DOCTORAL THESIS

From genes to transcripts: Integrative modeling and analysis of regulatory networks

A dissertation submitted towards the degree Doctor of Engineering Science of Faculty of Mathematics and Computer Science of Saarland University

by

Azim Dehghani Amirabad

Saarbrücken, 2019

Defense

Day of Colloquium: 11.06.2019
Dean of the Faculty: Prof. Dr. Sebastian Hack

Examination Committee

Chair: Prof. Volkard Helms
Reviewer, Advisor: Prof. Dr. Marcel H. Schulz
Reviewer: Prof. Dr. Andreas Keller
Academic Assistant: Dr. Peter Ebert

Abstract

Although all the cells in an organism posses the same genome, the regulatory mechanisms lead to highly specific cell types. Elucidating these regulatory mechanisms is a great challenge in systems biology research. Nonetheless, it is known that a large fraction of our genome is comprised of regulatory elements, the precise mechanisms by which different combinations of regulatory elements are involved in controlling gene expression and cell identity are poorly understood.

This thesis describes algorithms and approaches for modeling and analysis of different modes of gene regulation. We present POSTIT a novel algorithm for modeling and inferring transcript isoform regulation from transcriptomics and epigenomics data. POSTIT uses multi-task learning with structured-sparsity inducing regularizer to share the regulatory information between isoforms of a gene, which is shown to lead to accurate isoform expression prediction and inference of regulators. Furthermore, it can use isoform expression level and annotation as informative priors for gene expression prediction. Hence, it constitute a novel accurate approach applicable to gene or transcript isoform centric analysis using expression data. In an application to microRNA (miRNA) target prioritization, we demonstrate that it out-competes classical gene centric methods. Moreover, pinpoints important transcription factors and miRNAs that regulate differentially expressed isoforms in any biological system.

Competing endogenous RNA (ceRNA) interactions mediated by miRNAs were postulated as an important cellular regulatory network, in which cross-talk between different transcripts involves competition for joint regulators. We developed a novel statistical method, called SPONGE, for large-scale inference of ceRNA networks. In this framework, we designed an efficient empirical p-value computation approach, by sampling from derived null models, which addresses important confounding factors such as sample size, number of involved regulators and strength of correlation. In an application to a large pan-cancer dataset with 31 cancers we discovered protein-coding and non-coding RNAs that are generic ceRNAs in cancer.

Finally, we present an integrative analysis of miRNA and protein-based post-transcriptional regulation. We postulate a competitive regulation of the RNA-binding protein IMP2 with miRNAs binding the same RNAs using expression and RNA binding data. This function of IMP2 is relevant in the contribution to disease in the context of adult cellular metabolism. As a summary, in this thesis we have presented a number of different novel approaches for inference and the integrative analysis of regulatory networks that we believe will find wide applicability in the biological sciences.

Kurzfassung

Obwohl jede Zelle eines Organismus das gleiche Genom beinhaltet, führen regulatorische Mechanismen zu hoch spezialisierten Zelltypen. Die Aufdeckung dieser regulatorischen Mechanismen ist eine große Herausforderung für Forschung in der Systembiologie. Obwohl ein Großteil unseres Genoms aus regulatorischen Elementen besteht, sind die genauen Mechanismen, durch die verschiedene regulatorische Elemente miteinander kombiniert Genexpression und Zellidentität bestimmen, bislang noch unbekannt.

Diese Arbeit beschreibt Algorithmen und Ansätze zur Modellierung und Analyse verschiedener Aspekte der Genregulation. Wir stellen POSTIT als einen neuen Algorithmus zur Modellierung und Vorhersage der Regulierung von Transkript-Isoformen basierend auf Transcriptomics- und Epigenomics-Daten vor. POSTIT verwendet multi-task learning mit Regularisierung auf Grundlage der Transkriptannotation, um Information zwischen Isoformen des gleichen Gens zu teilen. Dies führt zur besseren Vorhersage der Isoformexpression, sowie der Detektierung von Regulatoren. Außerdem können Isoform-expression und -annotation als informativer Prior zur Genexpressionsvorhersage verwendet werden. Daher stellt es einen neuen, genaueren Ansatz dar, der zur Gen- oder Transkriptisoform-abhängigen Analyse von Genexpressiondaten verwendbar ist. In einer Anwendung auf miRNA target prioritization zeigen wir, dass es die Genauigkeit der Vorhersage durch das klassische genabhängige Modell übertrifft. Zusätzlich werden wichtige Transkriptionsfaktoren und miRNAs, die differentiell exprimierte Genisoformen in Krebs regulieren, erkannt.

Competing endogeneous RNA (ceRNA) Interaktionen, die durch miRNAs mediiert werden, bilden ein wichtiges regulatorischen Netzwerk, in welchem ein Wettkampf um die gemeinsame Regulation verschiedener Transkripte entsteht. Dafür entwickelten wir eine neue statistische Methode, SPONGE, zur Vorhersage großer ceRNA Netzwerke. Im Zuge dessen leiteten wir einen Ansatz ab, der empirische p-Werte durch mehrmaliges Ziehen von Stichproben aus einem abgeleiteten Nullmodell efizient berechnen kann. Dadurch können wichtige fehlleitende Faktoren, wie Stichprobengröße, Anzahl der Regulatoren oder Stärke der Korrelation adressiert werden. Es handelt sich hierbei um einen einheitlichen, schnellen Ansatz, der Koregulation verschiedener miRNAs betrachtet. Durch die Anwendung auf einen Krebsdatensatz mit 31 Krebstypen detektierten wir proteinkodierende und nicht-proteinkodierende RNAs, die generische ceRNAs in Krebs darstellen.

Abschließend präsentieren wir eine integrative Analyse von miRNAs und protein-basierter posttranslationaler Regulation. Wir postulieren eine kompetitive Regulation des RNA-bindenden Proteins IMP2 mit miRNAs, die die selbe RNA binden. Diese Funktion von IMP2 ist insbesondere für die Beteiligung an Krankheiten des adulten zellulären Metabolismus relevant. Zusammenfassend präsentieren wir hier verschiedene neuartige Ansätze zur integrativen Analyse regulatorischer Netzwerke, welche aus unserer Sicht in vielen Bereichen der Biologie Anwendung finden werden.

Acknowledgements

This work is carried out during the years 2015-2019 at the Max Planck Institute for Informatics. These years were a period of scientific, intellectual and personal growth. It is for the love for learning, curiosity, and constant changes of the ideas.

First and foremost, I would like to express my deep gratitude to Marcel H. Schulz for giving me the opportunity to join his group. Marcel gave me the freedom to independently figure out and pursue the topics I was interested in. He has been a source of inspiration, encouragement and advice. Thank you for everything, Marcel.

I would like to extend my special thanks to Andreas Keller who agreed to review this thesis on short notice.

I am also deeply grateful to Ernest Fraenkel whose lab I had the privilege of joining for a research internship. The time spent there has been highly instructive. Thanks for providing such a stimulating and rich environment, discussions and advice. I would like to offer special thanks to Thomas Lengauer, who attracted me to the world of statistical learning and guided me the ways to follow my passion. Thomas, you are a role model for enthusiasm and scientific curiosity.

I have been fortunate to work closely with many outstanding collaborators. Thank you to Martin Simon, Sonja M. Kessler, Dennis Kostka, Markus List, and Alexandra K. Kiemer.

I would like to thank the International Max Planck Research School for Computer Science, Cluster of Excellence, and Graduate School for Computer Science for funding my PhD., conference trips, and my stay at Boston.

I am grateful that I had an opportunity to work with two students, Eva Paul and Tobias Heinen, on different aspects of my main endeavor on gene regulation modeling. My friends and colleagues at Max Planck and Saarland University provided an enjoyable working environment. I am especially grateful for Ali Ghaffaari, Azin Ghazimatin, Michael, Lisa, Nora, Prabhav, and Matthias, Mateusz, Max Maria Losch, and Jan Hendrik. Special thanks to Michael, Lisa, Dilip, Max, and Barbara Koscielecka for proofreading this thesis. Your valuable comments helped me to shape this thesis.

Finally, I would like to express my sincere gratitude to my family, especially my parents for their love and support. I love you and I dedicate this thesis to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The twenty-first century is an exciting time for molecular biology. The DNA sequencing of the human genome and many other model organisms in early 2000 generated renewed hopes for novel therapeutics and treatments, but also left many fundamental questions unanswered. For example, what are the regulatory interactions that determine cell identity? Given that different cells have the same DNA sequence, how do they traverse different developmental trajectories with intricate spatial and temporal precision? How does the cell decide when to respond to an external stimulus? What are the changes in regulatory layers in different diseases prevent them from being appropriately decoded and executed?

Our body comprises a huge atlas of cells with different functions and phenotypes. All the cells possess the same set of genes, and how they are used determines cell fate. Orchestrating combinatorial usage of the genes at the precise time and location demands a vast amount of regulatory options. Hence, it is not surprising large fraction of our genome consists of regulatory regions. International consortia like the Encyclopedia of DNA Elements (ENCODE) consortium have assigned regulatory functions for around 80% of our genome (Roy et al., 2010).

1.1 Thesis scope

Learning new biological insights on gene regulation mechanisms from large-scale high-throughput data set is a grand challenge in systems biology. In this thesis, we focus on developing methods for modeling and analysis of different modes of gene regulation. We develop and apply methods for transcriptional and post-transcriptional regulatory network inference from epigenomics, transcriptomics data sets and demonstrate their value for gene regulation and gene expression prediction. We formulate the network inference problem in a supervised learning framework to predict the regulatory edges by integrating chromatin accessibility, transcription factor binding, sequence motifs, miRNA, and transcript isoform expression profiles.

1.1.1 Contributions

The key contributions of this thesis are as follows:

- *POSTIT*- a novel computational method for modeling and inferring transcript isoform regulatory networks from transcriptomics and epigenomics data. This method is based on multi-task learning, which encodes isoform annotation for inferring isoform regulatory networks. Moreover, it implements a scalable proximal gradient descent algorithm to solve the multi-task regression formulation for the entire human transcriptome.

Contributions: All method development and data analysis of this project was done by me, supervised by Marcel Schulz.

- *SPONGE*- a novel method for large-scale inference of competing endogeneous RNA networks. SPONGE uses *multiple sensitivity correlation*, a newly defined measure for which we can estimate a distribution under a null hypothesis. SPONGE can accurately quantify the contribution of multiple miRNAs to a ceRNA interaction with a probabilistic model that addresses previously neglected confounding factors and allows fast p-value calculation.
  Contributions: I had a major contribution to the design and development of the algorithm. I implemented the miRNA target prediction module, made all the derivations for the null model simulation, implemented the first working version of the algorithm and designed the simulation experiments. Furthermore, I contributed to data analysis, interpretation and writing the manuscript. Dennis Kostka suggested the idea of using the Schur complement for covariance matrix simulation. Moreover, he helped to derive the null model simulation for the case $microRNA = 1$. Markus List led the data analysis, wrote the Bioconductor package, and contributed to writing the manuscript. Marcel Schulz supervised the project and contributed to writing the manuscript.

- We present an integrative analysis of miRNA and protein-based post-transcriptional regulation. We postulate a competitive regulation of the RNA-binding protein IMP2 with miRNAs binding the same RNAs using expression and RNA binding data. This function of IMP2 is relevant in the contribution to disease in the context of adult cellular metabolism.
  Contributions: I did the miRNA and IMP2 target prediction and the integrative analysis of IMP2 and miRNA data. Moreover, I contributed to data analysis, interpretation and writing of the manuscript. Finally, I supervised Pathmanaban Ramasamy for miRNA expression quantification and differential expression analysis. Marina Wierz prepared all sequence libraries and was involved in data interpretation. Karl Nordström was involved in primary data generation. Sonja Kessler prepared the mice and was invovled in data interpretation and paper writing. Martin Simon and Marcel Schulz were involved in data analysis and paper writing.

1.1.2 Thesis outline

The remaining chapters of this thesis are summarized as follows:

- **Chapter 2:** Establishes the biological background of gene regulation. It provides a general overview of the different levels of gene regulation. Moreover, it introduces technologies used to profile the transcriptome, chromatin accessibility, and genome-wide binding occupancy of regulatory proteins.

- **Chapter 3:** Establishes the computational background on the methods that are used in this thesis. The first part provides a primer on convex optimization and describes a method for solving non-smooth convex functions. The second part provides a theoretical background on supervised learning and different types of sparsity-inducing norms. Finally, we review methods for modeling gene expression and regulation.

- **Chapter 4:** We present an integrative analysis of miRNA and protein-based post-transcriptional regulation. We show that overexpression of the RNA-binding protein IMP2 alters the regulatory capacity of miRNAs by competition for binding sites.

- **Chapter 5:** Presents a new paradigm for modeling and inference of gene regulation by incorporating transcript isoforms in a multi-task regression formulation. Moreover, it describes an efficient optimization framework for solving a multi-task learning approach for the entire human transcriptome. Finally, we discuss the biological insights that we learned from analyzing the transcript isoform regulatory network in cancer.

- **Chapter 6:** Presents a scalable algorithm for inferring large ceRNA regulatory networks. A statistical framework for significance analysis of multiple sensitivity correlation is described, in which confounding factors are addressed. Finally, a pan-cancer analyis of ceRNAs is presented as a use-case for the developed Bioconductor package. Chapter 6 will appear at processing of ISMB-ECCB 2019 conference.

- **Chapter 7:** In chapter 7, we present directions for future studies.

Chapter 2

Biological background

This chapter introduces the basic aspects of the gene regulation.

2.1 Chromatin organization

The human genome consists of 3×10^9 bases, organized into 23 chromosomes, accounting for a total length of 2 m DNA (Allis et al., 2007). Hence, it needs to be compressed about 10^4-folds to fit into a cell's nucleus (Figure 2.1). Histone proteins provide a smart solution to this packing problem. The DNA wrapped around histone proteins gives rise to a flexible polymer known as *chromatin*. Chromatin has a compact organization, where most of the genetic material is inaccessible, and, hence functionally inactive.

Histone tails are extensively post-translationally modified. These modifications often correlate with the functional activity of the specific genomics region. The combination of different histone modifications in a specific genomic region defines *chromatin state*. Chromatin state can be interpreated as the genome's indexing system. It serves as an efficient indexing platform, which it facilitates dynamic accessibility of a specific genomic region in a particular time and location (Allis et al., 2007). Chromatin does not have a uniform structure; it comes in different flavours from a less condensed region, *euchromatin*, to a more condensed one called *heterochromatin*. The irreducible subunit of the chromatin is the *nucleosome*, which contains about 200 base pairs of DNA, organized by an octomer of histone proteins. Regulatory codes (histone modifications) on the histone tails are involved in signaling the opening and compaction of the DNA. Moreover, the histone code cues the recruitment of cellular machinery to read and execute the genome (Allis et al., 2007).

2.2 Basic principles of gene regulation

Gene regulation refers to the mechanisms contributing to gene expression control. Regulation of gene expression is a complex process, and a diverse set of regulatory elements directs the regulatory controls at different levels. These regulatory levels include: (1) When and where a gene is transcribed, (2) How RNA is spliced or processed, (3) Which RNAs are transported to cytoplasm, (4) Which RNAs are stabilized or degraded by post-transcriptional regulators, (5) Which mRNAs are translated into protein, and (6) How proteins are post-translationally modified.

First, we describe the basic elements involved in gene regulation, then briefly review its mechanisms.

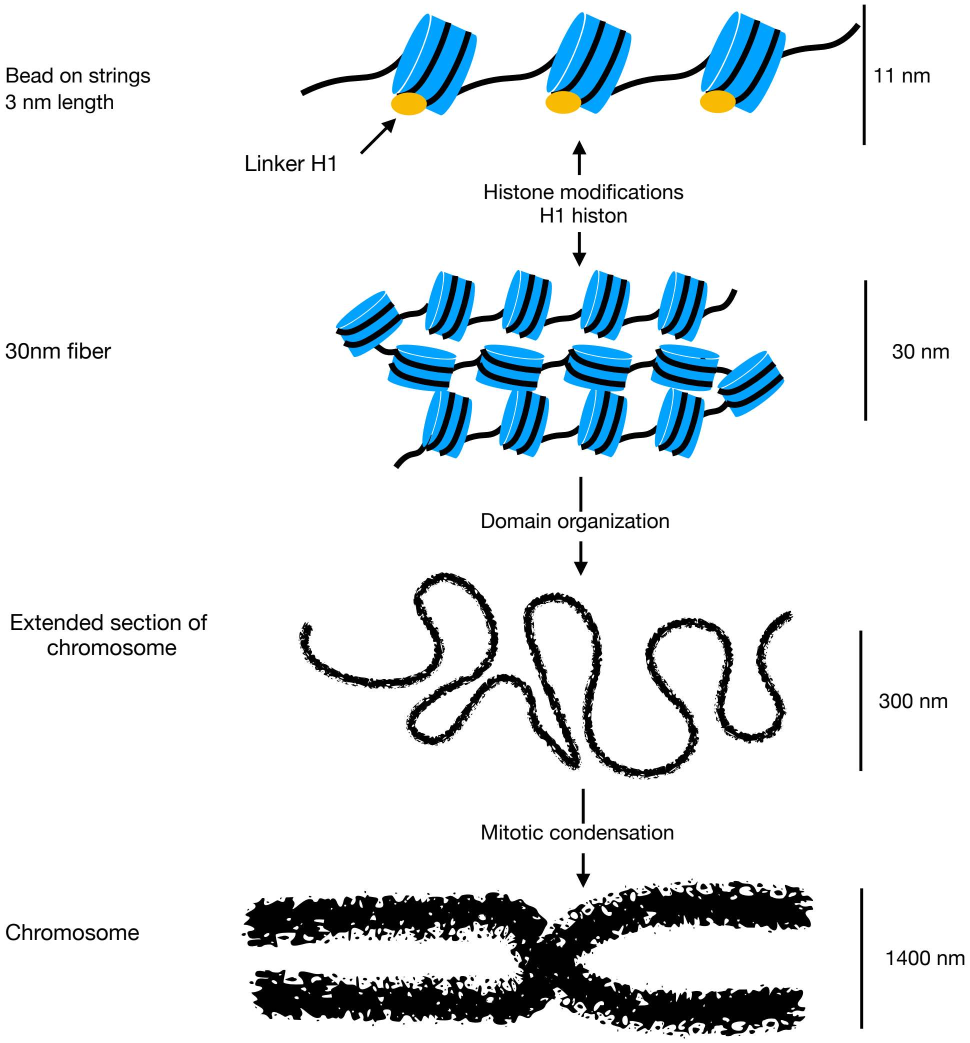


FIGURE 2.1: Different level of DNA compaction in the genome. 1) The nucleosome is the first level of compaction, which decreases the size sixfold over naked DNA. 2) The second level of compaction is the interaction between different nucleosomes, leading to a condensed nanofiber of chromatin, with approximately 30 nm in diameter. Linker histones stabilize the interaction between nucleosomes. The final level of compaction requires the folding of the chromatin fiber into 3D structures, which leads to 1,000-fold linear compaction in euchromatin. This structure is exchangeable with packing into mitotic chromosomes, which results into 10,0000-folds compaction. This figure is based on (Allis et al., 2007)

2.2.1 Gene regulatory elements

Regulatory elements are sequence motifs, which are involved in regulating gene expression. They serve as a landing pad for different regulators. They are usually 6-8 nucleotide (nt) and, sometimes, are degenerate. Motifs are scattered all around the genome. A collection of motifs constructs the regulatory element; combinatorial usage of regulatory motifs encodes specific regulatory functions for regulatory elements. These elements are also overlaid with epigenomic marks, which increase the regulatory capacity of the regulatory elements.

There are different types of gene regulatory elements including promoter, enhancer, silencer, insulator (Chatterjee and Ahituv, 2017).

Promoters are regulatory elements that are located close to the transcription start site (TSS) of a gene and act as landing pad for assembling the core trancriptional machinery.

Enhancers are so-called "promoters of the promoter" controlling when, where, and to what level expresses the gene. Enhancers can be located in the *cis*, or *trans* of the target gene. Multiple enhancers can target a single promoter, and a single enhancer can target many promoters. Enhancers are activated by the binding of sequence specific transcription factors (TFs) and co-activators. Active enhancers are characterized by p300 protein occupancy, DNaseI hypersensitivity, and H3K4me1, H3K4me2, and H3K27ac histone marks (Ong and Corces, 2011). Each enhancer can have a different sequence, and hence, different activities across different cell types. Enhancers specify tissue-specific gene expression.

Insulators are modular short sequence sets that act as a boundaries to define a region that most regulatory elements reside. They are generally marked by CTCF occupancy and function as barriers for chromatin interactions, hence preventing the propagation of epigenomics marks from one genomic region to another (Ong and Corces, 2014; Phillips and Corces, 2009).

Silencers act as negative regulators of the target promoters and suppress the target gene activity (Lanzuolo et al., 2007). The precise regulatory mechanisms of silencers have not yet beeen understood. It is hypothesized that they interact with the target promoter and establish repressive chromatin marks (Harris, Mostecki, and Rothman, 2005). Moreover, they may compete with the core promoter to sequester TFs (Li et al., 2004).

The interplay between different regulatory elements at a specific time and location creates complex regulatory circuitry that modulates gene expression at the transcriptional and post-transcriptional levels. It is not surprising that the precise understanding of regulatory circuitry is still limited.

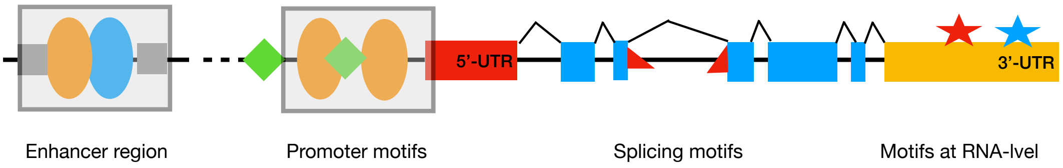


FIGURE 2.2: Different types of regulatory elements. They can be thousands of base pair away from the target gene like enhancer, or can reside within or in the proximity of the gene like proximal promoters, splicing codes, and 3'-UTR.

2.2.2 Regulatory factors

Gene expression level can be modulated by the regulatory molecules, which directly bind to a regulatory element at the DNA or RNA level. There are three main classes of regulatory molecules: TFs, RNA-binding proteins (RBPs), and regulatory RNAs.

Transcription factors

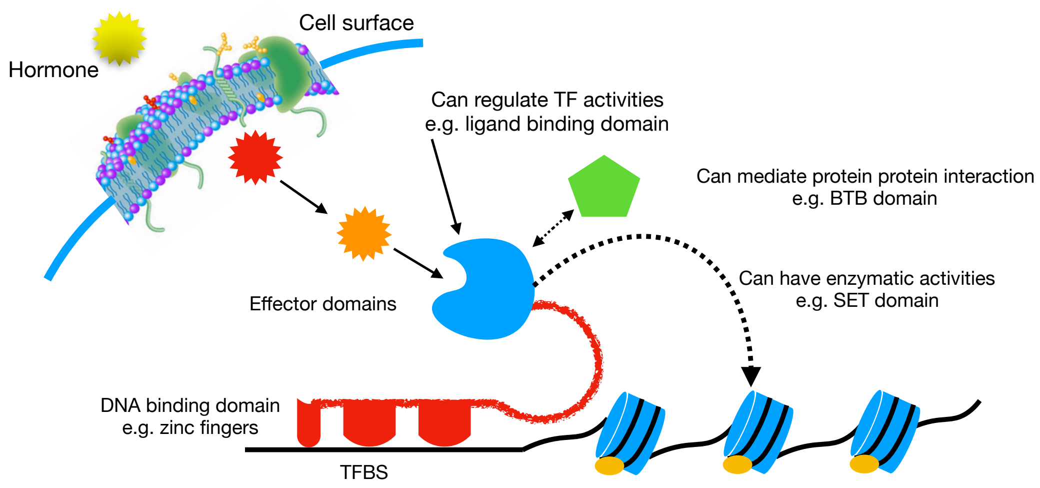

Hormone
Cell surface
Can regulate TF activities
e.g. ligand binding domain
Can mediate protein protein interaction
e.g. BTB domain
Can have enzymatic activities
e.g. SET domain
Effector domains
DNA binding domain
e.g. zinc fingers
TFBS

FIGURE 2.3: A typical prototype of TFs. TFs have a very modular structure. They have DNA binding doamins that interact with DNA in a sequence-specific way with one of the grooves of the DNA. They also typically contain domains, which have repression or activation function, and interact with other proteins. This figure is based on (Lambert et al., 2018).

Transcription factors are defined as DNA-binding proteins capable of binding to DNA in a sequence-specific manner, and modulating the chromatin and transcription (Figure 2.3). They are the working horses of the cell reading genome. TFs have modular structures and bind different motifs comprising 5-12 nt of DNA with different degrees of specificity (Spitz and Furlong, 2012). Many TFs recognize similar binding motifs, which leads to the definition of TF families. TFs selectively read a subset of regulatory elements and exert control over processes that specify cell types and developmental patterns (Lee and Young, 2013).

The functions of TFs are determined based on the target genes. From a functional perspective, TFs can be categorized into four distinct classes (Pope and Medzhitov, 2018).

- **Class-A TFs** target house keeping genes. Hence, these TFs are expressed in most of the cell types. Examples of this class include SP1, YY1, and NRF-1.
- **Class-B TFs** are broadly expressed, but inactive inside the cell, waiting for specific signals to become active. Examples of these TFs include STATs, SMADs, NF-kB, and p53. Usually, these TFs interact with the proteins involved in signaling pathways, or have allosteric binding sites for specific ligands (e.g., lipids and steroid hormones).
- **Class-C TFs** are inactive inside the cell. Their expression induced by class-B TFs. Examples of class-C TFs include FosB, JunB, and E2F.
- **Class-D TFs** are the lineage-specific TFs. They usually target cell type specific genes, and regulate cell differentiation. They usually bind to the enhancer regions and keep the cell-type-specific genes active. Moreover, they facilitate binding of class-B and class-C TFs to cell-type-specific enhancers.

RNA-binding proteins

RNA-binding proteins (RBP) bind to single or double-strand RNA (dsRNA) through one or multiple RNA-binding domains (RBDs), and modulate the fate or function of the target RNAs (Hentze et al., 2018). Approximately 800 distinct RBDs are known. RBPs have modular structures, and possess various structural motifs, such as dsRNA-binding domain and RNA recognition motifs. Multiple domains enable RBPs to recognize the target RNA sequences with different degrees of specificity and affinity (Lunde, Moore, and Varani, 2007). Recent reports suggest that the human genome contains between 1,072 and 1,540 RBP genes. Binding motifs for RBPs are usually 6-7 nt and very degenerate and, hence, they can target hundreds of mRNA.

RBPs have emerged as a key regulator of gene expression at co-translational and post-transcriptional levels. They can bind to exon, intron, or 3'-UTR sequences. The functional impact of RBP depend on where it binds. Usually, they control the localization, degredation, and translation of mRNA. Moreover, RBPs regulate the spatiotemporal rate of RNA splicing, ployadenlaytion, and stability (Keene, 2007). A recent study has shown that some RBPs have chromatin association properties, and may couple the transcription and co-transcriptional splicing mechanisms (Van Nostrand et al., 2018).

Non-coding RNA

Since the discovery of non-coding RNAs (ncRNA) and their genes, such as rRNA and tRNA, in the 1950s, new classes are constantly being discovered. Systematically focused efforts such as the ENCODE and modeENCOFE projects have revealed that around 70% of the eukaryotic genome is transcribed (Pennisi, 2012; Roy et al., 2010). This includes transcription from centromeres, telomeres, coding and non-coding strands of the gene regions, and the regions between the genes (lincRNA). More surprisingly, promoter (prRNAs) and enhancer (eRNAs) regions are also transcribed.

Different classes of ncRNAs have very similar regulatory functions, and pathways, and employ similar molecular machinery, which makes it difficult to assign them a distinct function and definition. Hence, they are generally are classified according to their size into large (rRNA size), medium (tRNA size), and microRNA sizes.

ncRNAs are used as a powerful system to modulate gene expression level on the transcriptional or post-transcriptional levels (Morris and Mattick, 2014; Guttman and Rinn, 2012). This regulation can be direct or indirect. The direct regulation occurs by interference with RNA polymerase or by forming an RNA-duplex of antisense RNA with the authentic gene transcript. The indirect regulation happens by changing the local chromatin signature of the gene and even nuclear organization of the gene in the nucleus.

Small RNAs have a length of less than 200 nt and are involved in RNA interference (RNAi) pathways that modulate gene expression via post-transcriptional gene silencing (PTGS) and chromatin-dependent gene silencing (CDGS) (Morris and Mattick, 2014). Although there are many classes of small RNA, based on their origin, structure, and biological role they are classified into three main categories: miRNAs, small interfering RNAs (siRNA) and piRNAs (Farazi, Juranek, and Tuschl, 2008). siRNAs have been primarily studied in plant and fungi. Their biogenesis is involved in processing from double-strand RNA via Dicer enzymes and loading into a silencing complex, which is transported into the cytoplasm. They cleave the target mRNA by complete sequence complementarity, thus have high target specificity. In the mature form, they are around 21-22 nt in length. piRNAs are germ-cell-specific small RNAs named according to their interactions with Piwi clade of Argonaute proteins. They have been associated with the epigenetic and post-transcriptional silencing of transposons (Siomi et al., 2011).

MicroRNA are small RNAs that are found in almost all eukaryotes. They are processed from longer transcripts, pre-RNA, that fold back and form hairpin structures. miRNAs play a central role in gene regulation, both at transcriptional and post-transcriptional levels. It has been shown that miRNAs regulate around 90% of genes. The human genome encodes around 1500 miRNAs. The primary transcript comes in three different forms: i) Only one hairpin made from a gene, ii) three distinct hairpins from polycistronic mRNA or iii) from the intronic region of the gene after splicing of pre-mRNA. Figure 2.4 shows the general miRNA biogenesis pathway. After the miRNA duplex is produced in the cytoplasm, one of the strands is loaded into the silencing complex. Then miRNA guides the silencing complex to the target transcript. Many microRNA are conserved across different species. Moreover, our genome contains hundreds of non-conserved microRNA expressed at a low level compared to conserved miRNAs. microRNAs regulate gene expression both at the transcriptional and post-transcriptional levels (Morris and Mattick, 2014). It has been reported that miRNAs affect transcription initiation by binding to the gene promoters (Place et al., 2008; Zhang et al., 2014), or promote heterochromatin formation.

Lewis et al. (Lewis, Burge, and Bartel, 2005) have reported that the evolutionary conservation of the miRNA seed target sequence is a strong indicator of miRNA target site functionality, and that methods that consider evolutionary selection of the target sites demonstrate better performance. Another factor that contributes to the functionality of the miRNA target site is the location along the mRNA. Generally, miRNA target sites at the beginning and end of the 3′-UTR are under strong evolutionary selection pressure, and the sequence around these target sites is optimized during evolution to enable efficient targeting by miRNAs. Moreover, miRNA binding sites in the coding region of the transcript have smaller effects on mRNA stability than those in the 3′-UTR (Gaidatzis et al., 2007). Recently, Meijer et al. (Meijer et al., 2013) have suggested that the distribution of the miRNA binding sites across the transcriptome is not random, and interaction between different regulatory subunits might dictate additional constraints on the spatial positioning of the miRNA target sites.

Long Noncoding RNAs (lncRNAs) are usually longer than 200 nt, and are classified according to their genomic contexts, i.e., from where in the genome these RNAs are transcribed. This classification includes stand-alone lncRNA which are distinct transcription units, antisense transcripts contrary to the sense DNA strands of annotated transcription units, pseudogenes, long intronic ncRNAs, promoter-associated transcripts, and enhancer RNAs. They add up to 9,200 lncRNAs, the gene for some

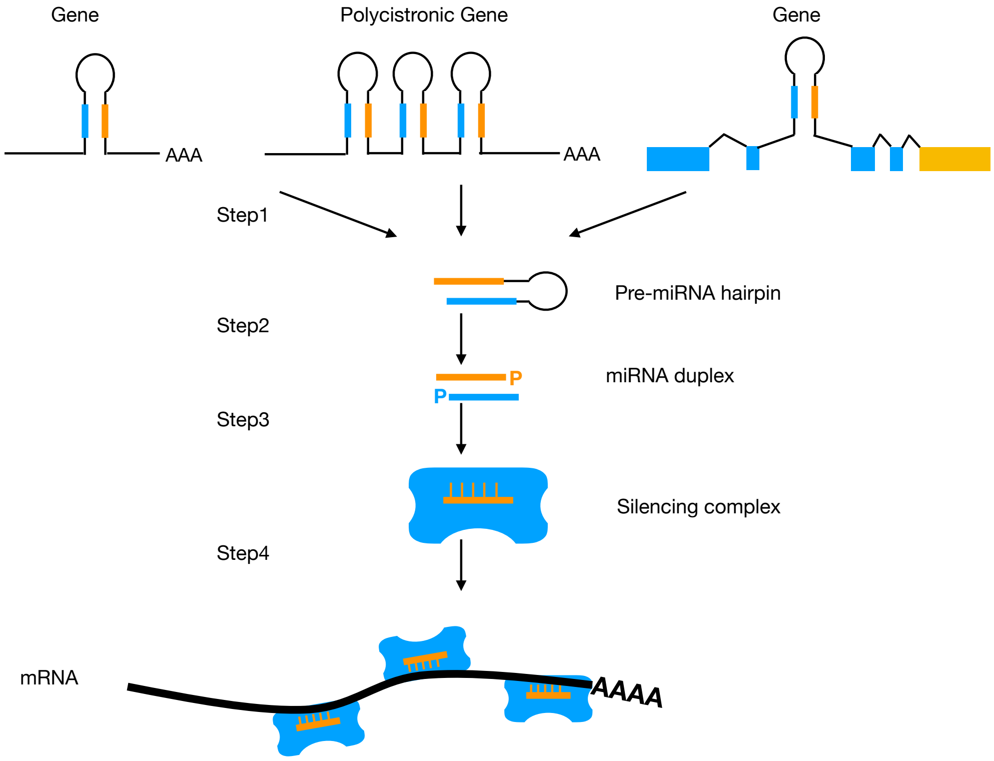


FIGURE 2.4: miRNA biogenesis. miRNA is transcribed by POL II. It is recognized by Pasha and Drosha enzymes at the nucleus. Drosha cleaves the primary transcript by one helical turn far from the hairpin. The pre-miRNA further processed by the Dicer enzyme at the cytoplasm and removes the loop, giving rise to the miRNA-duplex. Finally, one of the strands of the duplex is loaded into Argonaute protein and makes the silencing complex.

spans several kilobases and contain introns (Derrien et al., 2012). From the functional perspective, lncRNAs have been involved in an extraordinary number of regulatory processes. They have been shown to influence gene regulation at a different level. They can interact with epigenetic modifiers like DNMTs, PRC2, and trithorax complexes. They also act as a scaffold to recruit protein complexes to facilitate coordination of multiple layers of chromatin modifications. A prominent example the crucial role of lncRNA in development includes the XIST lncRNA. Xist coats inactive X-chromosome in females and acts as a major effector of the X inactivation process by recruiting an epigenetic silencing complex (Kung, Colognori, and Lee, 2013). In addition to the regulatory function of lncRNA at the epigenomics level, they can affect transcription as well. Some act as a sponge for TFs and change their cellular localization. Others may affect RNA processing, through interaction with splicing factors. Moreover, depending on the shared microRNA recognition elements between lncRNAs, they can compete with each other to sequester shared miRNAs and hence protect other transcripts from microRNA regulation (Kung, Colognori, and Lee, 2013).

2.2.3 Transcriptional gene regulation

Transcription is the process by which a particular segment of DNA is translated into RNA. The RNA polymerase (RNAP) enzyme binds to the accessible core promoter, opens up the double-strand DNA, and, while sliding along the DNA, synthesizes an RNA copy of the gene. To initiate transcription in eukaryotes, hundreds of proteins and subunits need to act in concert to assemble the transcriptional machinery at core promoters (Figure 2.5). The rate of recruitment of RNAP to specific genomic regions is modulated by TFs that bind to enhancer regions. These TFs recruit a series of general transcription factors (GTFs). GTFs enable RNAP to recognize the core promoter and form the transcription pre-initiation complex (PIC). After that, transcription beings.

The basic transcriptional machinery is similar in higher eukaryotes. Regulation at the transcriptional level is the most important step, because this ensures no other unnecessary intermediate products are synthesized. For this reason, many levels of regulatory control exist. Many of these regulatory control are mediated by *epigenetics*, which are the heritable changes in the genome not directly encoded in the DNA sequence.

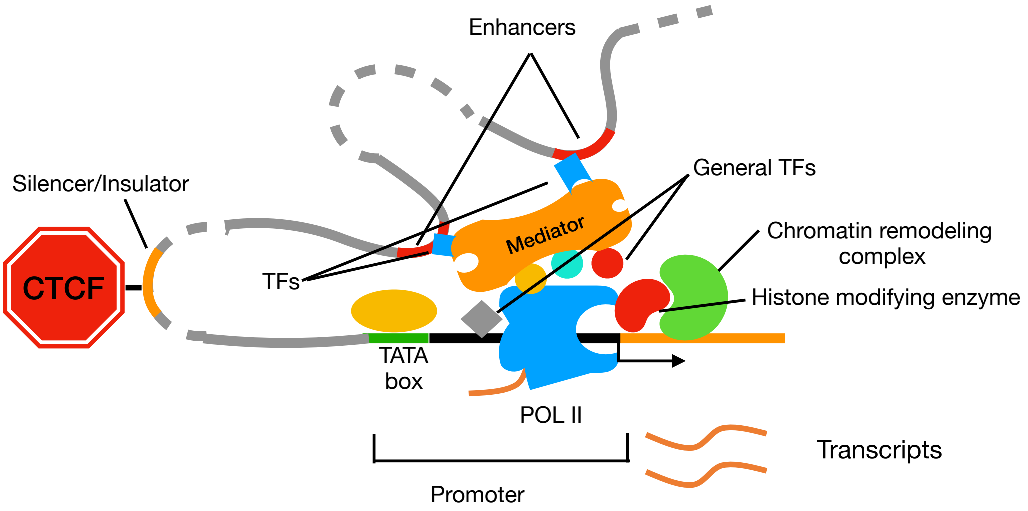


FIGURE 2.5: Transcriptional gene regulation. Transcriptional regulatory elements work together to control the expression of a eukaryotic gene. The core promoter consists of regulatory elements that provide landing pads for the general transcription factor. These TFs helps RNA POL II to recognize the promoter and initiate transcription.

Initiation of the transcription in eukaryotic cells depends on the chromatin structure. DNA is packed into nucleosome; and the core promoter sequence might not be accessible to the transcriptional machinery. For a gene to be transcribed it should have an active structure-, in other words, active promoters are associated with nucleosome-depleted regions (NDRs). The *Epigenetic state* controls the active or inactive structure of a gene. Hence, the expression of a gene is linked to the structure of the chromatin both locally (at the promoter) and in the surrounding regions (Krebs, Goldstein, and Kilpatrick, 2017). Acquisition of NDRs is the first step in transcription initiation. This raises an important question regarding the transcription initiation. Given that DNA is wrapped around nucleosomes, how are individual genes in a condensed chromatin region recognized and targeted for activation? It has been shown that some TFs can bind to their DNA target sequence in closed chromatin (Spitz and Furlong, 2012). These activator proteins recruit chromatin complexes and histone modifiers to start removing or sliding nucleosomes and clearing the promoter, which makes a gene potentially ready to be transcribed (Krebs, Goldstein, and Kilpatrick, 2017; Spitz and Furlong, 2012).

Epigenetic codes mark the regulatory region of the genome. Active promoters are usually characterized by the H3K4me3 histone mark and a low level of DNA methylation. Moreover, DNA methylation and the H3K36me3 mark in the gene body are associated with transcription elongation (Allis et al., 2007).

The basal transcriptional apparatus has very low efficiency and, in most cases, the activity of a promoter is increased by the presence of *distal regulatory elements*, mainly *enhancers*. Enhancers have a variable distance to the core promoter, some are located hundreds of kilobases from the core promoter, while others may lie quite close to it. The proximal enhancer might be located upstream, downstream, or in the intronic regions of the target gene. Active enhancers are characterized by p300 protein occupancy, DNaseI hypersensitivity, and H3K4me1, H3K4me2, and H3K27ac histone marks (Ong and Corces, 2011).

An enhancer might have binding sites for both activator or repressor transcription factors. Hence, depending on the cellular state at any given time, it will have a mixture of transcription factors that bind to an enhancer. If more repressors bind than activators, then this regulatory element will be a silencer (Krebs, Goldstein, and Kilpatrick, 2017).

Enhancer interact with the basal transcriptional machinery at the core promoter and usually exert their function by increasing the concentration of activator TFs near the core promoter. The mediator complex and cohesin stabilize this interaction. Often, there is no one-to-one relationship between an enhancer and a target gene. An enhancer can target multiple genes, and multiple enhancers can target a single gene in a context-specific manner. It has been shown that an enhancer also is transcribed into enhancer RNA, and is correlated with target gene expression level.

In a broad perspective, dynamic modification of histones and DNA, and its combination with nucleosome positioning, plays a key role in transcriptional gene regulation.

2.2.4 Post-transcriptional gene regulation

Post-transcriptional gene regulation refers to all the processes that start from RNA transcriptions and end at RNA translation and stability. These processes are mainly coordinated by miRNA and RBPs. While most of the processes for differential gene expression are achieved at the transcriptional level, there are still numerous post-transcriptional events that also regulate context-specific expression patterns (Corbett, 2018). The main post-transcriptional processes include splicing, 5′-capping, RNA editing and Polyadenylation.

Splicing is a form of RNA processing, in which introns of a primary transcript are spliced, and exons are joined together. This process is mainly catalyzed by the *spliceosome*, a complex of small nuclear ribonucleoproteins (snRNPs). Introns usually contain splicing regulatory signals. Combinatorial usage of these splicing signals leads to *alternative splicing*. Alternative splicing enables cells to generate more than one isoform from a single gene, hence expanding the post-transcriptional regulation and proteome complexity. Alternative splicing can generate isoforms which can differ in their 3′-UTR sequence, coding sequence, and intron retention pattern (Baralle and Giudice, 2017). These differences contribute to mRNA stability, localization, or translation. Alternative splicing contributes to cell differentiation, identity, development, and tissue-identity acquisition (Wang et al., 2008a). Moreover, mutations in splicing signals is associated with tissue-specific diseases (Scotti and Swanson, 2016)

Polyadenylation is another RNA processing mechanism, in which a 3′ end of nascent RNA is cleaved, followed by synthesis of a **polyA tail** on the 3′ end. The cleavage happens around a well-conserved regulatory sequence called a *polyadenylation site(PAS)*. A transcript can have multiple PASs; differential usage of these sites enables cells to produce distinct mRNA isoforms. This process is called *alternative polyadenylation(APA)*. Most APA sites reside in 3-UTRs. APA is a pervasive mechanism to expand transcriptome diversity and regulation. Recent reports show that around 70% of mammalian mRNA-encoding genes express APA isoforms and frequently accrue in the 3′-UTR (Hoque et al., 2012; Derti et al., 2012). As 3′-UTRs contain regulatory elements that determine the fate of isoforms, 3′-UTR-APA can considerably affect post-transcriptional isoform regulation in different aspects, including regulating stabilization, localization, and export(Tian and Manley, 2017). A 3′-UTR sequence contains a high density of miRNA binding sites. The well-known consequence of 3′-UTR-APA is differential targeting of 3-UTR-APA isoforms by miRNAs (Mayr and Bartel, 2009).

microRNA target reconition The seed region of the miRNAs is the most important factor for a miRNA to recognize its target. Structural studies of argonaut (AGO) with miRNA in complex revealed that the bases 2-6 of the miRNA are exposed for nucleating the interaction between miRNA and the target. Plant miRNAs seed sequences have perfect sequence complementarity with their target site but, in animals, partial sequence complementarity is often sufficient to regulate target genes (Wang et al., 2008b). Grimson et al. (Grimson et al., 2007) have reported that the higher the sequence complementarity of the seed region with the target site, the stronger the response of the target miRNA level to miRNA expression level changes. However, it is still not clear how many mismatches can be tolerated between the miRNA and the target sequence.

Consequence of microRNA target interaction Post-transcriptional gene repression is the most-studied consequence of miRNA target interactions, which leads to a decrease in the encoded protein level. One of the main targets of miRNAs are transcription factors (Hornstein and Shomron, 2006). In gene regulatory networks, miRNAs and transcription factors interact. One such example is the feedforward loop (FFL), where a miRNA and a transcription factor regulate a common target, and transcription of the miRNA is regulated by the same transcription factor. Through this regulatory module, miRNA counteract “leaky” transcription. These kinds of networks are believed to be effective in noise buffering (Hornstein and Shomron, 2006). Re et al. have identified hundreds of such network motifs in humans and hypothesized that one of the main activities of the miRNAs on a subset of their targets is to buffer transcriptional noise via FFLs and increase gene expression level precision (Re et al., 2009). Other consequences of the miRNA target interaction are the competing endogenous RNA (ceRNA) effect. The ceRNA hypothesis suggests that mRNAs with similar binding sites compete for shared miRNAs. Hence, a ceRNA with high abundance sponges the shared miRNA and decrease miRNA ability to suppress other targets (Arvey et al., 2010a). The advance of new experimental methods to measure miRNA and mRNA expression levels in single cells, as well as measuring their interaction affinities and rate of RNA-dependent responses, opens a new horizon to study their dynamic regulation in more details (Hausser and Zavolan, 2014).

2.3 Genomics data and high-throughput technologies

In recent years, significant progress has been made in the development of throughput technologies for studying various aspects of gene regulation. These technologies provide us the opportunity to reveal underlying biological mechanisms that contribute to a specific phenotype. We describe the high-throughput technologies used in the projects presented in this thesis.

2.3.1 Methods for genome-wide measurement of large and small RNA

The ability to measure the RNA expression level of thousands of genes simultaneously revolutionized biological research. RNA-seq is a next-generation sequencing technology that allows sequencing of the entire transcriptome of an organism, producing millions of small sequence reads per sequencing experiment (Torres et al., 2008; Marioni et al., 2008). It has several advantages over microarray technology, including very low background noise, high dynamic range of quantification, sensitivity to the genes with low and high expression levels, accurate estimation of expression levels, and a high level of reproducibility of results in biological and technical replicates (Wang, Gerstein, and Snyder, 2009).

Figure 2.6 shows a typical workflow for an RNA-seq experiment. In RNA-seq, the RNA molecules are first converted to cDNA fragments using RT-PCR. Then, short adaptor molecules are ligated to one or both ends of these cDNA molecules. The molecules are immobilized on the surface of the sequencer cell (in case of Illumina sequencing) and further amplified after immobilization. Each molecule is sequenced in parallel in a high-throughput manner [26]. Depending on the technology used, the sequencing can be done from one (single-end sequencing) or both ends of the cDNA fragments (paired-end sequencing). The length of the reads depends on the sequencing technology used, and is usually between 40-400 bp (Holt and Jones, 2008). In general, RNA-seq provides the opportunity to study the transcriptome of an organism at single base resolution, quantify the expression levels of transcripts, discover new genes and transcripts, and quantitatively examine the alternative splicing diversity of the transcriptome. To achieve these goals, RNA-seq technology faces three bioinformatic challenges: (i) read mapping, (ii) transcriptome reconstruction, and (iii) expression quantification (Garber et al., 2011). One of the main steps in RNA-seq data analysis is the mapping and alignment of the short RNA-seq reads, either to the reference transcriptome or the genome of the organism. RNA-seq read alignment is challenging because they are short, contain sequencing errors, and span exon junctions.

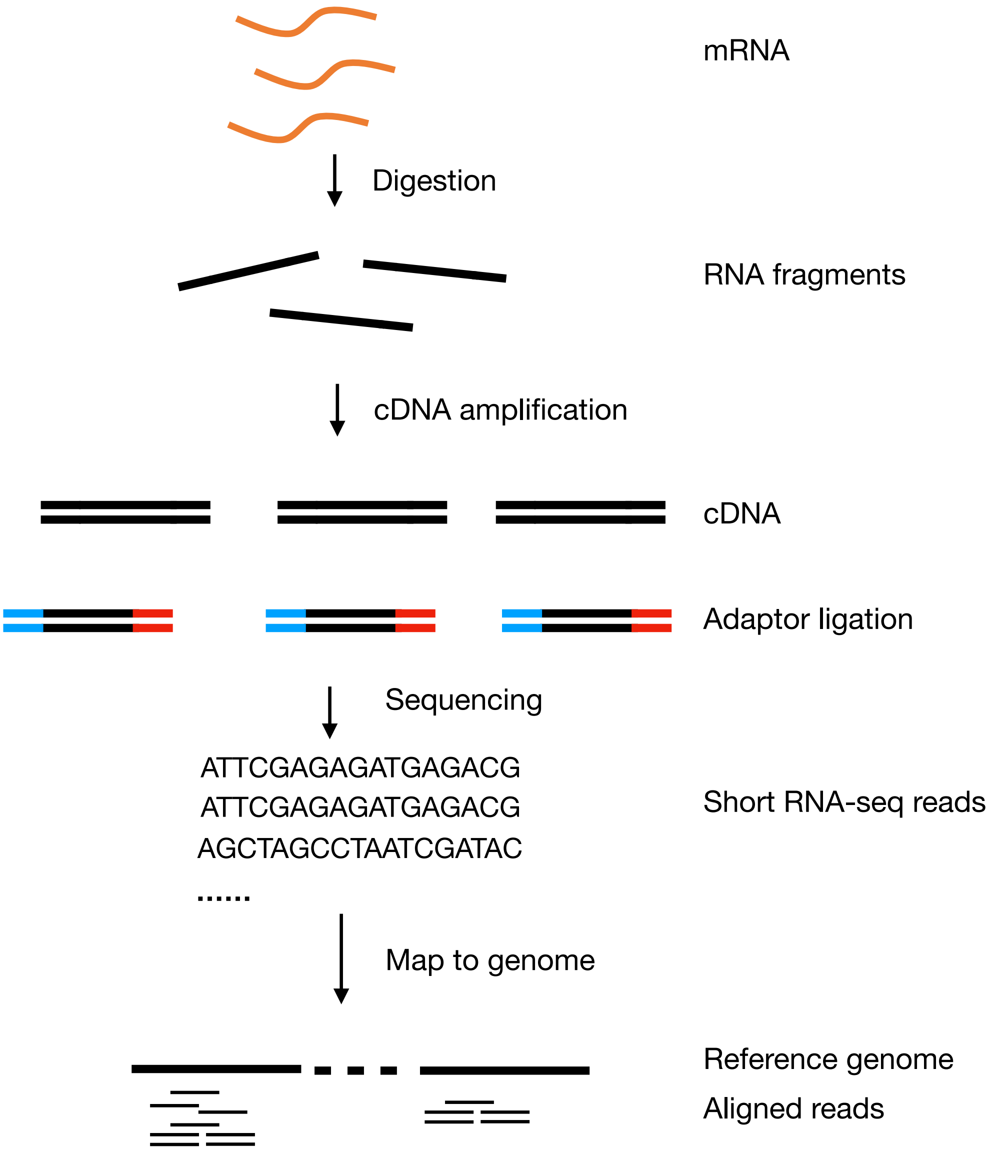


FIGURE 2.6: A typical RNA-seq workflow. Long RNAs are first fragmented and converted to cDNA. Sequencing adaptors (blue) are ligated to cDNA fragments. The using next generation sequencing technology, short reads are obtained. The sequencing reads are aligned to reference genome, transcriptome or loci, and used to estimate expression profiles.

2.3.2 Determining TF occupancy

Genome-wide mapping of regulator occupancy at regulatory sites is the essential step to fully understand the regulatory mechanisms of gene regulation. The method for genome-wide profiling of protein-DNA binding events is chromatin immunoprecipitation sequencing (ChIP-Seq) (Johnson et al., 2007). Binding events are estimated from the fraction of the DNA that is bound to a specific protein. To this end, using formaldehyde, all the proteins are cross-linked, and chromatin is sheared into fragments of size 100-300 bp. Subsequently, these cross-linked fragments are selectively enriched with factor specific antibodies by immunoprecipitation. Finally, the cross linking is reversed and the purified DNA is subjected to library preparation and high-throughput sequencing. The performance of ChIP-Seq depends on the quality of the antibody (sensitivity and specificity). For this reason, sensible quality control measures have been established.

The resulting short sequencing reads can be mapped to the reference genome using standard genome alignment tools such as STAR. In the end, we obtain sequence tag intensity for the whole genome, which represents the enrichment signal for the respective factor. This signal is normalized using the sequence tags derived from the same procedures omitting the immunoprecipitation step. Finally, standard peak calling softwares can be used to identify true factor binding sites.

2.3.3 Determining RBP occupancy

To profile the regulatory occupancy at the transcriptome, an analogous technology to ChIP-Seq has been developed. Photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) is a high-throughput method for profiling the binding sites of RBPS (Hafner et al., 2010). The steps are very similar to ChIP-Seq, with two main differences: First, UV-light induces cross-linking, and second, isolated RNA is converted into a cDNA library before deep sequencing.

2.3.4 Characterizing chromatin accessibility

Chromatin accessibility provides another useful lens to look at the genome from the gene regulatory perspective. Gene regulation requires access to regulatory elements. Hence, identifying accessible DNA in open chromatin is a crucial step in elucidating regulatory factor-binding events and other regulatory processes. DNaseI-Seq (Song and Crawford, 2010) is one of the assays for characterizing open segments of the genome. It is based on preferential cleavage of the accessible DNA compared to the condensed region by a nuclease enzyme. DNaseI cuts the DNA in locations where the DNA is open. Subsequently, DNaseI-digested fragments are collected, sequenced, and aligned to the reference genome. Quantifying the aligned sequencing reads allows for accurate and high resolution profiling of active regulatory sites and accessible DNA regions.

The DNaseI-Seq experiment provides more information beyond chromatin accessibility. Open chromatin regions frequently coincide with regulator occupancy. Moreover, each protein has its own particular profile of protection in the DNaseI-Seq experiment, and is reflected as small dips in the peaks. This enables elucidation of the regulators likely bound to these regions based on their motif profiles.

Chapter 3

Computational background

3.1 Optimization

The essence of most machine learning algorithms is optimization where, the goal is to optimize an objective function with respect to some constraints. There is an intricate interplay between optimization and machine learning. In the age of big data, scalability of learning algorithms are of paramount importance and, hence, there is a demand for constant design and efficient implementation of optimization algorithms. There are two main lines of research in this area (Bennett and Parrado-Hernández, 2006). The first is to design or extend current optimization algorithms for solving new learning models. The second is to solve existing machine learning algorithms more efficiently by exploiting learning model structure. Moreover, machine learning and optimization communities have different quality criteria of good optimization algorithms (Bennett and Parrado-Hernández, 2006). In the machine learning community, generalization, scalability, and fast convergence to the approximate solution of the model are the main factors. However, in the optimization community, accuracy of the solution, speed ,and numerical stability are the characteristics of good algorithms.

For the methods developed in this thesis, we make use of ideas and concepts from convex optimization. In this chapter, we review basic concepts of optimization based on (Boyd and Vandenberghe, 2004).

In general, a constrained optimization problem is defined as follows:

$$\begin{aligned} &\underset{x}{\text{argmin}} && f(x) \\ &\text{subject to} && g_i(x) \leq b_i,\ i = 1, \dots, m, \end{aligned} \tag{3.1}$$

where the $x \in \mathbb{R}^d$ is the *decision variable* or *optimization variable* of the program. $f : \mathbb{R}^d \mapsto \mathbb{R}$ is the objective function, $g_i : \mathbb{R}^d \mapsto \mathbb{R}, i = 1, ...m$ are the *constraints*, and $b_1, ..., b_m$ are bounds of the constraints. The goal is to find x, which minimize $f(x)$ while satisfying the constraints. In other words, the optimum solution, x^*, minimizes the function among all other feasible solutions to the program. The simplest interpretation of Eq.3.1 is that x represents a decision or action, and constraints impose limits on the actions.

Optimization arises everywhere, but the challenge is that most of the optimization problems are intractable, in other words, we can not solve them efficiently. Fortunately, for a subclass of optimization problems, convex functions, efficient algorithms and methods have been developed to solve them with polynomial complexity. Convex optimization has broad applicability in different fields, for example, machine learning, combinatorial optimization, finance, etc.

3.1.1 Convexity

We start with the definition of convex set and convex function. A set C is convex if any point in the set can be written as a convex combination of any other points in the set, which means that for any $\theta \in [0,1]$, they satisfy

$$\theta x + (1-\theta)y \in C.$$

In other words, a set is convex if any line segments that connect any pairs of points in the set lies in the set C (Figure 3.1). In convex function, both the objective and the constraint functions are convex. This means that, they satisfy the inequality:

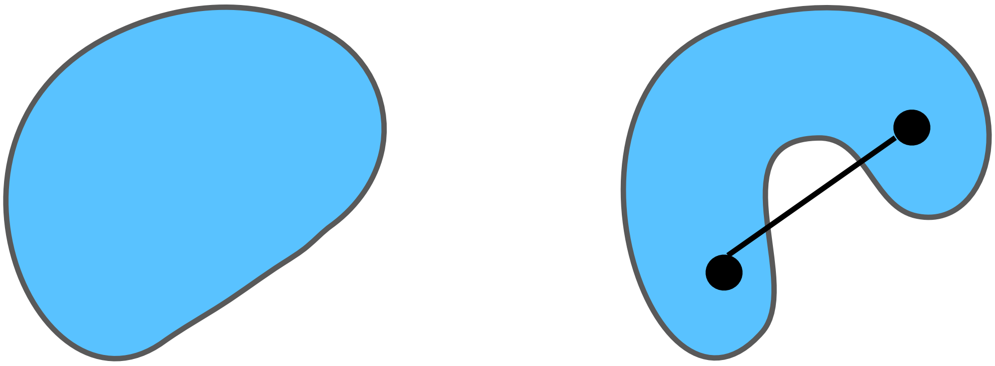


FIGURE 3.1: Simple convex and non-convex sets. *Left*. The round shaped which includes its boundary is a convex set. *Right*. The kidney-shaped set is not a convex set since a line segment that connects two points lies outside of the set.

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y),$$

for all $x, y \in R^d$ and for all scalars $\theta \in [0,1]$. A simple geometric interpretation of this inequality is that a line segment between $(x, f(x))$ and $(y, f(y)$ must entirely lie above (or on) the graph of the function f (Figure 3.2). Reformulating problems as convex optimization not only makes them attractive from the theoretical perspective

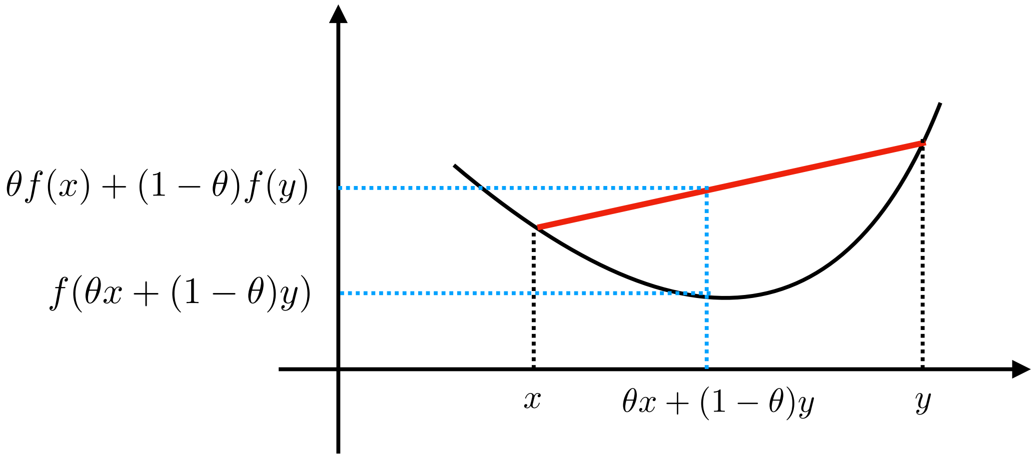


FIGURE 3.2: Graph of a convex function. A line segment $\alpha f(x) + (1-\alpha)f(y)$ always lies above the function value $f(\alpha x + (1-\alpha)y)$.

but also makes them actionable, which means we can solve them numerically.

Quadratic programs (QP) are one of the common optimization problem in machine learning. They consist of a quadratic objective and linear constraint. QPs play a central role in regularized risk minimization in machine learning. A QP can be expressed in the form of:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \frac{1}{2}x^T P x + q^T + r \\ \text{subject to} \quad & Gx \leq h \\ & Ax = b, \end{aligned}$$

where $P \in S^n_+$ is in the cone of positive semi definite matrices , $G \in \mathbb{R}^{m \times n}$, and $A \in \mathbb{R}^{p \times n}$.
In QP, we minimize a quadratic function over polyhedron, where the intersection of QP constraints forms a polyhedron (Figure 3.3).

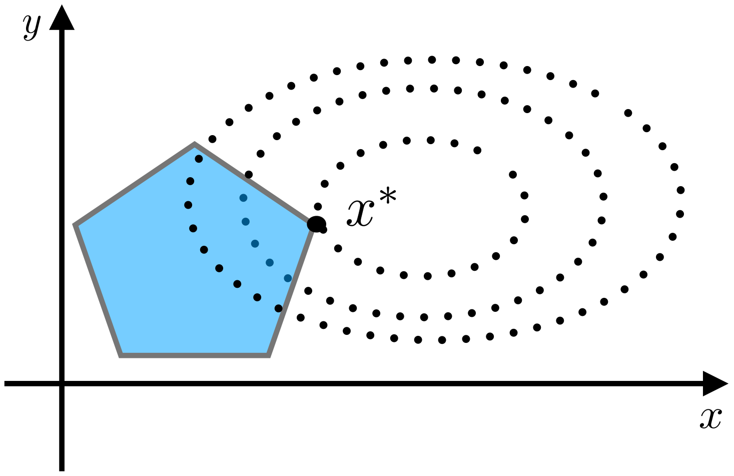


FIGURE 3.3: A quadratic program with two variables. The dashed curves represent contour lines of the objective function. The polyhedron defines the feasible solution region. The optimum value of the objective function is obtained at point x^*.

3.1.2 Optimality condition for contrained convex functions

Assume we want to minimize the smooth unconstrained function

$$\underset{x}{\text{minimize}} \quad f(x), \tag{3.2}$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a convex objective function and twice continuously differentiable. A necessary and sufficient condition for x^* to be the mimimizer of f is:

$$\nabla f(x^*) = 0. \tag{3.3}$$

Optimizing equation (3.2) is the same as finding the solution of (3.3). Except for a few cases where we can find the solution of (3.3) analytically, usually, the problem

must be solved by an iterative algorithm.
For a constrained optimization problem in the form

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq b_i,\ i = 1, \dots, m. \end{aligned}$$

We can convert it into an unconstrained convex problem by writing its associated Lagrangian function as follows:

$$L(x; \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x). \tag{3.4}$$

The non-negative weight $\lambda_i > 0$ is the Lagrangian multiplier, and it imposes a penalty whenever constraints $g_i(x) < 0$ are violated. Roughly speaking, Lagrangian provides us with a tool to solve constrained optimization problems by reducing them to an unconstrained optimization problem. According to the theory of Lagrangian duality, there exist an optimal vector λ^* such that the optimum value of f is a vector that minimizes the Lagrangian function of f. More specifically,

$$f^* = \underset{x \in \mathbb{R}}{\text{minimize}} \quad L(x; \lambda^*)). \tag{3.5}$$

To this end, any minimizer of equation (3.1) not only should satisfy the constraints but should also be a zero gradient of Lagrangian, and hence, it should satisfy the equation

$$0 = \nabla_x L(x^*; \lambda) = \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*). \tag{3.6}$$

3.1.3 Non-smooth functions and subgradients

Many interesting problems that occur in machine learning and statistics are convex but non-smooth and, hence, nondifferentiable. For example, ℓ_1-norm is convex but nondifferentiable at any point where at least one of the variables are zero. For this class of problems, the optimality conditions that we described for differentiable functions are not directly applicable. However, we can generalize the notion of the gradient to a non-differentiable function using a subgradient.

Definition 1. *(Subgradients). Given a convex function $f : \mathbb{R}^d \mapsto \mathbb{R}$ and a vector $\beta \in \mathbb{R}^d$, the subdifferential of f at β is defined as :*

$$\partial f(\beta) := \{z \in \mathbb{R}^d | \, f(\beta) + z^T(\acute{\beta} - \beta) \leq f(\beta) \quad for \quad all \quad vectors \quad \acute{\beta} \in \mathbb{R}^d\}. \tag{3.7}$$

Each element of $\partial f(\beta)$ is a subgradient of g at β.

A geomeric interpration of Eq. 3.7 is that any subgradient z in $\partial g(\beta)$ defines a linear function $\acute{\beta} \mapsto f(\beta) +^T (\acute{\beta} - \beta)$, which is tangent to the graph of f (Figure 3.4).

Subdiferentials are very useful for studying non-smooth convex functions because of the following property.

Proposition 1 (Subgradients at Optimality). *For any convex function $f : \mathbb{R}^d \mapsto \mathbb{R}$, a point $\beta \in \mathbb{R}^d$ is a global minimum of f if and only if the condition $0 \in \partial f(\beta)$ holds.*

This concept is mainly useful for nonsmooth functions. When f is differentiable, then $0 \in \partial f(\beta)$ reduces to first order optimality condition $\nabla f(\beta) = 0$.

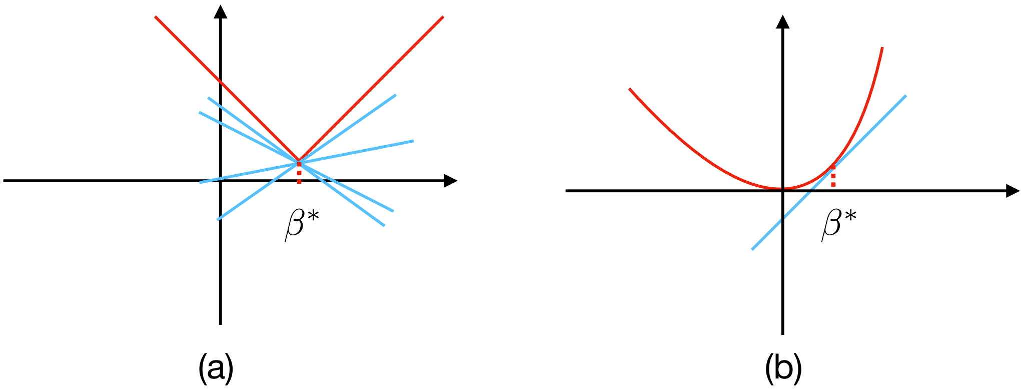


FIGURE 3.4: (a) Red curve represents the non-smooth convex function that is not differentialble at point β^*. Blue linear functions represent subgradients of the function f at point β^*. (b) Red curve represents smooth function and linear function f represents the gradient of f at point β^*.

3.1.4 Descent methods

Algorithms for solving convex functions can be categorized into two classes, namely *first order methods* and *second order methods*. A first order method is any numerical algorithm that requires the first gradient of the objective function. A second order method beyond of using first gradient information also leverage second gradient (Hessian) of the function. Hessian matrix provides information regarding the curvature of the function, but computing Hessian can be computationally prohibitive for large scale problems. For this reason, first order methods have become very popular in machine learning, since they only use gradient information; hence, they are scalable to large scale problems.
We have briefly reviewed optimality conditions for different types of convex functions. Now, we briefly review basic principles for two first order methods, gradient descent and proximal gradient descent, for solving convex functions.

Gradient descent algorithms solve an unconstrained objective function with producing a minimizing sequence of $\{\beta^t\}_{t=1}^{\infty}$ via the update

$$\beta^{t+1} = \beta^t - s^t \nabla f(\beta^t) \quad for \quad t = 0, 1, 2, \ldots, \tag{3.8}$$

where $s^t > 0$ is the step size. Intuitively, by computing the gradient, we choose the direction of steepest descent $-f(\beta)$, and we descend in this direction for a certain amount determined by s^t. A gradient descent algorithm is granted to converge to a local minimum with a rate of $\mathcal{O}(\frac{1}{k})$, where the k is the number of iterations. Since for convex functions, all local minima are a global minimum, convergence to global minimum is granted. Nevertheless, several extensions improve the convergence rate of the gradient descent (Qian, 1999; Kim and Fessler, 2016; Nesterov, 1983). In 1986, *Yurii Nesterov* developed an accelerated gradient descent that improves the convergence rate of gradient descent to $\mathcal{O}(\frac{1}{k^2})$ (Nesterov, 1983). Moreover, gradient descent has been tailored to solve certain constrained optimization problems as well (Calamai and Moré, 1987). Gradient descent is easily applicable to differentiable functions, but it get ill-posed when it encountered with non-smooth function.

Proximal gradient descent Gradient descent is a widely used algorithm for smooth convex optimization but is limited when it comes to non-smooth convex functions. Many interesting problems in machine learning translate into non-smooth convex functions. We often deal with these non-smooth functions as regularizers in machine learning, statistics, signal processing, etc. Now we review an algorithm, *proximal gradient descent*, that extends the ability of gradient descent to solve a subclass of non-smooth convex functions for large scale problems.

Proximal methods (forward-backward splitting methods) are a class of algorithms that uses the proximal operators of objective terms for solving a convex optimization problem.

We first define a basic element of proximal algorithms. *Proximal operator* of a function $g(x)$ is defined as the function that maps a vector $u \in \mathbb{R}^d$ to the unique solution of

$$\mathbf{Prox}_{\mu g}(w) = \underset{w \in R^p}{\text{minimize}} \quad \frac{1}{2}||u - w||_2^2 + \mu g(w). \tag{3.9}$$

It compromises between minimizing $g(w)$ and being close to point u , and μ controls a trade-off between these terms. For this reason, $prox_g(w)$: is called a proximal point of u with respect to function g. Under some assumptions (Parikh and Boyd, 2014), the proximal operator of function g can be viewed as a gradient step for g

$$\mathbf{Prox}(u) = v - \nabla g(v). \tag{3.10}$$

We now consider minimizing a non-smooth convex function $\ell(x)$, which we can decompose into a sum of two functions:

$$\text{minimize} \quad f(x) + \Omega(x), \tag{3.11}$$

where the f is smooth and differentiable with Lipschitz-continuous gradient, and $\Omega(x)$ is non-smooth and non-differentiable, for example an ℓ_1-norm. This decomposition may not be unique, and different decomposition can lead to a different implementation of the proximal gradient descent for the $\ell(x)$.

Proximal gradient descent defined as

$$x^{k+1} := \mathbf{Prox}^k_{\lambda\Omega}(x^k - \lambda^k \nabla f(x^k)), \tag{3.12}$$

where the k is the *k-ith* iteration of the algorithm. If we look carefully at equation 3.12, we realize that the proximal gradient descent algorithm consists of two main steps. First, $x^k - \lambda^k \nabla f(x^k)$ is a computing gradient descent step k on $f(x)$, and the second step is to evaluate the proximal operator of the $\Omega(x)$ on the solution that we get from the gradient descent step of $f(x)$. These two steps iterate until the algorithm converges.

Proximal algorithms generalize the notion of projecting a point into convex sets and are somewhat related to projected gradient descent algorithms. Generally, finding a proximal operator for the non-smooth part is equivalent to solving another optimization problem, but it turns out that, the proximal operator for many interesting problems, often accepts a closed formula solution or can be solved quickly with standard approaches. Proximal algorithms are practical when all the relevant proximal operators can be computed efficiently. Proximal methods are very suitable for solving non-smooth convex functions because of their convergence rate and their ability to deal with large-scale non-smooth functions. There are several extensions of proximal gradient descent that improve the convergence rate. For example, when it is coupled with Nesterov's acceleration method, we can achieve a convergence rate of $\mathcal{O}(\frac{1}{k^2})$. Moreover, since it allows a warm start, we can solve the objective function along the entire regularization path.

3.2 Supervised learning

Now we move on to introduce the basics of learning from data. We are given a sample of observations $T = (X_i, Y_i)_{i=1}^n$, coming from unknown probability measure $P(x, y)$. Also, we assume that data points are i.i.d, that means that they are drawn independently and identically distributed. X_i is the input feature vector, and Y_i is the output variable. Our goal is to learn a function $f_n : X \mapsto Y$ from a function class $\mathcal{F} = \{f | \, f : X \mapsto Y\}$, such that a specific objective function is minimized. More specifically, we assume that there is a function f which maps the input features to the output variable as follow:

$$y = f(x) + \epsilon, \tag{3.13}$$

where ϵ is an irreducible error and may contain unmeasurable variables that contain information about the output variable. However, the true underlying function is unknown, and our goal is to learn f from training data. More specifically, we want to fit a model to training data and make a prediction as accurate as possible on unseen data that comes from the same unknown probability distribution. In other words, given training data and a *loss function* $L(Y, f(X)$, we want to learn a function f which is as close to the optimum function f^* as possible. The term "closeness" is defined as the difference between the risk of the function f and the smallest possible risk.

Below, we discuss two frameworks for learning where almost all (semi) supervised learning algorithms can be formulated in one of these frameworks. This formulation creates a unified point of view on supervised learning algorithms, where the resulting learning algorithms has three main components: the loss function, the regularizer, and the employed function class.

3.2.1 Risk minimization

Recall that our goal is to select a function from a class of functions that, given a new input feature vector X_i, it best approximates the corresponding y_i value under the assumption that the data comes from the same but unknown underlying distribution. This raises an important question as to how we should select our function.

Essentially, the task of finding the "best" prediction function boils down to the task of finding a function that minimizes the "expected risk" or "generalization error" as follows:

$$R(f) = E[L(f(x), y)] = \int L(f(x), y) dP(x, y). \tag{3.14}$$

The loss function, $L(y, f(x))$, quantifies the discrepancy between prediction output $f(x)$ and true output y. The expected risk measures the expectation of loss concerning an infinite number of samples drawn from distribution $P(x, y)$.

Empirical risk minimization Unfortunately, in the real world, we almost never have the true probability distribution, $P(x, y)$. This makes it difficult (almost impossible) to compute the true value of the risk for a given loss function. Instead, we try to approximate P(x,y) using a given finite sample D, called *training data*. We assume to have i.i.d sample $(X_i, Y_i)_{i=1}^{n}$ input/output pairs drawn from data generating probability $p(x, y)$. The *empirical loss* is defined as

$$E_{P_n}[L(Y, f(X)] = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(x_i)). \tag{3.15}$$

Given a function class $\mathcal{F}$, *emprical risk minimization* is defined as

$$R_{emp}(f) = \underset{f \in \mathcal{F}}{min} E_{P_n}[L(Y, f(x)] = \underset{f \in \mathcal{F}}{min} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)). \tag{3.16}$$

The most often used loss function for regression is square loss. Given a function class $\mathcal{F}$, the optimization problem for the empirical risk minimization is defined as:

$$R_{emp}(f) = \underset{f \in \mathcal{F}}{min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2. \tag{3.17}$$

The standard function class $\mathcal{F}$ for linear regression is defined as

$$\mathcal{F} = \{f(x) = \langle \beta, X \rangle | \beta \in R^d\}. \tag{3.18}$$

If we choose a large functional class $\mathcal{F}$, we might get a very small empirical risk value. Necessarily, this does not mean that we have found a function f that will perform well on the unseen data. This means that we might overfit the data. For this reason, empirical risk underestimates the true risk, i.e., $R_{emp} \leq R$ generally holds.

Regularized empirical risk minimization (RERM) The main problem of empirical risk minimization is that it is more likely to overfit the data. There are two main strategies to overcome overfitting. The first solution is to restrict ourselves to a small function class $\mathcal{F}$, e.g., linear functions. But this creates another problem called *underfitting*, meaning that if the true underlying function is more complex than what our function class can model, we are always far from the best possible solution, irrespective of the size of the training data.

The other solution is to use regularization while allowing to have a large functional class. A regularization functional captures the complexity of the function. Hence, adding a regularization functional to the objective function of optimization problems creates a trade-off between fitting data and the complexity of the function. This form of regularization was developed by Tikhonov and is known as **Tikhonov regularization**.

Definition 2. *Given a training sample $(X_i, Y_i)_{i=1}^{n}$, a loss function $L(Y, f(X))$, a class of functions $\mathcal{F}$, and the regularization functional $\Omega : \mathcal{F} \mapsto R_+$, the* ***regularized empirical***

risk minimization *is defined as*

$$f_{n,\lambda} := \underset{f \in \mathcal{F}}{argmin} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda \Omega(f). \tag{3.19}$$

Another advantage of adding a regularization functional to the objective function of optimization problem is that it makes the problem often easy to optimize, since it makes the problem well posed. This regularization term is also connected to the concept of *bias-variance trade-off* in statistical learning (Geman, Bienenstock, and Doursat, 1992). Shortly, regularization leads to a biased estimator, while unbiased estimator might suffer from high variance and, hence, poor generalization. Currently, most of state-of-art machine learning algorithms use the principle of RERM.

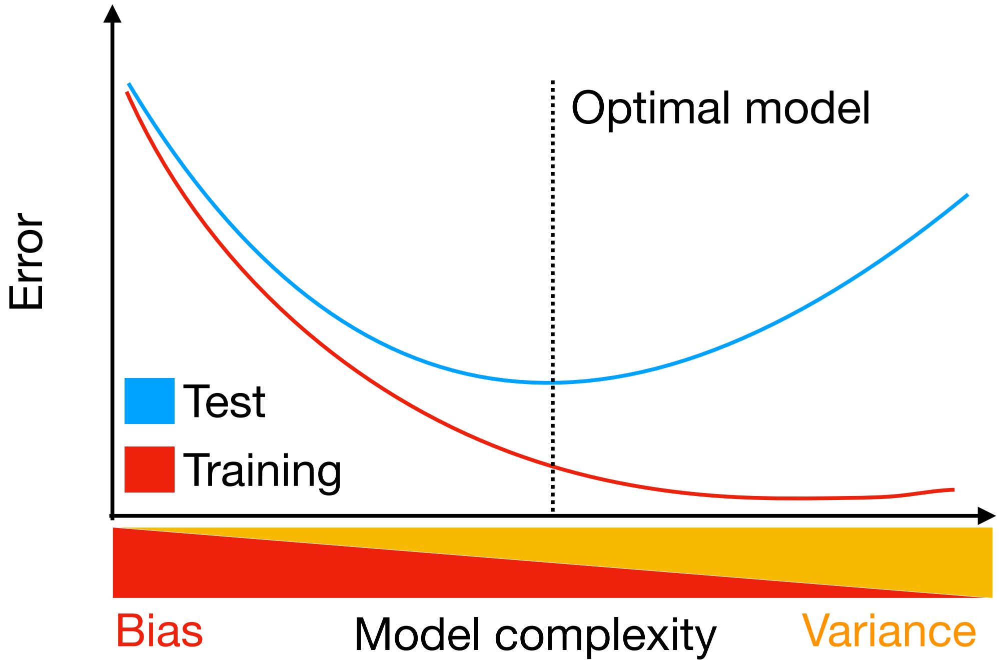


FIGURE 3.5: Behaviour of error for supervised learning models depending on the model's complexity. The training error shows a monotonic decrease with respect to the complexity of the model due to ability of the model to better model the training data. This can have a destructive effect on the test error, as the model shows poor generalization performance on unseen data.

The Bias-Variance Trade-Off The empirical solution β_n that we get by optimizing a RERM optimization function is an approximation to the optimal weight vector β^*. This approximation depends on the training sample $(X_i, Y_i)_{i=1}^n$ size n; hence, it can be random. Since our samples are i.i.d and come from same but unknown probability distribution $P(x, y)$, this raises two important questions about an estimator.
The first question is, if we compute β_n over all possible training samples size n, does $E[\beta_n] = \beta^*$? This is called *bias* and it is the systematic error we make when we do approximate the true underlying function.
The second question is, how much β_n will fluctuate around its expected value over

all possible training samples of size n? This quantity is called *variance*. If an estimator has high variance, then small changes in the training data sets can lead to large changes in f.

It turns out that we can decompose the *expected loss* into

$$E[R(f_n)] = \mathrm{Var}(f_n(x)) + (\mathrm{Bias}(f_n(x)))^2. \tag{3.20}$$

There is an interesting interplay between bias-variance decomposition and overfitting-underfitting (Figure 3.5). If we choose to have large a function class $\mathcal{F}$, it is more likely that our bias will be close to zero; however, the variance of our estimator will be large, since we need a large amount of data to learn the large model. On the other hand, if we choose a very small function class $\mathcal{F}$, we will have a large bias, because the model cannot fit well to the data (i.e., true underlying data generating models may be extremely complicated), but this will lead to low variance, since the fitted function does not vary much given different training data sets. Eq. (3.20) tells us that in order to minimize an expected test error, we need to have an optimum trade-off between the bias and variance of our model.

3.2.2 Regularized linear regression

Regularized linear regression is a very simple, yet powerful, class of methods for learning model $p(y_i|x_i)$. In high dimentional regression problems, our goal is to predict the output variable y from input vector $x \in \mathbb{R}^p$. This allows us to estimate parameters describing the dependence of output variable y_i on input features x_i.

Regularization enforces a preference over regression parameters, which can help us to address the issue of *bias-variance* trade-off, especially in very high dimensional data.

Given training sample $(x_i, y_i)_{i=1}^n$, the l_q regularized estimator can be obtained by solving the following optimization problem:

$$\underset{\beta \in R^p}{\mathrm{argmin}} f(\beta) = \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \Omega(\beta), \tag{3.21}$$

where the loss $\ell : \mathbb{R}^p \mapsto \mathbb{R}$ is assumed to be a smooth convex function with Lipschitz continuous gradient, and the regularizer $\Omega(\beta)$ is convex, typically non-smooth and non-Euclidean norm.

A typical example of loss function for linear regression is squared loss.

$$\ell(y_i, f(x_i)) = (y_i - \beta^T x_i)^2, \tag{3.22}$$

where parameters β_j describe magnitude and direction of *jth* feature on the target variable y.

Regularization not only addresses the issue of bias-variance trade-off but also is one of the practical frameworks to impose prior knowledge on the support of the predictor. For this reason, design and application of different sparsity-inducing norms is an active research field.

3.2.3 Sparsity inducing norms

The principle of parsimony, which states that the most straightforward description of a given phenomenon should be favored over more complicated ones, is central to different fields of science. In the context of learning, this translates into feature selection which has two main implications. First, make the model more interpretable or computationally efficient to use, even if the true underlying model is not sparse. Second, sparsity can be viewed as prior knowledge that our model should be sparse. We can achieve parsimony in a linear model by penalizing the empirical risk minimization or log-likelihood function by the size of the non-zero entries of the coefficient vector. However, reducing parsimony to the problem of learning a model with the lowest cardinality turns out to be insufficient and hence, *structured parsimony* has been emerged as an extension of it (Jenatton, Obozinski, and Bach, 2010; Jacob, Obozinski, and Vert, 2009).

Sparsity through ℓ_1 norm

In the last decade, numerous approaches have been developed to study ℓ_1 penalized regression from a theoretical perspective to efficient implementation (Zhao and Yu, 2006; Wainwright, 2009; Bickel, Ritov, and Tsybakov, 2009). When we know *a priori* that the optimum solution β^* for Eq. (3.21) should have a few non-zero entries, then ℓ_1 norm, i.e., $\Omega(\beta) = \sum_{i=1}^{p} |\beta_i|$ norm, is typical choice for $\Omega(\beta)$. Depending on the loss function, this penalization can lead to lasso (Tibshirani, 1996) or basis pursuit (Chen, Donoho, and Saunders, 2001). Regularization with ℓ_1 norm leads to a sparse solution meaning that, depending on the value of the regularizer, a number of coefficients β^* will be exactly equal to zero.

Elastic net

However standard Lasso penalty is a simple yet powerful regularizer for sparse learning, but it comes with limitations. For the case of high dimensional data $p \gg n$, it selects at most n variables before it saturates. Moreover, if there are any structured patterns in the input features (group structure or correlated variables), Lasso tends to be somewhat indifferent in selecting among a set of strong structured features. One of the ways to overcome this problem is to use a convex combination of ℓ_1 with ℓ_2 norms (Zou and Hastie, 2005). Adding the quadratic term to the regularizer makes the elastic net strictly convex, hence leading to a unique solution. The elastic net regularization function is defined as:

$$\Omega(\beta) := \lambda(\alpha||\beta||_1 + (1-\alpha)||\beta||_2^2). \tag{3.23}$$

The $||.||_2$ encourages highly correlated features to shrink toward each other, while the ℓ_1-norm induce a sparse solution over the coefficients of the shrunken features. The *elastic net* regularizer make the trade-off between these two effects. The elastic net penalty can be used with any loss function in the context of regression and classification.

Sparsity through ℓ_1/ℓ_q-norms

The standard ℓ_1 penalty does not assume any structure or dependencies between input features, which limits its applicability to structured high dimensional data in many real-world problems. In many situations, input features can be partitioned into *groups* of variables. This is typical when we are working with categorical variables, or we do have prior knowledge of group structures. It is desirable to simultaneously include or exclude all the variables forming a group. One can incorporate structured sparsity constraints by designing and applying a more complicated sparsity-inducing penalty that induces a joint sparsity pattern over related groups. It has been shown that regularization norms that explicitly exploit group structures improve the generalization performance and(or) interpretability of the models (Huang and Zhang, 2010; Lounici et al., 2009; Obozinski, Taskar, and Jordan, 2010). One of the simplest but powerful group norms is the so-called $\ell_1/\ \ell_2$ norm:

$$\Omega(\beta) := \sum_{g \in G} w_g ||\beta_g||_2, \tag{3.24}$$

where G is the partition of features $1, ..., p$, w_g is the group specific weight, and $\beta_g \in R^{|g|}$ denotes the coefficients of the features indexed by g in G. As defined in Eq.(3.24), Ω behaves like an ℓ_1-norm at group level $(||\beta_g||_2)_{g \in G}$ in $\mathbb{R}^{|G|}$, and hence, Ω induces group sparsity. In other words, each group β_g, is encouraged to be set to zero by ℓ_1 penalization, but combined with the ℓ_2-norm, the mixed ℓ_1/ℓ_2-norm plays the role of simultaneously setting all the weights within each group to zero or non-zero values. More specifically, if estimated $||\beta_g||_2 \neq 0$, then all the β_j for $j \in g$ will be non-zero. Hence the ℓ_1/ℓ_2-norm induces sparsity at group level, not within group sparsity. To achieve within group sparsity in the context of mixed ℓ_1/ℓ_2-nrom, *Friedman et al.* proposed to use additional ℓ_1 norm with $\Omega(\beta)$, which is known as the *sparse group lasso* penalty (Friedman, Hastie, and Tibshirani, 2010b) and is defined as

$$\Omega(\beta) := ||\beta||_1 + \sum_{g \in G} w_g ||\beta_g||_2. \tag{3.25}$$

In practice, the ℓ_1/ℓ_2 and ℓ_1/ℓ_∞-norm is used. The grouped ℓ_1-norm is usually used when we have the group structures as prior knowledge in multi-task learning (Obozinski, Taskar, and Jordan, 2010) and for multiple kernel learning (Xu et al., 2010).

3.2.4 A geometrical intuition for the sparsity-inducing properties

We have described how we could use different structured sparsity-inducing norms to incorporate our prior knowledge and learn a rich set of models with good generalization performance. The main question that arises here is how the geometry of the regularizer is directly related to the minimizer β^* of empirical regularized risk minimization.

The regularized formulation that we considered in Eq. (3.21) is the Lagrangian relaxation of the following constrained optimization problem:

$$\underset{\beta \in R^p}{min} \ell(\beta) \quad \text{such that} \quad \Omega(\beta) \leq \mu, \tag{3.26}$$

for some $\mathbb{R}_+$. Under some weak assumption on the $f(\beta)$ and Ω, from the Lagrangian multiplier theory, we can show that constrained formulation (Eq. 3.26) of the regularized risk regression (Eq. 3.21) is equivalent to each other for $\mu > 0$. This means that the set of solutions for these two formulations is equivalent for different values of μ and λ; however, there is no direct mapping between the corresponding value of λ and μ (Borwein and Lewis, 2010). In machine learning, the Lagrangian

form is preferred, since the solution is more robust to small changes in λ.
Now we focus on the constrained form of the $\ell(\beta)$ to describe the behavior of the sparsity-inducing norms.
The constraint $\Omega(\beta) < \mu$ defines the set of feasible solutions $\mathbb{B} = \{\beta \in \mathbf{R}^d; \Omega(\beta) \leq \mu\}$. At the optimum point, the gradient of f at any solution β to Equation 3.26 belongs to the normal cone of $\mathbb{B}$. This means that the level set of $f(\beta)$ is tangent to $\mathbb{B}$. Hence, depending on the $\Omega(\beta)$, the shape of the constraint will be different, and as a consequence, the geometry of the $\Omega(\beta)$ is directly related to the properties of the solution. For example, if $\Omega = ||.||_2$, then the resulting constraint will be a *round* ball that does not prefer any specific direction in the model parameters space. On the other hand, when Ω is taken to be ℓ_1-norm, then B is anisotropic (diamond shape pattern in two dimensions), and due to the non-smoothness, it will present some singular point on its surface. Furthermore, these non-smooth points are aligned in the direction of the features axis, and hence if the level set of the objective function $\ell(\beta)$ happens to be the tangent to one of these singular points, it will lead to the sparse solution (Figure 3.6).

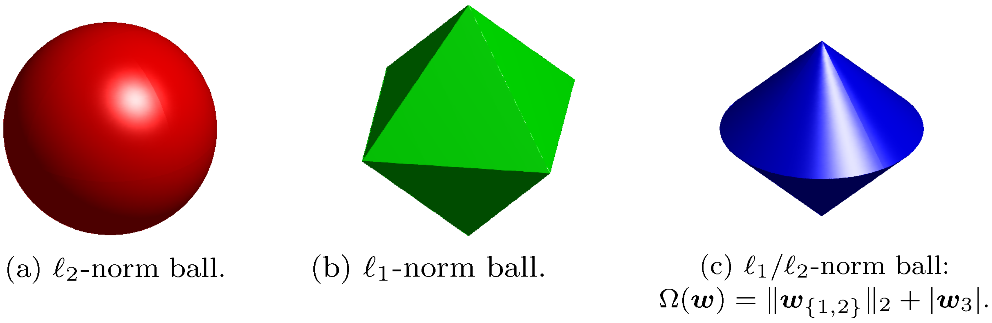

(a) ℓ_2-norm ball. (b) ℓ_1-norm ball. (c) ℓ_1/ℓ_2-norm ball: $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}_{\{1,2\}}\|_2 + |\boldsymbol{w}_3|$.

FIGURE 3.6: Geometry of different sparsity inducing norms. (a) shows the smooth ℓ_2-norm ball. (b) ℓ_1-nrom. Singular points are not differentiable and usually are in the direction of features axis, hence induce sparsity. (c) ℓ_1/ℓ_2 norm. This figure is reproduced from (Bach et al., 2012)

3.2.5 Multi-task learning

The common practice in machine learning is to learn one task/model each time. This methodology is sometimes sub-optimal because it does not leverage the rich domain-specific source of information available in the training data of related tasks. Multi-task learning is an approach in machine learning to simultaneously learn a group of related tasks, partly using a shared representation. This creates a framework to effectively transfer the information between related tasks by increasing the amount of data available per parameter and, hence, reduces the overfitting and improves the generalization.
This is especially important in the context of constructing a biological regulatory network from high-throughput data because the number of the parameters to learn is relatively way higher than the number of samples. Moreover, our model needs to be rich enough to capture the considerable complexity of biological mechanisms, which in turn requires reasonably sized training data. Hence, extensive data from one task may be able to compensate for the sparse noisy data in other related tasks. Many problems in computational biology can be formulated as multi-task learning problems (Qi et al., 2010). Different tasks might correspond to different cell lines,

tissues, pathways, or tumor subtypes (Widmer and Rätsch, 2012).
Recall that the regularized estimator for a single model is defined as

$$f(\beta) = \ell(y, \langle \beta_j, X \rangle) + \Omega(\beta). \tag{3.27}$$

In multi-task learning, we are interested in simultaneously learning multiple models parameterized by $\beta_1, \dots, \beta_M$, where the M is the number of tasks. We can easily extend the single task learning formulation (Eq.3.27) to a multi-task formulation by adding additional regularization terms that couple the model parameters as follows:

$$f(\beta_1, \dots, \beta_M) = \sum_{j=1}^{M} \ell(y, \langle \beta_j, X^j \rangle) + \sum_{j=1}^{M} \Omega(\beta_j) + \Omega_{\text{MTL}}(\beta_1, \dots, \beta_M). \tag{3.28}$$

The objective function for MTL (Eq.3.28) combines two main interrelated goals. First, like other single task learning algorithms, the first term in Eq. 3.28 minimizes the loss function for each task independently, while at the same time regularizing each task's parameters by $\Omega(\beta)$ avoids overfitting. Second, the regularizer $\Omega_{\text{MTL}}(\beta_1, \dots, \beta_M)$ shares the parameters across the tasks such that final parameters are similar to each other (Evgeniou, Micchelli, and Pontil, 2005).

3.2.6 Cross-validation

One of the fundamental tasks in machine learning is model selection, and the bottom line in model selection is generalization performance. In other words, the selected model should have a low *test error rate*. We can compute the test error if designated test data is available, which is usually not the case. In the absence of a large amount of test datasets to estimate the *test error rate*, a number of algorithms have been developed to estimate it using training data.
Cross validation (Golub, Heath, and Wahba, 1979; Stone, 1974) is generally the preferable method for approximating test error. In k-fold cross-validation, training data is randomly partitioned into k-disjoint subsets. In the ith cross-validation fold, the ith subset is used to compute prediction performance, while the rest of $k-1$ are used for training the model. The expectation of generalization performance in k folds used as an approximation of the test error rate. One needs to be careful in applying model selection approaches for tuning hyperparameters. *Cawley et al.* have shown that hyperparameter selection via cross-validation is prone to over-fitting in *model-selection*. Hence, in order to obtain an unbiased estimate of generalization performance, more rigorous procedures like nested-cross validation are needed (Stone, 1974).

3.3 Gene expression modeling and regulatory network inference

A regulatory network describes the connections between regulators and their target genes. These networks are an abstraction of condition-specific gene regulation. Inferring a condition-specific regulatory network is important because of their relevance to cell identity, diseases, and development. Regulatory networks are often represented as a directed graph, in which nodes represents regulators (TFs, miRNAs, RBPs, signaling proteins, and chromatin remodelers) and their targets. The edges represent directed regulatory relationship between regulator and the target. The expression level of the target gene is the function of all the regulators that binds, the activity level of regulators, as well as cis-regulatory elements features. The activity of the regulator depends on its abundance, location, and possible post-translational modifications of the regulator. The regulatory relationship can be direct or indirect. For example, a miRNA can target a TF, which can regulate gene on the transcriptional level. A regulatory network model has two main aspects: (I) structure or topology of the network, which represents the regulators of a target gene, and (II) the logic or regulatory function, which describes how combinations of different regulators specify a gene expression of targets. At a high level, methods for inferring regulatory networks can be divided into (I) Experimental, and (II) Computational approaches.

Experimental methods for inferring regulatory networks

The most widely used experimental approach for determining the transcriptional regulatory network is ChIP-Seq and ChIP-chip technology. This measures the genome-wide occupancy of a particular regulator. The main limitation of this technology is the need for a TF-specific antibody. Another approach is to knock-out a specific regulator and infer targets by measuring changes in gene expression. The limitation of this approach is its inability to discern direct vs. indirect target of the regulator. A more recent approach is the combination of chromatin accessibility assays (ATAC/DNaseI-seq, etc.) with TF binding site models such as a position weight matrix (PWM) to infer the genome-wide binding of the regulator. Overall the experimental approaches are accurate, expensive, time consuming, and less high-throughput.

Computational methods for inferring regulatory network

Here, we briefly review computational methods for both inferring regulatory network structure and regulatory functions.

Correlation-based approaches: Correlation is the most straightforward measure for quantifying the degree of co-expression for two genes, such as Pearson or rank-based correlation. The underlying assumption of correlation networks is that similar regulators might regulate genes with similar expression pattern in different samples. This idea led to the development of the concept of weighted gene coexpression network analysis (WGCNA) (Zhang and Horvath, 2005), which has been widely used and adapted. Correlation based-methods come with limitations however. First, they do not have expression predictive power. Second, they are limited to discern the indirect association from indirect association between a pair of genes, which can lead to false positive associations. *Partial correlation* networks offer an alternative to remove the indirect effects by other genes (Schäfer and Strimmer, 2004; Friedman, Hastie, and Tibshirani, 2008).

Information theory-based approaches: Another class of methods used to quantify the dependency between two variables is mutual information. Mutual information overcomes the limitation of correlation-based approaches in capturing non-linear dependencies between two variables. In the context of gene regulatory networks, the use of mutual information is refered as *relevance networks* (Butte and Kohane, 1999) which are widely used (Meyer et al., 2007).

Regression-based methods: The central idea behind this class of approaches is to formulate a regression problem, where information of the regulators is used to regress gene expression. The regression weight vector describes the direction and magnitude of the regulation between the gene and the respective regulators. This idea is the basis of several successful methods for network reconstruction (Haury et al., 2012; Lebre et al., 2010; Irrthum, Wehenkel, and Geurts, 2010). The regression-based methods are scalable and have predictive power of gene expression. Furthermore, they can capture higher-order interactions between different regulators. Moreover, one can incorporate different priors regarding the target genes or regulators using various sparsity inducing norms.

Furthermore, several other methods including Markov networks, factor graphs, and Bayesian networks are powerful paradigms for network inference. The readers are referred to (Wang and Huang, 2014; Sanguinetti, 2019) for comprehensives reviews of these methods.

The DREAM consortium (Marbach et al., 2012) assessed more than 35 different gene regulatory network inference algorithms, which covered all commonly used regulatory network inference algorithms in the field. However although most of the methods showed a performance better than random on simulated and *E.coli* gene expression data, overall the performance of these algorithms dropped significantly for eukaryotic organisms, such as yeast. This implies that expression data alone is not enough to infer a high quality regulatory network. One trend is to include auxiliary data sets, which are informative of the regulatory relationship between the regulator and the target gene. All expression based network inference algorithm can be categorized into per gene methods and per module methods.
Per gene methods consider genes as a random variable and estimate its regulators. The main advantage of per gene methods is the ability to reason at the level of individual genes. On the other hand, module-based methods group genes based on their co-expression pattern and they are learning the regulators for each module. Since module-based methods summarize information at the level of each module, they are more interpretable, but provide an approximation of the regulatory program for each gene. A more recent effort by (Roy et al., 2013; Koch et al., 2017) combines these two paradigms.

3.4 Network analysis

Here we describe some of the basic network analysis methods used in this thesis to identify important regulators.

Definitions

Definition 3. *(Graph). A graph $G = (V, E)$ comprises an arbitrary, finite set of vertices V(also called nodes) and set of edges $E \subseteq V \times V$ where each edge is assigned into two* *vertices. A graph G is undirected if there is no direction on the edges, i.e., for each edge $(v, \nu) \in E$ also $(\nu, v) \in E$ holds.*

Definition 4. *(In- and out-degree). Given a graph $G = (V, E)$, the in-degree d_v^- of a node v is the number of the head ends to v. Conversely, the out-degree d_v^+ of v is the number of the tail ends to v.*

In order to use the network to find the important nodes or edges based on their connectivity patterns relative to the neighbors, we can use network *centrality indices*. Centrality indices represent the relative importance of the nodes or edges by assigning a real number to them concerning network characteristics. Depending on the type of network, we can describe what *importance* means. Centrality indices can be categorized into *local* and *global* centrality matrices. Local centrality indices, like degree centrality, only consider the direct neighbor of the node of interest, whereas global centrality measure (eigenvector centrality, page rank) take into account the whole network (direct and indirect connections).
The simplest centrality measure is degree centrality, which is defined as follows:

Definition 5. *(degree centrality). Given an undirected graph $G = (V, E)$, the degree centrality is defined as:*

$$C_{deg}(v) := |\{e \mid e \in E \land v \in e\}|.$$

Degree centrality counts the numbers of edges that are directly connected to each node. For directed networks, we can extend the degree centrality definition for *in-degree* and *out-degree* centrality. Degree centrality has been successfully applied to identify essential proteins or genes for the survival of the organism in different biological networks (Jeong et al., 2001; Hahn and Kern, 2004; Bergmann, Ihmels, and Barkai, 2003).

The eigenvector centrality idea, first proposed by *Philip Bonacich* (Bonacich, 1972), states that the centrality value of a node directly depends on the centrality value of the connected neighbors.

Definition 6. *(eigenvector centrality) Given a strongly connected graph $G = (V, E)$ and A the adjacency matrix for G, the eigenvector centrality corresponds to the eignevector C_{eiv} of the largest eigenvalue λ_{max} of the following equation:*

$$\lambda C_{eiv} = A C_{eiv}.$$

In order to make the eigenvector centrality non-negative, we can select the eigenvector corresponding to principal component A high eigenvector value for node v_i means that it is connected to many nodes who themselves are connected to many others. Hence, this centrality measure takes into account the direct and indirect interactions between the nodes. This centrality score is very well suited for understanding the information propagation in the network.

Chapter 4

Integrative analysis of miRNA and protein-based post-transcriptional regulation

This chapter presents an integrative analysis of transcriptome and miRNAome-dynamics to elucidate the molecular function of IMP2, an RNA binding protein, and its contribution to disease in the context of adult cellular metabolism. Our analysis shows that IMP2 compete with miRNAs and alters the regulatory capacity of many miRNAs.

Contributions: I did the miRNA and IMP2 target prediction and the integrative analysis of IMP2 and miRNA data. Moreover, I contributed to data analysis, interpretation and writing of the manuscript. Finally, I supervised Pathmanaban Ramasamy for miRNA expression quantification and differential expression analysis. Marina Wierz prepared all sequence libraries and was involved in data interpretation. Karl Nordström was involved in primary data generation. Sonja Kessler prepared the mice and was invovled in data interpretation and paper writing. Martin Simon and Marcel Schulz were involved in data analysis and paper writing. The content of this chapter has been published at the Molecular Basis of Disease journal (Dehghani Amirabad et al., 2018).

4.1 Introduction

The insulin-like growth factor (IGF) 2 mRNA binding protein 2 (IGF2BP2 or IMP2) belongs to a conserved family of oncofetal proteins, which are usually expressed dominantly during embryogenesis but are associated with cancer as well. According to their naming, this protein binds to the IGF2 mRNA, thus enhancing its translation efficiency (Dai et al., 2011). IGF2 itself has structural similarity to insulin thus representing an important growth factor, whose overexpression was shown to be associated with poor prognosis in several cancers [2]. Consequently, studying the mechanisms of IMPs as a major control element for IGF2 and other genes may provide valuable insights into cancer development.

Although the naming of these IMPs implies specific binding to the IGF2 mRNA, all IMP family members target many more transcripts (Bell et al., 2013) . Analyzing IMP bound RNAs by PAR-CLIP revealed thousands of IMP2 targets in HEK cells (Hafner et al., 2010) and glioblastoma cells (Degrauwe et al., 2016). However, a clear picture of IMP targets in different cell types is missing as well as individual regulatory mechanisms.

Known IMP functions include effects on RNA stability, translational efficiency, localization, and transport (Hansen et al., 2004; Runge et al., 2000; Nielsen et al., 1999) by acting through three pairs of RNA-binding domains: RRM1–2, KH1–2, and KH3–4. Furthermore, IMP2 was described to associate with other proteins, i.e. eIF-4E, to initiate translation of its target RNAs (Bell et al., 2013). Interestingly, interactions of IMP2 with other proteins independent of any RNA fate have been described (Ren et al., 2015; Janiszewska et al., 2012).

An interesting mechanism linking miRNA function to IMPs was first shown in colorectal cancer cell lines: the stabilizing effect of IMP1 to the BTRC mRNA was absent in Dicer knockout cells (Noubissi et al., 2006). Elcheva et al. showed that this is due to counteraction of IMP1 and miRNA-183 in the coding region of the mRNA (Elcheva et al., 2009). In the same region TrCP1 mRNA associates only with Ago2, which is the only slicing Argonaute in humans (Nielsen, Gloggnitzer, and Martinez, 2009; Elcheva et al., 2009). More recently, another link between miRNAs and IMPs was reported in glioblastoma cells, which express the let-7 miRNA family apparently without inhibitory effect on their target genes, due to a transcriptome-wide competition effect of IMP2 and let-7 miRNAs (Degrauwe et al., 2016). Interestingly, IMP2 competition was not limited to let-7 targets, although these show the highest fold changes in response to IMP2 overexpression. It remains to be discovered to which extent such competition between miRNAs and IMPs contributes to transcriptome-wide alterations and disease phenotypes in other systems and models. Enhanced CLIP suggested for instance that IMP2 binding sites are enriched in 3-UTRs of coding genes in pluripotent stem cells (Conway et al., 2016). However, a complete omics point of view to mRNAs and miRNAs is missing for most IMP models.

The IMP2 splice variant IGF2BP2-2/IMP2-2 (also called p62), was originally identified as an autoantigen overexpressed in hepatocellular carcinoma, HCC (Zhang et al., 1999; Kessler et al., 2013; Kessler et al., 2015). IMP2-2 lacks exon 10 of the IMP2 gene, which does not affect the six characteristic RNA binding motifs (Christiansen, Kolte, and Nielsen, 2009). Overexpression of IMP2-2 has been shown to induce steatosis in mice (Tybl et al., 2011; Kessler et al., 2013; Laggai et al., 2014) and to promote progression of non-alcoholic steatohepatitis (NASH) (Simon et al., 2014). As IMP2-2 overexpression was also shown to promote hepatocarcinogenesis (Kessler et al., 2013; Kessler et al., 2015) the underlying molecular reasons for steatosis and carcinogenesis remain unknown.

To get an insight into the molecular mechanisms of this important RNA binding protein, we follow the hypothesis, that IMP2 may not simply bind translated transcripts but may moreover disturb miRNA function and thus RNA interference. We therefore analyze IMP2-2 overexpressing mouse livers for their alterations in the miRNAome and transcriptome. To allow for interpretation of functional associations, integrative analyses were carried out to compare the levels of miRNAs and their targets in IMP2 and WT livers. The full transcriptome (mRNA and long non coding RNA) in combination with the miRNAome of IMP2 induced steatosis gives new insights into the general mechanism of RNA binding proteins in this model system.

4.2 Methods

4.2.1 Transgenic animals

IMP2-2 transgenic mice were established as described by transgenic overexpression of the specific IMP2-2 transcript variant (cDNA) (Tybl et al., 2011). Euthanization of wild-type and IMP2-2 transgenic animals was carried out in week 5. Animal procedures were in accordance with the local animal welfare committee. Mice were kept with a 12h day/night rhythm and under stable humidity, temperature, and food supply.

4.2.2 RNA extraction, cDNA library generation, and Illumina sequencing

Total RNA was isolated from snap frozen livers tissue by immediate lysis in TriReagent (Sigma-Aldrich, Seelze, Germany). Because of the fatty composition of the tissue, twice the amount of TriReagent was used. After additional purification with acid phenol, the resulting RNA was additionally digested with DNAseI (Invitrogen, Karlsruhe, Germany) and again purified by acid phenol. After integrity check using the Agilent Bioanalyzer 2100, directional cDNA libraries were prepared of poly-A enriched RNA using the NEB Next Poly(A) magnetic isolation module and the NEBNext Ultra directional RNA library Prep Kit using 10 PCR cycles. Small RNA libraries were prepared as described before (Götz et al., 2016) by extraction of small RNAs (17˜30 nt) from denaturating PAGE and subsequent library preparation using the NEB Next-small RNA library preparation Kit involving overnight ligation to the 3-preadenylated adapter and 5-monophosphate dependent 5-ligation. PCR amplification was done using 10 PCR cycles and products were purified by gel extraction. Sequencing was carried out on the Illumina HiSeq2500 using high output mode for long RNA libraries (100 nt single end) and Rapid mode for miRNA libraries (30 nt, single end). Reads were demultiplexed with bcl2fastq (v1.8.4) and long RNAs were trimmed for adaptor contamination and low quality bases with the cutadapt (v1.4.1) wrapper trim galore (v0.3.3).

4.2.3 Gene expression analyses

Transcript isoform expression levels for each individual library were quantified using Sailfish an alignment free quantification algorithm (version 0.9.2) (Patro, Mount, and Kingsford, 2014). The transcript expression levels of the same gene are summarized to give a gene expression estimate. Gene ontology (GO) enrichment was done via Ontologizer (Bauer et al., 2008). Thereby, the differentially expressed genes were categorized into different GO terms. Those are concepts to describe gene molecular functions or biological processes. The GO annotation file was obtained from the Gene Ontology Consortium webpage(`http://geneontology.org`). Gene enrichment was determined for up- and down-regulated genes using the Parent-Child-Union method (Bauer et al., 2008).

4.2.4 qPCR

Quantitative PCR was performed as described in detail previously (Laggai et al., 2014)

4.2.5 Prediction of IGF2BP2/p62 binding partners

We used the CatRapid omics webserver for the prediction of RNAs that are bound by IGF2BP2-2/IMP2-2/p62 (Agostini et al., 2013). The mRNA sequence of the human IMP2-2/p62 transcript (ID: ENST00000346192) was obtained from the ENSEMBL database (version 88, assembly: GRCh38.p10) and used as input for CatRapid omics. Mouse mRNAs with an IMP2-2 interaction score were obtained and summarized at the gene level, taking the highest CatRapid score of all mRNAs for a gene. We defined different categories of IMP2 binding using the Star rating score of CatRapid. Using a score > 2.5 is denotated as IMP2+, a score > 2.75 is denotated as IMP2++.

In addition to the sequence based prediction of IMP2 target genes, we used eCLIP RNA-binding data from the ENCODE consortium (Consortium, 2012) for HepG2 liver cancer cells and K562 cells. eCLIP peaks were obtained from the ENCODE data portal (`https://www.encodeproject.org/`). Each annotated human gene in the Ensembl database, that had an eCLIP peak in at least one of the two cell lines, was denoted as a IMP2 target gene. All IMP2 target genes in human determined in this way were transferred to orthologous mouse genes according to Ensembl version 87. In this way a mouse IMP2 target gene set was determined.

4.2.6 Inferring miRNA regulatory effects

In order to predict the global regulatory effect of miRNAs in WT and IMP2 livers, we built linear models for individual samples to explain the expression levels of differentially expressed genes using miRNA expression levels. The overall procedure is illustrated in Figure 4.1.

In order to build the input feature matrix for the model, we retrieved sequence-based predicted miRNA targets in 3-UTRs and CDS regions of genes from the TargetScanMouse 7.1 (Agarwal et al., 2015a) and DIANA 5.0 (Paraskevopoulou et al., 2013) databases. In order to decrease the amount of missing predictions (false negatives), we merged predictions of these two databases for each gene. Then a feature matrix for each sample was built as follows: rows of the matrix denote DEGs, columns denote miRNAs that have a predicted binding site in at least one DEG. Every entry of the matrix represents the expression value of miRNAs that have a predicted binding site in a gene. The input feature matrices were log2 transformed and scaled before regression. We used the Elastic net algorithm (Zou and Hastie, 2005) from the glmnet package (Zhang and Heusdens, 2012) to learn linear regression models for each sample.

The objective function for Elastic net regression is defined as follow:

$$\mathrm{argmin}_{\beta}||y - \beta^t X||^2 + \lambda[(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1] \tag{4.1}$$

where y is the vector of DEG expression values. β is the miRNA regression coefficient vector and X is the input feature matrix. Model selection is done by 6-fold nested-cross validation on hold out datasets. Mean value for miRNA regulatory influence of DEGs for WT and IMP2 livers was obtained by computing the mean value for each miRNA over all regression vectors (Ws) obtained from the models built for each set of samples, respectively. Then a t-test was conducted for each miRNA to assess, whether a regression coefficient showed significant deviation between WT and IMP2 livers (FDR$\leq$ 0.05).

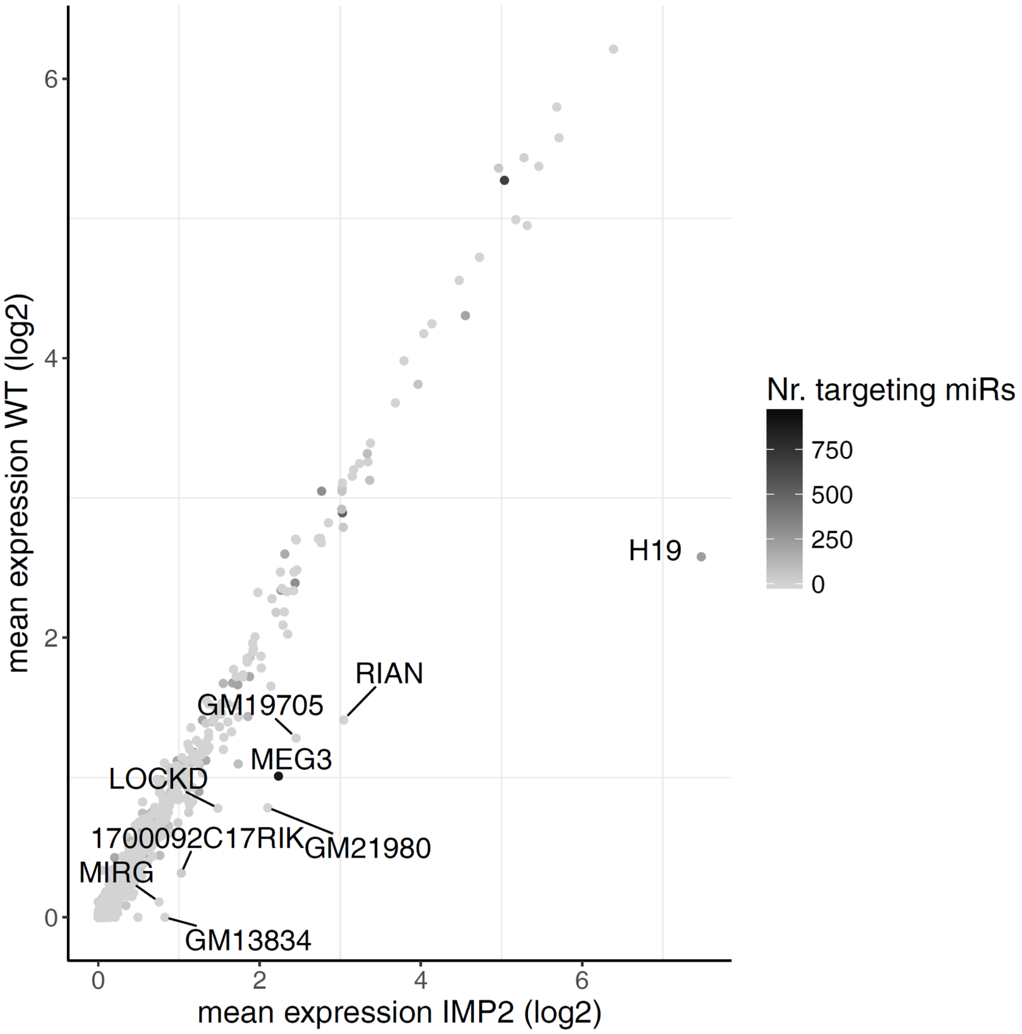


FIGURE 4.1: Detailed comparison of differentially expressed lncRNAs (FDR≤ 0.01). Scatterplot of mean expression values of lncRNAs in IMP2 overexpressing livers (x-axis) and wild type (y-axis) samples. Color scaling of elements indicates the number of expressed microRNAs that are predicted to target a gene (lightgrey: target of few exp. miRNAs, black: target of many exp. miRNAs). The top lncRNAs with the largest fold changes are labeled in the plot. H19, RIAN, Meg3, MiRG, Gm19705, Gm21980, Gm13834, 1700092C17RIK.

4.3 Results

4.3.1 Experimental setup

Five week-old wildtype (WT) and IMP2-2 transgenic (IMP2, p62) mice, which showed liver-specific overexpression of IMP2-2, were used for all analyses. As previously described, IMP2 mice developed a steatosis-like phenotype as indicated by Scharlach Red staining(Tybl et al., 2011; Simon et al., 2014) highlighting lipid droplets (Figure 4.2 A). Expression of the transgene was confirmed in all six IMP2 mice (Figure 4.2B).

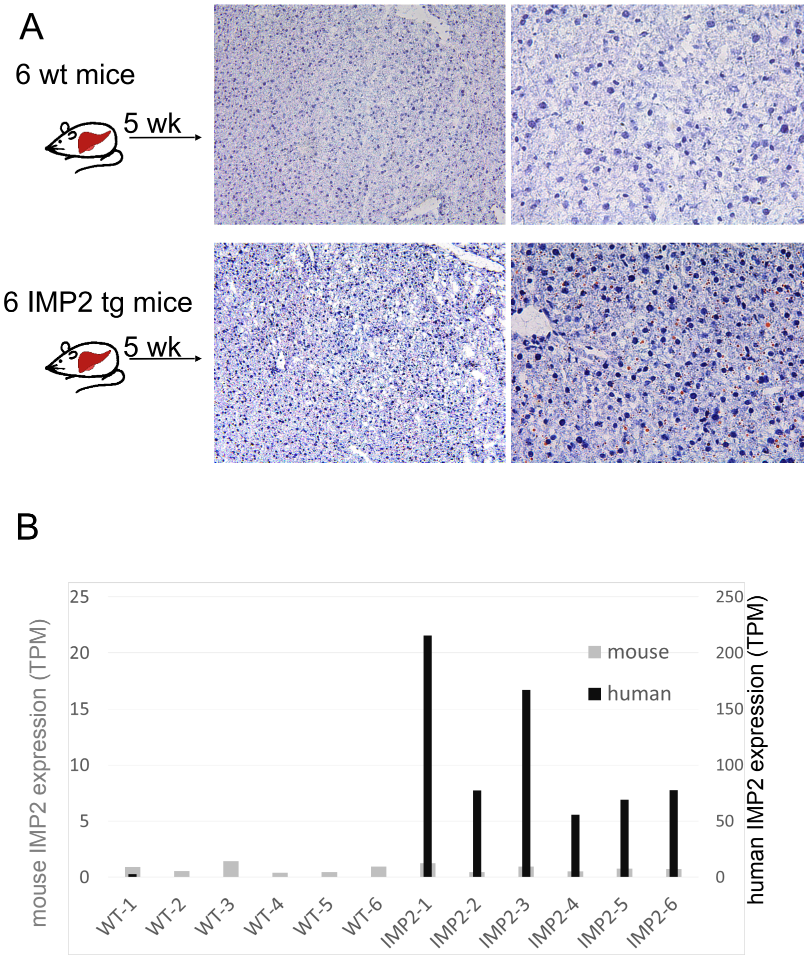


FIGURE 4.2: Experimental setup (A) Wildtype (wt) and IMP2 transgenic (IMP2 tg) mice were sacrificed at the age of 5 weeks. The figure shows a representative lipid staining of wt and IMP2 tg livers. Lipids were stained with Scharlach Red on cryo sections. Original magnification: 200 (left) and 500 (right). (B) IMP2 expression in the livers of all animals shown as individual TPM values for the endogenous mouse IMP2-2 (grey) and the transgenic human IMP2-2 (black). Note the different scales.

4.3.2 Characterization of differential gene expression indicates discrete differences

To characterize the different transcriptomes of WT and IMP2 induced fatty livers, poly(A)-RNA was isolated and libraries for deep sequencing were prepared. Bioinformatics analysis revealed that from 32,360 unique genes of the mouse annotation, 9434 genes had no read count. Within the group of expressed genes, our analysis shows 1801 genes to be deferentially expressed using a false discovery rate (FDR)$\leq$ 0.01. Within those, we identified the majority to be up-regulated (1140 genes) and only 661 genes were significantly down-regulated. Individual genes and lncRNAs have been validated by qPCR (Figure 4.3) supporting our bioinformatics analysis. The MA-plot in Figure 4.4A illustrates that the latter genes showed rather low fold-changes compared to WT, whereas the significantly up-regulated genes show a higher degree of regulation.

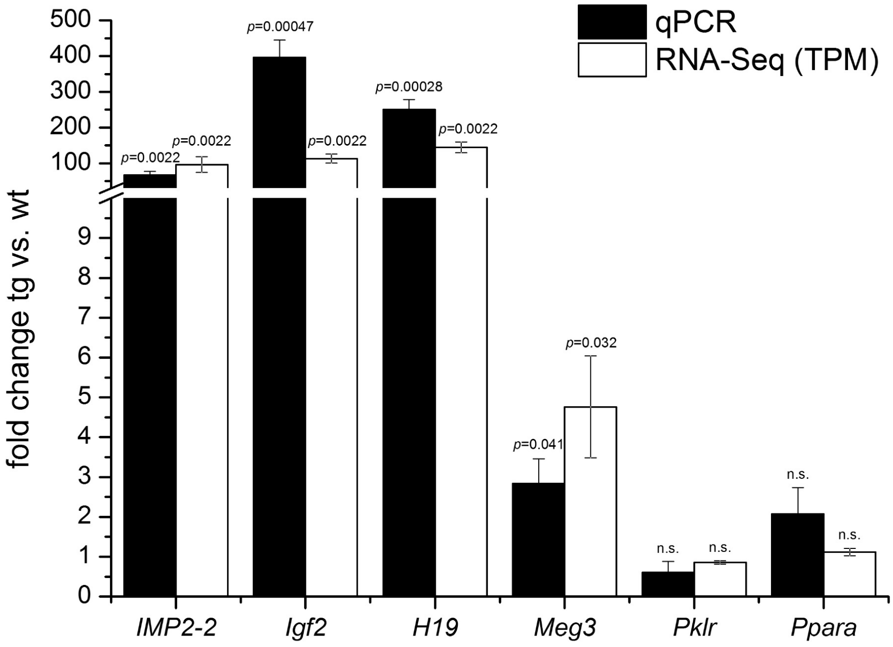


FIGURE 4.3: Comparison between gene expression determined by qPCR and by TMP normalization of NGS data. Expression fold changes of IMP2-2, Igf2, H19, Meg3, Pklr, and Ppara in IMP2 overexpressing livers normalized to wild type mice determined by qPCR and RNA-Seq are shown. *p*-values were evaluated by Student's t-test or Mann-Whitney U test depending on normal distribution of the data.

Comparing the two groups of WT and IMP2 livers, PCA analysis of the gene expression data indicated a common response of the liver tissue to IMP2 overexpression. A separate grouping of sample 1 3 within the transgenic livers was apparent, which correlates with a higher IMP2 expression level in these samples compared to other fatty livers (Figure 4.2B). Figure 4.5 shows, that the different IMP2-2 levels indeed have an effect on individual but not all Igf2 transcript isoforms.

4.3.3 IMP2 overexpression leads to stabilization of thousands of mRNAs

We sought to investigate the effect of IMP2 overexpression on the global gene expression program. It has been suggested that IMP2 binds to target mRNAs and leads to increased stabilization (Bell et al., 2013). We predicted target mRNAs of IMP2 using the following strategy. First, we obtained from the CatRapidOmics webserver human IMP2 (IMP2-2) binding predictions for 42,951 transcripts of 18,081 mouse Ensembl genes. Many of these binding sites had low probability of binding, as indicated by a low CatRapid score and were not considered for analysis. We noted that higher mean expression of genes after IMP2 overexpression leads to higher CatRapid scores suggesting an increased probability of IMP2 binding (Figure 4.4C). Therefore, we asked if there is a change in expression of IMP2 targets after IMP2 overexpression. For this we included publically available IMP2 CLIP data from human HepG2 and K562 cells and transferred these to mouse genes (see Methods). Then gene subgroups were dissected into genes with or without IMP2 binding, and the total of all expressed genes. Figure 4.4D shows that genes bound by IMP2 show significantly higher expression after IMP2 overexpression. Non-target genes (IMP2-) showed lower fold changes than the gene average. Similar results can be seen in

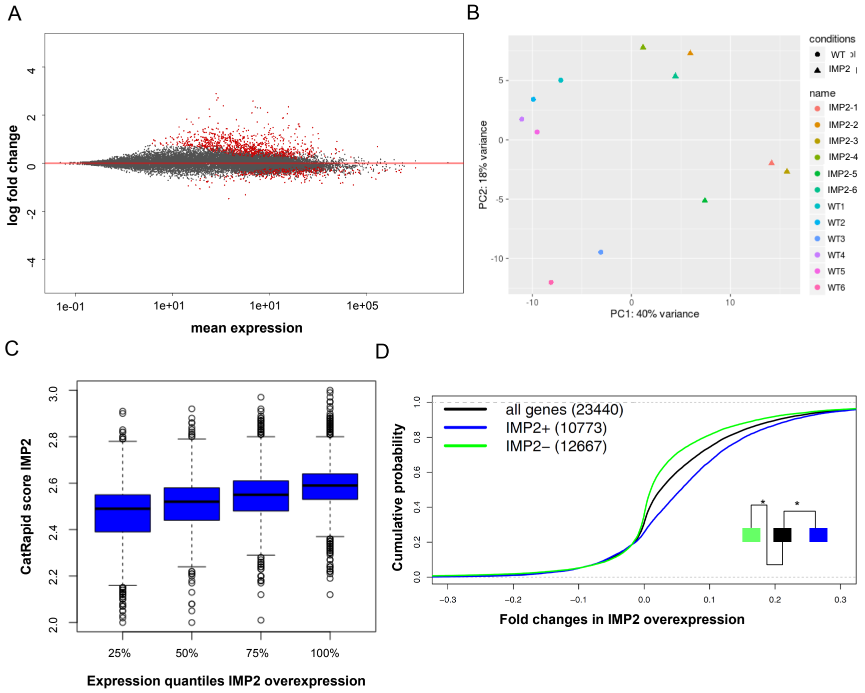


FIGURE 4.4: Differential gene expression in Wild Type and IMP2 induced fatty livers (A) MA plots of fold change expression levels (y-axis) against expression level (x-axis). Each point represents a gene, significantly expressed genes (FDR$\leq$ 0.01) are indicated in red. (B) Principal component analysis using two components (PC1 and PC2). Each dot represents an individual liver sample. The legend is shown on the right side. Difference in variance between the samples can be ascertained by the physical difference between circles on the plot. (C) Genes with predicted IMP2 binding where partitioned into 4 quantiles based on mean expression in IMP2 samples (x-axis). Boxplots of CatRapid scores (y-axis) are shown for each quantile. (D) Cumulative plot of gene expression (log2 fold change) in IMP2-2 overexpressing livers. Gene subgroups were built according to IMP2 binding determined by IMP2 CLIP data (see Methods) and representing all genes (black), IMP2 positive genes (IMP2+, blue) and IMP2 negative genes (IMP2-, green). Significant differences according to a Kolmogorov-Smirnoff test (p-value $\leq$ 0.05) are marked with (*). Numbers in brackets denote the number of genes in each category.

Suppl. Figure 4.6 analyzing altered gene expression of gene groups with positive and negative IMP2 binding prediction according to CatRapid. These analyses suggest that overexpressed IMP2 stabilizes thousands of mRNAs.

Since mRNA stability is regulated by miRNAs, we sought to investigate the effect of IMP2 overexpression on miRNAs and miRNA target genes. Figure 4.7 shows a scatter plot of mean expression values of all DEGs as well as their IMP2 binding prediction in combination with the number of miRNAs targeting an individual transcript. The most up-regulated gene was IGF2, which is a known target of IMP2. However, the fold change of IGF2 was much higher than the majority of genes that are predicted IMP2 targets according to CatRapid. Interestingly, IGF2 mRNA is also a target of 96 expressed miRNAs, which suggests that the high degree of up-regulation may not be solely due to IMP2 mediated stabilization (see below). According to Figure 4.7, the transcript with the second highest fold change was IMP2 itself.

Strikingly, the six genes with the highest up-regulation (Figure 4.7) have been

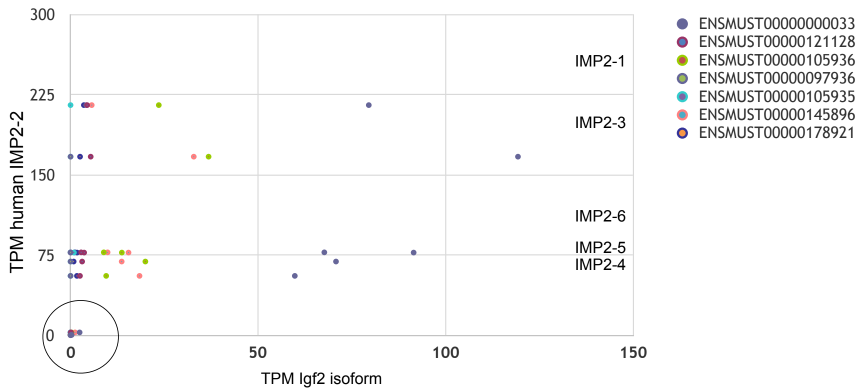


FIGURE 4.5: Scatter plot of TPM values for seven Igf2 isoforms and IMP2-2 (human transgenic isoform) of the twelve individual samples. All wild type samples show no or low expression and are summarized by the black circle. The individual IMP2-2 transgenic lines are indicated on the right.

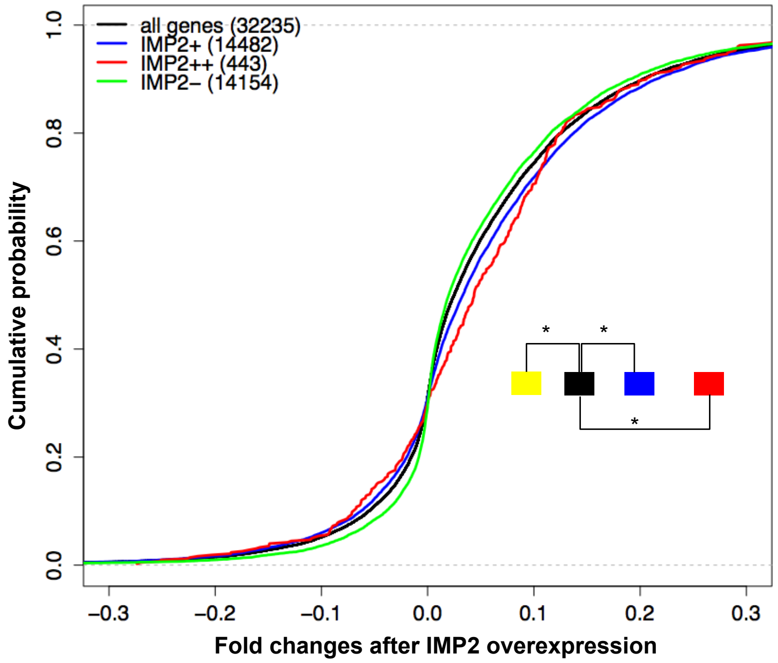


FIGURE 4.6: Cumulative distributions of log2 fold changes (IMP2/WT) after IMP2 overexpression (x-axis) are compared for different subsets of the data: all genes (black), genes with CatRapid score > 2.5 (IMP2+, blue) or > 2.75 (IMP2++, red) or no CatRapid prediction (IMP2-, green). Significant differences according to Kolmogorov-Smirnoff test (p-value$\leq$ 0.05) are marked with (*). Numbers in brackets denote the number of genes in each category.

described to be associated in HCC (including IMP2 and IGF2): the alpha-1 fetoprotein (AFP), which is expressed usually in the fetal liver as well as in cirrhosis and HCC (Zucman-Rossi et al., 2015); member C18 of the aldo-keto reductase family 1 (AKR1C18), a component of the lipid metabolism, which was shown to be strongly up-regulated in HCC (Ohrnberger et al., 2015), glypican-3 (GPC3) which controls cell division being highly up-regulated in HCC (Suzuki et al., 2010), and serine protease inhibitor Kazal-type 1, also known as tumor associated trypsin inhibitor, SPINK1, which is also overexpressed in HCC (Marshall et al., 2013). AFP is routinely used as a diagnostic marker for HCC and a combined usage of AFP and GPC3 to improve diagnosis has been suggested (Capurro et al., 2003).

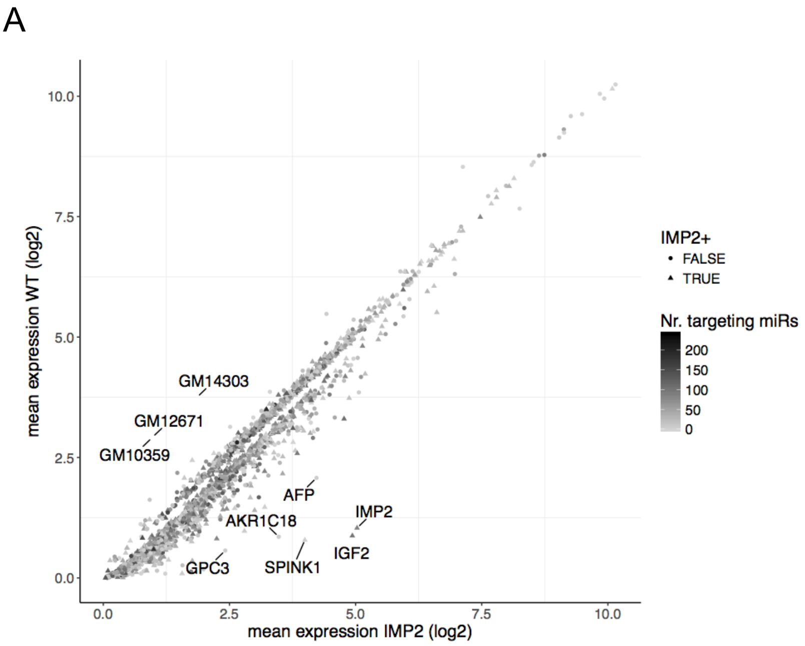


FIGURE 4.7: Detailed comparison of DEGs (FDR≤ 0.01). Scatterplot of mean expression values of DEGs in IMP2 overexpressing livers (x-axis) and wild type (y-axis) samples. Genes that are categorized as IMP2+ target genes according to CatRapid are denoted with a triangle. Color scaling of elements indicates the number of expressed microRNAs that are predicted to target a gene (lightgrey: target of few exp. miRNAs, black: target of many exp. miRNAs). The top 10 genes with the largest fold changes are labeled in the plot. IGF2, IMP2, SPINK1 (serine peptidase inhibitor, Kazal type1), AFP (alpha fetoprotein), AKR1C18 (Aldo-keto reductase family 1, member C18), GPC3 (Glypican 3), Gm12671 (predicted gene 12,671), Gm10359 (predicted gene 10,359), Gm14303 (predicted gene 14,303).

4.3.4 GO enrichment indicates activation of immune processes and cell cycle processes

To identify altered biological processes in IMP2 overexpressing livers, GO enrichment of DEGs was carried out. Figure 4.8 shows a graphical visualization of significantly altered biological processes in fatty livers (FDR≤ 0.01).

Figure 4.8A shows that GO terms enriched with up-regulated genes were dominated by terms correlated with cell divisions: "cell cycle", "cell cycle process", "chromosome segregation", and "organelle fission", thus indicating that mitotic divisions could occur more frequently in fatty liver tissue. However, IMP2 transgenic livers do usually not show increased weight compared to WT, but an increased liver to body weight ratio (Tybl et al., 2011). Other reports observed an altered cell cycle in steatosis models (Cui, Chen, and Hu, 2010). Interestingly, in pancreatic islet cells cell cycle highly correlates with IMP1 expression defining diabetes susceptibility (Keller et al., 2008).

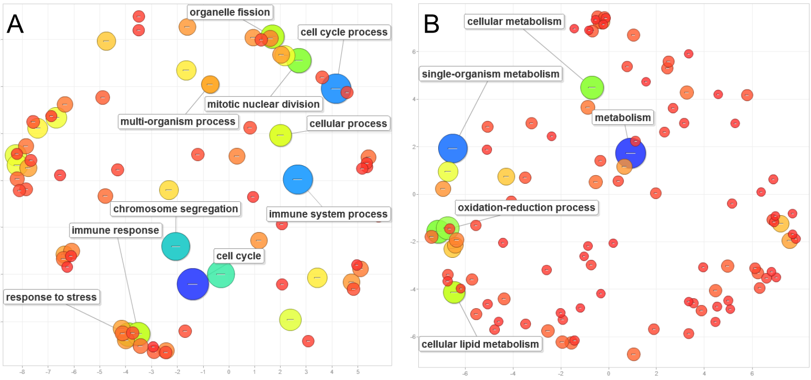


FIGURE 4.8: GO-enrichment of up-regulated (A) and down-regulated (B) GO terms. Plots are shown of GO representatives for biological processes generated by the REVIGO tool which summarizes the list of GO terms derived from differentially expressed genes (FDR$\leq$ 0.01) by removing redundant terms. The distance between circles (representing individual GO terms) indicates the relationship between terms: smaller distance means more similar relationships. Bubble color indicates significance of differential expression of an individual GO term (red low and blue high); the size (in log10 p-value) indicates the percentage of genes annotated with a term in the reference database (UniProt) and thus indicates more general terms (large) and more specific ones (small).

Also processes correlated to the immune system can be found among the up-regulated genes and the GO terms "immune system process" and "immune response" represented the second dominant group. As increased lipid content is often accompanied by inflammatory events in NASH (Day, 2006), and IMP2 overexpressing mice show a slightly inflammatory phenotype (Laggai et al., 2014), association with these processes makes sense and is further supported by the autoantigenic character of the transgene.

Among the down-regulated DEGs, enriched GO terms were related to energy metabolism and metabolism, such as "cellular metabolism", "metabolism" and "oxidation-reduction process" suggesting that the general energy turn-over and metabolic rate is lower in fatty livers compared to WT-liver-samples (Figure 4.8B). Also the process of "cellular lipid metabolism" was significantly down-regulated. In this context, the IMP2 overexpression phenotype has been shown to include not only the accumulation of lipids but also their different composition, as for instance the C18:C16 fatty acid ratio becomes significantly elevated(Laggai et al., 2013). This is in agreement with altered lipid compositions in human NAFLD (Puri et al., 2009; Puri et al., 2007). Still, down-regulated DEGs of the process "cellular lipid metabolism" mostly included genes leading to a reduction of lipids, such as enzymes of the fatty acid beta-oxidation or inhibitory enzymes of lipid synthesis, in turn resulting in hepatic lipid accumulation.

4.3.5 Genome-wide miRNA analysis shows global dysregulation

In addition to the DEG analysis, we also extracted small RNA fractions of the same RNA samples from which poly(A) RNA was prepared and subsequently analysed miRNAs by deep sequencing. Figure 4.8 reveals the global analysis of differentially regulated miRNAs. The MA-plot shows that only few miRNAs were significantly differentially expressed between WT and IMP2 induced fatty livers (Figure 4.9A).

Although the PCA analysis in Figure 4.9B shows that WT and IMP2 livers could be clearly separated based on their miRNA expression, a higher divergence of the samples becomes apparent as expected from the clustering of transcriptomes (Figure 4.4B).

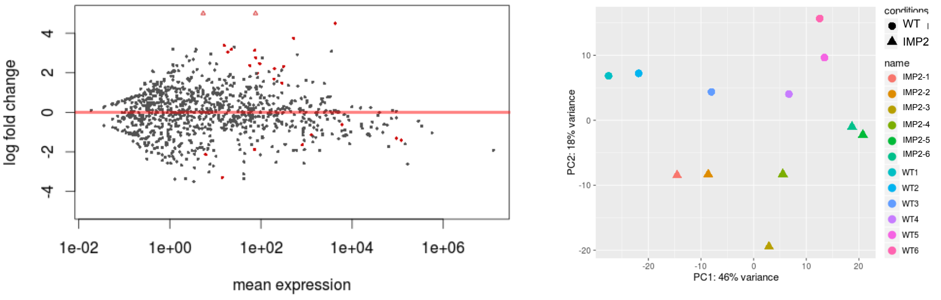
Going into detail which miRNAs are significantly expressed, Figure 4.9C lists up- and down-regulated miRNAs according to their significance indicated by the adjusted p-value. The figure shows that of a total of 24 significantly regulated miRNAs (FDR$\leq$ 0.01), 16 miRNAs are up-, and 8 miRNAs were down-regulated.

The two most significantly regulated miRNAs are miR151-5p and miR483-3p. The latter is encoded by the IGF2 mRNA, which we know to be up-regulated by IMP2 overexpression. We can consequently see that both strands of this miRNA accumulated in the IMP2 samples (Figure 4.9C).

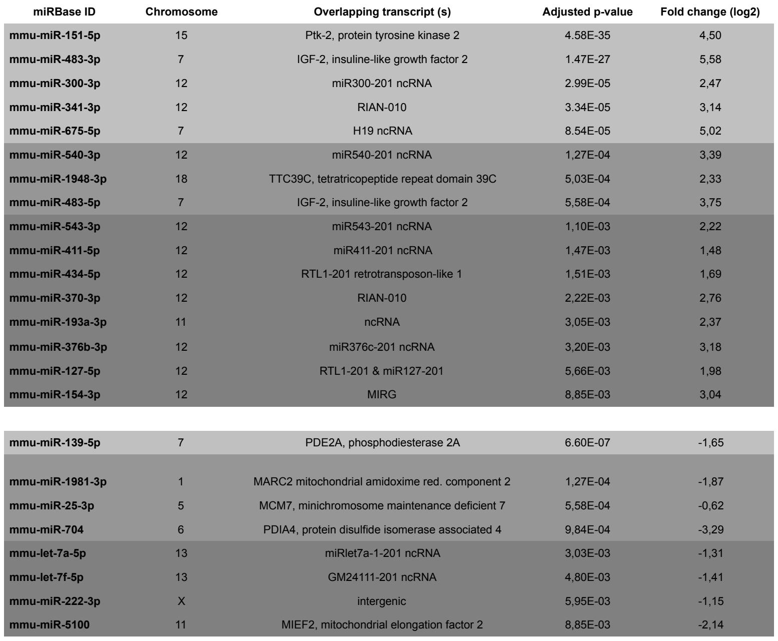
In the context of IGF2 and miR483 up-regulation, also miR675, which is processed from the H19 RNA was significantly accumulated (Figure 4.9C). Both, the H19 locus, encoding for a lncRNA, and the IGF2 locus are linked imprinted loci showing reciprocal monoallelic expression: H19 from the maternal, IGF2 from the paternal allele (Zemel, Bartolomei, and Tilghman, 1992). To clarify whether also the miR675 up-regulation could be due to increased precursor levels, genome wide lncRNA quantification was carried out comparing the IMP2 and the WT livers. Figure 4.1 shows that many lncRNAs are upregulated in IMP2-2 livers: these include H19 as well as three lncRNAs belonging to the DLK1/DIO3 domain. Therefore, the up-regulation of miR675, which is processed from the H19 transcript, is likely due to increased template RNA. Figure 4.9C also reveals that 10 out of 16 up-regulated miRNAs are located on chromosome 12. It is noteworthy that all of them are located in the DLK1/DIO3 domain, which contains a large and highly conserved miRNA cluster (Seitz et al., 2004). In agreement with the upregulation of three of its lncRNAs (Meg3, Rian, MiRG) we see nearly all of the encoded miRNAs enriched in IMP2-2 livers (Figure 4.10).

4.3.6 IMP2 alters miRNA regulation and stabilizes miRNA targeted transcripts

Our paired miRNA expression data allowed us to conduct integrative analyses to measure associations of miRNA expression and their potential target genes in WT and IMP2 samples. We developed an analysis framework that allowed us to quantify associations of miRNAs to DEGs. Briefly, we build a linear regression model for each WT and IMP2 sample for the prediction of DEG expression (Figure 4.1). The features for the regression are miRNA expression levels of miRNAs that target a gene obtained from sequence-based miRNA target predictions, see methods. This analysis allowed us to investigate the regulation of DEGs through changes in miRNA expression after IMP2 overexpression. For each sample we obtained a regression (association) coefficient for each miRNA. Figure 4.11A shows the Top 100 miRNAs with the largest change in mean regression coefficients between WT and IMP2 samples. We observed that miRNA regression coefficients are smaller (less influence on DEG expression) in IMP2 samples compared to WT samples. In total 8 miRNAs (t-test FDR$\leq$ 0.05) showed a significant difference in their mean regression coefficients. This suggests that expressed miRNAs have reduced regulatory influence on DEGs in IMP2 samples. We followed the possibility that IMP2 binding may interfere with genome-wide miRNA binding to transcripts as observed by (Degrauwe et al., 2016) in human glioblastoma stem cells. First, we arbitrarily partitioned all genes into four groups according to the amount of expressed miRNAs that are targeting a



C



miRBase ID	Chromosome	Overlapping transcript (s)	Adjusted p-value	Fold change (log2)
mmu-miR-151-5p	15	Ptk-2, protein tyrosine kinase 2	4.58E-35	4,50
mmu-miR-483-3p	7	IGF-2, insuline-like growth factor 2	1.47E-27	5,58
mmu-miR-300-3p	12	miR300-201 ncRNA	2.99E-05	2,47
mmu-miR-341-3p	12	RIAN-010	3.34E-05	3,14
mmu-miR-675-5p	7	H19 ncRNA	8.54E-05	5,02
mmu-miR-540-3p	12	miR540-201 ncRNA	1,27E-04	3,39
mmu-miR-1948-3p	18	TTC39C, tetratricopeptide repeat domain 39C	5,03E-04	2,33
mmu-miR-483-5p	7	IGF-2, insuline-like growth factor 2	5,58E-04	3,75
mmu-miR-543-3p	12	miR543-201 ncRNA	1,10E-03	2,22
mmu-miR-411-5p	12	miR411-201 ncRNA	1,47E-03	1,48
mmu-miR-434-5p	12	RTL1-201 retrotransposon-like 1	1,51E-03	1,69
mmu-miR-370-3p	12	RIAN-010	2,22E-03	2,76
mmu-miR-193a-3p	11	ncRNA	3,05E-03	2,37
mmu-miR-376b-3p	12	miR376c-201 ncRNA	3,20E-03	3,18
mmu-miR-127-5p	12	RTL1-201 & miR127-201	5,66E-03	1,98
mmu-miR-154-3p	12	MIRG	8,85E-03	3,04
mmu-miR-139-5p	7	PDE2A, phosphodiesterase 2A	6.60E-07	-1,65
mmu-miR-1981-3p	1	MARC2 mitochondrial amidoxime red. component 2	1,27E-04	-1,87
mmu-miR-25-3p	5	MCM7, minichromosome maintenance deficient 7	5,58E-04	-0,62
mmu-miR-704	6	PDIA4, protein disulfide isomerase associated 4	9,84E-04	-3,29
mmu-let-7a-5p	13	miRlet7a-1-201 ncRNA	3,03E-03	-1,31
mmu-let-7f-5p	13	GM24111-201 ncRNA	4,80E-03	-1,41
mmu-miR-222-3p	X	intergenic	5,95E-03	-1,15
mmu-miR-5100	11	MIEF2, mitochondrial elongation factor 2	8,85E-03	-2,14

FIGURE 4.9: GO-enrichment of up-regulated (A) and down-regulated (B) GO terms. Plots are shown of GO representatives for biological processes generated by the REVIGO tool which summarizes the list of GO terms derived from differentially expressed genes (FDR$\leq$ 0.01) by removing redundant terms. The distance between circles (representing individual GO terms) indicates the relationship between terms: smaller distance means more similar relationships. Bubble color indicates significance of differential expression of an individual GO term (red low and blue high); the size (in log_{10} p-value) indicates the percentage of genes annotated with a term in the reference database (UniProt) and thus indicates more general terms (large) and more specific ones (small).

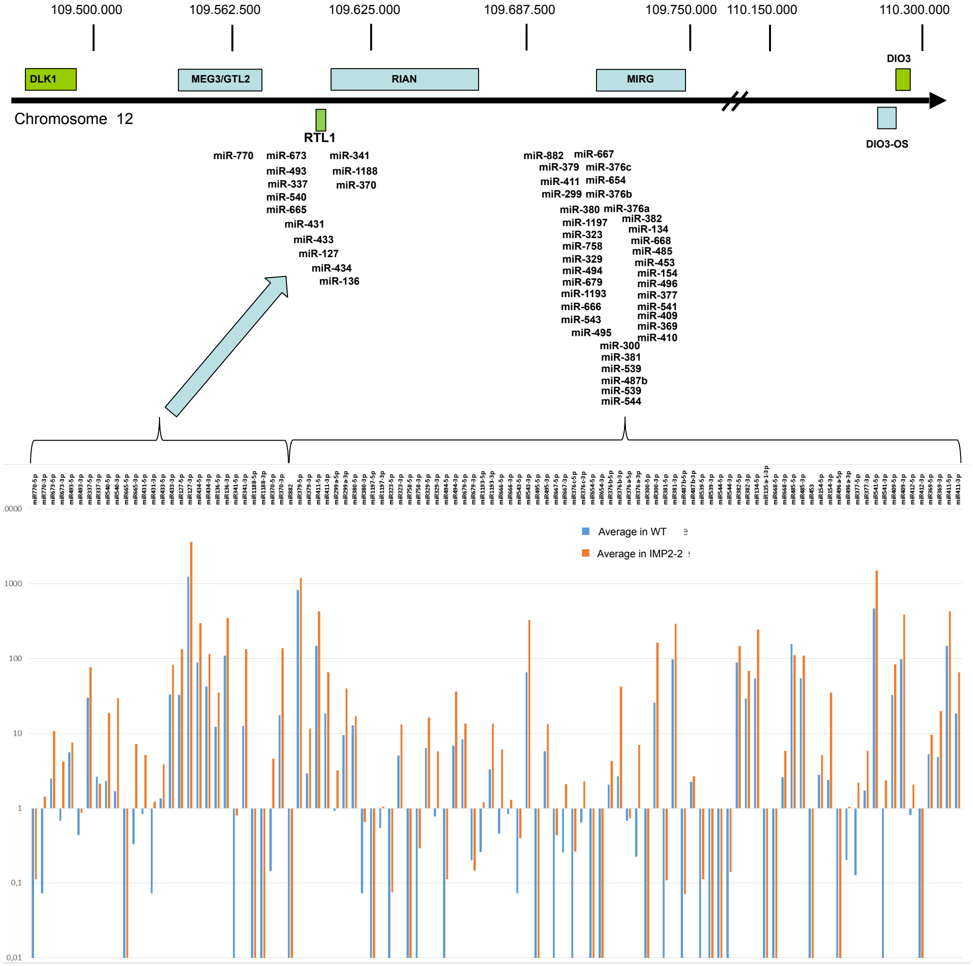


FIGURE 4.10: Regulation of miRNAs of the imprinted DLK1/DIO3 domain in IMP2 overexpressing livers. The scheme shows the murine DLK1/DIO3 domain. Maternally expressed loci are shown in blue, paternally expressed loci in green. Please note that the scheme is not drawn to scale. DLK1: delta like 1 Homolog, GTL2: glycosyltransferase-like2 (MEG3 in human), RTL1: Retrotransposon like, RIAN: RNA imprinted and accumulated in the nucleus, MIRG: miRNA containing gene, DIO3: Type III deionidase, DIO3-OS: Dio3 opposite strand. Below the scheme the approximate position of miRNAs is demonstrated. The graph shows average miRNA expression in WT (blue) and IMP2 livers. Expression of miRNAs with no reads are pseudocounted with 0.01.

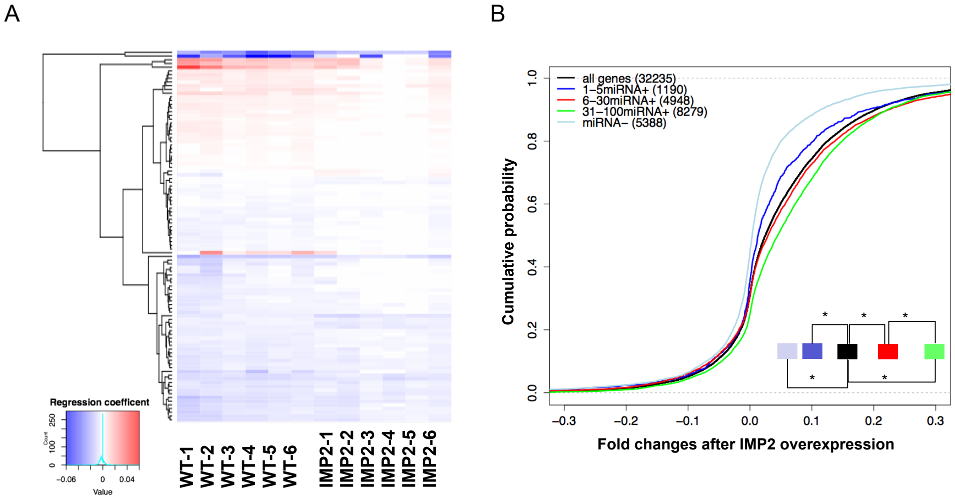


FIGURE 4.11: Integrative analysis of miRNA dysregulation and gene expression. (A) Heatmap of regression coefficients of TOP100 miRNAs (rows) that show a difference in regulation of DEGs. The strength of the miRNA association with all DEGs in a given sample (column) is indicated by a color coding (blue and red, negative and positive association, respectively). (B) Cumulative distributions of log2 fold changes (IMP2/WT) after IMP2 overexpression (x-axis) are compared for different subsets of the data: all genes (black), genes with binding sites for 1–5 expressed miRNAs (blue), for 6–30 expressed miRNAs (red), for 31–100 expressed miRNAs (green) and genes not targeted by any expressed miRNA (light blue). Significant differences according to Kolmogorov-Smirnoff test (p-value $\leq$ 0.05) are marked with (*). Numbers in brackets denote the number of genes in each category.

gene based on TargetScanMouse 7.1 (Agarwal et al., 2015a). Figure 4.11B shows that genes that are targeted by >5 expressed miRNAs (6–30miRNA+ and 31–100miRNA+ categories) show a significant up-regulation (KS-test p-value = 0.0091 and < 2.2e-16 respectively) compared to the whole set of genes. In addition, genes that had binding sites for >30 miRNAs showed a significantly stronger up-regulation (KS-test p = 8.038e-14). In contrast, genes with no predicted binding site for an expressed miRNA (miRNA-) showed the least up-regulation after IMP2 overexpression and significantly less than the average over all genes (KS-test $p < 2.2e - 16$).

To include a potential effect of IMP2-2 to the miRNA attacked transcripts, we combined the miRNA analyses with either the CatRapid IMP2 binding prediction (Figure 4.12) or the IMP2 CLIP data (Figure 4.11A). We stratified genes into several combinations of these two attributes (Figure 4.11C). We observed that genes that are bound by IMP2 and >50 miRNAs (IMP2 + miRNA++) showed the strongest up-regulation after IMP2 overexpression compared to all genes and to genes that are only targeted by >1 miRNA (IMP2-miRNA+). For genes that are not bound by miRNAs we observed reduced fold-changes compared to the majority once IMP2 binds them. To characterize the individual numbers of transcripts, which can be bound by either miRNAs, IMP2 or both, we used set intersection plots (Figure 4.13B). We observed that most of the genes that showed IMP2 binding also had miRNA binding sites. Moreover, genes with binding sites for at least 30 miRNAs were more likely bound by IMP2 (7,304 out of 12,327) suggesting that IMP2 binding and miRNA binding are not independent. These analyses strongly suggest that IMP2 competes with miRNA binding and thus leads to genome-wide stabilization and up-regulation of

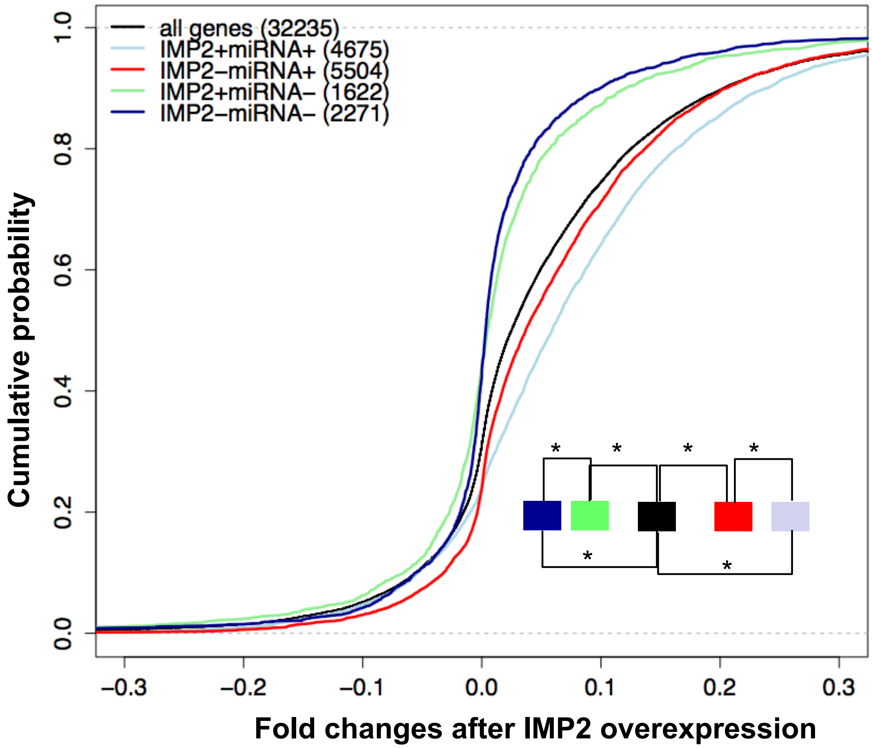


FIGURE 4.12: Cumulative distributions of log2 fold changes (IMP2/WT) after IMP2 overexpression (x-axis) are compared for different subsets of the data: IMP2 targets (IMP2+, CatRapid score > 2.5) or non-targets (IMP2, no CatRapid prediction), genes that have binding sites for 30 or more miRNAs (miRNA+) or have no binding site for an expressed miRNA (miRNA). All combinations of the four attributes are separated in the plot and compared to all genes (black). Significant differences according to Kolmogorov-Smirnoff test (p-value ≤ 0.05) are marked with (*). Numbers in brackets denote the number of genes in each category.

mRNAs. The large number of genes which could be regulated by this competing IMP2/miRNA system may suggest a co-evolution of binding sites.

4.4 Conclusion

It has been well documented that hepatocyte-specific overexpression of the RNA binding protein IMP2 in murine livers leads to a steatosis like phenotype (Kessler et al., 2015; Tybl et al., 2011; Laggai et al., 2014). To analyze the effect of IMP2 on the liver transcriptome and miRNAs, we analysed whole liver tissue samples of five week-old WT and transgenic mice. Although hepatocytes display the predominant hepatic cell type, the analysed Seq-data contained data not only from hepatocytes, but also from non-parenchymal cells such as Kupffer cells, stellate cells, and endothelial cells. Differences in cell composition cannot be excluded.

4.4.1 Genome wide effects of miRNA and IMP2 competition: Just stabilization?

We know several hundred RNA binding proteins in humans, but few of them have been implicated in progression of certain diseases (Lukong et al., 2008). However, precise mechanisms and genome-wide data are rare, especially for the growing understanding of RNA binding and miRNA competition. One exception is the RNA binding protein LIN28, which supresses maturation of the let7 miRNA family by binding to the precursor RNA thus blocking Drosha during development (Viswanathan and Daley, 2010). This is interesting also for our analyses of IMP2 as recent evidence that also IMP2 can prevent let7 silencing, not by inhibition of miRNA biogenesis but by competition with target sites (Degrauwe et al., 2016). In our study, we overexpress IMP2, thus mimicking aberrantly high expression of IMP2 as found in HCC (Zhang et al., 1999; Kessler et al., 2013; Kessler et al., 2015).

In this work, we dissected individual groups of genes by different criteria: their ability to bind miRNAs and their ability to bind IMP2. The latter argument was predicted either by CatRAPID or by the integration of human IMP2 CLIP data. Both analyses clearly show the same result: genes that are preferentially bound by IMP2 and many miRNAs are upregulated in IMP-2 overexpressing livers. We therefore conclude that both, IMP2-2 and miRNAs, compete for transcripts. This means that IMP2-2 overexpression tilts the balance towards IMP2: leading to widespread miRNA inhibition thus lowering the extent of genes regulated by miRNAs.

Similar effects have been shown on glioblastoma cells expressing endogenous up-regulated IMP2, only (Degrauwe et al., 2016). In addition, our data indicate a general decrease of miRNA regulation in IMP2 livers and this occurs for miRNAs with positive as well as negative effects on target abundance. For individual miRNAs, increasing knowledge shows the effect of individual RNA binding proteins to formation and function of the target bound RNA silencing complex (RISC) (Van Kouwenhove, Kedde, and Agami, 2011). Data for IMP2 are missing and our data do not allow for decisions whether competition inhibits miRNA repression or RISC and IMP2 interact to modulate translation. Likely, both happens which will have to be verified individually for each transcript by Ribo-footprinting of IMP2 livers. A non-exhaustive analysis of lipogenic enzymes in IMP2 overexpressing liver cells did not reveal significant changes in their mRNA abundance, although some components showed alterations of the protein expression levels in Western blots (Laggai

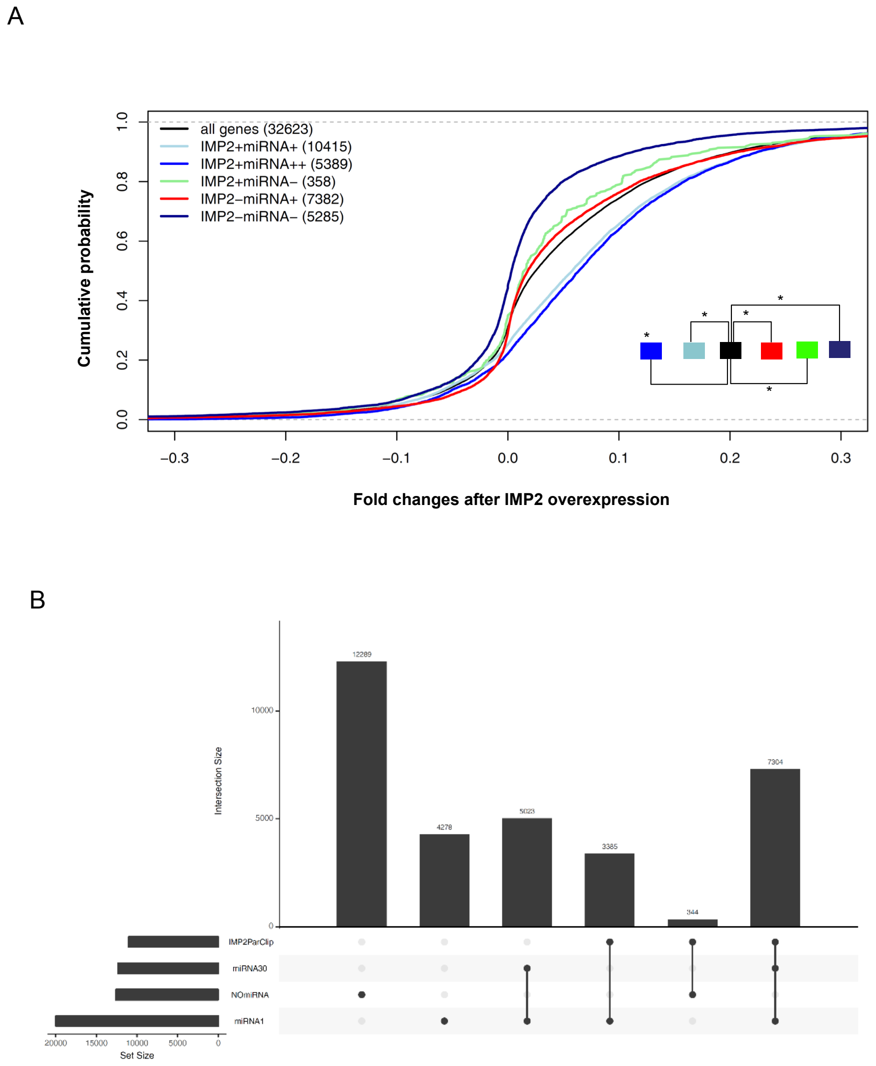


FIGURE 4.13: Integrative analysis of the effects of miRNAs and IMP2-2 to gene expression. (A) Cumulative plot of gene expression (log2 fold changes) in IMP2-2 overexpressing livers. Gene subgroups were dissected first according to IMP2 binding determined by IMP2 CLIP data with positive binding (+) and without () and second for miRNA binding sites, >1 miRNAs binding (miRNA+), $>$ 50 miRNAs binding (miRNA++) and all other genes (miRNA-). All combinations of the five attributes are separated in the plot and compared to all genes (black). Significant differences according to a Kolmogorov-Smirnoff test (p-value $\leq$ 0.05) are marked with (*). Numbers in brackets denote the number of genes in each category. (B) Set intersection plot of four different gene groups. With positive IMP2 CLIP, with >30 miRNA predicted binding sites (miRNA30), with more than one predicted miRNA binding site (miRNA1) and without (NOmiRNA).

et al., 2014). This supports the drawn conclusions that translational alterations of lipogenic genes contribute to the steatosis-like phenotype.

4.4.2 General characteristics of IMP2-2 induced transcriptome and miRNAome

With the identified GO enrichments, some aspects of the microsteatosis phenotype, such as the inflammation and alterations of the lipid metabolism, can be attributed to differential gene expression. Strikingly, the five highest elevated transcripts (IGF2, AFP, GPC3, AKR1C18, SPINK1) in IMP2 overexpressing livers are characteristic markers for HCC (Zucman-Rossi et al., 2015; Ohrnberger et al., 2015; Suzuki et al., 2010; Capurro et al., 2003).

Also in the analysis of miRNAs, an interesting aspect of the genome-wide comparison is that our study of IMP2-induced microsteatosis did not reveal significant changes in miRNAs, which have been previously described in NAFLD or NASH, such as miRNAs 34a, 155, 200b, 214-5p, 221, 29c, 122, 192, 203, and 467b (Ceccarelli et al., 2013). However, two miRNAs intensively studied in HCC, miR483 and miR151 (Callegari et al., 2015), were significantly up-regulated.

Both miRNAs, miR438 and miR151, are processed from the coding mRNAs IGF2 and PTK2, respectively. Our data indicates an up-regulation for the miR483 mRNA precursor only. The IGF2 mRNA becomes highly enriched in IMP2 overexpressing livers in agreement with previous findings (Tybl et al., 2011). Consequently, both strands of miR438 can be found to be enriched (Figure 4.9C). In contrast, the up-regulation of miR151-5p cannot be easily explained by increased abundance of its precursor mRNA (PTK2). Here, we can only speculate whether processing by the RNAi machinery is different or whether the miRNA becomes stabilized to a higher degree by increased occurrence of target sites in the altered transcriptome. In HCC, increased expression of miR151 is either linked to a copy number gain (Ding et al., 2010) or elevated levels of the hosting gene PTK2 (Luedde, 2010). Strikingly, the five most elevated transcripts and most significant miRNAs in IMP2 livers are HCC specific, which could explain the significantly increased risk to develop HCC after exposure to carcinogens in this model (Kessler et al., 2015).

4.4.3 miRNA dysregulation in imprinted loci

Our data show that many up-regulated miRNAs derive from reciprocally imprinted loci: IGF2, H19 and the DLK1/DIO3 domain (Figure 4.8D), raising the questions whether this is due to allele-specific regulation or loss of imprinting. Concerning the IGF2/H19 gene cluster, loss of imprinting was associated with developmental disorders as well as many tumors such as HCC (Kim and Lee, 1997; Li et al., 1997; Reik et al., 2000). However, loss of imprinting was not observed in IMP2-2-induced steatosis by us, instead a common transcriptional regulator Aire may be responsible for the concerted activation (Tybl et al., 2011).

The second imprinted region, which is affected by IMP2 overexpression, is the DLK1/DIO3 domain. Although IMP2 mice were shown to have elevated levels of the paternally expressed protein-coding gene DLK1 (Kessler et al., 2015) loss of imprinting cannot be excluded. A possible connection between both loci could be the H19 lncRNA, which was shown to be a central regulator of an imprinted gene network by binding MDB1 (methyl-CpG binding domain protein 1) (Gabory et al., 2009; Monnier et al., 2013). However, these pioneer studies clearly showed that H19 is a negative regulator for DLK1 during development, which is the opposite in our studies in five week-old animals. Also, alterations in H19 levels did not alter the entire DLK1/DIO3 domain as for instance expression of RIAN was not affected by H19 (Gabory et al., 2009). Clearly, the molecular triggers for regulation of imprinted loci are different in studies depleting H19 (Gabory et al., 2009; Monnier et al., 2013) and overexpressing IMP2 and therefore future studies have to clarify the epigenetic landscape of imprinted loci in IMP2 and H19 transgenic models.

4.4.4 Different aspects contribute to IGF2 mRNA accumulation

Still, the question remains open why exactly the IGF2 mRNA is highly accumulated in IMP2 overexpression lines. One aspect is of course the direct stabilization of the transcript by the RNA binding protein. Second, miR483 is processed from the transcript and our data shows accumulation of its 5p and 3p strand. Especially the increase of miR483-5p should lead to an additional increase of IGF2 transcriptional activity as the nuclear fraction of this miRNA was shown to bind to the 5-UTR of the IGF2 mRNA, thus increasing the transcriptional activity of the host gene by a feed-forward regulatory loop (Liu et al., 2013). Here, this would result in two sequential stabilizing effects i.e. (i) post-transcriptionally by IGF2 mRNA stabilization and (ii) on the transcriptional level increasing IGF2 transcription. As a third aspect, our prediction indicates that the IGF2 mRNA is target of 96 expressed miRNAs. Therefore, competition between IMP2 and miRNAs as predicted by our data could have an even stronger effect on this transcript.

Chapter 5

Multi-task learning of transcript isoform regulatory network

"Gene regulatory network" is a somewhat vague term. We use it to describe the connection between regulators and the target gene. However, the term "gene" by itself is virtual concept. This chapter introduces a new paradigm for gene regulatory network inference. We challenge the concept of a gene regulatory network as a gene is a virtual, but useful concept. We propose isoform regulatory network as an appropriate term for network inference.
Contributions: All method development and data analysis of this project was done by me. This chapter is the extension of our manuscript published at Proceedings of German Conference for Bioinformatics 2016 (Dehghani Amirabad and Schulz, 2016).

5.1 Introduction

The regulatory network describes the connection between regulators (TF, miRNA, RBP, etc.) and the target genes. These networks specify condition-specific expression of a gene. Inferring condition-specific regulatory network is important because of their relevance to cell identity, diseases, and development.

Substantial progress has been made in reconstruction of the regulatory network including regression models (Haury et al., 2012), graphical models (Schäfer and Strimmer, 2005) and Bayesian networks (Friedman et al., 2000). The readers are referred to (Wang and Huang, 2014; Sanguinetti, 2019) for comprehensives reviews of the methods.

Despite considerable progress in the field, regulatory network inference and analyses have yet been limited to inference from total gene expression levels in a sample cohort. Gene expression denotes the summed activity of all isoforms of a gene, and a gene may have a different number of isoforms due to alternative promoter usage, alternative splicing, or alternative polyadenylation (Figures 5.1 and 5.2). Furthermore, alternative isoforms are the ultimate effectors of RBPs and miRNA, therefore, isoform annotation should be taken into account.
Moreover, a regulator might only target a subset of isoforms of a gene, hence aggregation of the isoform expression to gene level might lead to loss of isoform-specific regulation, with possible false positive (FP) or false negative (FN) regulator target association.

In a pioneering study by Deng *et. al* (Deng *et. al*, 2011), who used Mutu I cells that overexpressed miR-155, it was shown that a simple transcript expression cut-off and seed enrichment strategy revealed more true miR-155 targets than using a cut-off on gene expression only. They had used RNA-seq data of Mutu I cells to estimate transcript expression levels using an Expectation Maximization (EM) algorithm. Over the last years, a number of EM-related approaches have been introduced which allow the quantification of transcript expression levels from RNA-seq data (Richard et al., 2010; Li and Dewey, 2011; Patro, Mount, and Kingsford, 2014). Therefore, we believe there is an opportunity to develop methods that allow regulatory model reconstruction on the level of individual isoforms, particularly in the light of a large number of available paired transcriptomics and epigenomics data sets, *e.g.* from the ENCODE(Sloan et al., 2015).

Here we propose a multitask learning (MTL) approach to fully exploit gene structure and cross isoforms commonalities. In this framework, as regulator binding site can be shared between different isoforms of a gene (see Figure 5.1, grey triangle miRNA), systematic information flows between different isoforms (tasks) of a gene, leading to more accurate isoform expression prediction, thus, reducing the chances of overfitting and improving model generalization performance. We investigate the difficulties of regulator-target prioritization for isoforms using different simulation studies and analysis of liver cancer RNA-seq data. We show that the new MTL approach is able to predict the expression level of target isoforms accurately than single task learning approaches, and leads to the most accurate prediction of experimentally validated miRNA targets (Dehghani Amirabad and Schulz, 2016).

Moreover, we extend our formulation to the joint inference of gene and isoform regulation, which allows to transfer the fine-grained biological variation at the isoform level to the gene level. We explicitly demonstrate that this leads to more accurate *gene* expression prediction in challenging scenarios (low sample size, high dimensionality). We shows that transcript isoforms is one of the richest sources of informative biological prior knowledge regarding transcriptional and post-transcriptional activities of a gene and our algorithm is the first and the only algorithm that fully exploit these priors.

In order to solve the multi-task regression for the entire human transcriptome, we derive and implement an efficient solver.
We apply our algorithm to integrate epigenomics and transcriptomics data from hepatocellular carcinoma (HCC), breast invasive cancer (BRCA), and prostate cancer. By analyzing the models, we highlight novel mechanisms of transcript isoform regulations.

5.2 Method

Transcript isoform regulatory network inference

We formulate the network inference problem as learning a single linear regression model for each transcript isoform using weighted linear combinations of TFs and miRNA (Figure 5.3). To this end, we integrate different high-throughput data including chromatin accessibility, paired transcriptomics (isoform and miRNA), and isoform annotation (3P-seq) into learning framework. The goal is to learn per isoform an interpretable model, and find sets of regulators for each isoform that explain observed variance in their expression.

Here, we describe our general setting for modeling transcript isoform regulation. Then, we describe a learning algorithm where we make full use of isoform annotations and expression level to transfer regulatory knowledge between related

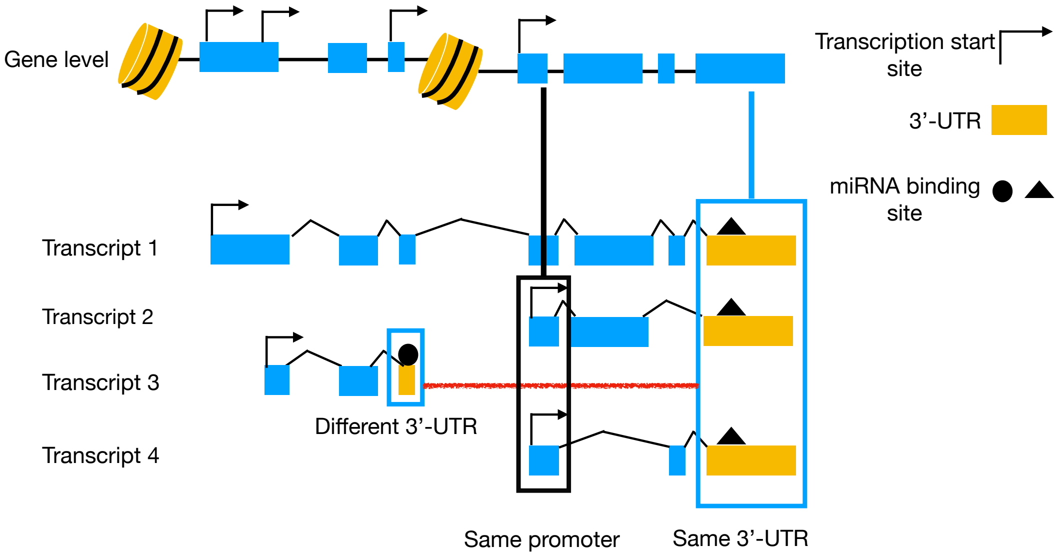


FIGURE 5.1: Structured information at isoforms of a gene. Different isoforms of a gene can share similar regulatory elements. Transcript isoforms 1, 2, and 4 have the same 3'-UTR, hence they posses the same miRNA and RBP binding sites in the 3'-UTR. Moreover, the transcript isoforms of 2 and 3 have the same promoter, hence they posses the same TF binding profile at the proximal promoter.

learning problems (tasks), which lead to an accurate isoform, and gene regulation inference. We evaluate the performance of models in terms of gene and isoform expression prediction accuracy and recovering the known underlying regulatory network.

Modeling isoform regulatory network inference

We model the steady-state expression of transcript isoform k at condition j, y_j^k, as a linear function of TFs ($X_r^k \beta_r^k$) and microRNAs ($X_m^k \beta_m^k$) as follows:

$$y_j^k = \sum_{r \in TFs} X_r^k \beta_r^k + \sum_{m \in miRNAs} X_m^k \beta_m^k. \tag{5.1}$$

For each transcript isoform model, we want to find a sparse solution ($\beta \in \mathbb{R}^{P \times 1}$) where non-zero elements of β indicate the contribution of each regulators in explaining observed variance in transcript isoform expression. We infer the parameters (β) of Eq. 5.1 in three different settings: single-task learning (STL), multi task learning (MTL) with isoforms of a gene, and multi-task learning of genes and the corresponding isoforms (GMTL).

5.2.1 Lasso formulation of regulatory network inference

Lasso is a sparse linear regression model for inferring the regression coefficients, β, and widely used in bioinformatics. The lasso estimate of Eq. 5.1 for isoform k is given by:

$$\min_{\beta^k \in \mathbb{R}^p} f(\beta^k) \equiv \frac{1}{2} ||y^k - X^k \beta^k||_F^2 + \lambda ||\beta^k||_1 , \tag{5.2}$$

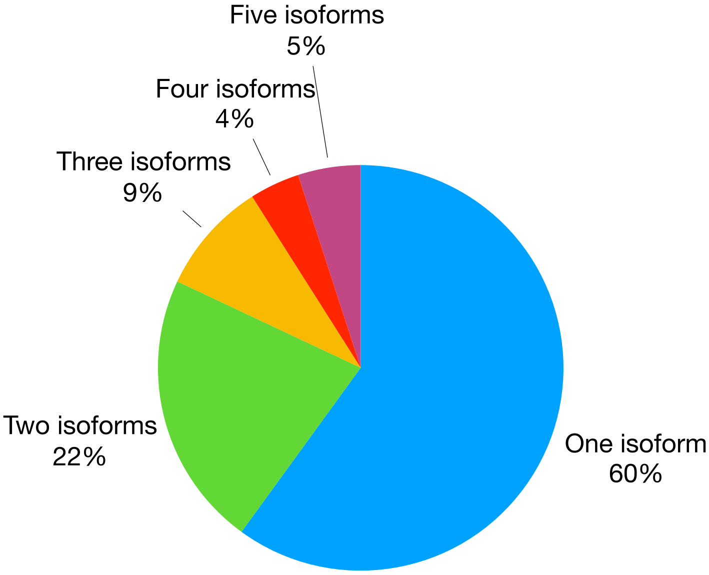


FIGURE 5.2: Isoform expression in BRCA data. The pie chart shows the distribution of the number of expressed isoforms per gene. Approximately 40% of genes express more than one isoform with predicted miRNA binding sites in their 3′-UTR.

where $y^k \in \mathbb{R}^{N\times 1}$ is the isoform expression level in N samples, $X^k \in \mathbb{R}^{N\times P}$ is the input feature matrix, $||\cdot||_F$ denotes the Frobenius norm, and $||\cdot||_1$ is the ℓ_1-norm. λ is the regularizer parameters, which controls the sparsity in β. We call this single-task learning, because parameters of each model are estimated separately.

5.2.2 Multi-task regression formulation of isoform regulatory network inference

Different isoforms of a gene may share regulatory elements, hence they can share similar regulators (Figure 5.1). For example, isoforms from the same gene may share miRNA binding sites because they have similar 3′-UTRs. In the case of Lasso estimation (Eq. 5.2), we optimize the parameter of individual models (tasks) independently (Eq. 5.2). This may lead to suboptimal inference and instability in regulator feature selection, especially in a low sample size setting or high dimensional data.

We aim to improve the generalization performance of individual isoform models by exploiting the shared and complementary information (structred sparsity pattern) between them. To this end, we formulate isoform expression level prediction as a multi-task regression problem as follows:

$$f(B) = arg\,min_B \frac{1}{2n}\sum_{k=1}^{K} ||y^k - X^k\beta^k||_F^2 + \Omega(B) \tag{5.3}$$

Where K, $X^k \in \mathbb{R}^{P\times N}$, and $\beta^k \in \mathbb{R}^{P\times 1}$ are the number of isoforms of a gene (tasks), isoform specific input feature matrix, and isoform specific regression coefficient vector respectively. The first term, is the fitting term and fits the parameters of the model independently. The second term $\Omega(B)$ is the sparse group lasso, which couples the models parameters by simultaneously shrinking the weights of the shared regulator

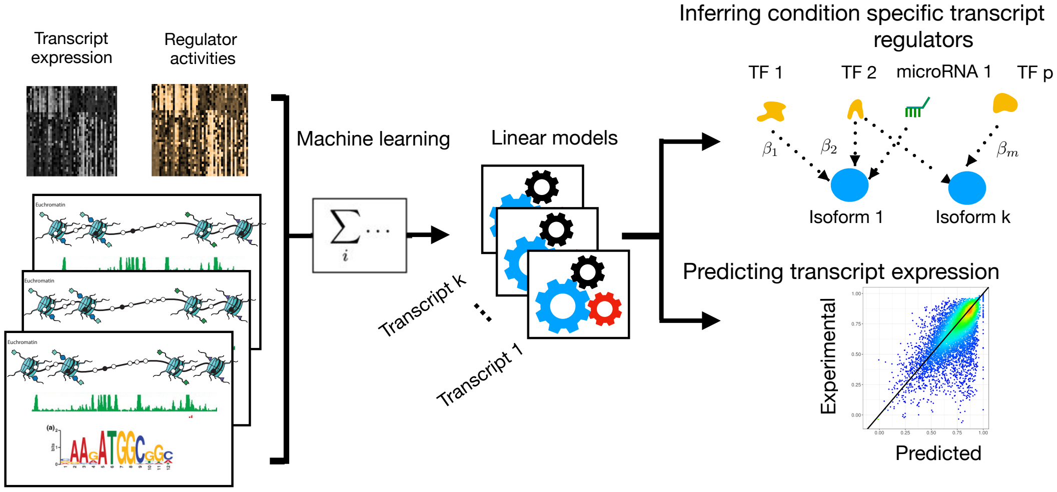


FIGURE 5.3: Isoform regulatory network inference workflow. Our network inference algorithm takes as input an isoform expression matrix, chromatin accessibility data, and priors for miRNA and TF target interactions. It returns a distinct regulation model for each isoform. Step I: Using priors and isoform expression data, we estimate TF activities in different samples. Subsequently, we expand our isoform-specific input features matrix by adding miRNA expression according to miRNA-isoform interaction priors. Step II: We learn a distinct linear model for each isoform as a function of TF activities and miRNA expression data. In this step, our algorithm incorporates isoform annotations (alternative promoter and 3′-UTR structure) to the model for effectively sharing regulatory information and a robust estimate of the isoform regulatory network. Step III: We construct the regulatory network of the entire transcriptome. Each model gives us the adjacency list of target node (isoform) and the input feature nodes (miRNA and TF). After statistical test corrections, we pool all the adjacency lists together and construct a directed regulatory network of the entire transcriptome. Weights of the individual edges correspond to the feature weight in the corresponding model. Morever, using our models, we predict the expression level of the isoforms in hold out data set

between isoforms (5.4). Its sparse multi-task group lasso is defined as:

$$\Omega(B) = \lambda||B||_1 + \gamma \sum_{p=1}^{P} g_p||G_p||_2$$

Where $B = [\beta^1, \ldots, \beta^K] \in \mathbb{R}^{P\times K}$ is the regression coefficien matrix for isforms of a gene. g_p is group-specific weight, which is defined as the square root of group size. G_p is the non-overlapping group lasso penalty and translates into the regression coefficient vector for the regulator p that has binding sites at subsets of isoforms of a gene (Figure 5.4).

One of our biological motivations behind this formulation is that a regulator with similar binding sites in different isoforms of a gene may also have similar regulatory effects. To this end, we use the non-overlapping group lasso to encourage the joint selection of regulators that have similar binding sites across different isoforms of a gene. This is achieved by structured sparsity-inducing property of the group lasso, which shrinks the coefficients within each group toward to a small value. Hence, if a group is selected by group lasso, then all entries of the group will have non-zero

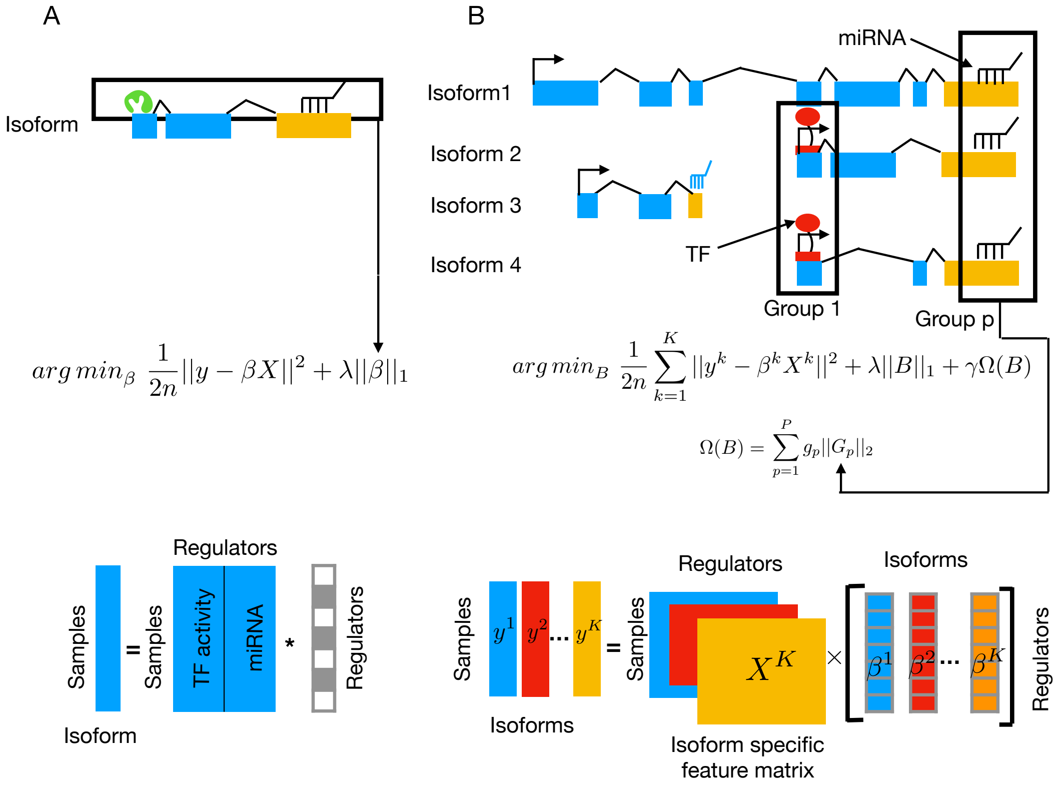


FIGURE 5.4: Comparison of single-task learning vs. multi-task learning. (A) In single-task learning, the output variable is the expression vector of isoform i measured in N samples. The input feature matrix consists of estimated TF activities and miRNA expression. By fitting lasso regression, we learn a weight vector β, which represents the contribution and direction of each regulator at explaining the observed variance in isoform expression. We learn a single model per isoform. In single-task learning, we do not encode any group information over input features. (B) In multi-task learning, the output variable is a matrix, where the rows are the samples, and the columns are the isoforms of a gene. Different to single-task learning, we have multiple input feature matrices (one matrix for each isoform). We infer a weight matrix B, where the columns are the isoforms, and the rows correspond to the regulators. Non-zero entries in the weight matrix correspond to the direction and magnitude of the regulators in explaining the observed variance in the corresponding isoform expression data. The black rectangles are the group structures that we define to incorporate isoform annotation into the model. Each regulator with similar binding sites in different isoforms of a gene forms a group (i.e., group 1 is the regression coefficient of the read TF in the isoform models 2, and 4). Structured sparsity-inducing penalty $\Omega(B)$ encourages the joint selection of the coefficient within each group.

values. In our context, this means that a regulator with similar binding sites in isoforms of a gene should have a similar regulatory effect. This might not be a realistic assumption for different reasons. For example, a binding site might be occluded depending on the secondary structure of the isoform. To relax this assumption, we add ℓ_1-norm to induce within group sparsity. This increases the power of our model to capture these subtle events as well.

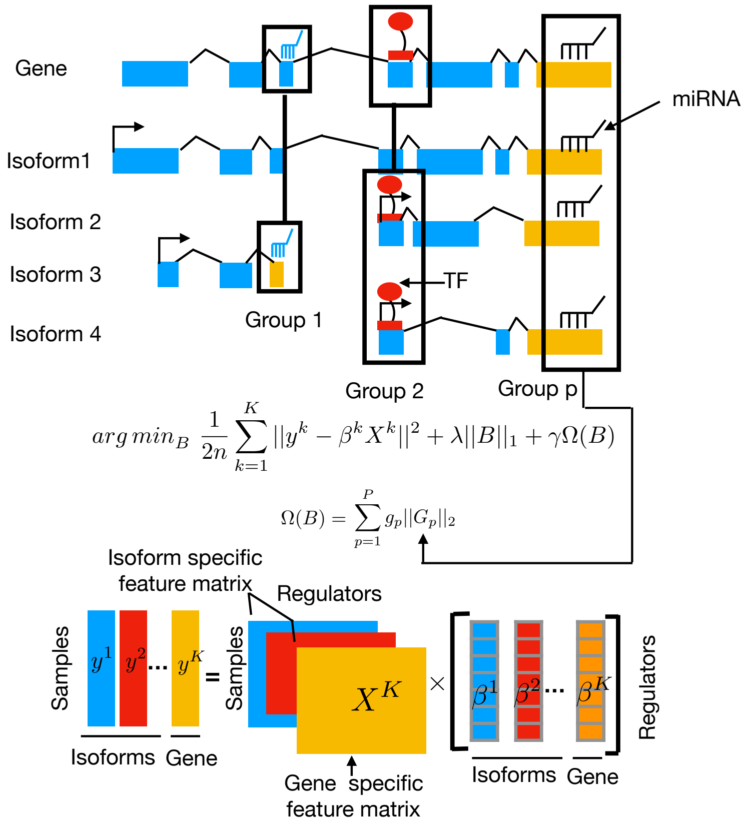

Gene
Isoform1
Isoform 2
Isoform 3
Isoform 4
miRNA
TF
Group 1
Group 2
Group p
$arg\,min_B \frac{1}{2n}\sum_{k=1}^{K}||y^k - \beta^k X^k||^2 + \lambda||B||_1 + \gamma\Omega(B)$
$\Omega(B) = \sum_{p=1}^{P} g_p||G_p||_2$
Isoform specific feature matrix
Regulators
Samples
y^1 y^2... y^K =
Samples
X^K
×
β^1 β^2... β^K
Regulators
Isoforms Gene
Gene specific feature matrix
Isoforms Gene

FIGURE 5.5: Simultaneous inference of gene and isoform regulation. Different to the previous formulation (Figure 5.4B), we incorporate gene expression as an extra task to the formulation. We project binding sites at isoform level to the gene level; hence group structure for each regulator includes the regression coefficients of the corresponding regulator at the gene level as well (See group 1, 2, and group p). $\Omega(B)$ encourages the joint selection of the coefficients within each group, which can lead to biologically meaningful regulator selection at the gene level, and subsequently accurate gene expression prediction.

5.2.3 Multi-task regression formulation for simultaneous inference of gene and isoform regulation

Another novel aspect we developed is to model gene expression by jointly learning gene and associated isoform expression level. This helps to transfer fine-grained biological knowledge at the isoform level to gene level, hence improving gene expression prediction accuracy. To this end, we add gene as an extra task to the previous formulation (Eq. 5.3) as follows.

$$arg\,min_B \frac{1}{2n}\sum_{k=1}^{K+1}||y^k - X^k\beta^k||_F^2 + \Omega(B) \tag{5.4}$$

Where K is the number of isoforms of a gene and $+1$ denotes gene as an extra task. To construct the input feature matrix for the gene task, we project predicted binding sites (miRNA and TFs) in isoforms to the gene level, and use activity of them as the input feature for training (i.e. isoform-specific task features are the subset of the gene task feature matrix). Moreover, we expand the definition of group structure for regulators. First, we project all the binding site information at isoform level to the gene level. Second, the size of each regulator group is increased by one to include the corresponding regression coefficient at gene level (Figure 5.5). This formulation enables us not only to learn an accurate model of gene and corresponding isoform regulations, but also highlight isoform specific regulators that may drive specific physiological function.

5.2.4 Model selection

In order to select the best model, we choose the optimal combination of hyper parameters (λ, γ) that minimizes sixfold cross-validation errors as estimated on validation set using a fixed parameter grid applied to the training samples.

5.2.5 Optimization and implementation

Although Eq. 5.3 defines a convex function, in which a globally optimal solution to β is attainable, the main challenge arises from the non-smooth penalty term. Widely used algorithm to solve regression problems with structured sparsity-inducing penalty is based on Nesterov's smoothing technique, which provides a smooth approximation of non-smooth structured sparsity term (Chen et al., 2012). Here, using accelerated proximal gradient descent algorithm, one can solve Eq. 5.3 for large-scale problems with convergence rate of $\mathcal{O}(\frac{1}{K^2})$. The idea is that we decompose Eq. 5.3 into a sum of two convex functions as follows:

$$f(B) = \ell(B) + g(B)\ , \tag{5.5}$$

where

$$\ell(B) = \sum_{k=1}^{K} ||y^k - \beta^k X^k||^2,$$

and

$$g(B) = \Omega(B).$$

Both $\ell(B)$ and $g(B)$ are optimized iteratively (Algorithm 1). Algorithm 1 consists of two main steps.

The first step is to compute the task-specific gradient descent using the following equation:

$$B_t[k] \leftarrow \beta_t^k - s^k \nabla l(\beta_t^k) \ \text{ for } k = 1, \ldots, K. \tag{5.6}$$

where K and s^k are the number of isoforms (tasks) and task-specific step size. Since $\ell(B)$ is Lipschitz continuous, the step size can be obtained from $s^k = \frac{1}{L^k}$, where $L^k = \lambda_{max}$ is the maximum eigenvalue of $(X^K)^T \times X^k$.

The second step is to optimize $g(B)$. Once we moved in the direction of gradient with step size s^k, we optimize non-smooth $g(B)$ at B_t using proximal operator as follows:

$$B_{t+1} \leftarrow arg\,min_W(g(W) + \frac{1}{2}||W - B_t||_F^2). \tag{5.7}$$

Eq. 5.6 and Eq. 5.7 iterate until the algorithm converges.
By solving Eq. 5.7, we want to find a matrix W which minimizes $g(W)$ and it is as close as possible to the solution that we get from the gradient step (B_t). The damping factor $\frac{1}{2}||W - B_t||_F^2$ enforces this closeness. Solving Eq. (5.7) is equivalent to solving another optimization problem. Here, we show that optimality of Eq. (5.7) is attainable using a closed formula solution.

Algorithm 1 Accelerated proximal gradient descent for multi-task regression with structured sparsity over input feature

```
Input: $\mathcal{X} \leftarrow [X^1, \ldots, X^K]$, Y, $\beta_0^k$, Lipschitz constants $L \leftarrow [L_1, \ldots, L_K]$, desired accuracy $\epsilon$
Initialization: $\beta_0^k = random, s^k = \frac{1}{L_k}$

for $t$ = 0, 1, ..., until convergences of $\beta_t^k$ do
    $B_t \leftarrow [\beta_t^1, ..., \beta_t^k]$
    #Compute task specific gradient
    for $k$= 1, ..., $K$ do
        $B_t[k] \leftarrow \beta_t^k - s^k \nabla l(\beta_t^k)$
    #Compute the proximal mapping using Eq. (5.13)
    $W_{t+1} \leftarrow \text{argmin}_W (g(W) + \frac{1}{2}||W - B_t||_F^2$          ▷

    #Compute momentum
    $B_{t+1} \leftarrow W_{t+1} + \frac{t}{t+3}(W_{t+1} - W_t)$.          ▷
Output: $\hat{B} \leftarrow W_{t+1}$.
```

To this end, we can write it as follows:

$$prox_g(B_{t+1}) = \text{argmin}_W(\lambda||W||_1 + \gamma \sum_{p=1}^{P} g_p||G_p||_2 + \frac{1}{2}||W - B_t||_F^2). \tag{5.8}$$

$$||W||_1 = \sum_{ij} |W_{ij}| \tag{5.9}$$

Since Eq. 5.9 is the element-wise l_1-norm, it can be decomposed either by columns or by rows. Similarly, $||W - U||_F^2$ can be decomposed into the sum of rows as shown above.

$$h(W) = \sum_{p=1}^{P} \lambda||W_{p*}||_1 + \gamma \sum_{p=1}^{P} g_p||G_p||_2 + \frac{1}{2}||W_{p*} - B_{p*}||_F^2. \tag{5.10}$$

Hence, we can optimize P functions independently. As this deals exclusively with vectors, $u = B_{p*}$ corresponds to the $p-$th row's of B and $f(w) = h(W_{p*})$, then each of P equations can written as:

$$f(w) = \lambda||w||_1 + \gamma g_p||G_p||_2 + \frac{1}{2}||w - u||_F^2 \ . \tag{5.11}$$

It is easy to show that (proof in the appendix A) the minimizer of $f(w)$ can be computed using the following closed formula solution:

$$Q_p = \text{argmin}_w f(w) = \begin{cases} (1 - \frac{\lambda}{||soft(u;\lambda)||})(soft(u;\lambda) & ||soft(u;\lambda)||_2 \geq \gamma g_p \\ 0 & ||soft(u;\lambda)||_2 < \gamma g_p, \end{cases} \tag{5.12}$$

where

$$soft(u;\lambda)\begin{cases}0 & if \quad |u_i| \leq \lambda \\ u_i - \lambda sign(u_i) & if \quad |u_i| > \lambda\ ,\end{cases}$$

Hence, the minimizer of proximal operator for Eq. 5.7 is obtained as follows:

$$\text{argmin}_W(g(W) + \frac{1}{2}||W - B_t||_F^2 = \sum_{p=1}^{P} Q_p. \tag{5.13}$$

5.2.6 Estimating regulatory network confidence score

Since different isoform models has different number of the features, the direct comparison of the regression coefficients between different isoforom models is not possible. Hence, we need a score to prioritize the predicted interactions comming from all isoform-specific models. Consequently, we compute F-test for each predicted interactions as follows:

$$F = \frac{(\frac{RSS_1 - RSS_2}{p_2 - p_1})}{(\frac{RSS_2}{N - p_2})}, \tag{5.14}$$

where RSS_1 is the residual sum of squared (RSS) errors produced by isoform model i, excluding feature j, and subsequently, the RSS_2 is the error from isoform model i, including all the selected features. p_1 and p_2 are the number of the parameters of the models (number of regulators + intercept), and N is the number of data point is used to learn the model. Our null hypothesis is that removing feature j from the model does not significantly increase the RSS error. F will have F-distribution, with $(p_2 - p_1, N - p_2)$ degrees of freedom. The p-value was calculated by $P[X > F(1, N - p_2)]$ and adjusted for multiple hypotheses testing over the regulators of an isoform using Benjamini-Hocherberg method (Benjamini and Hochberg, 1995a).

5.2.7 Expression data and preprocessing

Potential miRNA-isoform interactions are retrieved from TargetScan (v. 7.2) (Agarwal *et. al*, 2015). We used miRTarBase release 7.0 (Chou et al., 2017) to obtain experimentally validated miRNA target interactions.
We downloaded Sailfish-quantified gene, isoform, and miRNA expression values (TPM) (Patro, Mount, and Kingsford, 2014)for liver, breast, and prostate cancer from UCSC Xena (Vivian et al., 2016).
Differential expression of isoform in liver cancer determined with the DESeq2 algorithm version 3.3 (Love, Huber, and Anders, 2014) using un-normalized read counts per isoform. After performing multiple testing correction, isoforms with a false discovery rate (FDR) of ≤ 0.01 were considered differentially expressed isoforms.

The expression levels of gene, transcript isoforms, and miRNAs are log2 scaled. We filtered all miRNAs, transcripts, and genes if their expression was ≤ 0.5 RPKM in 90% of all samples. Moreover, we obtained DNaseI peak for HepG2, MCF-7, and PC-3 cell lines from Encode (https://www.encodeproject.org/).

Building TFs prior network from open chromatin data and sequence motifs

We defined the proximal promoter for each transcript isoform as a window of size 1.5 kb centered at each TSS.

In order to obtain an accessible promoter region for each transcript isoform, we intersected DNase1 peaks with core promoter regions of each transcript isoform using tools beetool (Quinlan and Hall, 2010). We used TEPIC to annotate and compute the binding affinities of TFs at open promoter regions (Schmidt et al., 2016).

Estimating the regulators activities

Transcription factor expression level is a not good proxy of their abundance in the cell (Arrieta-Ortiz et al., 2015). Instead of using TF expression level as the input feature, we do estimate their activities in different samples using RNA-seq, DNase-1 peak, and TF binding affinities as follows (Arrieta-Ortiz et al., 2015).
Let Y be the expression matrix of the isoforms, where rows are the isoforms and the columns represent samples. Moreover, P denotes a matrix of prior regulatory relationships between TFs (columns) and isoforms. The entries of the prior network are derived from scaled TF-binding affinities at open chromatin regions. In order to estimate the activities of TFs at each sample, we assume that we can write the expression level of isoform i in sample j as a linear combination of TFs activities A, in which we solve for

$$Y_{i,j} = \sum_{r \in TFs} P_{i,r} A_{r,j}. \tag{5.15}$$

We have more equations than number of variables (TFs), so it is an overdetermined system, hence there is no solution A. However, we can approximate activities by finding $\hat{A}$ that minimizes $||P\hat{A} - Y||^2$. Hence, using the pseudo inverse of $P^* = (P^T P)^{-1} P^T$, the solution is obtained by $\hat{A} = P^* Y$.

5.2.8 Simulation

The main assumption for creating the synthetic data is that there is a linear relationship between regulators and a target transcript:

$$y^k = X^k \beta^k + \epsilon^k \,.$$

Moreover, regulator (TFs and miRNAs) binding sites in different isoforms of the same gene, have a similar (not necessarily the same) regulatory strength. We simulated regulator activity levels (X) for 50 regulators by sampling from a univariate Gaussian distribution for each regulator. Mean and standard deviation of the regulator activity distribution is sampled from $\mu \sim \mathcal{N}(0.1, 2)$ and $\sigma \sim \mathcal{N}(0.01, 1)$, respectively. Moreover, we sampled the ground truth regulatory effect vector (β^*) from $\beta \sim \mathcal{N}(-0.5, 0.5)$. Then we created sparse regulator transcript binding priors such that isoforms of the gene share at least 50% of regulator binding sites with each other. For the simulation for every task (transcript) we defined 6 true regulator binding events (cf. Fig. 5.6A) and 42 regulators as false positive associations. Unless otherwise stated we simulated 35 samples in each setup.
We simulated transcript expression levles using the following linear model:

$$y^k = X^k \beta^k + \epsilon^k, \, \forall_k = 1, \ldots, 16 \tag{5.16}$$

where β^k is the vector of coefficient for regulators that have binding sites at transcript k. For each transcript, and $\epsilon \sim \mathcal{N}(0, 0.5)$ is the added random guassian noise. Both transcript and miRNA expression level scaled and considered model without intercept. We simulated synthetic data for three different setups with the above

setting
(I): We created 5 different simulation setups with 2, 6, 8, 10, and 16 isoforms using Eq. 5.16 with noise distribution $\epsilon \sim N(0, 0.5)$.
(II): We created 6 simulations with different noise levels using 16 isoforms with Eq. 5.16 and noise levels with $\mu = 0$ and $\sigma = 0.15, 0.5, 0.75, 1.0, 1.5,$ and 2.0 respectively.

Given the simulated expression level y^k we measure the root mean square error (RMSE) between predicted and test-set expression levels $\hat{y}^k$ as follows:

$$RMSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 , \tag{5.17}$$

where N is the number of test data samples (N=500).

Given true positive and true negative regulator binding events from simulated data, we normalized the estimated coefficients to the interval [0,1]. Then we used R package pROC (version 1.8., (Robin et al., 2011)) to compute Area Under the Curve (AUC) values for ROC curves of the models for each simulation setup.

5.3 Results

5.3.1 Evaluation of synthetic data

In order to assess the problem of transcript expression based inference of miRNA regulation we created a set of simulated datasets that allow us to explore the performance of the different approaches, see methods. In particular we are interested to test our hypothesis that the MTL approach can outcompete the disjoint model by borrowing information from several isoforms.

Effect of different number of transcripts on model performance

In the first setup we vary the number of transcripts (2,6,8,10,16) using the ground truth miRNA interaction matrix as shown in Fig. 5.6A, by taking different subsets of this matrix (columns from right to left). 8 miRNAs with different interaction strength are simulated, such that the interaction coefficient is the same among all regulated transcripts. 4 of the miRNAs are shared among all 16 transcripts and the other 4 are shared between 10 and 6 transcripts, respectively.

In Fig. 5.6B we show the normalized coefficient matrix for estimation with all 16 transcripts. Both the MTL and the disjoint model recover most of the true miRNA interactions, but select a number of false positive interactions. In this simulation $\sim$ 80% of the interactions in the ground truth set are FP interactions, which makes it a challenging problem, resembling the amount of FP interactions when relying on static sequencing based prediction methods. We noticed that the MTL method shows overall more low intensity FP coefficients compared to the disjoint model, albeit at higher frequency. We believe that this is an advantage as these low intensity coefficients are easy to filter. However, the disjoint model has a number of non-zero interactions that are false positives, but have equally high values compared to TP interactions.

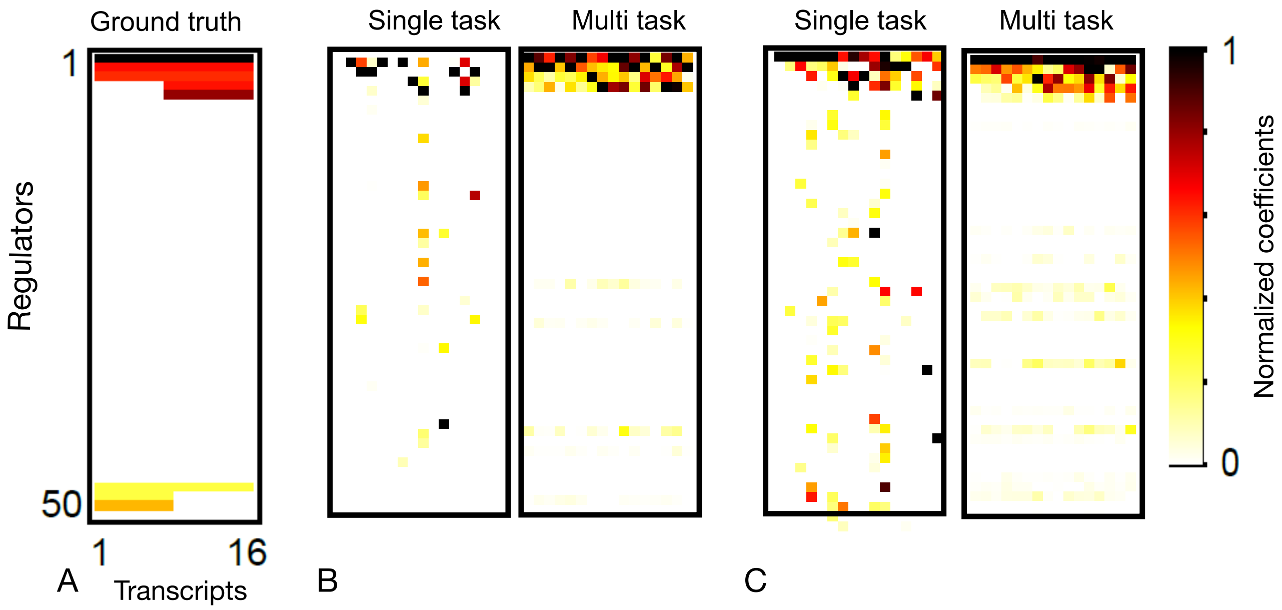


FIGURE 5.6: Comparison of recovering sparsity pattern for the 16 transcript simulation. (A) Ground truth values of regulator binding association for the complete set of 16 transcripts. Out of 50 regulatrors (rows) each transcript (columns) has 6 regulators that are associated with it, visible as a colored spot in the heatmap. The rows of the heatmap are clustered using hierarchical clustering and group regulators that show similar associations in the simulation. (B) Estimated regression coefficients with the single-task and multi-task methods on the simulation dataset with $N = 100$ for transcript expression levels. (C) Same as (B) with $N = 200$. All absolute values of simulated and estimated coefficients are normalized to [0,1].

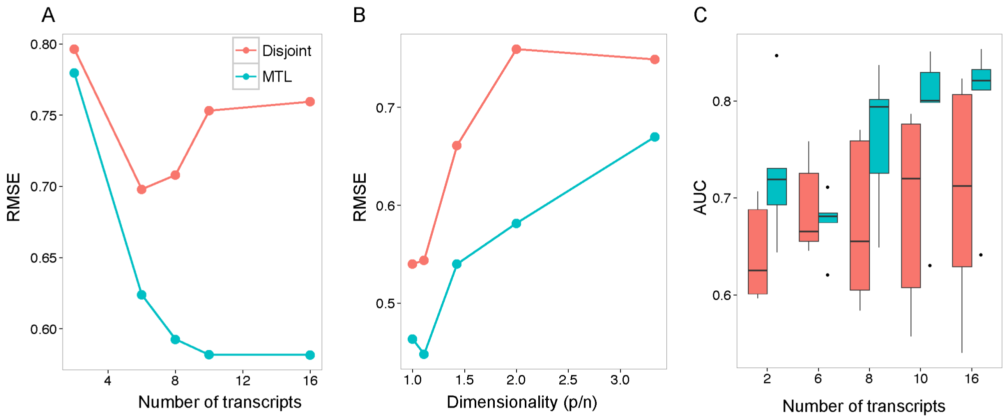


FIGURE 5.7: Comparison of the STL and MTL learning on different simulated datasets. Comparison of test RMSE (y-axis) obtained for an increasing number of transcripts per gene (A) or increasing dimensionality of the problem (ratio of features to samples) (B). (C) Distribution of AUC values for all simulations in setup (III), see Methods. The higher the number of transcripts that share a binding site, the more the MTL method outperforms the disjoint model

Figure 5.9A shows the RMSE (Equation 5.17) on test samples for all setups and compares the disjoint with the MTL model. Generally, the higher the number of coupled transcripts per miRNA binding site, the smaller the error for the MTL method. The error for MTL converges at 10 transcripts with no further improvement with

16 transcripts. In contrast, the disjoint model shows increased error with more than 10 transcripts. This is due to our simulation setup, where the 10 and 16 transcript problems have more miRNAs with smaller true binding coefficients, that are harder to estimate correctly (cf. Fig.5.9A).

Learning performance in a high dimensionality setting

Another important aspect of learning sparse models is the dimensionality (p/n) of the problem, where p denotes the number of parameters and n denotes the number of samples. Then if $p \gg n$ the problem is called high dimensional. We conducted additional simulations for learning with increased dimensionality (simulation setup III), as occurs often in practice when a limited number of paired miRNA and gene expression samples are available. In Fig. 5.9B we show the obtained error for the disjoint and MTL model. With higher dimensionality the error for both models increases as expected. For all values tested the MTL model shows smaller errors compared to the disjoint model. The disjoint model is more prone to overfitting in high dimensions.

While smaller RMSE values show, that the MTL model predicts true expression levels better, we used a ROC analysis in addition to evaluate the performance of true miRNA binding predictions (see Methods). In Figure 5.9C we show the distribution of AUC values for different ranges of problem dimensionality, stratified by the number of transcripts. Overall, we observed that for the MTL model, when the number of tasks increases the AUC values increases, with values up to 0.85 for 16 transcripts. Also the MTL model always shows higher AUC values and lower variance in the AUC compared to the disjoint model.

We further investigated the variance in the coefficient estimates in Fig. 5.8, studying all simulations (setup III). All 7 simulated miRNAs (features) that were consistently selected as non-zero by both models are shown. Note that the disjoint method failed to select one of the miRNA features most of the time and therefore we excluded it from this analysis. Over all feature values the variance of the MTL method is much smaller compared to the disjoint method. Features that are shared between all 16 transcripts show the largest reduction in variance.

Thus, if the assumption of shared miRNA binding strength for transcripts with the same site is true, then using MTL leads to better performance and can be trained with fewer samples.

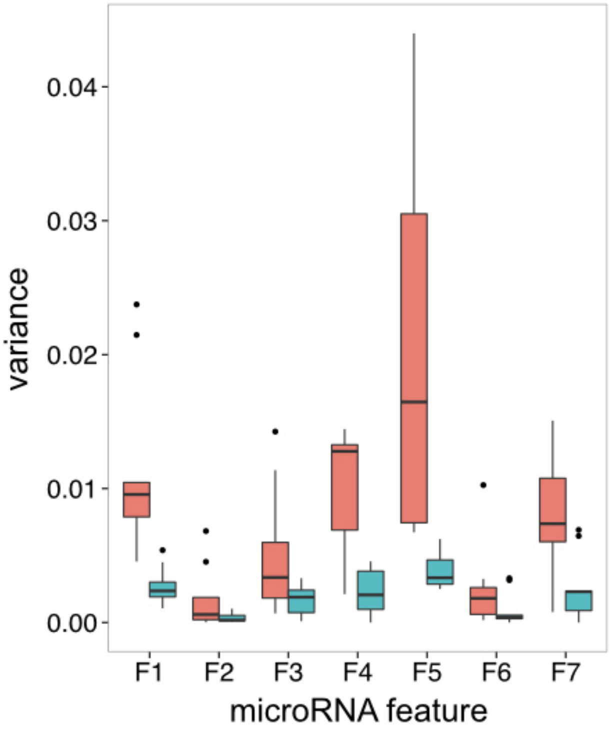


FIGURE 5.8: Variance in regulator binding strength estimation. Variance of estimated feature coefficients compared to the ground truth value from simulations (y-axis) over all simulation setups and for each miRNA feature and method (x-axis), same colors as in Fig. 3. MTL shows consistently smaller variance among all simulation setups.

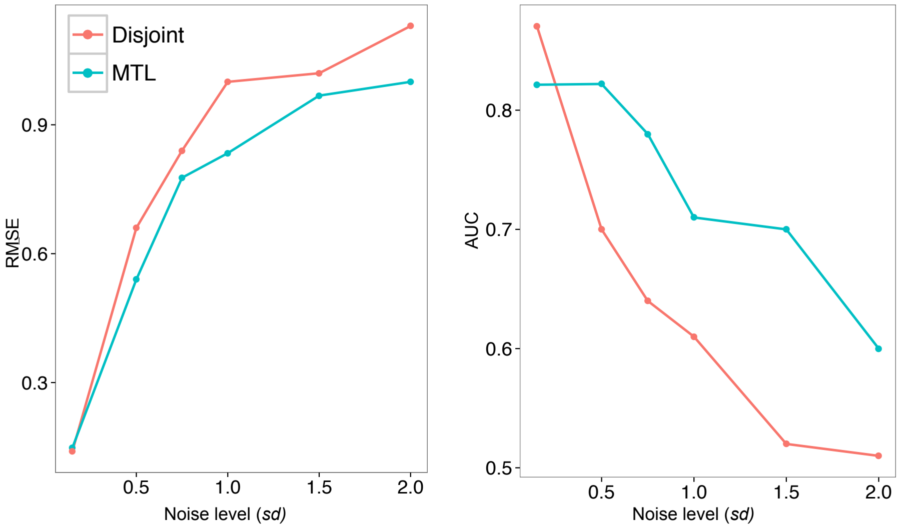


FIGURE 5.9: Comparison of the STL and MTL learning on different simulated datasets. Comparison of test RMSE (y-axis) obtained for an increasing number of transcripts per gene (A) or increasing dimensionality of the problem (ratio of features to samples) (B). (C) Distribution of AUC values for all simulations in setup (III), see Methods. The higher the number of transcripts that share a binding site, the more the MTL method outperforms the disjoint model.

Multiple transcripts of a gene may share very similar exonic structures. This can potentially make it hard to discern the origin of the read and lead to noisy estimation of transcript expression levels from RNA-seq data (Teng et al., 2016). In order to prioritize miRNA target at transcript level, the algorithm should be robust to high noise levels in transcript expression data.

5.3.2 Accurate isoform expression level prediction using TF and miRNA

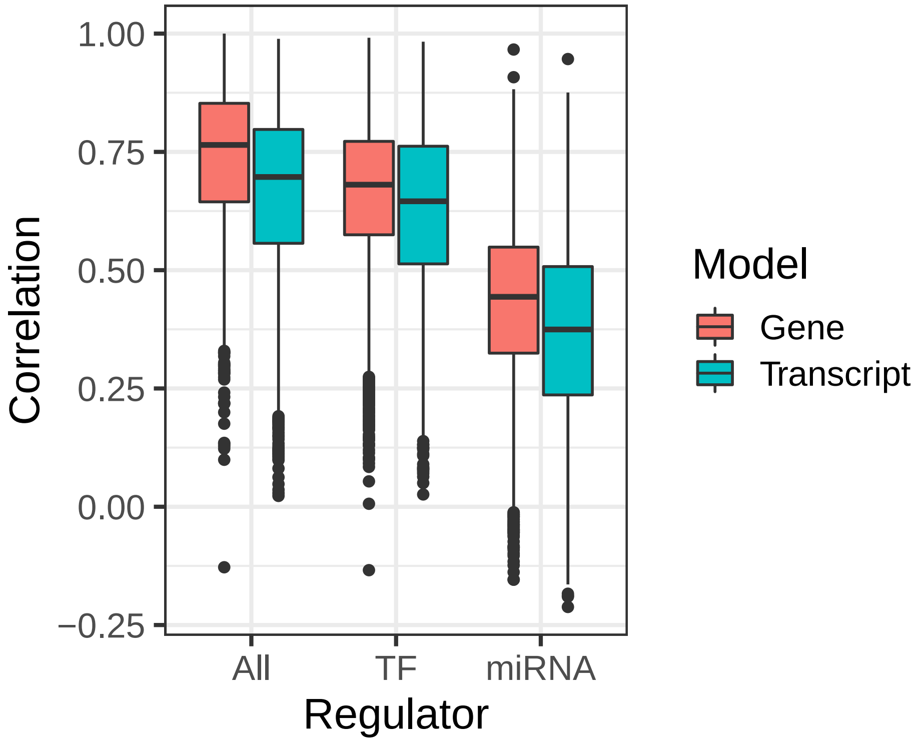


FIGURE 5.10: Comparison of gene and isoform expression prediction accuracy. Boxplot of Spearman correlation between the predicted and held-out genes (red) and transcript (blue) expression in sixfold cross-validation (CV). The All uses all input features (TF and miRNA) and is compared to the models exclusively trained using only one type of input variable (TF or miRNA). For each gene and respective isoforms, the optimum models are selected. The Box-plot shows the distribution of the correlation for all the models. The higher the reduction in correlation by excluding a specific variable, the more power that regulator has in explaining the variance in the expression data. Excluding TF from the models leads to a higher reduction in the model's performance relative to the full model; hence TFs provide more power in terms of explaining the variance in mRNA expression level.

In light of a large amount of throughput RNA-seq data, we aimed to learn predictive models of isoform expression in liver cancer. It remains to be determined whether we can predict isoform expression level with remarkable precision using miRNA expression level and TF binding profile at the core promoter region. Figure 5.10 shows the generalization performance measured as the correlation between predicted and hold-out isoform expression data for 14,000 isoform-specific models in liver cancer. On average, specific models explain the variance in the expression level to a large extent. Moreover, it highlights the contribution of the core promoter and 3'-UTR in controlling gene expression level. As expected, TF is a strong predictor of isoform expression compared to miRNA, which mainly inhibit translation of

the mRNA or fine-tune the expression level. As a result, they are not strong features in explaining observed variance in the expression data.
Another interesting aspect is the relative accuracy of the gene vs. isoform level expression prediction. Models trained at gene levels shows relatively better generalization performance compared to isoform-level models (Fig 5.10).

5.3.3 Multi-task learning improves miRNA target prioritization

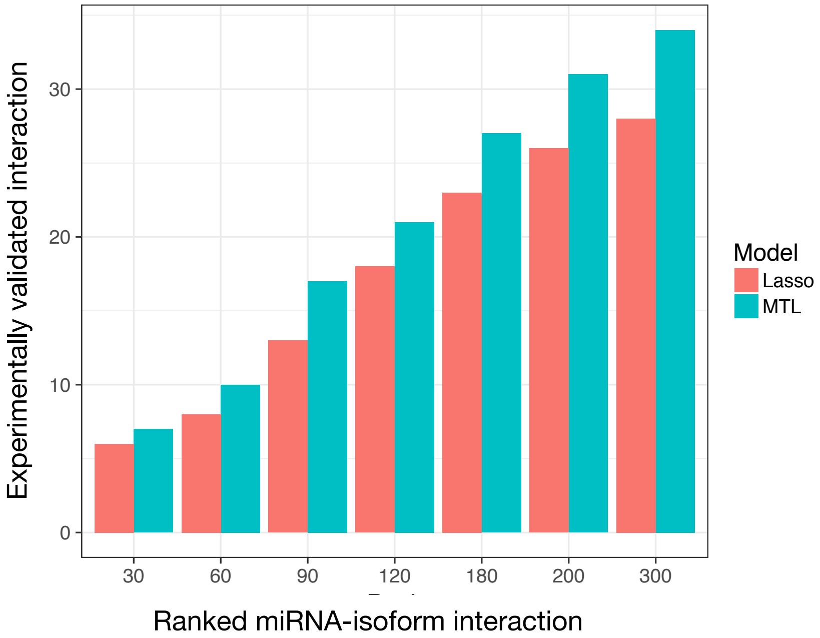


FIGURE 5.11: Comparison of the STL and MTL approach for prediction of miRNA-transcript target interactions in liver cancer RNA-seq data. Predicted interactions are ranked based on their adjusted p-value(x-axis) and at each interval the number of detected experimentally validated interactions is shown (y-axis). MTL consistently shows superior performance compared to the Lasso.

In order to test the ability of the methods to prioritize true miRNA interactions, we used Hits-Rank plots, using experimentally validated miRNA interactions from the miRTarBase database (Chou et al., 2017). We compared the MTL and disjoint methods in their ability to rank validated miRNA interactions. After fitting models to isoforms, all predicted miRNA-isoform interactions were pooled together and ranked by their p-value. Then, at each interval, the number of experimentally validated interactions (hits) was computed. As shown in Fig. 5.11, the MTL method is able to rank more true miRNA interactions in the top ranks, and also detects a higher number of true interactions among all non-zero coefficients.

5.3.4 Joint learning of gene and isoform models leads to accurate isoform expression prediction

We asked whether we can we improve isoform expression prediction accuracy in a low sample size setting through sharing regulatory information between transcripts. We compared the performance of MTL and the lasso method for isoform expression level prediction for a LumA liver cancer subtype dataset for 4,200 isoform models. Figure 5.12 shows the generalization performance obtained on test samples with

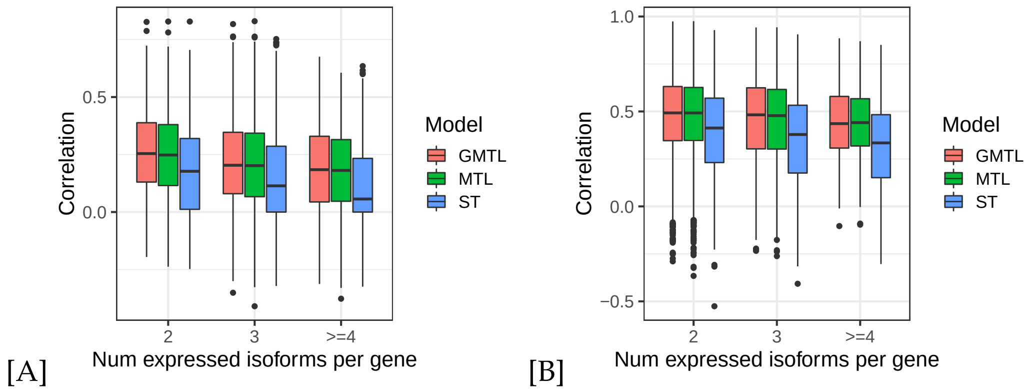


FIGURE 5.12: Comparison of isoform expression prediction accuracy. Sparse regression models predict transcript isoform expression level on holdout data for the LumA breast cancer subtype. ST and MTL represent lasso and multi-task regression, respectively. GMTL (Eq. 5.4) denotes the multi-task learning model, where the gene is added as an additional task in the formulation. Y-axis is the correlation between predicted and holdout expression data. (A) Models trained using expressed miRNA as input features. (B) Models trained using miRNAs and TFs as input feature.

both approaches. We observed that MTL outperforms lasso for most of the transcripts, as it does in the simulated data examples. Irrespective of the input features used (miRNAs or TFs), MTL models capture more of the observed expression behavior of individual transcripts in liver cancer. Moreover, lasso models are prone to overfitting when they are trained exclusively, using miRNAs input features.

Finally, we investigated the possibility of further improving the isoform expression level prediction by incorporating gene expression as an extra task to the formulation (GMTL in Fig 5.12). As in figure 5.12, the transcript models do not benefit remarkably from the gene as an auxiliary task in the multi-task learning setting.

5.3.5 Incorporating isoform-level estimates leads to more accurate gene expression prediction

As we mentioned, using gene expression for predictive modeling can mask transcript-level dynamics and lead to suboptimal regulatory network inference. Conversely, gene abundance estimates are more accurate, and downstream analysis at the gene level may be more interpretable than at the transcript level. A straightforward way to reconcile gene-centric vs. transcript-centric views is to construct a model at the isoform level, then aggregate that to the gene level. This approach may be able to capture biological variability at the isoform level, but does not leverage gene expression estimates in the learning procedure.

Keeping in mind that aggregating isoform expression levels cancels out the up and down regulation of isoform, we tried to address it to a large extent in network reconstruction. We hypothesized that, if isoforms of a gene show up and down regulation due to deregulation of regulators, then corresponding isoform tasks will more likely pick up the respective regulators in the learning procedure. Hence, if we add gene as an extra task to the formulation, the group lasso penalty will encourage the joint selection of the respective regulators at gene level as well. This might lead to a robust and accurate gene expression prediction, since captures the fine-grained biological variation at isoform level and shares it with gene level. To investigate this, we compared the expression prediction accuracy genes in LumA cancer subtype in

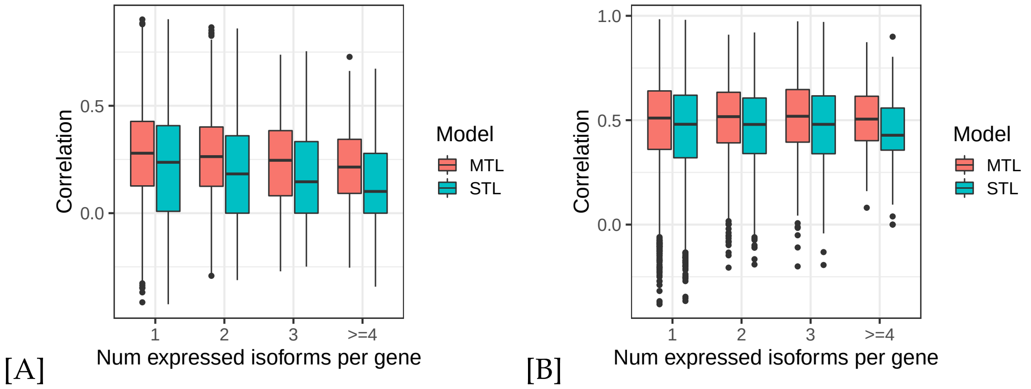


FIGURE 5.13: Comparison of gene expression prediction accuracy with and without incorporating isoform annotation and expression into the model.(A) STL is the lasso regression on gene expression across different samples. MTL is the joint modeling of gene and associated isoform expression values (Eq. 5.4). Models are trained using only miRNAs as input features and performance is separated by the number of expressed isoforms per gene (x-axis). (B) The same as A, but models are trained using TFs and miRNAs. In both cases, incorporating isoforms into the formulation boosts gene expression prediction accuracy, especially for genes with man y isoforms expressed.

two different settings. We trained the gene level models using the lasso, and multitask regression in two different settings; (I) Only expressed miRNA used as input feature (II) miRNA and TF used as an input feature. We assess accuracy of the prediction as a Spearman correlation between predicted and hold-out gene expression data sets. As figure 5.13 shows, incorporating isoform annotation and expression level improves the prediction accuracy of the gene expression in both settings.

5.3.6 Isoform expression regulation is associated with 3′-UTR length

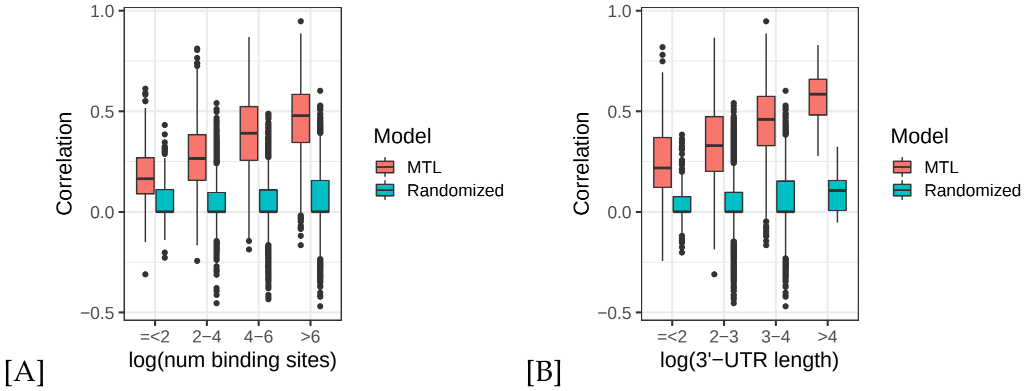


FIGURE 5.14: Association of 3′-UTR and model performance in explaining the variance in isoform expression level. (A) There is a positive association between the ability of the model to explain the isoform expression level with the number of miRNA binding sites at 3′-UTR. (B) This is similar to the trend with the length of 3′-UTR.

To get insights regarding the length of 3′-UTR and miRNA-mediated gene regulation, we investigated the connection between isoform expression level prediction accuracy and 3′-UTR length. We hypothesized that longer 3′-UTRs may mediate more regulatory functions, since they often harbor more regulatory elements, which may lead to more flexibility in combinatorial binding events. We categorized the performance of isoform models (I) by the number of miRNA binding sites in 3′-UTRs (Figure 5.14 A), or (II) the length of 3′-UTRs (Figure 5.14 B). Our analysis suggests that the accuracy of isoform expression prediction using miRNAs strongly depends on both the number of miRNA binding sites in the 3′-UTR and the length of 3′-UTRs (Figure 5.14).

5.3.7 3′-UTR is a major driver of tissue specific expression for many isoforms

As shown previously, 3′-UTRs can play a major role in driving of tissue-specific isoform expression (Merritt et al., 2008; Spies, Burge, and Bartel, 2013). We investigated whether we can find a subset of isoforms, where regulation heavily relies on the 3′-UTR region whereas the promoter shows a more constant gene expression phenotype. To this end, we partitioned models based on generalization performance into: (I) only post-transcriptional regulation (PTR) models, and (II) both transcriptional and post-transcriptional regulation (TR-PTR) models.

PTR corresponds to the isoforms, where TF-based models fail to predict expression level reliably (correlation <0.1 on holdout data), but miRNA-based models predict expression more accurately. The rest of the isoform models refer to TR-PTR, where both TF-based and miRNA-based models work given the correlation threshold

We asked if there is any subset of isoforms, where TF-based models (transcriptional models) fail to explain the observed variance in expression values. Surprisingly, we found 2,300 expressed isoforms in the liver cancer dataset, where the corresponding core promoter was accessible, but TF-based models failed to predict their expression level. This suggests that the core promoter of these isoforms are only permissive of transcription. Moreover, when we looked at the performance of miRNA-based (PTR) models, these were able to predict expression accurately, compared to the miRNA-based models for the rest of isoforms (TR-PTR) models (Figure 5.15 A). This suggests that the 3′-UTR sequence is the major driver of expression variation for these isoforms (PTR) compared to the rest of the expressed isoforms in the data.

When we compare the expression level of PTR isoforms with TR-PTR isoforms, PTR isoforms tend to have a low expression level (Figure 5.15 B). Moreover, the gap between the expression prediction accuracy for TPR and TR-PTR increases with the increase in the length of the 3′-UTR (Figure 5.15 A). This evidence supports the idea that the 3′-UTR sequence is the major driver of expression variation for PTR isoforms. Furthermore, recent reports (Wang et al., 2012) indirectly support our findings. It was shown that isoforms with longer 3′-UTRs are generally unstable, have a low expression level, and harbor high density of conserved miRNA binding sites.

5.3.8 On the tissue specificity of transcriptional and post-transcriptional regulation

Many computational methods model the expression level of individual genes from genomics and epigenomics data. One interesting aspect is whether models, trained in a particular tissue, can accurately predict the expression values of individual genes in other tissues. When we train a linear model for one individual isoform based on chromatin accessibility of the regulatory elements and miRNA expression, we are estimating a model of regulation for each isoform. How transferable is this information? Can we predict isoform expression in a different tissue? In other words, are the models of the regulation in different tissues similar, or are they

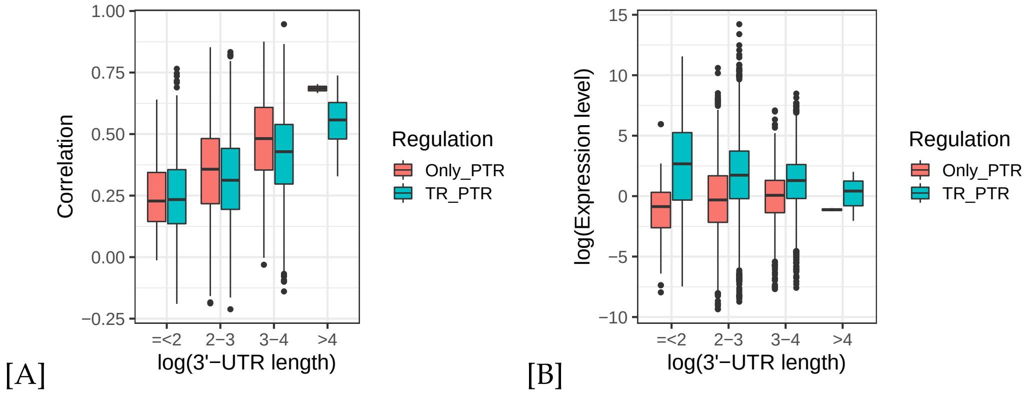


FIGURE 5.15: 3′-UTR as major driver of isoform expression level. Correlation between predicted hold-out expression (y-axis) is compared for different bins of 3′ UTR length (y-axis,log). (A) There is a positive association between the ability of the model to explain isoform expression with the number of miRNA binding sites at 3′-UTRs. (B) A similar trend is observed with the length of 3′-UTRs.

tissue-specific?
We investigated this for both TF-based and miRNA-based models. More specifically, we trained isoform-specific models in one tissue (5.16 A and B) and used pre-trained models to predict the expression level of the corresponding isoform in other tissues (here referred to as target tissues). We did this experiment for the isoforms expressed in both tissues, with an accessible corresponding proximal promoter. Surprisingly, on average isoform models show very poor generalization performance to other tissues (5.16). More interestingly, the relative performance of miRNA-based models is higher than TF-based models.

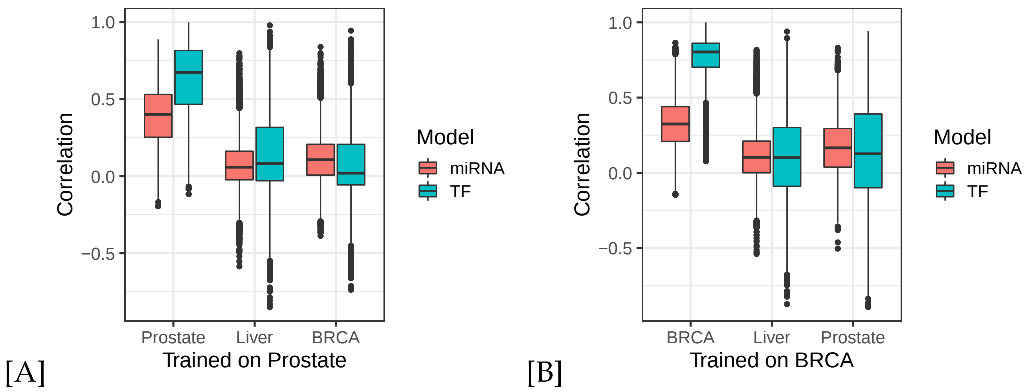


FIGURE 5.16: Isoform-specifc model generalization performance across tissues. (A) Two sets of isoform-specific models (TF and miRNA) are trained on prostate cancer, and pre-trained models are used to predict the expression level of the isoforms in liver and BRCA. Y-axis shows the correlation between predicted and holdout expression data. Every point in the distribution is the performance of an individual model. Both TF and miRNA based models show high reduction in their expression prediction performance in the target tissues (Liver and BRCA). (B) Similar to A, but models are trained using the BRCA data set and applied to predict the expression level of isoforms for liver and prostate cancer data. Relative reduction in TF based models is higher than for miRNA based models.

Regulator	Centrality	Node degree	Regulator	Centrality	Node degree
TFDP1	0.4	205	hsa-miR-5589-5p	0.3	159
TCFL5	0.37	193	hsa-miR-153-5p	0.28	157
TEAD2	0.36	187	hsa-miR-766-3p	0.27	153
NRF1	0.34	177	hsa-miR-590-3p	0.28	143
GMEB1	0.33	170	hsa-miR-340-5p	0.26	141
ZBTB14	0.33	169	hsa-miR-335-3p	0.25	141
KLF6	0.33	166	hsa-miR-576-5p	0.24	132
ETV5	0.32	165	hsa-miR-3074-5p	0.23	130
SP2	0.31	156	hsa-miR-629-3p	0.25	129
KLF7	0.3	153	hsa-miR-423-5p	0.22	129

TABLE 5.1: Top TFs and miRNA that are associated with differentially expressed isoforms in liver cancer. They are ranked according to eigenvector centrality.

5.3.9 Regulatory network analysis of differentially expressed isoforms in hepatocarcinoma

After demonstrating that POSTIT outperforms gene-centric methods in the prioritization of miRNA target interactions in liver cancer, we proceeded with network analysis to find important regulators that target many of the differentially expressed isoforms in liver cancer. Therefore, after FDR correction of the predicted interactions, we computed both TF-isoform and miRNA-isoform networks. In these networks we computed the eigenvector centrality and node degree for all the regulators. Table 5.1 shows the top TFs and miRNAs that regulate most of the differentially expressed isoforms in liver. To further examine the functional implication of the top regulators identified in the isoform regulatory network in hepatocarcinoma, we performed an in-depth literature survey. Interestingly, the top regulators prioritized by POSTIT are enriched in cancer-related TFs and miRNAs. To name a few by order of eigenvector score, the TFDP1 TF regulates the activity of genes involved in cell cycle progression from G1 to the S phase. Overexpression of TFDP1 is associated with progression of hepatocellular carcinomas (Yasui et al., 2003). TEAD2 (Liu et al., 2016) controls the tumor initiation and progression by regulating tumor progression-inducing genes such as CTGF, Cyr61, Myc, and Gli2. KLF6 (Sirach et al., 2007) regulates the cell cycle in the liver and it has been reported that inhibition of KLF6 helps hepatocellular carcinoma (HCC) cells to evade from apoptosis. It has been reported that overexpression of hsa-miR-153-5p promotes β-catenin transcriptional activity, which leads to cell cycle progression and colony formation in HCC cells (Wu et al., 2016). Furthermore, miR-153-5p anti-miR suppressed hepatocellular carcinogenesis in a murine liver cancer model. (Hua et al., 2015). Further, hsa-miR-766-3p (You et al., 2018) acts as tumor suppressor miRNA by targeting Wnt3a in HCC. Additionally, hsa-miR-590-3p (Dong and Qiu, 2017) is overexpressed in HCC and promotes cell proliferation, tumor growth, and metastasis. Finally, hsa-miR-423-5p (Stiuso et al., 2015) has been reported to promote autophagy in cancer cells and was increased in the blood serum of the HCC patients. This highlights that our isoform specific analysis, was able to prioritize many known important regulators of hepatocarcinoma. The isoform resolution level predictions generate novel hypotheses for downstream experiments.

5.4 Discussion

We have introduced a new paradigm for regulatory network inference that works on the level of individual isoforms instead of the (virtual) gene. We have introduced three sparse learning setups that allow inference of isoform regulation from paired transcriptomic and epigenomics data. The MTL approach in particular makes full use of available isoform expression data and annotations to estimate regulator interaction coefficients by borrowing information from all isoforms of the same gene that share this binding site. If no binding sites are shared between isoforms of a gene, then the MTL approach reduces to the lasso and, therefore, there is no limitation in practice.

We conclude that the reduced variance on coefficients with the MTL approach leads to lower errors in simulated and real data. However, we note that our simulations are oversimplified, disregarding many other relevant aspects of regulator target prediction, for example, RNA secondary structure, and assuming that expression levels follow a Gaussian distribution. Further, one could explore more complex simulation setups with a varying sparsity pattern and number of regulators, for instance.

In real data, although most of the genes we tested only had two expressed isoforms, the MTL approach showed a clear advantage, as sharing between more isoforms reduces the variance of coefficient estimates due to the group lasso norm.

It has been reported that isoform expression estimation from RNA-seq data can be noisy, which is a hard learning problem. Therefore, we evaluated the robustness of our method on different noise levels on synthetic data and showed that the MTL formulation increases robustness to noisy transcript expression estimation. Studying how biases in transcript expression estimation impact regulatory network inference further would be helpful to understand the limits of the suggested approach when applied to commonly available protocols for RNA-seq data sets (Griebel et al., 2012). However, more recent technologies like PacBio or Oxford Nanopore often allow for complete sequencing of the transcript, therefore mitigating the above mentioned biases in transcript expression estimation.

Gene vs. isoform expression prediction accuracy

Another interesting aspect is the relative accuracy of the gene vs. isoform level expression prediction. Models trained for gene expression show relatively better generalization performance compared to isoform-level models (Fig 5.10). This may be due to the difference in the total number of features used for training the models. More specifically, depending on the annotation of the 3′-UTR and core promoter of isoforms, different isoforms of a gene can have different input features, and the union of these feature sets are used as input for training the respective gene model. Hence, gene models have more input features compared to most isoform models. Therefore, gene models have more explanatory variables compared to isoform models, which may contribute to the difference in the prediction accuracy compared to isoform models. Another reason might be that the variance for isoform expression is higher than that observed at the gene level, since up and down regulation of isoforms may cancel out at the gene level. This averaging in the gene expression level, may contribute to the robustness of the gene model.

Isoform annotation and expression are informative biological priors for gene regulation modeling and expression prediction

Accurate gene regulation inference from high-throughput data is a big challenge in systems biology, for which many existing algorithms reach limited accuracy. This is mainly because of the dimensionality of the data, low sample size, inherent noise, and sparsity in high-throughput biological assays. Exploiting other sources of prior knowledge is one of the ways to address these issues (Ghanbari, Lasserre, and Vingron, 2015; Schäfer and Strimmer, 2005; Friedman et al., 2000). Here, we demonstrate that isoform expression level and annotation is a rich source of prior knowledge regarding transcriptional and post-transcriptional activities of a gene. Our new approach is the first one to make full use of this information.

POSTIT fully exploits isoform level priors, consequently addressing the dimensionality and sample size problem to a large extent and leading to more accurate gene expression prediction in challenging settings. To this end, it integrates gene expression, isoform expression, and annotation into a single learning framework for accurate gene regulation modeling and expression prediction (Eq. 5.4). Depending on the application, it provides an opportunity to conduct the downstream analysis at the gene level, transcript level, or both. This is important for two reasons. First, it leads to more accurate and robust gene expression prediction in a low sample size setting by incorporating informative biological priors, thus increasing the practical applicability of our algorithm. Second, it combines and leverages two complementary views (gene vs. isoform centric analysis) in an elegant way, and increases the interpretability and resolution of regulatory network analysis.

Association of 3′-UTR and expression prediction accuracy

The power of 3′-UTR as a driver of tissue-specific gene expression may be undervalued. We showed that there is a strong connection between isoform expression level prediction accuracy and 3′-UTR length (Figure 5.14). One possible explanation is that the transcripts with longer 3′-UTRs harbor a high density of miRNA binding sites, and are hence under tight regulation. Moreover, a recent study by (Wang et al., 2012) shows that unstable isoforms usually have longer 3′-UTRs and harbor a high density of miRNA and RBP binding sites. To our knowledge, this is the first systematic study to reveal the connection between 3′-UTR length and isoform expression prediction accuracy using miRNAs. We speculate that a similar trend might exist for isoform expression regulation through RBPs.

However, this trend might not hold for all the isoforms with long 3′-UTRs. It has been found that miRNA binding sites located in the middle of 3′-UTRs are less responsive to miRNA regulation (Kim, Kim, and Baek, 2014), since the secondary structure of the 3′-UTR might occlude the accessibility of the binding sites and decrease the regulation by miRNA. However, conversely, having multiple binding sites in a long 3′-UTR can have a synergistic effect on isoform regulation.

Moreover, we found 2,300 isoforms in liver cancer, whose expression level is predictable from miRNA seeds in the 3-UTR sequence, but not using TF binding in its promoter sequence (Figure 5.15). It has been reported that, for most genes in *C. elegans*, 3′-UTRs are sufficient for regulation, and promoters are just permissive for expression (Merritt et al., 2008). It would be interesting to understand how these dynamics are controlled in disease and cell development. Multiple methods have been developed to annotate or quantify 3′-UTRs in different tissues (Jan et al., 2011). Hence, our method can make full use of 3′-UTR data, and provide an opportunity to elucidate novel regulatory functions of 3′-UTRs in controlling transcriptome and proteome dynamics.

Regulatory connections are more tissue-specific than trans acting regulatory factors and target isoforms

Another novel aspect that we investigated is predictive power of the models across different cancers. Isoform-specific models which are trained in one cancer do not generalize well to other cancer (Figure 5.16). We demonstrated this for the transcriptional and miRNA-mediated post-transcriptional models. We observed significant reduction in their generalization performance (Figure 5.16). This indicates that although the regulators are shared between different cancers that compared, the regulatory connections are more tissue-specific than regulators and target isoforms. A recent coexpression-based network analysis, which integrated GTEx RNA-seq data across 38 different tissues and protein-protein interaction indirectly confirms our finding (Sonawane et al., 2017). We showed that pre-trained linear models reach limited accuracy in the corresponding isoform expression level in other cancers. Among them, the miRNA-based model demonstrated relatively better performance as compared to TF-based model. Many factors might contribute to the poor generalization of the pre-trained models. Regarding the transcriptional models, our modeling approach only considers TF binding profile at the proximal promoter region of each isoform. However, it is possible to accurately predict isoform expression level using accessible proximal promoter within a cancer, but it does not generalize to cross cancers. Transcriptional gene regulation are controlled at multiple levels, and our modeling approach does not consider epigenetic modification at core promoters. DNA methylation and histone marks also influence TF binding preferences to core promoter and enhancer region, thus contributing to the gene expression changes. Moreover, distal regulatory elements such as enhancers show very tissue-specific activities. Thus they modulate gene expression in a very tissue-specific manner. We speculate that the generalization performance of pre-trained transcriptional models might improve if these factors could be integrated into the model.

Regarding miRNA mediated post-transcriptional model, we also observe a high reduction in generalization performance; however the models we considered, at least 80% of their associated miRNAs were expressed between two tissues. We can think of many reasons including the unknown complexity of the miRNA target regulation, transitivity effect of TFs, and tissue-specific distribution of the miRNA. It is possible that the underlying distribution that generates miRNAs expression level is very tissue-specific as well, hence pre-trained models fail to explain isoform expression level using miRNA because they are trained on tissue-specific input data distribution.

MiRNA-based models explain the variance in the isoform expression level to large extent (Figure 5.16). Although this might initially look surprising, most of the transcriptome is coexpressed. Hence, it is more likely that the same set of TFs coregulates miRNA and its associated targets. Moreover, this coregulation might also be very tissue-specific, and pre-trained models are already overfitted to these coregulation effects, which could lead to poor generalization performance. Currently, we do not remove transitivity effects (indirect regulations) from our modeling approach. We expect that some part of the isoform prediction accuracy by miRNA is due to transitivity effects by TFs. We speculate that, if we correct our predictive models for these confounding factors, maybe the cross-cancer generalization performance of miRNA-based model will improve.

Nevertheless, the generalization performance of the miRNA-based models is higher than transcriptional models. Nevertheless, this might means suggest that transcriptional regulation is more tissue-specific than miRNA-mediated post-transcriptional regulation.

5.5 Conclusion and outlook

POSTIT is a fast and scalable algorithm for modeling and predicting gene and associated isoforms expression level. It fully exploits the gene and transcript isoform structure and expression level and shares the regulatory information between different tasks, thus leading to to more accurate gene and isoform regulation inference. The POSTIT approach is successful when the annotation of isoforms is neither too similar nor too different. If they are too different, it might be difficult transfer knowledge between them, and it reduces to single-task learning. On the other hand, if isoforms have the same annotations (same alternative 3'-UTR and promoter), then it might be a good idea to combine the training data and learn a single model for these isoforms.

Another critical factor is the difficulty of learning the isoform model concerning the number of parameters and available training data. More specifically, if the problem is hard to learn, then sharing regulatory information between transcripts of a gene helps. One way to test this is to compute the expression prediction accuracy as a function of a number of the training data point. If the performance of the isoform models saturates very fast, means that the problem is easy to learn. Therefore, exploiting cross isoform commonalities will not help much. A similar idea can be applied to check the similarity of regulation between the two isoforms. Accordingly, we can train an isoform-specific model and test on the data from another isoform of the gene. Subsequently, we can plot the expression prediction saturation curve for each isoform. This can give us a useful measure to test whether transferring regulatory information between two isoforms is beneficial.

Our approach proposed approach can be extended in multiple directions. First, from the algorithmics perspective, one can try to infer the task similarities from data and, hence, enforce the similarity of the cross task parameters accordingly. Moreover, it is easy to modify our optimization framework to learn whole regularization path, which can speed up hyper-parameters selection.

Second, from modeling perspective, it is not really clear which aspects of isoform expression level is controlled by proximal promoter, distal enhancers, and 3'-UTR. In oue anlysis, we showed that there are 2300 transcript isoforms in human liver whose expression is mainly regulated by 3'-UTR. To what extent these isoforms are tissue-specific and their distrubtion in other tissues yet to be understood.

Although gene is a useful concept, it is an abstraction of the underlying biological processes because regulatory mechanism works at the level of isoform and not the gene. Different isoforms of a gene might have various functions, which are even sometimes oppose to each other. We believe that reconstruction and analysis of the regulatory network at isoform level will provide an enhanced picture of gene regulation in different tissues, diseases, development, etc. There are many regulatory mechanisms that can not be identified if the analysis is conducted gene level. To mention a few, recent reports have demonstrated that, many diseases-associated variants change the ratio of the isoforms of a gene (Li et al., 2013b) rather than gene expression. Moreover, alternative transcript isoform usage might alter the gene function in different tissue, but might not significantly change the associated gene expression level, hence it is more likely that we will miss these regulatory mechanisms at the gene level.

We presented an algorithm that reconstructs the transcript isoform regulatory network by integrating different high-throughput assay, but unlike gene regulatory network, isoform regulatory network faces a unique challenge: the lack of functional annotation for isoforms. Currently available resources such as gene ontology and pathway annotation, provide only gene level function annotation, and hence lose the fine-grained functional information at the isoform level. Second, the annotation of the transcript isoforms is also far from perfect, and different transcript assembly algorithms output might not agree to a large extent with each other. Nevertheless, future researches will fill in the gaps (Li et al., 2013b; Li, Omenn, and Guan, 2015), thus facilitating the functional analysis of isoform regulatory network at transcript isoform and hopefully will further our understanding of regulatory mechanisms that contribute to a specific cellular phenotype.

Chapter 6

Large-scale inference of competing endogeneous RNA networks with sparse partial correlation

The main **contributions** in this chapter are the following:

- We propose a scalable algorithm for inference of competing endogenous RNA (ceRNA) network from transcriptomics data.
- We derive a theory for null distribution simulation for quantifying miRNA mediated correlation.
- We present an easy to use Bioconductor package for ceRNA network inference.

Contributions: I had a major contribution to the design and development of the algorithm. I made all the derivations for the null model simulation and implemented the first working version of the algorithm and designed the simulation experiments. Furthermore, I contributed to data analysis, interpretation and writing the manuscript. Dennis Kostka suggested the idea of using the Schur complement for covariance matrix simulation. Moreover, he helped to derive the null model simulation for the case $microRNA = 1$. Markus List led the data analysis, wrote the Bioconductor package, and contributed to writing the manuscript. Marcel Schulz supervised the project and contributed to writing the manuscript. This chapter will appear at ISMB-ECCB Proceedings 2019.

6.1 Introduction

In the previous chapter, we introduced an algorithm for modeling and analysis of transcriptional and post-transcriptional isoform regulation. We showed how we could improve condition specific miRNA target prediction, and how regulatory elements in 3'-UTR play roles in miRNA mediated regulation. In this chapter, we will build on top of that to study another mode of RNA regulation.
However, the common belief is that miRNA regulates the target transcript expression; the relationship between transcript and miRNA could be reciprocal. This means the level of one transcript can control the level of other transcripts through shared miRNA. More specifically, RNA transcripts with miRNA recognition elements (MREs) compete for shared miRNA for binding, hence create transcriptome-wide RNA cross-talk.

Notably, genes sharing binding sites for the same miRNA(s) compete over a limited pool of miRNA molecules, giving rise to a complex gene-regulatory network of

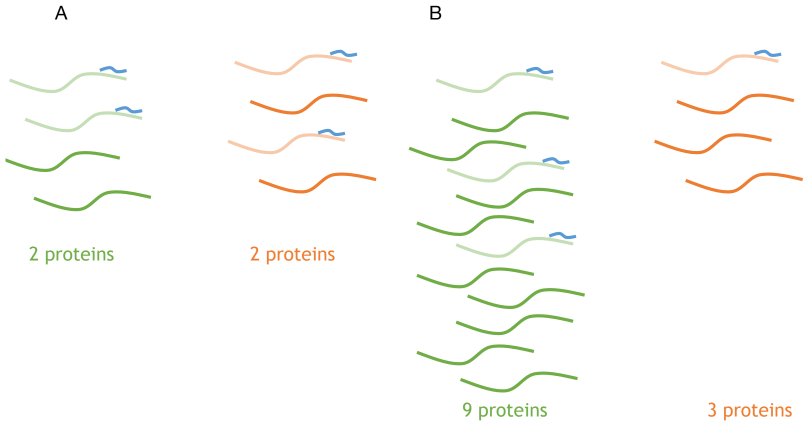


FIGURE 6.1: ceRNA interaction between two genes that have the same miRNA binding sites. (A) both green and red genes have the same expression level; hence they are regulated by miRNA by a similar amount. (B) Green gene is overexpressed, therefore binds more miRNAs and protects red transcript from translational repression, hence protein output of red transcript increases.

competing endogenous RNAs (ceRNAs) (Tsang, Ebert, and Oudenaarden, 2010). A number of cancer-associated genes have been shown to act as ceRNAs (Arvey et al., 2010b; Salmena et al., 2011; Tay, Rinn, and Pandolfi, 2014), including *PTEN* (Poliseno et al., 2010), *CD44* (Jeyapalan et al., 2011), *ESR1* (Chiu et al., 2015), *BRAF* (Karreth et al., 2015), *KRAS* (Poliseno et al., 2010), *MYCN* (Powers et al., 2016) and *HULC* (Wang et al., 2010). This evidence has sparked interest in developing systematic methods for inferring ceRNA interactions from gene and miRNA expression data, reviewed in (Le et al., 2016).

The different methods can be broadly categorized into methods that use (i) only static information, such as the number of miRNA binding sites or binding energy or (ii) methods that use condition-specific information in addition such as expression or Clip-data. One of the most commonly used methods in category (i) is based on the idea to assess the probability that two mRNAs share miRNAs and their binding sites, and to then highlight cases where this probability is much higher than expected by chance, for example by using the hypergeometric test (Li et al., 2013a).

With the emergence of large-scale studies providing gene and miRNA expression data for hundreds of samples, a number of methods of category (ii) have been developed (Le et al., 2016). (Sumazin et al., 2011) proposed the use of conditional mutual information (CMI) for estimating the effect of a miRNA on a gene-gene interaction in their method HERMES. The advantage of this approach is that it measures nonlinear associations, but estimation of significance is done using permutations, later implemented as part of the CUPID software (CUPID step III) (Chiu et al., 2015). Recently the JAMI software has improved the runtime of the extensive CMI computation compared to CUPID (Hornakova et al., 2018), but runtime is still a limiting factor for this approach in applications to very large datasets.

This issue has motivated the use of conceptually simpler and fast linear correlation-based methods, for instance, restricting to only gene-gene correlation values (Du et al., 2016; Xu et al., 2015), gene-miRNA correlation (Zhang et al., 2017) or correlations within triplets of two genes and one miRNA (Wang et al., 2015; Liu et al., 2017). However, in contrast to CUPID, these approaches do not quantify the contribution of the miRNA to the ceRNA interaction in a unified model.

(Paci, Colombo, and Farina, 2014) overcame this issue with the definition of sensitivity correlation (*scor*), which has similarities to the CMI-based approach. Linear partial correlation can be used to quantify the remaining correlation between two genes after accounting for the effect of one miRNA. *scor* is then defined as the difference between gene-gene correlation and partial correlation and thus quantifies the contribution of the miRNA in the regulation of two genes. Similar to CMI, *scor* considers the impact of miRNA regulation on both genes in a single mathematical model and is thus more powerful than the methods proposed in (Wang et al., 2015; Zhang et al., 2017; Du et al., 2016; Xu et al., 2015; Liu et al., 2017). Unlike CMI, however, *scor* computation is based on efficient estimators of covariance matrices for computing partial correlation and thus allows large-scale ceRNA network inference as demonstrated by (Zhang et al., 2016), who inferred lncRNA-mRNA related ceRNA networks for 12 different cancer types.

While the estimation of the *scor* coefficients is efficient, no theory for the computation of the null distribution of these values exists. Therefore, previous work relied on *ad hoc* approaches. (Paci, Colombo, and Farina, 2014) selected the top 5% of *scor* coefficients for downstream analysis, disregarding significance testing. (Zhang et al., 2016) addressed this issue by generating a null distribution using permutations based on randomly selected lncRNA-miRNA-mRNA triplets. This null distribution was then used to obtain empirical *p*-values.

We have identified a number of issues with the current approaches that use *scor*. First, current correlation-based approaches assume independence between *scor* values and the gene-gene correlation. However, as we show in this work, the distribution of *scor* coefficients is strongly affected by gene-gene correlation (Fig. 6.2A and Supp. Fig. 1). Thus, previous studies that have used *scor* values have been biased.

Second, we note that many ceRNAs are regulated by several miRNAs. Neglecting joint contributions, many significant ceRNA interactions may be missed. The CUPID approach considers that ceRNA interactions may be mediated by several miRNAs in conjunction. To accommodate this, CUPID pools *p*-values obtained from individual ceRNA triplets (Chiu et al., 2015). We propose that the contributions of multiple miRNAs should be part of the ceRNA inference model to optimally account for miRNA covariance effects.

Here we present a unified mathematical approach that addresses the above issues. We have developed a Bioconductor/R package called Sparse Partial correlation ON Gene Expression (SPONGE). At the core of SPONGE is a new mathematical framework that is a generalization of *scor* values for more than one miRNA, which we call multiple miRNA sensitivity correlation (*mscor*). Assessing the significance of *mscor* coefficients is difficult due to biases of gene-gene correlation, number of samples, and the number of miRNAs. Therefore, we have developed a novel strategy for simulating background distributions that accommodate the aforementioned factors and for inferring *p*-values for *mscor* coefficients efficiently. Due to SPONGE's efficiency, we were able to perform an analysis of the complete human transcriptome across 31 different cancer types combining over 10,000 paired gene and miRNA expression samples using data from The Cancer Genome Atlas (TCGA). Our analysis highlights the potential of ceRNA network inference for hypothesis generation by revealing extensive ceRNA cross-talk. Some of the key regulators have already been reported as ceRNAs while others potentially represent novel biomarkers and drug target candidates.

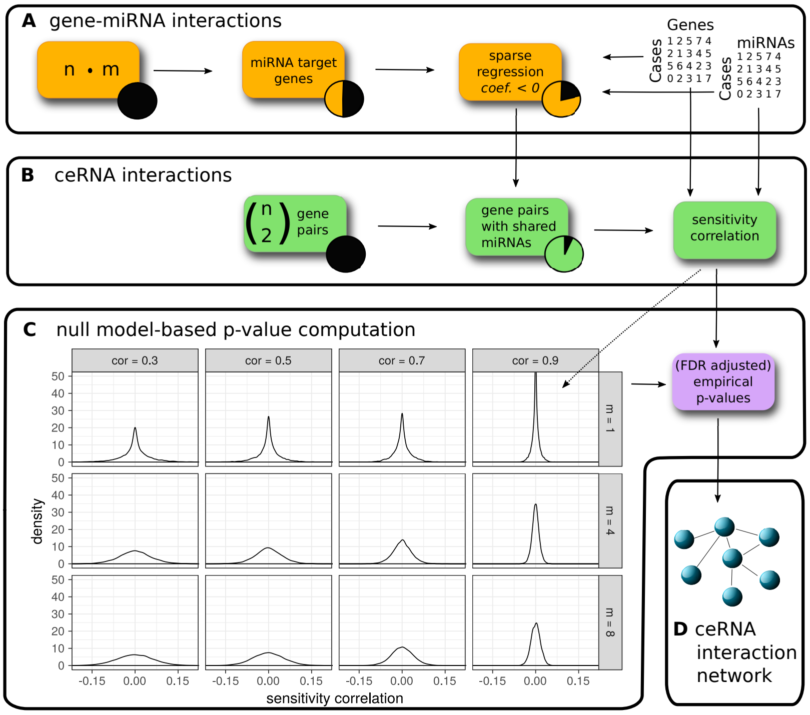


FIGURE 6.2: Overview of the SPONGE workflow. (A) Predicted and/or experimentally validated gene-miRNA interactions are subjected to regularized regression on gene and miRNA expression data. Interactions with negative coefficients are retained since they indicate miRNA induced inhibition of gene expression. (B) We compute sensitivity correlation coefficients for gene pairs based on shared miRNAs identified in (A). (C) Given the sample number, we compute empirical null models for various gene-gene correlation coefficients (k) and number of miRNAs (m). Sensitivity correlations coefficients are assigned to the best matching null model and a p-value is inferred. (D) After multiple testing correction, significant ceRNA interactions can be used to construct a genome-wide, disease or dataset-specific ceRNA interaction network.

6.2 Methods

SPONGE overview

The objective of SPONGE is to infer a ceRNA interaction network from gene and miRNA expression data of paired samples. In theory, inferring a genome-wide ceRNA network with n genes entails considering $\binom{n}{2}$ interactions for all pairwise combinations. In practice, only gene pairs with shared miRNAs need to be considered. First, SPONGE identifies for each gene those miRNAs that are likely to have a regulatory effect (Fig. 6.2A). Second, we filter for gene pairs with shared miRNAs and determine their ceRNA interaction scores (Fig. 6.2B). Third, we assess the significance of each ceRNA interaction using a series of null models (Fig. 6.2C) adjusting for confounders. Finally, significant interactions are retained for constructing a ceRNA interaction network (Fig. 6.2D). In the following, we describe each of these steps in detail.

Step 1: Identifying relevant miRNA-gene interactions

SPONGE identifies relevant gene-miRNA interactions in two stages. First, we retain only miRNA-gene pairs for which we have general evidence from external predictive or experimental sources. SPONGE allows for an arbitrary number of data sources to be combined.

Second, we test if the gene and miRNA expression data provides support for these interactions, since we expect many of the putative miRNA-gene interactions in particular to be false positives (Pinzón et al., 2017). Negative correlation of gene and miRNA expression can provide evidence for a miRNA-gene regulation. However, many miRNAs might target a single gene. To take this into account, and to identify the most likely miRNA regulators of each gene, we use regularized regression.

We build an Elastic net regularized linear regression model with the expression of gene g as the dependent variable and the expression of miRNAs $Z' \in Z$ as explanatory variables, where Z' are miRNAs predicted or experimentally shown to target g. Elastic net balances lasso (L1) and ridge (L2) penalties using a linear combination of both denoted as a weight factor α. We build a range of Elastic net models to optimize the parameters for $\alpha = 0.1, 0.2, ..., 1.0$ and the optimal shrinkage parameter λ via 10-fold cross validation using the *glmnet* package (Friedman, Hastie, and Tibshirani, 2010a). We select the best model based on the residual sum of squares. Since miRNAs with positive coefficients are likely caused by effects other than miRNA repression, we retain only miRNAs with negative coefficients, which was previously shown to work well (Muniategui et al., 2012; Schulz et al., 2013). Moreover, SPONGE offers a user-definable coefficient threshold for discarding miRNAs with negligible impact on gene expression (default < -0.05).

In summary, we identify for each gene condition-specific miRNA regulators. This leads to a dramatic reduction of gene pairs that share miRNAs (Fig. 6.2B) compared to using all predicted miRNA-gene interactions and reduces the runtime of SPONGE. In the next step, we determine the effect strength of ceRNA interactions.

Step 2: Computing miRNA induced correlation coefficients

In order to compute ceRNA interaction between gene 1 and gene 2, we are interested to quantify how of much observed correlation between them are induced by shared miRNA between these two genes. To quantify miRNA induced correlation (miRNA effect), we use the different of observed correlation with partial correlation between these gene 1, and gene 2 as follows(Paci, Colombo, and Farina, 2014):

$$scor(g_1, g_2, m) \quad = \quad cor(g_1, g_2) - pcor(g_1, g_2|m)\,. \tag{6.1}$$

$pcor(g_1, g_2|m)$ explains away the effect of miRNA m in explaining the observed correlation between g_1 and g_2. Hence, different of correlation and partial correlation quantify the miRNA contoribution in inducing correlation between g_1 and g_2.

Note that this approach does not account for a combinatorial effect of several miRNAs. Consequently, strong ceRNA interactions mediated by several moderate miRNA regulators can not be detected.

We thus propose to extend the definition of sensitivity correlation considering the effect of multiple miRNAs M for the computation of the partial correlation. In this way, we implicitly incorporate the effect of miRNA-miRNA cross-correlation. We call this multiple miRNAs sensitivity correlation (*mscor*):

$$mscor(g_1, g_2, M) \quad = \quad cor(g_1, g_2) - pcor(g_1, g_2|M)\,, \tag{6.2}$$

where $M = m_1, ..., m_i$ and i is the number of shared miRNAs between g_1 and g_2. We compute *mscor* coefficients efficiently using the R package *ppcor* (Kim, 2015). In the next step, we establish the significance of each *mscor* coefficient.

Step 3: Sampling from the *mscor* null distribution with respect to important parameters

(Zhang et al., 2016) proposed to establish the significance of *scor* coefficients by means of sampling a background distribution from random triplets. This approach, however, disregards that correlation coefficients have smaller variance when the coefficient is high (Fisher, 1915). Moreover, it can be expected that the significance of sensitivity correlation values is linked to the number of samples and the number of miRNAs involved.

To accommodate these biases, we propose a novel algorithm to study the null distribution of *mscor* coefficients.

6.3 A null model for sparse partial correlation

SPONGE systematically computes the difference between the correlation of two genes $g1$ and $g2$ and the partial correlation of those genes after accounting for the expression of miRNAs that potentially regulate both genes. This measure, also called multiple miRNA sensitivity correlation (*mscor*), serves to quantify the strength of the ceRNA relationship between two genes, i.e.,

$$mscor = cor(g_1, g_2) - pcor(g_1, g_2|mir_1, ..., mir_m)$$

Our null hypothesis is that the miRNAs have no effect on the correlation of the two genes, i.e. $mscor = 0$ and thus we are interested in the significance of observing $mscor > 0$. Here, we illustrate how numerical simulations can be used to model the expected distribution of *mscor* under this null hypothesis for varying numbers of miRNAs m, varying numbers of samples n and different correlations of $g1$ and $g2$. We will demonstrate that all of these parameter influence the null distribution and thus have to be taken into account. The result of the simulations are used in *SPONGE* to compute p-values for *mscor* coefficients given the number of samples in the data set and given the number of miRNAs considered for each pair of genes.

6.3.1 Computing conditional covariance matrix using Schur complement

Given a partitioned vector $Z = [g_1, g_2, mir_1, ..., mir_m]$, we can write its correlation matrix as

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$$

where

$$R_{11} = \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix}$$

is the correlation matrix between the first two entries of Z. In order to compute the conditional co-variance between the first two entries (g_1, g_2) of Z given $(mir_1, ..., mir_m) \in Z$, we compute the Schur complement of R/R_{22} as follows:

$$R/R_{22} = R_{11} - R_{12}R_{22}^{-1}R_{12}^T = \begin{pmatrix} 1-a & r_{12}-b \\ r_{12}-b & 1-c \end{pmatrix}$$

where

$$\begin{aligned} a &= v_1^T R_{22}^{-1} v_1 \\ b &= v_1^T R_{22}^{-1} v_2 \\ c &= v_2^T R_{22}^{-1} v_2 \end{aligned} \tag{6.3}$$

We define the vectors v_1 and v_2 as correlation vector of $g1$ and $g2$ with $[mir_1, ..., mir_m]$, respectively. Hence $(R/R_{22})^{-1}$ gives us partial correlation as follows:

$$r_{12.m} = (r_{12} - b)(1-a)^{-1/2}(1-c)^{-1/2} \tag{6.4}$$

For the null model, we need to show that partial correlation is equal to correlation. In other words, we need to solve:

$$r_{12.m} = r_{12}$$

Our strategy to achieve this is to first sample $K = r_{12}$ and R_{22} with $-1 \leq K \leq 1$ and $\forall x \in R_{22} : -1 \leq x \leq 1$. We then look for sensible values for v_1 and v_2 such that sensitivity correlation $(r_{12} - r_{12.m})$ is equal to zero. Replacing $r_{12.m}$ with r_{12} yields

$$r_{12} = (r_{12} - b)(1-a)^{-1/2}(1-c)^{-1/2} \tag{6.5}$$

6.3.2 Sampling Correlation Matrices

In the following, we consider two sampling strategies, one for the case of $m = 1$ and one for the case of $m > 1$:

The case of $m = 1$:

With only one miRNA, Eq. (6.5) can be simplified as

$$r_{12} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \tag{6.6}$$

Substituting a, b and c from (6.3) yields

$$r_{12} - r_{13}r_{23} = r_{12}\sqrt{(1-r_{13}^2)(1-r_{23}^2)} \tag{6.7}$$

We further substitute $K = r_{12}$ as well as $Z = r_{13}$:

$$
\begin{aligned}
K^2 - 2KZr_{23} + Z^2r_{23}^2 &= K^2 - K^2r_{23}^2 - Z^2K^2 + K^2Z^2r_{23}^2 \\
Z^2r_{23}^2 + K^2r_{23}^2 - K^2Z^2r_{23}^2 - 2KZr_{23} + K^2 - Z^2K^2 &= 0 \\
(Z^2 + K^2 - K^2Z^2)r_{23}^2 - (2KZ)r_{23} + Z^2K^2 &= 0
\end{aligned} \tag{6.8}
$$

We can fix Z by uniform sampling from $[-1, 1]$ and solve above quadratic equation:

$$
\begin{aligned}
A &= K^2 + Z^2 - K^2Z^2 \\
B &= -2KZ \\
C &= Z^2K^2
\end{aligned} \tag{6.9}
$$

This allows us to compute r_{23}:

$$
r_{23} = \frac{-B + / -\sqrt{4AC}}{2A} \tag{6.10}
$$

and to construct valid examples of correlation matrices R in which $mscor = 0$.

The case of $m > 1$:

When considering several miRNAs obtaining valid instances of R is considerably more complex. Instead of considering individual values r_{13}, r_{23} and r_{12} we have much more degrees of freedom. As before, we sample $K = r_{12}$ from $[-1, 1]$. Moreover, we sample a positive semi-definite correlation matrix R_{22} that expresses the correlation between different miRNAs. Diagonal elements of R_{22} are 1 and off-diagonal elements are $-1 \leq r_{ij} \leq 1$. More importantly R_{22} should be positive definite which means that:

$$
x^t R_{22} x > 0
$$

$$
\forall x \in R^m
$$

To obtain sensible values for $v1$ and $v2$ we modify our problem. Let the vectors u_1 and u_2 be defined as

$$
\begin{aligned}
u_1 &= R_{22}^{-1/2} v_1 \\
u_2 &= R_{22}^{-1/2} v_2
\end{aligned} \tag{6.11}
$$

This allows us to rewrite a, b and c from (6.3) as

$$
\begin{aligned}
a &= ||u_1||^2 \\
b &= u_1^T u_2 = \sqrt{a}\sqrt{c}\cos\theta \\
c &= ||u_2||^2
\end{aligned} \tag{6.12}
$$

We also define $\lambda = \cos\theta$ and substitute a, b, and c in (6.3) in the form of :

$$
[(K^2 + \lambda^2)||u_1||^2 - K^2]||u_2||^2 + (2\lambda K||u_1||)||u_2|| - K^2||u_1||^2 = 0 \tag{6.13}
$$

where K, λ are scalars. Our strategy is the following: we sample $u_1 \in \mathcal{R}^m$, which allows us to compute $||u_1||$. This allows us to determine $||u_2||$ since, knowing that $||u_1||^2 = u_1^t u_1 = constant$, we can simplify (6.13) and solve a quadratic function w.r.t the $||u_2||$.

We define:

$$\begin{aligned} A &= [(K^2 - \lambda^2)||u_1||^2 - K^2] \\ B &= 2\lambda K||u_1|| \\ C &= -K^2||u_1||^2 \end{aligned} \tag{6.14}$$

Then we can rewrite equation (6.13) as :

$$A||u_2||^2 + B||u_2|| + C = 0\,. \tag{6.15}$$

Solving this equation w.r.t $||.||$ of u_2 will give us the following solution:

$$||u_2|| = \frac{-B + / -\sqrt{B^2 - 4AC}}{2A} \tag{6.16}$$

Suppose F is one of solutions to above equation. To construct a valid instance of R, we now have to sample a vector u_2 that respects the norm we just determined and has the correct angle to u_1. Finally, since elements of v_2 are correlations, we have to make sure to impose the following constraint on u_2

$$-R_{22}^{-1/2}1 \leq u_2 \leq R_{22}^{-1/2}1 \tag{6.17}$$

To test if our chosen $u2$ has the correct angle, we can use the two alternative definitions of the dot product

$$\begin{aligned} \cos\theta||u_1||||u_2|| &= \sum_{i=1}^{n} u_{1_i} u_{2_i} \\ \lambda||u_1||||u_2|| &= \sum_{i=1}^{n} u_{1_i} u_{2_i} \end{aligned} \tag{6.18}$$

We can now obtain a valid solution of u_2 by randomly sampling $m - 1$ elements of u_2 to obtain $u_{2_{\setminus m}}$. The last elements of u_1 and u_2 are defined as u_{1_m} and u_{2_m}, respectively.

To compute a valid element u_{2_m}, we rewrite (6.18) as follows

$$u_{2_m} = \frac{\lambda||u_1||||u_2|| - \sum_{i=1}^{m-1} u_{1_i} u_{2_i}}{u_{1_m}} \tag{6.19}$$

Since $||u_2||$ depends on u_{2_m} we substitute it with

$$||u_2|| = \sqrt{\sum_{i=1}^{m-1} {u_{2_i}}^2 + u_{2_m}^2} \tag{6.20}$$

This yields

$$u_{2_m} = \frac{\lambda||u_1||\sqrt{\sum_{i=1}^{m-1} {u_{2_i}}^2 + u_{2_m}^2} - \sum_{i=1}^{m-1} u_{1_i} u_{2_i}}{u_{1_m}} \tag{6.21}$$

We define

$$\beta = \sum_{i=1}^{m-1} u_{1_i} u_{2_i} u_{1_m} \tag{6.22}$$

and rewrite (6.21) as

$$(u_{1_m}^2 - \lambda^2||u_1||^2)u_{2_m}^2 + (2\beta u_{1_m})u_{2_m} + \beta^2 - \lambda^2||u_1||^2 \sum_{i=1}^{m-1} {u_{2_i}}^2 = 0 \tag{6.23}$$

This quadratic equation can be solved as before with

$$A = u_{1_m}^2 - \lambda^2||u_1||^2 \tag{6.24}$$

$$B = 2\beta u_{1_m} \tag{6.25}$$

$$C = \lambda^2||u_1||^2 \sum_{i=1}^{m-1} {u_{2_i}}^2 \tag{6.26}$$

yielding

$$u_{2_m} = \frac{-B + / -\sqrt{B^2 - 4AC}}{2A} \tag{6.27}$$

We now can construct a vector u_2' with the desired angle to u_1. However, it does not have the correct norm yet. We thus need to scale u_2 to the norm of F we have previously determined in (6.15. This operation will not affect the angle between both vectors.

$$u_2 = \frac{u_2'}{||u_2'||} F \tag{6.28}$$

Finally, we have to check if the final vector u_2 is still compatible with R with respect to the correlation constraint [-1,1]. We can test this by rewriting (6.17) as

$$\begin{bmatrix} -c_1 \\ -c_2 \\ . \\ . \\ -c_m \end{bmatrix} \leq \begin{bmatrix} u_{2_1} \\ u_{2_2} \\ . \\ . \\ u_{2_m} \end{bmatrix} \leq \begin{bmatrix} c_1 \\ c_2 \\ . \\ . \\ c_m \end{bmatrix} \tag{6.29}$$

Should u_2 violate these constraints, we need to repeat the sampling procedure. The same is true if no solutions can be obtained for the quadratic equations in (6.10) or (6.16) as well as (6.27).

Our sampling procedure for the case $m > 1$ can be summarized in the following steps

1. Fix v_1, R_{22} and λ by random sampling.

2. Compute a valid solution for $F = ||u_2||$ via (6.16).

3. If the previous step does not yield F, go to step 1.

4. Randomly select $m - 1$ components of vector u_2 and compute a solution for u_{2_m} via (6.27).

5. If u_{2_m} is without valid solution go to step 1.

6. Construct u'_{2_m} and scale it to F to obtain u_2.

7. Test if u_2 fulfills the constraints in (6.29).

8. Construct and return R.

6.4 Simulation Results

We used the above algorithm to sample sets of covariance matrices for all combinations of the parameters gene-gene correlation (0.2 - 0.9 in steps of 0.1) and number of shared miRNAs (1 to 8). These are incorporated in the SPONGE R package and can be used to construct null models for a given number of samples. Here, we generated null models for varying numbers of samples to study how the three parameters, gene-gene correlation, number of shared miRNAs and number of samples affect the random distribution of *mscor* coefficients. The results are depicted in 6.3.

Our null hypothesis is that the shared miRNAs M do not affect the correlation of two genes $g1$ and $g2$. Hence,

We have devised sampling strategies that enable us to find values a, b and c such that these conditions are fulfilled, allowing us to construct random covariance matrices under the null.

Most importantly, we can control the gene-gene correlation r_{12} and the number of miRNAs (via the dimensions of R_{22}) to construct a series of covariance matrices with respect to these important parameters. SPONGE uses these covariance matrices to draw random samples which are subsequently used to estimate empirical p-values for *mscor* values computed on experimental data. The details of this approach and of our sampling strategy can be found in the Supplemental Material.

The SPONGE R package provides precomputed covariance matrices for a range of gene-gene correlations and number of miRNAs. Given the number of samples in the expression data, SPONGE can efficiently construct a series of null distributions from these covariance matrices. Next, we assign each *mscor* coefficient to the closest matching null model and infer its p-value via its rank in the random distribution (Fig. 6.2D). The number of data points sampled for the null distribution determines the maximal precision of this p-value ($p > 1e6$ by default). Finally, p-values are adjusted for multiple testing within each null model using the method by (Benjamini and Hochberg, 1995b).

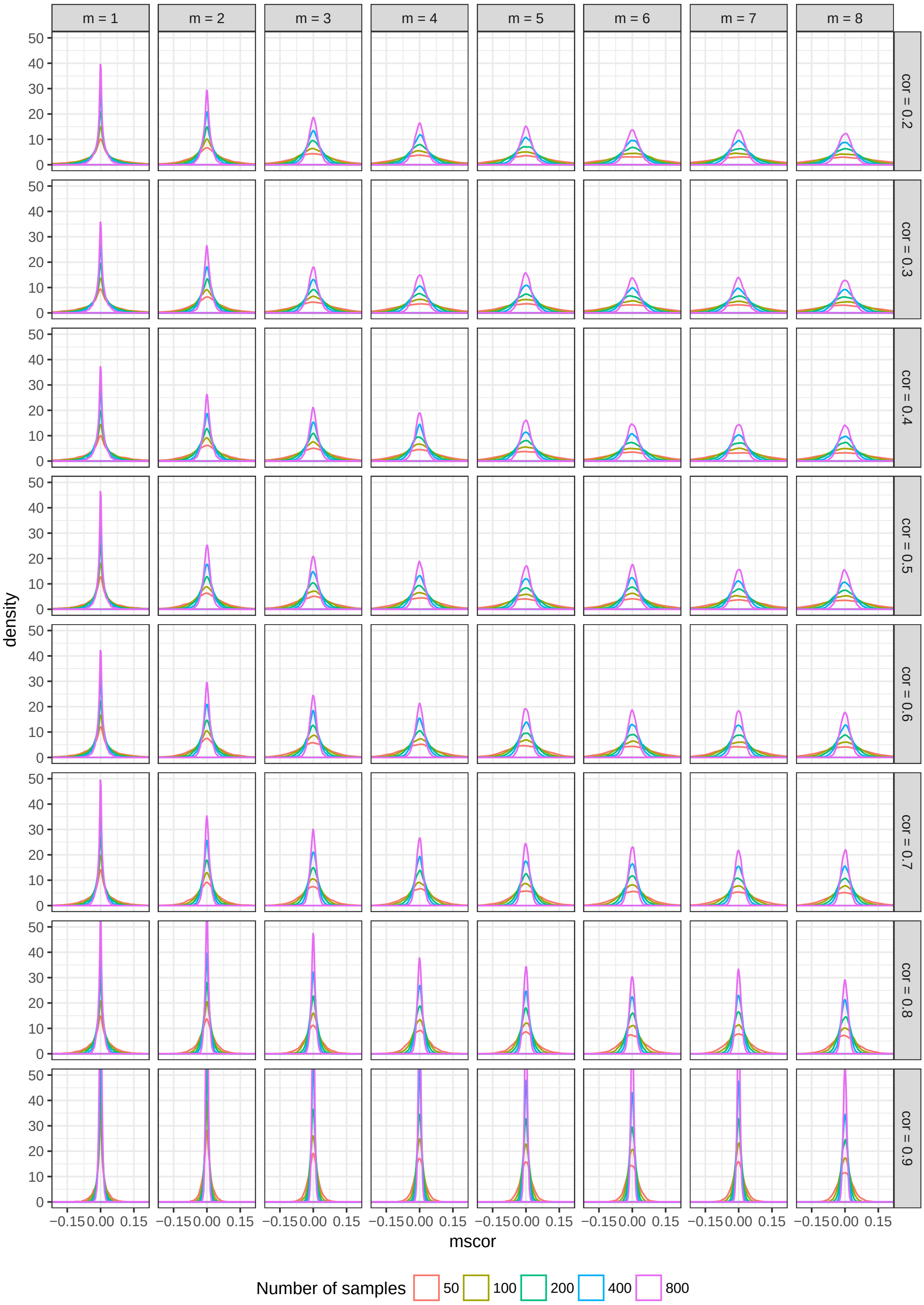


FIGURE 6.3: Density plots of mscor values randomly sampled in the process of generating null models for SPONGE. Different panels show that the distribution depends strongly on the number of shared miRNAs (1-8) and gene-gene correlation (0.2 - 0.9 in steps of 0.1) as well as on the sample number (50, 100, 200, 400 and 800).

Step 4: Constructing a ceRNA network

We filter ceRNA interactions returned by SPONGE by a user-defined significance threshold (FDR < 0.01 by default) and subsequently construct a ceRNA interaction network $N = (V, E)$, where nodes V correspond to genes participating in significant ceRNA interactions and edges correspond to significant ceRNA interactions between two genes.

Using SPONGE to construct a pan-cancer ceRNA network

We downloaded reprocessed TCGA pan-cancer data from the TOIL project (Vivian et al., 2017) via the UCSC Xena Browser (Goldman et al., 2018). We identified 10,019 samples for which both gene and miRNA expression data were available. Next, we performed log2 transformation and discarded genes and miRNAs not expressed in more than 80% of samples as well as genes and miRNAs with expression variance < 0.5.

To consider both coding and non-coding miRNA-gene interactions, we downloaded sequence-based predictions of two methods, namely TargetScan (Agarwal et al., 2015b) (v.7.1, downloaded 10/03/2017) and miRcode (Jeggari, Marks, and Larsson, 2012) (v.11, downloaded 10/03/2017). We included the latter since it also considers non-coding RNAs which have been shown to act as ceRNAs.

TargetScan and miRcode predict target genes for miRNA families. We thus downloaded suitable miRNA family definitions for both data sets (available at the TargetScan website `http://www.TargetScan.org/`). Note that miRcode uses the miRNA family definitions corresponding to TargetScan v.6. After mapping family ids to miRBase mature miRNA ids (MIMATs) we generated integer matrices in which genes are listed as rows and miRNAs are listed as columns. Each entry of the matrix represents the number of binding sites for the corresponding interaction.

To consider experimental evidence for miRNA-gene interactions, we obtained data sets from miRTarBase (v.6, downloaded 13/03/2017) (Chou et al., 2016) for coding and lncBase (v.2, downloaded 13/03/2017) (Paraskevopoulou et al., 2016) for non-coding genes and generated input matrices as described above.

Matrices, for gene and miRNA expression and miRNA-gene interactions were analyzed with SPONGE. Significant ceRNA interactions were used to construct the pan-cancer ceRNA network.

Runtime Analysis

In order to compare against the runtime of the CMI-based approach of CUPID (Chiu et al., 2015), which similarly uses paired gene and miRNA expression to estimate gene-miRNA-gene triplets, we used the JAMI software (Hornakova et al., 2018). JAMI is a fast reimplementation of CUPID step III that leverages parallelization. We compared the runtime of JAMI to that of SPONGE (without step1: the regression filter) for a fair comparison. We used a subset of the pan-cancer dataset with 200 genes, which form ca. 80,000 gene-miRNA-gene triplets. We ran both tools with default parameters in parallel mode with 16 cores with varying number of samples and genes.

Survival Analysis

For assessing the impact of gene or miRNA expression on the survival probability, we downloaded right-censored TCGA survival data of TCGA patients from the

UCSC Xena Browser (Goldman et al., 2018). We divided patients into two groups based on the 50% quantile of the expression vector. Survival probability was computed in R using the survfit function in the R package *survival* (Therneau, 2015). *p*-values were computed using the function survdiff in the same package. survdiff tests for significant differences of survival curves using the χ^2 statistic. Kaplan Meier plots were generated using the ggsurv function of package *survminer*.

To test for the enrichment of survival genes in a list of top candidates ranked by degree, we used the following strategy. First, we computed survival *p*-values based on expression data for all genes as outlined above using the survdiff function. Second, we classified genes into survival-associated and -unassociated (background) genes (FDR < 0.001, (Benjamini and Hochberg, 1995b) for the purpose of enrichment analysis. Third, we computed enrichment of the candidate gene set in survival-associated compared to background using the hypergeometric test in R.

6.5 Results

We have devised a method for the statistical evaluation of condition-specific ceRNA interactions from paired miRNA and gene expression data considering contributions for multiple miRNAs: called multiple miRNA sensitivity correlation (*mscor*). *mscor* is a generalization of *scor* previously defined for one miRNA by (Paci, Colombo, and Farina, 2014) (see Methods for details).

Simulated data reveals dependency of sensitivity correlation on several factors

As mentioned above, no theory existed to describe the distribution of sensitivity correlation values (Paci, Colombo, and Farina, 2014). However, we wanted to understand how the *mscor* measure is influenced by confounding factors present in ceRNA relationships: (i) the correlation of two genes, (ii) the number of miRNAs involved in the ceRNA interaction and (iii) the number of samples that are available for estimation. We developed an efficient simulation approach to explore null models in which miRNAs have no effect on the correlation of two genes, hence *mscor* is zero (see Methods and Supplemental Material for details). Our method is able to compute random covariance matrices that fulfill this null hypothesis. This allowed us to simulate data sets for a range of gene-gene correlation coefficients (0.2 - 0.9 in steps of 0.1), shared miRNAs (1 to 8) and number of samples (50, 200, 800) and thus to approximate the random distribution of the *mscor* coefficients under the null hypothesis that *mscor* is zero.

Fig. 6.2 and Supp. Fig. 1 show our simulation results, which reveal that the null distribution is strongly affected by all three tested parameters. Our findings indicate that large *mscor* coefficients are more likely to occur by chance when the gene-gene correlation is low and when the number of miRNAs increases. As expected, it is more difficult to obtain significant *mscor* coefficients with few samples as higher *mscor* values are obtained with smaller samples sizes by chance. Thus, comparing *mscor* values without proper adjustment for these parameters would prioritize low gene-gene correlation pairs, interactions with many miRNAs and lead to a bias when tests between studies with different sample numbers are compared.

The above insights led us to develop SPONGE, an R/Bioconductor package to infer ceRNA interactions between pairs of genes. We briefly outline how SPONGE facilitates this in two steps (see Methods and Fig. 6.2). First, we estimate condition-specific miRNA-gene associations from a large set of putative miRNA-gene interactions. This is done using sparse regression of paired gene and miRNA expression data obtained from many samples. Second, ceRNA interactions are predicted using *mscor* values estimated for all gene-gene pairs that share at least one miRNA from the first step. Statistical significance of *mscor* values is efficiently computed using the simulation approach described above.

Considering multiple miRNAs leads to information gain

To demonstrate the advantages of *mscor* measure over *scor*, we selected a subset of the TCGA data with 364 liver cancer samples and 1,000 randomly selected genes. *mscor* allows us to incorporate multiple miRNAs in the model and thus to detect ceRNA interactions that only become significant when several miRNAs act in concert. Fig. 6.4A shows that considering all miRNAs lead in most but not all cases to a higher *mscor* coefficient compared to the individual miRNA with highest *scor*. However, when also considering significance (FDR < 0.01), the signal to noise ratio increased and led to a clear gain in information, namely consistently higher *mscor* coefficients for multiple miRNAs. Consequently, SPONGE is able to assess the joint regulatory effect of several miRNAs in a ceRNA relationship in a condition-specific way.

Our approach correctly adjusts the p-value to the number of miRNAs involved (see Fig. 6.2C). As CUPID uses a meta analysis strategy on individual gene-miRNA-gene triplets (Chiu et al., 2015) to obtains one p-value for a set of miRNAs per gene-gene interaction, we sought to compare to such an approach for our measure. We used Fisher's popular meta analysis approach to combine p-values (Fischer, 1925) of individual miRNA triplets. Fig. 6.4B shows that aggregated p-values tend to be considerably higher in meta analysis, illustrating the loss of information and sensitivity compared to assessing significance in a joint model via *mscor*.
Our simulation suggested that ranking ceRNA interactions by the *scor* or *mscor* values would introduce a bias towards interactions with low gene-gene correlation (see Fig. 6.2C). In Fig. 6.4C, we compared the gene-gene correlation values of the top 5% ceRNA interactions sorted according to *mscor* with our FDR corrected set of ceRNA interactions. (Paci, Colombo, and Farina, 2014) used 5% as an arbitrary cutoff. We observed that SPONGE selected ceRNA interactions showed significantly higher gene-gene correlation values on average (t-test p-value $< 2.2e^{-16}$) underlining that sorting without proper correction leads to a bias.

Runtime comparison with a conditional mutual information-based approach

CMI is an alternative to partial correlation for estimating the effect and significance of a gene-miRNA-gene interaction. We compared the performance of JAMI (Hornakova et al., 2018), a fast implementation of the CMI-based approach of CUPID (Chiu et al., 2015), and SPONGE on a subset of the pan-cancer dataset. Fig. 6.5 illustrates that the SPONGE workflow can be computed fast even for large sample numbers and large number of triplets, while the runtime of JAMI increases dramatically due to the need to rank expression values and due to computationally intensive permutations that are needed for assessing the significance of CMI values. In addition, SPONGE does normally not evaluate each triplet individually, but considers all shared miRNAs in a joint model, giving rise to an additional speedup. SPONGE is thus uniquely suited to infer a genome-wide ceRNA network even on large-scale datasets such as the TCGA pan-cancer data.

The empirical null model allows strict control over the false positive rate

To study ceRNA interactions in a pan-cancer setting, we applied SPONGE to paired miRNA and gene expression data for 10,019 samples from TCGA (see Methods) combining data from 31 cancer types. A comprehensive set of putative miRNA-gene interactions was obtained by combining several sources: sequence-based predictions from TargetScan (Agarwal et al., 2015b) and miRcode (Jeggari, Marks, and Larsson, 2012) as well as experimentally validated miRNA-gene interactions from mirTarBase (Chou et al., 2016) and LncBase (Paraskevopoulou et al., 2016).

Considering all possible pairwise combinations of genes, ca. 10^9 putative ceRNA interactions can be formed. Fig. 6.6 shows how the three-step approach of SPONGE reduces this large set of putative interactions. In the first step, condition-specific gene-miRNA interactions are inferred, which reduces the set of considered ceRNA interactions to 10^8. However, many of these denote spurious ceRNA interactions that do not pass our selected significance threshold ($FDR < 1e-5$) in the second filter step. Finally, ca. 10^6 significant ceRNA interactions are predicted by SPONGE and used to construct a pan-cancer ceRNA interaction network.

SPONGE estimates ceRNA interaction significance based on simulated null distributions. To determine if this estimation is accurate when applied to real data, we devised a random scenario in which SPONGE should not be able to find significant interactions. We devised a true-negative setting by using only miRNAs as features for a particular gene, which do not have a predicted miRNA binding site in the target gene in any of our considered databases, *i.e.* miRNAs that have no seed match in the gene (blue bars, Fig. 6.6). Here only 66 interactions remained significant. Thus, our assumed $FDR < 1e-5$ appears conservative, which demonstrates the efficacy of SPONGE in filtering for significant miRNA-mediated interactions between genes.

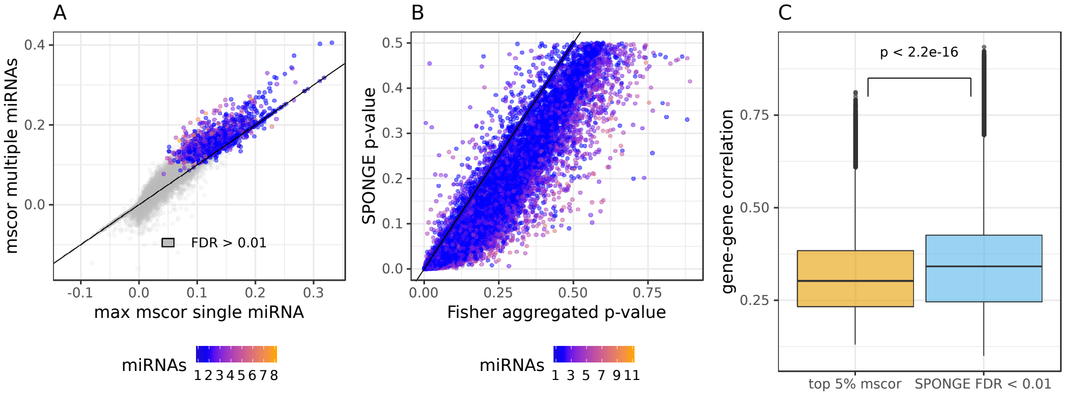


FIGURE 6.4: Comparison of sensitivity correlation and SPONGE FDR control on liver cancer data. (A) *mscor* values (y-axis) compared to maximal *scor* values (x-axis) for the same gene-gene interaction. (B) *mscor* p-values obtained from sampling compared to p-value summarization of *scor* values using Fisher's method. (C) Boxplot of gene-gene correlations for gene-miRNA-gene triplets obtained after selecting the top 5% ceRNA interactions according to the raw *scor* values (orange) or based on FDR corrected p-values from SPONGE (blue). t-test p-value between both distributions is shown on top.

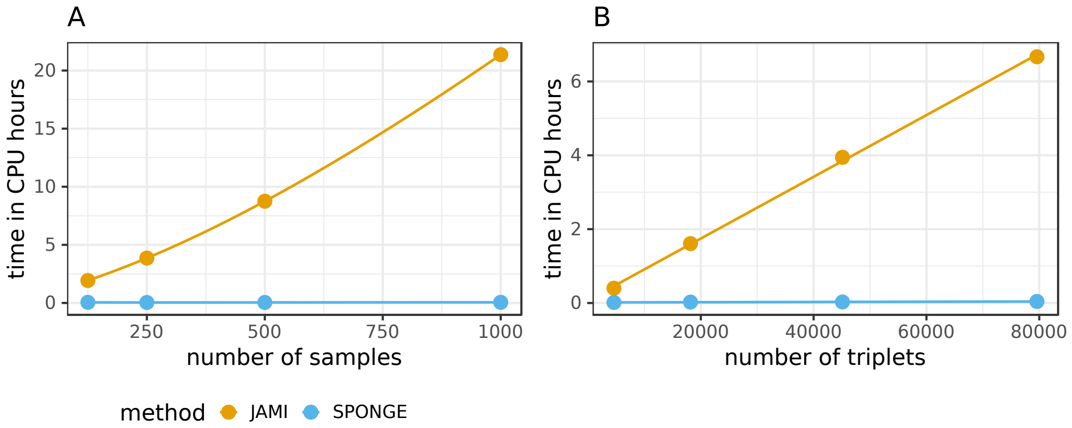


FIGURE 6.5: Runtime comparison between SPONGE and JAMI, a fast method for computing ceRNA interactions based on CMI. (A) Runtime for varying number of samples on a fixed set of ca. 80,000 triplets. (B) Runtime for varying number of triplets on a fixed number of samples. Time was measured in CPU hours (y-axis).

gene type	number of genes
protein coding	12776
pseudogenes	1529
lincRNA	1086
antisense	1025
processed transcript	207
sense intronic	69
sense overlapping	67

TABLE 6.1: Number of genes participating in significant ceRNA pan-cancer interactions (FDR < 1e-5) divided by Ensembl gene type.

Pan-cancer ceRNA network analysis

After demonstrating that most of the ceRNA interactions in the pan-cancer network are statistically sound, we proceeded with a more detailed analysis. After processing expression data from 60,498 genes and 2,463 mature miRNAs, SPONGE reported 95,541,095 gene-gene interactions after step one from which we retained 914,165 at an FDR threshold of 1e-5 (Fig. 6.6). 16,935 genes participated in ceRNA cross-talk with a median of 29 interactions per gene and a median of six miRNAs per ceRNA interaction with a maximum of 36 miRNAs per interaction. Table 6.1 shows the number of genes in different Ensembl gene categories, highlighting that ceRNA interaction is not limited to protein-coding genes with a 3′ UTR. Interestingly, we found a large number of pseudogenes in this pan-cancer analysis, including the two previously reported pseudogenes PTENP1 and BRAFP1 (Sanchez-Mejias and Tay, 2015).

We further investigated which microRNAs facilitate ceRNA cross-talk by counting how many interactions they participate in. These results are shown in Supp. Fig. 2. Our results highlight that a few miRNAs mediate most of the ceRNA interactions in the network. We observe that these miRNAs have comparably high expression levels, which is in line with what we would expect since ceRNA competition only plays a role if sufficient miRNA copies are present in a cell.

The ceRNA network is based on the pan-cancer TCGA dataset which contains cancer as well as tumor-adjacent samples. To identify which of the key ceRNA regulators are associated with cancer, we filtered for genes which showed high mean

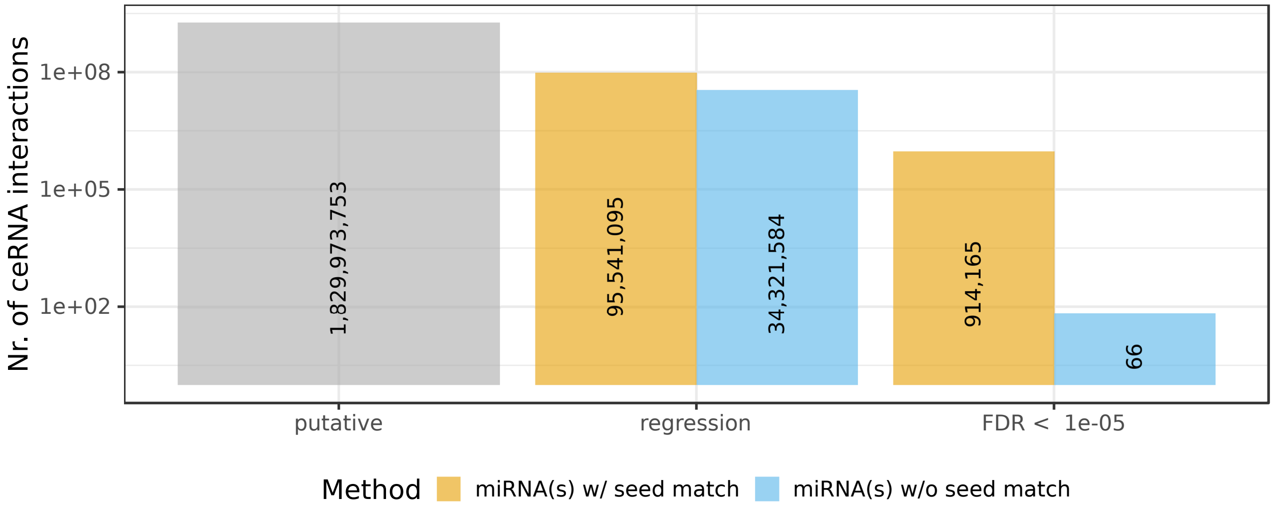


FIGURE 6.6: Analysis of SPONGE ceRNA interactions on the pan-cancer dataset. Barplots show the number of interactions (y-axis) that are initially analyzed (grey), obtained after the regression filter (Step 1) and after computing *mscor* values and FDR correction of empirical p-values (Step 3). The analysis is shown for miRNA-gene relationships for which miRNA binding sites (seeds) have been predicted (orange bars) and for a large set of true-negative miRNA-gene relationships, investigating miRNAs without seed matches in a given gene (blue bars).

expression levels (TPM > 100) and were differentially expressed between cancer and tumor-adjacent samples (t-test, FDR < 0.01 (Benjamini and Hochberg, 1995b) and log_2 fold change > 1). Our rationale was to determine genes that are present at sufficient copy numbers to exert cell-relevant ceRNA effects concentrated on genes that are overexpressed in the pan-cancer samples, thus likely mediating oncogenic effects. A total of 141 unique genes were obtained using these criteria (Supp. Table 1). The 10 genes with the highest number of interactions are shown in Table 6.2 and in Fig. 6.7.

The gene with the largest number of significant ceRNA interactions is *VCAN*, which is an established ceRNA (Tay, Rinn, and Pandolfi, 2014; Sanchez-Mejias and Tay, 2015). In fact, previous work has shown that overexpression of the *VCAN* 3'UTR sequence alone is able to induce cancer growth in liver cancer cells (Fang et al., 2013). Similarly *FN1* is a known ceRNA (Sanchez-Mejias and Tay, 2015).

We used clinical data from TCGA to assess if the expression of the genes identified here is significantly associated with survival probability. Fig. 6.7B shows that among the top 10 genes, 8 are significant ($p < 0.05$, see Methods). Among all 141 genes in this analysis (Supp. Table 1) we find a significant enrichment for survival related genes according to a hypergeometric test (p = 3.75e-10) comparing against the background of other genes.

An intriguing candidate in this list is the linc-RNA *LINC00511*, which has the highest expression of all non-coding genes in this set and is associated with survival (Fig. 6.7C). Interestingly, a recent paper has shown that *LINC00511* is an oncogenic ceRNA and regulates VEGFA gene expression in pancreatic adenocarcinoma (Zhao et al., 2018). Further, it was found that *LINC00511* is a ceRNA for E2F1 and is involved in breast cancer tumourigenesis (Lu et al., 2018). Also, it was found that *LINC00511* drives tumourigenesis in non-small-cell lung cancer (Sun et al., 2016). This suggests that *LINC00511* is an oncogenic ceRNA that plays an important role

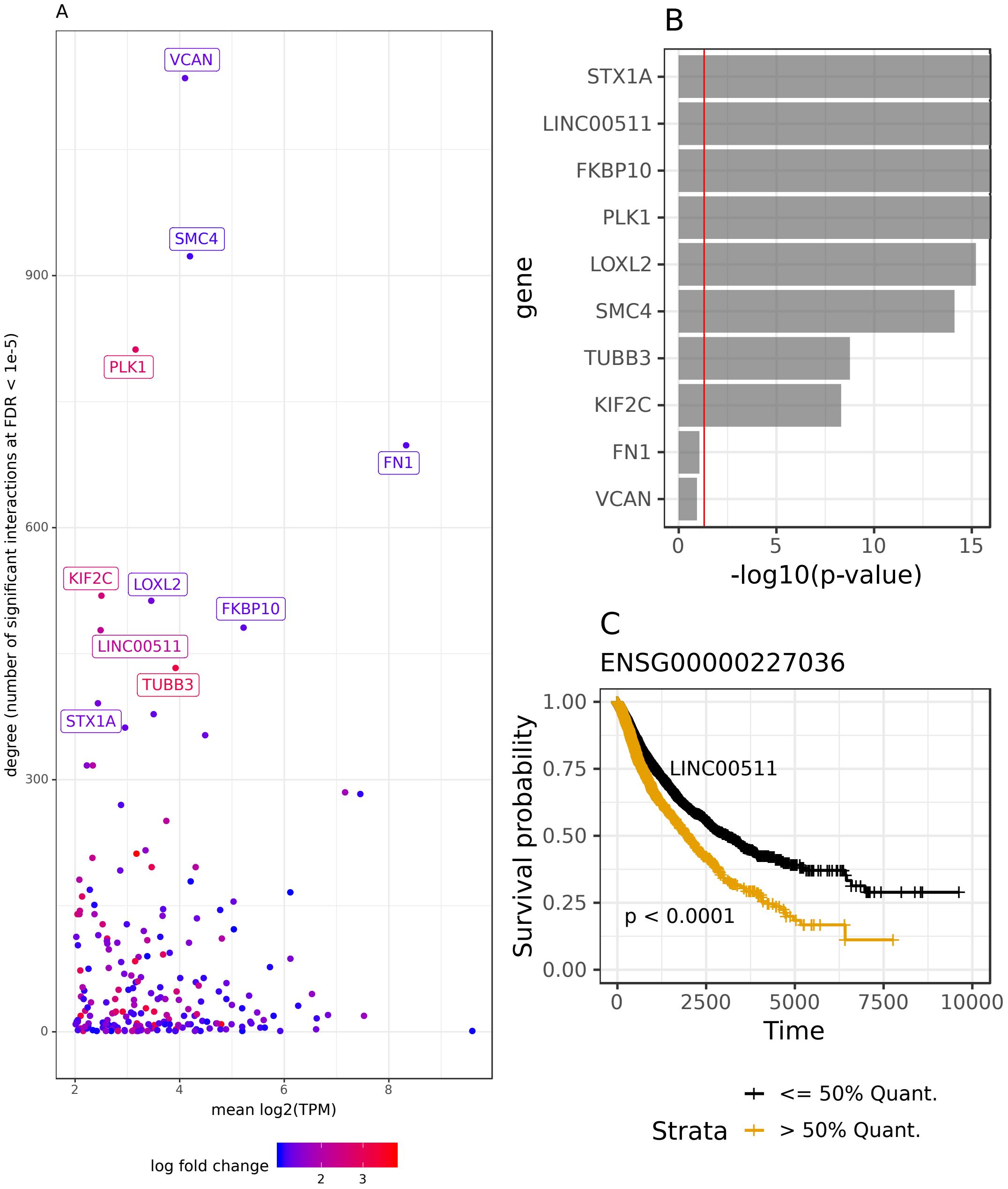


FIGURE 6.7: (A) Degree of ceRNA genes with mean expression (TPM > 100) and differential expression between cancer and tumor-adjacent samples (FDR < 0.01 and log fold change > 1). Number of ceRNA interactions (y-axis) is compared to mean expression (x-axis). Differential expression magnitude is shown as color code in the plot. (B) The 10 genes with highest degree ranked by their survival analysis *p*-value. (C) Kaplan Meier survival plot of the non-coding RNA LINC00511.

in diverse cancer types, as experimental evidence for *LINC00511* mediated ceRNA regulation in two cancers already exists. Thus, *LINC00511* qualifies as an interesting pan-cancer drug target.

	ensembl gene id	HGNC gene symbol	degree
1	ENSG00000038427	VCAN	1135
2	ENSG00000113810	SMC4	923
3	ENSG00000166851	PLK1	812
4	ENSG00000115414	FN1	698
5	ENSG00000142945	KIF2C	519
6	ENSG00000134013	LOXL2	513
7	ENSG00000141756	FKBP10	481
8	ENSG00000227036	LINC00511	478
9	ENSG00000258947	TUBB3	433
10	ENSG00000106089	STX1A	391

TABLE 6.2: Top 10 ceRNA regulating genes with highest node degree among genes differentially expressed between cancer and tumor-adjacent samples. The full table with 141 differentially expressed genes is shown in Supp. Table 1.

6.6 Discussion

We identified two major obstacles that prevent the efficient inference of a comprehensive genome-wide ceRNA interaction network. One of the first approaches, CUPID (Sumazin et al., 2011; Chiu et al., 2015; Hornakova et al., 2018), does not scale to the genome-wide level (see Fig.6.5) due to the use of permutation-based empirical p-value computation for establishing significance and the complexity of estimating CMI. Partial correlation-based approaches employing *scor* (Paci, Colombo, and Farina, 2014), on the other hand, are fast but do not accurately determine significance of the estimated effects.

To overcome these two issues, we designed an efficient empirical p-value computation approach by sampling from null models that describe the random distribution of *mscor* values. Moreover, this approach enabled us to accommodate possible biases introduced by several parameters, namely the number of samples, the gene-gene correlation and the number of shared miRNAs. Our results highlight that the current practice of ranking ceRNA interactions by *scor* coefficients introduces a bias towards gene pairs with low correlation, which are also more abundant. Furthermore, it became evident that *scor* can not be directly compared across datasets with different sample numbers, suggesting that previous studies on unbalanced data sets, where ceRNA network comparisons between cancer and related normal samples were conducted, have likely been biased. We note that null model-based significance analysis is fast, which entails that SPONGE can compute p-values at high numerical precision ($p >$ 1e6) compared to permutation-based approaches that often limit precision ($p >$ 1e3) due to excessive runtime.

In this work we have presented a statistical approach to jointly estimate the significance of multiple miRNAs in a ceRNA interaction between two genes. Most genes are regulated by several miRNAs and it can thus be expected that potential ceRNA interaction partners share more than one miRNA between them. This suggests that there is an advantage in considering joint effects of several miRNAs. As far as we are aware only CUPID considers such combinatorial effects when using paired expression data. However, CUPID integrates these effects at the level of triplets, where p-values of triplets involving the same genes are pooled (Chiu et al., 2015), which, as we have shown, results in a loss of sensitivity (Fig. 6.4B). In contrast, our approach captures the contribution of several miRNAs and their co-expression in a single mathematical model. To this end, we extended the concept of sensitivity correlation to multiple miRNA sensitivity correlation (*mscor*).

To make this approach broadly available, we developed SPONGE, a R/Bioconductor package which provides a general framework for analyzing sensitivity correlation beyond its current application in ceRNA network inference. SPONGE enabled us to construct the first pan-cancer ceRNA network that systematically infers interactions between all genes within a few days on a typical compute cluster. Notably, close to 16,000 genes are involved in ceRNA regulation. Roughly 12,000 of these are protein-coding genes highlighting that this is a genome-wide phenomenon as proposed by (Salmena et al., 2011) and not limited to non-coding RNAs. However, association may not be confused with causation. We can not rule out that some of the effects we observe are caused by the activity of the proteins encoded by the tested ceRNA genes. For instance, transcription factors or RNA binding proteins may affect the expression of ceRNA interaction partners directly or indirectly.

To further investigate to what extend our results are biased by non-miRNA-mediated regulatory effects, we conducted an *in silico* control experiment where we observed that almost no significant ceRNA interactions remained when miRNAs were tested for which an actual regulation is unlikely as they have no seed match in either of the genes. This suggests that the majority of SPONGE reported ceRNA interactions can be attributed to miRNA-based association.

Network analysis in which we focused on genes that show moderate to high average expression and that are differentially expressed between cancer and tumor-adjacent samples revealed ceRNA genes with hundreds of interactions, many of which also show a significant association with survival probability. Our findings suggest that many protein-coding genes auch as VCAN and FN1 have an additional regulatory function as a ceRNA. Moreover, SPONGE suggests ceRNA regulation as a potential mechanism to explain why non-coding RNAs such as LINC00511 have a significant impact on survival. This straight-forward analysis thus illustrates the potential of ceRNA networks for hypothesis generation and biomarker discovery.

We note that results might vary depending on the choice and quality of miRNA target interaction databases. To alleviate this issue, we selected datasets based on sequence-based predictions as well as experimentally validated miRNA target interactions. Most of the sequence-based prediction methods focus exclusively on the 3′ UTR of protein-coding genes for detecting miRNA binding sites. Our results indicate that non-coding RNAs make a substantial contribution to miRNA cross-talk such that future miRNA-target annotations should be adapted.

It is important to emphasize that statistical significance does not equal biological relevance. While we have ensured that the pan-cancer ceRNA interactions predicted in this work are likely true associations with respect to our model and its assumptions, understanding which of those individual interactions are of physiological relevance, is another important problem. Large-scale validation of ceRNA interactions is challenging and new methods are needed. One interesting approach is the work by Rzepiela et al., in which miRNA target sensitivity values were estimated using mathematical modelling of miRNA overexpression coupled to single cell expression analyses and may provide a way to prioritize ceRNA targets of functional biological relevance (Rzepiela et al., 2018).

6.7 Conclusion and Outlook

The TCGA pan-cancer analysis performed here provides unique insights into global ceRNA cross-talk in cancer. However, cancer-specific networks will be needed to draw a more comprehensive map of ceRNA regulation where sophisticated network alignment methods are employed to reveal commonalities and differences between cancer types. Generating paired gene and miRNA expression data for healthy tissues in databases like GTEx (Lonsdale et al., 2013) will become crucial for gaining an understanding of tissue-specific ceRNA cross-talk which will in turn present a baseline for detecting cancer-specific aberrations in the network presented here. Recently, single cell protocols that facilitate measurements of multi-omics have become available (Macaulay, Ponting, and Voet, 2017). We envision that a protocol supporting parallel measurement of microRNA and gene expression will particularly benefit from fast correlation-based approaches like SPONGE for celltype-specific ceRNA network inference.

Note that, to our knowledge, we have devised the first generalized algorithm for sampling covariance matrices in which the partial correlation is equal to the correlation. We envision that this might be relevant beyond the inference of ceRNA interaction networks with possible applications in other scientific disciplines.

Chapter 7

Conclusions and future directions

7.1 Future directions

In this section we talk about the extension of our work.

7.1.1 POSTIT

In the model we presented, we have the parameter γ Eq. (5.3), which controls the similarity of regression coefficients for a regulator between different isoforms. Another alternative would be to define a kernel function and learn the task similarity from the data directly. This might lead to a more accurate estimation of regression coefficients but would make the computational algorithm more involved. Moreover, one could use the Bayesian sparse group lasso, where we could incorporate our priors regarding regulators and epigenomics priors of the regulator binding site. This may be a more tractable formulation and elegant solution to learn the fine-grained complexity of gene regulatory mechanisms.

POSTIT reconstructs a transcript isoform regulatory network by integrating different high-throughput assays. Unlike a gene regulatory network, an isoform regulatory network faces a unique challenge: the lack of functional annotation for isoforms. Currently available resources such as the gene ontology or pathway annotations, provide only annotation on the level of genes. Hence, these resources lose the fine-grained functional information at the isoform level. There is a big need for resources that curate functional annotation on the isoform level of transcripts, acknowledging the complexity of human transcription and post-transcriptional regulation. If more of these resources are being developed, downstream analyses of prediction of POSTIT will become easier and will be more revealing.

7.1.2 SPONGE

In many genes, alternative splicing gives rise to a large number of transcripts, many of which differ strongly in their expression. Some of these transcripts are not translated and vary in the miRNA binding sites they carry. Thus, similar to transcripts originating from non-coding genes, they have no apparent biological role but may potentially contribute to ceRNA cross-talk. Considering transcript-level expression data will improve the quality of ceRNA network inference and allow for identifying disease-relevant changes in alternative splicing that act through ceRNA effects.

7.1.3 Modeling competition between miRNAs and RBPs

In Chapter 4 we showed that over expression of IMP2 alters the regulatory capacity of miRNAs by competing for binding sites. As a result, IMP2 induces a steatosis-like phenotype and enhances the risk of developing hepatocellular carcinoma (Deghghani Amirabad et al., 2018). Currently, there is no computational method that integrates transcriptomics and par-clip data to prioritize RBPs that control the regulatory capacity of miRNA in biological systems and diseases. Moreover it is poorly understood to what extent RBPs interfere with miRNA mediated regulation.

Game theoretic approach: In this approach, self-interested agents (RBPs and miRNAs) compete to bind to limited resources (transcripts). We can define the strategy profiles and pay-off functions of players from sequence motifs and binding affinities of the regulators. The goal is to solve the game and find the Nash equilibrium that represents the optimum allocation of regulatory elements between regulators.

Probabilistic graphical model: In this approach, we can integrate miRNAs, RBPs and isoform expression as observed nodes. Furthermore, competition is encoded as a binary latent variables. The goal is to compute the posterior probability distribution for a latent variable given noisy priors of RBP and microRNA motif overlaps on the transcripts

These kinds of computational models have the potential to prioritize RPBs that add another layer of complexity to post- transcriptional gene regulation and, hence, might contribute to disease mechanisms. Both of these models can be easily extended to simultaneously model both cooperative and competitive modes of regulation. Furthermore, these modeling approaches can applicable to modeling the competition of other types of regulators, such as TFs binding, to open chromatin regions.

Appendix A

Appendix A

A.1 Proximal operator of sparse group lasso for multi-task regression

Since Eq. 5.9 is the element-wise ℓ_1-norm, it can be decomposed either by columns or by rows. Similarly, $||W - U||_F^2$ can be decomposed into the sum of rows.

$$h(W) = \sum_{p=1}^{P} \lambda ||W_{p*}||_1 + \gamma \sum_{p=1}^{P} g_p ||G_p||_2 + \frac{1}{2} ||W_{p*} - B_{p*}||_F^2. \tag{A.1}$$

Hence, we can optimize P functions independently. Because we are dealing with vectors exclusively, $u = B_{p*}$ corresponds to the p^{th} row's of B and $f(w) = h(W_{p*})$, then each of P equations can written as:

$$f(w) = \lambda ||w||_1 + \gamma g_p ||G_p||_2 + \frac{1}{2} ||w - u||_F^2 \ . \tag{A.2}$$

In order to find the optimality condition for the Eq. A.2, we derive $\nabla f(w)$, which leads to the following:

$$0 \in \lambda \partial ||w||_1 + \gamma g_p \partial ||G_p||_2 + w - u \ . \tag{A.3}$$

Using dual-norm theory of vectors we can write the subdifferential of the norms as follows:

$$\partial ||w||_1 = \{D_1 \ s.t \ ||D||_\infty \leq 1, \langle D, w \rangle = ||w||_1\} \tag{A.4}$$

$$\partial ||G_p||_2 = \{D_2 \ s.t. \ ||D||_2 \leq 1, \langle D, w \rangle = ||w||_2\} \ . \tag{A.5}$$

Now Equation A.3 can be written as:

$$0 \in \lambda D_1 + \gamma g_p D_2 + w - u \tag{A.6}$$

$$w^* \in u - \lambda D_1 - \gamma \ g_p D_2 \ , \tag{A.7}$$

where w^* denotes the optimal solution. First we will simplify $u - \lambda D_1$ in the next Lemma.

Lemma A.1.1. *$u - \lambda D_1$ is equal to $soft(u; \lambda)$, known as soft-thresholding:*

$$soft(u; \lambda) \begin{cases} 0 & if \quad |u_i| \leq \lambda \\ u_i - \lambda sign(u_i) & if \quad |u_i| > \lambda \ , \end{cases}$$

where u_i denotes the i-th element of u.

Proof. We define $s(u) = \lambda||u||_1$, proximal of s is defined as follows:

$$prox_s(u) = argmin_z(\lambda||z||_1 + \frac{1}{2}||z-u||_2^2)\ . \tag{A.8}$$

We can then write the optimality condition as:

$$0 \in \nabla(||z-u||_2^2) + \partial(\lambda||z||_1) \tag{A.9}$$

$$0 \in z - u + \lambda\partial||z||_1 \Rightarrow z* = u - \lambda\partial||z||_1 \tag{A.10}$$

Thus $u - \lambda\partial||z||_1$ is the proximal operator of l_1 norm at point u. We can derive the solution for Equation A.10 by considering each component separately, due to separability of l_1 norm:

- **Case I:**

  $$z_i \neq 0 \Rightarrow \partial||z_i|| = sign(z_i),$$

  then

  $$0 = z_i - u_i + \lambda sign(z_i)$$

  $$z_i = u_i - \lambda sign(z_i)$$

  $$z_i \neq 0 \begin{cases} if \quad z_i < 0 & \Rightarrow u_i - \lambda sign(z_i) < 0 \Rightarrow u_i < -\lambda \\ if \quad z_i > 0 & \Rightarrow u_i - \lambda sign(z_i) > 0 \Rightarrow u_i > \lambda \end{cases}$$

  We can conclude that $|u_i| \geq \lambda$.

- **Case II:**

  $$z_i = 0 \Rightarrow \partial||z_i||_1 \in [-1, 1],$$

  then we can write A.10 as follow:

  $$0 \in 0 - u_i + \lambda[-1, 1]$$

  $$u_i \in [-\lambda, \lambda] \Rightarrow |u_i| < \lambda.$$

Putting case I and II together, we have:

$$[prox_s(u_i)] = z_i^* = \begin{cases} 0 \qquad if \quad |u_i| \leq \lambda \\ u_i - \lambda sign(z_i) \qquad if \quad |u_i| > \lambda \end{cases}$$

Above equation is called soft-thresholding. So we can conclude that:

$$u - \lambda\partial||z||_1 = u - \lambda D_1 = soft(u; \lambda).$$

□

Lemma 1 reduces optimality condition for A.7 to:

$$w^* \in soft(u; \lambda) - \gamma g_p D_p. \tag{A.11}$$

Lemma A.1.2. *Equation A.11 is equivalent to evaluating the proximal sum of ℓ_2-norm at point $soft(u;\lambda)$.*

Proof. let's define $A = soft(u,\lambda)$ and $k(A) = \gamma A$. Proximal operator of $k(A)$ is defined as follow:

$$prox_k(A) = argmin_w(k(w) + \frac{1}{2}||w - A||_2^2). \tag{A.12}$$

We can write optimality condition for equation A.12 as follow:

$$0 \in \gamma\partial||w||_2 + w - A \tag{A.13}$$

then

$$w^* \in A - \gamma D_2 = soft(u,\lambda) - \gamma g_p D_2. \tag{A.14}$$

□

Lemma 2 reduces optimality condition for A.7 to:

$$\gamma g_p D_2 + w \in soft(u;\lambda) \Rightarrow w^* \in soft(u;\lambda) - \gamma g_p D_2. \tag{A.15}$$

Since D_2 is a point in the inner product of dual space of ℓ_2-norm (which is again the ℓ_2-norm), we can apply *Moreau's decomposition theorem* to the ℓ_2-norm to solve Equation A.15.

Theorem A.1.3. *Moreau's decomposition theorem: Given a function $g(x)$, the following holds regarding proximal of g*

$$\forall\lambda > 0 \quad prox_{\lambda g}(X) = X - \lambda prox_{g*/\lambda}(\frac{x}{\lambda}), \tag{A.16}$$

where $g*$ is convex conjugate of g(x).

So, for any $g = ||.||$ and B is unit ball of dual norm, then

$$prox_{\lambda g}(x) = x - \lambda \prod_B (x/\lambda). \tag{A.17}$$

*When $g = ||.||_2$ and **B** is unit ℓ_2 ball, then:*

$$\prod_B(x) = \begin{cases} \frac{x}{||x||_2} & ||x||_2 > 1 \\ x & ||x||_2 \leq 1, \end{cases}$$

which helps us to rewrite the proximal operator of ℓ_2-norm in the following form:

$$prox_{\lambda g}(x) = \begin{cases} (1 - \frac{\lambda}{||x||_2})x & ||x||_2 \geq \lambda \\ 0 & ||x||_2 < \lambda \end{cases}$$

We already showed that the optimum value for $w*$ is given by

$$w* \in soft(u;\lambda) - \gamma g_p D_2. \tag{A.18}$$

Which is equivalent to proximal for sum of ℓ_2-norms at $soft(u;\lambda)$ point. Thus we conclude for group p is defined as:

$$Q_p = \begin{cases} (1 - \frac{\lambda}{||soft(u;\lambda) \circ I_i||_2})(soft(u;\lambda)) & ||soft(u;\lambda)||_2 \geq \gamma g_i \\ 0 & ||soft(u;\lambda)||_2 < \gamma g_i \end{cases}$$

Thus we conclude that:

$$argmin_w f(w) = Q_p. \tag{A.19}$$

This concludes the proof.

Bibliography

Agarwal, Vikram et al. (2015a). “Predicting effective microRNA target sites in mammalian mRNAs”. In: *elife* 4, e05005.

— (2015b). “Predicting effective microRNA target sites in mammalian mRNAs”. In: *eLife* 4. Ed. by Elisa Izaurralde, e05005. ISSN: 2050-084X. DOI: 10.7554/eLife.05005. URL: https://dx.doi.org/10.7554/eLife.05005.

Agarwal *et. al*, Vikram (2015). “Predicting effective microRNA target sites in mammalian mRNAs”. In: *eLife* 4. Ed. by Elisa Izaurralde. DOI: 10.7554/eLife.05005.

Agostini, Federico et al. (2013). “cat RAPID omics: a web server for large-scale prediction of protein–RNA interactions”. In: *Bioinformatics* 29.22, pp. 2928–2930.

Allis, C David et al. (2007). *Epigenetics*. CSHL Press.

Arrieta-Ortiz, Mario L et al. (2015). “An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network”. In: *Molecular systems biology* 11.11, p. 839.

Arvey, Aaron et al. (2010a). “Target mRNA abundance dilutes microRNA and siRNA activity”. In: *Molecular systems biology* 6.1, p. 363.

— (2010b). “Target mRNA abundance dilutes microRNA and siRNA activity”. In: *Molecular Systems Biology* 6, pp. 363–363. DOI: 10.1038/msb.2010.24. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2872614/.

Bach, Francis et al. (2012). “Optimization with sparsity-inducing penalties”. In: *Foundations and Trends® in Machine Learning* 4.1, pp. 1–106.

Baralle, Francisco E and Jimena Giudice (2017). “Alternative splicing as a regulator of development and tissue identity”. In: *Nature Reviews Molecular Cell Biology* 18.7, p. 437.

Bauer, Sebastian et al. (2008). “Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration”. In: *Bioinformatics* 24.14, pp. 1650–1651.

Bell, Jessica L et al. (2013). “Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression?” In: *Cellular and molecular life sciences* 70.15, pp. 2657–2675.

Benjamini, Yoav and Yosef Hochberg (1995a). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300.

— (1995b). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300. ISSN: 00359246. URL: http://www.jstor.org/stable/2346101.

Bennett, Kristin P and Emilio Parrado-Hernández (2006). “The interplay of optimization and machine learning research”. In: *Journal of Machine Learning Research* 7.Jul, pp. 1265–1281.

Bergmann, Sven, Jan Ihmels, and Naama Barkai (2003). “Similarities and differences in genome-wide expression data of six organisms”. In: *PLoS biology* 2.1, e9.

Bickel, Peter J, Ya'acov Ritov, Alexandre B Tsybakov, et al. (2009). "Simultaneous analysis of Lasso and Dantzig selector". In: *The Annals of Statistics* 37.4, pp. 1705–1732.

Bonacich, Phillip (1972). "Factoring and weighting approaches to status scores and clique identification". In: *Journal of mathematical sociology* 2.1, pp. 113–120.

Borwein, Jonathan and Adrian S Lewis (2010). *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media.

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Butte, Atul J and Isaac S Kohane (1999). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements". In: *Biocomputing 2000*. World Scientific, pp. 418–429.

Calamai, Paul H and Jorge J Moré (1987). "Projected gradient methods for linearly constrained problems". In: *Mathematical programming* 39.1, pp. 93–116.

Callegari, Elisa et al. (2015). "MicroRNAs in liver cancer: a model for investigating pathogenesis and novel therapeutic approaches". In: *Cell death and differentiation* 22.1, p. 46.

Capurro, Mariana et al. (2003). "Glypican-3: a novel serum and histochemical marker for hepatocellular carcinoma". In: *Gastroenterology* 125.1, pp. 89–97.

Ceccarelli, Sara et al. (2013). "Dual role of microRNAs in NAFLD". In: *International journal of molecular sciences* 14.4, pp. 8437–8455.

Chatterjee, Sumantra and Nadav Ahituv (2017). "Gene regulatory elements, major drivers of human disease". In: *Annual review of genomics and human genetics* 18, pp. 45–63.

Chen, Scott Shaobing, David L Donoho, and Michael A Saunders (2001). "Atomic decomposition by basis pursuit". In: *SIAM review* 43.1, pp. 129–159.

Chen, Xi et al. (2012). "Smoothing proximal gradient method for general structured sparse regression". In: *The Annals of Applied Statistics* 6.2, pp. 719–752.

Chiu, Hua Sheng et al. (2015). "Cupid: Simultaneous reconstruction of microRNA-target and ceRNA networks". In: *Genome Research* 25.2, pp. 257–267. ISSN: 15495469. DOI: 10.1101/gr.178194.114.

Chou, Chih-Hung et al. (2016). "miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database". In: *Nucleic Acids Research* 44.D1, pp. D239–D247.

Chou, Chih-Hung et al. (2017). "miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions". In: *Nucleic acids research* 46.D1, pp. D296–D302.

Christiansen, Jan, Astrid M Kolte, Finn C Nielsen, et al. (2009). "IGF2 mRNA-binding protein 2: biological function and putative role in type 2 diabetes". In: *Journal of molecular endocrinology* 43.5, pp. 187–195.

Consortium, ENCODE Project et al. (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, p. 57.

Conway, Anne E et al. (2016). "Enhanced CLIP uncovers IMP protein-RNA targets in human pluripotent stem cells important for cell adhesion and survival". In: *Cell reports* 15.3, pp. 666–679.

Corbett, Anita H (2018). "Post-transcriptional regulation of gene expression and human disease". In: *Current opinion in cell biology* 52, pp. 96–104.

Cui, Wei, Stephen L Chen, and Ke-Qin Hu (2010). "Quantification and mechanisms of oleic acid-induced steatosis in HepG2 cells". In: *American journal of translational research* 2.1, p. 95.

Dai, Ning et al. (2011). “mTOR phosphorylates IMP2 to promote IGF2 mRNA translation by internal ribosomal entry”. In: *Genes & development*.

Day, CP (2006). “Genes or environment to determine alcoholic liver disease and non-alcoholic fatty liver disease”. In: *Liver International* 26.9, pp. 1021–1028.

Degrauwe, Nils et al. (2016). “The RNA binding protein IMP2 preserves glioblastoma stem cells by preventing let-7 target gene silencing”. In: *Cell reports* 15.8, pp. 1634–1647.

Dehghani Amirabad, Azim and Marcel Holger Schulz (2016). “Multitask regression for condition-specific prioritization of miRNA targets in transcripts”. In: *PeerJ Preprints* 4.

Dehghani Amirabad, Azim et al. (2018). “Transgenic expression of the RNA binding protein IMP2 stabilizes miRNA targets in murine microsteatosis”. In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1864.10, pp. 3099–3108.

Deng *et. al*, Nan (2011). “Isoform-level microRNA-155 target prediction using RNA-seq”. In: *Nucleic Acids Research* 39.9, e61–e61. DOI: 10.1093/nar/gkr042. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3089486/.

Derrien, Thomas et al. (2012). “The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression”. In: *Genome research* 22.9, pp. 1775–1789.

Derti, Adnan et al. (2012). “A quantitative atlas of polyadenylation in five mammals”. In: *Genome research* 22.6, pp. 1173–1183.

Ding, Jie et al. (2010). “Gain of miR-151 on chromosome 8q24. 3 facilitates tumour cell migration and spreading through downregulating RhoGDIA”. In: *Nature cell biology* 12.4, p. 390.

Dong, Yang and Guang-Bin Qiu (2017). “Biological functions of miR-590 and its role in carcinogenesis”. In: *Frontiers in Laboratory Medicine* 1.4, pp. 173–176.

Du, Zhou et al. (2016). “Integrative analyses reveal a long noncoding RNA-mediated sponge regulatory network in prostate cancer”. In: *Nature communications* 7.

Elcheva, Irina et al. (2009). “CRD-BP protects the coding region of βTrCP1 mRNA from miR-183-mediated degradation”. In: *Molecular cell* 35.2, pp. 240–246.

Evgeniou, Theodoros, Charles A Micchelli, and Massimiliano Pontil (2005). “Learning multiple tasks with kernel methods”. In: *Journal of Machine Learning Research* 6.Apr, pp. 615–637.

Fang, Ling et al. (2013). “Versican 3′-untranslated region (3′-UTR) functions as a ceRNA in inducing the development of hepatocellular carcinoma by regulating miRNA activity”. In: *The FASEB Journal* 27.3. PMID: 23180826, pp. 907–919. DOI: 10.1096/fj.12-220905. eprint: https://doi.org/10.1096/fj.12-220905. URL: https://doi.org/10.1096/fj.12-220905.

Farazi, Thalia A, Stefan A Juranek, and Thomas Tuschl (2008). “The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members”. In: *Development* 135.7, pp. 1201–1214.

Fischer, Ronald.A (1925). *Statistical Methods for Research Workers. The Science of Microfabrication*. Oliver and Boyd.

Fisher, R. A. (1915). “Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population”. In: *Biometrika* 10.4, p. 507. ISSN: 00063444. DOI: 10.2307/2331838. URL: http://www.jstor.org/stable/2331838?origin=crossref.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010a). “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software, Articles* 33.1, pp. 1–22. ISSN: 1548-7660. DOI: 10.18637/jss.v033.i01. URL: https://www.jstatsoft.org/v033/i01.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2008). "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3, pp. 432–441.

— (2010b). "A note on the group lasso and a sparse group lasso". In: *arXiv preprint arXiv:1001.0736*.

Friedman, Nir et al. (2000). "Using Bayesian networks to analyze expression data". In: *Journal of computational biology* 7.3-4, pp. 601–620.

Gabory, Anne et al. (2009). "H19 acts as a trans regulator of the imprinted gene network controlling growth in mice". In: *Development* 136.20, pp. 3413–3421.

Gaidatzis, Dimos et al. (2007). "Inference of miRNA targets using evolutionary conservation and pathway analysis". In: *BMC bioinformatics* 8.1, p. 69.

Garber, Manuel et al. (2011). "Computational methods for transcriptome annotation and quantification using RNA-seq". In: *Nature methods* 8.6, p. 469.

Geman, Stuart, Elie Bienenstock, and René Doursat (1992). "Neural networks and the bias/variance dilemma". In: *Neural computation* 4.1, pp. 1–58.

Ghanbari, Mahsa, Julia Lasserre, and Martin Vingron (2015). "Reconstruction of gene networks using prior knowledge". In: *BMC systems biology* 9.1, p. 84.

Goldman, Mary et al. (2018). "The UCSC Xena Platform for cancer genomics data visualization and interpretation". In: *bioRxiv*. DOI: 10.1101/326470. eprint: https://www.biorxiv.org/content/early/2018/08/28/326470.full.pdf. URL: https://www.biorxiv.org/content/early/2018/08/28/326470.

Golub, Gene H, Michael Heath, and Grace Wahba (1979). "Generalized cross-validation as a method for choosing a good ridge parameter". In: *Technometrics* 21.2, pp. 215–223.

Götz, Ulrike et al. (2016). "Two sets of RNAi components are required for heterochromatin formation in trans triggered by truncated transgenes". In: *Nucleic acids research* 44.12, pp. 5908–5923.

Griebel, Thasso et al. (2012). "Modelling and simulating generic RNA-Seq experiments with the flux simulator". In: *Nucleic Acids Research* 40.20, pp. 10073–10083. DOI: 10.1093/nar/gks666. eprint: http://nar.oxfordjournals.org/content/40/20/10073.full.pdf+html. URL: http://nar.oxfordjournals.org/content/40/20/10073.abstract.

Grimson, Andrew et al. (2007). "MicroRNA targeting specificity in mammals: determinants beyond seed pairing". In: *Molecular cell* 27.1, pp. 91–105.

Guttman, Mitchell and John L Rinn (2012). "Modular regulatory principles of large non-coding RNAs". In: *Nature* 482.7385, p. 339.

Hafner, Markus et al. (2010). "Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP". In: *Cell* 141.1, pp. 129–141.

Hahn, Matthew W and Andrew D Kern (2004). "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks". In: *Molecular biology and evolution* 22.4, pp. 803–806.

Hansen, Thomas VO et al. (2004). "Dwarfism and impaired gut development in insulin-like growth factor II mRNA-binding protein 1-deficient mice". In: *Molecular and cellular biology* 24.10, pp. 4448–4464.

Harris, Miera B, Justin Mostecki, and Paul B Rothman (2005). "Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function". In: *Journal of Biological Chemistry* 280.13, pp. 13114–13121.

Haury, Anne-Claire et al. (2012). "TIGRESS: trustful inference of gene regulation using stability selection". In: *BMC systems biology* 6.1, p. 145.

Hausser, Jean and Mihaela Zavolan (2014). "Identification and consequences of miRNA–target interactions—beyond repression of gene expression". In: *Nature Reviews Genetics* 15.9, p. 599.

Hentze, Matthias W et al. (2018). “A brave new world of RNA-binding proteins”. In: *Nature Reviews Molecular Cell Biology*.

Holt, Robert A and Steven JM Jones (2008). “The new paradigm of flow cell sequencing”. In: *Genome research* 18.6, pp. 839–846.

Hoque, Mainul et al. (2012). “Analysis of alternative cleavage and polyadenylation by 3 region extraction and deep sequencing”. In: *Nature methods* 10.2, p. 133.

Hornakova, Andrea et al. (2018). “JAMI-Fast computation of Conditional Mutual Information for ceRNA network analysis”. In: *Bioinformatics* 34.17, pp. 3050–3051.

Hornstein, Eran and Noam Shomron (2006). “Canalization of development by microRNAs”. In: *Nature genetics* 38, S20.

Hua, Hong-Wei et al. (2015). “MicroRNA-153 promotes Wnt/β-catenin activation in hepatocellular carcinoma through suppression of WWOX”. In: *Oncotarget* 6.6, p. 3840.

Huang, Junzhou, Tong Zhang, et al. (2010). “The benefit of group sparsity”. In: *The Annals of Statistics* 38.4, pp. 1978–2004.

Irrthum, Alexandre, Louis Wehenkel, Pierre Geurts, et al. (2010). “Inferring regulatory networks from expression data using tree-based methods”. In: *PloS one* 5.9, e12776.

Jacob, Laurent, Guillaume Obozinski, and Jean-Philippe Vert (2009). “Group lasso with overlap and graph lasso”. In: *Proceedings of the 26th annual international conference on machine learning*. ACM, pp. 433–440.

Jan, Calvin H et al. (2011). “Formation, regulation and evolution of Caenorhabditis elegans 3′ UTRs”. In: *Nature* 469.7328, p. 97.

Janiszewska, Michalina et al. (2012). “Imp2 controls oxidative phosphorylation and is crucial for preserving glioblastoma cancer stem cells”. In: *Genes & development*.

Jeggari, Ashwini, Debora S. Marks, and Erik Larsson (2012). “miRcode: a map of putative microRNA target sites in the long non-coding transcriptome”. In: *Bioinformatics* 28.15, pp. 2062–2063. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts344. URL: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts344.

Jenatton, Rodolphe, Guillaume Obozinski, and Francis Bach (2010). “Structured sparse principal component analysis”. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 366–373.

Jeong, Hawoong et al. (2001). “Lethality and centrality in protein networks”. In: *Nature* 411.6833, p. 41.

Jeyapalan, Zina et al. (2011). “Expression of CD44 3′-untranslated region regulates endogenous microRNA functions in tumorigenesis and angiogenesis”. In: *Nucleic Acids Research* 39.8, pp. 3026–3041. ISSN: 03051048. DOI: 10.1093/nar/gkq1003.

Johnson, David S et al. (2007). “Genome-wide mapping of in vivo protein-DNA interactions”. In: *Science* 316.5830, pp. 1497–1502.

Karreth, Florian A. et al. (2015). “The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo”. In: *Cell* 161.2, pp. 319–332. ISSN: 10974172. DOI: 10.1016/j.cell.2015.02.043.

Keene, Jack D (2007). “RNA regulons: coordination of post-transcriptional events”. In: *Nature Reviews Genetics* 8.7, p. 533.

Keller, Mark P et al. (2008). “A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility”. In: *Genome research* 18.5, pp. 706–716.

Kessler, SM et al. (2015). “IMP2/p62 induces genomic instability and an aggressive hepatocellular carcinoma phenotype”. In: *Cell death & disease* 6.10, e1894.

Kessler, Sonja M et al. (2013). “IGF2 mRNA binding protein p62/IMP2-2 in hepatocellular carcinoma: antiapoptotic action is independent of IGF2/PI3K signaling”. In: *American Journal of Physiology-Gastrointestinal and Liver Physiology* 304.4, G328–G336.

Kim, Donghwan and Jeffrey A Fessler (2016). “Optimized first-order methods for smooth convex minimization”. In: *Mathematical programming* 159.1-2, pp. 81–107.

Kim, Doyeon, Jongkyu Kim, and Daehyun Baek (2014). “Global and local competition between exogenously introduced microRNAs and endogenously expressed microRNAs”. In: *Molecules and cells* 37.5, p. 412.

Kim, Kyoung-Sook and Young-Ik Lee (1997). “Biallelic expression of the H19 and IGF2 genes in hepatocellular carcinoma”. In: *Cancer letters* 119.2, pp. 143–148.

Kim, Seongho (2015). “ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients”. In: *Communications for Statistical Applications and Methods* 22.6, pp. 665–674. ISSN: 2383-4757. DOI: 10.5351/CSAM.2015.22.6.665. URL: http://www.csam.or.kr/journal/view.html?doi=10.5351/CSAM.2015.22.6.665.

Koch, Christopher et al. (2017). “Inference and evolutionary analysis of genome-scale regulatory networks in large phylogenies”. In: *Cell systems* 4.5, pp. 543–558.

Krebs, Jocelyn E, Elliott S Goldstein, and Stephen T Kilpatrick (2017). *Lewin's Genes XII*. Jones & Bartlett Learning.

Kung, Johnny TY, David Colognori, and Jeannie T Lee (2013). “Long noncoding RNAs: past, present, and future”. In: *Genetics* 193.3, pp. 651–669.

Laggai, Stephan et al. (2013). “Rapid chromatographic method to decipher distinct alterations in lipid classes in NAFLD/NASH”. In: *World journal of hepatology* 5.10, p. 558.

Laggai, Stephan et al. (2014). “The IGF2 mRNA binding protein p62/IGF2BP2-2 induces fatty acid elongation as a critical feature of steatosis”. In: *Journal of lipid research*, jlr–M045500.

Lambert, Samuel A et al. (2018). “The human transcription factors”. In: *Cell* 172.4, pp. 650–665.

Lanzuolo, Chiara et al. (2007). “Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex”. In: *Nature cell biology* 9.10, p. 1167.

Le, Thuc Duy et al. (2016). “Computational methods for identifying miRNA sponge interactions”. In: *Briefings in Bioinformatics* 18.4 (4), bbw042. ISSN: 1467-5463. DOI: 10.1093/bib/bbw042. URL: http://bib.oxfordjournals.org/lookup/doi/10.1093/bib/bbw042.

Lebre, Sophie et al. (2010). “Statistical inference of the time-varying structure of gene-regulation networks”. In: *BMC systems biology* 4.1, p. 130.

Lee, Tong Ihn and Richard A Young (2013). “Transcriptional regulation and its misregulation in disease”. In: *Cell* 152.6, pp. 1237–1251.

Lewis, Benjamin P, Christopher B Burge, and David P Bartel (2005). “Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets”. In: *cell* 120.1, pp. 15–20.

Li, Bo and Colin N Dewey (2011). “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”. In: *BMC bioinformatics* 12.1, p. 323.

Li, Hong-Dong, Gilbert S Omenn, and Yuanfang Guan (2015). “MIsoMine: a genome-scale high-resolution data portal of expression, function and networks at the splice isoform level in the mouse”. In: *Database* 2015.

Li, Jun-Hao et al. (2013a). "starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein?RNA interaction networks from large-scale CLIP-Seq data". In: *Nucleic Acids Research* 42.D1, pp. D92–D97. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1248. eprint: http://oup.prod.sis.lan/nar/article-pdf/42/D1/D92/25891302/gkt1248.pdf. URL: https://dx.doi.org/10.1093/nar/gkt1248.

Li, Lin et al. (2004). "Gene regulation by Sp1 and Sp3". In: *Biochemistry and Cell Biology* 82.4, pp. 460–471.

Li, Wenyuan et al. (2013b). "High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method". In: *Nucleic acids research* 42.6, e39–e39.

Li, Xuri et al. (1997). "Disrupted IGF2 promoter control by silencing of promoter P1 in human hepatocellular carcinoma". In: *Cancer research* 57.10, pp. 2048–2054.

Liu, Chenglin et al. (2017). "Cancer-Related Triplets of mRNA-lncRNA-miRNA Revealed by Integrative Network in Uterine Corpus Endometrial Carcinoma". In: *BioMed research international* 2017.

Liu, Mingzhu et al. (2013). "The IGF2 intronic miR-483 selectively enhances transcription from IGF2 fetal promoters and enhances tumorigenesis". In: *Genes & development* 27.23, pp. 2543–2548.

Liu, Y et al. (2016). "Increased TEAD4 expression and nuclear localization in colorectal cancer promote epithelial–mesenchymal transition and metastasis in a YAP-independent manner". In: *Oncogene* 35.21, p. 2789.

Lonsdale, John et al. (2013). "The Genotype-Tissue Expression (GTEx) project". In: *Nature Genetics* 45, 580 EP –. URL: https://doi.org/10.1038/ng.2653.

Lounici, Karim et al. (2009). "Taking advantage of sparsity in multi-task learning". In: *arXiv preprint arXiv:0903.1468*.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12, p. 550.

Lu, Guanming et al. (2018). "Long noncoding RNA LINC00511 contributes to breast cancer tumourigenesis and stemness by inducing the miR-185-3p/E2F1/Nanog axis". In: *Journal of experimental & clinical cancer research* 37.1, pp. 289; 289–289. DOI: 10.1186/s13046-018-0945-6.

Luedde, Tom (2010). "MicroRNA-151 and its hosting gene FAK (focal adhesion kinase) regulate tumor cell migration and spreading of hepatocellular carcinoma". In: *Hepatology* 52.3, pp. 1162–1164.

Lukong, Kiven E et al. (2008). "RNA-binding proteins in human genetic disease". In: *Trends in Genetics* 24.8, pp. 416–425.

Lunde, Bradley M, Claire Moore, and Gabriele Varani (2007). "RNA-binding proteins: modular design for efficient function". In: *Nature reviews Molecular cell biology* 8.6, p. 479.

Macaulay, Iain C, Chris P Ponting, and Thierry Voet (2017). "Single-cell multiomics: multiple measurements from single cells". In: *Trends in Genetics* 33.2, pp. 155–168.

Marbach, Daniel et al. (2012). "Wisdom of crowds for robust gene network inference". In: *Nature methods* 9.8, p. 796.

Marioni, John C et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays". In: *Genome research*.

Marshall, Aileen et al. (2013). "Global gene expression profiling reveals SPINK1 as a potential hepatocellular carcinoma marker". In: *PloS one* 8.3, e59459.

Mayr, Christine and David P Bartel (2009). "Widespread shortening of 3 UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells". In: *Cell* 138.4, pp. 673–684.

Meijer, HA et al. (2013). "Translational repression and eIF4A2 activity are critical for microRNA-mediated gene regulation". In: *Science* 340.6128, pp. 82–85.

Merritt, Christopher et al. (2008). "3′ UTRs are the primary regulators of gene expression in the C. elegans germline". In: *Current Biology* 18.19, pp. 1476–1482.

Meyer, Patrick E et al. (2007). "Information-theoretic inference of large transcriptional regulatory networks". In: *EURASIP journal on bioinformatics and systems biology* 2007, pp. 8–8.

Monnier, Paul et al. (2013). "H19 lncRNA controls gene expression of the Imprinted Gene Network by recruiting MBD1". In: *Proceedings of the National Academy of Sciences* 110.51, pp. 20693–20698.

Morris, Kevin V and John S Mattick (2014). "The rise of regulatory RNA". In: *Nature Reviews Genetics* 15.6, p. 423.

Muniategui, Ander et al. (2012). "Quantification of miRNA-mRNA Interactions". In: *PLOS ONE* 7.2, pp. 1–10. DOI: 10.1371/journal.pone.0030766. URL: https://doi.org/10.1371/journal.pone.0030766.

Nesterov, Yurii (1983). "A method for unconstrained convex minimization problem with the rate of convergence O (1/kˆ 2)". In: *Doklady AN USSR*. Vol. 269, pp. 543–547.

Nielsen, Anne F, Jiradet Gloggnitzer, and Javier Martinez (2009). "MicroRNAs cross the line: the battle for mRNA stability enters the coding sequence". In: *Molecular cell* 35.2, pp. 139–140.

Nielsen, Jacob et al. (1999). "A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development". In: *Molecular and cellular biology* 19.2, pp. 1262–1270.

Noubissi, Felicite K et al. (2006). "CRD-BP mediates stabilization of βTrCP1 and c-myc mRNA in response to β-catenin signalling". In: *Nature* 441.7095, p. 898.

Obozinski, Guillaume, Ben Taskar, and Michael I Jordan (2010). "Joint covariate selection and joint subspace selection for multiple classification problems". In: *Statistics and Computing* 20.2, pp. 231–252.

Ohrnberger, Stefan et al. (2015). "Dysregulated serum response factor triggers formation of hepatocellular carcinoma". In: *Hepatology* 61.3, pp. 979–989.

Ong, Chin-Tong and Victor G Corces (2011). "Enhancer function: new insights into the regulation of tissue-specific gene expression". In: *Nature Reviews Genetics* 12.4, p. 283.

— (2014). "CTCF: an architectural protein bridging genome topology and function". In: *Nature reviews Genetics* 15.4, p. 234.

Paci, Paola, Teresa Colombo, and Lorenzo Farina (2014). "Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer". In: *BMC Systems Biology* 8.1, p. 83. ISSN: 1752-0509. DOI: 10.1186/1752-0509-8-83. URL: http://www.biomedcentral.com/1752-0509/8/83http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-8-83.

Paraskevopoulou, Maria D et al. (2013). "DIANA-microT web server v5. 0: service integration into miRNA functional analysis workflows". In: *Nucleic acids research* 41.W1, W169–W173.

Paraskevopoulou, Maria D et al. (2016). "DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts". In: *Nucleic Acids Research* 44.D1, pp. D231–D238.

Parikh, Neal, Stephen Boyd, et al. (2014). "Proximal algorithms". In: *Foundations and Trends® in Optimization* 1.3, pp. 127–239.

Patro, Rob, Stephen M Mount, and Carl Kingsford (2014). “Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms”. In: *Nature biotechnology* 32.5, p. 462.

Pennisi, Elizabeth (2012). *ENCODE project writes eulogy for junk DNA*.

Phillips, Jennifer E and Victor G Corces (2009). “CTCF: master weaver of the genome”. In: *Cell* 137.7, pp. 1194–1211.

Pinzón, Natalia et al. (2017). “microRNA target prediction programs predict many false positives”. In: *Genome Research* 27.2, pp. 234–245.

Place, Robert F et al. (2008). “MicroRNA-373 induces expression of genes with complementary promoter sequences”. In: *Proceedings of the National Academy of Sciences* 105.5, pp. 1608–1613.

Poliseno, Laura et al. (2010). “A coding-independent function of gene and pseudogene mRNAs regulates tumour biology”. In: *Nature* 465.7301, pp. 1033–1038. ISSN: 0028-0836. DOI: 10.1038/nature09144. URL: http://dx.doi.org/10.1038/nature09144http://www.nature.com/doifinder/10.1038/nature09144.

Pope, Scott D and Ruslan Medzhitov (2018). “Emerging principles of gene expression programs and their regulation”. In: *Molecular cell* 71.3, pp. 389–397.

Powers, John T et al. (2016). “Multiple mechanisms disrupt the let-7 microRNA family in neuroblastoma”. In: *Nature* 535.7611, pp. 246–251. ISSN: 0028-0836. DOI: 10.1038/nature18632. URL: http://dx.doi.org/10.1038/nature18632http://dx.doi.org/10.1038/nature18632{\%}5Cnhttp://10.1038/nature18632{\%}5Cnhttp://www.nature.com/nature/journal/v535/n7611/abs/nature18632.html{\#}supplementary-information.

Puri, Puneet et al. (2007). “A lipidomic analysis of nonalcoholic fatty liver disease”. In: *Hepatology* 46.4, pp. 1081–1090.

Puri, Puneet et al. (2009). “The plasma lipidomic signature of nonalcoholic steatohepatitis”. In: *Hepatology* 50.6, pp. 1827–1838.

Qi, Yanjun et al. (2010). “Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins”. In: *Bioinformatics* 26.18, pp. i645–i652.

Qian, Ning (1999). “On the momentum term in gradient descent learning algorithms”. In: *Neural networks* 12.1, pp. 145–151.

Quinlan, Aaron R and Ira M Hall (2010). “BEDTools: a flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6, pp. 841–842.

Re, Angela et al. (2009). “Genome-wide survey of microRNA–transcription factor feed-forward regulatory circuits in human”. In: *Molecular BioSystems* 5.8, pp. 854–867.

Reik, WOLF et al. (2000). “Igf2 imprinting in development and disease”. In: *Chromosomes today*. Springer, pp. 93–104.

Ren, Liping et al. (2015). “The Cdc15 and Imp2 SH3 domains cooperatively scaffold a network of proteins that redundantly ensure efficient cell division in fission yeast”. In: *Molecular biology of the cell* 26.2, pp. 256–269.

Richard, Hugues et al. (2010). “Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments”. In: *Nucleic Acids Research* 38.10, e112–e112.

Robin, Xavier et al. (2011). “pROC: an open-source package for R and S+ to analyze and compare ROC curves”. In: *BMC Bioinformatics* 12.1, pp. 1–8. DOI: 10.1186/1471-2105-12-77. URL: http://dx.doi.org/10.1186/1471-2105-12-77.

Roy, Sushmita et al. (2010). “Identification of functional elements and regulatory circuits by Drosophila modENCODE”. In: *Science*, p. 1198374.

Roy, Sushmita et al. (2013). “Integrated module and gene-specific regulatory inference implicates upstream signaling networks”. In: *PLoS computational biology* 9.10, e1003252.

Runge, Steffen et al. (2000). “H19 RNA binds four molecules of insulin-like growth factor II mRNA-binding protein”. In: *Journal of Biological Chemistry* 275.38, pp. 29562–29569.

Rzepiela, Andrzej J et al. (2018). “Single-cell mRNA profiling reveals the hierarchical response of miRNA targets to miRNA induction”. In: *Molecular Systems Biology* 14.8. DOI: 10.15252/msb.20188266. eprint: http://msb.embopress.org/content/14/8/e8266.full.pdf.

Salmena, Leonardo et al. (2011). “A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?” In: *Cell* 146.3, pp. 353–8. ISSN: 1097-4172. DOI: 10.1016/j.cell.2011.07.014. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3235919{\&}tool=pmcentrez{\&}rendertype=abstract.

Sanchez-Mejias, Avencia and Yvonne Tay (2015). “Competing endogenous RNA networks: tying the essential knots for cancer biology and therapeutics”. In: *Journal of hematology & oncology* 8, pp. 30; 30–30. DOI: 10.1186/s13045-015-0129-1.

Sanguinetti, Guido et al. (2019). “Gene regulatory network inference: an introductory survey”. In: *Gene Regulatory Networks*. Springer, pp. 1–23.

Schäfer, Juliane and Korbinian Strimmer (2004). “An empirical Bayes approach to inferring large-scale gene association networks”. In: *Bioinformatics* 21.6, pp. 754–764.

— (2005). “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics”. In: *Statistical applications in genetics and molecular biology* 4.1.

Schmidt, Florian et al. (2016). “Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction”. In: *Nucleic acids research* 45.1, pp. 54–66.

Schulz, Marcel H. et al. (2013). “Reconstructing dynamic microRNA-regulated interaction networks”. In: *Proceedings of the National Academy of Sciences* 110.39, pp. 15686–15691. DOI: 10.1073/pnas.1303236110. URL: http://www.pnas.org/content/110/39/15686.abstract.

Scotti, Marina M and Maurice S Swanson (2016). “RNA mis-splicing in disease”. In: *Nature Reviews Genetics* 17.1, p. 19.

Seitz, Hervé et al. (2004). “A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain”. In: *Genome research* 14.9, pp. 1741–1748.

Simon, Yvette et al. (2014). “Elevated free cholesterol in a p62 overexpression model of non-alcoholic steatohepatitis”. In: *World journal of gastroenterology: WJG* 20.47, p. 17839.

Siomi, Mikiko C et al. (2011). “PIWI-interacting small RNAs: the vanguard of genome defence”. In: *Nature reviews Molecular cell biology* 12.4, p. 246.

Sirach, E et al. (2007). “KLF6 transcription factor protects hepatocellular carcinoma-derived cells from apoptosis”. In: *Cell death and differentiation* 14.6, p. 1202.

Sloan, Cricket A et al. (2015). “ENCODE data at the ENCODE portal”. In: *Nucleic acids research* 44.D1, pp. D726–D732.

Sonawane, Abhijeet Rajendra et al. (2017). “Understanding tissue-specific gene regulation”. In: *Cell reports* 21.4, pp. 1077–1088.

Song, Lingyun and Gregory E Crawford (2010). “DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells”. In: *Cold Spring Harbor Protocols* 2010.2, pdb–prot5384.

Spies, Noah, Christopher B Burge, and David P Bartel (2013). “3’ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts”. In: *Genome research*.

Spitz, François and Eileen EM Furlong (2012). “Transcription factors: from enhancer binding to developmental control”. In: *Nature reviews genetics* 13.9, p. 613.

Stiuso, Paola et al. (2015). “MicroRNA-423-5p promotes autophagy in cancer cells and is increased in serum from hepatocarcinoma patients treated with sorafenib”. In: *Molecular Therapy-Nucleic Acids* 4, e233.

Stone, Mervyn (1974). “Cross-validatory choice and assessment of statistical predictions”. In: *Journal of the royal statistical society. Series B (Methodological)*, pp. 111–147.

Sumazin, Pavel et al. (2011). “An extensive MicroRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma”. In: *Cell* 147.2, pp. 370–381. ISSN: 00928674. DOI: 10.1016/j.cell.2011.09.041. URL: http://dx.doi.org/10.1016/j.cell.2011.09.041.

Sun, Cheng-Cao et al. (2016). “Long Intergenic Noncoding RNA 00511 Acts as an Oncogene in Non-small-cell Lung Cancer by Binding to EZH2 and Suppressing p57”. In: *Molecular therapy. Nucleic acids* 5.11, e385–e385. DOI: 10.1038/mtna.2016.94.

Suzuki, Masahiro et al. (2010). “Up-regulation of glypican-3 in human hepatocellular carcinoma”. In: *Anticancer research* 30.12, pp. 5055–5061.

Tay, Yvonne, John Rinn, and Pier Paolo Pandolfi (2014). “The multilayered complexity of ceRNA crosstalk and competition”. In: *Nature* 505.7483, p. 344.

Teng, Mingxiang et al. (2016). “A benchmark for RNA-seq quantification pipelines”. In: *Genome Biology* 17.1, pp. 1–12. DOI: 10.1186/s13059-016-0940-1. URL: http://dx.doi.org/10.1186/s13059-016-0940-1.

Therneau, Terry M (2015). *A Package for Survival Analysis in S*. version 2.38. URL: https://CRAN.R-project.org/package=survival.

Tian, Bin and James L Manley (2017). “Alternative polyadenylation of mRNA precursors”. In: *Nature reviews Molecular cell biology* 18.1, p. 18.

Tibshirani, Robert (1996). “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Torres, Tatiana Teixeira et al. (2008). “Gene expression profiling by massively parallel sequencing”. In: *Genome research* 18.1, pp. 172–177.

Tsang, John S, Margaret S Ebert, and Alexander van Oudenaarden (2010). “Genome-wide dissection of microRNA functions and co-targeting networks using gene-set signatures”. In: *Molecular cell* 38.1, pp. 140–153. DOI: 10.1016/j.molcel.2010.03.007. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3110938/.

Tybl, Elisabeth et al. (2011). “Overexpression of the IGF2-mRNA binding protein p62 in transgenic mice induces a steatotic phenotype”. In: *Journal of hepatology* 54.5, pp. 994–1001.

Van Kouwenhove, Marieke, Martijn Kedde, and Reuven Agami (2011). “MicroRNA regulation by RNA-binding proteins and its implications for cancer”. In: *Nature Reviews Cancer* 11.9, p. 644.

Van Nostrand, Eric L et al. (2018). “A large-scale binding and functional map of human RNA binding proteins”. In: *bioRxiv*, p. 179648.

Viswanathan, Srinivas R and George Q Daley (2010). “Lin28: A microRNA regulator with a macro role”. In: *Cell* 140.4, pp. 445–449.

Vivian, John et al. (2016). “Rapid and efficient analysis of 20,000 RNA-seq samples with Toil”. In: *bioRxiv*, p. 062497.

Vivian, John et al. (2017). “Toil enables reproducible, open source, big biomedical data analyses”. In: *Nature biotechnology* 35.4, pp. 314–316. DOI: 10.1038/nbt.3772.

Wainwright, Martin J (2009). “Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso)”. In: *IEEE transactions on information theory* 55.5, pp. 2183–2202.

Wang, Chengyang et al. (2012). “Computational inference of mRNA stability from histone modification and transcriptome profiles”. In: *Nucleic acids research* 40.14, pp. 6414–6423.

Wang, Eric T et al. (2008a). “Alternative isoform regulation in human tissue transcriptomes”. In: *Nature* 456.7221, p. 470.

Wang, Jiayi et al. (2010). “CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer”. In: *Nucleic Acids Research* 38.16, pp. 5366–5383. ISSN: 13624962. DOI: 10.1093/nar/gkq285.

Wang, Peng et al. (2015). “Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer”. In: *Nucleic Acids Research* 43.7, pp. 3478–3489. ISSN: 13624962. DOI: 10.1093/nar/gkv233.

Wang, Yanli et al. (2008b). “Structure of the guide-strand-containing argonaute silencing complex”. In: *Nature* 456.7219, p. 209.

Wang, YX Rachel and Haiyan Huang (2014). “Review on statistical methods for gene network reconstruction using expression data”. In: *Journal of theoretical biology* 362, pp. 53–61.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). “RNA-Seq: a revolutionary tool for transcriptomics”. In: *Nature reviews genetics* 10.1, p. 57.

Widmer, Christian and Gunnar Rätsch (2012). “Multitask learning in computational biology”. In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pp. 207–216.

Wu, Xiaowei et al. (2016). “MiR-153 promotes breast cancer cell apoptosis by targeting HECTD3”. In: *American journal of cancer research* 6.7, p. 1563.

Xu, Juan et al. (2015). “The mRNA related ceRNA–ceRNA landscape and significance across 20 major cancer types”. In: *Nucleic Acids Research* 43.17, pp. 8169–8182.

Xu, Zenglin et al. (2010). “Simple and efficient multiple kernel learning by group lasso”. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. Citeseer, pp. 1175–1182.

Yasui, Kohichiroh et al. (2003). “Association of over-expressed TFDP1 with progression of hepatocellular carcinomas”. In: *Journal of human genetics* 48.12, p. 609.

You, Yu et al. (2018). “MicroRNA-766-3p Inhibits Tumour Progression by Targeting Wnt3a in Hepatocellular Carcinoma”. In: *Molecules and cells* 41.9, p. 830.

Zemel, Sharon, Marisa S Bartolomei, and Shirley M Tilghman (1992). “Physical linkage of two mammalian imprinted genes, H19 and insulin–like growth factor 2”. In: *Nature genetics* 2.1, p. 61.

Zhang, Bin and Steve Horvath (2005). “A general framework for weighted gene co-expression network analysis”. In: *Statistical applications in genetics and molecular biology* 4.1.

Zhang, Guoqiang and Richard Heusdens (2012). “Linear coordinate-descent message passing for quadratic optimization”. In: *Neural computation* 24.12, pp. 3340–3370.

Zhang, Jian-Ying et al. (1999). “A novel cytoplasmic protein with RNA-binding motifs is an autoantigen in human hepatocellular carcinoma”. In: *Journal of Experimental Medicine* 189.7, pp. 1101–1110.

Zhang, Junpeng et al. (2017). "Inferring miRNA sponge co-regulation of protein-protein interactions in human breast cancer". In: *BMC bioinformatics* 18.1, p. 243.

Zhang, Yijun et al. (2014). "Cellular microRNAs up-regulate transcription via interaction with promoter TATA-box motifs". In: *Rna*.

Zhang, Yunpeng et al. (2016). "Comprehensive characterization of lncRNA-mRNA related ceRNA network across 12 major cancers". In: *Oncotarget* 7.39, p. 64148.

Zhao, Peng and Bin Yu (2006). "On model selection consistency of Lasso". In: *Journal of Machine learning research* 7.Nov, pp. 2541–2563.

Zhao, Xiaohui et al. (2018). "Linc00511 acts as a competing endogenous RNA to regulate VEGFA expression through sponging hsa-miR-29b-3p in pancreatic ductal adenocarcinoma". In: *Journal of cellular and molecular medicine* 22.1, pp. 655–667. DOI: 10.1111/jcmm.13351.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.

Zucman-Rossi, Jessica et al. (2015). "Genetic landscape and biomarkers of hepatocellular carcinoma". In: *Gastroenterology* 149.5, pp. 1226–1239.