related genes, and ceRNA modules in the inferred ceRNA networks were involved in cancer-associated biological processes. Using LINCS-L1000 shRNA-mediated gene knockdown experiment in breast cancer cell line to assess accuracy, Cancerin was able to predict expression outcome of ceRNA genes with high accuracy.

2.2 Motivation and Related Work

As a miRNA can regulate multiple targets, and a target be simultaneously be regulated by multiple miRNAs, the regulatory network formed by miRNAs and their target genes is complex. Many layers of biological knowledge still is hidden in this complex miRNA-target gene regulatory network. Proposed by Pandofi in 2011, the ceRNA hypothesis, which posited that two RNA transcripts can indirectly regulate each other via their direct interactions with common miRNAs, has gained much attention, since this novel gene regulation can be involved in many crucial biological processses [Salmena et al., 2011]. CeRNA interactions have been shown to regulate important biological processes such as muscle differentiation [Cesana et al., 2011], self-renewal capability of embryonic stem cells [Jovanovic and Hengartner, 2006], and inhibition of cancer cell differentiation [Zhou et al., 2014a]. Aberrance of ceRNA interactions has been reported to be associated with tumorge-

nesis in multiple types of cancer [Li et al., 2017, Yang et al., 2014, Sanchez-Mejias and Tay, 2015, Karreth and Pandolfi, 2013, Tay et al., 2014].

The existence and strength of ceRNA interactions may vary significantly in different physiological and cellular settings (i.e., normal cells versus tumor cells). As ceRNA interaction is considered as a new layer of gene regulation, identification and construction of genome-wide and condition-specific ceRNA interaction networks could facilitate better understanding of ceRNA regulatory mechanisms and their biological significance. While experimental studies are of great importance to confirm ceRNA interactions, inference of ceRNA interaction networks by only experimental methods is time- and cost-prohibitive. Thus, computational tools are needed to infer ceRNA interaction networks and to generate new hypotheses for further experimental validation.

Since ceRNA interactions are mediated via miRNAs, identifying interactions between miRNAs and their targets is a prerequisite to infer ceRNA interactions. Sequence-based miRNA target prediction algorithms such as TargetScan [Agarwal et al., 2015] and miRanda [John et al., 2004] have been employed to search for miRNA-response-element (MREs) in 3'UTR of mRNAs, and miRNA-mRNA interaction databases such as StarBase [Li et al., 2013] and miRWalk [Dweep et al., 2011] store computationally and experimentally verified miRNA-mRNA interactions. Expression profiles of both mRNAs and miRNAs also were used to identify

condition-specific miRNA-mRNA interactions. As miRNAs were mostly known to repress the expression of its targets, expression levels of miRNAs and their targets were often required to be negatively correlated [Zhou et al., 2014b, Shao et al., 2015].

After predicting miRNA-target gene interactions, existing ceRNA inference methods differed in how they related expression of miRNAs and their coregulated genes to decide which genes can establish ceRNA interactions. Pair-wise gene expression correlation often were considered as the main criteria to select ceRNA interactions. Two ceRNAs were required to have positively correlated expression, and the ceRNAs and their miRNA regulators were required to have negatively correlated expression [Zhou et al., 2014b, Shao et al., 2015]. However, miRNA expression data were used to model the mediating effect of miRNAs in regulating ceRNA interaction. Partial Pearson correlation (PPC) [Paci et al., 2014] and conditional mutual information (CMI) [Sumazin et al., 2011, Chiu et al., 2015] metrics have been used to measure linear or nonlinear dependence of candidate ceRNAs' expression on their shared miRNAs' expression. Applying CMI to identify and construct a glioblastoma-specific ceRNA interaction network, Sumazin et al. found experimentally-validated interactions between PTEN and their known ceRNAs in the ceRNA network [Sumazin et al., 2011]. In [Paci et al., 2014], a new metric called sensitivity partial correlation was proposed to quantify the expression cor-

relation dependency between two ceRNAs conditioned on their shared miRNAs' expression. The researchers applied this metric to gene and miRNA expression of normal and tumor breast samples to construct normal-specific and tumor-specific ceRNA interaction networks. They observed that multiple cancer hallmarks such as tumor inflammation were only enriched in the tumor-specific ceRNA network. A detailed review of computational methods to infer ceRNA interactions can be found in [Le et al., 2016].

In existing ceRNA studies, most computational methods considered miRNAs as the only type of gene regulators, while overlooking other important types of gene regulators (e.g., TF, DM, and CNA). Not considering other types of regulators might lead to spurious miRNA-gene interactions, which would cause false positive predictions of ceRNA interactions. Notably, lack of experimental studies to confirm ceRNA interactions posed a big challenge to validate the accuracy and significance of inferred ceRNA interactions.

This chapter presents a computational pipeline called Cancerin, which infers ***Can***cer-associated ***ceR***NA ***i***teraction ***n***etworks. A cancer-associated ceRNA interaction is defined as an interaction between two DE RNAs (between normal and cancer samples), and the interaction is mediated by some DE miRNAs that regulate both RNAs. besides mRNAs, non-coding RNAs such as long non-coding RNAs (lncRNAs) have been shown to actively participate in functionally important

ceRNA interactions in both normal and cancer cells [Sanchez-Mejias and Tay, 2015, Tay et al., 2014]. Thus, Cancerin considers both mRNAs and lncRNAs as potential ceRNAs. Cancerin employs knowledge from both putative miRNA-RNA interactions and miRNA/RNA expression profiles. In addition, Cancerin incorporates other types of gene expression regulatory factors, namely CNA, DM, and TF.

In brief, input data for Cancerin include expression of miRNAs, lncRNAs, miRNAs, CNA and DNA of each mRNAs, and putative interactions between miRNA-mRNA, miRNA-lncRNA, and TF-mRNA. Cancerin outputs inferred ceRNA interactions of mRNA-mRNA, mRNA-lncRNA, and lncRNA-lncRNA. R software for Cancerin is freely available (MIT license) at https://github.com/bozdaglab/Cancerin. We believe that Cancerin is an easy-to-use method for both biologists and bioinformaticians to infer ceRNA interactions.

Cancerin was applied to three cancer datasets. The result indicate that the ceRNAs in the obtained ceRNA interaction networks were significantly enriched with cancer-related genes. Additionally, closely connected ceRNAs in the ceRNA networks were associated with cancer cell formation and development processes. Compared to non-ceRNA genes, expression change of predicted ceRNAs had a higher association with cancer survival outcomes. To validate the effect of ceRNA interactions to expression change on an external dataset, we used the LINCS perturbation

dataset [Liu et al., 2015a] and observed that the knockdown of ceRNAs was associated with the expression change of their ceRNA partners.

The rest of the chapter is organized as follows. Section 2.3 recaps the input data and the data processing procedure. Section 2.4 provides a detailed description of each computational step in Cancerin. Section 2.5 discusses the results obtained from applying Cancerin to the three cancer datasets. Section 2.6 summarizes the main components of Cancerin and the key biological findings discussed in the chapter.

2.3 Input Data

As described in Section 1.6, the input data for Cancerin are genome-wide expression of miRNAs, mRNAs, and lncRNAs. Each mRNA is also associated with a CNA value and a DM value. In addition, putative miRNA-mRNA, miRNA-lncRNA, and TF-mRNA interactions are also employed to select candidate regulators for each target RNA transcript. Cancerin was applied to the breast cancer dataset (BRCA), the kidney cancer dataset (KIRC), and the head and neck cancer dataset (HNSC) from TCGA. Section 1.6 describes for how the data were preprocessed to be used as input for Cancerin.

2.4 Cancerin pipeline

Cancerin is a computational pipeline to identify genome-wide cancer-associated ceRNA interaction networks. It consists of three main steps (Fig. 2.1). Using putative miRNA-mRNA and miRNA-lncRNA interactions, the first step constructs an interaction network between DE miRNAs and DE RNAs. In the second step, only the miRNAs that are associated with their targeted RNAs' expression change are kept. In the final step, several filtering layers are applied to infer ceRNA interactions between RNAs that are targeted by common miRNAs. The details in each step in Cancerin are described in the following.

Step 1: Identifying putative regulatory interactions between DE miRNAs and DE mRNAs based on sequence binding

As discussed in Section 1.6, the putative interactions between miRNAs and mRNAs came from the TargetScan 7.1 [Agarwal et al., 2015] and StarBase v2.0 [Li et al., 2013] databases. The putative interactions between miRNAs and lncRNAs came from Starbase v2.0 [Li et al., 2013], DIANA-LncBase v2 [Paraskevopoulou et al., 2015] and LnCeDb [Das et al., 2014]. The output for Step 1 are putative interactions between DE miRNAs and DE RNAs.

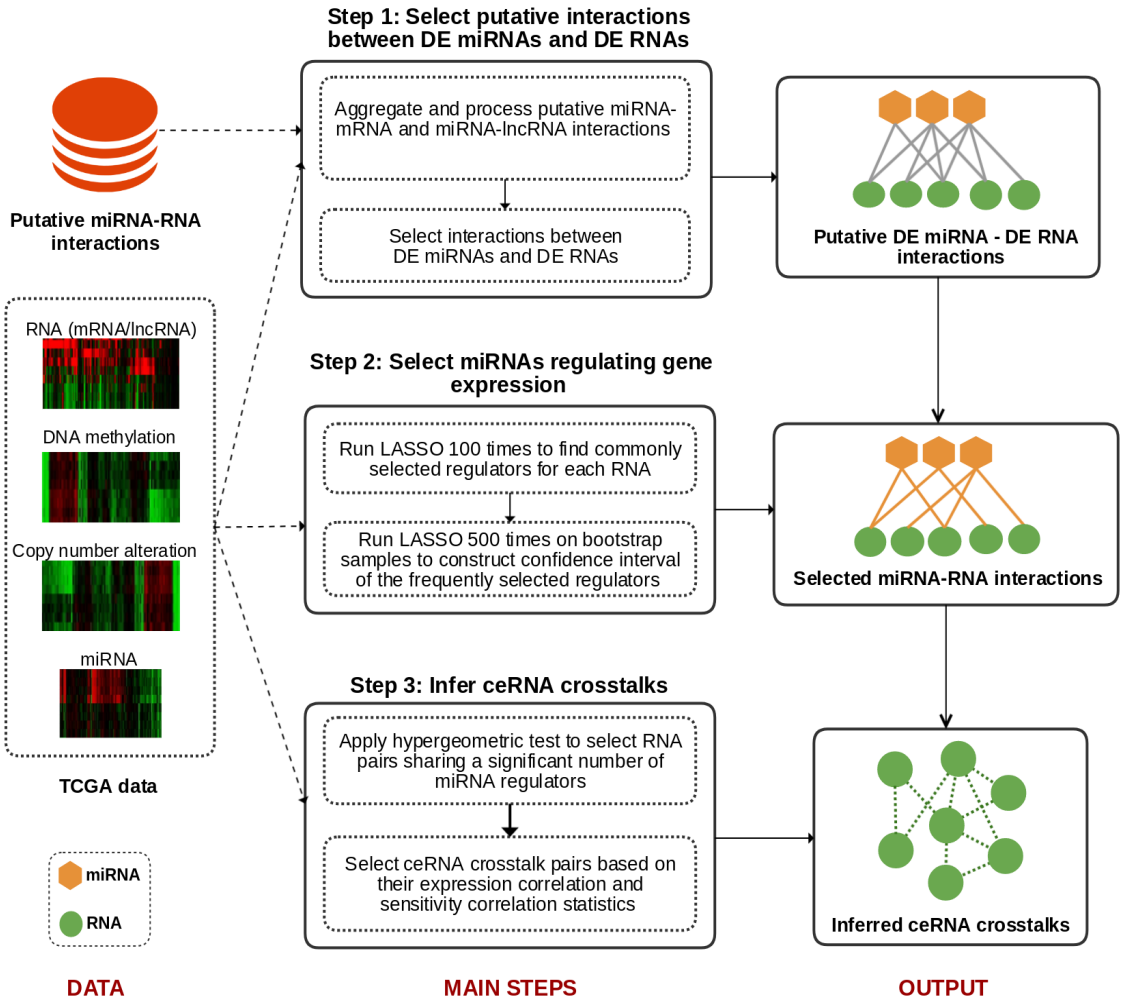

Figure 2.1: **Cancerin pipeline to infer cancer-associated ceRNA interaction networks**. Cancerin consists of three steps. In Step 1, for each DE RNA, Cancerin selects its candidate DE miRNA regulators based on sequence binding results. In Step 2, Cancerin applies a LASSO-based variable selection procedure to select a subset of miRNA regulators that contribute to the expression variation of the DE RNA. In Step 3, Cancerin applies multiple filtering conditions to infer ceRNA interactions between the RNAs that are regulated by common miRNAs.

Step 2: Selecting miRNAs associated with expression change of their predicted RNA targets

Given the putative interactions between DE miRNAs and DE RNAs ob-

tained from Step 1, for each DE RNA and its putative DE miRNA regulators, Cancerin identified which miRNAs contributed to the RNA's expression variation. As mRNA expression can be controlled by TFs, CNA, DM, and miRNA, our LASSO-based variable selection procedure to infer cancer-specific miRNA-mRNA interactions considered all of those regulatory factors as candidate regulators for mRNA expression.

LASSO is a regularized regression method that penalizes the sum of absolute value of the regression coefficients, so that it shrinks some covariates' coefficients to be exactly zero. Hence, it can be used for variable selection purposes [Tibshirani, 1996]. LASSO regression was applied for each RNA. For each mRNA, its expression was used as the response variable's value, and its CNA, DNA methylation, and the expression of its candidate miRNAs and TFs were used as values of the independent variables' values. For each lncRNA, its expression was used as the response variable, and its candidate miRNAs' expression were used as the independent variables' value.

Training a LASSO model requires selecting the regularization hyperparameter λ. To select the optimal λ value, we applied 10-fold cross validation to find the λ value that provided the simplest model such that its cross-validation error was within one standard error of the minimum cross-validation error. Thus, for each RNA_{j}, out of all of its candidate predictors (independent variables), LASSO re-

gression selected a set of non-zero coefficient predictors. We employed R package glmnet [Friedman et al., 2010] to perform LASSO regression.

However, independent variables selected by LASSO have been shown to be inconsistent, especially when sample size gets large [Tibshirani et al., 2013]. To address this problem, we ran the LASSO regression 100 times for each RNA. Only the non-zero coefficient predictors that were selected more than 75 times were considered as frequently selected regulators of the RNA.

Unlike in linear multiple regression, where each independent variable's regression coefficient is associated with a p-value testing the null hypothesis that its coefficient is equal to zero, coefficients of LASSO-selected predictors are not associated with any statistical significance test, so we employed a bootstrap procedure to construct a confidence interval for the frequently selected predictors that were obtained above. Suppose a regulator R_{i} is a frequently selected predictor for RNA_{j}. From the 100 LASSO runs, we used the median of R_{i}'s coefficients to represent its regression coefficient and called it $\overline{\alpha}_{\mathrm{ij}}$. To estimate the confidence interval of $\overline{\alpha}_{\mathrm{ij}}$, for RNA_{j}, we fit LASSO regression 500 times, each time on a set of bootstrapped samples, to generate a bootstrap regression coefficient distribution $\{\alpha_{\mathrm{bootstrap\_ij}}\}$. R_{i} would be kept as one of the RNA_{j}'s regulators if its $\overline{\alpha}_{\mathrm{ij}}$ was within the 95% confidence interval of $\{\alpha_{\mathrm{bootstrap\_ij}}\}$, and the 95% confidence interval did not include 0. As miRNAs are mostly known to repress the expression level of its RNA target, for

each RNA, out of all the retained variables, we only selected the miRNAs that had negative $\overline{\alpha}_{ij}$ coefficients. In brief, after Step 2, each RNA is associated with a set of miRNA regulators selected by the LASSO procedure.

Step 3: Identifying cancer-associated ceRNA interaction networks

Using the miRNA-RNA interactions obtained from Step 2, we generated all possible RNA-RNA pairs such that the constituent RNA in each pair share at least one miRNA regulator. Those pairs were considered as candidate ceRNA pairs. Following the ceRNA hypothesis, we only kept the candidate ceRNA pairs with high positive Pearson expression correlation (correlation ≥ 0.5, p-value < 0.05).

Given the number of miRNAs regulating each RNA, to assess whether the two RNAs in each candidate ceRNA pair shared a significant number of miRNA regulators, we applied a hypergeometric test to each of the candidate ceRNA pairs. Let N be the total number of all DE miRNAs. For a ceRNA pair consisting of RNA_i and RNA_j, let N_i and N_j be the total number of miRNAs regulating RNA_i and RNA_j respectively, and N_{ij} be the number of common miRNAs regulating both RNA_i and RNA_j. The p-value of the hypergeometric test was calculated using the formula in Eq. 2.1. Based on the hypergeometric test results, a candidate

ceRNA pair was selected if its adjusted Bonferroni-Hochberg p-value was smaller than 0.05.

$$p - value = 1 - \sum_{k=0}^{N_{ij}-1} \frac{\binom{N_j}{k}\binom{N-N_j}{N_i-k}}{\binom{N}{N_i}} \tag{2.1}$$

To further eliminate potentially spurious ceRNA pairs, we employed the sensitivity correlation (SC) metric proposed in [Paci et al., 2014] to estimate the ceRNA interaction strength for each ceRNA pair. Let $\{miRNA_{ij}\}$ be the set of common miRNAs regulating both RNA_i and RNA_j. Let $Corr(RNA_i, RNA_j)$ be the expression correlation between RNA_i and RNA_j and $PC(RNA_i, RNA|\{miRNA_{ij}\})$ be the partial expression correlation between RNA_i and RNA_j conditioned on $\{miRNA_{ij}\}$. Sensitivity correlation $SC(RNA_i, RNA_j|\{miRNA_{ij}\})$ is

$$\begin{aligned} SC(RNA_i, RNA_j \mid miRNA_{ij}) = Corr(RNA_i, RNA_j) \\ - PC(RNA_i, RNA_j \mid \{miRNA_{ij}\}). \end{aligned} \tag{2.2}$$

The the R package bnlearn [Scutari, 2009] was used to compute partial correlation (PC) for each candidate ceRNA pair. Since $PC(RNA_i, RNA_j|\{miRNA_{ij}\})$ computed the correlation of the RNAs' expression while controlling/eliminating the effect of their shared miRNAs' expression, $SC(RNA_i, RNA_j \mid \{miRNA_{ij}\})$ quantifies the contribution of the shared miRNAs to the linear relation between

expression of the two RNAs. A high SC value signifies a strong indirect interaction between the two RNAs mediated by their shared miRNA regulators. Thus, we selected the ceRNA pairs with positive SC values and p-values from partial correlation test smaller than 0.05. Additionally, to estimate the statistical significance of SC, we computed the SC empirical p-value for each candidate ceRNA pair. For the pair (RNA_i,RNA_j), suppose the $\{miRNA_{ij}\}$ was of size N_{ij}. Then we randomly selected N_{ij} miRNAs to compute the pair's sampled SC value. For each ceRNA pair, the resampling procedure was repeated 1000 times. An empirical SC p-value was assigned as the percentage of iterations in which the sampled SC value exceeded the original SC value. A ceRNA pair was kept if its empirical SC p-value was smaller than 0.05.

2.5 Results

Our Cancerin pipeline used different types of cancer genomics data to infer cancer-associated ceRNA interaction networks. In assessing the effectiveness of Cancerin, we used Cancerin to infer ceRNA networks in three cancer types: breast (BRCA), kidney (KIRC), and head and neck cancer (HNSC). We obtained the RNAseq, miRNAseq, DNA methylation, and CNA datasets for BRCA, KIRC, and HNSC samples from TCGA [Grossman et al., 2016]. The numbers of normal/tumor

tissue samples in each cancer type were 47/193 (BRCA), 20/243 (KIRC), and 20/413 (HNSC).

2.5.1 Putative DE miRNA - DE RNA interactions (Step 1)

The first step in Cancerin involved aggregating the putative interactions between miRNAs and RNAs from various data sources. The candidate miRNA-mRNA interactions were downloaded from the StarBase and TargetScan databases. Using mRNAs' and miRNAs' FASTA sequences, we selected only the mRNAs whose 3'UTR sequences and the miRNAs whose mature sequences were specified. To further refine those putative interactions, the miRanda algorithm was used to check for the existence of MRE(s) on the mRNAs' 3'UTR and to estimate the thermodynamic folding energy between the miRNAs and their predicted mRNA targets. The lower the energy, the higher chance that an interaction will actually occur [Mathews et al., 1999]. A miRNA-mRNA interaction was kept if there was at least one MRE on the mRNA as miRNA's binding site, and the miRNA-mRNA interaction's folding energy was lower than 140 kcal/mol (default value). After applying miRanda, there remained 465,049 interactions between 473 miRNAs and 13,932 mRNAs. Putative miRNA-lncRNA interactions were aggregated from Starbase v2.0, DIANA-LncBase v2, and and LnCeDb, resulting in 3,961,135 interactions between 2,695 miRNAs and 24,215 lncRNAs.

Given all the putative miRNA-RNA interactions, since we aimed to infer cancer-associated miRNA-RNA interactions, we only kept the interactions between DE miRNAs and DE RNAs. Table 2.1 reported the number of DE miRNA - DE RNA interactions in each cancer type. There were 66 common DE miRNAs and 2,147 common DE RNAs across all the three cancer types. The putative DE miRNA - DE RNA interactions were specific to each cancer type. There were only 15,591 common putative interactions that are included in all the three cancer types.

Table 2.1: Number of putative DE miRNA-DE RNA interactions and number of DE miRNAs and DE RNAs included in those interactions (output for Cancerin - Step 1).

	BRCA	KIRC	HNSC
No. of putative DE miRNA - DE mRNA interactions	153,465	107,348	94,980
No. of DE miRNAs [1]	215	164	201
No. of DE mRNAs [1]	7,502	6,690	5,005
No. of putative DE miRNA - DE lncRNA interactions	60,935	18,589	17,350
No. of DE miRNAs [2]	215	164	201
No. of DE lncRNAs [2]	3,111	1,335	896

[1]: included in putative DE miRNA - DE mRNA interactions.

[2]: included in putative DE miRNA - DE lncRNA interactions.

To identify cancer-associated ceRNA interactions, Cancerin employed the putative miRNA-RNA interactions and RNA expression as input data for the next two steps, which included applying a LASSO-based variable selection procedure to select cancer-specific miRNA-RNA interactions and using that information to identify ceRNA interactions.

2.5.2 Analysis of miRNA-RNA interactions obtained from the LASSO-based variable selection procedure (Step 2)

The LASSO-based variable selection procedure (Cancerin Step 2) was applied to identify cancer-specific miRNA-RNA interactions while also taking into account other types of gene regulators including TF, DM, and CNA. Table 2.2 summarizes the number of miRNA-RNA interactions selected by the variable selection procedure in each cancer type. We found only 44 common miRNA-RNA interactions across all the three cancer types. The result is expected because there were already few common putative miRNA-RNA interactions selected in the previous step.

Table 2.2: Number of selected miRNA-RNA interactions obtained after applying the variable selection procedure (output of Cancerin - Step 2).

	BRCA	KIRC	HNSC
No. of miRNA-mRNA interactions	6,616	8,408	9,893
No. of miRNAs [1]	196	154	190
No. of mRNAs [1]	2,814	2,971	3,020
No. of miRNA-lncRNA interactions	502	217	467
No. of miRNAs [2]	134	93	141
No. of lncRNAs [2]	210	91	175

[1]: included in the selected miRNA - mRNA interactions.

[2]: included in the selected miRNA - lncRNA interactions.

2.5.2.1 Many miRNA-RNA interactions were only identified when different types of gene expression regulators were taken into account

The Cancerin pipeline was constructed under the premise that different

types of gene regulators were important to infer miRNA-RNA interactions correctly. Out of all the RNA targets that were found to have at least one miRNA regulator (3,024 (BRCA), 3,062 (KIRC), and 3,195 (HNSC)), we computed the percentage of those targets that were also under regulation of at least one additional regulatory factor such as CNA, DNA methylation, or TF (Table 2.3). Not surprisingly, those additional regulatory factors, especially CNA, were observed to be associated with expression change in majority of target RNAs.

Table 2.3: Percentage of RNA targets regulated by miRNAs and also by at least one additional type of regulators.

	BRCA	KIRC	HNSC
Percentage of RNA targets under CNA regulation	76.2%	69.2%	77.2%
Percentage of RNA targets under DNA Methylation regulation	30.4%	26.3%	35.0%
Percentage of RNA targets under TF regulation	54.1%	59.3%	48.0%

To check the impact of those additional regulators in inferring miRNA-RNA interactions, we performed a comparative analysis between the miRNA-RNA interactions that were selected in two different cases depending on whether the different regulatory factors besides miRNA (i.e., CNA, DNA methylation, and TF) were present in the LASSO-based variable selection procedure. In the first case when those regulators were incorporated, we referred it as "Cancerin (original)." The second case, in which miRNAs were the only type of regulators to be considered, was

referred as "Cancerin (only_miRNA)." Table 2.4 shows the number of miRNA-RNA interactions and their constituent miRNAs and RNA targets selected in the two cases.

Table 2.4: Number of miRNA-RNA interactions and their constituent miRNAs and RNAs selected in "Cancerin (original)" and "Cancerin (only_miRNA)". The first, second, and third value in each cell refer to the results from "Cancerin (original)", "Cancerin (only_miRNA)", and the common results between the two cases, respectively.

	BRCA	KIRC	HNSC
No. of miRNA-RNA interactions	7,118/4,071/3,242	8,625/6,524/5,085	10,360/8,648/6,619
No. of miRNAs	204/201/198	155/153/153	195/196/195
No. of RNAs	3,024/1,763/1,523	3,062/2,219/2,068	3,195/2,520/2,404

While the two cases selected similar miRNAs that have at least one RNA target (row 2 in Table 2.4), many miRNA-RNA interactions and RNA targets could only be found in "Cancerin (original)" (row 1 and 3 in Table 2.4). To check how the additional regulatory factors besides miRNAs played a role in that distinction, we looked at the common RNA targets that were included in both "Cancerin (original)" and "Cancerin (only_miRNA)", and compared them with the RNA targets that were uniquely found in "Cancerin (original)." Among the common RNA targets, the percentage of RNAs that had at least one additional regulator in "Cancerin (original)" results was 78.2% (BRCA), 83.8% (KIRC), and 85.2% (HNSC).

Among the RNA targets unique to "Cancerin (original)," the percentage values increased to 97.6% (BRCA), 96.7% (KIRC), and 97.1% (HNSC). These results suggest that while "Cancerin (only_miRNA)" could still discover some RNA targets that were regulated by an additional regulatory factor besides miRNAs, there were RNAs that could only be found to be regulated by miRNAs when different types of regulatory factors were incorporated in the variable selection step. Thus, while inferring miRNA-RNA interactions, it is important to include the different types of regulatory factors since certain miRNA-RNA interactions can only be found when the other regulatory factors are considered.

2.5.2.2 Hub miRNA regulators were known to be associated with cancer

In all three cancer types, there were miRNAs that regulated many RNA targets, which made those miRNAs common mediators in multiple ceRNA interactions. The miRNA regulators with highest number of RNA targets in each cancer type were let-7a-5p (BRCA), miR-106b-5p (KIRC), and miR-9-5p (HNSC), which contributed to 2.5%, 3.6%, and 2.5% of total miRNA-RNA interactions, respectively. Let-7a-5p was downregulated in the BRCA dataset (log fold change (FC) = -0.42, False Discovery Rate (FDR) = 7e-4). Known as a tumor-suppressor, let-7a-5p downregulation was shown to cause disruption of crucial signaling pathways,

including Janus protein tyrosine kinase (JAK) and signal transducer [Wang et al., 2012], which can lead to tumor cell migration and invasion in breast cancer [Kim et al., 2012, Liu et al., 2015b]. In the KIRC dataset, miR-106b-5p was upregulated (logFC = 1.5, FDR = 6e-19). Upregulation of this miRNA can enhance activation of the PI3K signaling pathway and promote tumor cell metastasis in KIRC [Zhang et al., 2015]. In the HNSC dataset, miR-9-5p was highly upregulated (logFC = 3.37, FDR = 5e-06). Upregulation of the miR-9 family was known to activate oncogenic pathways in multiple cancers such as leukemia, breast, and colon cancer [Chen et al., 2013]. Interestingly, miR-130-3p was among the top five miRNAs that had highest number of RNA targets in all the three cancer types. Aberration in gene regulation by the miR-130 family was known to drive tumorgenesis in many cancer types including BRCA, KIRC, and HNSC [Hamilton et al., 2013].

2.5.2.3 Selected miRNA-mRNA interactions included cancer-associated miRNA-mRNA interactions

To test if our variable selection procedure to identify miRNA-mRNA interactions was able to detect known cancer-associated miRNA-mRNA interactions, we retrieved 2,259 cancer-related miRNA-mRNA interactions from the oncomiRDB database [Wang et al., 2014]. Each miRNA-target interaction curated in oncomiRDB meets two conditions: (1) the miRNA is invo0ved in at least one cancer-related phenotype or cellular process, and (2) the mRNA is a known oncogene or tumor-

suppressor. As our method only used DE miRNAs and DE mRNAs as input, we only selected the interactions in oncomiRDB in which both miRNAs and mRNAs were also DE miRNAs and DE mRNAs.

We observed that several miRNA-mRNA interactions in oncomiRDB also were included in the miRNA-mRNA interactions inferred by Cancerin (Step 2). We performed a hypergeometric test between the oncomiRDB interactions and inferred miRNA-mRNA interactions to test whether they shared a significant number of interactions. For each cancer type, the background sets in the hypergeometric test consisted of all possible pairs between DE mRNAs and DE miRNAs. The numbers of overlapping interactions and their p-values from the hypergeometric test in BRCA, KIRC, and HNSC were 50 (p-value = $1.75E^{-39}$), 40 (p-value = $4.6E^{-24}$), and 49 (p-value = $1.7E^{-32}$), respectively. We also performed the same hypergeometric test between the sequence-based miRNA-mRNA interactions (Cancerin - Step 1) and the oncomiRDB interactions. The sequence-based interactions also had significant enrichment in oncomiRDB interactions (p-values ≈ 0 in all three cancer types).

2.5.3 Analysis of the inferred ceRNA networks (Step 3)

In Cancerin (Step 3), given all the miRNA-RNA interactions obtained after applying the LASSO-based variable selection procedure, we identified all the

candidate ceRNA interactions in which both the constituent RNAs were regulated by at least one common miRNA. Then we applied several filtering layers to select the final ceRNA interactions out of those candidate ceRNA pairs. Two RNAs were considered to have a ceRNA interaction if they had a significant number of shared miRNAs, and their expression profiles were both significantly correlated (correlation ≥ 0.5, p-value < 0.05) and had significantly positive sensitivity correlation (empirical p-value < 0.05). Table 2.5 summarizes the number of ceRNA interactions and the constituent ceRNAs in those interactions for each cancer type.

Overall, the selected ceRNA interactions were very specific to each cancer type. We found only one common ceRNA interaction in all the three cancer types. The number of common ceRNA interactions between any two cancer types was also very low (9 between BRCA and KIRC, 22 between BRCA and HNSC, and 32 between KIRC and HNSC). In all three cancer types, almost all ceRNA interactions were between mRNAs (84% (BRCA), 99% (KIRC), and 95% (HNSC)). In BRCA and HNSC, many lncRNAs that were involved in lncRNA-lnRNA ceRNA interactions also participated in mRNA-lncRNA ceRNA interactions. Specifically, out of 57 lncRNAs (BRCA) and 20 lncRNAs (HNSC) involved in lncRNA-lncRNA ceRNA

Table 2.5: Number of inferred ceRNA interactions and number of ceRNAs in those interactions (output of Cancerin - Step 3).

	BRCA	KIRC	HNSC
No. of all ceRNA interactions	4,115	4,639	2,725
No. of mRNA-mRNA ceRNA interactions[1]	3,674	4,614	2,589
No. of mRNA-lncRNA ceRNA interactions[1]	394	25	121
No. of lncRNA-lncRNA ceRNA interactions[1]	47	0	15
No. of all ceRNAs	1,593	1,081	1,110
No. of mRNAs as ceRNAs[2]	1,491	1,071	1,063
No. of lncRNAs ceRNAs[2]	102	10	47

[1]: subset of all ceRNA interactions (Row 1)

[2]: subset of all ceRNAs (Row 5)

interactions, 41 (BRCA) and 14 (HNSC) of those lncRNAs also participated in mRNA-lncRNA ceRNA interactions.

2.5.3.1 Inferred ceRNA networks were scale-free and independent from protein-protein interactions (PPI) and TF-gene interactions

Biological networks usually exhibit a scale-free property, meaning that some nodes have more connections than the others [Ma'ayan, 2011]. To check if the inferred ceRNA networks were scale-free, we computed the degree probability distribution function of each ceRNA network. Following the power-law rule [Girvan and Newman, 2002], we fitted linear regression of log(ceRNA's degree probability) to log(ceRNA's degree). Log-log plots of all three ceRNA networks had negative slope

with high fitness, which clearly indicated that the inferred ceRNA networks were scale-free as shown in Fig. 2.2.

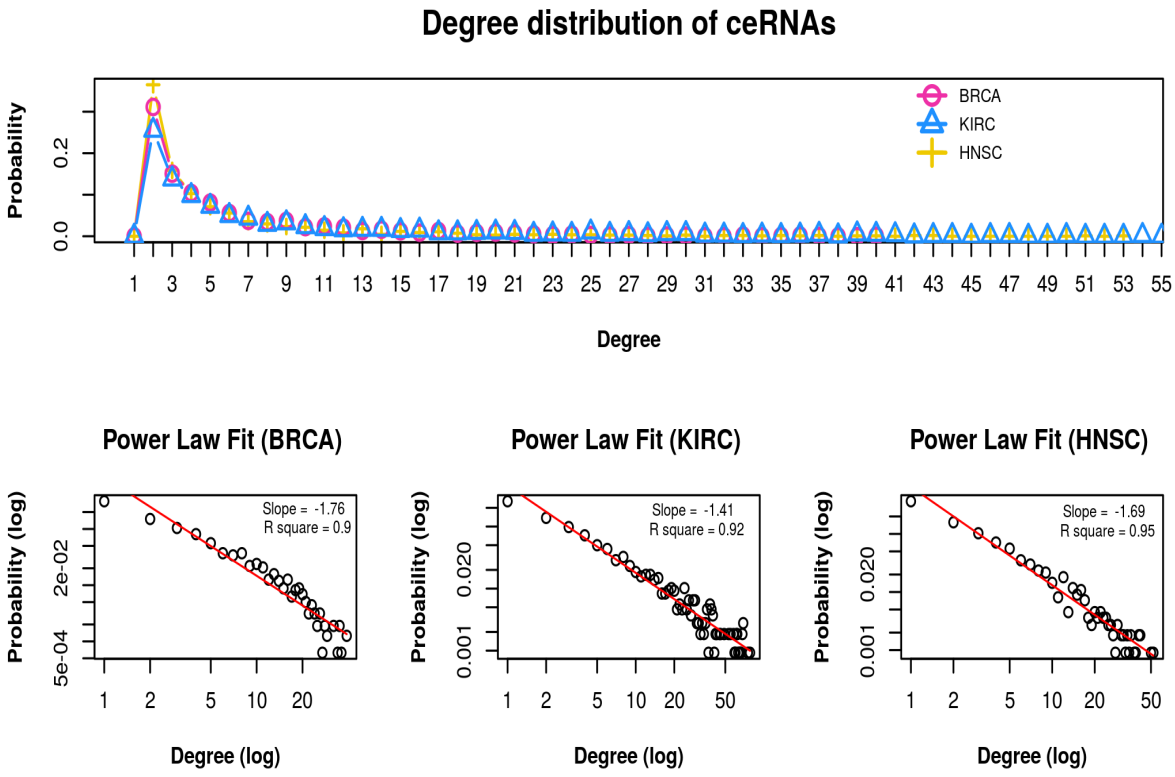

Figure 2.2: **Degree distribution and power-law statistics of the inferred ceRNAs.** (A) Degree distribution of ceRNAs for each cancer type. Linear regression statistics between log(ceRNA's degree) and log(ceRNA's degree probability) in (B) BRCA, (C) KIRC, and (D) HNSC cancer types.

Two genes can interact and thereby regulate each other via different regulatory layers (e.g., protein-protein interactions (PPIs) and TF-gene interactions). To test the specificity of Cancerin to identify ceRNA interactions, we checked whether the inferred ceRNA interaction networks also contained TF-gene interactions or PPIs. We collected 410,337 PPIs from the BioGrid database version 3.4.159 [Stark

et al., 2006]. Within the total number of inferred ceRNA interactions in each cancer network, very few interactions were PPI (0.85% (BRCA), 0.63% (KIRC), and 0.73% (HNSC)). Similarly, we also found very few ceRNA interactions that were also TF-gene interactions (0.78% (BRCA), 0.09% (KIRC), and 0.18% (HNSC)).

2.5.3.2 CeRNAs were significantly associated with cancer-related genes

To test whether the ceRNAs in the inferred ceRNA networks were enriched in cancer-associated genes, we compiled a list of cancer-related genes (oncogenes and tumor-suppressor genes) from the Cancer Gene Census in COSMIC v83 [Forbes et al., 2016], the Bushman lab's Cancer Gene List v3 [Bushman], and the Network of Cancer Genes 5.0 [An et al., 2015], resulting in 2,944 cancer-related genes in total. We performed a hypergeometric test between the inferred ceRNAs in each cancer type with the cancer-related gene list. The results showed that ceRNAs were significantly enriched in the cancer-related genes (p-values were 4.3e-4 (BRCA), 5.0e-3 (KIRC), and 1.9e-5 (HNSC)). We also performed a hypergeometric test between the DE RNAs that were not predicted to be ceRNAs (i.e., non-ceRNAs) and the cancer-related genes. In all three cancer types, unlike the ceRNAs, the non-ceRNAs did not show significant enrichment with the cancer-related genes (p-values ≈ 1 in all three cancer types).

To explore the significance of lncRNAs which were ceRNAs, we analyzed the degree of connection of lncRNAs in the ceRNA networks. A hub ceRNA in the network was defined as the ceRNAs which had high degree (i.e., top 90% edge connection) in the ceRNA network. Within of hub ceRNAs in each cancer, we found a small number of hub lncRNAs (11 (BRCA), 0 (KIRC), and 2 (HNSC)). Interestingly, MAGI2-AS3 was a hub lncRNA in both BRCA and HNSC, and it was also the lncRNA with the highest degree in both BRCA and HNSC ceRNA interaction networks. Among the MAGI-AS3's ceRNA partners, 25% (BRCA) and 35% (KIRC) of them were cancer-associated genes. Recently, MAGI2-AS3 was shown to play an important role in tumorigenesis and tumour progression in breast cancer [Yang et al., 2018]. These result suggests that while lncRNAs contributed to a small number of ceRNA interactions, the hub lncRNAs may hold important functions in cancer biology.

2.5.3.3 CeRNAs were potential biomarkers for cancer prognosis

To assess the prognostic power of the ceRNAs, we tested if the ceRNAs were better than the non-ceRNAs (i.e., DE genes not in the ceRNA network) at predicting survival status of cancer patients. A Univariate Cox proportional hazard model was fit for each DE RNA, which was either a ceRNA or a non-ceRNA. The response variable was the number of days until death for each patient. The patients

who were alive or had no death record were censored, and their last follow-up dates were used.

After hazard model fitting, each DE RNA was associated with a hazard ratio and a p-value (from testing the null hypothesis that its hazard ratio equals to 1). A hazard ratio > 1 implies that an increase of expression of the gene increases the risk of death, while a hazard ratio < 1 implies that an increase of the gene expression decreases the risk of death. Thus, the prognostic power of a gene is reflected through how much its hazard ratio is deviated from 1 (i.e., |hazard ratio - 1|).

A DE RNA was considered as potential prognostic biomarker if its Cox proportional hazard ratio's p-value was smaller than 0.05. Fig. 2.3 shows the hazard ratio distribution of prognostic ceRNAs versus prognostic non-ceRNAs for each cancer type. The Wilcoxon rank-sum test was applied to test if the hazard ratio of prognostic ceRNAs and non-ceRNAs came from the same distribution. In BRCA, we observed a marginal Wilcoxon p-value (0.10). However, the median ceRNAs' hazard ratio was high (1.54), signifying that an increase of BRCA ceRNAs' expression was associated with increased risk of death event. The Wilcoxon p-values for KIRC (1.4e-35) and HNSC (0.03) were both significant. Notably, in all three cancer types, compared to non-ceRNAs' hazard ratios, ceRNAs' hazard ratios were deviated from 1 with higher magnitude, which suggests that ceRNAs hold higher

prognostic power than non-ceRNAs. We observed that the hazard ratio of prognostic ceRNAs were smaller than 1 in KIRC while prognostic ceRNAs in BRCA and HNSC were higher than 1. This result indicates that prognostic ceRNAs in KIRC were more likely to be involved in tumor suppressor-related activities, while prognostic ceRNAs in BRCA and HNSC were more likely to be involved in oncogene-related activities.

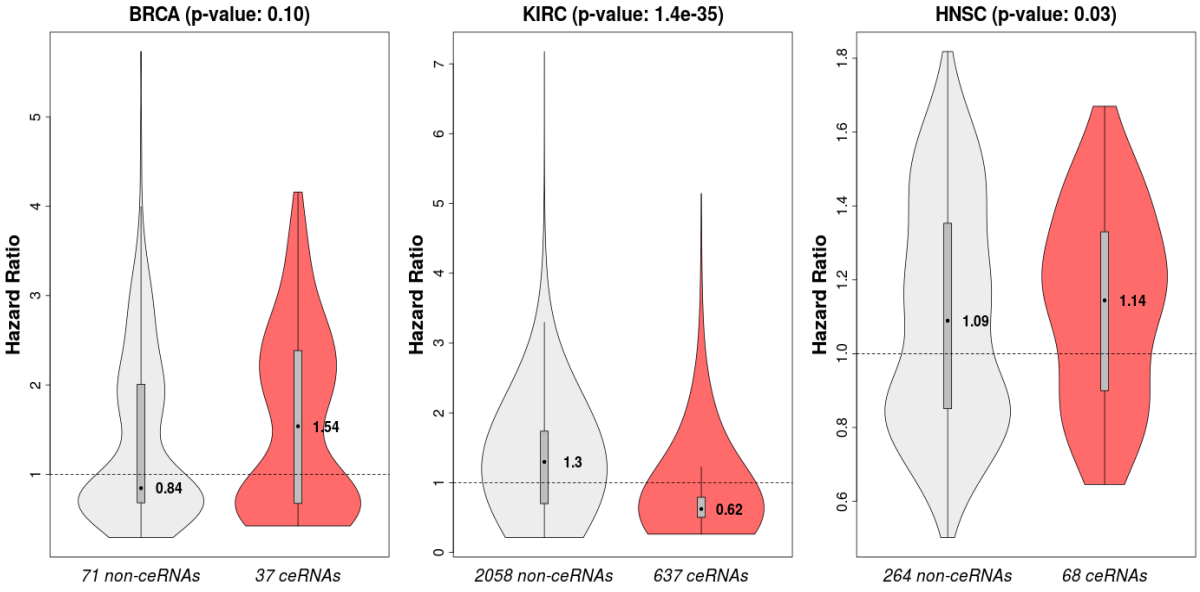

Figure 2.3: **Hazard ratio distribution of prognostic ceRNAs and non-ceRNAs.** A prognostic RNA was defined as a DE RNA whose p-value from univariate Cox regression was smaller than 0.05. For each cancer type, prognostic RNAs were categorized into ceRNAs and non-ceRNAs. The p-values shown in the plot were from the Wilcoxon rank-sum test between hazard ratios of prognostic ceRNAs and non-ceRNAs.

2.5.3.4 CeRNA modules were enriched with cancer processes

To examine the biological significance of the inferred ceRNA networks, we clustered each ceRNA network into modules and performed functional enrichment on each module. A ceRNA module was defined as a sub-network of densely connected ceRNAs. We hypothesized that the ceRNA modules, which were extracted from the inferred ceRNA networks, may act as functional units and play an important role in cancer development. To identify ceRNA modules in each ceRNA network, we employed the R package igraph [Csardi and Nepusz, 2006] to implement the multilevel graph clustering algorithm [Djidjev, 2007]. The algorithm identifies densely-connected modules within a network by using a greedy approach that aims to maximize the module's modularity, which measures the density of connections inside the modules compared to connections between the modules. In each iteration, each vertex is assigned/reassigned to a module to maximize the module's modularity. When no vertex can be reassigned, each module is considered a vertex, and the process is restarted and would be stopped when only a single vertex is left or when the modularity could not be increased. Therefore, the algorithm does not require users to specify the number of modules in advance. When applied to large networks (>100k nodes), the algorithm was able to return modules of high modularity without over-merging or over-dividing those modules [Djidjev, 2007].

To explore the modules' functional importance, we performed enrichment

analysis between the ceRNAs in each ceRNA module and Cancer Hallmark (CH) terms, Gene Ontology (GO) terms, and KEGG/REACTOME pathways. To make an enrichment test statistically feasible, only modules with at least 10 ceRNAs were used for this analysis. The R package clusterProfiler [Yu et al.] was used to perform the enrichment analysis.

The number of ceRNA modules containing more than 10 ceRNAs for each cancer type was 18 (BRCA), 11 (KIRC), and 14 (HNSC). The average number of ceRNAs in each module was 74 (BRCA), 87 (KIRC), and 55 (HNSC). Table 2.6 lists the CH terms that were enriched with the ceRNA modules in each cancer type. Notably, the CH term "Epithelial To Mesenchymal Transition" was enriched in all three cancer types. The CH terms that were enriched in at least two cancer types included "G2M checkpoint," "E2F targets," "TGF beta signaling," and "MYC Targets V1." In all the three cancer types, there existed ceRNA modules that were associated with multiple CH terms (i.e., modules 3 and 7 in BRCA, modules 4 and 11 in KIRC, and modules 4 and 7 in HNSC). The same ceRNA modules were also enriched in GO terms and pathways related to regulation of cell division, development, and activation processes. Interestingly, while some ceRNA modules that were not enriched in any CH terms, they were enriched in GO terms and pathways associating with disease development and progression processes. For instance, module 15 in BRCA was enriched in the KEGG pathways related to Parkinson,

Alzheimer, and Huntington diseases and module 2 in KIRC was enriched in GO Terms involving in negative regulation of metabolic process and molecular function.

In brief, we observed multiple cancer hallmark terms, biological processes, and pathways that were significantly enriched in the ceRNA modules across all the three cancer types. The result indicated the functional significance of the ceRNA interaction networks obtained by Cancerin.

Table 2.6: Cancer hallmark terms that were enriched in the ceRNA modules.

Cancer type	Cancer hallmark geneset	Description	Enriched Module
BRCA	Epithelial Mesenchymal Transition	Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis	2, 4, 14
	E2F Targets	Genes encoding cell cycle related targets of E2F transcription factors	3, 7, 13
	Estrogen Response Early	Genes defining late response to estrogen	1, 11
	G2M Checkpoint	Genes involved in the G2/M checkpoint, as in progression through the cell division cycle	3, 7
	TGF Beta Signaling	TGF-beta signaling pathway	6
	Spermatogenesis	Genes up-regulated during production of male gametes (sperm), as in spermatogenesis	7
	IL-6/JAK/STAT3 Signaling	Genes up-regulated by IL6 via STAT3, e.g., during acute phase response	12
	Interferon Gammaresponse	Genes up-regulated in response to IFNG	12

	UV Response Up	Genes up-regulated in response to ultraviolet (UV) radiation	17
KIRC	Epithelial Mesenchymal Transition	Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis	4
	UV Response DN	Genes down-regulated in response to ultraviolet (UV) radiation	4
	Oxidative Phosphorylation	Genes encoding proteins involved in oxidative phosphorylation	11
	MYC Targets V1	A subgroup of genes regulated by MYC - version 1 (v1)	11
	Adipogenesis	Genes up-regulated during adipocyte differentiation (adipogenesis)	11
HNSC	Epithelial Mesenchymal Transition	Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis	4, 5
	TGF Beta Signaling	TGF-beta signaling pathway (UV) radiation	4
	MYC Targets V1	A subgroup of genes regulated by MYC - version 1 (v1)	6
	G2M Checkpoint	Genes involved in the G2/M checkpoint, as in progression through the cell division cycle	7
	E2F Targets	Genes encoding cell cycle related targets of E2F transcription factors	7

2.5.4 Modification of individual steps in the Cancerin pipeline substantially changed the selected ceRNA interactions

In this section, we examine the technical importance of the two major steps in the Cancerin pipeline. The LASSO-based variable selection to select miRNA-mRNA interactions (Step 2) and sensitivity correlation-based filtering to select ceRNA interactions (Step 3) were two key components in Cancerin. To assess the importance of those two steps, we modified/deactivated those steps to see how it would alter the final ceRNA interaction network topology. Specifically, we kept Steps 1 and 3 in Cancerin, but in Step 2, we replaced the LASSO-based variable selection procedure by ordinary least square (OLS) multiple regression. For each RNA, its candidate miRNA regulators were selected if their coefficients from OLS were negative and p-values < 0.05. We termed this method "Cancerin (OLS regression)." We also kept Steps 1 and 2 in Cancerin, but in Step 3, we deactivated the ceRNA filtering criterion based on sensitivity correlation. We termed this method "Cancerin (sensitivity correlation filtering step deactivated)." The Cancerin pipeline with no modification is referred to as "Cancerin (original)."

To compare Cancerin to other existing methods, we replicated the method used in [Zhou et al., 2014b, Shao et al., 2015], which inferred ceRNA interactions based on negative expression correlation between miRNA and RNA targets and positive expression correlation between RNA targets. We referred to this method as

the "Correlation-based" method. The method did not consider the other types of regulators besides miRNA (i.e., TF, CNA, and DNA methylation) as potential regulators of gene expression and it also did not take into account the additive effects of multiple regulators on controlling gene expression.

Table 2.7 summarizes the number of selected ceRNA interactions obtained by applying the "Cancerin (original)," "Cancerin (OLS regression)," "Cancerin (sensitivity correlation filtering step deactivated)," and "Correlation-based method." As expected, using only expression correlation to infer ceRNA interactions resulted in many ceRNA pairs. Compared to Cancerin, the number of correlation-based ceRNA interactions was more than 6-fold higher in BRCA, 10-fold higher in KIRC, and 6-fold higher in HNSC. All ceRNA interactions found by "Cancerin (original)" were included in the "Correlation-based" method. There were also more ceRNA interactions found by "Cancerin (OLS regression)" than by "Cancerin (original)" but the increased size was in smaller compared to the "Correlation-based" method. There is a low overlap between the ceRNA interactions found in "Cancerin original" and the those from "Cancerin (OLS regression)." Specifically, with respect to interactions found in "Cancerin (original)," the percentages of common interactions that were also found in "Cancerin (OLS regression)" were 26.8% (BRCA), 40% (KIRC), and 33.2% (HNSC). Compared to "Cancerin original," deactivation of sensitivity correlation filtering step also increased the number of ceRNA interac-

tions. The fold-change increase in each cancer type was 1.7 (BRCA), 4.1 (KIRC), and 3.0 (HNSC). Overall, this comparative analysis indicated that due to several filtering layers used in "Cancerin (original)," the pipeline is more selective than other methods in selecting ceRNA interactions.

We also checked the number of PPIs and TF-gene interactions that were also inferred ceRNA interactions obtained by modifying particular steps in Cancerin or using the "Correlation-based" method. As expected, compared to ceRNA interactions obtained by "Cancerin (original)," with other methods we observed an increase of ceRNA interactions that were also PPI or TF-gene interactions. Especially the ceRNA interactions inferred by the "Correlation-based" method contained a consistently higher percentage of PPI and TF-gene interactions. These results suggest that the ceRNA interaction predictions obtained from pairwise expression correlation methods could have high false positive rate.

Table 2.7: Number of selected ceRNA interactions by applying different methods.

	BRCA	KIRC	HNSC
Cancerin (original)	4,115	4,639	2,725
Cancerin (OLS regression)	6,039	19,202	6,262
Cancerin (sensitivity correlation filtering step deactivated)	7,018	18,976	8,179
Correlation-based method	25,853	46,518	16,908

2.5.5 Inferred ceRNA interactions were able to predict gene expression change

To assess the accuracy of the inferred ceRNA interactions to predict gene expression change, we employed shRNA-mediated perturbation assays data obtained from the Library of Integrated Network-based Cellular Signature (LINCS) database [Liu et al., 2015a]. In the LINCS-L1000 shRNA-perturbation database, gene knockdown experiments using shRNAs were conducted on multiple disease cell lines, making the database a valuable resource to assess gene-gene interactions inferred from computational methods. Each experiment reported gene expression changes of 978 genes as response to the knockdown of a specific gene, which was targeted by a specific shRNA. We referred to the knocked down genes as upstream genes and to the 978 expression-profiled genes as downstream genes. Details of how we used the LINCS-L1000 dataset to evaluate the accuracy of inferred ceRNA interactions in predicting gene expression change are described in Appendix B. In brief, if an upstream ceRNA is silenced, the upstream ceRNA's miRNA regulators become more available to bind and thereby downregulate the downstream ceRNA partners. Thus, given a downstream ceRNA, its expression level should be lower in response to the silencing of upstream ceRNA partners in comparison to the silencing of other upstream genes. Ratio Fold Change (RFC) of a downstream ceRNA is defined as ratio of its expression fold change following the knockdown of its ceRNA partners to its

expression fold change following the knockdown of upstream genes that are not its ceRNA partners. A downstream ceRNA's RFC was expected to be smaller than 1. A lower value of RFC indicated better prediction of gene expression change due to ceRNA interactions.

Table 2.8: Accuracy of the ceRNA networks inferred by different methods based on the LINCS-L1000 (MCF7) dataset.

	Accuracy (96h)	Accuracy (144h)	Overall Accuracy (96h + 144h)
Cancerin (original)	71.4%	69.6%	**70.7%**
Hermes	**77.2%**	60.0%	70.2%
Cancerin (only_miRNA)	67.1%	**73.9%**	69.6%
Cancerin (OLS regression)	66.1%	58.1%	62.9%
Cancerin (sensitivity correlation filtering step deactivated)	66.3%	66.1%	66.2%
Correlation-based method	62.8%	68.2%	65.0%

Chiu et al. [Chiu et al., 2017] used LINCS shRNA-mediated perturbation assays to assess the Hermes algorithm, their genome-wide ceRNA interaction prediction tool [Sumazin et al., 2011]. We also used the same LINCS dataset (L1000-MCF7) that had been used in [Chiu et al., 2017] to validate our results and to compare the accuracy of Cancerin with that of Hermes. We defined the accuracy of a ceRNA network as the percentage of downstream ceRNAs whose RFCs were smaller than 1. As gene expression in the MCF7 dataset was measured in two different

time points (96h and 144h), our analysis was applied at each time point (Table 2.8). At 96h, out of all downstream ceRNAs (77 in Cancerin and 22 in Hermes), the number of ceRNAs whose RFC was smaller than 1 was 55 in Cancerin (accuracy 71.4%) and 17 in Hermes (accuracy 77.2%). At 144h, out of all downstream ceRNAs (46 in Cancerin and 15 in Hermes), the number of ceRNAs whose RFC was smaller than 1 was 32 in Cancerin (accuracy 69.6%) and 9 in Hermes (accuracy 60%). While overall accuracy (i.e., percentage of total downstream ceRNAs whose RFC was smaller than 1 at both time points) between Cancerin and Hermes was approximately equal (70.7% in Cancerin and 70.2% in Hermes), Cancerin showed consistent accuracy at both time points. We also computed the RFC values for the downstream ceRNAs obtained when the individual steps in Cancerin pipeline were modified and when only miRNAs were used as potential regulators in the variable selection step (i.e., Cancerin (only_miRNA)). Cancerin outperformed those methods based on the overall accuracy (see Table 2.8).

2.6 Summary

In this chapter, we introduced Cancerin, a tool to infer genome-wide cancer-associated ceRNA interaction networks and applied it in three types of cancer. Unlike existing ceRNA inference tools that considered miRNAs as the only type of gene regulator, Cancerin considered other types of gene regulators besides miR-

NAs, namely TFs, DM, and CNA. In addition, using sensitivity correlation metric proposed in [Paci et al., 2014], the method directly modeled the ceRNA hypothesis, which posited that the expression profiles of two ceRNAs should be positively correlated, and that correlation was conditioned on the expression of their shared miRNA regulators.

The inferred ceRNA networks in all the three cancer types were scale-free networks as the ceRNAs' degree distribution followed power-law with high fitness. There were very few overlapping interactions between the inferred ceRNA interactions and the PPIs or TF-gene interactions.

Only a subset of input DE RNAs were selected as ceRNAs in the final ceRNA networks. In all three cancer types, the ceRNAs were significantly enriched with cancer-related genes whereas DE RNAs that were not in the ceRNA networks did not have a significant enrichment. To explore the biological importance of our inferred ceRNA networks, we clustered ceRNA networks into modules and performed functional enrichment on each module. Various cancer hallmark terms, biological processes, and pathways were enriched in the ceRNA modules across all the three cancer types. In addition, some ceRNA modules were associated with multiple cancer hallmark terms, making the ceRNAs in such module valuable biomarkers to be further investigated. In brief, the results shows that Cancerin found cancer-associated ceRNAs and the ceRNA modules were involved in cancer-related biologi-

cal processes. Thus, Cancerin can be used to explore the functional roles of ceRNAs in cancer.

To examine the prognostic capability of the inferred ceRNA networks, we performed univariate Cox proportional hazard models for each ceRNA and non-ceRNA. In all three cancer types, compared to non-ceRNAs, ceRNAs exhibited higher association with cancer outcome. We also observed that KIRC ceRNAs had low hazard ratios indicating that they might act as tumor-suppressors. Since the ceRNAs found by Cancerin are prognostic of cancer outcome, they can be valuable targets for cancer therapy.

We also examined the functional importance of the miRNAs that mediated ceRNA interactions. The miRNAs that mediated the highest number of ceRNA interactions (i.e., let-7a-5p, miR-106b-5p, and miR-9-5p) are well-known in the cancer literature [Barh et al., 2010, Ivanovska et al., 2008, Coolen et al., 2013, Barbano et al., 2017]; however, their prevalent roles in mediating ceRNA interactions could suggest a novel role in cancer pathogenesis.

Validation of computationally predicted ceRNA interactions is challenging due to the low number of experimentally-validated ceRNA interactions. To address this challenge, we used the LINCS-MCF7 dataset [Liu et al., 2015a] to check if the knockdowns of ceRNAs would cause downregulation of their predicted ceRNA partners. We also compared Cancerin's accuracy with that of Hermes [Sumazin et al.,

2011], a ceRNA inference tool based on mutual information criteria. Based on the prediction of gene expression change using the inferred ceRNA interactions, Cancerin achieved approximately equal accuracy as Hermes; however, the accuracy values from Cancerin at different experimental time points were more consistent.

In summary, Cancerin is a computational method that integrates genomic, transcriptomic, and epigenetic regulatory factors to infer genome-wide ceRNA interactions in cancer. Analysis of the inferred ceRNA networks constructed by Cancerin would provide novel insights on the biological functions of this novel layer of gene regulation, especially on how it contributes to cancer pathogenesis.

CHAPTER 3

CANMOD: A COMPUTATIONAL METHOD TO IDENTIFY FUNCTIONAL MIRNA-TRANSCRIPTION FACTOR-TARGET MODULES

(This chapter is adapted from a manuscript that we are preparing to submit to conference IEEE International Conference on Bioinformatics and Biomedicine 2018. We have full permission to reuse the manuscript's contents in this chapter.)

3.1 Abstract

Transcription factors (TFs) and microRNAs (miRNAs) are two important classes of gene regulators that govern many critical biological processes. Dysregulation of TF-gene and miRNA-gene interactions can lead to the development multiple diseases including cancer. Many studies aimed to identify interactions between target genes and their regulators in both normal and disease settings. However, few studies attempted to elucidate the collaborative relationship between TFs and miRNAs in regulating genes involved in cancer-associated biological processes. Identification of the coregulatory functions of those regulators in cancer would provide a better understanding of gene regulation at different layers and may also suggest better approaches for targeted therapy. This study proposes a computational pipeline called CanMod to identify cancer-associated gene regulatory modules. CanMod was designed so that it could infer gene regulatory modules that meet three criteria. First, within a module, target genes should involve in similar

biological processes; thus, the modules are distinguishable based on their biological functions. Second, the expression of target genes in a module should be collectively dependent on the expression of their regulators. Third, a regulator and a target should be allowed to be included in multiple modules to reflect the diverse biological roles that the genes and the regulators may be responsible for. Unlike some existing methods, our proposed pipeline also considered copy number alteration (CNA) and DNA methylation (DM) data while inferring regulator-target gene interactions with higher accuracy.

We applied CanMod on the breast cancer dataset (BRCA) from The Cancer Genome Atlas (TCGA). We found that modules found by CanMod were associated with distinguishable biological functions and the expression of target genes in the modules were significantly correlated. In addition, many hub regulators in CanMod were known cancer genes, and CanMod was able to find experimentally validated regulator-target interactions.

3.2 Motivation and Related Work

While both miRNAs and TFs are important gene regulators, there is still much unknown about their collaborative relationship in gene regulation. Moreover, it is unclear what biological functions and processes are coregulated by miRNAs and TFs, especially in complex diseases such as cancer. This chapter aims to iden-

tify cancer-associated gene regulatory modules, which consist of miRNAs, TFs, and their coregulated target genes involved in similar biological processes. Identification of such modules can generate important understandings of gene regulation activities involving miRNAs and TFs and how they are implicated in cancer.

The chapter is inspired by two streams of studies that are closely relevant to the research goal. The first stream involves integrating miRNA-target and TF-target interactions into an unified gene regulatory network and analyzing the network to explore the coregulatory relationship between miRNAs and TFs. The second stream involves identification of miRNA-target modules, in which multiple miRNAs collectively regulate the expression of coregulated genes.

In the first stream of studies, the predicted interactions between TF-target, miRNA-target, and TF-miRNA were retrieved from different public databases and merged to construct the miRNA-TF-target (MTT) networks [Shalgi et al., 2007, Delfino and Rodriguez-Zas, 2013]. Using gene expression, only certain edges in the MTT networks were kept to model the expression dependency of the targets on their regulators [Qin et al., 2014]. The miRNAs and TFs that regulate many similar targets were analyzed to specify their coregulatory relationship as well as their biological and clinical significance. Several hub regulators were found to be able to classify different breast cancer subtypes [Qin et al., 2014] and had high prognostic power to predict ovarian cancer recurrence [Delfino and Rodriguez-Zas, 2013].

Several studies applied Bayesian network modeling approaches to learn the topology of MTT networks [Zacher et al., 2012, Roqueiro et al., 2012], which could be used to identify the coregulatory relationship between TFs and miRNAs based on their common targets. While the above studies are helpful in exploring potential relationships between TFs and miRNAs, the inferred MTT networks are often very complex; thus it is challenging to interpret the biological significance of the relationships between those coregulators.

The second stream of studies focused on identifying functional miRNA-target gene modules (MTMs), which were defined as groups of miRNAs and their coregulated target genes holding important biological functions [Liu et al., 2009, Tran et al., 2008]. Some studies considered an MTM as a maximal biclique, which is a network composed of two sets of nodes (i.e., miRNAs and their targets) and each node of one set is connected to the all nodes in the other set and form an all-to-all connection pattern [Peng et al., 2009, Uhlmann et al., 2012]. After constructing miRNA regulatory networks by leveraging putative miRNA-target interactions and their expression profiles, different maximal biclique algorithms were applied to identify MTMs [Peng et al., 2009, Uhlmann et al., 2012]. However, requiring an MTM to be a maximal-biclique led to the unnecessary splitting of a reasonable MTM into separate modules, thus creating many modules that contained a single miRNA. Several studies relaxed the all-to-all connection requirement to be most-

to-most and considered such MTMs as quasi-bicliques [Mukhopadhyay and Maulik, 2014]. The algorithms to find such MTMs often required users to predetermine the number of MTMs, which was unknown in most cases. In addition, most of the existing computational methods to infer MTMs only allow a regulator and a target gene to belong to a single gene module, which did not reflect the diverse biological roles that the regulators and the target genes may hold. Moreover, it is challenging to extend the existing methods to incorporate TFs as an additional type of regulators.

This study proposes a computational pipeline called CanMod, which aims to identify cancer-associated gene regulatory modules. A module consists of a group of regulators (i.e., miRNAs and TFs) that coregulate a set of target genes with similar biological functions in cancer. CanMod is different from previous computational methods to infer gene regulatory modules in several aspects. First, CanMod requires the target genes in a module to participate in similar biological processes so that the modules are biologically significant and interpretable. Second, CanMod does not require the number of modules to be predetermined. Third, to reflect the diverse biological roles that a regulator or a target gene may have, CanMod allows a regulator or a target gene to appear in different modules. Besides miRNAs and TFs, there are other regulatory factors controlling gene expression such as CNA and DM [Jones, 2015]. Thus, besides using gene expression, CanMod also employs

CNA and DM data to infer regulator-gene interactions. R software for CanMod is freely available (MIT license) at https://github.com/bozdaglab/CanMod.

We applied CanMod to the breast cancer dataset (BRCA) from The Cancer Genome Atlas (TCGA) [Network et al., 2012]. We found that the regulators that were included in many modules were previously known to be associated with cancer. In addition, the modules obtained by CanMod contained a significant number of experimentally validated regulator-target interactions. Functional enrichment analysis of the modules revealed that target genes in the modules were strongly enriched with cancer-associated biological processes, pathways, and cancer hallmark terms.

The rest of the chapter is organized as follows. Section 3.3 recaps the input data and the data processing procedure. Section 3.4 provides a detailed description of each computational step in CanMod. Section 3.5 discusses the results obtained from applying CanMod to the breast cancer dataset from TCGA. Section 3.6 summarizes the main components of Cancerin and the key biological findings discussed in the chapter.

3.3 Input Data

As described in Section 1.6, the input data for CanMod are genome-wide expression of miRNAs and mRNAs. Each mRNA is also associated with a CNA

value and a DM value. In addition, putative miRNA-target gene and TF-target gene interactions are employed to select candidate regulators for each target target gene. CanMod was applied to the breast cancer dataset from TCGA. Section 1.6 provides the details for how all of the data were preprocessed to be used as input for CanMod.

3.4 CanMod pipeline

CanMod is a computational pipeline to identify cancer-associated gene regulatory modules. A module comprises two groups, a group of genes that have similar biological functions and a group of regulators that coregulate the target genes. CanMod allows a target gene or a regulator (i.e., TF or miRNA) to belong to multiple modules. CanMod consists of six main steps, which are illustrated in Fig. 3.1. The details of each step are described in the following sections.

Step 1: Select regulators associated with expression change of each target gene

One important criterion for gene regulatory modules obtained by CanMod is that the expression of target genes in a module is dependent on the expression of the regulators in the same module. To infer group-wise expression dependence between regulators and targets, CanMod first identifies the pair-wise relation between

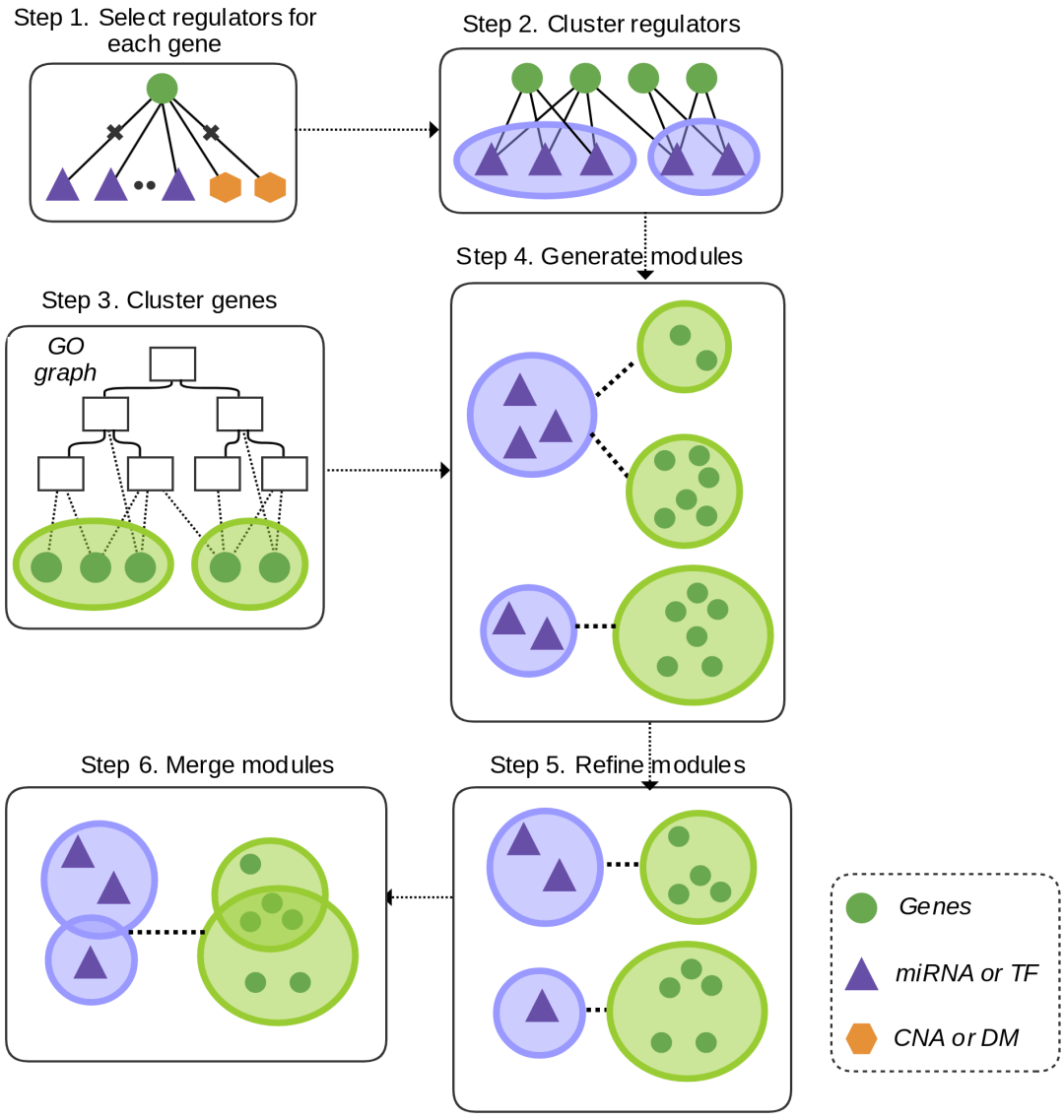

Figure 3.1: **CanMod pipeline.** Step 1: Select regulators associated with the expression change of each target gene. Step 2: Identify regulator clusters (RCs) such that regulators within an RC regulate similar target genes. Step 3: Identify target gene clusters (GCs) so that genes within a GC share similar biological functions. Step 4: Obtain candidate modules for each RC by recruiting co-expressed target genes from the same GCs. Step 5: Select correlated regulators and targets in each candidate module. Step 6: Merge candidate modules whose regulators have similar target genes. (CNA: Copy Number Alteration, DM: DNA Methylation)

a target gene and its candidate regulators. In other words, for each gene, CanMod aims to select which regulators contribute to the expression change of the gene.

First, to reduce the search space of regulator-gene interactions, CanMod employs putative regulator-gene interactions. As described in Section 1.6 (Datasets and Data Preprocessing), the putative miRNA-gene interactions were retrieved from StarBase v2.0 [Li et al., 2013] and TargetScan 7.1 [Agarwal et al., 2015]. Putative TF-gene interactions were retrieved from TRED [Zhao et al., 2005] and TTRUST (version 2) [Han et al., 2015].

CanMod also employs a similar LASSO-based variable selection procedure used in the computational method Cancerin (Chapter 2) to identify the miRNA and TF regulators for each gene.

Step 2: Cluster regulators with similar targets into regulator clusters (RCs)

A regulator cluster (RC) is defined as a group of regulators that regulate a high number of similar targets. In this step, CanMod identifies such RCs based on the regulator-target pairs obtained in the previous step. First, CanMod constructs a 2D matrix $Z^{M \times N}$, where M is the number of possible regulators (i.e., TFs and miRNAs), and N is the number of all the possible target genes. The value of each cell (Z_{ij}) equals 1 if $regulator_i$ (i^{th} row) is the regulator of $gene_j$ (j^{th} column) and 0 otherwise.

Using $\mathrm{Z}^{M \times N}$ as an input and considering target genes as attributes, CanMod computes the square distance matrix of regulators, called $\mathrm{RD}^{M \times N}$, which quantifies the dissimilarity between any two regulators based on what target genes they regulate. Given two regulators $R_{\mathrm{i}1}$ and $R_{\mathrm{i}2}$, let O be the number of common targets of $R_{\mathrm{i}1}$ and $R_{\mathrm{i}2}$, P be the number of targets of $R_{\mathrm{i}1}$ but not of $R_{\mathrm{i}2}$, and Q be the number of targets of $R_{\mathrm{i}2}$ but not of $R_{\mathrm{i}1}$. CanMod applies the Jaccard distance (Eq. 3.1) to compute the dissimilarity between regulators $R_{\mathrm{i}1}$ and $R_{\mathrm{i}2}$.

$$\mathrm{Jaccard\_dist}(R_{\mathrm{i}1}, R_{\mathrm{i}2}) = 1 - \frac{O}{O + P + Q} \tag{3.1}$$

Given the Jaccard distance matrix of the regulators, CanMod applies agglomerative hierarchical clustering to identify the RCs. In agglomerative hierarchical clustering, each regulator is considered initially as a single-element cluster. At each iteration, two clusters with the lowest dissimilarity are merged, and the algorithm stops when all clusters are merged into a single big cluster. The distance between any two clusters is their average linkage value, which is the average of dissimilarity values of all pairwise elements between the two clusters. Hierarchical clustering produces a dendrogram. CanMod cuts the dendrogram at the top to obtain RCs, which represent groups of regulators that regulate many similar target genes.

Step 3: Cluster target genes with similar Gene Ontology - Biological Process (GO-BP) terms into gene clusters (GCs)

One distinguishable characteristic of CanMod compared to existing gene module inference methods is that we require modules found by CanMod to be biologically interpretable and functionally important. Specifically, genes in the same module should have similar biological functions. To meet that criterion, CanMod employed Gene Ontology - Biological Process (GO-BP) terms to cluster target genes. For the rest of the chapter, GO-BP terms are referred to briefly as GO terms.

The ontology of a gene describes of its general biological roles such as what molecular events the gene participate in. A gene can have diverse roles, and those roles can be intricately related to one another, making annotating gene ontology a challenging task. The Gene Ontology project [Day-Richter et al., 2007] aims to address the challenge and provides a comprehensive and consistent vocabulary to describe gene ontology. The outcome of the Gene Ontology project is a directed acyclic graph (GO DAG) whose nodes are GO terms and whose edges represent relationships between GO terms. A GO-BP term signifies a biological objective, which could be broad such as cell proliferation (GO:0008283) or specific such as regulation of neuroblast proliferation (GO:1902692). Specific biological objectives (i.e., low-level GO terms in the GO DAG) are inherently parts of broad biological

objectives (i.e., high-level GO terms in the GO DAG), which explains the hierarchical structure of the GO DAG.

Since a single gene can participate in multiple biological processes, it is associated with multiple GO terms. In CanMod, genes with highly similar biological functions are grouped into a gene cluster (GC). Using GO DAG, CanMod computes the biological similarity between two genes based on the similarity of the GO terms associated with those genes. CanMod employs a graph-based approach to compute semantic similarity between any two GO terms [Wang et al., 2007]. Loosely speaking, two GO terms are considered to be similar if they are close to each other in the GO DAG, and two GO terms that are close to each other at a lower level are considered more similar than those at a higher level in the GO DAG.

Semantic similarity between two GO terms is computed using the semantic value (S-value) of each GO term [Wang et al., 2007]. Computing the S-value of a GO term A (i.e., $SV(A)$) is based on traversing a subgraph of the GO DAG $DAG_{\mathrm{A}} = (A, T_{\mathrm{A}}, E_{\mathrm{A}})$. T_{A} is a set of GO terms that includes term A and its ancestor terms, and E_{A} is a set of edges connecting the terms in DAG_{A}. $SV(A)$ is de-

fined as the sum of all terms in DAG_{A}, and the terms closer to term A have more weights in $SV(A)$. For each term t in DAG_{A}, its S-value to term A (i.e., $S_{\mathrm{A}}(t)$) is

$$S_{\mathrm{A}}(t) = \begin{cases} 1 \text{ if } t = A \\ \max\{w_{\mathrm{e}} \times S_{\mathrm{A}}(t') | t' \in \text{children}(t)\} \text{ if } t \neq A. \end{cases} \tag{3.2}$$

Thus,

$$SV(A) = \sum_{t \in T_{\mathrm{A}}} S_{\mathrm{A}}(t). \tag{3.3}$$

The semantic similarity between term A and term B is

$$\mathrm{sim}(A, B) = \frac{\sum_{t \in T_{\mathrm{A}} \cap T_{\mathrm{B}}} S_{\mathrm{A}}(t) + S_{\mathrm{B}}(t)}{SV(A) + SV(B)}. \tag{3.4}$$

Suppose two genes g_1 and g_2 are associated with the set $\{term_{11}, term_{12}, ..., term_{1\mathrm{m}}\}$ and the set $\{term_{21}, term_{22}, ..., term_{2\mathrm{n}}\}$, respectively. Semantic similarity between g_1 and g_2 is computed as the average similarity of all pairs of GO terms between these two sets (Eq. 3.5):

$$\mathrm{sim}(g_1, g_2) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{sim}(term_{1\mathrm{i}}, term_{2\mathrm{j}})}{m \times n} \tag{3.5}$$

CanMod employs the R package GoSemSim [Yu et al., 2010] to compute a similarity matrix between all pairs of genes. Then, to obtain gene clusters based on their biological similarity, the affinity propagation (AP) clustering algorithm is applied [Frey and Dueck, 2007] to the similarity matrix. AP clustering is based on the idea of "message passing" between data points to iteratively find "exemplars," which are some specific data points that are representative of clusters. In each iteration, the algorithm applies a heuristic approach to update "exemplars" to maximize the distances among "exemplars," and to minimize the distances between "exemplars" and their corresponding clusters' members. Unlike clustering algorithms such as k-means or k-medoids, AP clustering does not require the number of clusters to be determined before applying the algorithms. Thus, AP clustering is applicable to our case because it is not possible to determine in advance how many clusters of genes have similar biological functions. In brief, after applying AP clustering on the GO-based gene similarity matrix, CanMod obtains GCs such that genes in the same GC are involved in similar biological processes.

Step 4: Obtain candidate modules for each regulator cluster (RC) by recruiting co-expressed targets in the same gene clusters (GCs)

Generation of candidate modules starts with the RCs that are obtained in Step 2. Each module is comprised of a group of regulators (i.e., RC) and a group of target genes (i.e., GC). Suppose RC_{i} consists of H regulators $\{R_1, ..., R_{\mathrm{H}}\}$. For

RC_{i}, its candidate target genes are the union of all target genes of the regulators in RC_{i} (results obtained in Step 1). Let us refer to this union set of target genes of RC_{i} as $\{G\}_{\mathrm{i}}$ and call the genes in $\{G\}_{\mathrm{i}}$ as "seed" genes.

Next, using the GC assignment obtained in Step 3, CanMod splits $\{G\}_{\mathrm{i}}$ into different clusters such that each cluster includes seed genes belonging to the same GC. Suppose the seed genes $\{G\}_{\mathrm{i}}$ belong to k GCs. Thus, after the splitting procedure, the mapping $(RC_{\mathrm{i}} \rightarrow \{G\}_{\mathrm{i}})$ becomes $\{(RC_{\mathrm{i}} \rightarrow GC_{\mathrm{i1}}), ..., (RC_{\mathrm{i}} \rightarrow GC_{\mathrm{ik}})\}$. A mapping $(RC_{\mathrm{i}} \rightarrow GC_{\mathrm{ij}})$ is considered as a candidate module. Next, for each candidate module (i.e., $(RC_{\mathrm{i}} \rightarrow GC_{\mathrm{ij}})$), additional candidate target genes are added to GC_{ij}. A target gene is added if it meets two conditions. First, it belongs to the same GC (Step 3) as the seed genes in GC_{ij}. Second, its expression correlation to at least one of the seed genes is in the top 90% correlation in the correlation distribution of all possible gene pairs. We require these two conditions to ensure that the target genes in the candidate modules are co-expressed and have similar biological functions.

Step 5: Select regulators and targets that are co-expressed in each candidate module

One important criterion for modules obtained by CanMod is that regulators within each module should regulate the expression of their target genes collectively. To model the group-based expression dependency between the regulators

and the targets genes, for each module, CanMod applies Sparse Canonical Correlation (SCCA) to select a subset of regulators and target genes such that expression of the selected target genes exhibit (canonical) correlation with the selected regulators. Let $R = [R_1, R_2, ..., R_{\mathrm{P}}]$ and $G = [G_1, G_2, ..., G_{\mathrm{Q}}]$ be expression of P regulators and Q target genes in a module M_{i}. SCCA aims to maximize the canonical correlation ρ between canonical variates Ru and Gv, where u and v are weight vectors $u = (u_1, , u_{\mathrm{m}})$, $v = (v_1, , v_{\mathrm{m}})$. Thus

$$\rho = \frac{v'R'G'u}{\sqrt{v'R'Rv}\sqrt{u'G'Gu}}. \tag{3.6}$$

While maximizing ρ, SCCA also applies regularization terms to u and v, which shrinks some weights to zero and yields to p regulators ($p < P$) and q targets ($q < Q$) to be selected. In brief, SCCA is applied on each candidate module to select a subset of regulators and targets exhibiting high expression dependency.

Step 6: Merge modules whose regulators have similar target genes

The procedure used in Step 4 to generate candidate modules creates a likely scenario that many modules may share a high number of similar target genes. Even when several target genes might be removed from each module in Step 5, many modules still may have many similar targets. Thus, to ensure the specificity of different modules while still allowing a target gene to be able to assigned in multiple

modules, CanMod applies hierarchical clustering to merge modules that share many similar target genes. The method details are similar to those in Step 4.

First, using results obtained in Step 5, CanMod computes a Jaccard distance matrix of modules based on their shared targets. Then CanMod applies agglomerative hierarchical clustering and uses average linkage to construct a dendrogram. CanMod cuts the dendrogram at the top to obtain the final merged modules. In brief, the final modules come from merging constituent modules obtained in Step 5.

3.5 Results

CanMod uses various data types, namely putative interactions between target genes and their candidate regulators (miRNAs or TFs), the expression of the regulators and the target genes, and the CNA and DM of the target genes. CanMod was applied to the breast cancer dataset from TCGA. CanMod considers only the regulators and target genes that were DE between normal and tumor breast samples. Between 47 normal and 193 tumor samples, we found 215 DE miRNAs, 1,185 DE TFs, and 7,502 DE genes. Among those, there were 158,819 putative miRNA-gene interactions and 33,638 TF-gene interactions. The following sections will discuss and analyze the results.

3.5.1 Results from each step in CanMod

The CanMod pipeline consists of six steps. In Step 1, CanMod computes regulator-target interactions. For each DE target gene, out of all of its candidate DE regulators, CanMod applies a LASSO-based variable selection procedure to select a subset of regulators that are significantly associated with the expression change of the target gene. After Step 1, CanMod obtained 6,616 miRNA-gene interactions between 196 miRNAs and 2,814 target genes, and 11,017 TF-gene interactions between 944 TFs and 3,208 target genes. On average, a target gene was regulated by two miRNAs and three TFs.

Using the results of Step 1, in Step 2, CanMod clusters regulators based on their target similarity. Regulator clustering resulted in 343 regulator clusters (RCs) and each RC had three regulators on average.

In Step 3, CanMod clusters target genes into gene clusters (GCs) based on their GO term similarity. Gene clustering resulted in 251 GCs, and each GC contained 17 genes on average.

Using all the results obtained in the previous steps, in Step 4, CanMod computes candidate modules. Each candidate module consists of regulators in an RC, their target genes from the same GC and other genes that are co-expressed with target genes and are in the same GC with target genes. After Step 4, CanMod ob-

tained 5,880 candidate modules and on average, each module had three regulators and nine target genes.

In Step 5, CanMod filters regulators and target genes in candidate modules based on their expression correlation. Specifically, for each candidate module, CanMod applies SCCA to select a subset of regulators and target genes that exhibited linear expression correlation. After regulator/target filtering, candidate modules that have no regulators or target genes are eliminated. After this step, there remained 4708 candidate modules, which had two regulators and five target genes on average.

To ensure that modules are distinctive in terms of their target genes' biological functions, in Step 6, CanMod merges modules that share a high number of target genes. After module merging, there were 912 final modules. On average, each module consisted of eight regulators and seven target genes (Fig. 3.2A). While a module often contained more TFs than miRNAs, on average a miRNA appeared in more modules than a TF did (Fig. 3.2B).

3.5.2 Hub regulators were associated with cancer

A hub regulator was defined as a regulator included in many modules, thus having a high module degree value. Hub regulators may hold important biological roles. We required a hub TF and a hub miRNA to have module degree in the

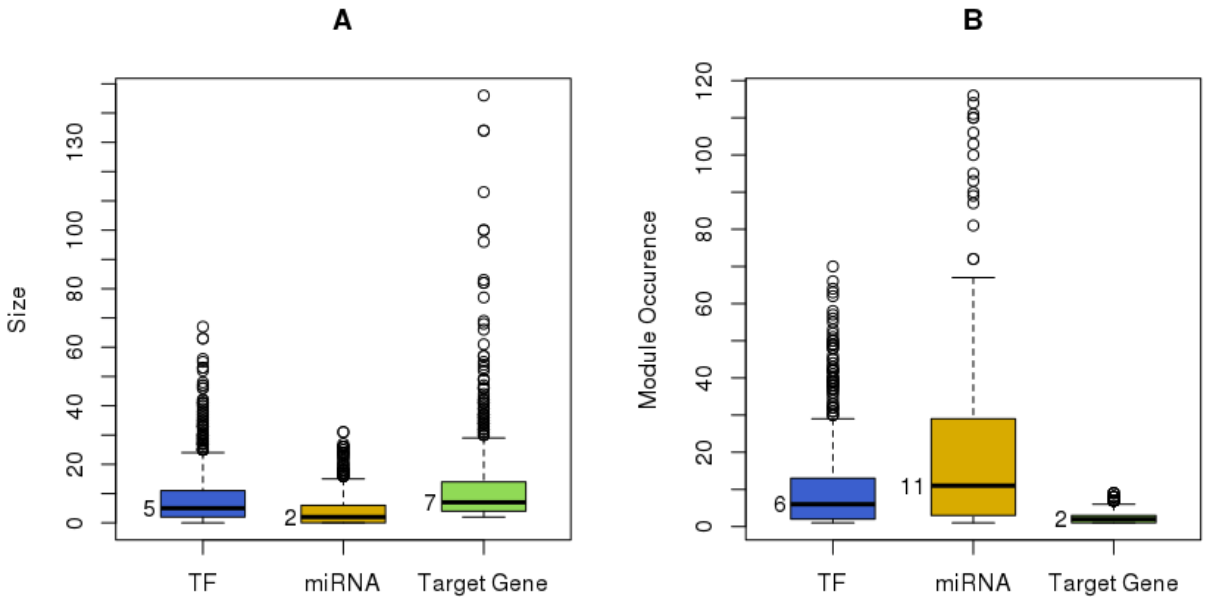

Figure 3.2: **Size and module degree of TFs, miRNAs, their target genes included in the inferred modules**. (A) Number of TFs, miRNAs, and target genes in 912 modules obtained by CanMod. (B) Number of modules (module degree) with which a TF, miRNA, or target gene was associated.

top 10 percentile of module degree across all the TFs and miRNAs, yielding 61 hub TFs and 13 hub miRNAs. On average, a hub TF and hub miRNAs had module degree 41 and 95, respectively. To evaluate the functional relevance of the hub regulators, we performed a hypergeometric test between the hub miRNAs/TFs found by CanMod and a list of cancer-related genes and miRNAs.

There were 2,944 cancer-related genes retrieved from the Cancer Gene Census in COSMIC v83 [Forbes et al., 2016], the Bushman lab's Cancer Gene List v3 [Bushman], and the Network of Cancer Genes 5.0 [An et al., 2015]. From the oncomiRDB [Wang et al., 2014] database we retrieved 314 cancer-related miRNAs.

The hypergeometric results between the hub TFs versus the cancer-related

genes and hub miRNAs versus the cancer-related miRNAs indicated that both the hub TFs and the hub miRNAs were significantly associated with cancer (p-value = 2.5e-07 for hub TFs and p-value = 0.048 for hub miRNAs). In contrast, the TFs and miRNAs that were included in only one module (i.e., module degree = 1) did not show significant association with cancer (p-value = 0.073 for such TFs and p-value = 0.91 for such miRNAs).

3.5.3 Experimentally validated regulator-target gene interactions were found in the modules

To check the ability of CanMod to discover experimentally validated regulator-target interactions, we collected validated miRNA-gene interactions from miRTarBase [Chou et al., 2017] and validated TF-gene interactions based on ChIP-seq data of breast cancer cell line (MCF7) from ENCODE project [ENCODE Project Consortium, 2004]. MiRTarBase is a database that curates experimentally validated miRNA-target interactions found by different methods such as reporter assay, western blot, microarray, and next-generation sequencing experiments. The ChIP-seq data (MCF7) from ENCODE provides direct binding validation between TFs and their target genes. We only kept the interactions between DE regulators and DE targets, which resulted in 29,693 miRNA-target interactions between 215 miRNAs and 6,123 targets in miRTarBase, and 47,506 TF-gene interactions between 41 TFs and 6,642 targets in ENCODE.

For each module in CanMod, we checked whether it included at least one validated interaction. We applied a sampling procedure to evaluate the significance of the number of validated interactions found in each module. Suppose module M_{i} contained S validated interactions between H regulators $\{R_{\mathrm{i1}}, .., R_{\mathrm{iH}}\}$ and K targets $\{T_{\mathrm{i1}}, .., T_{\mathrm{iK}}\}$. From all DE targets, we generated a set of K randomly selected targets $\{T'_{\mathrm{i1}}, .., T'_{\mathrm{iK}}\}$. Then we counted the number of validated interactions S' between $\{R_{\mathrm{i1}}, .., R_{\mathrm{iH}}\}$ and $\{T'_{\mathrm{i1}}, .., T'_{\mathrm{iK}}\}$, and checked if $S \leq S'$. The procedure was repeated 1,000 times. The empirical p-value of S was the number of times $S \leq S'$ over 1,000.

Significance of regulator-target gene interactions can be assessed in modules that have at least one regulator and one target that also appeared in the miRTarBase or ENCODE datasets. Among 912 final modules, 667 and 257 modules had regulators and targets in miRTarBase and ENCODE datasets, respectively. Out of the 667 modules, 373 modules (56%) contained validated miRNA-target interactions. Out of the 257 modules, 180 modules (70%) contained validated TF-target interactions. Several examples of validated interactions between cancer-related miRNAs and TFs with their cancer-related target genes are shown in Table 3.1. Notably, in all modules that had validated interaction(s), the number of their validated interactions were significant (p-value < 0.005).

Table 3.1: Examples of validated interactions between cancer-related regulators and targets found in CanMod modules.

Cancer-related regulators	Cancer-related targets	Regulator description
hsa-miR-106a-5p	CAV1, HBP1, TP53INP1, FBXO31	Promote cell migration and invasion [Wang et al., 2014]
hsa-miR-130b-3p	EGR2, SIX4, HBP1, DLC1	Promote tumor aggression and reduce multidrug resistance [Wang et al., 2014]
hsa-miR-16-5p	TM4SF1, UBR3, USP7, WDR75, GOLGA5	Suppress cell self-renewal and cell growth [Patki et al., 2013]
MYC	CCND1, GOLGA5, FBXO31, CUX1	Activate angiogenesis and suppress of the host immune response [Safran et al., 2010]
RAD21	CAV1, EGR2, MEF2D, RPN1	Repair DNA double-strand break [Safran et al., 2010]
ELK1	CDK4, CORO1C, E2F3, UBE2C	Regulate many genes responsible for cell growth functions [Safran et al., 2010]

3.5.4 Expression of target genes in large modules were significantly correlated

We expected that expression of the target genes within each modules should be correlated because in Step 4, to generate candidate modules, only genes that were highly correlated with the "seed" genes were incorporated into the candidate modules. However, because of the elimination of some genes in Step 5 and the module merging procedure in Step 6, expression of the target genes in the final 912

modules were not guaranteed to be correlated. To measure the expression correlation of target genes within each module and assess its statistical significance, we employed the procedure used in [Jin and Lee, 2015] to assign an empirical p-value for the mean correlation of expression of the target genes within each module.

Briefly, for a module M_{i}, we computed its mean expression correlation $\mathrm{cor}(M_{\mathrm{i}})$, i.e., the average of absolute pairwise expression correlation among the genes in M_{i}. If M_{i} had A target genes, we generated a random target gene set of size A and computed the sampled mean correlation of the generated target set. The procedure was repeated 1,000 times, which produced a sampling distribution of the mean correlation associated with $\mathrm{cor}(M_{\mathrm{i}})$. P-value of $\mathrm{cor}(M_{\mathrm{i}})$ was computed using Eq. 3.7:
(3.7)

$$\text{p-value}(\mathrm{cor}(M_{\mathrm{i}})) = \frac{\sum_{j=1}^{1000} F(\mathrm{cor}(M_{\mathrm{i}}) < \mathrm{cor}(random\_M_{\mathrm{ij}}))}{1,000}$$

$F = 1$ if $\mathrm{cor}(M_{\mathrm{i}}) < \mathrm{cor}(random\_M_{\mathrm{ij}})$ and $F = 0$ otherwise. $\mathrm{Cor}(M_{\mathrm{i}})$ is considered to be significant if the $\text{p-value}(\mathrm{cor}(M_{\mathrm{i}})) < 0.05$. Out of 912 modules, 693 of them (76%) had significant mean correlation. The mean correlation distribution of the 912 modules is presented in Fig. 3.3. As shown in Fig. 3.3, the means of ex-

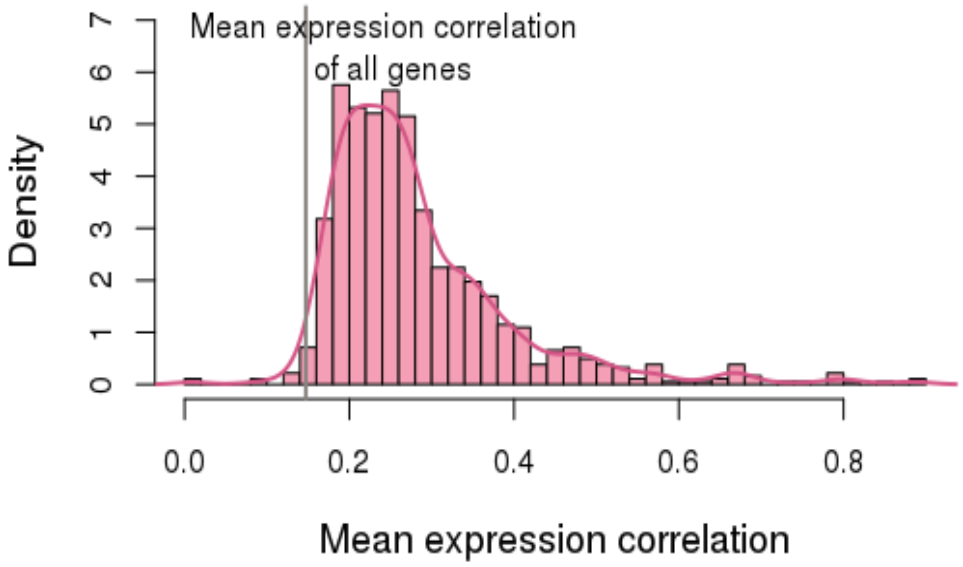

Figure 3.3: **Distribution of the average absolute correlations among target genes across the inferred modules.**

pression correlations of the target genes included in the modules (pink bars) are higher than the mean of expression correlations of all genes (vertical line).

We hypothesized that the number of target genes included in each module was correlated with its mean expression correlation significance. To check that, we employed the Wilcoxon rank sum test to compare the distribution of the number of target genes in the 693 modules that had significant mean correlations to those of the remaining 219 modules that did not have significant mean correlations. P-value from the Wilcoxon rank sum test was smaller than 2.2e-16. The median number of target genes in the 693 modules and the 219 modules was nine and four, respec-

tively. The result indicated that expression of target genes in the larger size modules were more likely to be significantly correlated.

3.5.5 Functional enrichment of the modules

The gene regulatory modules inferred by CanMod were based on the premise that the regulated genes in each module participated in similar biological processes. To assess the functional importance of the modules, we performed enrichment analysis between the target genes in each module and the GO terms, Cancer Hallmark (CH) terms, and KEGG pathway terms. To make the enrichment test statistically feasible, only the modules having at least five target genes were used as input for this enrichment analysis. A term or pathway was considered to be enriched in a module if its adjusted p-value from the enrichment test was smaller than 0.01. The GO, CH, and KEGG pathway terms were retrieved from [Liberzon et al., 2011]. We employed the R package clusterProfiler [Yu et al.] to perform the enrichment analysis.

There were 2,098 significantly enriched GO terms across all the CanMod modules. The top three most commonly enriched GO terms were GO:0009057 macromolecule catabolic process (44 modules), GO:0006396 RNA processing (40 modules), and GO:0016070 RNA metabolic process (39 modules). This enrichment result is not surprising, as these terms are broad biological processes and are associated with many genes. In contrast, the GO terms that were enriched in only

one module such as GO:0021517 ventral spinal cord development and GO:0007035 vacuolar acidification refer to specific biological processes and are associated with a small number of genes (22 genes in human). Among all the enriched GO terms, 31% of those (664 terms) were enriched in only one module. The median number of modules that a GO term was enriched in was only two. This result indicated that modules obtained by CanMod were distinctive in term of their biological functions.

We also observed a number of cancer-related KEGG pathway terms in the CanMod modules. For instance, KEGG:04110 Cell cycle, KEGG:05200 Pathways in cancer, KEGG:03040 P53 signaling pathway, and KEGG:04010 MAPK signaling pathway were enriched in 13, 8, 7, and 6 modules, respectively. Those pathways are known to be activated in breast cancer [Gasco et al., 2002, Santen et al., 2002]. KEGG:04710 circadian rhythm mammal term was enriched in only one module. Deregulation of circadian rhythm genes are shown in breast cancer [Chen et al., 2005].

We observed that certain CH terms were enriched in multiple modules obtained by CanMod. The most commonly enriched CH terms included CH:5926 MYC Targets V1 (17 modules), CH:5898 DNA Repair (15 modules), and CH:5901 G2M checkpoint (15 modules). On the other hand, CH terms such as CH:5903 Notch signaling, CH:9539 P53 pathway, CH5942: UV Response DN, CH:5895 WNT

beta catenin signaling, and CH:5934 xenobiotic metabolism occurred in only one module.

In overall, the modules obtained by CanMod were associated with important biological processes and pathways associated with cancers. Thus, CanMod is a useful tool to explore the functional significance of gene modules coregulated by TFs and miRNAs. We also inferred gene modules by applying three other methods to compare their functional enrichment results with those obtained by CanMod. Unlike in CanMod, the three methods did not employ regulator-gene interaction data and only used gene expression to infer gene modules. In the first method, we applied the K-means algorithm to cluster genes, and each cluster was considered as gene module. We set K to 1000 to make sure the number of modules found by K-means was comparable with the number of modules obtained by CanMod. In the second method, we applied the agglomerative HC algorithm to construct the dendrogram of gene clusters, and then applied dynamic cut tree algorithm [Langfelder et al., 2007] to the dendrogram to construct the gene clusters, which were considered as gene modules. In the third method, to obtain gene modules, we applied a functional gene module detection algorithm (namely FGMD) [Jin and Lee, 2017], which was based on hierarchical clustering but modified so that a gene was allowed

Table 3.2: Summary of enrichment analysis results obtained by four methods (CanMod, K-means, HC, and FGMD).

	CanMod	Kmeans	HC	FGMD
No. of input modules for enrichment analysis	643	763	169	149
Total no. of genes across all input modules	4,695	7,502	7,502	479
Avg. occurrence of genes across all input modules	2.3	1.0	1.0	6.9
GO-ES KEGG-ES CH-ES	6.06 4.82 5.54	4.29 3.40 3.72	5.13 4.57 4.46	4.19 4.82 5.92
No. of BP-enriched modules No. of KEGG-enriched modules No. of CH-enriched modules	511 (79.5%) 250 (38.9%) 152 (23.7%)	106 (13.9%) 103 (13.5%) 88 (11.5%)	42 (24.9%) 29 (17.2%) 36 (21.3%)	146 (98.0%) 116 (77.9%) 131 (87.9%)
Avg. occurrence of an enriched GO term Avg. occurrence of an enriched KEGG term Avg. occurrence of an enriched CH term	4.1 4.0 5.1	1.6 1.7 3.2	1.4 1.2 2.1	14.0 12.5 21.4

to belong to multiple modules. The enrichment analysis results obtained by the CanMod, K-means, HC, and FGMD are summarized in Table 3.2.

To quantify the enrichment levels of all enriched terms from all the modules, for each method we computed its GO enrichment score (GO-ES), KEGG enrichment score (KEGG-ES), and CH enrichment score (CH-ES). Suppose there were

totally T enriched GO terms across all the modules obtained in a method, then GO-ES of the method was defined as $GO\text{-}ES = \frac{1}{T}\sum_{i=1}^{T}$ -$\log_{10}$ pvalue(T_{i})

CanMod had the highest GO-ES among the four methods, and nearly 80% of modules in CanMod were associated with at least one BP term. The result indicated that gene targets found by CanMod were strongly associated with GO terms. The result was not surprising because CanMod employed GO terms to cluster target genes, which contributed to the high possibility that the genes associated with similar GO terms were together in the final modules. As shown in Table 3.2, CanMod had comparable high KEGG-ES and CH-ES values to those of FGMD.

FGMD outperformed the other methods in terms of the percentage of modules that were enriched with at least one BP, KEGG, or CH term. The result could be explained by examining the average occurrence of genes across all input modules in FGMD and the average occurrence of an enriched GO/KEGG/CH term across all the enriched modules. In FGMD on average a gene was included in seven modules compared to two modules in CanMod. In addition, the number of input modules for enrichment analysis and the total number of genes across all input modules were much smaller in FGMD compared to CanMod (see Table 3.2). It led to our hypothesis that that similar groups of genes would appear frequently in modules in FGMD, which increased the chance that similar terms would be enriched in different modules in FGMD. The high average occurrences of an enriched terms

across all modules in FGMD validated our hypothesis. It is clear that compared to FGMD, CanMod was better at finding distinctive enriched terms, but not as good as K-means and HC. The result is expected because K-means and HC did not allow a gene to be assigned in multiple modules, which lowered the chance a term would be enriched in different modules. In general, as we required different modules to be distinguishable based on their biological functions while allowing a gene to be associated with different modules, compared to other methods, CanMod achieved the best trade-off between the two criteria.

3.6 Summary

In this chapter, we presented CanMod, a computational method to infer cancer-associated gene regulatory modules. Each module found in CanMod contains genes that have similar biological functions and regulators that coregulate the genes. Unlike some existing methods, CanMod does not require users to specify the number of modules in advance. In CanMod, the target genes and regulators are allowed to belong to multiple modules, which follows the biological fact that a single gene or regulator may be involved in multiple biological processes.

We applied CanMod to infer gene regulatory modules in breast cancer using the BRCA dataset from TCGA. We found that the regulators that appeared in many modules were known cancer genes. We also observed that a significant num-

ber of experimentally validated regulator-target interactions occurred in the modules obtained by CanMod. Expression of target genes in large size modules were significantly correlated. Functional enrichment analysis applied on the target genes indicated that the coregulated genes in the modules were significantly enriched with cancer-related biological processes. Given the results, CanMod is a valuable tool to help deciphering the complex coregulatory relationship between TFs and miRNAs in cancer biology.

CHAPTER 4

CONCLUSION AND FUTURE WORK

MiRNAs play important roles in regulating genome-wide gene expression, thereby controlling many crucial biological processes. Dysregulation of miRNA regulation can lead to tumor formation and progression in many types of cancer [Jansson and Lund, 2012, Peng and Croce, 2016]. Many functional roles of miRNAs, especially in complex diseases such as cancer, are still unknown. This dissertation presented two computational tools that integrated multiple types of biological data to model two miRNA-mediated gene regulation mechanisms. In Chapter 2, we presented the computational tool Cancerin, which identifies genome-wide cancer-associated ceRNA interaction networks. In Chapter 3, we presented the computational tool CanMod, which identifies cancer-associated gene regulatory modules consisting of miRNAs, TFs, and their coregulated target genes. This chapter summarizes the contributions of the two computational tools and the important biological findings obtained by applying the tools to different cancer datasets. We conclude the dissertation by pointing out multiple research directions to extend the work discussed in the dissertation.

4.1 Summary of Cancerin and CanMod

CeRNA interaction is a post-transcriptional gene regulation that involves interactions between RNAs competing for common miRNA regulators. Dysregulation of ceRNA interactions have been implicated in multiple diseases, including cancer. In Chapter 2, we described the computational pipeline Cancerin, which infers genome-wide ceRNA interaction networks in cancer. Unlike existing ceRNA identification methods that consider miRNAs as the only factor regulating gene expression, Cancerin takes into account other types of gene regulators besides miRNAs, which include transcription factor (TF), copy number alteration (CNA), and DNA methylation (DM). Taking into account other types of gene regulators helps avoid spurious inference of miRNA-target gene interactions, which would result in spurious inference of ceRNA interactions. Cancerin is able to find ceRNA interactions among mRNAs, between mRNAs and lncRNAs, and among lncRNAs. By applying the sensitivity correlation metric proposed in Paci et al. [2014], Cancerin directly models the ceRNA hypothesis, which posited that two ceRNAs participating in a ceRNA interaction should have a positive expression correlation, and that correlation is conditioned on the expression of their common miRNA regulators.

To identify miRNA regulators for each gene, Cancerin incorporates a LASSO-based variable selection procedure that leverages both sequence-based and gene expression information. Then multiple expression-based filtering conditions are em-

ployed to select ceRNA interactions. Cancerin was applied to three cancer datasets from TCGA. Functional analysis indicated that the inferred ceRNAs were enriched with cancer-related genes, and ceRNAs within ceRNA modules (densely-connected ceRNAs) were involved in cancer-associated biological processes. Survival analysis showed that compared to non-ceRNAs, ceRNAs hold better prognostic power to predict survival outcomes. Our results showed that Cancerin can be used to identify genome-wide and functionally important ceRNA interactions, which makes it a valuable tool to explore the roles of this recently discovered gene regulation mechanism in cancer biology.

In Chapter 3, we described the computational tool CanMod to infer cancer-associated gene regulatory modules composed of miRNAs, TFs, and their coregulated genes. We require the modules obtained by CanMod to satisfy several conditions. First, different modules should have distinctive biological functions, which makes them biologically meaningful and interpretable. Second, the expression of target genes in a module is dependent on the expression of their regulators. Third, a regulator (i.e., miRNA or TF) and a target gene should be allowed to be included in more than one module. As different modules are distinguishable by their biological functions, this requirement reflects a biological fact that a regulator or a gene can carry multiple biological functions. In addition, in CanMod there is no need to specify the number of modules in advance.

CanMod was applied to the breast cancer dataset from TCGA. CanMod was able to rediscover experimentally validated regulator-target interactions. We found that the expression of target genes in the large modules were significantly correlated, and the hub regulators in CanMod were known cancer-associated genes. By applying functional enrichment analysis to the target genes in each module, we observed that different modules were associated with distinctive biological processes, and many modules were associated with cancer-related hallmarks. Given the results, CanMod is a valuable tool to help deciphering the complex coregulatory relationship between TFs and miRNAs in cancer.

4.2 Future Work

While our understanding of gene regulation by miRNAs continues to develop, many functional roles of miRNAs are yet to be discovered. Due to the complex interaction between miRNAs and their target genes and the intricate interplay between miRNAs and the other types of gene regulators such as TFs, computational methods are important tools to study miRNAs. This section points out several research directions in which computational approaches could be used to study miRNA regulation.

In Chapter 2, we described the computational tool Cancerin, which infers cancer-associated ceRNA interactions. Besides mRNAs and lncRNAs, which were

considered as potential ceRNAs in Cancerin, other types of RNAs such as pseudogenes and circular RNAs have also been found to participate in ceRNA interactions [Yamamura et al., 2017]. When data of those ncRNAs become more available in the future, they can be easily incorporated as additional input for Cancerin, and Cancerin can be used to explore their biological importance in the context of ceRNA regulation.

In each ceRNA interaction found by Cancerin, the constituent RNAs are regulated directly by a common set of miRNAs. However, multiple RNAs also can communicate, therefore regulate, each other through multiple layers of ceRNA interactions, and each layer is mediated by different set of miRNAs. Such indirect ceRNA interactions probably involve many miRNAs and RNA targets. The RNAs involved in such chain-based multilayered ceRNA interactions can govern crucial molecular functions and biological pathways. The indirect ceRNA interactions form a complex gene regulation mechanism that can control important biological functions. Identification of such interactions would be of great interest.

Recent studies have suggested several important factors for determining the strength of ceRNA interactions such as the number of shared miRNA-response-elements (MREs), the number of shared miRNAs, and the miRNA-RNA binding affinity [Denzler et al., 2014, 2016]. However, the optimal configuration of those factors for ceRNA interactions to occur in different cellular conditions (e.g., normal

cells versus tumor cells) remains unclear. Machine learning approaches could be employed to address the challenge. For example, supervised machine learning approaches can be used to predict if two RNAs can establish ceRNA interactions in different cellular settings. Under that scheme, the features for the supervised learning algorithms are the above mentioned factors (e.g., number of shared MREs and number of shared miRNAs). The outcome variables are binary values indicating the existence of ceRNA interactions based on the experimentally validated results.

Admittedly, supervised machine approaches require historical data of validated ceRNA interactions. However, a repository of those interactions is not yet available. Thus, a comprehensive database of experimentally validated ceRNA interactions in different cellular conditions will be extremely helpful to build predictive models for ceRNA interactions. Constructing a such database will be of great significance.

In Chapter 3, we described a computational method called CanMod, which infers cancer-associated gene regulatory modules consisting of miRNAs, TFs, and their coregulated genes. While specifying the cooperation of miRNAs and TFs in coregulating their commonly targeted genes, CanMod does not examine the regulatory relation among the regulators (e.g., miRNA regulates ($\rightarrow$) TF and TF $\rightarrow$ miRNA). "Directional gene regulatory modules" can be thought of as extensions of the modules obtained by CanMod. A directional module has all the characteristics

of a CanMod module. In addition, within the directional modules, the regulatory direction among the regulators is specified. Multiple interesting biological findings can be obtained from the directional modules. For example, the directionality between nodes (i.e., regulators → regulators, and regulators → targets) in a module can present some specific biological pathways. Within a directional module, there will be some hub regulators, which are the regulators that target many other regulators and target genes. Since a gene module is associated with specific biological functions, the hub regulators in a module can be thought of as the trigger of the functions; thus, hub regulators can have critical biological significance and are worthy of further examination.

Probabilistic graphical approaches can be used to infer such directional gene regulatory modules [Friedman et al., 1997, Yu et al., 2004]. For instance, the authors in [Gosline et al., 2016] introduced a probabilistic network modeling approach to model the how miRNAs regulates gene expression via TFs (i.e., miRNAs → TFs → genes). Biological knowledge (i.e., priors) can be used to facilitate inference of the module structure. An inference algorithm can assign a high probability that a specific miRNA will regulate specific TF if the TF's sequence contains many binding sites for the miRNA and there is also a CLIP-seq experiment showing that the miRNA/RISC complex can bind to the TF's 3'UTR. In addition to using the

regulator-target binding information, gene expression can also be used to update the network structure and learn the network parameters.

Advances in sequencing technologies will yield more time-series sequencing data. Hence, it is possible to learn the directional gene regulatory module at different time points to study how the regulatory interactions evolved. It will be interesting to compare a regulatory module at a tumor-free timepoint with the module itself at a tumor-formation timepoint to see what alters in the module at different timepoints that triggers the tumor creation.

Most existing computational methods to study miRNA-gene regulatory networks are based on using the gene expression of a group of samples with similar cellular condition. However, samples collected from patients diagnosed with a similar disease can still exhibit a highly heterogeneous expression profiles [Hu et al., 2013, Norton et al., 2016]. In such cases, a single gene regulatory network constructed by those samples is not informative enough, or even misleading, to represent the gene regulatory network of the individual patients. Thus, we need to infer gene regulatory networks for individual patients. Such patient-specific gene regulatory models can be very helpful for precision medicine and targeted therapy research. For instance, patient-specific miRNA-gene regulatory models can facilitate high-resolution investigation on how miRNA's activities are different among different pa-

tients. Hence, they can enable the identification of new cancer types or novel cancer subtypes.

However, building patient-specific gene regulatory networks is challenging because the inference of connection between nodes (i.e., regulators and targets) in a network often is based on measurements such as gene expression correlation or mutual information from a group of samples, which does not satisfy the patient-specific constraints. To address that challenge, one computational approach proposed by [Liu et al., 2016] can be used as a framework to construct patient-specific gene regulatory networks. Their basic idea is that to construct a network for a patient P_i, we first construct a network using the data from all patients including P_i, which is referred to as network $\{P\}$. Then we construct another network called $\{P\}_{\text{-i}}$, which uses the data from all the patients except the patient P_i. Thus, the network for the patient P_i is inferred by comparing the two networks $\{P\}$ and $\{P\}_{\text{-i}}$. The edges that are significantly different between the two networks are used to construct the network for P_i. While this approach is able to discover gene interactions that are distinctive to P_i, it is not able to rediscover the common interactions between network P_i and network $\{P\}$. In addition, it is not clear to how to infer interaction directionality using this approach. Hence, there is a need for more sophisticated computational approaches to infer patient-specific models.

In conclusion, this dissertation described two computational methods to

model the two miRNA-mediated gene regulation mechanisms. As we gain deeper understanding of miRNA regulation mechanisms and at the same time, and more genomic and clinical data are collected, computational methods are indispensable tools for us to acquire new insights of miRNA functions in cancer.

BIBLIOGRAPHY

Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, 4:e05005, 2015.

Sarah Aldridge and James Hadfield. Introduction to miRNA profiling technologies and cross-platform comparison. In *Next-Generation MicroRNA Expression Profiling Technology*, pages 19–31. Springer, 2012.

Victor Ambros. A hierarchy of regulatory genes controls a larva-to-adult developmental switch in *C. elegans*. *Cell*, 57(1):49–57, 1989.

Omer An, Giovanni M Dall'Olio, Thanos P Mourikis, and Francesca D Ciccarelli. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. *Nucleic Acids Research*, 44 (D1):D992–D999, 2015.

Yang An, Kendra L Furber, and Shaoping Ji. Pseudogenes regulate parental gene expression via ceRNA network. *Journal of Cellular and Molecular Medicine*, 21 (1):185–192, 2017.

Prema Arasu, Bruce Wightman, and Gary Ruvkun. Temporal regulation of lin-14 by the antagonistic action of two other heterochronic genes, lin-4 and lin-28. *Genes & Development*, 5(10):1825–1833, 1991.

Sanghamitra Bandyopadhyay and Ramkrishna Mitra. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics*, 25(20):2625–2631, 2009.

Raffaela Barbano, Barbara Pasculli, Michelina Rendina, Andrea Fontana, Caterina Fusilli, Massimiliano Copetti, Stefano Castellana, Vanna Maria Valori, Maria Morritti, Paolo Graziano, et al. Stepwise analysis of MIR9 loci identifies miR-9-5p to be involved in Oestrogen regulated pathways in breast cancer patients. *Scientific Reports*, 7:45283, 2017.

D Barh, R Malhotra, B Ravi, and P Sindhurani. MicroRNA let-7: an emerging next-generation cancer therapeutic. *Current Oncology*, 17(1):70, 2010.

David P Bartel. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, 2004.

David P Bartel. Metazoan MicroRNAs. *Cell*, 173(1):20–51, 2018.

Stephen B Baylin. DNA methylation and gene silencing in cancer. *Nature Reviews Clinical Oncology*, 2(S1):S4, 2005.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A prac-

tical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Rameen Beroukhim, Craig H Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S Boehm, Jennifer Dobson, Mitsuyoshi Urashima, et al. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899, 2010.

Daniel Bushman. Cancer Gene List. Visited on 2017-09-01. URL `http://www.bushmanlab.org/links/genelists`.

George Adrian Calin, Calin Dan Dumitru, Masayoshi Shimizu, Roberta Bichi, Simona Zupo, Evan Noch, Hansjuerg Aldler, Sashi Rattan, Michael Keating, Kanti Rai, et al. Frequent deletions and down-regulation of microRNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*, 99(24):15524–15529, 2002.

Caterina Catalanotto, Carlo Cogoni, and Giuseppe Zardo. MicroRNA in control of gene expression: An overview of nuclear functions. *International Journal of Molecular Sciences*, 17(10):1712, 2016.

James WF Catto, Antonio Alcaraz, Anders S Bjartell, Ralph De Vere White, Christopher P Evans, Susanne Fussel, Freddie C Hamdy, Olli Kallioniemi, Lourdes Mengual, Thorsten Schlomm, et al. MicroRNA in prostate, bladder, and kidney cancer: A systematic review. *European Urology*, 59(5):671–681, 2011.

Marcella Cesana, Davide Cacchiarelli, Ivano Legnini, Tiziana Santini, Olga Sthandier, Mauro Chinappi, Anna Tramontano, and Irene Bozzoni. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147(2):358–369, 2011.

Ping Chen, Colles Price, Zejuan Li, Yuanyuan Li, Donglin Cao, Anissa Wiley, Chunjiang He, Sandeep Gurbuxani, Rejani B Kunjamma, Hao Huang, et al. MiR-9 is an essential oncogenic microRNA specifically overexpressed in mixed lineage leukemia–rearranged leukemia. *Proceedings of the National Academy of Sciences*, 110(28):11511–11516, 2013.

Shou-Tung Chen, Kong-Bung Choo, Ming-Feng Hou, Kun-Tu Yeh, Shou-Jen Kuo, and Jan-Gowth Chang. Deregulated expression of the PER1, PER2 and PER3 genes in breast cancers. *Carcinogenesis*, 26(7):1241–1246, 2005.

Jill Cheng, Philipp Kapranov, Jorg Drenkow, Sujit Dike, Shane Brubaker, Sandeep Patel, Jeffrey Long, David Stern, Hari Tammana, Gregg Helt, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308 (5725):1149–1154, 2005.

Sung Wook Chi, Julie B Zang, Aldo Mele, and Robert B Darnell. Argonaute HITS-

CLIP decodes microRNA–mRNA interaction maps. *Nature*, 460(7254):479, 2009.

Hua-Sheng Chiu, David Llobet-Navas, Xuerui Yang, Wei-Jen Chung, Alberto Ambesi-Impiombato, Archana Iyer, Hyunjae Ryan Kim, Elena G Seviour, Zijun Luo, Vasudha Sehgal, et al. Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. *Genome Research*, 25(2):257–267, 2015.

Hua-Sheng Chiu, María Rodríguez Martínez, Mukesh Bansal, Aravind Subramanian, Todd R Golub, Xuerui Yang, Pavel Sumazin, and Andrea Califano. High-throughput validation of ceRNA regulatory networks. *BMC Genomics*, 18(1):418, 2017.

Chih-Hung Chou, Sirjana Shrestha, Chi-Dung Yang, Nai-Wen Chang, Yu-Ling Lin, Kuang-Wen Liao, Wei-Chi Huang, Ting-Hsuan Sun, Siang-Jyun Tu, Wei-Hsiang Lee, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 46(D1):D296–D302, 2017.

Matthew Cobb. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biology*, 15(9):e2003243, 2017.

Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71–e71, 2016.

Marion Coolen, Shauna Katz, and Laure Bally-Cuif. MiR-9: A versatile regulator of neurogenesis. *Frontiers in Cellular Neuroscience*, 7:220, 2013.

Geoffrey M Cooper and Robert E Hausman. The development and causes of cancer. *The cell: A Molecular Approach*, pages 725–766, 2000.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

Shaoli Das, Suman Ghosal, Rituparno Sen, and Jayprokas Chakrabarti. lnCeDB: database of human long noncoding RNA acting as competing endogenous RNA. *PloS ONE*, 9(6):e98965, 2014.

John Day-Richter, Midori A Harris, Melissa Haendel, and Suzanna Lewis. Obo-edit—an ontology editor for biologists. *Bioinformatics*, 23(16):2198–2200, 2007.

Kristin R Delfino and Sandra L Rodriguez-Zas. Transcription factor-microRNA-target gene networks associated with ovarian cancer survival and recurrence. *PloS ONE*, 8(3):e58608, 2013.

Ahmet M Denli, Bastiaan BJ Tops, Ronald HA Plasterk, René F Ketting, and Gregory J Hannon. Processing of primary microRNAs by the microprocessor complex. *Nature*, 432(7014):231–235, 2004.

Rémy Denzler, Vikram Agarwal, Joanna Stefano, David P Bartel, and Markus Stoffel. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Molecular Cell*, 54(5):766–776, 2014.

Rémy Denzler, Sean E McGeary, Alexandra C Title, Vikram Agarwal, David P Bartel, and Markus Stoffel. Impact of microRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression. *Molecular Cell*, 64(3):565–579, 2016.

Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G Knowles, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*, 22 (9):1775–1789, 2012.

Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101, 2012.

Hristo Djidjev. A fast multilevel algorithm for graph clustering and community detection. *arXiv preprint arXiv:0707.2387*, 2007.

Duc Do and Serdar Bozdag. Cancerin: A computational method to identify cancer-associated competing endogenous RNA interactions mediated by miRNA regulation. *PLoS Computational Biology*, 2018.

Harsh Dweep, Carsten Sticht, Priyanka Pandey, and Norbert Gretz. MiRWalk database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *Journal of Biomedical Informatics*, 44(5):839–847, 2011.

Charles F Ehret and Gérard De Haller. Origin, development, and maturation of organelles and organelle systems of the cell surface in paramecium. *Journal of Ultrastructure Research*, 9:1–42, 1963.

Semih Ekimler and Kaniye Sahin. Computational methods for microRNA target prediction. *Genes*, 5(3):671–683, 2014.

ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799, 2007.

Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. MicroRNA targets in Drosophila. *Genome Biology*, 5(1):R1, 2003.

Agustin F Fernandez, Yassen Assenov, Jose Ignacio Martin-Subero, Balazs Balint, Reiner Siebert, Hiroaki Taniguchi, Hiroyuki Yamamoto, Manuel Hidalgo, Aik-

Choon Tan, Oliver Galm, et al. A DNA methylation fingerprint of 1628 human samples. *Genome Research*, 22(2):407–419, 2012.

Simon A Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45 (D1):D777–D783, 2016.

Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

Meital Gabay, Yulin Li, and Dean W Felsher. MYC activation is a hallmark of cancer initiation and maintenance. *Cold Spring Harbor Perspectives in Medicine*, 4(6):a014241, 2014.

Milena Gasco, Shukri Shami, and Tim Crook. The P53 pathway in breast cancer. *Breast Cancer Research*, 4(2):70, 2002.

Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

Sara JC Gosline, Allan M Gurtan, Courtney K JnBaptiste, Andrew Bosson, Pamela Milani, Simona Dalin, Bryan J Matthews, Yoon S Yap, Phillip A Sharp, and Ernest Fraenkel. Elucidating icroRNA regulatory networks using transcriptional, post-transcriptional, and histone modification measurements. *Cell Reports*, 14(2):310–319, 2016.

Richard I Gregory, Kai-ping Yan, Govindasamy Amuthan, Thimmaiah Chendrimada, Behzad Doratotaj, Neil Cooch, and Ramin Shiekhattar. The Microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014):235, 2004.

Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

Minju Ha and V Narry Kim. Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology*, 15(8):509, 2014.

Mark P Hamilton, Kimal Rajapakshe, Sean M Hartig, Boris Reva, Michael D McLellan, Cyriac Kandoth, Li Ding, Travis I Zack, Preethi H Gunaratne,

David A Wheeler, et al. Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. *Nature Communications*, 4:2730, 2013.

Heonjong Han, Hongseok Shim, Donghyun Shin, Jung Eun Shim, Yunhee Ko, Junha Shin, Hanhae Kim, Ara Cho, Eiru Kim, Tak Lee, et al. TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific Reports*, 5, 2015.

KD Hansen. IlluminaHumanMethylation450kanno. ilmn12. hg19: Annotation for Illumina's 450k methylation arrays. *R package, version 0.2*, 1, 2015.

Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, 2012.

Josie Hayes, Pier Paolo Peruzzi, and Sean Lawler. MicroRNAs in cancer: biomarkers, functions and therapy. *Trends in Molecular Medicine*, 20(8):460–469, 2014.

Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

Jie Hu, Jason W Locasale, Jason H Bielas, Jacintha O'Sullivan, Kieran Sheahan, Lewis C Cantley, Matthew G Vander Heiden, and Dennis Vitkup. Heterogeneity of tumor-induced gene expression changes in the human metabolic network. *Nature Biotechnology*, 31(6):522–529, 2013.

Alexander Hüttenhofer, Peter Schattner, and Norbert Polacek. Non-coding RNAs: hope or hype? *Trends in Genetics*, 21(5):289–297, 2005.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931, 2004.

Marilena V Iorio, Manuela Ferracin, Chang-Gong Liu, Angelo Veronese, Riccardo Spizzo, Silvia Sabbioni, Eros Magri, Massimo Pedriali, Muller Fabbri, Manuela Campiglio, et al. MicroRNA gene expression deregulation in human breast cancer. *Cancer Research*, 65(16):7065–7070, 2005.

Irena Ivanovska, Alexey S Ball, Robert L Diaz, Jill F Magnus, Miho Kibukawa, Janell M Schelter, Sumire V Kobayashi, Lee Lim, Julja Burchard, Aimee L Jackson, et al. MicroRNAs in the miR-106b family regulate p21/CDKN1A and promote cell cycle progression. *Molecular and Cellular Biology*, 28(7):2167–2174, 2008.

Ritu Jain, Tiffany Devine, Ajish D George, Sridar V Chittur, Timothy E Baroni, Luiz O Penalva, and Scott A Tenenbaum. RIP-Chip analysis: RNA-binding pro-

tein immunoprecipitation-microarray (Chip) profiling. In *RNA*, pages 247–263. Springer, 2011.

Martin D Jansson and Anders H Lund. MicroRNA and cancer. *Molecular Oncology*, 6(6):590–610, 2012.

Ashwini Jeggari, Debora S Marks, and Erik Larsson. miRcode: A map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics*, 28 (15):2062–2063, 2012.

Daeyong Jin and Hyunju Lee. A computational approach to identifying gene-microRNA modules in cancer. *PLoS Computational Biology*, 11(1):e1004042, 2015.

Daeyong Jin and Hyunju Lee. FGMD: A novel approach for functional gene module detection in cancer. *PloS ONE*, 12(12):e0188900, 2017.

Bino John, Anton J Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, and Debora S Marks. Human microRNA targets. *PLoS Biology*, 2(11):e363, 2004.

Bryony Jones. Gene expression: layers of gene regulation. *Nature Reviews Genetics*, 16(3):128, 2015.

Peter A Jones. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484, 2012.

M Jovanovic and MO Hengartner. MiRNAs and apoptosis: RNAs to die for. *Oncogene*, 25(46):6176–6187, 2006.

Florian A Karreth and Pier Paolo Pandolfi. ceRNA cross-talk in cancer: when ce-bling rivalries go awry. *Cancer Discovery*, 3(10):1113–1121, 2013.

Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39(10): 1278, 2007.

Seok-Jun Kim, Ji-Young Shin, Kang-Duck Lee, Young-Ki Bae, Ki Woong Sung, Seok Jin Nam, and Kyung-Hee Chun. MicroRNA let-7a suppresses breast cancer cell migration and invasion through downregulation of CC chemokine receptor type 7. *Breast Cancer Research*, 14(1):R14, 2012.

V Narry Kim. MicroRNA precursors in motion: exportin-5 mediates their nuclear export. *Trends in cell biology*, 14(4):156–159, 2004.

Ulf Klein, Marie Lia, Marta Crespo, Rachael Siegel, Qiong Shen, Tongwei Mo, Alberto Ambesi-Impiombato, Andrea Califano, Anna Migliazza, Govind Bhagat, et al. The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell*, 17(1):28–40, 2010.

Ana Kozomara and Sam Griffiths-Jones. miRBase: annotating high confidence mi-

croRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1):D68–D73, 2013.

Jan Krüger and Marc Rehmsmeier. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research*, 34(suppl_2):W451–W454, 2006.

Kimberly R Kukurba and Stephen B Montgomery. RNA sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11):pdb–top084970, 2015.

Jennifer YY Kwan, Pamela Psarianos, Jeff P Bruce, Kenneth W Yip, and Fei-Fei Liu. The complexity of microRNAs in human cancer. *Journal of Radiation Research*, 57(S1):i106–i111, 2016.

Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24(5): 719–720, 2007.

Nelson C Lau, Lee P Lim, Earl G Weinstein, and David P Bartel. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans* . *Science*, 294(5543):858–862, 2001.

Thuc Duy Le, Junpeng Zhang, Lin Liu, and Jiuyong Li. Computational methods for identifying miRNA sponge interactions. *Briefings in Bioinformatics*, page bbw042, 2016.

Rosalind C Lee, Rhonda L Feinbaum, and Victor Ambros. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

Cheng Li, Lufeng Zheng, Yu Xin, Zhoulin Tan, Yan Zhang, Xia Meng, Zhigang Wang, and Tao Xi. The competing endogenous RNA network of CYP4Z1 and pseudogene CYP4Z2P exerts an anti-apoptotic function in breast cancer. *FEBS Letters*, 591(7):991–1000, 2017.

Jun Li, Leng Han, Paul Roebuck, Lixia Diao, Lingxiang Liu, Yuan Yuan, John N Weinstein, and Han Liang. TANRIC: An interactive open platform to explore the function of lncRNAs in cancer. *Cancer Research*, 75(18):3728–3737, 2015.

Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-seq data. *Nucleic Acids Research*, 42(D1):D92–D97, 2013.

Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

Bing Liu, Jiuyong Li, and Anna Tsykin. Discovery of functional miRNA–mRNA regulatory modules with computational methods. *Journal of Biomedical Infor-*

matics, 42(4):685–691, 2009.

Chenglin Liu, Jing Su, Fei Yang, Kun Wei, Jinwen Ma, and Xiaobo Zhou. Compound signature detection on lincs l1000 big data. *Molecular BioSystems*, 11(3): 714–722, 2015a.

Kui Liu, Chunfu Zhang, Tao Li, Yanling Ding, Tao Tu, Fangfang Zhou, Wenkai Qi, Huabiao Chen, and Xiaochun Sun. Let-7a inhibits growth and migration of breast cancer cells by targeting HMGA1. *International Journal of Oncology*, 46 (6):2526–2534, 2015b.

Xiaoping Liu, Yuetong Wang, Hongbin Ji, Kazuyuki Aihara, and Luonan Chen. Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Research*, page gkw772, 2016.

E Lund and JE Dahlberg. Substrate selectivity of Exportin 5 and Dicer in the biogenesis of microRNAs. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 71, pages 59–66. Cold Spring Harbor Laboratory Press, 2006.

Natalia J Martinez and Albertha JM Walhout. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *Bioessays*, 31(4): 435–445, 2009.

David H Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.

Avi Ma'ayan. Introduction to network analysis in systems biology. *Science Signaling*, 4(190):tr5, 2011.

Susanne Motameny, Stefanie Wolters, Peter Nürnberg, and Björn Schumacher. Next generation sequencing of miRNAs–strategies, resources and methods. *Genes*, 1(1):70–84, 2010.

Anirban Mukhopadhyay and Ujjwal Maulik. Network-based study reveals potential infection pathways of Hepatitis-C leading to various diseases. *PloS ONE*, 9(4): e94029, 2014.

Genta Nagae, Takayuki Isagawa, Nobuaki Shiraki, Takanori Fujita, Shogo Yamamoto, Shuichi Tsutsumi, Aya Nonaka, Sayaka Yoshiba, Keisuke Matsusaka, Yutaka Midorikawa, et al. Tissue-specific demethylation in CpG-poor promoters during cellular differentiation. *Human Molecular Genetics*, 20(14):2710–2721, 2011.

Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.

Nadine Norton, Pooja P Advani, Daniel J Serie, Xochiquetzal J Geiger, Brian M Necela, Bianca C Axenfeld, Jennifer M Kachergus, Ryan W Feathers, Jennifer M

Carr, Julia E Crook, et al. Assessment of tumor heterogeneity, as evidenced by gene expression profiles, pathway activation, and gene copy number, in patients with multifocal invasive lobular breast tumors. *PloS ONE*, 11(4):e0153411, 2016.

Susumu Ohno. So much "junk" DNA in our genome. In *Brookhaven Symposium in Biology*, volume 23, pages 366–370, 1972.

Paola Paci, Teresa Colombo, and Lorenzo Farina. Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. *BMC Systems Biology*, 8(1):83, 2014.

Alexander F Palazzo and Eliza S Lee. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics*, 6:2, 2015.

Maria D Paraskevopoulou, Georgios Georgakilas, Nikos Kostoulas, Ioannis S Vlachos, Thanasis Vergoulis, Martin Reczko, Christos Filippidis, Theodore Dalamagas, and Artemis G Hatzigeorgiou. DIANA-microT web server v5. 0: service integration into miRNA functional analysis workflows. *Nucleic Acids Research*, 41(W1):W169–W173, 2013.

Maria D Paraskevopoulou, Ioannis S Vlachos, Dimitra Karagkouni, Georgios Georgakilas, Ilias Kanellos, Thanasis Vergoulis, Konstantinos Zagganas, Panayiotis Tsanakas, Evangelos Floros, Theodore Dalamagas, et al. Diana-lncbase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Research*, 44 (D1):D231–D238, 2015.

Amy E Pasquinelli. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics*, 13(4):271–282, 2012.

Mugdha Patki, Venkatesh Chari, Suneethi Sivakumaran, Mesfin Gonit, Robert Trumbly, and Manohar Ratnam. The ETS domain transcription factor ELK1 directs a critical component of growth signaling by the androgen receptor in prostate cancer cells. *Journal of Biological Chemistry*, 288(16):11047–11065, 2013.

Xinxia Peng, Yu Li, Kathie-Anne Walters, Elizabeth R Rosenzweig, Sharon L Lederer, Lauri D Aicher, Sean Proll, and Michael G Katze. Computational identification of Hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC Genomics*, 10(1):373, 2009.

Yong Peng and Carlo M Croce. The role of microRNAs in human cancer. *Signal Transduction and Targeted Therapy*, 1:15004, 2016.

Sarah M Peterson, Jeffrey A Thompson, Melanie L Ufkin, Pradeep Sathyanarayana, Lucy Liaw, and Clare Bates Congdon. Common features of microRNA target prediction tools. *Frontiers in Genetics*, 5:23, 2014.

Theresa Phillips. The role of methylation in gene expression. *Nature Education*, 1 (1):116, 2008.

Leslie Pray. Discovery of DNA structure and function: Watson and Crick. *Nature Education*, 1(1):100, 2008.

Sheng Qin, Fei Ma, and Liming Chen. Gene regulatory networks by transcription factors and microRNAs in breast cancer. *Bioinformatics*, 31(1):76–83, 2014.

Brenda J Reinhart, Frank J Slack, Michael Basson, Amy E Pasquinelli, Jill C Bettinger, Ann E Rougvie, H Robert Horvitz, and Gary Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403 (6772):901, 2000.

Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

Damian Roqueiro, Lei Huang, and Yang Dai. Identifying transcription factors and microRNAs as key regulators of pathways using Bayesian inference on known pathway structures. *Proteome Science*, 10(1):S15, 2012.

Sarah Roush and Frank J Slack. The let-7 family of microRNAs. *Trends in Cell Biology*, 18(10):505–516, 2008.

Marilyn Safran, Irina Dalah, Justin Alexander, Naomi Rosen, Tsippi Iny Stein, Michael Shmoish, Noam Nativ, Iris Bahir, Tirza Doniger, Hagit Krug, et al. GeneCards Version 3: the human gene integrator. *Database*, 2010, 2010.

Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, 146(3):353–358, 2011.

Avencia Sanchez-Mejias and Yvonne Tay. Competing endogenous RNA networks: tying the essential knots for cancer biology and therapeutics. *Journal of Hematology & Oncology*, 8(1):30, 2015.

Richard J Santen, Robert Xinde Song, Robert McPherson, Rakesh Kumar, Liana Adam, Meei-Huey Jeng, and Wei Yue. The role of mitogen-activated protein (MAP) kinase in breast cancer. *The Journal of steroid biochemistry and molecular biology*, 80(2):239–256, 2002.

Marco Scutari. Learning Bayesian networks with the bnlearn R package. *arXiv preprint arXiv:0908.3817*, 2009.

Reut Shalgi, Daniel Lieber, Moshe Oren, and Yitzhak Pilpel. Global and local architecture of the mammalian microRNA–transcription factor regulatory network. *PLoS Computational Biology*, 3(7):e131, 2007.

Tingting Shao, Aiwei Wu, Juan Chen, Hong Chen, Jianping Lu, Jing Bai, Yongsheng Li, Juan Xu, and Xia Li. Identification of module biomarkers from the dysregulated ceRNA–ceRNA interaction network in lung adenocarcinoma. *Molecular Biosystems*, 11(11):3048–3058, 2015.

Adam Shlien and David Malkin. Copy number variations and cancer. *Genome medicine*, 1(6):62, 2009.

T Smith and M Waterman. Identification of common molecular subsequences. *Molecular Biology*, 147:195–197, 1981.

François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012.

Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl_1):D535–D539, 2006.

Pavel Sumazin, Xuerui Yang, Hua-Sheng Chiu, Wei-Jen Chung, Archana Iyer, David Llobet-Navas, Presha Rajbhandari, Mukesh Bansal, Paolo Guarnieri, Jose Silva, et al. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell*, 147(2):370–381, 2011.

Yvonne Tay, Lev Kats, Leonardo Salmena, Dror Weiss, Shen Mynn Tan, Ugo Ala, Florian Karreth, Laura Poliseno, Paolo Provero, Ferdinando Di Cunto, et al. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell*, 147(2):344–357, 2011.

Yvonne Tay, John Rinn, and Pier Paolo Pandolfi. The multilayered complexity of ceRNA crosstalk and competition. *Nature*, 505(7483):344, 2014.

Barry S Taylor, Jordi Barretina, Nicholas D Socci, Penelope DeCarolis, Marc Ladanyi, Matthew Meyerson, Samuel Singer, and Chris Sander. Functional copy-number alterations in cancer. *PloS ONE*, 3(9):e3179, 2008.

Daniel W Thomson, Cameron P Bracken, and Gregory J Goodall. Experimental strategies for microRNA target identification. *Nucleic Acids Research*, 39(16): 6845–6853, 2011.

Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Ryan J Tibshirani et al. The LASSO problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Dang Hung Tran, Kenji Satou, and Tu Bao Ho. Finding microRNA regulatory modules in human genome using rule induction. *BMC Bioinformatics*, 9(12): S5, 2008.

Nham Tran, Tessa McLean, Xiaoying Zhang, Chuan Jia Zhao, John Michael Thomson, Christopher O'Brien, and Barbara Rose. MicroRNA expression profiles in head and neck cancer cell lines. *Biochemical and Biophysical Research Communications*, 358(1):12–17, 2007.

Stefan Uhlmann, Heiko Mannsperger, Jitao David Zhang, Emöke-Ágnes Horvat, Christian Schmidt, Moritz Küblbeck, Frauke Henjes, Aoife Ward, Ulrich Tschulena, Katharina Zweig, et al. Global microRNA level regulation of EGFR-driven cell-cycle protein network in breast cancer. *Molecular Systems Biology*, 8(1):570, 2012.

Harm Van Bakel, Corey Nislow, Benjamin J Blencowe, and Timothy R Hughes. Most "dark matter" transcripts are associated with known genes. *PLoS Biology*, 8(5):e1000371, 2010.

Ty C Voss and Gordon L Hager. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics*, 15(2):69–81, 2014.

Dongfang Wang, Jin Gu, Ting Wang, and Zijian Ding. OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics*, 30(15):2237–2238, 2014.

James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, 2007.

Xirui Wang, Lei Cao, Yingyi Wang, Xiefeng Wang, Ning Liu, and Yongping You. Regulation of let-7 and its target oncogenes. *Oncology Letters*, 3(5):955–960, 2012.

Yuka Watanabe, Masaru Tomita, and Akio Kanai. Computational methods for microRNA target prediction. *Methods in Enzymology*, 427:65–86, 2007.

Bruce Wightman, Ilho Ha, and Gary Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell*, 75(5):855–862, 1993.

Tian Xia, Qi Liao, Xiaoming Jiang, Yongfu Shao, Bingxiu Xiao, Yang Xi, and Junming Guo. Long noncoding RNA associated-competing endogenous RNAs in gastric cancer. *Scientific Reports*, 4:6088, 2014.

Soichiro Yamamura, Mitsuho Imai-Sumida, Yuichiro Tanaka, and Rajvir Dahiya. Interaction and cross-talk between non-coding RNAs. *Cellular and Molecular Life Sciences*, pages 1–18, 2017.

Jue Yang, Tong Li, Chao Gao, Xiaobo Lv, Kunmei Liu, Hui Song, Yingying Xing, and Tao Xi. FOXO1 3'UTR functions as a ceRNA in repressing the metastases of breast cancer cells via regulating miRNA activity. *FEBS Letters*, 588(17):3218–

3224, 2014.

Yong Yang, Hong Yang, Miao Xu, Haibin Zhang, Mingtao Sun, Peng Mu, Tongbao Dong, Shanmei Du, and Kui Liu. Long non-coding RNA (lncRNA) MAGI2-AS3 inhibits breast cancer cell growth by targeting the Fas/FasL signalling pathway. *Human Cell*, pages 1–10, 2018.

Soraya Yekta, I-hung Shih, and David P Bartel. MicroRNA-directed cleavage of HOXB8 mRNA. *Science*, 304(5670):594–596, 2004.

Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: An R package for comparing biological themes among gene clusters.

Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.

Jing Yu, V Anne Smith, Paul P Wang, Alexander J Hartemink, and Erich D Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594–3603, 2004.

Benedikt Zacher, Khalid Abnaof, Stephan Gade, Erfan Younesi, Achim Tresch, and Holger Fröhlich. Joint Bayesian inference of condition-specific miRNA and transcription factor activities from combined gene and microRNA expression data. *Bioinformatics*, 28(13):1714–1720, 2012.

Guang-jun Zhang, Jian-shui Li, He Zhou, Hua-xu Xiao, Yu Li, and Tong Zhou. microRNA-106b promotes colorectal cancer cell migration and invasion by directly targeting DLC1. *Journal of Experimental & Clinical Cancer Research*, 34 (1):73, 2015.

Guangde Zhang, Haoran Sun, Yawei Zhang, Hengqiang Zhao, Wenjing Fan, Jianfei Li, Yingli Lv, Qiong Song, Jiayao Li, Mingyu Zhang, et al. Characterization of dysregulated lncRNA-mRNA network based on ceRNA hypothesis to reveal the occurrence and recurrence of myocardial infarction. *Cell Death Discovery*, 4(1): 35, 2018.

J Zhang. CNTools: Convert segment data into a region by sample matrix to allow for other high level computational analyses. *R Package (Version 1.6.0.)*, 2016.

Fang Zhao, Zhenyu Xuan, Lihua Liu, and Michael Q Zhang. TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies. *Nucleic Acids Research*, 33(suppl_1):D103–D107, 2005.

Xi Zhou, Qin Gao, Jianzhong Wang, Xin Zhang, Kaige Liu, and Zhao Duan. Linc-RNA-RoR acts as a "sponge" against mediation of the differentiation of endometrial cancer stem cells by microRNA-145. *Gynecologic Oncology*, 133(2):333–339, 2014a.

Xionghui Zhou, Juan Liu, and Wei Wang. Construction and investigation of breast-cancer-specific ceRNA network based on the mRNA and miRNA expression data. *IET Systems Biology*, 8(3):96–103, 2014b.

APPENDIX A

METHODS TO IDENTIFY MIRNA-TARGET INTERACTIONS

This appendix describes the principles behind common experimental and computational methods to identify miRNA-target interactions. Understanding the principles behind both the computational and experimental methods is crucial to deciding which method to use and when/how to use them together to generate the best set of putative miRNA-target interactions for different research purposes.

Most of the computational methods that are used to predict miRNA-target interaction explore the sequence complementary rules between miRNAs and their target genes. Using publicly available transcript sequences of miRNAs and candidate genes as input, computational methods are able to generate many candidate miRNA-target interactions. While not as time-efficient or as cost-effective as computational methods, experimental methods can not only provide strong evidence of miRNA-target interaction, but also can generate understanding of miRNA regulation mechanisms. It is also worth noting that every experimental method includes some computational parts; thus there is not always a clear line between experimental and computational methods.

A.1 Experimental methods

Common experimental methods to identify miRNA-target interaction follow two main approaches. The first approach is based on profiling of gene expression change as a response to miRNA overexpression/inhibition. MiRNAs repress translation and cause degradation of their targets, resulting in an inverse correlation between the abundance of the miRNAs and their target transcripts. Thus, this approach often is employed when researchers have a specific miRNA to study. They transfect (introduce) the miRNA mimics or inhibitors into the cell of interest. The expression of mRNAs is measured before and after the transfection; the genes that show significant change of expression are considered as candidate targets of the miRNA.

The methods to predict miRNA-target interactions based on gene expression profiling are unable to distinguish whether the target is a direct or an indirect target of the miRNA. For example, expression of a gene could be changed not because it is bound by the miRNA (as the miRNA's direct target), but because the gene's transcription factor was bound by the miRNA, which in turn alters the gene expression at the transcriptional level.

The second experimental approach to identify miRNA-target interaction is based on identifying the binding mark of RISC complexes, which are guided by

miRNAs, on their target transcripts [Thomson et al., 2011]. This approach is able to identify direct binding between miRNAs and their target transcripts. As RISC complexes contain one of the AGO proteins, one experimental technique following this approach involves using an antibody of the AGO protein associating with a miRNA of interest. To predict targets of the miRNA, the antibody is injected into the cells before and after the cells are transfected with the synthetic miRNAs. By introducing the antibody before and after the miRNA transfection, a directly bound target gene is detected based on measuring the antibody signals before and after the miRNA transfection. Methods such as RIP-ChIP (Ribonucleoprotein Immunoprecipitation followed by microarray chip analysis), or RIP-Seq (Ribonucleoprotein Immunoprecipitation followed by high-throughput sequencing) have been used to quantify that differential binding information [Jain et al., 2011]. A limitation of these techniques is that they cannot provide the precise binding location between RISCs and their target transcripts.

Recent high-throughput methods based on AGO cross-linking and immunoprecipitation (AGO CLIP) overcome this limitation [Chi et al., 2009]. This technique uses ultraviolet radiation to form a covalent bond (aka., cross-link) between the RISC complex and the target genes. Therefore after being immunopreciptated, the binding areas still contain the sequence information of the target genes, which can be mapped back into the reference genome to decide the exact binding location.

Thus, the AGO CLIP technique allows identification of genome-wide miRNA-target direct bindings. Combined with the miRNA over-expression/inhibition method, this technique can infer if the direct binding causes the gene expression change of the target genes.

In summary, experimental methods are indispensable tools to find and confirm targets of miRNA regulation. The ability of the CLIP-based method to discover direct bindings between miRNAs and their targets also facilitates our understanding of miRNA regulation mechanisms. However, performing experimental procedures is costly and requires specialized wet-lab skills and knowledge. Computational methods have shown to be an invaluable alternative to experimental methods when we want to quickly generate many candidate miRNA-target interactions for various research purposes [Watanabe et al., 2007, Ekimler and Sahin, 2014].

A.2 Computational methods

While all existing computational methods to predict miRNA-target interactions take sequences of miRNAs and their candidate target genes as input, they are different in the way they convert the sequence information into different features to be used for interaction predictions. The common features include seed match, conservation, free energy, and site accessibility [Peterson et al., 2014].

Seed match: As mentioned above, the seed region of an miRNA contains nu-

cleotides from position 2 to 8 from 5' to 3' ends. A seed region composed of *seed matches* between the miRNA and an area of its target transcript (i.e., miRNA-response-element (MRE) on 3'UTR of mRNAs). A seed matches is a complementary base-pair: adenosine (A) matches with uracil (U) or guanine (G) matches with cytosine (C).

Some algorithms and tools such as TargetScan [Agarwal et al., 2015], PITA [Kertesz et al., 2007], and RNAhybrid [Krüger and Rehmsmeier, 2006] require the miRNAs to have perfect seed matches with their targeted genes. A perfect seed match includes 6mer seeds, which are perfect matches from nucleotides 2–7 of the miRNA to a region in the targetted gene's 3'UTR region. As imperfect seed matches are prevalent in animals, some algorithms such as MiRanda [Enright et al., 2003] allow some exceptions of miRNA-target sequence matching in addition to perfect seed matching. MiRanda allows *GU wobble* in the seed match, which refers to G pairing with U instead of C. Also unlike other miRNA-target prediction tools that only focus on miRNA seed region, MiRanda considers potential matching between entire miRNA and target transcript bodies, but weigh the miRNA seed region matching more heavily.

Free Energy: *Free energy* (or Gibbs free energy) in the miRNA-target interaction context indicates the thermodynamic stability of a miRNA-target binding. If a miRNA is predicted to have a stable thermodynamic binding with its target,

the miRNA-target interaction is also more likely to be true [Watanabe et al., 2007]. The Vienna RNA package [Hofacker et al., 1994] is a popular tool for estimating the minimum energy (kcal/mol) needed to make the binding possible. It uses dynamic programming to compute the thermodynamic stability between a miRNA and its predicted target based on sequence complementary between the miRNA and the target. ΔG refers to the change in the thermodynamic stability before and after the binding. More negative ΔG indicates that the system has less energy available to react in the future, resulting in a more stable system and increasing the possibility of the binding.

Site Accessibility: *Site accessibility* measures how easily a miRNA can locate its mRNA target. It has been shown that the secondary structure of a mRNA can interfere with the ability of its miRNA regulator to bind to mRNAs' MREs [Ekimler and Sahin, 2014]. To ease the interference, the miRNA to first binds to a short region of the mRNA, which enables the mRNA secondary structure to be unfolded to complete the binding. The site accessibility feature is available in tools such as miRanda [Enright et al., 2003], DIANA-microT-CDS [Paraskevopoulou et al., 2013], TargetMiner [Bandyopadhyay and Mitra, 2009], PITA [Kertesz et al., 2007], and RNAhybrid [Krüger and Rehmsmeier, 2006]

APPENDIX B

VALIDATION OF THE INFERRED CERNA NETWORKS USING LINCS-L1000 DATASET

This appendix describes how we employed LINCS-L1000 to assess the accuracy of the inferred ceRNA interactions in predicting gene expression change. In the LINCS-L1000 shRNA-perturbation database, each gene knockdown experiment involved using a specific shRNA to target and thereby silenced a gene [Liu et al., 2015a]. The shRNAs were designed to target and silence their predetermined targets (i.e., to avoid off-target matching and unwanted miRNA effects). For each experiment, expression of 978 landmark genes were profiled before and after the gene knockdown. Thus, in response to a gene knockdown experiment, for each of the 978 genes, its expression fold change (EFC) and p-value from differential expression analysis were reported. As mentioned in Chapter 2, we refer to the targeted/knocked-down genes as upstream genes and to the 978 genes as downstream genes.

We employed data from LINCS-L1000 shRNA-perturbation performed on the breast cancer cell MCF7. In the MCF7 data set, expression changes of the 978 downstream genes were recorded at two different time points (96h and 144h). Thus, our analysis was specific for each time point. One upstream gene could be silenced by multiple shRNAs (on average by 3 shRNAs). Consequently, a downstream gene would have multiple EFC records corresponding to the silencing of the upstream

gene. In such cases, we used the downstream gene's EFC average to represent its overall EFC. The number of upstream-downstream pairs in each time point were 2,578,986 pairs (96h) and 1,022,988 pairs (144h). The number of upstream genes in each time point were 2,637 (96h) and 1,046 (144h).

We used LINCS-L1000 (MCF7) shRNA-perturbation data to assess if the inferred ceRNA crosstalks can be used to predict gene expression patterns. We expected that if a downstream gene is an inferred ceRNA, its EFC would be lower in response to the silencing of its upstream ceRNA partners, compared to the silencing of its upstream non-ceRNAs. In other words, for a downstream ceRNA gene, its ratio of expression fold change is expected to be smaller than 1 (see Eq. B.1).

Given the inferred ceRNA crosstalk results, a downstream $ceRNA_i$ in MCF7 has M upstream ceRNA partners and N upstream non-ceRNAs. Let $EFC(ceRNA_i \leftarrow ceRNA_m)$ and $EFC(ceRNA_i \leftarrow RNA_n)$ be the expression fold change of $ceRNA_i$ caused by silencing of its ceRNA partner $ceRNA_m$ and the non-ceRNA RNA_n, respectively. The ratio of expression fold change $RFC(ceRNA_i)$ is

$$RFC(ceRNA_i) = \frac{\frac{1}{M}\sum_{m=1}^{M} EFC(ceRNA_i \leftarrow ceRNA_m)}{\frac{1}{N}\sum_{n=1}^{N} EFC(ceRNA_i \leftarrow RNA_n)}. \tag{B.1}$$

Lower RFC indicates better prediction of gene expression change due to inferred ceRNA crosstalks.

In the LINCS-L1000 (MCF7) dataset, a subset of all upstream genes and a subset of all downstream genes were also inferred ceRNAs. A downstream ceRNA was selected for this analysis if it had at least one upstream ceRNA in the MCF7 dataset. As the Cancerin method only selected ceRNA crosstalk out of all possible pairs between DE mRNAs, we only kept the upstream genes that were also DE mRNAs in the TCGA-BRCA dataset.

APPENDIX C

ACRONYMS

AGO	Argonaute (proteins)
CNA	Copy Number Alteration
CPM	Count Per Million
DM	DNA Methylation
ENCODE	Encyclopedia of DNA Elements
EFC	Expression Fold Change
FGMD	Functional Gene Module Detection
GC	Gene Cluster
HC	Hierarchical Clustering
LASSO	Least Absolute Shrinkage and Selection Operator
mRNA	Messenger RNA
miRNA	Micro RNA
MRE	MiRNA Response Element
MTM	MiRNA Target Module
MTT	MiRNA-Transcription factor-Target (network)
PC	Partial Correlation
RC	Regulator Cluster
RFC	Ratio of Fold Change
RISC	RNA-Induced Silencing Complex
RPKM	Reads Per Kilobase Million
SC	Sensitivity Correlation
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
UTR	Untranslated Region