INVESTIGATING THE ROLE OF RNA-BINDING PROTEINS (RBPS) IN EXPLAINING DIFFERENTIAL GENE EXPRESSION IN CANCER

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ATEFEH LAFZI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE
IN
BIOINFORMATICS

FEBRUARY 2016

Approval of the thesis:

INVESTIGATING THE ROLE OF RNA-BINDING PROTEINS (RBPS) IN EXPLAINING DIFFERENTIAL GENE EXPRESSION IN CANCER

submitted by **ATEFEH LAFZI** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics , Middle East Technical University** by,

Prof. Dr. Nazife Baykal
Director, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics** _____

Assoc. Prof. Dr. Yesim Aydin Son
Supervisor, **Health informatics, METU** _____

Assist. Prof. Dr. Hilal Kazan
Co-supervisor, **Computer Engineering, AIU** _____

Examining Committee Members:

Assoc. Prof. Dr. Rengul Çetin Atalay
Health Informatics, METU _____

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, METU _____

Assist. Prof. Dr. Hilal Kazan
Computer Engineering, AIU _____

Assist. Prof. Dr. Aybar Can Acar
Health Informatics, METU _____

Assist. Prof. Dr. Bala Gür Dedeoğlu
Biotechnology Institute, Ankara University _____

Date: **02.02.2016**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ATEFEH LAFZI

Signature :

ABSTRACT

INVESTIGATING THE ROLE OF RNA-BINDING PROTEINS (RBPS) IN EXPLAINING DIFFERENTIAL GENE EXPRESSION IN CANCER

Lafzi, Atefeh

M.S., Bioinformatics Program

Supervisor : Assoc. Prof. Dr. Yesim Aydin Son

Co-Supervisor : Assist. Prof. Dr. Hilal Kazan

February 2016, 50 pages

Most of the studies on cancer have tried to explain the observed differential gene expression considering only transcriptional regulation. However, post-transcriptional regulation (PTR) has been increasingly recognized as a complex mechanism that also controls various steps of gene expression regulation. Post-transcriptional regulation is governed by the interactions of RNA-binding proteins (RBPs) and microRNAs (miRNAs) with their target genes. In this thesis, having found that several RBPs are differentially expressed in Lung squamous cell carcinoma (LUSC), we developed a statistical model which incorporates copy number variation, DNA methylation and the regulatory effects of transcription factors, miRNAs and RBPs to predict gene expression in cancer. Including RBP-based regulation in addition to other features significantly increased the Spearman rank correlation between predicted and measured expression of held-out genes. Using a feature selection procedure we identified the candidate RBP regulators in LUSC and confirmed that many of them are also differentially expressed. We also determined the targets of these RBPs and compared them with CLIP-determined targets. Lastly, we performed Kaplan-Meier survival analysis, and showed that some of our candidate RBP regulators have prognostic power in LUSC. Our results suggest that the regulatory effects of RBPs have to be considered to explain gene expression in cancer.

Keywords: Post-transcriptional regulation, RNA-binding proteins, microRNA, Lung squamous cell carcinoma, linear regression

ÖZ

RNA'YA BAĞLANAN PROTEİNLERİN KANSERDE GEN İFADE DEĞİŞİMİNE OLAN ETKİLERİNİN İNCELENMESİ

Lafzi, Atefeh
Yüksek Lisans, Biyoenformatik Programı
Tez Yöneticisi : Doç. Dr. Yesim Aydin Son
Ortak Tez Yöneticisi : Yrd. Doç. Dr. Hilal Kazan

Şubat 2016 , 50 sayfa

Kansere yol açan etmenleri bulmayı amaçlayan çalışmalar özellikle kanserli ve normal hücreler arasında farklı ifadesi olan genlerin regülasyonunu incelemektedir. Şu ana kadarki çalışmaların büyük bir kısmı sadece transkripsiyonel kontrolle ilgili etmenleri dikkate alarak bu ifade değişimlerini açıklamaya çalışmıştır. Son çalışmalar, transkripsiyon sonrası kontrolün (TSK) de gen ifadelerini kontrol eden önemli bir mekanizma olduğunu göstermiştir. Transkripsiyon sonrası kontrol RNA'ya bağlanan proteinler (RBP) ve mikroRNA'ların (miRNA) hedef genlere bağlanmasıyla gerçekleştirilmektedir. Bu tez kapsamında, ifadesi değişen RBP'lerin sayısının en fazla olduğu LUSC (akciğer sküamoz karsinomu) kanserinde, gen kopya sayılarını, DNA metilasyonunu, transkripsiyon faktörlerin, miRNA'ların ve RBP'lerin etkilerini göz önüne alarak tahmin eden bir istatistiksel model geliştirildi. Diğer özniteliklere ek olarak RBPlerin kullanılması bu modelle tahmin edilen ifadelerle bilinen gen ifadeleri arasındaki Spearman korelasyonunu önemli ölçüde arttırdı. Öznitelik seçimiyle LUSCde önemli rol oynayan RBP'ler bulunmuş ve bu RBPlerin ifadelerinin değişim gösterdiği tespit edilmiştir. Modelde öğrenilen parametreler incelenerek bu RBPlerin hedefleri bulunmuş ve CLIP-deneyiyle bulunan hedeflerle karşılaştırılmıştır. Son olarak Kaplan-Meier analizi ile bu RBPlerin bazılarının kurtulma olasılığını tahmin edebildiği bulunmuştur. Bu sonuçlar kanserde gen ifade değişimlerinin daha iyi anlaşılması için RBPlerin de göz önüne alınması gerektiğini göstermektededir.

Anahtar Kelimeler: Transkripsiyon sonrası kontrol (TSK), RNA'ya bağlanan protein-

ler (RBP), mikroRNA'lar, Akciğer kanseri, lineer regresyon

Dedicated to my family...

ACKNOWLEDGMENTS

I owe my gratitude to many great people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my co-supervisor Dr. Hilal Kazan. I have been so fortunate to get the chance to involve in her TÜBITAK project (Project No: 113E159) as a research assistant during my master studies. Working on this project and under her supervision would be always the most pivotal point of my academic life which transformed me from an undergraduate student to a master-level researcher. She taught me how to explore on my own while she was always available whenever I needed an advice. I really want to thank her for her support and patience.

I would also thank my supervisor Dr. Yeşim Aydın Son for her endless encouragement, continuous support and friendly attitude in every step and each difficulty throughout my master studies.

I am grateful to Dr. Tolga Can, Dr. Rengul Çetin Atalay, Dr. Aybar Can Acar, Dr. Bala Gür Dedeoğlu and Dr. Nurcan Tunçbağ for their valuable courses, each of which helped me to enrich my knowledge for this work, and also for serving on my examining committee.

I also want to thank my dear friends Sahar Habibiabad and Reza Soleimani for their precious friendship, Saber Hafezqorani for being a great colleague during all happy and sad days of project, Esma Pala for being a joyful and lovely friend who always made me laugh during my intense working days, and finally my special thanks goes to Majid Biazaran for all his emotional and technical support, care and companion during last months of my hard work.

Finally, I would like to thank my dear family for their endless support through my entire life. My mom Azar Goudarzi for enduring all the lonesomeness for having me so far from her during these years, my dad Mahdi Lafzi, for being the most supportive, kind, intellectual and open-minded father of all time and my beloved brother, Alireza Lafzi who is the most trustable person in my life.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Aberrant gene expression is a main feature of cancer development. Characterizing the regulatory events that lead to gene expression changes during cancer development is critical for cancer research. Differential gene expression in cancer can occur due to several factors including copy-number variation (CNV), DNA methylation changes, and alterations in transcriptional and post-transcriptional regulatory mechanisms. Among these factors, post-transcriptional regulation (PTR) has gained significant importance due to its emerging roles in cancer biology.

PTR is mediated by the interactions of RNA-binding proteins (RBPs) and microRNAs (miRNAs) with target mRNAs through short sequence and/or structure motifs. Recent studies have found that RBPs are key regulators controlling every step of RNA metabolism including RNA splicing, transport, localization, decay and translation. More than 850 RBPs have been identified in the human genome [8, 17]. Recent advances in experimental methods that characterize the binding sites of RBPs have significantly expanded our knowledge of *in vivo* and *in vitro* RBP binding preferences [38, 85]. This recent explosion of knowledge on RBP binding sites provide opportunities to study RBP-mediated regulation in greater detail.

Several RBPs have been found to be implicated in cancer [55]. For example, overexpression of KHDRBS1 (Sam68) has been revealed in various cancer types including breast, prostate, colorectal and lung cancer cells [12, 33, 62]. KHDRBS1 is found to mediate the alternative splicing of oncogenes. ELAVL1 is another well-known RBP that is found to be associated with tumorigenesis by regulating the stability and translation of key growth factors and proto-oncogenes [1, 103]. Overexpression of ELAVL1 has been observed in many cancer types [11, 13]. Recently, FXR1 is found to regulate tumor progression in lung cancer, and is identified as a driver of the 3q amplicon, the most frequent genomic alteration in squamous cell lung cancers [83]. eIF4E is another RBP which studies show that its overexpression increases translational initiation of specific mRNAs in cancer. A recent study by Mitzutani et al also reveals the functional linkage between eIF4E and IGF2BP3, known as another RBP regulator of tumorigenesis, in cancer [74]. Also some other studies indicate QKI and SRSF1 as fundamental players in tumorigenesis [3, 112]. These and many other examples indicate that dysregulation of the function or the expression of RBPs has profound implications for cancer development. Although these single RBP studies help to uncover the PTR effect on cancer, there is a need for a more comprehensive study that takes into account transcriptional and post-trancriptional regulation as well

as other genetic and epigenetic factors like copy number variation (CNV) and DNA methylation to explain gene regulation in cancers.

Recently developed computational models that study gene expression in cancer have mainly focused on transcriptional regulation and miRNA-mediated regulation. For instance, Setty et al predicted expression changes in glioblastoma (GBM) with a lasso-regularized regression [93] . In addition to copy number variation (CNV) and methylation changes, they included features that correspond to transcription factor (TF) binding sites from TRANSFAC filtered by DNA hypersensitive regions, and miRNA binding sites obtained from scanning with 7-mer seed sequences. Their model predicted a number of key regulators from TFs and miRNAs that are predictive of survival rate in GBM. Jacobsen et al focused on miRNA-based regulation, ignoring transcriptional-regulation [46]. They looked at the relation between the expression of miRNAs and mRNAs in tumors from 11 human cancer types in TCGA, and identified a pan-cancer miRNA-mRNA network. Li et al proposed a two-stage regression framework that combines data from TCGA and ENCODE to predict gene expression in Acute Myeloid Leukemia (AML) [61] . Their model revealed a number of TFs and miRNAs as candidate regulators of AML. To the best of our knowledge, there is still no study that investigates the effects of RBP-mediated regulation in cancer.

In this study, we propose to explain gene expression in cancer with a statistical model that incorporates RBP-based regulation in addition to CNV, DNA methylation and the regulatory effects of transcription factors and miRNAs. As a case study, we applied our model to Lung squamous cell carcinoma (LUSC) dataset, as we found that there are a large number of differentially expressed RBPs in this cancer type. By comparing the performance of our full model with partial models that exclude one of the feature groups (e.g. TFs, CNV etc.), we show that the added predictive value of RBPs is higher than all the other feature groups. Following up on this result, we applied a feature selection procedure to identify the RBPs as well as other factors that play a key role in LUSC. Importantly, majority of our predicted candidate regulators are previously found to be associated with lung cancer, and are differentially expressed. Subsequently, we determined the targets of these candidate regulators, and compared against experimentally determined targets. Lastly, we performed Kaplan-Meier survival analysis in order to reveal the prognostic power of our candidate RBP regulators in LUSC. The results of this study suggest that future studies of gene regulation must consider the effects of RBP-mediated regulation.

This thesis is organized into five chapters as follows:

- **Chapter 2** provides background about cancer biology. We review several factors that are associated with cancer development: CNV, DNA methylation, TFs, miRNAs and RBPs. Here we discuss the experimental and computational approaches to identify the binding sites of TFs, miRNAs and RBPs. This chapter also reviews previous work about modeling ragulatory network in cancer.
- **Chapter 3** describes the methods and datasets used in this thesis. In particular, we explain how we mapped the genome-wide binding sites of RBPs, miRNAs and TFs. We also describe the details of the statistical model that we developed to predict gene expression in cancer.
- **Chapter 4** includes our results. In particular, we show that our model can pre-

dict gene expression in cancer accurately, and incorporation of RBPs is critical to achieve this performance. Lastly we go through the candidate regulators that we find with our model.

- **Chapter 5** summarizes the results and discusses possible future directions of this work.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we briefly explain the necessary background that is required to understand the overall study. We start by reviewing the genetic background of cancer and the effect of gene regulation in tumorigenesis. We then discuss the factors that govern differential gene expression in cancer that range from copy number variation to transcriptional and post-transcriptional regulation. We continue by introducing RNA binding proteins (RBPs) as major regulators of post-transcriptional regulation and we review the studies in which the effects of RBPs have been studied experimentally. Then, we introduce recent studies that use computational models to infer candidate regulators of gene expression in cancer. Finally we identify the missing piece in the literature, that is, no study has taken into account the effect of RBPs together with other regulators like transcription factors and miRNAs to explain gene expression in cancer.

2.1 Cancer

Cancer is caused by dynamic changes in the genome [39]. These changes could be on the genomic level involving mutations or copy number changes. They could be also in the expression level of the genes which control the way our cells function, especially how they grow and divide. Although some changes in the expression of cancer-related genes arise through genome level mutations in its early stages, most cancer-specific gene expression patterns involve genes that are not mutated, but differentially expressed [91]. So, in addition to conventional selection of only mutated genes as candidate cancer genes, cancer phenotypes resulting from altered gene expression should also be considered. Differential expression of genes in cancer can be explained by DNA methylation, gene copy number changes, transcriptional and post-transcriptional regulation. Now, we review each of these mechanisms below.

2.2 DNA copy number variation

As we mentioned above, technological advances in genome scanning, has revealed the fact that Single Nucleotide Polymorphisms (SNPs) are not the main source of genetic and phenotypic human variation. These technologies have uncovered unexpectedly large extent of 'structural variation' in the human genome. These structural variations

include deletions and duplications, as well as insertions, inversions and translocations [30]. Although the result of genome-wide studies have identified many common SNPs associations with cancer, these studies mostly ignore the inter-individual genetic variation provided by copy number variations (CNVs), which affect more than 10% of human genome [94]. Consequently, recent studies appreciate the role of CNV as risk factors for cancer and characterization of the location and extent of these regions in genomes becomes and important issue which is now achievable by the help of high-resolution SNP arrays.

2.3 Transcriptional regulation

Transcriptional regulation is one of the most studied part of gene regulation. Regulation at this level occurs at the state where mRNA is being transcribed to transfer the information to ribosomes, protein building machineries. The two most important processes that control the expression of gene at this stage are: 1) Binding of transcription factors (TFs) to promoter or enhancer regions of genes which initiates a program of increased or decreased gene transcription, 2) Addition of methyl groups to DNA in a gene promoter region and activating that gene, called DNA methylation.

2.3.1 Transcription Factors

Transcription factors are proteins that bind to specific regions on DNA and control which genes are turned on or off in the genome. These factors are so essential for regulation of genes so that through their action, various cells of the body that all have identical DNA sequence can function differently. Transcription factors bind to their specific motif on promoter or enhancer region of DNA, interact with each other to form complexes, and recruit RNA polymerase II [58]. Enhancers are usually up to 500 base pairs long and they can contain multiple binding sites for two or more transcription factors [58]. For example when two TFs bind near each other on the DNA strand, they can form a dimer and bend the DNA which is believed to be an activating process. So they can perform individually or form complexes with other TFs to regulate the expression of transcripts.

Though the identification of putative transcription factor binding sites has been one of the most challenging problems, recent novel experimental techniques like chromatin immunoprecipitation (ChIP), has permitted the direct genome-wide identification of TF-DNA interactions. The length of the binding sites of TFs range from 8-10 to 16-20 nts. TFs bind to DNA in a sequence-specific manner which means that they are able to recognize sequences that are similar but not identical, differing in a few nucleotides from one another [106]. Currently, ENCODE [78] is a large project that tries to systematically map regions of transcription, transcription factor association, chromatin structure and histone modification. Also JASPAR [68] is a large database that contains matrix-based nucleotide profiles (Position Frequency Matrices, PFMs) that describes the binding preference of TFs in multiple species. These PFMs are based on published experiments from diverse sources.

2.3.2 DNA methylation

DNA methylation is one of the epigenetic factors that regulates gene expression in transcriptional level. This is the process of conversion of cytosine bases of DNA to 5-methylcytosine by methyltransferase enzymes. It typically occurs in CpG sequences throughout the entire genome while CpG islands (regions with elevated CpG content) which are often found in gene promoters, remains unmethylated in normal tissues [4]. Methylation of these sequences can lead to inappropriate silencing of genes like silencing of tumor suppressors in cancer [82]. In addition, pioneer studies in cancer cells reports that global and gene specific loss of DNA methylation is also associated with neoplastic transformation [26, 27]. All these data show that there is a massive disruption of DNA methylome in tumor cells compared to normal cells [56]. Therefore, early detection of hypermethylation can serve as a biomarker for cancer.

The major advance in measuring the methylation state of cytosines is Bisulfite sequencing which modifies sodium bisulfite of DNA to convert unmethylated cytosines to uracil and leaving methylated cytosines unchanged [19]. After PCR amplification and conversion of unmethylated cytosines to thymines, by mapping these bisulfite treated DNA to the original reference genome, it is possible to determine the methylation state of individual cytosines [28]. This method is used to measure DNA methylation at specific loci. However, the advances in next-generation sequencing enables to determine the methylation state of an entire genome using quantitative methods. These methods usually work by preparing bisulfite treated libraries, and sequencing them using sequencers like *Illumina DNA methylation* or *Human Methylation 27/450* platforms. Most of these results are available through data portals like TCGA [6].

2.4 Post-transcriptional regulation

Post-transcriptional regulation (PTR) is the second level of control that happens after the pre-mRNA has been transcribed. Regulation at this level includes the control of splicing, localization, degradation and translation. Several recent studies support the observation that post-transcriptional control is a pervasive and complex system. For example, 90% of mammalian genes undergo alternative splicing [102], and 60% of the variation between mRNA and protein levels can be explained by post-transcriptional effects [37, 92]. PTR is governed by trans-acting factors that bind to short regions of mRNAs and control their fates. Most important trans-acting factors in PTR are microRNAs (miRNAs) and RNA-binding proteins (RBPs). Recent studies have shown that miRNAs and RBPs can bind to a common set of mRNAs, and they can also act in cooperative or competitive interactions.

2.4.1 MicroRNAs

MicroRNAs (miRNAs) are small (22 nt long) non-coding RNAs that function in various biological pathways. miRNAs are transcribed by RNA polymerase II as pri-microRNAs and then processed by RNase III enzyme Drosha to form pre-microRNA. This precursor is then transported to the cytoplasm where it is processed by DICER

to form a mature microRNA. This mature microRNA mediates gene silencing by forming RNA-induced silencing complex, RISC. miRNAs regulate gene expression by binding to target mRNAs containing complementary sequences. Recent studies show that over half of the human transcripts are subject to miRNA regulation through degradation or translation inhibition [9].

Hundreds of miRNAs have been found in the human genome. They bind to their target sites by either complete or partial complementary base pairing on 3'UTRs of target mRNAs. The base pairing usually happens in the region which is positioned at 2-7 nt of miRNAs and is named as the 'seed region'. By complete complementary base paring of seed region of miRNA with 3'UTR of mRNA, the target mRNA will be downregulated. miRNAs and mRNAs can have a many-to-many relationship. A miRNA can bind to hundreds of mRNAs and an mRNA can contain target sites of several miRNAs. miRNAs have shown effects in various cancer types, such as lung, breast and prostate cancer. They also play important role in some neurological disorders such as schizophrenia and Alzheimer's diseases.

MiRNA targets can be identified through techniques such as qRT-PCR, western blot and luciferase reporter assays. These methods can only be used to identify a small number of targets of the miRNA of interest, while high throughput approaches have been also developed to identify the genome-wide targets of a miRNA. Measuring the changes of gene expression upon miRNA transfection or inhibition using microarray or RNA-seq techniques are examples of high throughput techniques. Also there are biochemical approaches to find miRNA targets involved in the immunoprecipitation of the RISC component using an antibody against the Argonaute (AGO) protein [51]. Then, either by microarray or sequencing, precipitates are analyzed to identify the targets. Cross-linking by ultraviolet radiation technique is another alternative to identify miRNA sites. For instance, PARCLIP has been applied in HEK293 cells to identify miRNA target sites [38].

Currently, there are large number of databases that contain miRNA targets either identified experimentally or computationally. For example miRTarBase contains a large number of experimentally validated miRNA-target interactions. These targets are compiled by data-mining existing literature [43]. These miRNA-target interactions are grouped based on the type of evidence. Namely, targets identified with reporter assay, western blot and qPCR provide strong evidence; whereas targets identified with approaches such as microarray, NGS-based methods, pSILAC provide weak evidence. On the other hand, TargetScan is another widely used methods which considers complementarity to seed region and conservation information to predict miRNA target sites [59]. Pictar and GenMiR are among the popular methods to predict miRNA target sites. Several factors play a role on the regulatory effect of miRNAs. The location and secondary structure of binding sites of miRNAs are examples of these factors which may enhance the chance of miRNAs to access their targets. Studies show that miRNAs prefer to bind to AU-rich regions and unstructured areas in 3'UTRs. Another factor is the number of miRNA binding sites that plays a crucial role in the regulation of mRNAs. Transcripts that contain more binding sites have higher probability of being regulated by the miRNA of interest. Besides, miRNAs may involve in competitive or collaborative interactions with RBPs. Previous studies have investigated the effect of miRNAs or RBPs independently. However, recent studies show that miRNAs and RBPs are involved in complex gene regulation mech-

anisms.

2.4.2 RNA-binding proteins

RBPs are another important factors in regulation of gene expression at the mRNA level. They bind to 3'UTR end of their target mRNAs and control several steps of RNA processing including splicing, stability, localization, and degradation [24]. RBPs bind to their targets through one or more RNA-binding domains (RBDs). Some RBPs prefer binding to a single-stranded RNA by direct readout of the primary sequence, while others recognize the structure of the RNA [23, 111]. There are also some RBPs which recognize their targets by both the sequence and the secondary structure [50]. As it was for miRNAs, RBP-mRNA interactions can form a multi-dimensional network such that an mRNA can be bound by several RBPs and each RBP can have hundreds of targets. More than 800 RBPs have been identified in the human genome. Alternation in the activity of RBPs may cause many diseases such as neurodegenerative disorders and cancer.

In recent years, several *in vitro* and *in vivo* experimental methods have been developed to identify binding specificities of RBPs. While *in vitro* studies are conducted in some controlled environment, *in vivo* experiments are carried out within the cell which is advantageous due to modeling RBP-RNA interactions in natural environment. On the other hand, *in vitro* methods identify RBP binding sites in non-biological conditions but this also enables querying non-genomic sequences such as variants of the wild-type binding sites and testing a wide range of interesting conditions (e.g. salts, pH).

In vitro methods: One of the recently developed *in vitro* methods to identify binding specificities of RBPs is RNAcompete [85]. This method identifies binding specificities of RBPs in three main steps: (i) generation of a custom-designed RNA pool containing short sequences; (ii) a single binding reaction to identify the RNAs bound by the tagged RBP of interest; and (iii) analysis of the microarray data to determine binding preferences of the RBP. Recently, this method is used to characterize the binding specificities of more than 200 RBPs from 24 diverse eukaryotes. Figure 2.1 summarizes the steps in RNAcompete method. There are other *in vitro* methods as well like SELEX [25], but we will use RNAcompete for the purpose of this thesis.

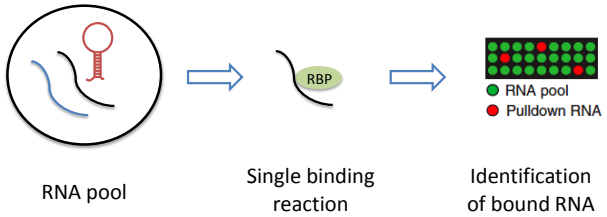


Figure 2.1: Steps of RNAcompete methods are summarizes in this figure (Adopted from RNAcompete [85])

In vivo methods: *in vivo* methods are carried out within the cell. RNA immunoprecipitation (RIP) and Cross-linking and immunoprecipitation (CLIP or HITS-CLIP) are among the *in vivo* methods of identifying binding specificities of RBPs. RNA immunoprecipitation detects the association of an individual RBP with specific RNA.

RIP purifies RBP-mRNA complexes from cellular extracts and identifies protein-bound mRNAs using either a microarray (RIP-chip) or high-throughput sequencing (RIP-seq) [54, 110]. The absence of cross-linking step in this method may cause dissociation of RBPs from their targets [73]. CLIP method aims to eliminate the dissociation of RBPs mentioned in the previous method. Figure 2.2 shows the steps in PARCLIP and HITS-CLIP methods. CLIP approach uses an ultraviolet light (UV) cross-linking step before immunoprecipitation. This further step causes a more stringent washing procedure to reduce contaminants and eliminates interactions that occur after cell lysis [100]. PARCLIP is a modified version of CLIP technique which uses photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation. In PARCLIP, cross-linked sites are enriched with a thymidine to cytidine transition [38]. This method identifies RBP binding sites by scoring T to C transitions in sequenced cDNA. CLIP experiments identify the binding sites of a single RBP at a time while a recently proposed method called gPARCLIP (global PARCLIP) is able to identify the binding sites of all the RBPs in the cell [8]. However, this method is unable to detect and identify the particular RBPs that bind to a site.

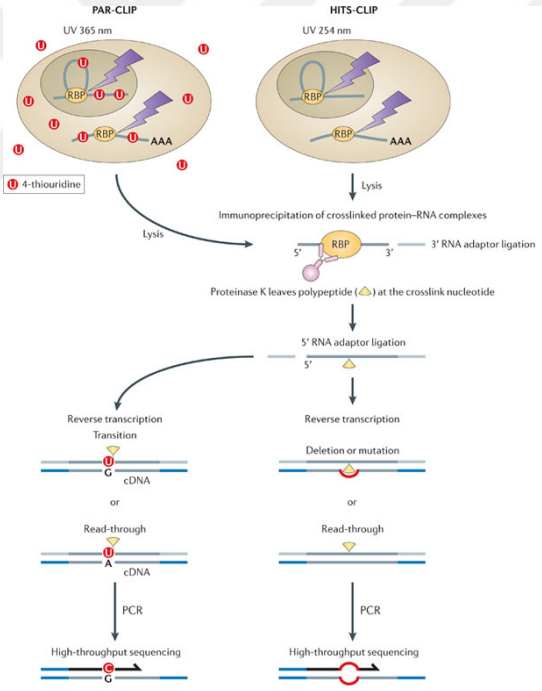


Figure 2.2: Various steps in PARCLIP and HITS-CLIP methods (Figure and description adopted from König et al. [57])

In silico methods: These methods try to predict the binding sites of the RBPs computationally. There are tools available like MEME [7] and HOMER [40] which try to identify the binding motifs from a set of RNA sequences that are known to be bound by the RBP. Recently developed methods infer RNA motifs that include both sequence and structure features [41, 53, 69] but all of these methods require a set of bound RNA sequences so that they are based on the results of experimental methods.

Researchers also try to predict RBP-RNA interactions using features such as 3'UTR characteristics, RBP properties, di-nucleotide content, RNA secondary structure statistics. Pancaldi et. al. [81], uses this de novo method in their study and integrates more

than 100 features to predict the target mRNAs of RBPs in yeast. RPI-seq is another method by Muppirala et. al. [76] which uses the 4-mer composition of the RNA sequence and the 3-mer composition of the RBP sequence to train a statistical model. Muppirala et. al. try machine learning algorithms such as Support Vector Machine (SVM) and Random Forest (RF) to predict whether the RNA-RBP pair will interact or not.

2.5 RBPs in cancer

Studies show that RBPs are master regulators of gene expression in cancer but much remains ahead in understanding how RBPs and their regulatory network contribute to tumor initiation and progression. Some researchers try to identify the effect of individual RBPs in cancer experimentally but this method reveals a small number of RBPs that may contribute directly to cancer progression.

IGF2BP protein family is an evolutionary conserved family of RBPs that also has been reported as a useful diagnostic marker for various cancers. This family is formed from three RBPs: IGF2BP1, IGF2BP2 and IGF2BP3. The first family member, IGF2BP1, is known to promote tumor cell proliferation and survival in various cancer contexts [10]. It controls the subcellular sorting of the ACTB mRNA in primary fibroblasts and neurons by binding to the cis-acting zipcode in the 3'UTR of ACTB [89]. By doing so, it enhances neurite outgrowth and axonal guidance by controlling the spatially restricted translation of the ACTB mRNA [45]. The second member, IGF2BP2, has been proposed as an auto-antigen (i.e., any antigen that stimulates autoantibodies in the organism that produced it) in hepatocellular carcinoma [107]. Also SNPs that are identified in IGF2BP2 gene were correlated with an elevated risk of type 2 diabetes [18]. Finally, the last member of this RBP family, IGF2BP3, has been reported to be the mainly expressed family member in human cancer [31]. Studies suggest IGF2BP1 and IGF2BP3 as *bona fide* oncofetal proteins [79], however, they are found to be re-expressed in several aggressive cancer types. Moreover, this protein family is highly associated with cancer metastasis and the expression of oncogenic factors (KRAS, MYC and MDR1) [10]. A recent study by Mitzutani et. al. also reveals the functional linkage between IGF2BP3 and eIF4E, known as another RBP regulator of tumorigenesis, in cancer [74]. SRSF1 is another protein which has been demonstrated as an oncogene, and its oncogenic activity is mediated by controlling alternative splicing of the tumor suppressor BIN1 and the kinases Mnk2 and S6K1 which induces the expression of pro-tumorigenic isoforms [52]. This protein is a prototypical SR protein that functions as a splicing factor [3]. It also plays an important role in translation [72], mRNA export [44] and mRNA decay [108]. Studies on this protein suggest that this RBP is frequently overexpressed in cancer [3]. It has a role in development of mammary tumors in human [96] and it is overexpressed in most of breast tumor panel. Another example is QKI, which is one of the conserved STAR (Signal Transduction and Activation RNA) family proteins. STAR family proteins are usually involved in various aspects of RNA metabolism like mRNA splicing, localization and transport [29], and have been recently identified as key regulators of alternative splicing in lung cancer [112]. Zong et. al. have shown that QKI is frequently down-regulated in cancer and its down-regulation is significantly associated

with a poorer prognosis. Their results illustrate that QKI-5 regulates the alternative splicing of NUMB by binding to two RNA elements in its pre-mRNA, which in turn suppresses cell proliferation and prevents the activation of the Notch signaling pathway.

Sam68 (KHDRBS1) is another RBP found to be associated with cancer and regulates alternative splicing of cancer-related mRNAs [12]. Sam68 is known to be overexpressed in breast, prostate, renal, and cervical cancer cells [15, 109] and is also frequently upregulated in tumors [84]. eIF4E is another RBP which its overexpression is highly associated with malignancy and poor prognosis [20, 35]. Studies show that its overexpression increases translational initiation of specific mRNAs in cancer [105]. Finally, ELAVL1 (HuR) is the most prominent RBP known to be implicated in tumorigenesis [1]. Overexpression of HuR has been observed in lymphomas, gastric, breast, pancreatic, prostate, oral, colon, skin, lung, ovarian, and brain cancers [11, 13] and it is known to regulate the stability and translational of transcripts involved in cancer.

Although these single RBP studies help to uncover the PTR effect on cancer, there is a need for a more comprehensive study that takes into account both transcriptional and post-trancriptional regulation as well as other genetic and epigenetic factors like copy number variation (CNV) and DNA methylation to explain gene regulation in cancer.

2.6 Inferring candidate regulators of cancer

Unlike the simple assumption that cancer development is dictated mostly by aberrant transcriptional events, it is now clear that post-transcriptional regulation of gene expression also controls cell proliferation, differentiation, invasion, metastasis, apoptosis, and angiogenesis which influence initiation and progression of cancer. The large amount of data explaining genome wide signals of many regulatory factors in online resources like TCGA and ENCODE gives us the opportunity to study cancer comprehensively by considering many factors that govern differential gene expression in cancer both at the transcriptional and post-transcriptional level.

Recent studies have used statistical models to utilize these data in order to model the regulatory network of cancer. Setty et al [93] model the regulatory network of glioblastoma (GBM) using a regularized linear regression model. They use genomic and epigenomic features like copy number variation and DNA methylation, as well as binary TF-binding sites from TRANSFAC filtered by DNA hypersensitive regions from the ENCODE data at transcriptional level and miRNA binding sites obtained from scanning 7-mer seed matches on 3'UTR at post-transcriptional level. They have applied their approach on 320 GBM samples and have identified miR-124 and miR-132 as drivers of the proneural subtype of GBM. Another study by Jacobsen et al [46] predicts miRNA-mRNA interactions across 11 cancer types proposing similar model, but does not consider transcriptional regulation effect of transcription factors which may be susceptible of overestimating the influence of the miRNAs [61]. More recently Li et al [61] have developed a two-stage regression framework (RACER) which infers sample-specific TF and miRNA activities. It obtains ENCODE TF binding data

derived from a generic cell-line and conserved miRNA binding sites from TargetScan. v6. and uses the estimated regulatory activities to infer miRNA/TF-gene regulatory relationships across samples in Acute Myeloid Leukemia (AML).

All of the studies described above ignore the effect of RBPs in gene regulation. In this study, we fill this gap by proposing a regression model which incorporates copy number variation, DNA Methylation and the regulatory effects of transcription factors, miRNAs and RBPs to predict gene expression in cancer.

CHAPTER 3

MATERIALS AND METHODS

3.1 Data integration and preprocessing

We used *The Cancer Genome Atlas (TCGA)* as our main source of data. TCGA is a project started in 2006 that aims to utilize advanced genomic technologies, to generate comprehensive and multi-dimensional maps of the key genomic changes in major types and subtypes of cancer. This project has finalized tissue collection with matched tumor and normal tissues from 11,000 patients, allowing for the comprehensive characterization of 33 cancer types and subtypes, including 10 rare cancer types. This project has been split into two parts, I) Genome characterization center (GCCs) which performs the sequencing, II) Genome data analysis center (GDACs) which performs the bioinformatic analysis and tries to process data obtained from first part. TCGA is sequencing whole genome of tumors including at least 6000 genes and microRNAs. It provides various genomic data like gene expression profiles, copy number variation profiles, DNA methylation profiles, microRNA profiles, SNP genotypes and exon sequences for at least 1200 genes for each cancer type.

TCGA have sequenced whole genome of tumors including at least 6000 genes and microRNA. It provides various genomic data like gene expression profiles, copy number variation profiles, DNA methylation profiles, microRNA profiles , SNP genotypes and exon sequences for at least 1200 genes for each cancer type.

On the other hand, organizations like Broad Institute aim to facilitate the extraction of these data by developing applications and quantitative algorithms. *Firehose Broad GDAC* tries to systematically analyse data produced by TCGA and provides more than 55 terabytes of analysis-ready TCGA data and also executes thousands of pipelines per month. These analyses include identifiying genomic regions that are significantly gained or lost across a set of tumors, calculating correlations between DNA methylation and gene expression profiles, inferring putative direct gene targets of miRNAs based on miRseq and mRNAseq expression profiles across multiple samples and etc. *UCSC Cancer Genomics Browser* is another web-based application that integrates datasets from TCGA and other related data sources and provides easy access and analysis of these data to researchers in order to discover and share their research observations.

3.1.1 mRNA and miRNA expression

All mRNA and miRNA expression data were obtained from TCGA data portal [6]. We downloaded RNASeq V2 (level 3) gene expression data and used normalized read counts. Also we filtered mRNAs that are not expressed across most samples in RNA-seq data sets, by removing those having less than 4 reads in more that 70 percent of samples. To allow log transformation, we added 1 to all read counts across all samples. The RNA-seq mRNA expression values were log2 transformed for all subsequent analysis. For miRNA expression we combined level 3 Illumina HighSeq (387 samples) and Illlumina GA (136 samples) data and we processed the miRNA expression data similar to RNA-seq data.

3.1.2 Copy number variation and DNA methylation

SNP6 Copy number level 4 data processed by GISTIC2 algorithm [71] were obtained from Firehose [1]. Firehose employs the GISTIC2 algorithm to generate copy number estimates for the genes mapped in the genome. This algorithm has been upgraded to a number of analytical improvements to the standard copy number analysis workflow. These developments increase the sensitivity and specificity with which driver genes may be localized and enable data-driven estimation of the background rates of somatic copy number alterations (SCNA) and how these rates vary with features of the SCNA, such as length or amplitude.

For DNA methylation, we also obtained the data from Firehose [2]. We used the results of correlation analysis between mRNA expression and DNA methylation with methylation probes mapped to gene promoters. Then we selected for each gene, data corresponding to the methylation probe showing strongest negative correlation (Pearson correlation coefficient) of methylation beta-value and mRNA expression across all samples in each cancer type.

3.1.3 RBPs and TFs expression

To determine the RBPs and TFs that are expressed in each cancer type, we downloaded the cancer tissue atlas from Human Proteome Atlas [5]. This data contains a multitude of human cancer types representing the 20 most common forms of cancer, including breast-, colon-, prostate-, lung-, urothelial-, skin-, endometrial- and cervical cancer. Altogether 216 different cancer samples are used to generate protein expression profiles for all proteins using immunohistochemistry. For each cancer type, 12 samples are available. RBPs and TFs that show expression in at least one of the 12 samples were considered as expressed in that cancer type.

[1] Firehose analyses__2014_10_17
[2] Firehose analyses__2014_10_17

3.1.4 Clinical data

UCSC Cancer Genomics Browser provides easy to manipulate formats of data available in TCGA. Recently, users are able to download clinical and genomic columns of interest for analysis. The file formats provided by UCSC Cancer Genomic Browser are more easily recognizable by most spreadsheets and advanced analysis tools such as R [34]. We downloaded the clinical data for patients in order to perform survival analysis for the RBPs that we found as candidate regulators.

3.2 Motif prediction of regulatory elements

3.2.1 Transcription factors

In order to map transcription factors (TFs) on mRNAs, we downloaded promoter regions of mRNAs from UCSC genome browser with a 200 fllank on each side. Position frequency matrices of human transcription factors are obtained from JASPAR database (JASPAR CORE 2016) [67]. This database contains a non-redundant and curated set of experimentally defined TF binding site profiles in eukaryotes which are derived from published collections.

We then used individual motif scanning tool FIMO (Find Individual Motif Occurrences) from MEME-SUITE [36] to scan the binding sites of these transcription factors on promoter regions of mRNAs. FIMO is a software tool for scanning DNA or protein sequences with a set of given motifs in the format of position specific scoring matrices. This program computes a log-likelihood ratio score for each match in a given sequence and uses established dynamic programming method to convert this score to a p-value. It then applies false discovery rate analysis to estimate a q-value for each match in the given sequence. We select matches with p-values< 1e-4.

In order to strengthen the support of TF binding sites that we identified with FIMO, we intersected these sites with DNaseI hypersensitive sites (DHSs) for A549 cells. DHSs are the regions where chromatin has lost its condensed structure, where DNA gets more accessible in that region for binding of proteins like TFs. This data is obtained from *The ENCODE Project Consortium (2004)*, the project that aims to map all of DHSs in the human genome to catalog human regulatory DNA. Using massive sequencing techniques, they obtained DHSs of every cellular type.

3.2.2 microRNAs

The proportion of the isoforms of a 3'UTR is specific to a tissue or cell line [98]. This proportion can affect 3'UTR length related features such as the number of binding sites of an RBP or miRNA on 3'UTR. To avoid these complications we used 3'UTR annotations curated by Agarwal et al [2], where they compiled 3'UTR isoform quantifications previously measured for HeLa cells [77] using poly(A)-position profiling by sequencing (3P-seq) [47]. Agarwal et al selected the genes for which $\geqslant 90\%$ of the 3P-seq tags corresponded to a single 3'UTR isoform. They downloaded the 3'UTR

sequences from Gencode version 19 (hg19), and extended the end of the 3'UTRs with information from 3P-seq annotations.

We then downloaded conserved human miRNA targets from TargetScan 7.0 (Agarwal et al) and mapped those targets on 3'UTRs. TargetScan predicts these miRNA targets by searching for 8mer, 7mer and 6mer sites that match the seed region of the miRNA. These predictions are then ranked based on the predicted efficacy of targeting calculated as the cumulative weighted context++ scores of the sites [2]. For human, TargetScan considers matches to human 3'UTR and their orthologs, as defined by UCSC whole-genome alignmnets. TargetScan 7.0 has improved in comparison with previous releases of TargetScan. This new version predicts targeting efficacy more accurately and uses 3'UTR profiles that indicate the fraction of mRNA containing each site as described above, and also uses updated miRNA families.

3.2.3 RBP binding sites

To map the binding sites of RBPs, we downloaded 103 position frequency matrices (PFMs) that correspond to 85 human RBPs from the RNAcompete paper [85]. These PFMs (which are of length seven or eight) are generated from the alignment of top 10 7-mers determined using all data (i.e. both setA and setB of RNAcompete pool). Rather than using these top 10 7-mers directly, we generated the top 10 nmers from the PFMs. In this way, we were able to scan for motifs that are longer than seven. An example is the FXR1 RBP for which the PFM inferred by RNAcompete is of length eight. By using the top 10 8-mers in our motif search, we can represent the binding preferences to all eight positions of this PFM. In addition to RNAcompete motifs, we downloaded the motifs (consensus motifs and the top 10 nmer when a PFM is available) for the following well-known RBPs from RBPDB database [21]: HNRNPAB, PUM1, PUM2, ELAVL2, KHSRP, ZFP36, AUF1 and CUGBP.

We downloaded human 3'UTRs from TargetScan and determined the genome-wide binding sites of each RBP by finding matches to its top 10 n-mers or consensus motifs on human 3'UTR sequences. We downloaded CLIP-seq data for a list of RBPs (HuR, FMR1, FUS, FXR1, FXR2, hnRNPC, IGF2BP1-3, LIN28A, PTBP1, PUM2, QKI, SRSF1, TIA1) from starBase database [60]. starBase is a database that has developed to decipher protein-RNA and miRNA-RNA interactions from 108 CLIP-seq (HITS-CLIP, PAR-CLIP, iCLIP, CLASH) datasets. In addition to these RBP specific CLIP datasets, we downloaded gPARCLIP-determined peaks [8]. gPARCLIP dataset contains genome-wide protein-occupied regions bound by any RBP in HEK293 cells. However, the identity of a particular RBP that binds to a region is unknown.

We first intersected the peaks identified with CLIP or gPARCLIP with human 3'UTRs. To correct for the background binding bias in CLIP-based techniques as identified by Friedersdorf et al. [32], we excluded parts of peaks that overlap with regions that correspond to background binding. Finally we created a RBP feature matrix which contains number of confident (CLIP or gPARCLIP supported) binding sites of each RBP on each mRNA.

RBP and miRNA sites using AIR score: Alternatively, we also tried incorporating alternative polyadenylation (APA) in mapping binding sites of RBPs and miRNAs.

To this end, we downloaded *3P-seq tag info* from TargetScan 7.0, and used a metric called Affected Isoform Ratios (AIRs) that indicates for each RBP/ miRNA binding site the fraction of mRNA transcripts containing that site [77]. We then used the average AIR score of each RBP / miRNA along each mRNA to create another feature matrix. So, rather than counting simply the number of RBP/ miRNA sites on an mRNA, we averaged the AIR values of each site.

3.3 Differentially expression of RBPs in cancer

In order to find RBPs that are differentially expressed in various cancer types, we used datasets of those cancer types having sufficient paired (tumor and its corresponding normal) samples (>15 pairs). According to this criteria, we selected the following 13 cancer types: BLCA, BRCA, COAD, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, THCA and UCEC (3.1)

In order to calculate log fold-change (logFC) (tumor/normal) of genes in each cancer, we used edgeR method from Bioconductor package. This method is used for differential expression analyses of read counts arising from RNA-Seq, SAGE or similar technologies [86]. edgeR uses empirical Bayes method that permits the estimation of gene-specific biological variation, even for experiments with minimal levels of biological replication. The package implements both exact statistical methods for multi-group experiments [87, 88] and generalized linear models (GLMs) which are suitable for multi-factor experiments of any complexity [64, 70].

Table 3.1: This table lists the abbreviations and full names of the analyzed cancer types together with the corresponding number of total and paired samples downloaded from TCGA database.

Cancer type	Description	All samples	Paired samples
BLCA	Bladder urothelial carcinoma	408	19
BRCA	Breast invasive carcinoma	1095	113
COAD	Colon adenocarcinoma	459	41
HNSC	Head and neck squamous cell carcinoma	521	43
KICH	Kidney chromophobe	66	25
KIRC	Kidney renal clear cell carcinoma	533	72
KIRP	Kidney renal papillary cell carcinoma	290	32
LIHC	Liver hepatocellular carcinoma	371	50
LUAD	Lung adenocarcinoma	516	58
LUSC	Lung squamous cell carcinoma	501	51
PRAD	Prostate adenocarcinoma	497	52
THCA	Thyroid carcinoma	505	59
UCEC	Uterine corpus endometrial carcinoma	545	23

Using edgeR, we performed pairwise comparison of matched tumor and normal sam-

ples to calculate the logFC of genes within each cancer type. In particular, we determined differential expression using the generalized linear model likelihood ratio test (using *glmFit* and *glmLRT* functions). We defined differentially expressed RBPs as those genes with FDR cutoff < 0.05 and $|logFC| > 0.5$. We clustered the logFCs of profiles of RBPs across the cancers using *heatmap.2* function in R (hierarchical clustering performed with Euclidean distance and complete linkage). Differentially expressed miRNAs are found using edgeR similarly.

3.4 Regression analysis

Our goal was to computationally model the regulatory network of gene expression in cancer to identify the predominant regulators. Our dependent variable is the expression of the gene, and our independent variables are the factors that affect gene expression. So we developed a Generalized Linear Model which incorporates copy number variation, DNA Methylation and the regulatory effects of transcription factors, miRNAs and RBPs to predict gene expression in cancer and estimate sample specific TF, miRNA and RBP activities (Figure 3.1).

In order to avoid over-fitting of the model due to the large number of features we included a LASSO penalty term [99]. The LASSO constraint enforces most of the regression coefficients to be zero to reduces the number of features included in the model, leading to higher prediction accuracy and more interpretable results. By this way we remained only with small number of expressed TFs, miRNAs and RBPs that best explains the global changes in expression.

$$y_g = w_0 + w_C C_g + w_M M_g + \sum_{TF} w_{TF} N_g^{TF} + \sum_{miR} w_{miR} N_g^{miR} + \sum_{RBP} w_{RBP} N_g^{RBP} \quad (3.1)$$

where:

- y_g is the expression of gene g,
- C_g is the CNV of gene g,
- M_g is the gene's methylation level
- N_g^{TF} is the counts of binding sites of TFs gene g
- N_g^{miR} is the counts of binding sites of miRNAs gene g
- N_g^{RBP} is the counts of binding sites of RBPs for gene g

This regression was performed in R using *glmnet* with $\alpha = 1$ except that the best λ was chosen using cross-validation function cv.glmnet.

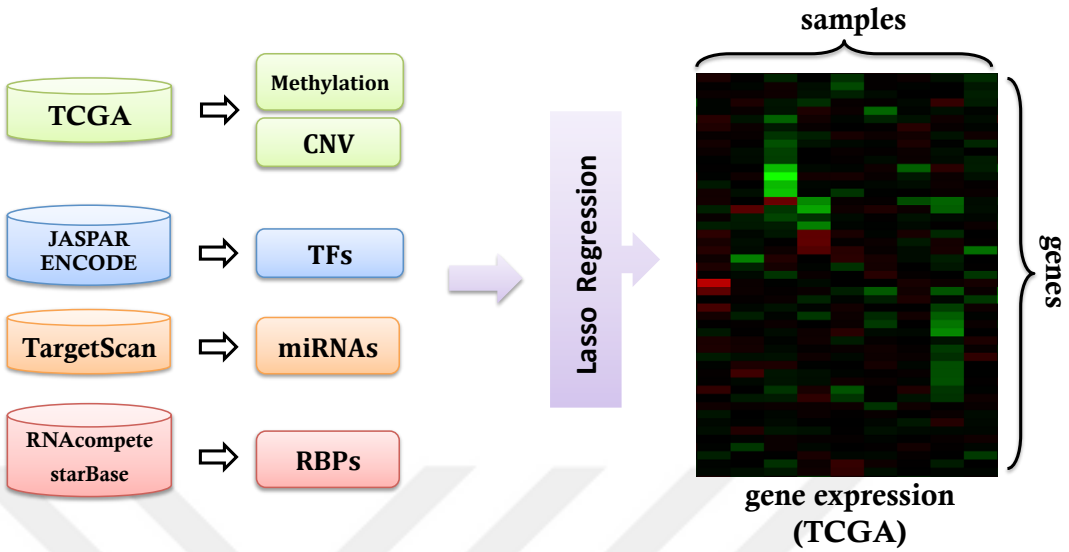


Figure 3.1: DNA methylation, copy number variation and regulatory effects of transcription factors, miRNAs and RBPs are input to a lasso regularized regression model to predict gene expression in LUSC.

3.5 Feature selection analysis

In order to determine candidate regulators in LUSC, we performed a feature selection procedure specifically developed for lasso-regularized regression models. We downloaded the Selective Inference package from R and used the *fixedLassoInf* function to calculate selective p-values for a given lambda (i.e. regularization constant) value. Here, we used the lambda that is selected with *cv.glmnet* function. We repeated this procedure for each sample independently, and calculated, for each feature, the number of times a significant p-value (p-val < 0.05) is obtained.

3.6 Identification of target gene sets of candidate regulators

We summed the changes in prediction error of each gene across the samples when a regulator is removed. To estimate the significance of the error changes, we repeated this calculation with shuffled feature matrices 2000 times [93]. The shuffling was done for each column independently. We calculated an empirical p-value (for a gene-regulator pair) by comparing the error change obtained from the original feature matrix with the distribution of error changes that are obtained from shuffled feature matrices. The target gene set of a regulator is defined as the genes with FDR-corrected p-value $<= 5e-4$. We compared the predicted target genes of RBPs against CLIP-based targets. To evaluate the predicted targets of miRNAs, we downloaded experimentally verified targets from MirTarBase database [43]. We grouped the experimentally verified targets based on the type of evidence. Namely, targets identified with reporter assay, western blot and qPCR provide strong evidence; whereas targets identified with approaches such as microarray, NGS-based methods, pSILAC provide less strong or weak evidence.

3.7 Random Forest

We also tried to model the regulatory network of gene expression in cancer using the popular ensemble learning method, random forest. Random forest is a collection of decision trees. In contrast to single decision trees which are likely to suffer from high bias based on how they are tuned, random forest outputs mean prediction of the individual trees which corrects for overfitting the training set [14].

For this purpose, we used the R package "RandomForest". Although higher number of trees gives more accurate predictions, it also significantly increases the computational cost. A study by Oshiro et al investigates whether there is an optimal number of trees within a random forest, i.e., a threshold after which increasing the number of trees would bring no significant performance gain. They suggest a range between 64 and 128 trees in a forest [80]. Taking this into account, in order to decide for the best number of trees, we looked at the out of bag error rate (MSE in regression) of the model in range of 1 to 150 trees and we chose the number of trees where the error does not change significantly (Figure 3.2). The other parameter that can be tuned is the number of variables randomly sampled as candidates at each split, *mtry*. We used the RandomForest package's function *tuneRF* to decide for the best mtry. This function starts with the default value of mtry (number of total features/3) and searches for the optimal value (with respect to Out-of-Bag error estimate) of mtry for random forest. Finally by selecting 100 as number of trees and 64 as mrty, we ran the model sample by sample for each of 362 samples.

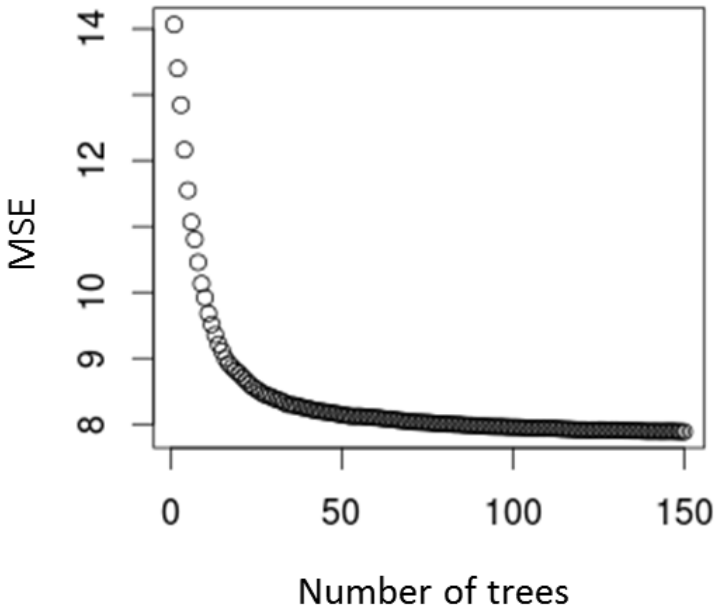


Figure 3.2: This figure shows the out of bag error rate (MSE in regression) of the model in range of 1 to 150 trees.

Then we calculated the correlation between predicted and observed gene expression and compared these correlations with correlations we obtained from regression. Also, we looked for candidate regulators of each sample using the function *importance*. This function extracts the feature importance measures as produced by randomForest

method. We selected the top 10 most important features as candidate regulators for that sample. We then chose those sample specific candidates that are found in more than 50 percent of all samples as LUSC candidate regulators.

3.8 Kaplan-Meier survival analysis

We performed Kaplan-Meier survival analysis using the candidate regulators we obtained from previous step. We downloaded clinical data from UCSC cancer genomics browser (which is based on TCGA) and we used R Bioconductor's *survival* package and *Survfit* function. We investigated the associations of patient survival time with parameter values that represent the activities of regulators and with corresponding mRNA expression profiles of the regulators.

The optimal cut-off level for expression or parameter value was chosen by the "Maximally selected rank statistics", *MaxStat* package in R. This package is commonly used in medical statistics to determine a simple cut-point of predictors between two groups pf observations.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Differentially expressed RBPs in cancer

Cancer is commonly characterized by the differential expression of several master regulators. In particular, aberrant expression of RBPs have been found to be associated with cancer initiation and progression [105]. To investigate the expression changes of RBPs in cancer systematically, we downloaded matched tumor-normal samples for 13 cancer types (see 3.1 for a summary of these datasets). We used edgeR to identify differentially expressed genes across the matched samples for each cancer. Fig. 4.1 shows the log fold changes (LFCs) of RBPs that are differentially expressed in at least one of these cancer types (LFC > 0.5 or LFC < -0.5, FDR-corrected p-value < 0.05). We observed that a number of well-known RBPs (e.g. PTBP1, KHSRP, ELAVL1, PABPC1, PABPC3, HNRNPAB) display increased expression in majority of cancer types. Among these RBPs, ELAVL1 has been previously found to have elevated levels of expression in cancer [55]. On the other hand, RBPs such as CPEB4 and A2B1 show decreased expression across the majority of cancer types. Interestingly, FXR1 is found to be overexpressed most in LUSC compared to the other cancer types. Indeed, a recent study revealed FXR1 as a driver for non-small cell lung cancer (NSCLC), and showed that increased FXR1 promotes tumor progression, and is associated with poor survival [83]. Lastly, we observed that IGF2BP2 and IGF2BP3 display strong up- or down-regulation of expression among the different cancer types. In particular, both IGF2BP2 and IGF2BP3 are overexpressed significantly in LUSC. IGF2BP proteins are known to be expressed mainly in the embryo; however, they have been found to be re-expressed in several cancer types including lung cancer [10].

4.2 Predicting gene expression in LUSC

The amount of gene expression in different cellular conditions is due to copy number variations (CNV), DNA methylation and activities of distinct regulators in transcriptional and post-transcriptional level. The combined effect of these elements and transcription factors at transcriptional level and microRNAs at post-transcriptional level in cancer have been studied previously in different ways [46, 61, 93]. The goal of our study is to dissect the effect of RBPs on top of the other factors that govern mRNA abundance in cancer. To this aim we developed a Lasso-regularized regression model

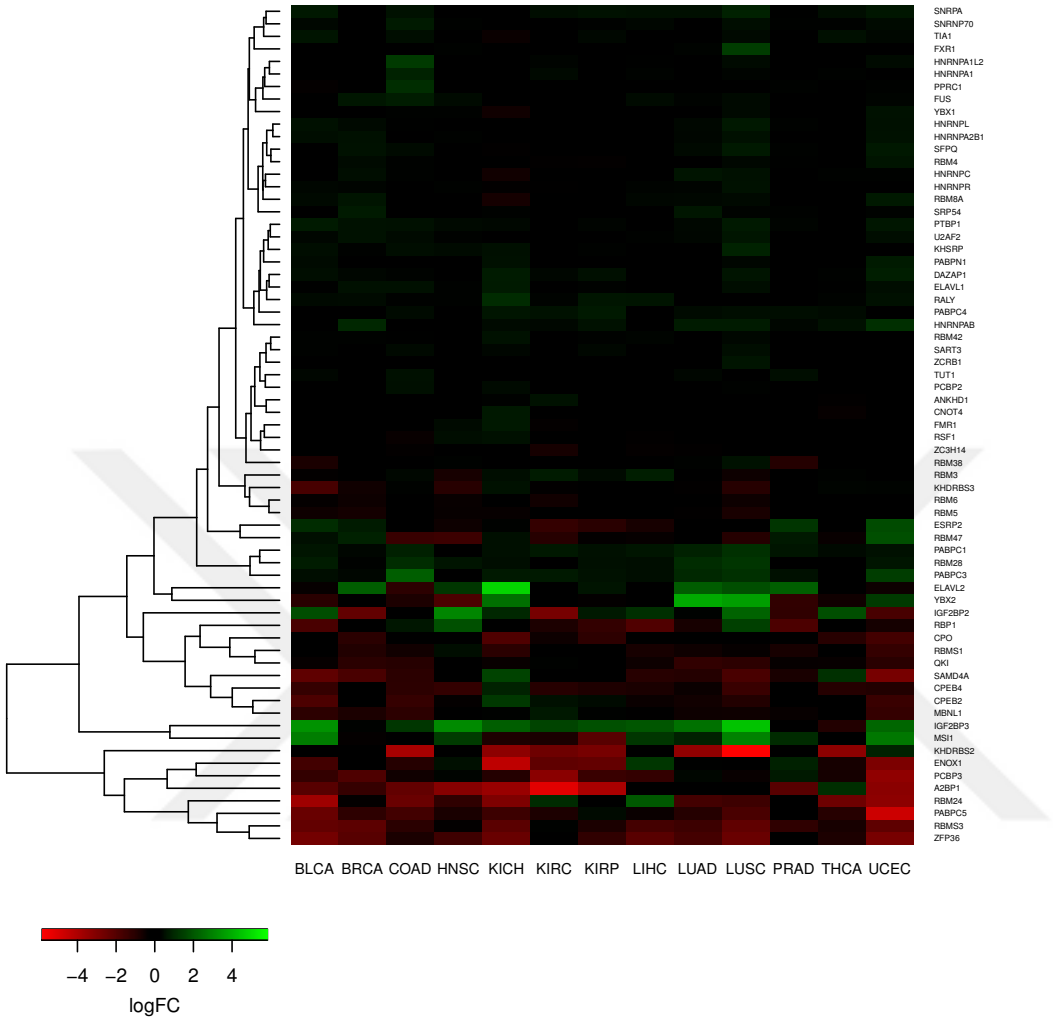


Figure 4.1: This heatmap shows the log fold expression changes of RBPs across matched tumor-normal samples (calculated with edgeR [86]). Rows are the RBPs that are differentially expressed in at least one cancer type. Columns correspond to different cancer types. Rows are clustered with hierarchical clustering.

which incorporates copy number variation, DNA Methylation and the regulatory effects of transcription factors, miRNAs and RBPs to predict gene expression. The RBP features here is defines as their *Number of binding sites* on each gene.

Having found that many RBPs are differentially expressed in LUSC, we set out to investigate the regulatory effects of RBPs in this cancer type in more detail, 4.2. Also we were able to collect different dataset (CNV, methylation, mRNA and miRNA expression) with high number of samples for this cancer type.

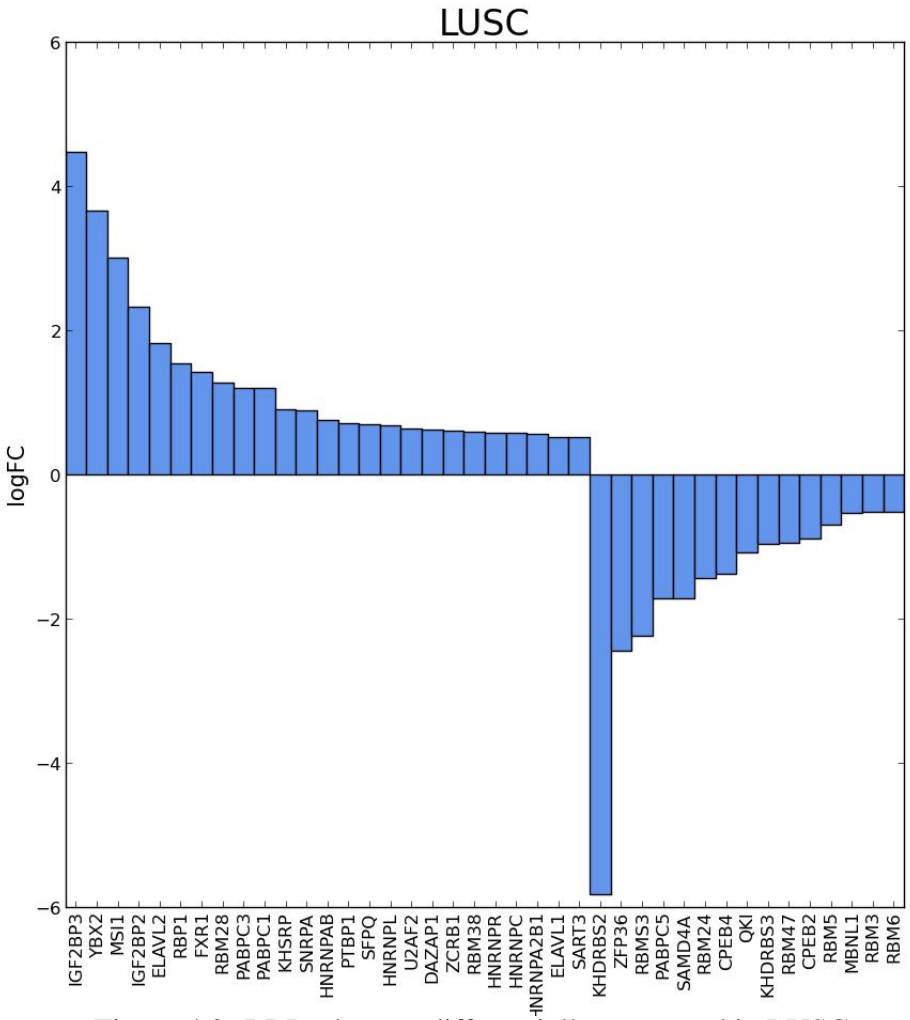


Figure 4.2: RBPs that are differentially expressed in LUSC

4.3 Performance evaluation

To evaluate the performance of the model, we fit a regression model for each sample separately, and performed 10-fold CV. For each CV run, we calculated the Spearman rank correlation between predicted and observed expression of genes in the held-out set. We then averaged these correlation values across the CV folds, and then across the samples. When we used all the features described above, we obtained a Spearman rank correlation of 0.39. To determine the predictive value of features, we compared the full model with partial models that exclude one of the regulatory classes. Fig. 4.3 shows how average Spearman rank correlation changes when one type of regulatory class (i.e., CNV, DM, TFs, miRNAs and RBPs) is removed from the model. This comparison revealed that RBPs show the greatest added predictive value (14% reduction when omitted) followed by TFs (10% reduction). miRNAs (5% reduction), DNA methylation (5% reduction) and CNV (3% reduction) contribute relatively less to the predictive performance (4.1). The strong association between TFs and gene expression have been previously observed several times, whereas the

remarkably high effect of RBP-mediated regulation in explaining gene expression in cancer is a novel result.

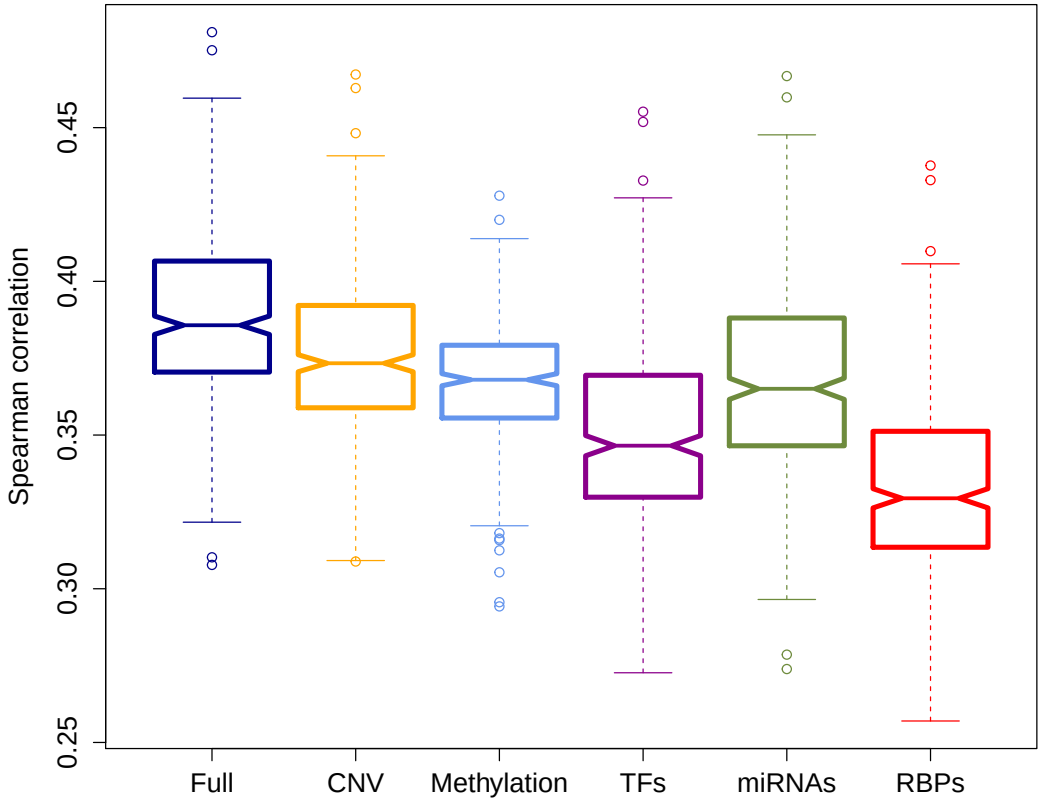


Figure 4.3: This box plot displays the Spearman rank correlation between the predicted and the actual held-out genes in 10-fold cross- validation (CV) averaged across all the samples. The full model that uses all features was compared with the partial models that lack one of the regulator groups: CNV (copy number variation), DM (DNA methylation), TFs, miRNAs, RBPs.

Table 4.1: Comparison of models using glmnet: This table displays the Spearman correlation coefficient of the full model and the five partial models where one feature group is removed. P-values indicate the significance of the difference between models (Wilcoxon sign-rank test).

	Spearman correlation	p-value (compared to Full model)
Full	0.38	N/A
CNV excluded	0.37	8.15e-61
Methylation excluded	0.36	5.12e-61
TFs excluded	0.35	4.45e-61
miRNAs excluded	0.36	4.45e-61
RBPs excluded	0.33	4.45e-61

We also tested whether different methods to define RBP binding sites change the predictive performance. We used the methods below to create our RBP feature matrix:

- All sites by scanning 3'UTR sequences with top 10 kmers without intersecting with CLIP or gPARCLIP-determined peaks
- CLIP supported or gPAR-CLIP supported sites by scanning 3'UTR sequences with top 10 kmers
- All sites by scanning 3'UTR sequences with PFMs obtained from RNAcompete using FIMO from MEME Suite
- All sites calculated by 3P-seq AIR scores (see section 3.2.3 from Materials and Methods)

Table 4.2 shows the correlations obtained using each of these methods. When we compiled the features of RBPs and miRNAs by taking alternative polyadenylation (APA) into account (Materials and Methods) we obtained a Spearman correlation of 0.34. This surprising result might be due to the fact that APA has been measured in other cell types than lung cells. Also, more complex models might need to be developed to incorporate the measurements of abundance of alternative 3'UTR isoforms to the definition of binding sites. We also tried counting the RBPs sites without intersecting with CLIP or gPARCLIP-determined peaks. This decreased the performance slightly (Spearman correlation: 0.37) indicating that CLIP or gPARCLIP-determined binding sites are likely to be more functional.

Table 4.2: This table shows that using RBP sites that are CLIP or gPARCLIP supported, gives us more accurate predictions

Definition of RBP sites	Spearman correlation
All top 10 kmer sites	0.37
Confident top 10 kmer sites	0.39
FIMO scanned sites	0.34
Average AIR scores	0.34

We also fit a random forest model to confirm the predictive value of RBPs. We used the same features as in the regression model and learned a random forest model for each sample independently. As before, we calculated the Spearman correlation between predicted and observed expression of genes in the held-out set. We then averaged these values across all the samples. As we can see in table 4.3, random forest gives very similar results to the regression model. Importantly, we also confirm with random forest that RBPs increase the performance significantly.

4.3.1 Candidate regulators of LUSC

Having found that regulatory factors can explain a significant portion of gene expression in LUSC, we used a feature selection procedure to determine the predominant regulators. Although Lasso selects important features by penalizing unimportant

Table 4.3: Comparison of model using ransom forest: This table displays the Spearman correlation coefficient of the full model and the five partial models where one feature group is removed. P-values indicate the significance of the difference between models (Wilcoxon sign-rank test).

	Spearman correlation	p-value (compared to Full model)
Full	0.40	N/A
CNV excluded	0.39	1.50e-58
Methylation excluded	0.38	4.45e-61
TFs excluded	0.37	4.45e-61
miRNAs excluded	0.39	1.45e-49
RBPs excluded	0.35	4.45e-61

ones, the average number of features having nonzero coefficient across all samples in our model was 178 which was quiet high. Another common approach to test the importance of an additional feature between two nested linear models is to compare the change in error to a chi-square distribution or F-distribution. However, this approach becomes invalid when the additional feature is chosen adaptively as in lasso regularized regression [63]. Despite this fact, F-test has been previously used for lasso [61]. Here, we improve over previous studies by applying a significance test (i.e., covariance test) that accounts for adaptivity [97]. We applied the covariance test for each sample independently, and counted the number of times a significant p-value (p-value < 0.05) is obtained for each feature. When we sorted the features based on this, DNA methylation is ranked the first as it was selected as a significant predictor in 44% of all samples. Copy-number variation is ranked fourth (%37 of all samples). Table 4.4 shows this ranking for all the regulatory factors that are selected in more than 20% (i.e., 70 samples) of the samples. In addition to the name and type of the regulator, the log fold change and the associated FDR-corrected p-value are also displayed if the regulator is found to be differentially expressed. This table reveals that many of the candidate regulators are indeed differentially expressed between cancer and normal samples.

We see that RBPs are ranked on top of this list. For instance, SFPQ, which is selected as a significant regulator in the largest number of samples, has been recently found to interact with a long non-coding RNA called MALAT1 (metastasis-associated lung adenocarcinoma transcript 1) [49]. MALAT1 is overexpressed in several human cancers including non-small cell lung cancer, and has been identified as a critical regulator of metastasis in lung cancer cells [48, 49]. ELAVL1, which ranks third in our list of candidate regulators, has key functions in mRNA stability and translation. In fact, cytoplasmic ELAVL1 expression has been previously found to be associated with high tumor grade and poor survival rate in non-small cell lung carcinoma [104]. Indeed, we found that ELAVL1 is upregulated in LUSC (LFC=0.52). YY1 (Ying Yang 1) is the top ranking regulator among the TFs (LFC = 0.76). YY1 is highly expressed in various cancer types, and its depletion inhibited tumor formation of breast cancer cells [16, 101]. We found that miR-1 is the top ranking miRNA regulator. A recent study revealed that miR-1 was significantly reduced in lung squamous cell carcinoma, and its restoration significantly reduced cancer cell progression [66]. The second

ranking miRNA, miR-218, is significantly down regulated in lung squamous cell carcinoma and has been identified as candidate tumor suppressor [22]. The LFCs that are calculated with our differential expression analysis are in agreement with these studies (LFCs -3.44 and -2.07 for miR-1 and miR-218 respectively). Lastly, though located on the lower part of the list, FXR1 is one of the identified candidate regulators. A previous study has identified FXR1 as a key regulator of tumor progression and found that its overexpression is critical for nonsmall cell lung cancer (NSCLC) cell growth [83]. Similarly, we found that FXR1 is significantly upregulated in LUSC (LFC is 1.42). Altogether, the high correspondence between our predicted candidate regulators and previous literature indicates that our model is accurate in inferring the key regulators of LUSC.

Table 4.4: Candidate regulators of LUSC using glmnet feature selection

Regulator	Type	Selection %	logFC	p.value
DNA methylation	-	44	-	-
SFPQ	RBP	38	0.7	2.06e-12
LIN28A	RBP	38	-	-
Copy Number Variation	-	37	-	-
ELAVL1	RBP	32	0.52	2.10e-12
CPEB4	RBP	31	-1.37	1.21e-26
miR-1	miRNA	31	-3.44	2.13e-36
miR-218	miRNA	30	-2.07	5.93e-29
YY1	TF	30	0.76	6.56e-25
ZC3H14	RBP	29	-0.21	1.5e-2
PABPN1	RBP	29	0.4	8.27e-7
HNRNPC	RBP	29	0.57	8.18e-11
REST	TF	25	0.32	1.1e-2
ETV6	TF	25	0.45	3e-2
miR-142	miRNA	25	0.44	2.8e-2
PCBP2	RBP	25	0.32	7e-2
HNRNPH2	RBP	24	-0.18	4.9e-2
GCM2	TF	24	-	-
miR-145	miRNA	23	-1.43	1.46E-20
PUM2	RBP	23	0.4	5.10e-7
miR-29	miRNA	22	-	-
miR-15	miRNA	22	-0.70	1.39e-13
MAFF	TF	21	-1.1	4.16e-11
miR-140	miRNA	21	-1.01	5.02e-16
miR-381	miRNA	21	-0.49	3.6e-2
RBM6	RBP	21	-0.52	7.48e-11

As we did in previous section, we also found candidate regulators using random forest. For each sample, we ranked the importance value of regulators calculated with the *importance* method of randomForest R package. Then, we averaged the ranks of

each regulator over all samples and sorted the regulators based on their average ranks. Top 60 candidate features obtained from random forest model is in Table 4.5. We intersected these regulators with the candidate regulators we got from glmnet feature selection procedure. We found that the following 9 RBPs are common: ELAVL1, LIN28A, CPEB4, SFPQ, ZC3H14, HNRNPC, PABPN1, PCBP2, PUM2.

Table 4.5: Candidate regulators of LUSC using Random Forest importance method

Regulator	Importance	Regulator	Importance
Met	1	SNRNP70	30.58
CNV	2.04	RBM6	30.89
ELAVL1	2.96	G3BP2	31.49
PTBP1	4.95	SNRPA	33.81
CPEB4	5.09	E2F4	34.38
ZC3H14	6.4	SP4	36.66
HNRNPC	6.6	RBMS1	37.23
RALY	8.71	SP1	38.13
PCBP2	9.44	ZNF638	38.15
LIN28A	10.57	FMR1	38.28
PABPC4	11.48	RBM4	38.56
SART3	12.58	RBM28	40.92
PABPC1	12.58	SP2	41.44
HNRNPL	13.77	DAZAP1	41.61
KHDRBS1	14.11	HNRNPA1	43.94
HNRNPH2	14.8	HNRNPA2B1	43.94
RBM5	14.94	ELK4	46.06
SRSF9	15.35	FXR2	46.62
PABPN1	17.1	SRSF2	47.02
ESRP2	19.07	TARDBP	49.31
IGF2BP2	20.07	NRF1	49.6
IGF2BP3	21.33	RBM45	50.91
PUM2	21.36	KLF5	51.16
YBX1	23.65	REST	52.31
HNRNPK	23.73	KLF14	52.86
SFPQ	24.42	SP3	53.73
CNOT4	26.32	ZIC4	55.02
PCBP1	27.04	KHDRBS3	55.72
MATR3	29.8	EGR1	56.48
MSI1	30.41	FXR1	56.64

4.3.2 Target analysis of candidate regulators

The input feature matrix that we compiled by counting the number of binding sites of each regulator in each gene provides a noisy approximation of functional targets of regulators. To identify the targets of the regulators from our model robustly, we identified the genes for which the squared prediction error increases when a regulator is removed. We determined the significance of an increase in error by comparing it against a distribution of error changes that are obtained when the feature matrix is randomized (see section 3.6 from Materials and Methods).

We evaluated our predicted target gene sets by comparing against experimentally verified interactions, when available. For RBPs, our validation set consists of the genes that are identified by CLIP experiment. As such, we could evaluate the target sets of RBPs with CLIP data: LIN28A, ELAVL1, HNRNPC, PUM2 and IGF2BP2. We evaluated the target predictions for our top ranking miRNAs miR-1 and miR-218, by compiling experimentally verified targets (either with strong evidence or weak evidence) from MirTarBase database. Fig.4.4 shows the number of genes that are shared between the set of our predicted targets and the set of experimentally verified targets, for RBPs and miRNAs. We see a high overlap between the two sets for RBPs. In particular, almost 30% of the predicted target genes for ELAVL1 are also CLIP targets. The intersection is much smaller for miRNAs. A similar result has been previously obtained when miRNA target prediction methods were compared based on the number of validated targets in miRTarBase [61].

Next, we utilized a previously published ELAVL1 knockdown dataset that includes genome-wide measurements of transcripts upon ELAVL1 depletion in HEK293 cells [75]. In Fig. 4.5 we plotted the cumulative distribution of transcripts in two groups:

- Predicted target gene set
- CLIP-based target gene set

This analysis revealed that transcripts in predicted target gene set are more destabilized upon ELAVL1 knockdown. We observed that transcripts in the first group are significantly more destabilized upon ELAVL1 depletion than transcripts in second group. As ELAVL1 is known to stabilize its targets, these results show that ELAVL1 targets predicted by our model show greater effect than targets having CLIP-based ELAVL1 sites. This result indicates the accuracy of our model in identifying the functional targets of ELAVL1.

4.3.3 Survival results

Next, we assessed whether the candidate regulators that we identified with our statistical model are predictive of survival time in LUSC. We performed Kaplan-Meier survival analysis (see section 3.7 from Materials and Methods) for the top ranking RBP regulators shown in Table4.4. Using clinical data from TCGA, we looked at the associations of patient survival time with parameter values that represent the activities of regulators and with corresponding mRNA expression profiles of the regulators.

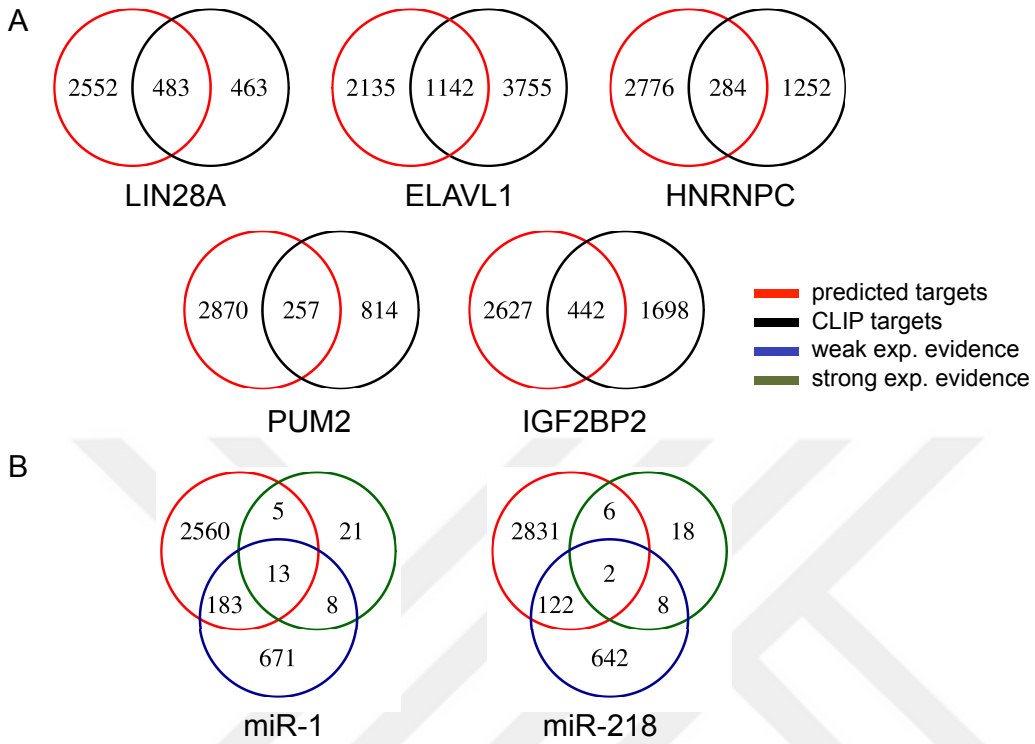


Figure 4.4: A. Predicted target gene sets of RBPs are intersected with targets determined with CLIP method. B. Predicted targets of miRNAs are intersected with experimentally verified (with either weak or strong evidence) miRNA targets downloaded from miRTarBase.

We found that the top ranking candidate regulators, ELAVL1 and SFPQ, are predictive of survival rate when we stratified the patients based on expression level (Fig.4.6a and Fig.4.7a). We confirmed the same finding when we grouped patients based on parameter values learned with our model (Fig.4.6b and Fig.4.7b). For SFPQ, stratification based on parameter values gives a much significant difference.

For both RBPs (ELAVL1 and SFPQ), patients with high expression levels or high parameter values showed a trend toward better survival indicating that the activities of these RBPs are positively correlated with survival rate. We repeated the same analysis for miR-1 and miR-218, and found that miR-1 but not miR-218 activity is associated with survival rate. Both the expression levels and the fitted parameter values indicate that low miR-1 activity (i.e., low expression or high model model parameters) is correlated with survival (Fig.4.8a and b).

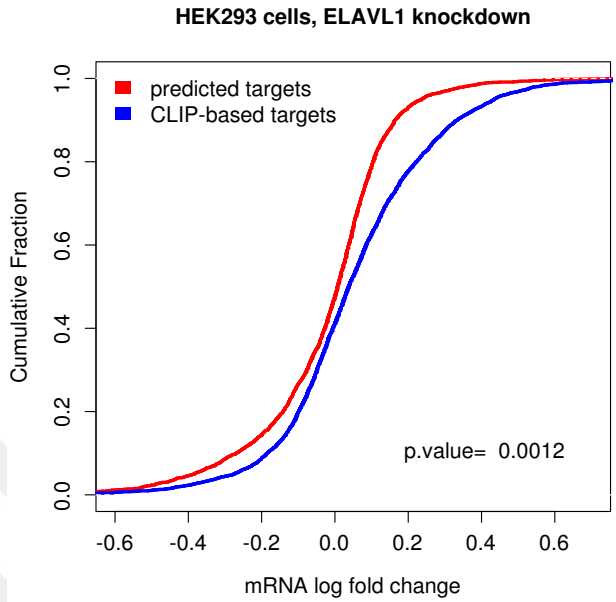
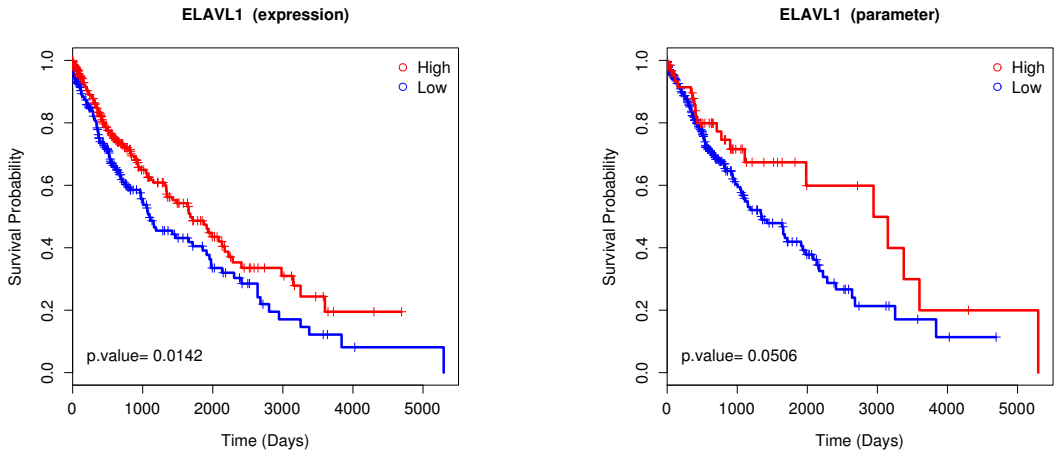
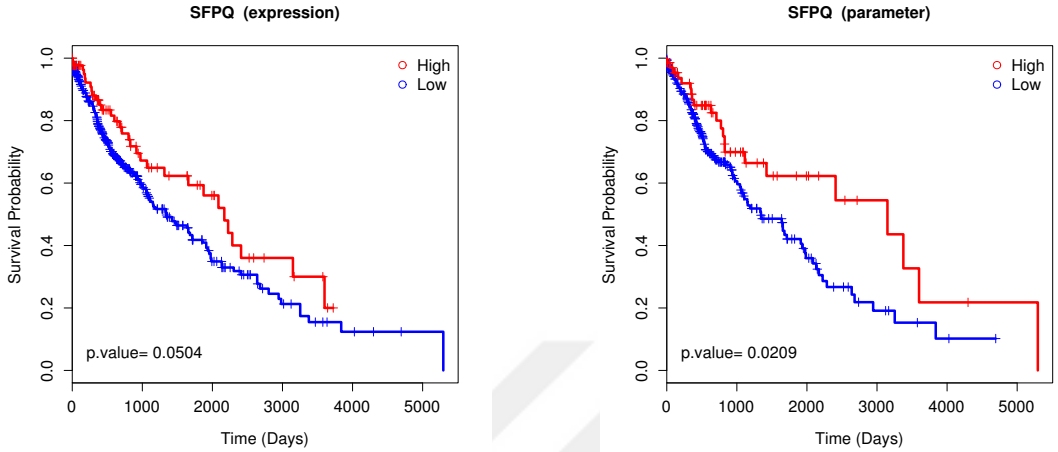


Figure 4.5: The distribution of LFCs of predicted targets is compared with the distribution of LFCs of CLIP-based targets. Predicted targets display increased destabilization which indicates that they are likely to be functional ELAVL1 targets. The difference is significant according to Mann Whitney U test (p-value = 0.0012).



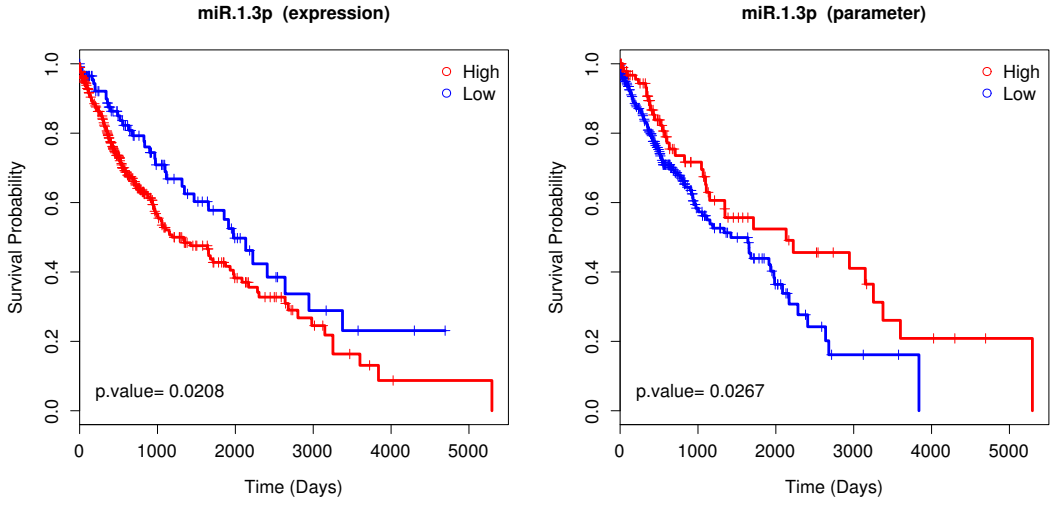
(a) (b)

Figure 4.6: Expression and activity of ELAVL1 are predictive of survival rate in LUSC. a. Patients with high ELAVL1 expression show a significantly higher median survival time. b. Stratification based on model parameters for ELAVL1 confirms the same finding as in a.



(a) (b)

Figure 4.7: Expression and activity of SFPQ are predictive of survival rate in LUSC. a. Patients with high SFPQ expression show a significantly higher median survival time. b. Stratification based on model parameters for SFPQ confirms the same finding as in a.



(a) (b)

Figure 4.8: Expression and activity of miR-1 are predictive of survival rate in LUSC. Patients with low miR-1 activity (i.e., low expression (a) or high model model parameters(b)) show a significantly higher median survival time.

CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

5.1 Conclusion

In this study, we investigated the mechanisms that account for gene expression regulation in LUSC. We initially assessed the alterations in expression of genes that encode for RBPs across a number of cancer types. To our knowledge, this is the first time differentially expressed RBPs are searched using a method that accounts for matched samples across several cancer types (i.e. edgeR). The results of this analysis revealed that several RBPs are differentially expressed with district profiles of up- or down-regulation across the cancers. Having found that the number of differentially expressed RBPs is largest in LUSC, we developed a lasso-regularized regression model to predict gene expression in LUSC by incorporating several features including the regulation mediated by RBPs. We were able to accurately predict the expression of genes in held-out sets by incorporating a comprehensive set of regulatory elements that are bound by TFs, miRNAs and RBPs, as well as genetic and epigenetic alterations as features in our statistical model. Importantly, compared to other regulatory classes, exclusion of RBPs results in the largest decrease in predictive performance revealing the influence of RBP-mediated regulation. This is one of the key novel observations of the current study that indicates the importance of RBPs in regulation gene expression in LUSC.

Next, we identified key regulators of LUSC by calculating the significance of each feature using a recently proposed statistical test that accounts for the adaptive nature of fitting lasso models. Inference of statistically significant features in adaptive models is an active research area in statistics, and we believe that our study will be instrumental in dissemination of this recent result to bioinformatics community. We found that the majority of the top ranking candidate regulators are differentially expressed in LUSC, and have been previously identified to be associated with lung cancer. We have also identified additional regulators such as LIN28A and CPEBP4 that were not previously studied in the context of lung cancer. Also, the fact that many of the candidate regulators are RBPs agrees with our previous result on the added predictive value of RBPs. In fact, two RBPs, ELAVL1 and SFPQ have been found to be associated with survival rate in LUSC patients.

Apart from the results on RBP regulation, our study is also amongst the first to incorporate the recently released JASPAR and TargetScan databases in predicting TF and miRNA binding sites, respectively. Identification of TF and miRNA target sites

can become more accurate with the availability of ChiP-Seq and CLIP-seq datasets in lung cells. Similarly, CLIP experiments have been performed for a small number of RBPs, and increase in the number of such experiments would improve the definition of RBP target sets.

Lung cancer is one of the most difficult cancers to treat. Recently developed molecular therapies can be targeted to adenocarcinoma of the lung [65]. Such a treatment has not been proposed for squamous cell carcinoma yet. Therefore, identification of novel therapeutic agents is vital for this cancer type. Here, we applied our novel statistical model to infer gene regulatory mechanisms in LUSC, and identified a number of candidate regulators including RBPs. Further studies of these candidate regulators will provide insights into the molecular mechanisms of cancer development in LUSC.

5.2 Future directions

In our model, we made a simplifying assumption that TFs, RBPs and miRNAs can bind to mRNA independently. However, multiple TFs can bind to the same promoter in a competitive or collaborative fashion. Similarly, recent studies show that RBPs and miRNAs can act in competition or collaboration with each other [42]. Increased knowledge on these interactions will be instrumental in developing more accurate models of regulatory networks in the future.

In this study we limited our model to LUSC. One future step is to apply our model to other cancer types to identify the candidate regulators. In this way, we can also compare the candidate regulators across different cancer types.

Furthermore, RNA secondary structure, which is an important factor for target recognition of some RBPs has been ignored in the current study. RNA secondary structure can be considered in the identification of RBP binding sites as more RBPs have characterized secondary structure preferences. Also, recent advances in experimental techniques to query secondary structure *in vivo* [90, 95] promise to generate a more accurate set of mRNA secondary structures compared to the computational prediction methods.

Currently, the knowledge of binding preferences of factors on mRNAs are limited and there are limited number of experimentally validated binding sites for TFs, RBPs and miRNAs. As more experiments are performed to find binding sites of these regulatory factors, our features will become more accurate.

Also, there are limited datasets that measure genome-wide effect of factors upon their depletion or transfection. We used ELAVL1 knockdown dataset to investigate the accuracy of our model in identifying the functional targets of ELAVL1 since it is a well characterized RBP which its effect on mRNA expression is known. But as the effect of many other factors on mRNA expression is not very well known, we cannot reason based on them. In the future, we can repeat our analyses using knockdown or transfection dataset of other RBPs.

Finally, we fit regression models to each sample independently. A possible future direction is to use group-Lasso model with all the samples to identify subtypes of

cancer.

REFERENCES

[1] K. Abdelmohsen and M. Gorospe. Posttranscriptional regulation of cancer traits by HuR. *Wiley Interdisciplinary Reviews*, 1(2):214–229, 2011.

[2] V. Agarwal, G. W. Bell, J. W. Nam, and D. P. Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4:e05005, 2015.

[3] O. Anczukow, A. Z. Rosenberg, M. Akerman, S. Das, L. Zhan, R. Karni, S. K. Muthuswamy, and A. R. Krainer. The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. *Nature Structural and Molecular Biology*, 19:220–228, 2012.

[4] F. Antequera, J. Boyes, and A. Bird. High levels of de novomethylation and altered chromatin structure at CpG islands in cell lines. *Cell*, 62:503–514, 1990.

[5] H. P. Atlas. Human Protein Atlas. `http://www.proteinatlas.org`, 2014. Accessed: 2014-12-30.

[6] T. C. G. Atlas. The Cancer Genome Atlas Research Network (2011) TCGA data portal. `https://tcga-data.nci.nih.gov/tcga/`, 2011. Accessed: 2010-09-30.

[7] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34:W369–W373, 2006. [PubMed:16845028] [PubMed Central:PMC1538909] [doi:10.1093/nar/gkl198].

[8] A. G. Baltz, M. Munschauer, B. Schwanhäusser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, and M. Landthaler. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell*, 46:674–690, 2012.

[9] D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.

[10] J. L. Bell, K. Wachter, B. Muhleck, N. Pazaitis, M. Kohn, and M. Lederer. Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell Mol Life Sci*, 70:2657–75, 2013.

[11] J. Bergalet, M. Fawal, and C. Lopez. HuR-Mediated control of C/EBPbeta mRNA stability and translation in ALK-Positive anaplastic large cell lymphomas. *Molecular Cancer Research*, 9(4):485–496, 2011.

[12] P. Bielli, R. Busa, M. P. Paronetto, and C. Sette. The RNA-binding protein Sam68 is a multifunctional player in human cancer. *Endocrine-Related Cancers*, 18(4):R91–R102, 2011.

[13] F. Bolognani, A. I. Gallani, and L. Sokol. mRNA stability alterations mediated by HuR are necessary to sustain the fast growth of glioma cells. *Journal of Neuro-Oncology*, 106(3):531–542, 2012.

[14] L. Breiman. Random Forests. *Mach. Learn.*, 45(1):5–23, 2001.

[15] R. Busa, M. P. Paronetto, and D. Farini. The RNA-binding protein Sam68 contributes to proliferation and survival of human prostate cancer cells. *Oncogene*, 26(30):4372–82, 2007.

[16] E. Castellano G., Torrisi, G. Ligresti, M. G., L. Militello, A. Russo, J. McCubreyi, S. Canevari, and M. Libra. The involvement of the transcription factor yin yang 1 in cancer development and progression. *Cell Cycle*, 8(9):1367–1372, 2009.

[17] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, N. E. Davay, D. T. Humphreys, T. Preiss, L. M. Steinmetz, J. Krijgsveld, and M. W. Hentze. Insights into RNA Biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 149(6):1393–1406, 2012.

[18] J. Christiansen, A. M. Kolte, T. Hansen, and F. C. Nielsen. IGF2 mRNA-binding protein 2: biological function and putative role in type 2 diabetes. *J Mol Endocrrinol*, 43:187–195, 2009.

[19] S. J. Clark, J. Harrison, C. L. Paul, and M. Frommer. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.*, 22:2990–97, 1994.

[20] L. J. Coleman, M. B. Peter, and T. J. Teall. Combined analysis of eIF4E and 4E-binding protein expression predicts breast cancer survival and estimates eIF4E activity. *British Journal of Cancer*, 100(9):1393–99, 2009.

[21] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res*, 39:D301–D308, 2011.

[22] M. R. Davidson, J. E. Larsen, I. A. Yang, N. K. Hayward, B. E. Clarke, E. E. Duhig, L. H. Passmore, R. V. Bowman, and K. M. Fong. Microrna-218 is deleted and downregulated in lung squamous cell carcinoma. *PLoS One*, 5(9):e12560, 2010.

[23] D. E. Drapper. Themes in RNA-protein recognition. *J Mol Biol*, 293(2):255–270, 1999.

[24] G. Dreyfuss, V. N. Kim, and N. Kataoka. Messenger-RNA-binding proteins and the messages they carry. *Nat Rev Mol Cell Biol*, 3(3):195–205, 2002.

[25] A. D. Ellington and J. W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346(6248):818–822, 1990.

[26] A. P. Feinberg and B. Vogelstein. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301:89–92, 1983.

[27] A. P. Feinberg and B. Vogelstein. Hypomethylation of ras oncogenes in primary human cancers. *Biochem. Biophys. Res. Commun*, 111:47–54, 1983.

[28] S. Feng, L. Rubbi, S. E. Jacobsen, and M. Pellegrini. Determining DNA methylation profiles using sequencing. *Methods Mol Biol*, 733:223–38, 2011.

[29] Y. Feng and A. Bankston. The star family member QKI and cell signaling. *Adv Exp Med Biol*, 693:25–36, 2010.

[30] L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7:85–97, 2006.

[31] J. J. Findeis-Hosey and H. Xu. The use of insulin like-growth factor II messenger RNA binding protein-3 in diagnostic pathology. *Hum Pathol*, 42:303–314, 2011.

[32] M. B. Friedersdorf and J. D. Keene. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol*, 15:R2, 2014.

[33] P. Frisone, D. Pradella, A. Di Matteo, E. belloni, C. Ghigna, and M. P. Paronetto. SAM68: Signal Transduction and RNA Metabolism in Human Cancer. *BioMed Research International*, 2015:14, 2015.

[34] M. Goldman, B. Craft, T. Swatloski, M. Cline, O. Morozova, M. Diekhans, D. Haussler, and J. Zhu. The UCSC Cancer Genomics Browser: update 2015. *Nucleic Acids Res.*, 43(Database issue):D812–D817, 2015.

[35] J. R. Graff, B. W. Konicek, and R. L. Lynch. eIF4E activation is commonly Elevated in advanced human prostate cancers and significantly related to reduced patient survival. *Cancer Research*, 69(9):3866–73, 2009.

[36] C. E. Grant, T. L. Bailey, and W. S. Nobel. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–18, 2011.

[37] D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein. Comparing protein abundance and mrna expression levels on a genomic scale. *Genome Biol*, 4(9):117, 2003.

[38] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A. C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010.

[39] D. Hanahan and R. A. Weinberg. The Hallmarks of Cancer. *Cell*, 100:57–70, 2000.

[40] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38(4):576–589, 2010.

[41] M. Hiller, R. Pudimat, A. Busch, and R. Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res*, 34(17):e117, 2006. [PubMed:16987907] [PubMed Central:PMC1903381] [doi:10.1093/nar/gkl544].

[42] J. J. Ho and P. A. Marsden. Competition and collaboration between rna-binding proteins and micrornas. *Wiley Interdiscip. Rev. RNA*, 5(1):69–86, 2014.

[43] S. D. Hsu, Y. T. Tseng, S. Shrestha, Y. L. Lin, A. Khaleel, C. H. Chou, C. F. Chu, H. Y. Huang, C. M. Lin, S. Y. Ho, T. Y. Jian, F. M. Lin, T. H. Chang, S. L. Weng, K. W. Liao, I. E. Liao, C. C. Liu, and H. D. Huang. mirtarbase update 2014: an information resource for experimentally validated mirna-target interactions. *Nucleic Acids Res.*, 42:D78–85, 2014.

[44] Y. Huang, R. Gattoni, J. Stevenin, and J. A. Steitz. SR splicing factors serve as adapter proteins for TAP-dependent mRNA export. *Mol. Cell*, 11:837–843, 2003.

[45] S. Huttelmaier, D. Zenklusen, M. Lederer, J. Dictenberg, M. Lorenz, and X. Meng. Spatial regulation of beta-actin translation by Src-dependent phosphorylation of ZBP1. *Nature*, 438:512–515, 2005.

[46] A. Jacobsen, J. Silber, G. Harinath, J. T. Huse, N. Schultz, and C. Sander. Analysis of microRNA-target interactions across diverse cancer types. *Nature Structural and Molecular Biology*, 20:1325–32, 2013.

[47] C. H. Jan, R. C. Friedman, J. G. Ruby, and D. P. Bartel. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature*, 469:97–101, 2011.

[48] P. Ji, S. Diederichs, W. Wang, S. Böing, R. Metzger, P. M. Schneider, N. Tidow, B. Brandt, H. Buerger, E. Bulk, et al. Malat-1, a novel noncoding rna, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, 22(39):8031–8041, 2003.

[49] Q. Ji, L. Zhang, X. Liu, L. Zhou, W. Wang, Z. Han, H. Sui, Y. Tang, Y. Wang, N. Liu, et al. Long non-coding rna malat1 promotes tumour growth and metastasis in colorectal cancer through binding to sfpq and releasing oncogene ptbp2 from sfpq/ptbp2 complex. *British journal of cancer*, 111(4):736–748, 2014.

[50] P. E. Johnson and L. W. Donaldson. RNA recognition by the Vts1p SAM domain. *Nat Struct Mol Biol*, 13(2):177–178, 2006.

[51] F. V. Karginov, C. Conaco, Z. Xuan, B. H. Schmidt, J. S. Parker, G. Mandel, and G. J. Hannon. A biochemical approach to identifying microRNA targets. *Proc Natl Acad Sci USA*, 104(49):19291–19296, 2007.

[52] R. Karni and et al. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat. Struct. Mol. Biol*, 14:185–193, 2007.

[53] H. Kazan, D. Ray, E. T. Chan, T. R. Hughes, and Q. Morris. RNAcontext: A new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comp Biol*, 6(7):e1000832, 2010.

[54] J. Keene, J. Komisarow, and M. Friedersdorf. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nature Prot.*, 1(1):302–307, 06 2006.

[55] m. Y. Kim, J. Hur, and S. Jeong. Emerging roles of RNA and RNA-binding protein network in cancer cells. *BMB Rep*, 42(3):125–30, 2009.

[56] M. Kulis, A. C. Queiros, R. Beekman, and J. I. Martin-Subero. Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochim Biophys Acta.*, 1829(11):1161–74, 2013.

[57] J. König, K. Zarnack, N. M. Luscombe, and J. Ule. Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet*, 13(2):77–83, 2012.

[58] M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–151, 2003.

[59] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120:15–20, 2005.

[60] J. H. Li, S. Liu, H. Zhou, L. H. Qu, and J. H. Yang. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.*, 42:D92–D97, 2014.

[61] Y. Li, M. Liang, and Z. Zhang. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput Biol*, 10(10):e1003908, 2014.

[62] W. T. Liao, J. L. Liu, Z. G. Wang, Y. M. Cui, L. Shi, T. T. Li, X. H. Zhao, X. T. Chen, Y. Q. Ding, and L. B. Song. High expression level and nuclear localization of sam68 are associated with progression and poor prognosis in colorectal cancer. *BMC Gastroenterol*, 13:126, 2013.

[63] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[64] S. Lund, D. Nettleton, D. J. McCarthy, and G. K. Smyth. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology*, 11:Article 8, 2012.

[65] M. Maemondo, A. Inoue, K. Kobayashi, S. Sugawara, S. Oizumi, H. Isobe, A. Gemma, M. Harada, H. Yoshizawa, I. Kinoshita, et al. Gefitinib or chemotherapy for non–small-cell lung cancer with mutated egfr. *New England Journal of Medicine*, 362(25):2380–2388, 2010.

[66] H. Mataki, H. Enokida, T. Chiyomaru, K. Mizuno, R. Matsushita, Y. Goto, N. R., I. Higashimoto, T. Samukawa, M. Nakagawa, H. Inoue, and N. Seki. Downregulation of the microrna-1/133a cluster enhances cancer cell migration and invasion in lung-squamous cell carcinoma via regulation of coronin1c. *J Hum Genet*, 53-61:2120–2133, 2015.

[67] A. Mathelier, O. Fornes, D. J. Arenillas, C. Y. Chen, G. Denay, J. Lee, W. Shi, C. Shyr, G. Tan, R. Worsley-Hunt, A. W. Zhang, F. Parcy, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, page gkv1176, 2015.

[68] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 42:D142–7, 2014.

[69] D. Maticzka, S. J. Lange, F. Costa, and R. Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*, 15(1):R:17, 2014.

[70] D. J. McCarthy, Y. Chen, and G. K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, 40:4288–97, 2012.

[71] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*, 12(4):R41, 2011.

[72] G. Michlewski, J. R. Sanford, and J. F. Cáceres. The splicing factor SF2/ASF regulates translation initiation by enhancing phosphorylation of 4E–BP1. *Mol. Cell*, 30:837–843, 2008.

[73] S. Mili and J. A. Steitz. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA*, 10(11):1692–1694, 2004.

[74] R. Mizutani, N. Imamachi, Y. Suzuki, H. Yoshida, N. Tochiqi, T. Oonishi, Y. Suzuki, and N. Akimitsu. Oncofetal protein IGF2BP3 facilitates the activity of proto-oncogene protein eIF4E through the destabilization of EIF4E-BP2 mRNA. *Oncogene*, 2(Nov):Epub ahead of print, 2015.

[75] N. Mukherjee, D. L. Corcoran, J. D. Nusbaum, D. W. Reid, S. Georgiev, M. Hafner, M. J. Ascano, T. Tuschl, U. Ohler, and J. D. Keene. Integrative Regulatory Mapping Indicates that the RNA-Binding Protein HuR Couples Pre-mRNA Processing and mRNA Stability . *Mol. Cell*, 43(3):327–339, 2011.

[76] U. K. Muppirala, V. G. Honavar, and D. Dobbs. Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics*, 12:489, 2011.

[77] J. W. Nam, O. S. Rissland, D. Koppstein, C. Abreu-Goodger, C. H. Jan, V. Agarwal, M. A. Yildirim, A. Rodriguez, and D. P. Bartel. Global analyses of the effect of different cellular contexts on microRNA targeting. *Molecular Cell*, 53:1031–43, 2014.

[78] J. Nature. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[79] J. Nielsen, J. Christiansen, J. Lykke-Andersen, A. H. Johnsen, U. M. Wewer, and C. Nielsen, F. A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development. *Mol Cell Biol*, 19:1262–70, 1999.

[80] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas. How Many Trees in a Random Forest? *Machine Learning and Data Mining in Pattern Recognition*, Volume 7376 of the series Lecture Notes in Computer Science:154–168, 2012.

[81] V. Pancaldi, O. S. Saraç, C. Rallis, J. R. McLean, M. Převorovský, K. Gould, A. Beyer, and J. Bähler. Predicting the fission yeast protein interaction network. *G3*, 2(4):453–467, 2012.

[82] T. Phillips. The Role of Methylation in Gene Expression. *Nature Education*, 1(1):116, 2008.

[83] J. Qian, M. Hassanein, M. D. Hoeksema, B. K. Harris, Y. Zou, H. Chen, P. Lu, R. Eisenberg, J. Wang, A. Espinosa, X. Ji, F. T. Harris, S. M. Rahman, and P. P. Massion. The RNA binding protein FXR1 is a new driver in the 3q26-29 amplicon and predicts poor prognosis in human cancers. *Proc Natl Acad Sci U S A*, 106(3):3469–74, 2012.

[84] P. Rajan, L. Gaughan, and C. Dalgliesh. Regulation of gene expression by the RNA-binding protein Sam68 in cancer. *Biochemical Society Transactions*, 36(3):505–507, 2008.

[85] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, and et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499:172–177, 2013.

[86] M. Robinson, D. McCarthy, and G. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, 2010.

[87] M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23:2881–87, 2007.

[88] M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9:321–332, 2008.

[89] A. F. Ross, Y. Oleynikov, E. H. Kislauskis, K. L. Taneja, and R. H. Singer. Characterization of a beta-actin mRNA zipcode-binding protein. *Moll Cell Biol*, 17:2158–65, 1997.

[90] S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, and J. S. Weissman. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, 505:701–705, 2014.

[91] R. Sager. Expression genetics in cancer: shifting the focus from DNA to RNA. *Proc Natl Acad Sci USA. (PNAS)*, 94:952–955, 1997.

[92] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011.

[93] M. Setty, K. Helmy, A. A. Khan, J. Silber, A. Arvey, F. Neezen, P. Agius, J. T. Huse, E. C. Holland, and C. S. Leslie. Inferring transcriptional and microrna-mediated regulatory programs in glioblastoma. *Molecular systems biology*, 8(1):605, 2012.

[94] A. Shlien and D. Malkin. Copy number variation and cancer. *Genome Med*, 1(6):62, 2009.

[95] R. C. Spitale, R. A. Flynn, Q. C. Zhang, P. Crisalli, B. Lee, J. Jung, H. Y. Kuchelmeister, P. J. Batista, E. A. Torre, E. T. Kool, and H. Y. Chang. Structural imprints in vivo decode rna regulatory mechanisms. *Nature*, 519:486–490, 2015.

[96] E. Stickeler, F. Kittrell, D. Medina, and S. M. Berget. Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis. *Oncogene*, 18:3574–82, 1999.

[97] J. Taylor and R. Tibshirani. Statistical learning and selective inference. *PNAS*, 112(25):7629–7634, 2015.

[98] B. Tian, J. Hu, H. Zhang, and C. S. Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, 33:201–212, 2005.

[99] R. Tibshirani. Regression shrinkage and selection via the lasso. *J R Stat Soc Series*, 58:267–228, 1996.

[100] J. Ule, K. Jensen, A. Mele, and R. B. Darnell. CLIP:A method for identifying protein-RNA interaction sites in living cells. *Methods*, 37:376–386, 2005.

[101] M. Wan, W. Huang, T. Kute, L. Miller, Q. Zhang, H. Hatcher, J. Wang, D. Stovall, G. Russell, P. Cao, Z. Deng, W. Wang, Q. Zhang, M. Lei, S. Torti, S. A. Akman, and G. Sui. Yin yang 1 plays an essential role in breast cancer and negatively regulates p27. *Am J Pathol*, 180(5):2120–2133, 2012.

[102] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[103] J. Wang, Y. Guo, H. Chu, Y. Guan, J. Bi, and B. Wang. Multiple functions of the RNA-binding protein HuR in cancer progression, treatment responses and prognosis. *Int J Mol Sci*, 14(5):10015–41, 2013.

[104] J. Wang, B. Wang, J. Bi, and C. Zhang. Cytoplasmic hur expression correlates with angiogenesis, lymphangiogenesis, and poor outcome in lung cancer. *Med Oncol*, 28:S577–S585, 2011.

[105] L. Wurth. Versatility of RNA-Binding Proteins in Cancer. *Comparative and Functional Genomics*, Article ID 178525:11, 2012.

[106] F. Zambelli, G. Pesole, and G. Pavesi. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, 1:1–13, 2012.

[107] J. Y. Zhang, E. K. Chan, X. X. Peng, and E. M. Tan. A novel cytoplasmic protein with RNA-binding motifs is an autoantigen in human hepatocellular carcinoma. *J Exp Med*, 189:1101–1110, 1999.

[108] Z. Zhang and A. R. Krainer. Involvement of SR proteins in mRNA surveillance. *Mol. Cell*, 16:597–607, 2004.

[109] Z. Zhang, J. Li, H. Zheng, C. Yu, J. Chen, Z. Liu, M. Li, M. Zeng, F. Zhou, and L. Song. Expression and cytoplasmic localization of sam68 is a significant and independent prognostic marker for renal cell carcinoma. *Cancer Epidemiology Biomarkers & Prevention*, 18 (10):2685–2693, 2009.

[110] J. Zhao, T. K. Ohsumi, J. T. Kung, Y. Ogawa, D. J. Grau, K. Sarma, J. J. Song, R. E. Kingston, M. Borowsky, and J. T. Lee. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol cell*, 40(6):939–953, 12 2010. [PubMed:21172659] [PubMed Central:PMC3021903] [doi:10.1016/j.molcel.2010.12.011].

[111] D. Zhu, C. R. Stumpf, J. M. Krahn, M. Wickens, and T. M. Hall. A 5' cytosine binding pocket in Puf3p specifies regulation of mitochondrial mRNAs. *Proc Natl Acad Sci USA*, 106(48):20192–20197, 2009.

[112] F. Y. Zong, X. Fu, W. J. Wei, Y. G. Luo, M. Heiner, L. J. Cao, Z. Fang, R. Fang, D. Lu, H. Ji, and J. Hui. The RNA-Binding Protein QKI Suppresses Cancer-Associated Aberrant Splicing. *PLoS Genet.*, 10(4):e1004289, 2014.