INFORMATICS APPROACHES FOR IDENTIFYING DRUG-SPECIFIC MARKERS AND GENETIC REGULATION MECHANISM IN CANCERS

A Dissertation
Presented to
The Academic Faculty

By

Shuting Lin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biological Sciences

Georgia Institute of Technology

May 2025

INFORMATICS APPROACHES FOR IDENTIFYING DRUG-SPECIFIC MARKERS AND GENETIC REGULATION MECHANISM IN CANCERS

Thesis committee:

Dr. Peng Qiu
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Soojin Yi
Department of Ecology, Evolution and Marine Biology
University of California, Santa Barbara

Dr. Xiuwei Zhang
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Saurabh Sinha
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Yong Teng
Department of Hematology and Medical Oncology
Emory University

Date approved: March 5th, 2025

For my family and friends

ACKNOWLEDGMENTS

Pursuing a Ph.D. was never part of my life plan. I didn't have a detailed roadmap or a strong desire to achieve it—it wasn't the result of ambition, but something that unfolded naturally. By the time I realized it, I was already on this path. Finishing this journey has shaped me in many ways, both in how I think and in what I believe. Along the way, I've been met with so much kindness and support.

I am incredibly grateful to my Ph.D. advisor, Dr. Peng Qiu, whom I feel truly fortunate to have met. He gave me the freedom and space to explore the questions I was curious about, while consistently offering support along the way. I know I'm not someone with a natural gift for research, but he never lost patience with me. He constantly encouraged me to think deeper, and was always honest and generous in sharing his ideas and feedback. His guidance helped shape the way I approach problems and gradually strengthened my ability to think. Through his example, I've come to understand what it means to be a dedicated and professional scholar. Although I still have a long way to go, I hope that one day I can become someone like him—doing meaningful work in a field I love, and bringing encouragement and energy to those around me.

I also want to sincerely thank all my committee members for their guidance and thoughtful reviews. Although they are all incredibly busy, they always responded to my emails with kindness and gave thoughtful feedback that helped me improve my work.

Appreciation also goes to my lab mates of the past and now, who have been great colleagues, supporters, and friends. Thank you for the conversations, the help, and the laughter. There was much to learn from working alongside such colleagues, and I sincerely wish them all the best in the future.

I am deeply grateful to my parents. I know I've given them plenty to worry about over the years, but they have always shown me unwavering love and understanding. After I started college, my life changed in many ways. Even if they didn't fully understand the

situations I faced, they supported every decision I made. I'm especially thankful for their encouragement and support in helping me study abroad, and for all the sacrifices they made along the way. It is because of them that I've had the confidence and freedom to pursue the life I want.

I also want to thank my fiancé, Xinze Wang, who has been with me through this entire Ph.D. journey. He was always willing to listen and helped me rehearse my research presentations—even when he didn't fully grasp the science behind them, he still tried to understand and offered thoughtful suggestions. He stood by me during the most difficult times, offering comfort and encouragement when I needed it most. Beyond the work, he reminded me to appreciate the beauty in life—both in the everyday moments and in the places we explored together. I feel incredibly fortunate to have had him throughout this chapter of my life.

To all my friends, thank you for being part of this time in my life. I feel truly lucky to have had your presence along the way. To all my family members, thank you for your support and blessings. And to everyone who has seen my imperfections and still chosen to care, your kindness means more than words can say. A special thank you also goes to my cat, Cengceng, whose quiet companionship brought me comfort and joy through long days and late nights.

Lastly, I want to thank myself—not the smartest, but brave enough to keep going. Thank you for not giving up during moments of doubt, and for staying curious about the world. I hope to move forward with the same curiosity and courage, and continue chasing a life that is meaningful, joyful, and true to who I am.

TABLE OF CONTENTS

Acknowledgments iv

List of Tables x

List of Figures xii

List of Acronyms xv

Summary xvii

Chapter 1: Introduction 1

1.1 Background 1

1.1.1 Cancer treatment 2

1.1.2 The role of molecular biomarkers in cancer treatment 4

1.1.3 The role of genetic regulation in cancer treatment 5

1.2 Thesis outline 7

Chapter 2: Identifying biomarkers that are predictive of drug-specific survival outcome in cancer 8

2.1 Introduction 8

2.2 Methods 11

2.2.1 Data access 11

2.2.2 Data preprocessing 12
2.2.3 Survival analysis 12
2.2.4 Literature survey 13
2.3 Results 13
2.3.1 Identify miRNAs predictive of survival in cancer-drug groups . . . 13
2.3.2 Literature survey of identified miRNA markers 18
2.3.3 Correlation between drug-specific miRNAs and their target genes . 20
2.3.4 Prognostic performance of drug-specific miRNAs in independent datasets 24
2.3.5 Significant proteins predictive of drug-specific survival 25
2.3.6 Literature support of predictive protein markers 27
2.3.7 Correlation between predictive proteins and their coding genes . . . 30
2.4 Discussion 32

Chapter 3: Establishing the correlation patterns between CpG methylation and alternative splicing in cancer 35

3.1 Introduction 35
3.2 Methods 38
3.2.1 Data access 38
3.2.2 Data preprocessing 38
3.2.3 Correlation analysis 38
3.2.4 Survival analysis 39
3.2.5 Literature search 39
3.3 Results 39

3.3.1 Correlations between DNA methylation and alternative splicing . . 39

3.3.2 CpG sites correlated with exons tended to be more predictive of survival outcomes 43

3.3.3 CpG sites are more strongly correlated with exons than their correlation with isoforms 46

3.4 Discussion 47

Chapter 4: Predicting potential miRNA target genes 50

4.1 Introduction 50

4.2 Methods 52

4.2.1 Data access 52

4.2.2 Correlation analysis between miRNAs and genes 52

4.2.3 Construction of prediction model 53

4.2.4 Literature survey 53

4.2.5 Differential expression analysis 53

4.3 Results 53

4.3.1 Construction of miRNA-gene prediction models 53

4.3.2 Validation of the identified miRNA-gene pairs in the validation databases and literature 56

4.3.3 Validation of the identified miRNA-gene pairs in independent datasets 58

4.4 Discussion 61

Chapter 5: Clustering analysis of multiple omics data types identifies cancer patients with consistent survival outcomes 63

5.1 Introduction 63

5.2 Methods 65

5.2.1 Data acquisition and preprocessing 65

5.2.2 Clustering analysis 65

5.2.3 Survival analysis 66

5.2.4 Literature survey 66

5.2.5 Construction of DE features network and analysis 66

5.3 Results 66

5.3.1 Clustering analysis of individual omics data types for each cancer type 66

5.3.2 Significant survival differences were observed among patient clusters defined by multiple omics data types in various cancers 68

5.3.3 Patient clusters with the most prominent survival outcomes showed a high degree of consistency across omics data types 69

5.3.4 Differentially expressed features associated with the most prominent survival outcomes 72

5.3.5 Network analyses identified molecular modules associated with patient survival outcomes in cancers 74

5.4 Discussion 76

Chapter 6: Conclusions and Future works 79

References 82

LIST OF TABLES

2.1 Predictive miRNAs identified in BRCA-drug groups 12

2.2 Significant miRNAs identified in cancer-drug groups. 15

2.2 This table represents the number of significant miRNAs identified in all cancer-drug groups. The Literature column represents the number of literature found by PubMed search for each combination. 18

2.3 Significant miRNA-cancer-drug combinations, and corresponding target genes that are also predictive of survival outcoms. 22

2.3 This table represents the miRNA-mediated gene-cancer-drug combinations in all cancer-drug groups. 24

2.4 Additional GEO datasets that served as independent validation datasets . . . 25

2.5 65 Significant protein-cancer-drug combinations identified in cancer-drug groups 28

3.1 Top 30 multi-cancer CpG-exon correlations across all the cancer types . . . 41

3.2 CpG sites that correlate with both exon expression and survival outcome . . 43

4.1 Number of miRNA-gene pairs consistently identified as positives by predictive models with different frequency thresholds, and the proportion of these pairs reported in the validation database 57

4.2 Comparison of predicted miRNA targets of different methods and known miRNA targets based on differentially expressed genes in independent datasets. 58

4.3 Examination of the relationship between the correlation direction of the predicted miRNA-gene pairs and their regulatory direction 60

5.1 Number of patient clusters produced by different omics data types for each cancer type 67

5.2 The number of DE features identified in each of the 4 cancer types 72

LIST OF FIGURES

2.1 Heatmap of number of patients in each cancer-drug group. 14

2.2 Kaplan-Meier curves of overall survival for patients treated with gem-citabine at low or high expressed classes stratified by miR-330 in the BLCA or SARC. 19

2.3 Kaplan-Meier curves of overall survival for patients treated with cis-platin or gemcitabine at low or high expressed classes stratified by miR-296 inthe CESC or SARC. 19

2.4 Kaplan-Meier curves of overall survival for patients treated with cis-platin at low or high expressed classes stratified by miR-483 in the BLCA or CESC. 20

2.5 Kaplan–Meyer survival curves of BRCA patients in additional GEOdatasets, stratified by drug-specific miRNA marker identified from TCGA data.(A) BRCA patients in GSE40267 stratified by miR-363. (B) BRCA patients inGSE19783 stratified by miR-20b. (C) BRCA patients in GSE19783 stratified by miR-628. (D) BRCA patients in GSE22220 stratified by miR-577. . 26

2.6 Kaplan-Meier curves of overall survival for patients treated with fluorouracil at low or high expressed classes stratified by PTEN or SRC in the PAAD. . 29

2.7 Correlation between Bcl-2 and BCL2 in expression level and survival outcomes. A-B. Kaplan-Meier curves of overall survival for patients treated with fluorouracil at low or high expressed classes stratified by Bcl-2 or BCL2 in the HNSCC. C. Consistency of Bcl-2 and BCL2 correlation in expression levels for patients in HNSCC-carboplatin group. 30

3.1 flowchart of data analysis used in this study. First, we selected gene-specific samples that highly expressed the gene for each of the genes we analysed, and then we performed correlation analysis between DNA methylation and exon/isoform expression based on gene-specific samples. A threshold of FDR<0.05 was required to detect significant correlations. In addition, we performed survival analysis for each CpG sites and identified a list of CpG sites that are predictive of patients' survival outcome. Finally, a literature survey was conducted to search for existing evidences of identified correlations. 37

3.2 A: Cluster heatmap of the methylation data of the CpG sites associated with CD302 in brain cancer. Patients were separated into two groups based on the clustering results. B: Kaplan–Meyer survival curves of the two patient group stratified by the CpGs associated with CD302. 44

3.3 CpGs correlated with exons and survival. This Venn diagram shows the overlapping between the CpGs correlated with the exons and the CpGs that are predictive of survival. 45

3.4 Comparison of correlation strength between CpG-exon and CpG-isoform. This histogram shows the differences between CpG-exon correlations and the corresponding strongest CpG-isoform correlations, which indicates that CpG-exon correlations tend to be stronger than the corresponding CpG-isoform correlations. 47

4.1 Classification accuracies for the prediction models on the testing datasets across 1000 experiments. 55

5.1 Overview of the analysis workflow in this study. The steps include the collection of three omics data types (miRNA expression, gene expression, and DNA methylation) from TCGA, followed by clustering analysis for each omics type. Subsequent analyses involve survival analysis of patient clusters and cross-omics cluster comparison to identify patients with consistent survival outcomes across multiple omics data for each cancer type. Finally, differential expression analysis and network analysis are performed on the identified patient groups to discover molecular modules associated with survival outcomes. 64

5.2 Log-rank analysis of patient clusters across 20 cancer types in TCGA. Summary of log-rank analysis to assess survival differences among patient clusters defined by different omics data types for each cancer type. Colors indicate statistical significance, highlighting clinically relevant clusters. . . . 69

5.3 Kaplan-Meier survival curves depicting the survival outcomes of patient clusters derived from miRNA, gene expression, and DNA methylation data for each of the six cancer types where significant survival differences were observed in multiple omics. The rows represent different cancer types, and the columns represent different omics data types. 71

5.4 Gene Ontology (GO) enrichment analysis results for differentially expressed genes in BLCA, BRCA, KIRC, and LGG. The enriched biological processes are shown for each cancer type, with color intensity representing the significance level of enrichment. The analysis reveals key biological processes linked to cancer development and progression, underlining the molecular characteristics associated with different patient clusters. 73

5.5 Densely connected network modules of DE features associated with consistently clustered patients in BLCA, BRCA, KIRC, and LGG. Nodes are color-coded to represent the type of feature, and edges indicate known interactions among the DE features based on miRNA-gene and gene-gene relationships. These network modules suggest potential molecular interactions that may influence patient survival outcomes. 75

LIST OF ACRONYMS

BLCA Bladder urothelial carcinoma

BRCA Breast invasive carcinoma

CESC Cervical squamous cell carcinoma and endocervical adenocarcinoma

DAVID Database for Annotation, Visualization and Integrated Discovery

DE Differential expression

DFS Disease-free survival

DRFS Distant-relapse free survival

ESCA Esophageal carcinoma

FDR False Discovery Rate

GDC Genomic Data Commons

GO Gene Ontology

HNSC Head and Neck squamous cell carcinoma

KEGG Kyoto Encyclopedia of Genes and Genomes

KIRC Kidney renal clear cell carcinoma

LGG Brain Lower Grade Glioma

LUAD Lung adenocarcinoma

miRNA microRNA

OS Overall survival

OV Ovarian serous cystadenocarcinoma

PAAD Pancreatic adenocarcinoma

READ Rectum adenocarcinoma

SARC Sarcoma

scRNA-seq Single-cell RNA sequencing

SNF Similarity Network Fusion

TCGA The Cancer Genome Atlas

TSCC Tongue squamous cell carcinoma

UCEC Uterine Corpus Endometrial Carcinoma

XGBoost eXtreme Gradient Boosting

SUMMARY

Focusing on developing informatics approaches to advance precision oncology, this research work encompassed the identification of molecular markers predictive of drug response and the investigation of genetic regulatory mechanisms in cancers through multi-omics data integration, aiming to contribute to a deeper understanding of cancer biology and support the development of personalized therapeutic strategies.

In Chapter 2 of this dissertation, we presented an integrative analysis to identify miRNAs and proteins that are predictive of drug response in cancer. By integrating transcriptomic profiling data, drug treatment information, and survival outcomes, we identified markers whose expression levels are correlated with drug-specific survival outcome. Many of these markers showed consistent relationships with their target or coding genes in terms of expression levels and correlations with survival outcomes. Additional validation in independent datasets further demonstrated their potential to predict survival outcomes in a cancer- and drug-specific manner. These findings provide critical insights into the molecular features of therapeutic responses and highlight their promise as tools for precision oncology.

In Chapter 3 of this dissertation, we explored the interplay between CpG methylation and alternative splicing in cancer, an area that remains poorly discussed in the literature. By integrating methylation and transcriptomic data, this study systematically examined the associations between CpG methylation and exon expression across various cancer types. In particular, the findings demonstrated that CpG sites associated with exon expression were more likely to correlate with patient survival outcomes than those not linked to exon expression. Furthermore, we performed pairwise correlation analysis between CpG sites and isoforms, and compared the correlation strength between CpG-exon and CpG-isoform. Our analysis demonstrated stronger correlations between CpGs and exons compared to the correlations between CpGs and isoforms that contain the exons, indicating that CpG

methylation may be associated with alternative splicing via regulating the inclusion or exclusion of exons, which subsequently impacts the relative usage of various isoforms. These findings provide novel insights into the relationship between DNA methylation and transcriptional regulation, advancing our understanding of genetic regulatory mechanisms in cancer biology.

In Chapter 4 of this dissertation, we developed machine learning models to predict novel miRNA target genes that have not been previously reported. Although several databases curate known miRNA target genes, they likely represent only a subset of all possible interactions. To address this gap, we conducted a comprehensive correlation analysis between miRNAs and genes across multiple cancer types. Using curated miRNA target databases as a reference, we defined class labels for each miRNA-gene pair and trained machine learning models to predict potential miRNA-gene relationships, offering a systematic approach to uncovering previously unidentified interactions. Our findings were further validated by held-out miRNA target databases and a literature survey. Additionally, the analysis of independent datasets demonstrated a consistent alignment between the correlation directions of miRNA-gene pairs and their regulatory patterns. This work provides a framework for uncovering previously unreported miRNA-gene interactions, contributing to the broader understanding of miRNA-mediated regulation in cancer.

In Chapter 5 of this dissertation, we explored cancer subtyping across various cancer types by performing clustering analyses on multi-omics data. Our analysis revealed that, in several cancer types, specific patient groups consistently clustered together across different omics data layers and exhibited either the most favorable or unfavorable survival outcomes. This observation suggests that patients with most prominent survival outcomes may exhibit distinct and robust expression patterns spanning multiple genomic dimensions, detectable through clustering analyses of different omics data types. These findings underscore the potential of integrative multi-omics clustering to identify biologically meaningful patient subtypes in cancer research.

CHAPTER 1
INTRODUCTION

1.1 Background

Cancer, a complex and heterogeneous disease, remains a leading cause of death worldwide, accounting for over 10 million deaths annually [1, 2]. Beyond its high mortality rates, a the clinical challenge of cancer lies in its pronounced heterogeneity. This heterogeneity is evident not only across different cancer types but also within tumors of the same type, where variations in molecular and genetic characteristics are often observed [3]. Even tumors with similar histological features can exhibit distinct genetic and epigenetic profiles, driving differences in progression, therapeutic resistance, and overall behavior [4]. Additionally, the tumor microenvironment, comprising immune cells, stromal elements, and signaling molecules, further influences the clinical outcomes and treatment responses [5]. These multifaceted factors collectively contribute to the unpredictable nature of therapy responses, highlighting the urgent need for a deeper understanding of cancer biology. Such insights are essential for developing precise therapeutic strategies that can improve patient outcomes.

While traditional treatments such as surgery, chemotherapy, and radiation therapy have long been central to cancer care, their generalized approaches often lead to limited efficacy and significant side effects due to a lack of precision [6]. In contrast, personalized treatment leverage molecular insights to tailor therapies to individual patients, offering the potential to improve outcomes and minimize toxicity [7]. As a key aspect of personalized treatment, molecular markers offer valuable insights into tumor biology and guide clinical decisions. These markers are regulated by complex transcriptional, epigenetic, and post-transcriptional processes that drive cancer progression and influence therapeutic re-

sponses [8, 9]. Investigating these regulatory mechanisms enhances our understanding of the complex pathways underlying cancer development and provides opportunities to identify novel therapeutic targets. However, investigating these intricate processes requires comprehensive approaches to analyze the diverse molecular interactions. Integrative analysis of multi-omics cancer data, which incorporates information from multiple molecular layers, enables us to explore different molecular layers simultaneously [10]. By capturing the complex interplay between different molecular mechanisms, such analyses reveal drivers of cancer progression and therapy resistance, advancing our understanding of cancer biology and enabling the development of more effective treatments for individual patients.

This dissertation employs informatics approaches to address challenges in cancer treatment by identifying drug-specific markers and elucidating genetic regulatory mechanisms. By integrating multi-omics data with computational approaches, it aims to advance knowledge in cancer research and contribute to the development of precision oncology strategies.

1.1.1 Cancer treatment

For decades, cancer treatment has been dominated by three pillars: surgery, chemotherapy, and radiation therapy [11]. Surgery remains a cornerstone for localized cancers, offering potential cures when complete tumor removal is achieved [12]. Chemotherapy extends the scope of treatment to systemic and advanced cancers by targeting rapidly dividing cells throughout the body, while radiation therapy is primarily localized, focusing on destroying rapidly dividing cells within specific tumor sites [13, 14]. However, these approaches often lack precision, as they struggle to differentiate between cancerous and healthy tissues, resulting in significant side effects. Additionally, not all patients respond equally to these treatments due to resistance mechanisms and individual differences in tumor biology, highlighting the growing need for more precise, patient-specific treatment strategies.

Personalized treatment represents a paradigm shift in cancer care, involving medical strategies tailored to each patient to maximize treatment efficacy while minimizing side

effects [15]. The concept of personalized treatment has been discussed for decades, but it gained significant attention in the early 2000s with advances in genomics and molecular biology that allowed for a deeper understanding of individual variability in disease and treatment response [16]. Unlike traditional approaches that treat all patients similarly, personalized treatment acknowledges that individual differences in genetics and phenotypes can significantly impact how a person responds to medical treatment. One of the cornerstones of personalized cancer treatment is targeted therapy, which uses drugs specifically designed to attack cancer cells by targeting molecules essential for their growth and survival [17]. For example, therapies like imatinib and trastuzumab target BCR-ABL and HER2-positive cancer cells, respectively, offering significant improvements in outcomes for cancer patients [18, 19]. Another transformative approach is immunotherapy, which stimulates the immune system's natural ability to identify and destroy cancer cells [20]. Immune checkpoint inhibitors, such as pembrolizumab and nivolumab, have revolutionized the treatment landscape for cancers like melanoma and lung cancer [21, 22, 23]. Additionally, bone marrow transplantation serves as a critical treatment option for hematologic cancers, allowing for the replacement of damaged or diseased bone marrow with healthy stem cells to restore normal blood cell production [24].

The rise of personalized treatment has been driven by rapid advancements in technologies and the availability of large-scale biological data. Innovations such as polymerase chain reaction (PCR) [25], microarray technologies [26], next-generation sequencing [27], RNA-sequencing [28], and Single-cell RNA sequencing (scRNA-seq) [29] have transformed the ability to analyze cancer at the molecular level. These technologies, from the introduction of PCR in 1983 to single-cell RNA sequencing in the early 2010s, have enabled researchers to dissect the molecular complexity of tumors and identify actionable therapeutic targets.

At the same time, the extensive availability of genomic, proteomic, and metabolomic datasets has further fueled personalized treatment. Resources like The Cancer Genome At-

las (TCGA) [30] have provided comprehensive molecular and clinical data, encompassing mutation, copy number variation, gene expression, microRNA (miRNA) profiles, methylation patterns, and protein expression. The clinical data include details such as cancer stage, survival outcomes, drug treatments, and various other clinical features. This wealth of multi-omics data has enabled the identification of molecular features associated with cancer and the development of therapies tailored to the unique characteristics of individual patients.

1.1.2 The role of molecular biomarkers in cancer treatment

While advancements in high-throughput technologies underscore the promise of personalized treatment, several critical challenges remain in improving treatment efficacy and patient outcomes. Central to personalized treatment are molecular biomarkers, which provide critical insights into the molecular mechanisms driving cancer and play a pivotal role in guiding clinical decisions [31]. These biomarkers encompass a wide range of molecules, including genes, miRNAs, CpG sites, and proteins, are broadly categorized as diagnostic, prognostic, or monitoring biomarkers. Diagnostic biomarkers are used to identify the presence or specific subtype of cancer, facilitating early detection and accurate classification [32, 33]. Prognostic biomarkers, on the other hand, provide information about disease progression and outcomes independent of therapy, helping to assess the likely course of the disease. Monitoring biomarkers track changes in disease status or the effectiveness of treatment over time, enabling clinicians to adjust therapeutic strategies as needed [34].

Despite their importance, many current biomarkers lack precision in predicting drug responses, limiting their utility in fully realizing the potential of personalized treatment. A critical challenge remains in the development of drug-specific markers, which are essential for predicting how individual patients will respond to specific therapies. In many cancer studies, variations in drug treatment among patients have been largely overlooked [35]. This oversight can be attributed to two primary reasons. First, the need for statistical power

typically shapes study designs, prompting researchers to include as many patients as possible to enhance survival analyses, regardless of variations in treatment. Second, drug treatment data is frequently missing or inconsistently formatted in publicly available cancer genomics and survival datasets, making it difficult to integrate into biomarker discovery efforts [36]. Consequently, most existing biomarkers are generalized to the cancer type or subtype under study, rather than being specific to a particular drug treatment. While studies have successfully identified molecular biomarkers in cancers, the discovery of drug-specific survival biomarkers would provide even greater clinical value, enabling more precise and effective treatment strategies tailored to individual patients.

1.1.3 The role of genetic regulation in cancer treatment

The limited understanding of genetic regulatory mechanisms continues to challenge the development of targeted therapeutic strategies. Genetic regulation plays an important role in cancer progression, affecting tumor initiation, growth, metastasis, and resistance to treatment [37]. These processes include transcriptional and post-transcriptional mechanisms that influence how gene are expressed and interact [38, 39]. A deeper understanding of these processes is critical for identifying novel therapeutic targets and advancing precision oncology.

One key area of genetic regulation that remains incompletely understood is the role of DNA methylation in transcriptional processes, particularly alternative splicing. DNA methylation, which modulates gene expression by adding methyl groups to CpG sites in the genome, can silence tumor suppressor genes or activate oncogenes [40, 41]. Beyond its established role in gene expression, emerging evidence highlights its influence on alternative splicing [42, 43]. This regulation is particularly significant in cancer, where disrupted methylation patterns and aberrant splicing events frequently result in oncogenic or non-functional protein variants. Investigating the interplay between DNA methylation and alternative splicing offers promising opportunities to improve cancer treatment. Such insights

could drive the development of more precise therapies, advancing the goals of personalized treatment.

Another crucial aspect of genetic regulation is the interaction between miRNAs and their target genes. MiRNAs are small, non-coding RNAs that regulate gene expression post-transcriptionally by binding to messenger RNAs, leading to their degradation or inhibiting their translation [44]. This regulatory process is vital for various cellular functions, including cell growth, differentiation, and apoptosis [45]. MiRNAs regulate a broad spectrum of genes, forming an extensive and intricate network that influences critical cellular pathways. Each miRNA can target multiple genes, and individual genes may be regulated by several miRNAs, resulting in a highly interconnected and multilayered regulatory system. This network is dynamic and context-dependent, with miRNA activity influenced by cell type, developmental stage, and pathological state. In cancer, the disruption of these regulatory networks amplifies or suppresses signals that promote malignancy. The intricate interplay between miRNAs and their target genes highlights their pivotal role in maintaining cellular homeostasis and their potential as therapeutic targets in oncology [46]. However, despite significant progress in understanding miRNA biology, our knowledge of the full regulatory landscape remains incomplete. Although several existing databases have cataloged known miRNA-gene interactions, they likely represent only a fraction of the complete miRNA regulatory network. Expanding our understanding of these relationships is essential for elucidating the molecular mechanisms underlying cancer and holds significant potential for identifying novel therapeutic targets.

To gain a more comprehensive understanding of genetic regulatory mechanisms, it is necessary to integrate data from multiple omics layers, such as genomic, transcriptomic, epigenomic, and proteomic data. Each of these layers provides a unique perspective on regulatory mechanisms, and combining them allows for a more holistic view of cancer biology. Furthermore, multi-omics integration has proven especially valuable in identifying distinct patient subtypes or clusters based on shared molecular characteristics [47]. These

subtypes often reveal important differences in tumor biology, disease progression, and therapeutic responses, offering a foundation for personalized treatment strategies. Leveraging multi-omics approaches facilitates the discovery of novel biomarkers and the identification of molecular vulnerabilities, helping the development of precise therapeutic interventions tailored to specific patient groups.

1.2 Thesis outline

The remaining of the thesis is divided into five chapters. Chapter 2 focuses on the identification of biomarkers that predict drug-specific survival outcomes in cancer patients. By integrating multi-omics data, drug treatment records, and survival outcome data, this chapter identifies miRNA and protein biomarkers associated with patient responses to specific therapies, which can be promising tools toward personalized treatment. Chapter 3 investigates the relationship between CpG methylation and alternative splicing in cancers by integrating multiple data types, this chapter successfully identifies significant CpG-exon correlation patterns in various cancer contexts and explores the regulatory mechanism of CpG methylation on alternative splicing. Chapter 4 presents a method for predicting potential miRNA target genes. Building on known miRNA-target interactions and transcriptomic data, this chapter identifies novel miRNA-gene relationships not previously reported in the literature. Chapter 5 conducts a clustering analysis of multi-omics data to identify cancer patient subgroups with consistent survival outcomes. This chapter highlights the shared biological features across various omics layers that are associated with the most prominent survival outcomes. Finally, Chapter 6 summarizes the findings of the thesis and outlines potential future directions.

CHAPTER 2
IDENTIFYING BIOMARKERS THAT ARE PREDICTIVE OF DRUG-SPECIFIC SURVIVAL OUTCOME IN CANCER

2.1 Introduction

One important aspect of personalized medicine in cancer therapy is the ability to predict an individual patient's response to drug treatments. Currently, there is an unmet need for such drug-specific predictive biomarkers, which can spare patients from ineffective toxic agents and optimize treatment for individual patients. Utilizing various powerful sequencing technologies, significant efforts have been spent in genomics and proteomics research to profile cancer patients. Such molecular data of cancer enabled studies of the relationship between molecular expression profiles and clinical outcomes, and revealed genetic and epigenetic alterations as biomarkers predictive of survival. However, in existing studies, variations in drug treatment among patients were often overlooked, even though it has been reported that drug exposure could affect specific histones, modify gene expression, and impact survival outcomes. In our opinion, there are two main reasons why drug treatment was often not considered in the search for survival biomarkers for cancer. One reason is statistical power. Since larger sample size often leads to improved statistical power for identifying cancer survival biomarkers, studies usually chose to include as many relevant patients as possible, while ignoring the fact that not all included patients received the same drug treatment. The other reason is data availability. In publicly available datasets of cancer genomics and survival studies, drug treatment data of patients is often unavailable. In rare cases where drug treatment data of patients is provided, the data is often in inconsistent and non-standardized formats, making it difficult to use. As a result, the existing biomarkers are often general to the cancer or cancer subtype being studied, but not specific to any drug treatment. Although

existing studies have identified useful survival biomarkers, it would be more beneficial if drug-specific survival biomarkers can be identified.

As a valuable data resource for cancer research, TCGA provides an opportunity for exploring biomarkers predictive of drug-specific survival outcomes. The TCGA data resources contain comprehensive molecular characterizations of over 11,000 cancer patients across 33 different cancer types, as well as drug treatment data and clinical outcomes of the patients. The molecular characterizations include mutation, copy number variation, methylation, gene expression, miRNA expression and protein expression. Although the drug treatment data in TCGA contains a lot of variations in naming conventions and spelling issues, our group have previously manually standardized the data to eliminate the nomenclature problems [48]. This previous effort enabled integration of the drug treatment data into bioinformatics pipelines for genomics and survival analysis. In this study, we focused on two critical molecular classes, miRNAs and proteins, to investigate their potential as drug-specific predictive biomarkers for cancer survival.

miRNAs are small non-coding RNAs that regulate gene expression post-transcriptionally and are known to play pivotal roles in cancer biology. It has been showed that miRNA involvements in the development of cancer, including cell proliferation, metastasis, differentiation, and evasion of apoptosis [49]. Moreover, it is proposed that expression of more than one third of human genes are under miRNAs control, which means many drug target genes may be regulated by miRNAs [50]. In addition, the functional pattern of miRNA–mRNA interaction network has been shown in the onset and development of cancer [51, 52, 53]. Two previous studies took the advantage of network-based models and explored the regulatory mechanism between miRNA and mRNA in colorectal and pancreatic cancer by using the matched specimens of human cancer tissue and adjacent non-tumorous mucosa. Such network-based analysis identified cancer-specific biomarkers, as well as the relationship between deregulated miRNA expression and biological pathways involved in the cancers [54, 55]. Since miRNAs play regulatory roles on numerous genes involved in

oncology and pharmacology, differences in the levels of circulating miRNAs can contribute to inter-individual variability in response to cancer therapies [56].

Proteins, as the functional products of gene expression, are central to cellular processes and the development of cancer. It has already been hypothesized that proteomic profiling more directly addresses biological and pharmacologic problems in cancer [57]. In recent years, proteomics efforts have led to proteins that can serve as cancer biomarkers. Several lines of evidence have shown that the expression level of proteins is frequently associated with drug response. One example is MRP1, which is associated with drug resistance or poor patient outcomes in a variety of cancers [58]. MRP3 is the ABC transporter that is most closely related to MRP1. For both MRP3 and MRP1, their protein expression levels correlated with decreased sensitivity of lung cancer cell lines to doxorubicin [59]. Another well-characterized example is eight protein signatures that were identified for the prediction of drug response to 5-fluorouracil, including CDH1, CDH2, KRT8, ERBB2, MSN, MVP, MAP2K1, and MGMT. All of these proteins, except for KRT8, are involved in the pathogenesis of colon cancer [60].

To identify miRNAs and proteins that can serve as drug-specific survival biomarkers, we first performed survival analysis on cancer-stratified and drug-stratified subpopulations of patients, and identified miRNAs and proteins that significantly correlated with survival outcome in a drug-specific manner. Further, in order to assess the prognostic power of miRNA and protein markers, we investigated the regulatory mechanisms of significant miRNAs by examining their target genes, and of identified protein markers by analyzing their corresponding coding genes. Moreover, we performed a literature survey to assess whether there are existing evidences that supported the correlations between the identified miRNAs and drug responses, as well as proteins and drug responses. Finally, we examined several additional breast cancer datasets that provided miRNA expression and survival outcomes, and showed that our drug-specific miRNA survival markers for breast cancer were able to effectively stratify the prognosis of patients in those additional datasets.

2.2 Methods

2.2.1 Data access

miRNA, protein, and gene expression data were downloaded from Genomic Data Commons (GDC) using the GDC Data Transfer Tool. These expression data covered more than 11,000 patients across 33 cancer types. Clinical data were also downloaded from GDC, which included patients' drug treatment records and survival outcomes. The drug treatment data contains 10,863 treatment entries for 4,328 patients. The drug names in the treatment records were standardized to remove inconsistency in terms of naming conventions and spelling errors. The standardization was performed based on a manually curated list previously created by our group. The survival data contains the survival outcome for 11,082 patients across 33 cancer types. After removing duplicates in the molecular data and filtering for samples with treatment and survival data, we narrowed down to a total of 9,559 patients across 31 cancer types and 264 unique drugs in this study.

To examine the functional relevance of the identified miRNA markers, we obtained experimentally verified miRNA-target pairs from four databases including mir2Disease, miRecords, TarBase and miRTarBase. The miRNA-target pairs obtained from the four databases include: 96 pairs from mir2Disease, 518 pairs from miRecords, 26388 pairs from TarBase, and 50381 pairs from miRTarBase. These amount to a total of 57,863 human specific miRNA-target gene pairs, involving 14,652 genes and 579 miRNAs.

We also downloaded and log-transformed independent miRNA expression datasets from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) database to assess the prognostic power of identified drug-specific miRNAs. A total of 490 patients from 3 independent datasets were analysed in this study. Detailed information about these 3 datasets is shown in Table 2.1.

Table 2.1: Predictive miRNAs identified in BRCA-drug groups

Cancer type	Drug	miRNA
BRCA	Doxorubicin	miR-20b
	Cyclophosphamide	miR-20b; miR-363; miR-628; miR-7.2
	Docetaxel	miR-577
	Tamoxifen	miR-577
	Anastrozole	miR-3677; miR-577

2.2.2 Data preprocessing

The miRNA expression data and gene expression data downloaded from TCGA have been normalized by FPKM-UQ [61], and we subsequently transformed the expression data by log-transformation. The protein expression data available from TCGA have already been properly normalized and transformed.

For each miRNA, protein, and gene feature, we used StepMiner [62] to compute a global threshold based on all patients across all cancer types. We first sorted the expression data of a given feature for all patients and then fitted a step function to minimize the square error between the original and the fitted values. Since this threshold is derived based on the data of all patients across all cancer types, it is able to robustly define high and low expression. The threshold of each miRNA, protein, or gene was used to binarize the data, so that the patients could be divided into two classes (highly-expressed class vs. lowly-expressed class) according to the binarized data of an individual feature.

2.2.3 Survival analysis

We grouped patients based on cancer type and drug exposure. Patients with a given cancer and exposed to a given drug were assigned into one cancer-drug group. Within each cancer-drug group, patients were further stratified into highly expressed or lowly expressed classes for each miRNA or protein based on its binarized data. Log-rank test was used to assess the statistical significance of the survival difference between highly- and lowly-expressed

classes. P-values of the log-rank test were adjusted for multiple testing using Benjamini-Hochberg method with a false-discovery rate (FDR) <0.1. Kaplan-Meier analysis and log-rank test were performed using the R package 'survival'.

2.2.4 Literature survey

A literature search was performed using PubMed, accessed via the National Library of Medicine PubMed interface (http://www.ncbi.nlm.nih.gov/pubmed). We programmatically searched the PubMed database with custom Python scripts, using keywords of drugs AND identified drug-specific miRNAs or proteins. We searched through PubMed for all keywords in all filed, including the title, abstract and main texts of the articles.

2.3 Results

2.3.1 Identify miRNAs predictive of survival in cancer-drug groups

We took patients who suffered from the same cancer type and were exposed to the same drug, and assigned them into one cancer-drug group. A patient can be a member of multiple cancer-drug groups, if the patient received multiple different drugs as reflected in the drug treatment records. The cancer-drug groups containing at least 15 patients with miRNA expression data available were selected for the subsequent analysis to identify miRNA markers for drug-specific survival. A total of 110 cancer-drug groups were selected, which involved 23 cancer types and 43 different drugs. The heatmap in Figure 2.1 shows the number of patients in each of the 110 cancer-drug groups.

For each of the 110 cancer-drug groups, we applied survival analysis for the expression data of each of the 1,881 miRNAs measured in TCGA, which amounted to a total of 206,910 miRNA-cancer-drug combinations examined. For each miRNA, patients in the cancer-drug group were stratified into a highly-expressed class and a lowly-expressed class by applying StepMiner [63] to the expression data of the miRNA. Specifically, we sorted the expression data across all patients for an miRNA, and then fitted a step function to the

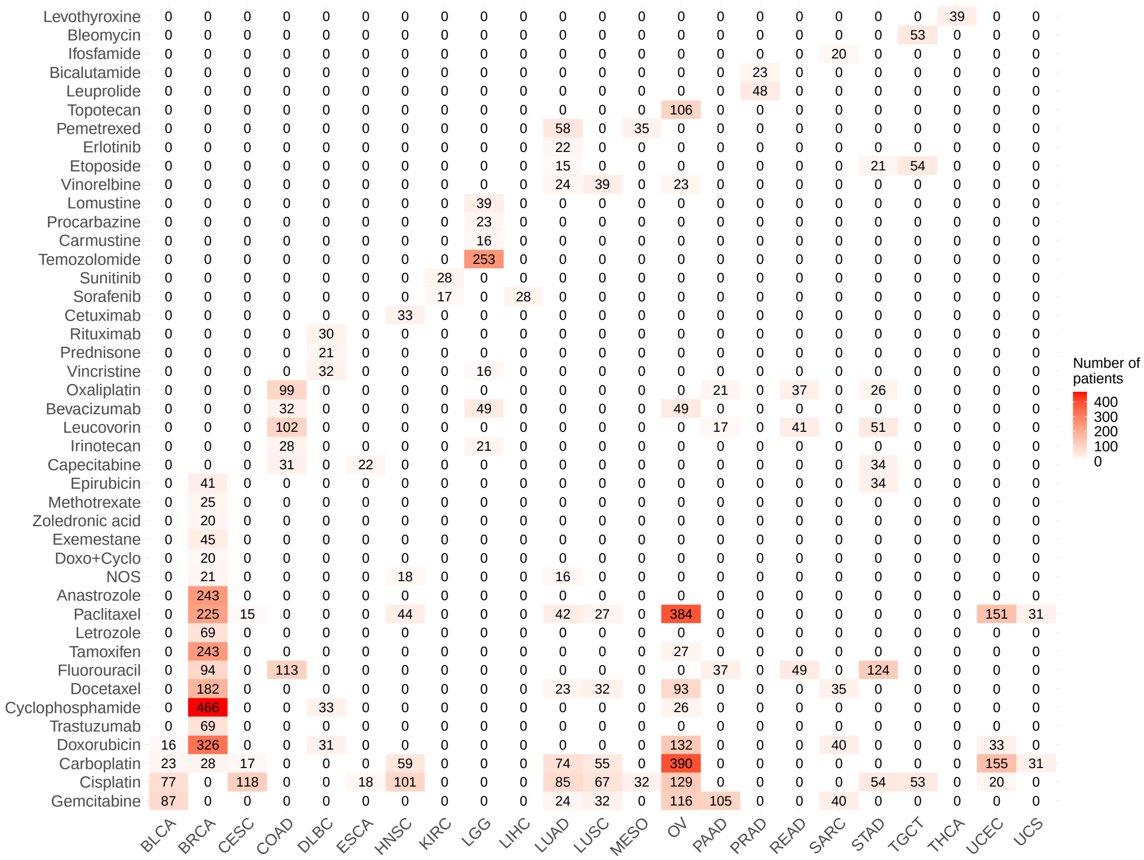

Figure 2.1: Heatmap of number of patients in each cancer-drug group.

sorted data that minimizes the square error between the original and the fitted values, which provides a global threshold to binarize the expression of the miRNA. Since this threshold is derived based on the data of all patients across all cancer types, it is able to robustly define high and low expression of the miRNA. In order to avoid situations where the sample size is too small to extrapolate significant findings, we only performed survival analysis on those miRNAs with at least 5 low-expressed patients and 5 high-expressed patients within the cancer-drug group, which excluded 16,535 out of 206,910 miRNA-cancer-drug combinations in our analysis. Log-rank test was applied to evaluate whether there is significant difference in survival between the two classes. Benjamini-Hochberg multiple tests were used to calibrate the False Discovery Rate (FDR). MiRNAs with adjusted FDR < 0.1 were selected as predictive markers whose expression levels were predictive of patients' survival outcome in the corresponding cancer-drug groups. In order to assess whether the identified

miRNAs are specifically associated to the drug, we performed the same analysis on all patients in the corresponding cancer type, and only selected the miRNA markers significant in the cancer-drug groups, but not significant in the corresponding cancer type.

In total, we identified 115 significant miRNA-cancer-drug combinations in 44 cancer-drug groups Table 2.2 with an FDR threshold of < 0.1, which involved 71 unique miRNAs that were significant in at least one cancer–drug group. All the identified miRNAs significantly associated with survival outcome of specific cancer in a drug-specific manner. We found 10 significant miRNAs in the Sarcoma (SARC)-gemcitabine group and 9 significant miRNAs in the Ovarian serous cystadenocarcinoma (OV)-paclitaxel group. 31 identified miRNAs were found to be significant in more than one cancer-drug group. These miRNAs are able to predict the survival of patients with different cancers and treated with different drugs, which may be promising biomarkers for drug response in multiple cancer contexts. Among these 31 repeated miRNA markers, two miRNAs were significantly associated with survival in four cancer-drug groups. We observed that the expression level of miR-577 was related to the outcomes of patients with Breast invasive carcinoma (BRCA) and treated with docetaxel, tamoxifen, anastrozole, and NOS, respectively. Another observation was miR-576, which was found to be associated with drug response to fluorouracil, leucovorin in Rectum adenocarcinoma (READ), and gemcitabine, docetaxel in SARC.

Table 2.2: Significant miRNAs identified in cancer-drug groups.

Cancer	**Drug**	**miRNA**	**Literature**	**PubMed**
BLCA	Carboplatin	1	1	28459431
	Cisplatin	2	4	25843291;26801660;27554045; 30885939
	Gemcitabine	5	5	32104074;26036346;22132977;

Table 2.2 – continued from previous page

Cancer	Drug	miRNA	Literature	PubMed
				26606261;30887178
BRCA	Anastrozole	2	0	
	Carboplatin	3	1	25824335
	Cyclophosphamide	4	1	30824586
	Docetaxel	1	0	
	Doxorubicin	1	0	
	Epirubicin	1	0	
	Fluorouracil	3	3	32162886;22955854;24460313
	Methotrexate	1	0	
	NOS	5	0	
	Tamoxifen	1	0	
CESC	Cisplatin	2	9	31647948;28751441;21248297; 22939244;24462518;25843291; 26801660;27554045;30885939
COAD	Fluorouracil	1	0	
	Oxaliplatin	1	1	34477047
HNSC	Carboplatin	4	3	31487677;31886905;1411628
	Cetuximab	4	0	
	Paclitaxel	3	5	31352515;27338043;27338042; 31063487;31496800

Table 2.2 – continued from previous page

Cancer	Drug	miRNA	Literature	PubMed
LGG	Bevacizumab	1	0	
	Carmustine	1	0	
	Lomustine	1	0	
	Temozolomide	3	0	
LUAD	Docetaxel	2	0	
	Gemcitabine	1	0	
LUSC	Gemcitabine	6	3	31632071;21106054;27129291
MESO	Cisplatin	1	0	
OV	Carboplatin	3	3	27873337;23229111;24314246
	Cyclophosphamide	1	0	
	Doxorubicin	1	0	
	Gemcitabine	4	1	32765654
	Paclitaxel	9	4	25155039;29632436;25973036; 24060847
	Topotecan	1	0	
PRAD	Leuprolide	1	0	
READ	Fluorouracil	2	0	
	Leucovorin	2	0	
SARC	Docetaxel	5	0	
	Gemcitabine	10	6	31563901;32104074;26036346;

Table 2.2 – continued from previous page

Cancer	Drug	miRNA	Literature	PubMed
				22132977;26606261;30304549
	Ifosfamide	1	0	
STAD	Leuprolide	7	1	27840964
TGCT	Cisplatin	2	3	29565481;32440152;28096802
	Etoposide	2	0	
UCEC	Carboplatin	2	1	29899543
	Paclitaxel	1	0	

Table 2.2: This table represents the number of significant miRNAs identified in all cancer-drug groups. The Literature column represents the number of literature found by PubMed search for each combination.

2.3.2 Literature survey of identified miRNA markers

The significant drug-specific survival miRNA markers included miRNAs that have been previously implicated, as well as the novel ones that have never been mentioned in the literature. The PubMed database was used to search for potentially relevant studies for each of the 115 identified miRNA-cancer-drug combinations, and we have found supportive literature for multiple predictive miRNAs associate with drug responses in various cancer types. The number of papers linked to each identified miRNA-cancer-drug combination is shown in Table 2.2. For example, two previous studies [64, 65] have reported that the expression of miR-330-5p is related to the promotion of gemcitabine response, one in the context of pancreatic cancer and the other in the context of colon cancer. Coincidentally, we found that increased miR-330 was associated with prolonged survival outcomes of Bladder urothelial carcinoma (BLCA) patients exposed to gemcitabine (log-rank, p-

value= 0.003), and worse survival outcomes of SARC patients exposed to gemcitabine (log-rank, p-value$<$ 0.001)(Figure 2.2). As another example, our analysis suggested that low expression of

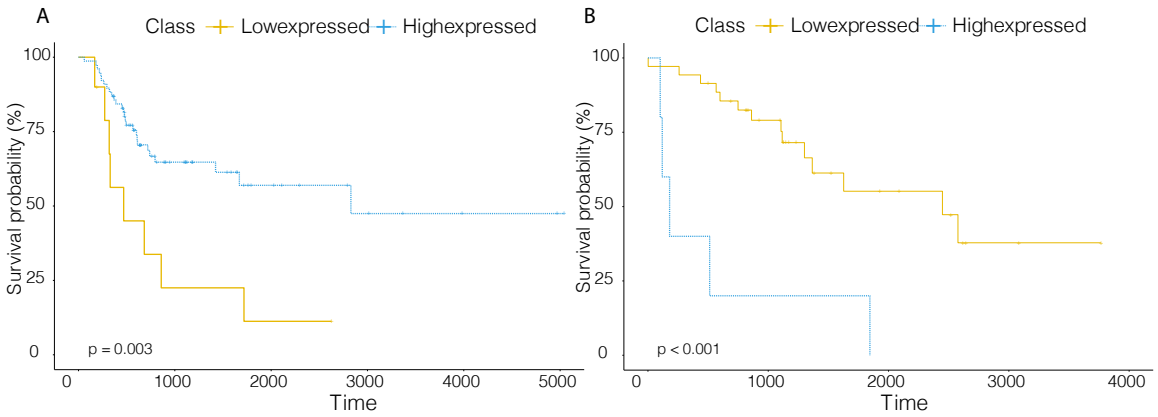

Figure 2.2: Kaplan-Meier curves of overall survival for patients treated with gem-citabine at low or high expressed classes stratified by miR-330 in the BLCA or SARC.

miR-296 led to increased cisplatin sensitivity in Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) (log-rank, p-value= 0.002) and increased gemcitabine sensitivity in SARC (log-rank, p-value= 0.031), as showed in Figure 2.3. This is supported by several previous

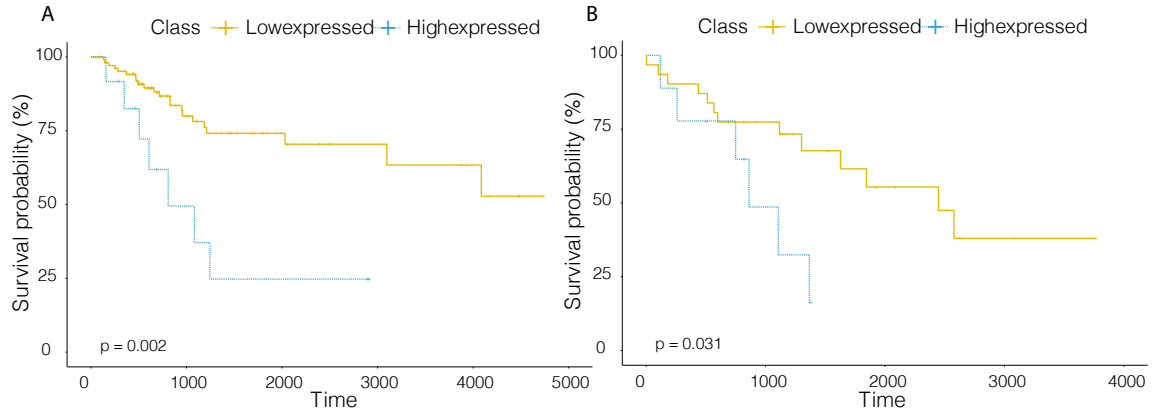

Figure 2.3: Kaplan-Meier curves of overall survival for patients treated with cis-platin or gemcitabine at low or high expressed classes stratified by miR-296 inthe CESC or SARC.

studies on the role of miR-296 in drug response [66, 67, 68, 69]. One of those studies indicated that miR-296 was able to sensitize Lung adenocarcinoma (LUAD) cells to cisplatin in vitro and in vivo [70]. The inhibition of miR-296 also resulted in increased chemosensitivity in Esophageal carcinoma (ESCA) cell lines to standard chemotherapeutic agents such as 5-fluorouracil and cisplatin. Another study proved that miRNA-296-5p could promote drug resistance by targeting Bcl2-related

ovarian killer, leading to a poor prognosis in pancreatic cancer [71]. As a third example, we found that the over-expression of miR-483 resulted in decreased cisplatin sensitivity in BLCA and CESC (log rank, p-value= 0.003; p-value< 0.001) (Figure 2.4). This is consistent with previous studies

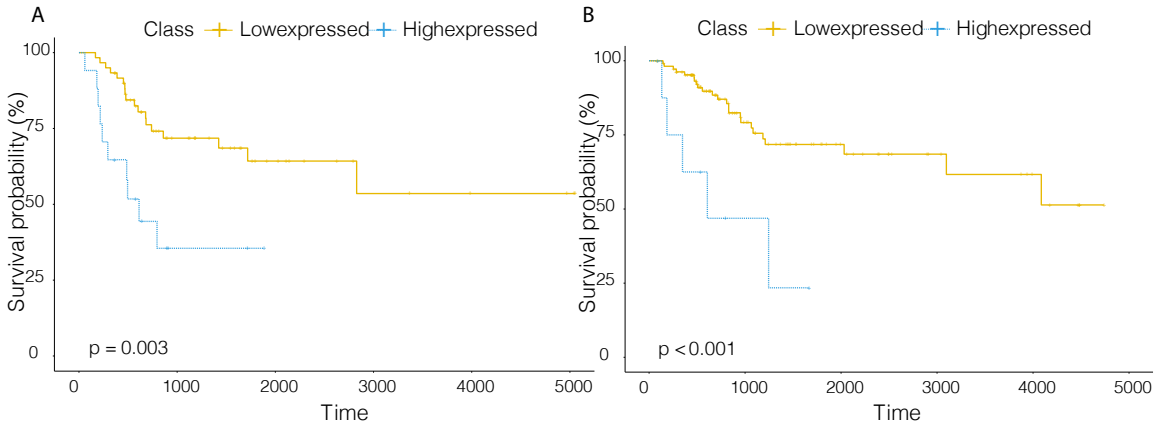

Figure 2.4: Kaplan-Meier curves of overall survival for patients treated with cis-platin at low or high expressed classes stratified by miR-483 in the BLCA or CESC.

showing that miR-483-5p is significantly associated with cisplatin sensitivity and with overall survival in patients with Tongue squamous cell carcinoma (TSCC) [72, 73]. In addition, there has been evidence that miR-483-3p expression is increased in platinum drug-resistant ovarian carcinoma cells and indicated an association between increased miR-483-3p expression and cisplatin resistance [74, 75].

2.3.3 Correlation between drug-specific miRNAs and their target genes

To examine the correlation between the identified drug-specific miRNAs and their associated genes in the cancer-drug groups where the miRNAs were identified, we performed drug-specific survival analysis on the expression data of genes that are targets of the identified miRNAs. For each of the 115 identified miRNA-cancer-drug combinations, we extracted the gene expression data of their target genes for patients in that cancer-drug group. We binarized the gene expression data using the StepMiner, the same method we used to binarize the miRNA expression data. The target genes were determined according to the experimentally verified miRNA-target gene pairs incorporated from multiple databases [76, 77, 78, 79]. Similar to the analysis of miRNAs, for each cancer-drug group and each gene under consideration, we stratified patients in the cancer-drug group to

highly-expressed and lowly-expressed classes according to the binarized data for the gene, and only performed survival analysis on gene-cancer-drug combinations with at least 5 lowly-expressed patients and 5 highly-expressed patients. A total of 8,513 genes targeted by identified miRNAs were examined in the corresponding cancer-drug groups. Log-rank test was performed to examine whether each of the target genes is significantly correlated with survival outcome in the corresponding cancer-drug groups. We applied FDR adjusted significance p-value < 0.01 as thresholds to identify genes whose expression were also predictive of drug-specific survival, in the same context as its associated miRNA. 48 gene-cancer-drug combinations exceeded the FDR threshold, which involved 42 unique genes and 9 cancer-drug groups. Based on the 48 significant gene-cancer-drug combinations, $\sim$37% of the identified miRNA-cancer-drug combinations showed significance in their corresponding genes, leading to 59 miRNA-mediated gene-cancer-drug combinations in total, as shown in Table 2.3. In order to examine the biological role of the 42 target genes, we performed Gene Ontology (GO) [80] enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) [81] pathway enrichment analysis by using The Database for Annotation, Visualization and Integrated Discovery (DAVID) [82], searching against the entire Homo sapiens genome. We used a P-value threshold of ¡0.05 to identify enriched processes and pathways. However, due to the small number of target genes identified, this functional enrichment analysis may not have sufficient statistical power to be conclusive, even though it was encouraging to observe the target genes were involved in relevant biological processes including ”developmental growth”, ”beta-catenin-TCF complex assembly”, and ”negative regulation of transcription from RNA polymerase II promoter”, as well as general pathways such as ”MicroRNAs in cancer” and ”Pathways in cancer”.

In each of these 59 miRNA-mediated gene-cancer-drug combinations, since both the miRNA and the target gene were predictive of survival difference in the corresponding cancer-drug group, these combinations may reflect the functional role of the miRNA markers in affecting drug response through modulating the expression of their target genes. Interestingly, we observed miR-18a and its target gene SMAD4 were predictive of carboplatin response in Head and Neck squamous cell carcinoma (HNSC), which was consistent with our previous study on the relationship between SMAD4 and carboplatin response [83]. Additionally, we also identified miR-18a to be predictive of the prognosis of HNSC patients treated with cetuximab. In the literature, it has been showed that SMAD4

down-regulation leads to cetuximab resistance in HNSC cell lines [84], therefore, we can reasonably hypothesize that the miR18a-SMAD4 axis may be involved in the resistance mechanism of cetuximab response in HNSC.

Table 2.3: Significant miRNA-cancer-drug combinations, and corresponding target genes that are also predictive of survival outcoms.

Cancer Type	Drug	Gene Symbol	FDR	miRNA	FDR
BRCA	Carboplatin	TRIM4	5.392E-03	miR-96	1.004E-03
	Carboplatin	MED1	5.392E-03	miR-96	1.004E-03
HNSC	Carboplatin	SMAD4	2.811E-03	miR-18a	8.442E-02
	Carboplatin	SON	2.811E-03	miR-18a	8.442E-02
	Carboplatin	TNRC6B	1.458E-03	miR-18a	8.442E-02
	Carboplatin	PRR12	2.954E-04	miR-18a	8.442E-02
	Carboplatin	DPF1	6.489E-03	miR-18a	8.442E-02
	Carboplatin	SENP1	2.376E-03	miR-18a	8.442E-02
	Carboplatin	CACNA2D3	8.704E-03	miR-18a	8.442E-02
	Carboplatin	CIC	4.742E-03	miR-18a	8.442E-02
	Carboplatin	UBXN7	6.489E-03	miR-18a	8.442E-02
	Carboplatin	NFIC	6.145E-03	miR-18a	8.442E-02
	Carboplatin	NIPBL	4.624E-03	miR-18a	8.442E-02
	Carboplatin	EDF1	8.793E-03	miR-18a	8.442E-02
	Carboplatin	HNRNPUL1	1.458E-03	miR-18a	8.442E-02
	Carboplatin	TLE3	1.458E-03	miR-18a	8.442E-02
	Carboplatin	TCP1	8.234E-03	miR-18a	8.442E-02
	Carboplatin	TRAPPC1	6.145E-03	miR-18a	8.442E-02
	Carboplatin	SON	4.128E-03	miR-744	1.568E-02
	Carboplatin	TNRC6B	1.673E-03	miR-744	1.568E-02
	Carboplatin	PSMD4	7.077E-03	miR-744	1.568E-02
	Carboplatin	PRR12	2.712E-04	miR-744	1.568E-02

Table 2.3 – continued from previous page

Cancer Type	Drug	Gene Symbol	FDR	miRNA	FDR
	Carboplatin	CIC	7.077E-03	miR-744	1.568E-02
	Carboplatin	TLE3	1.673E-03	miR-744	1.568E-02
	Carboplatin	TNRC6B	5.805E-04	miR-1293	3.407E-02
	Paclitaxel	KRAS	4.355E-04	miR-18a	9.724E-02
	Paclitaxel	DICER1	6.334E-03	miR-18a	9.724E-02
	Paclitaxel	TNRC6B	3.988E-04	miR-18a	9.724E-02
	Paclitaxel	MED13L	1.747E-03	miR-18a	9.724E-02
	Paclitaxel	TRAPPC10	2.869E-03	miR-18a	9.724E-02
	Paclitaxel	CACNA2D3	1.295E-03	miR-18a	9.724E-02
	Paclitaxel	NIPBL	5.656E-04	miR-18a	9.724E-02
	Paclitaxel	EDF1	3.024E-04	miR-18a	9.724E-02
	Paclitaxel	TLE3	5.656E-04	miR-18a	9.724E-02
	Paclitaxel	RPL9	4.617E-04	miR-18a	9.724E-02
	Paclitaxel	ZNF770	8.679E-03	miR-18a	9.724E-02
	Paclitaxel	MATR3	6.334E-03	miR-18a	9.724E-02
	Paclitaxel	TRAPPC1	3.024E-04	miR-18a	9.724E-02
	Paclitaxel	HERC1	8.679E-03	miR-18a	9.724E-02
	Paclitaxel	TNRC6B	3.103E-04	miR-362	5.662E-02
	Paclitaxel	MED13L	1.359E-03	miR-362	5.662E-02
	Paclitaxel	HERC1	6.100E-03	miR-362	5.662E-02
	Paclitaxel	TNRC6B	1.724E-04	miR-1293	2.016E-02
	Paclitaxel	KMT2D	9.635E-03	miR-1293	2.016E-02
	Paclitaxel	ATXN2L	8.052E-03	miR-1293	2.016E-02
LGG	Bevacizumab	TMEM205	2.373E-04	miR-935	1.084E-03
	Bevacizumab	CREBBP	4.019E-04	miR-935	1.084E-03
	Carmustine	SHCBP1	8.172E-03	miR-33a	9.814E-02
	Temozolomide	MAFB	1.541E-03	miR-135b	1.196E-03

Table 2.3 – continued from previous page

Cancer Type	Drug	Gene Symbol	FDR	miRNA	FDR
	Temozolomide	SLC7A1	1.541E-03	miR-135b	1.196E-03
	Temozolomide	CHPF2	1.541E-03	miR-671	6.440E-02
	Temozolomide	CEP85	3.072E-04	miR-671	6.440E-02
	Temozolomide	ZFP36L2	5.837E-04	miR-671	6.440E-02
	Temozolomide	SLC7A1	1.761E-03	miR-671	6.440E-02
	Temozolomide	R3HDM2	1.781E-04	miR-671	6.440E-02
OV	Paclitaxel	PPIL1	7.804E-03	miR-504	4.365E-02
	Paclitaxel	CASP2	4.042E-03	miR-130b	7.904E-02
PRAD	Leuprolide	RRP1B	1.011E-04	miR-331	1.696E-02
SARC	Docetaxel	GNAS	1.666E-04	miR-331	4.551E-02

Table 2.3: This table represents the miRNA-mediated gene-cancer-drug combinations in all cancer-drug groups.

2.3.4 Prognostic performance of drug-specific miRNAs in independent datasets

The prognostic power of the identified drug-specific miRNA biomarkers in survival prediction was further tested in several independent miRNA expression datasets generated by microarray platforms. When searching for independent datasets, we considered cancer-drug groups that contained at least 150 patients and produced at least 1 significant miRNA survival marker. Four cancer types were considered, which were BRCA, Brain Lower Grade Glioma (LGG), OV and Uterine Corpus Endometrial Carcinoma (UCEC). After an extensive search on GEO for datasets that contained both miRNA expression data and survival outcome data of these four cancer types, three appropriate datasets were found (GSE40267, GSE19783, and GSE22220), all of which were on BRCA. Using these datasets, we examined the drug-specific miRNAs we identified in BRCA-Doxorubicin, BRCA-Cyclophosphamide, BRCA-Docetaxel, BRCA-Tamoxifen, and BRCA-Anastrozole groups. In these five BRCA-drug groups, 6 drug-specific miRNAs were identified, which were miR-20b,

miR-363, miR-628, miR-7.2, miR-577, and miR-3677 as shown in Table 2.4. Since expression data for miR-7.2 and miR-3677 were not available in these three datasets, we only examined the remaining 4 drug-specific miRNAs for their ability to predict survival outcomes. For each dataset, we performed survival analysis for each of the 4 drug-specific miRNAs. Patients in the datasets were divided into highly-expressed class and lowly-expressed class by the median of expression value of individual miRNAs, and log-rank test was applied to examine the ability of the drug-specific miRNA markers for stratifying patients with different prognosis.

Table 2.4: Additional GEO datasets that served as independent validation datasets

Datasets	Platform	Number of patients	Outcome
GSE40267	GPL8227	181	OS
GSE19783	GPL10850	101	DFS
GSE22220	GPL8178	210	DRFS

Abbreviations: OS, Overall survival; DFS, Disease-free survival; DRFS, Distant-relapse free survival.

Specifically, in the GSE40267 dataset, stratification of patients by miR-363 expression was significantly correlated with patients' Overall survival (OS) (Figure 2.5A, log-rank test, p-value = 0.01), where patients with high miR-363 expression had significantly longer overall survival than patients with low expression of miR-363. In the GSE19783 dataset, Disease-free survival (DFS) was used as the survival endpoint. High expression of miR-20b and miR-628 were significantly associated with improved survival outcome (Figure 2.5B-C, log-rank test, p-value = 0.021 and 0.026, respectively). In the GSE22220 dataset with Distant-relapse free survival (DRFS) as the survival endpoint, miR-577 expression was significantly correlated with distant-relapse free survival, where low expression of miR-577 resulted in prolonged survival outcomes of patients (Figure 2.5D, log-rank test, p-value = 0.003). Overall, Figure 2.5 shows the Kaplan–Meyer curves comparing highly- vs. lowly-expressed patients according to these 4 drug-specific miRNAs. Among the 4 drug-specific miRNAs examined in the three independent datasets, all of them turned out to be significantly predictive of survival outcomes.

2.3.5 Significant proteins predictive of drug-specific survival

Similar to our analysis of miRNA, we grouped patients with the same cancer and receiving the same drug for the protein expression data. To ensure robust statistical analysis, we imposed a

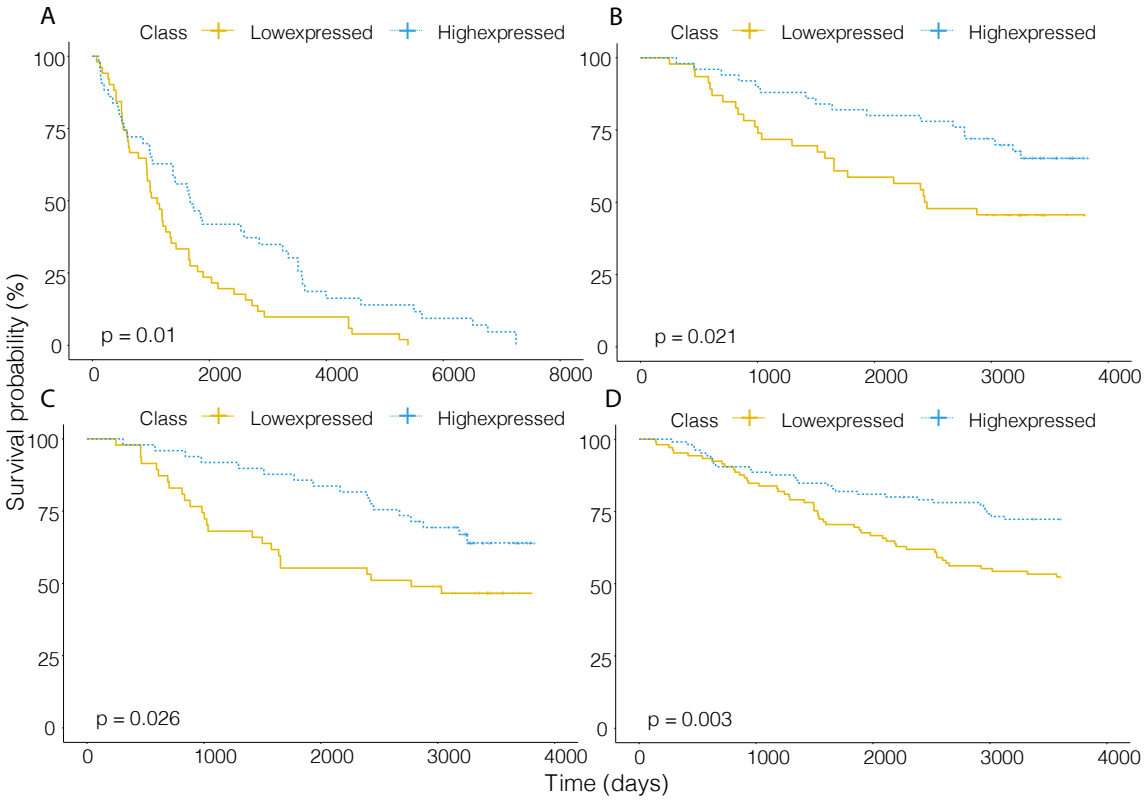

Figure 2.5: Kaplan–Meyer survival curves of BRCA patients in additional GEOdatasets, stratified by drug-specific miRNA marker identified from TCGA data.(A) BRCA patients in GSE40267 stratified by miR-363. (B) BRCA patients inGSE19783 stratified by miR-20b. (C) BRCA patients in GSE19783 stratified by miR-628. (D) BRCA patients in GSE22220 stratified by miR-577.

minimum sample size requirement of 15, including only cancer-drug groups with more than 15 patients. Therefore, a total of 98 cancer-drug groups were considered for the subsequent analysis to identify protein markers for drug-specific survival. Next, we binarized the protein expression data in TCGA, which was needed in our survival analysis. For each of the 336 proteins measured by TCGA, we applied StepMiner [85] to binarize its expression data across all patients in all cancer types. Finally, we performed survival analysis to evaluate each protein's ability to predict the survival outcome of patients in each cancer-drug group. Patients in the cancer-drug group were stratified into a highly-expressed class and a lowly-expressed class based on the binarized data of the protein. To minimize undesired statistical bias, we only performed survival analysis on proteins in cancer-drug groups with at least 5 lowly-expressed patients and 5 highly-expressed patients.

In total, 17,812 protein-cancer-drug combinations were tested in our analysis, which involved

23 cancer types and 41 drugs. We applied log-rank test to determine the statistical significance of survival difference between highly-expressed class and lowly-expressed class. 90 proteins exceeding a FDR threshold of < 0.1 were selected as predictive markers whose expression levels were related to patients' survival outcome in a drug-specific manner. In order to identify proteins that are specifically related to individual drugs, we performed the same analysis on all patients in each cancer type, and identified proteins that are predictive of cancer-specific survival irrespective of drug treatment. Among the 90 proteins significant for drug-specific survival, 25 were also identified in the cancer-specific analysis. In our subsequent analysis, we excluded these 25, and only included the protein markers that were significant in cancer-drug groups but not significant in the corresponding cancer types.

As showed in Table 2.5, a total of 65 significant protein-cancer-drug combinations were identified in 21 cancer-drug groups, which involved 55 unique proteins. We found 13 significant proteins predictive of cisplatin response in CESC, and 9 protein markers are associated with fluorouracil response in Pancreatic adenocarcinoma (PAAD). Interestingly, there are 11 proteins that turned out to be significant in multiple cancer-drug groups, which may potentially serve as key biomarkers to drug responses in multiple cancer types. Among the proteins that were significant in more than one cancer-drug group, we observed that CDH1 was related to the sensitivity of cisplatin in CESC and also associated with the overall survival of Doxorubicin-treated patients in OV. We also found that CDK1 was correlated with drug response to gemcitabine in BLCA, doxorubicin in BRCA, and vinorelbine in OV.

2.3.6 Literature support of predictive protein markers

To assess whether there are previous research that supported the identified protein markers predictive of drug response, we conducted a comprehensive literature survey on the PubMed database for each of the 65 protein-drug combinations. We found supportive evidence for multiple protein-drug combinations in various cancer contexts.

In particular, our analysis suggested that high-expressed CDKN1B was able to increase drug response to gemcitabine in pancreatic adenocarcinoma (PAAD). This is consistent with previous studies that the re-expression of CDKN1B was related to the sensitization of pancreatic cancer cells

Table 2.5: 65 Significant protein-cancer-drug combinations identified in cancer-drug groups

Cancer-drug groups	**Protein markers for drug-specific survival**	**Number of patients**
BLCA-Gemcitabine	YWHAE,AKT1-3,CDK1,MAPK1	75
BRCA-Carboplatin	AKT1	24
BRCA-Doxorubicin	CDK1	282
CESC-Cisplatin	TP53BP1,Tubulins,BAP1,CTNNB1,COL6A1, CDH1,EIF4G1,HSPA1A,KU80,MRE11,CDH2, SERPINE1,TSC2	61
COAD-Oxaliplatin	EIF4EBP1	82
CES-Carboplatin	BCL2,CLDN7,FOXM1,MSH2	22
KIRC-Sorafenib	KDR	16
LGG-Bevacizumab	EIF4EBP1,ETS1,FASN,TIGAR,TSC2	41
LGG-Irinotecan	ETS1,KU80,SRSF1,FYN	20
LGG-Lomustine	CTNNA3,AR,CDKN1A,BRAF	31
LUAD-Cisplatin	NF2	66
LUSC-Gemcitabine	ABL1	26
LUSC-Cisplatin	CCNE2,PEA15	54
OV-Docetaxel	CCNE1,RBM15	69
OV-Carboplatin	CASP7,RICTOR	290
OV-Doxorubicin	CDH1,GAB2,RBM15	106
OV-Vinorelbine	AKT1-3,CDK1,ERCC5	19
PAAD-Gemcitabine	CDKN1B,MAPK11	73
PAAD-Fluorouracil	BECN1,GAPDH,ERBB3,CDKN1B,MAPK11, MAPK12,PTEN,SRC,PARP1	24
STAD-Cisplatin	DPP4	48
STAD-Etoposide	INPP4B	19

to gemcitabine leading to a significant induction of apoptosis, which could be a superior potential treatment for pancreatic cancer [86, 87]. Another literature support is about PTEN. PTEN was first discovered as a tumor suppressor, and its loss of function is strongly associated with tumor growth and survival. Figure 2.6A shows how PTEN expression correlated with PAAD patients in TCGA, that PTEN over-expression led to increased sensitivity of fluorouracil. This observation is supported

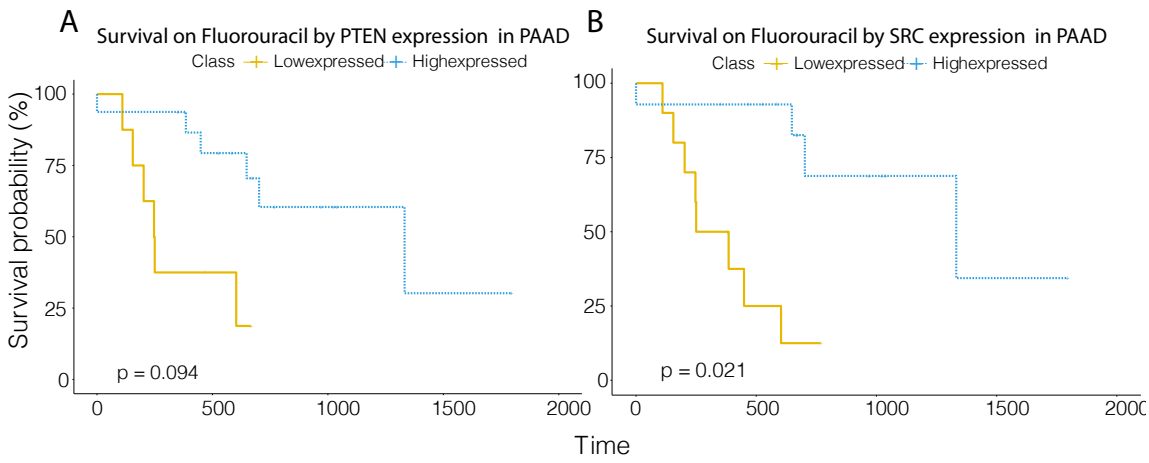

Figure 2.6: Kaplan-Meier curves of overall survival for patients treated with fluorouracil at low or high expressed classes stratified by PTEN or SRC in the PAAD.

by previous studies which showed that PTEN was involved in promoting 5-Fluorouracil-induced apoptosis, and the reduced expression of PTEN was associated with increased malignancy grade in PAAD, whereas maintenance of PTEN expression showed a trend toward a longer survival [88]. In addition, it has been shown that the inhibition of TAP subsequently promoted the expression of PTEN that increase sensitivity to chemotherapeutic agents in cancer [89]. A third example is VEGFR2, which was previously reported to be predictive of sorafenib efficacy in patients with metastatic renal cell carcinoma (mRCC) and was associated with longer overall survival of patients those treated with sorafenib [90]. In our analysis, we found that the repressed VEGFR2 resulted in prolonged survival outcomes of patients exposed to sorafenib in Kidney renal clear cell carcinoma (KIRC), which reveals the potential prediction of VEGFR2 on gemcitabine in other diseases.

We also found literature that showed inconsistent direction of survival impact compared to two of the protein-drug combinations we identified. Our analysis suggested that decreased SRC was related to poor overall survival of patients in PAAD and treated with Fluorouracil, shown in Figure 2.6B. However, a recent study that involved fluorouracil and a few other drugs showed that SRC expression up-regulation in some Pancreatic ductal adenocarcinoma (PDAC) patients was associated to relatively poor patient outcome [91]. The second inconsistency was related to BCL2. Our analysis demonstrated that the over-expression of BCL2 resulted in better survival outcomes of patients with HNSC and exposed to carboplatin (Figure 2.7A), and the high-expressed BCL2-coding gene was also associated with prolonged overall survival of patients with HNSCC (Figure 2.7B).

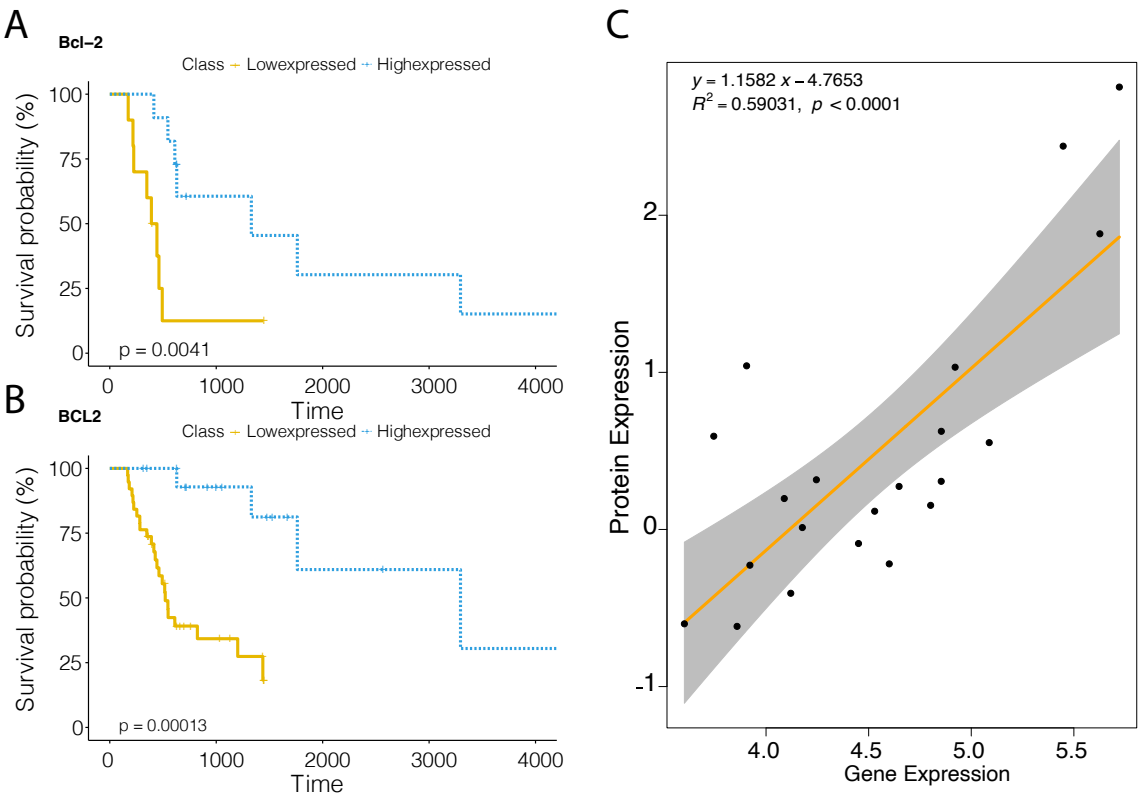

Figure 2.7: Correlation between Bcl-2 and BCL2 in expression level and survival outcomes. A-B. Kaplan-Meier curves of overall survival for patients treated with fluorouracil at low or high expressed classes stratified by Bcl-2 or BCL2 in the HNSCC. C. Consistency of Bcl-2 and BCL2 correlation in expression levels for patients in HNSCC-carboplatin group.

In contrast, a previous study observed that BCL2 could inhibit apoptosis induced by cisplatin, carboplatin and paclitaxel, making HNSCC that express BCL2 resistant to rapamycin, carboplatin and paclitaxel [92]. Despite of these inconsistencies in the direction of correlation with survival, the literature did indicate the relevance of our identified proteins to drug responses in cancer patients.

2.3.7 Correlation between predictive proteins and their coding genes

To understand the roles of drug-specific proteins during carcinogenesis and pharmacotherapy, we investigated the regulatory mechanism of identified protein markers by examining their corresponding coding genes. We performed survival analysis on the genes coded the protein markers, in the same cancer-drug context where the protein markers were identified. Specificity, for each of the 65 identified protein-cancer-drug combinations, we extracted the binarized gene expression data of the

corresponding gene for patients in that cancer-drug group, stratified the patients to high-expressed and low-expressed classes according to the binarized gene expression data, and performed log-rank test to examine whether there is a significant difference in survival outcome between the two classes. We applied p-value threshold of < 0.05 to identify genes whose expression were also predictive of drug-specific survival outcomes. Similar to our analysis of proteins, survival analysis was only performed on the corresponding genes if there were at least 5 highly-expressed patients and 5 lowly-expressed patients in the corresponding cancer-drug group. 7 genes were identified whose expression were also predictive of drug-specific survival, in the same context as its associated protein-cancer-drug combinations. Therefore, this result suggests $\sim$11% of the identified proteins also showed significance in their corresponding genes.

To elucidate the relationships between significant proteins and their corresponding genes in each cancer-drug group, we examined the correlation between the expression levels of protein and gene in each of the 7 significant protein-gene pair. Correlation analysis was performed between log-transformed gene expression data and protein expression data by using R package 'lm'. Among the 7 protein markers whose gene expression also correlated with survival in the same cancer-drug groups, we noticed that 4 (BCL2, CCNE2, ETS1, GAB2) showed positive correlation between gene expression level and protein expression level, while the remaining 3 (MAPK3, TIGAR, CTNNB1) showed negative correlation.

We also examined the consistency between the survival analyses based on the proteins and the genes. For example, for a particular protein-cancer-drug combination whose corresponding gene was also predictive of drug-specific survival, we examined the direction of their correlation with survival outcome. If high expression of the protein led to better survival in the cancer-drug group, we considered the protein to be positively correlated with survival outcome. If high expression of the corresponding gene also led to better survival in the cancer-drug group, the gene was also positively correlated with survival outcome. In this case, the protein and its corresponding gene showed consistency in terms of their directions of the survival outcome. However, if high protein expression and low gene expression led to better survival, or low protein expression and high gene expression led to better survival, the protein and its corresponding gene were inconsistent in their directions of the survival outcome. Similar to the correlation analysis above, out of the 7 proteins

whose corresponding genes were also predictive of drug-specific survival outcomes, 4 showed consistent survival directions between genes and proteins, whereas the remain 3 showed inconsistent survival direction. This is not surprising, given mixed reports in the literature on the concordance and discordance between gene expression and protein expression in various contexts [93, 94].

2.4 Discussion

This study represents an integrative analysis using TCGA data. By performing survival analysis on miRNA and protein expression data for patients with the same cancer type and exposed to the same drug, we identified several miRNAs and proteins significantly associated with survival outcomes in a drug-specific manner.

For miRNA markers, We further investigated the regulatory mechanism of identified miRNAs by examining the relationship between drug-specific miRNA markers and their target genes, which revealed consistencies between the identified miRNAs and their target genes in terms of their expression levels and their correlations with survival outcomes. For each miRNA-cancer-drug combination being examined, we considered the regulation direction of the miRNA on its target genes, as well as the directions of correlations with the survival outcome. Using the R package 'lm', we examined the correlation between the expression levels of miRNAs and their target genes in each of the 59 miRNA-mediated gene-cancer-drug combinations in Table 2.3, and observed that the correlations of all 59 are significant (p-value <0.05). We then examined the consistency of significant miRNAs and their target genes by comparing both their survival directions and regulatory directions. For example, for a given miRNA-mediated gene-cancer-drug combinations where both the miRNA and the target gene were predictive of survival outcome in the same cancer-drug group, we compared the direction of their associations with the survival outcome. If the miRNA inhibits the expression level of the target gene, and high miRNA expression and low gene expression led to better survival, or low miRNA expression and high gene expression led to better survival, then we considered that the miRNA and its target gene were consistent in their directions of correlation with survival outcome. Otherwise, the particular pair of miRNA and target gene showed inconsistency in the directions of correlation with survival outcome. In contrast, if the miRNA activates the target gene, and both the high miRNA expression and high gene expression, or both the low miRNA ex-

pression and low gene expression resulted in improved survival outcome, then the miRNA and the target gene were consistent in the directions of survival outcome. Based on the experimentally verified miRNA-target gene pairs we obtained from multiple databases, among the 59 miRNA-mediated gene-cancer-drug combinations, 32 were reported in these databases that the target genes were inhibited by the miRNAs, whereas the regulatory relationships for the rest 27 are unknown. Here, 22 ($\sim$69%) of the 32 with inhibition relationship showed consistency in the survival directions with the miRNAs and target genes, which illustrated the consistency between the drug-specific miRNAs and the target genes regarding their correlations with survival outcome. Although the remaining 10 ($\sim$31%) showed inconsistent survival direction, this may be because gene expression is not only regulated by miRNAs alone, but can be controlled at various steps,including transcription, pre-mRNA splicing and export, mRNA stability, translation, protein modification, and protein half-life [95].

Some of the identified drug-specific miRNAs (such as miR-296, miR-330 and miR-483) have been previously reported to be potential biomarkers that are highly correlated with drug sensitivity and resistance in various cancer contexts. In several independent datasets on breast cancer, we observed that multiple of our drug-specific miRNA markers for breast cancer (i.e., miR-20b, miR-268, miR-363 and miR-577) were predictive of overall survival and other measures of clinical outcome. Notably, we observed that miR-577 was predictive of survival in multiple BRCA-drug groups (Table 2.4), indicating that miR-577 may potentially serve as key biomarkers to multiple drug responses in breast cancer. We found that decreased miR-577 expression leads to prolonged survival outcomes of BRCA patients treated with docetaxel, tamoxifen, or anastrozole. However, there is previously literature indicating that down-regulation of miR-577 was correlated with increased invasion and metastasis in breast cancer [96]. Despite this inconsistency in miR-577's impact on survival between the previous literature and the observations in this study, miR-577 can be a promising biomarker for both predicting clinical outcomes and studying drug mechanisms.

For protein markers, we identified several candidates whose expression levels were significantly associated with drug-specific survival outcomes across various cancer types. Notably, our results included proteins that have been previously reported to be predictive biomarkers for drug sensitivity and resistance in cancers, as well as the novel ones that have not been proposed in the literature. In addition, by using the gene expression data from TCGA, we found the gene expression of $\sim$11% of

the drug-specific proteins also showed significant correlation with drug-specific survival, and most of these drug-specific proteins and their corresponding genes are strongly correlated.

In summary, this study demonstrates that miRNAs and proteins can be effective biomarkers predictive of drug-specific survival outcomes in cancer. The miRNA and protein markers we identified are promising and may be useful tools toward personalized medicine.

CHAPTER 3

ESTABLISHING THE CORRELATION PATTERNS BETWEEN CPG METHYLATION AND ALTERNATIVE SPLICING IN CANCER

3.1 Introduction

DNA methylation is one of the most extensively studied epigenetic mechanisms and is essential for normal development and many key biological processes, including embryogenesis, genome imprinting, and regulation of gene transcription [97, 98, 99]. Although DNA methylation is known to be perturbed in various cancers, its role in tumorigenesis is not fully understood [100], motivating numerous studies to explore the mechanism of DNA methylation in human cancers [101, 102]. A main question in this regard is the role of CpG methylation and gene expression. Many studies have characterized DNA methylation as a silencer or activator of gene transcription because methylation of the promoter region often represses gene expression, while methylation of some CpG sites in the gene body is correlated with increase of transcriptional activities [103, 104, 105, 98]. However, emerging evidence shows that DNA methylation not only affects transcription but also regulates alternative splicing [42, 43].

Alternative splicing is known to have significant impact on cancer [106, 107]. Yet, few studies have explored the correlation between methylation and alternative splicing, specifically the regulatory mechanisms of DNA methylation on alternative splicing in cancer. We speculate that this could be due to methylation being an epigenetic event that occurs on DNA, while alternative splicing occurs during the processing of RNA molecules, and therefore the possibility of a direct relationship between DNA methylation and alternative splicing has received less attention than it should have. The lack of explorations of the possible role of DNA methylation in alternative splicing presents an opportunity for further research, as it could contribute toward new understanding of DNA methylation during transcriptional processes, especially in tumorigenesis.

With advances in high-throughput genomics, extensive cancer databases such as The Cancer Genome Atlas (TCGA) have become publicly available to examine the relationship between DNA

methylation and alternative splicing. Using TCGA data resources, some in vitro studies in the literature have incorporated DNA methylation data and transcript level quantification data to explore the potential link between alternative splicing and DNA methylation. For example, one previous study has used association analysis to characterize the underlying relationship between cancer-specific alternative splicing and DNA methylation, and observed several different patterns of methylation-alternative splicing correlations [108]. Another study developed a splicing decision model to identify actionable methylation loci potentially affecting splicing events, which revealed that intragenic methylation status is important for splicing regulation [109]. However, most of the previous studies were limited to specific cancers, or focused exclusively on differentially methylated CpGs between tumor and normal samples. In this study, we aimed to comprehensively integrate and analyze different molecular data types in TCGA, including data on DNA methylation, exon expression, isoform expression, and gene expression, to unravel the correlation pattern between methylation and alternative splicing across multiple cancer types.

Here, we present a pan-cancer analysis to examine the relationship between DNA methylation and alternative splicing on an individual gene basis (Figure 3.1). For one cancer and one gene, we selected patients with high expression of that specific gene, henceforth defined as "gene-specific patients". We used the data on those patients to compute the correlations between DNA methylation levels of CpG sites associated with the gene and the expression data of exons of that gene. Thus, we identified genes whose DNA methylation was significantly correlated with alternative splicing in cancers. Next, we reviewed the literature to find evidence supporting the relationship between DNA methylation and alternative splicing in the identified genes. Furthermore, for each cancer, we stratified CpG sites into two classes (significant CpGs vs. non-significant CpGs), depending on whether the CpG sites were significantly associated with at least one of the exons of the corresponding gene. In order to evaluate whether there are differences in these two classes in terms of their ability to predict survival outcomes, we applied survival analysis to all the individual CpG sites and found CpG sites that correlate with exons are more likely to be correlated with survival, compared to the CpG sites that do not correlate with exons. In addition, we applied a correlation analysis between the CpG sites and isoforms for each gene in each cancer type. Interestingly, we observed that CpG sites are more strongly correlated with exon expressions than their correlations with isoform expres-

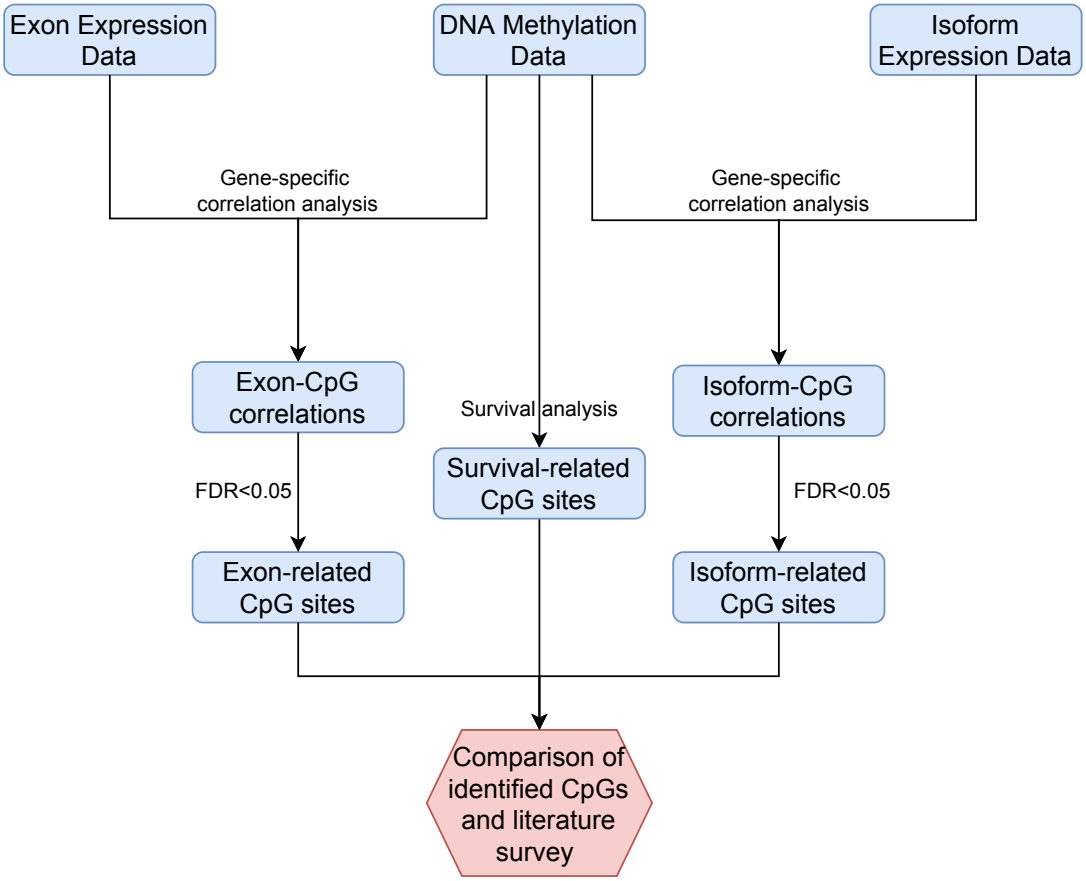

Figure 3.1: flowchart of data analysis used in this study. First, we selected gene-specific samples that highly expressed the gene for each of the genes we analysed, and then we performed correlation analysis between DNA methylation and exon/isoform expression based on gene-specific samples. A threshold of FDR$<$0.05 was required to detect significant correlations. In addition, we performed survival analysis for each CpG sites and identified a list of CpG sites that are predictive of patients' survival outcome. Finally, a literature survey was conducted to search for existing evidences of identified correlations.

sions. This observation indicates that CpG methylation may affect alternative splicing by regulating the inclusion or exclusion of exons, which subsequently impacts the expression and usage at the isoform level.

3.2 Methods

3.2.1 Data access

Methylation, exon expression, isoform expression, and gene expression data were downloaded from TCGA Genomic Data Commons (GDC) using the GDC Data Transfer Tool. The methylation data used in this study were acquired by the Illumina HumanMethylation450K array, which integrates 485,577 CpG sites on the Illumina chip. Exon, isoform, and gene expression data were measured using the IlluminaHiSeq technology, which included 239,322 exons, 198,619 isoforms, and 23,548 genes, respectively. In addition, this analysis covered 33 different cancer types and data of more than 11,000 cancer patients available in TCGA. We also download the clinical data from GDC, which provided survival outcome data for 11,082 patients.

3.2.2 Data preprocessing

The gene expression data downloaded from TCGA were normalized by FPKM-UQ [61], and we subsequently transformed the expression data by log-transformation. For each gene, we used StepMiner [63] to compute a global threshold based on all patients across all cancer types. We first sorted the expression data of a given feature for all patients and then fitted a step function to minimize the square error between the original and the fitted values. Since this threshold is derived based on the data of all patients across all cancer types, it is able to robustly define high and low expression. The threshold of each gene was used to binarize the data, so that the patients could be divided into two groups (highly-expressed group vs. lowly-expressed group) according to the binarized expression data of an individual gene.

3.2.3 Correlation analysis

Correlation analysis was performed using Pearson's correlation with a Bonferroni correction to the p-values based on the number of correlations computed for each cancer type. A correlation analysis was performed between DNA methylation data and expression data of either exons or isoforms, with an FDR-corrected p-value threshold of 0.05. All statistical tests were performed using standard Python functions.

3.2.4 Survival analysis

For each gene and each cancer type, we performed survival analysis on all the individual CpG sites associated with that gene based on the methylation data of patients who highly expressed the gene. Log-rank test was used to assess the statistical significance of the survival difference between patients with the CpG site methylated or unmethylated. P-values of the log-rank test were adjusted for multiple testing using Benjamini-Hochberg method with a false-discovery rate (FDR) <0.05. Kaplan-Meier analysis and log-rank tests were performed using the R package "survival".

3.2.5 Literature search

A literature search was performed using PubMed, accessed via the National Library of Medicine PubMed interface (http://www.ncbi.nlm.nih.gov/pubmed). We programmatically searched the PubMed database using custom Python scripts. We searched through PubMed for all keywords in all fileds, including the title, abstract and main texts of the articles.

3.3 Results

3.3.1 Correlations between DNA methylation and alternative splicing

For each of the 33 cancer types in TCGA and each gene, we performed a correlation analysis of the DNA methylation data for each CpG site and the expression data of each exon of the corresponding gene, which involved with a total of 485,577 CpG sites and 239,322 exons in 20,880 genes. On average, one gene was associated with 22 CpG sites and had 9.65 exons, which amounted to examining a total of 134,299,062 CpG-exon correlations.

For each cancer type and each gene, patients belonging to the cancer type were stratified into a highly-expressed group and a lowly-expressed group by applying StepMiner [63] to binarize the gene expression data of the gene. Specifically, we sorted the expression data of the gene for all patients in all cancer types, and then fitted a step function to the sorted data that minimized the squared error between the sorted data and the fitted step function, which provided a global threshold to binarize the expression of the gene. Since this threshold was derived based on the data of all patients across all cancer types, it robustly defined high and low expression of the gene. Since

we were interested in studying alternative splicing, we focused on patients in the highly-expressed group for the specific gene of interest in the cancer type. We then calculated the correlations between DNA methylation of each of the CpG sites and expression of each of the exons for the given gene based on those patients who highly expressed the gene in that cancer type. In addition, we required that correlation analysis can only be performed when there were at least 10 patients for a given gene and a given cancer type to avoid bias due to small sample size. After computing CpG-exon correlations for each gene in each cancer type, we identified significant CpG-exon correlations using a 5% false discovery rate (FDR) threshold. The correlation analysis in 33 cancer types led to 6,049,182 significant CpG-exon correlations associated with 280,838 unique CpG sites and 190,577 unique exons in 19,984 genes, which we called "significant CpG sites" and "significant exons", respectively. Since the correlation analysis was performed for each cancer and each gene separately, it is possible that a CpG site may correlate with multiple exons in the corresponding gene, and their association may appear in different cancer contexts. We focused on CpG-exon correlations that were significant in multiple cancer types in this study, because these CpG sites may play important role in exon usage during transcriptional processes and carcinogenesis. A total of 36,470 CpG-exon correlations were significant in more than one cancer type, which we called multi-cancer CpG-exon correlations. The 30 strongest correlations we found are presented in Table 3.1.

To better understand the association between DNA methylation and exon expression, we evaluated the consistency of correlation direction in each of the 36,470 CpG-exon pairs that showed significant correlation in multiple cancers. Specifically, for each of the 36,470 CpG-exon pairs, we extracted its corresponding CpG-exon correlation values that showed significance, and stratified these significant correlation values. This stratification defined two groups of values for a CpG-exon pair, a majority group consisting of more than half of the corresponding correlation values that shared the same correlation direction, and a minority group consisting of the remaining correlation values for the CpG-exon pair. After that, we computed a consistency score for the CpG-exon pair as the ratio between the number of correlations in the majority group and the total number of significant correlations for the CpG-exon pair. The consistency score ranged from 0.5 to 1, with 1 indicating a perfect consistency where all of the corresponding CpG-exon correlation values shared the same correlation direction, and 0.5 indicating poor consistency. Among the 36,470 CpG-exon

pairs with significant correlations in multiple cancers, 78.71% showed a perfect consistency score of 1, and 84.52% showed a consistency score of >0.7. Therefore, when a CpG-exon pair showed significant correlations in multiple cancers, the directions of the correlations tended to be the same.

Table 3.1: Top 30 multi-cancer CpG-exon correlations across all the cancer types

CpG	**Exon**	**Gene**	**Cancer**	**Corr**	**Majority Sign**
cg23538703	chr3:118864997-118866434:+	RP11-484M3.5	BRCA;CESC	0.845	−
cg02425416	chr11:2170356-2170833:-	INS-IGF2	CHOL;LIHC	0.822	+
cg02425416	chr11:2170356-2170575:-	INS-IGF2	CHOL;LIHC	0.821	+
cg02425416	chr11:2168796-2169037:-	INS-IGF2	CHOL;LIHC	0.818	+
cg05777976	chr11:2170356-2170833:-	INS-IGF2	CHOL;LIHC	0.807	+
cg05777976	chr11:2170356-2170575:-	INS-IGF2	CHOL;LIHC	0.803	+
cg05777976	chr11:2168796-2169037:-	INS-IGF2	CHOL;LIHC	0.802	+
cg13167664	chr11:2170356-2170575:-	IGF2	CHOL;LIHC	0.792	+
cg04083712	chr7:99158156-99158318:+	GS1-259H13.10	COAD;ESCA	0.786	−
cg13167664	chr11:2170356-2170833:-	IGF2	CHOL;LIHC	0.782	+
cg01814130	chr3:118906622-118906822:+	RP11-484M3.5	BLCA;CESC	0.773	−
cg14458615	chr3:118864997-118866434:+	RP11-484M3.5	BRCA;CESC	0.766	−
cg12389423	chr3:118864997-118866434:+	RP11-484M3.5	BRCA;CESC	0.765	−
cg10066151	chr19:57874879-57875071:+	AC003002.4	CESC;LGG	0.756	−
cg15744005	chr10:104632854-104632990:+	C10orf32-ASMT	BRCA;CESC	0.755	−
cg20675391	chr3:118864997-118866434:+	RP11-484M3.5	BRCA;CESC	0.751	−
cg15744005	chr10:104632205-104632355:+	C10orf32-ASMT	BRCA;CESC	0.750	−
cg27331871	chr11:2168796-2169037:-	INS-IGF2	CHOL;LIHC	0.749	+

CpG	Exon	Gene	Cancer	Corr	Majority Sign
cg05777976	chr11:2170356-2170833:-	IGF2	CHOL;LIHC	0.747	+
cg05777976	chr11:2170356-2170575:-	IGF2	CHOL;LIHC	0.746	+
cg02166532	chr11:2170356-2170575:-	IGF2	CHOL;LIHC	0.744	+
cg15744005	chr10:104629841-104629968:+	C10orf32-ASMT	BRCA;CESC	0.744	−
cg04334121	chr11:18257383-18257477:-	SAA4	CHOL;LIHC	0.743	−
cg16817891	chr11:2170356-2170575:-	INS-IGF2	CHOL;LIHC	0.743	+
cg04334121	chr11:18253942-18254080:-	SAA4	CHOL;LIHC	0.742	−
cg13167664	chr11:2168796-2169037:-	IGF2	CHOL;LIHC	0.741	+
cg16817891	chr11:2170356-2170833:-	INS-IGF2	CHOL;LIHC	0.740	+
cg27331871	chr11:2170356-2170833:-	INS-IGF2	CHOL;LIHC	0.740	+
cg20844262	chr20:62365997-62366176:+	ZGPAT	CHOL;LIHC	0.740	−

In Table 3.1, the multi-cancer CpG-exon pairs were sorted by the mean of absolute values of their significant correlations. We observed that these top CpG-exon pairs all showed relatively high correlations in various cancers. These top correlations are dominated by a few CpG positions and their associations with several genes. A literature search based on PubMed database was performed to further investigate the significance of these top CpG-exon correlations. Interestingly, we found supporting evidence for two of the top genes. One is insulin-like growth factor 2 (IGF2), an imprinted gene with the parental allele expressed and the maternal allele silenced [110]. Previous studies have reported that epigenetic features such as histone modifications and DNA methylation are associated with different levels of IGF2 transcription regulation such as alternative splicing [111]. The same CpG sites or nearby CpG sites also showed strong correlations with the exon expression of INS-IGF2, which is a read-through transcript composed of exons from proinsulin precursor (INS) and IGF2 [112]. In the literature, it was observed that hypomethylation of INS-IGF2 was correlated with increase of INS-IGF2 transcripts in four breast cancer cell lines [113]. Our analysis, using

Table 3.2: CpG sites that correlate with both exon expression and survival outcome

CpG	Gene	Cancer	Pvalue	FDR
cg20351640	CD302	brain	1.70E-08	0.000138852
cg03903831	RP11-307N16.6	brain	3.51E-08	0.000181116
cg27048140	ATP5J2-PTCD1	brain	8.93E-08	0.000289902
cg19857457	RPL17-C18orf32	breast	1.09E-08	0.000337161
cg03903831	SPATA13	brain	1.88E-07	0.000432463
cg05032848	RFFL	brain	2.45E-07	0.000497251
cg20477147	NPEPL1	brain	3.95E-07	0.000600647
cg24859623	CD302	brain	3.95E-07	0.000600647
cg09088496	RP11-307N16.6	brain	3.86E-07	0.000600647
cg04735129	CD302	brain	4.81E-07	0.000657035
cg08347373	CD302	brain	5.69E-07	0.000711062
cg08278937	SERINC4	brain	7.10E-07	0.000783884
cg16992627	TNXB	brain	8.01E-07	0.000830607
cg21870038	RFFL	brain	8.07E-07	0.000830823
cg13586610	TNXB	brain	1.07E-06	0.000969694
cg22158248	ALDH2	brain	1.23E-06	0.001040217
cg13799005	RP11-644F5.10	brain	1.25E-06	0.001049216
cg09088496	SPATA13	brain	1.97E-06	0.00123508
cg06017559	RP11-161M6.2	brain	2.34E-06	0.001332273
cg07204711	CDK3	brain	2.32E-06	0.001332273

*P-value and FDR are for log-rank test of CpG association with survival

an entirely different method, has identified these previously recognized relationships between DNA methylation and expression of these genes.

3.3.2 CpG sites correlated with exons tended to be more predictive of survival outcomes

To explore the differences between CpG sites with and without significant correlation with exons, we compared their ability to predict cancer survival outcomes. Specifically, for each cancer type and each CpG site, we focused on patients that highly expressed the gene associated with that CpG site, and performed survival analysis and log-rank test to examine whether methylation status of that

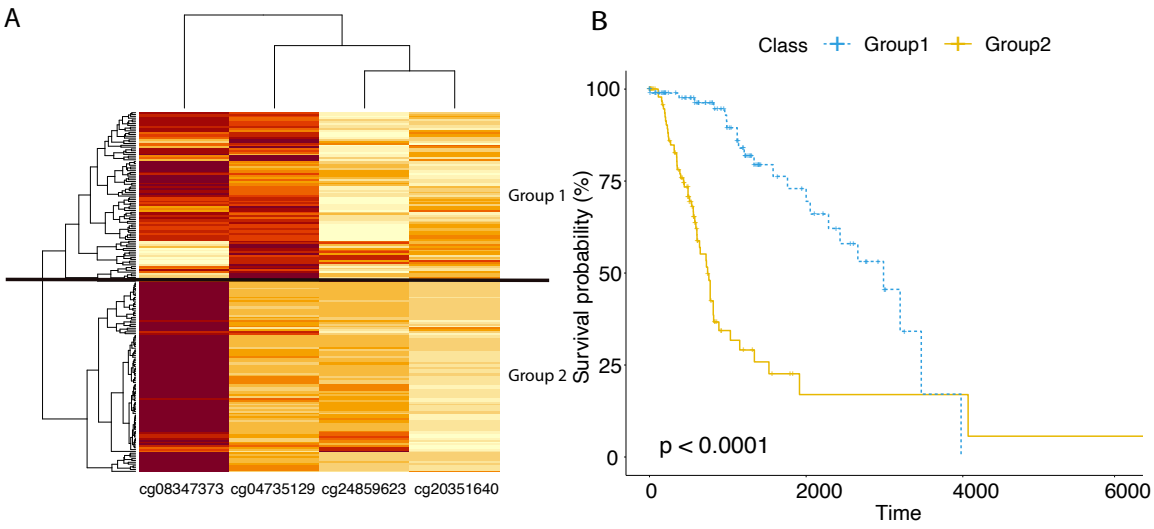

Figure 3.2: A: Cluster heatmap of the methylation data of the CpG sites associated with CD302 in brain cancer. Patients were separated into two groups based on the clustering results. B: Kaplan–Meyer survival curves of the two patient group stratified by the CpGs associated with CD302.

CpG site was predictive of survival of patients who expressed the corresponding gene. This survival analysis in 33 cancer types identified 31,045 CpG sites that correlated with patient's survival after a FDR correction for p-values (<0.05). Interestingly, when we compared the list of CpG sites correlated with the exons and the list of CpG sites associated with survival, we observed that 643 CpG sites were overlapped between the two lists, all of which were related to brain and breast tumors. According to the significance of survival differences, Table 3.2 shows the top CpG sites that correlated with both exon expression and survival outcomes. We performed a literature search on the genes associated with the overlapping CpG sites and found previous studies linked to them. For example, our results showed that multiple CpG sites associated with CD302 were predictive of survival in brain tumors (Figure 3.2), which is consistent with another study showing that the expression level of CD302 was upregulated in brain regions and may be involved in biological processes such as development, differentiation, and immunological responses [114]. Specifically, we identified four CpGs associated with CD302 and are predictive of survival in brain cancer, which are cg08347373, cg04735129, cg24859623, and cg20351640. In order to examine their prognostic ability, we clustered 199 patients with brain cancer using the methylation data of the four survival-related CpGs (Figure 3.2A), and observed two well-separated clusters. Next, we stratified patients

into two groups, which are group1 and group2, according to the clustering results. Survival analysis was performed to examine the ability of the identified CpGs for stratifying patients with different prognoses (Figure 3.2B).

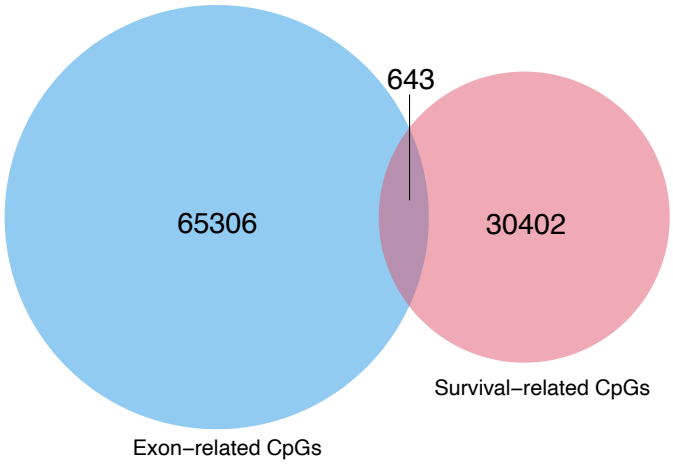

Figure 3.3: CpGs correlated with exons and survival. This Venn diagram shows the overlapping between the CpGs correlated with the exons and the CpGs that are predictive of survival.

To assess whether the number of overlapping CpG sites that correlated with both exon and survival was statistically significant, we computed the expected number of overlaps in cases where the two relationships were independent. Since we analyzed data for 33 cancer types, the total number of CpG-survival relationships examined is the number of CpG sites (485,577) multiplied by 33. As mentioned above, we observed 31,045 significant CpG-survival relationships, which is roughly 0.19% of the total number of CpG-survival relationships examined. If the CpG-exon correlation was independent of the CpG-survival relationships, among the 65,949 CpG sites correlated to exons, only 124 are expected to be significantly correlated with survival. However, we observed 643 overlapping CpGs in both CpG-exon correlation analysis and survival analysis, which was 5.2 times greater than expected (Figure 3.3). This observation showed that the CpG sites correlated with exons tended to be more correlated with cancer survival outcomes than expected, which could be an indication that CpG sites correlated with exons tend to have functional consequences on clinical outcomes [106, 107]. This result was consistent with those of previous studies regarding the ability of methylation-related alternative splicing events to predict survival outcomes in cancer [106, 107, 108]

3.3.3 CpG sites are more strongly correlated with exons than their correlation with isoforms

We examined the relationship between DNA methylation and isoform expression. Similar to the analysis of CpG-exon correlation, for each cancer type and each gene, we first stratified patients into highly-expressed and lowly-expressed groups according to the binarized expression data for the gene. We only used patients in the highly-expressed group to calculate correlations between the methylation level of CpG sites associated with the gene and the expression data of isoforms of that gene. For a given cancer type, we computed the correlation for a given gene only if the number of patients in highly-expressed group exceeded 10. We applied FDR adjusted significance p-value < 0.05 as a threshold in each cancer type to identify significant CpG-isoform correlations in the 33 cancer types. A total of 1,929,993 CpG-isoform correlations exceeded the threshold, which involved 276,554 CpG sites, 131,735 isoforms, and 19,368 genes.

We compared the strengths between CpG-exon correlations and CpG-isoform correlations, where the same CpG was significantly correlated with the expression of an exon as well as with an isoform containing the exon. Specifically, in each cancer type, for each of the CpG sites significantly correlated with an exon, we examined all the isoforms that correlated with the same CpG site and contained the exon. We then calculated the difference between the CpG-exon correlation and the corresponding CpG-isoform correlation to determine which of them was stronger. Since an exon can belong to multiple isoforms, it is possible that a CpG-exon correlation has multiple corresponding CpG-isoform correlations in the same cancer context, and these CpG-isoform correlations may have different strengths or even different directions. Therefore, for a given CpG-exon correlation and its corresponding CpG-isoform correlations, we selected the strongest CpG-isoform correlation with a direction consistent with the CpG-exon correlation. For example, if a CpG-exon pair was positively correlated, we selected the CpG-isoform that showed a positive correlation and had the largest correlation value. In contrast, if a CpG-exon pair was negatively correlated, we selected the negative CpG-isoform correlation with the smallest correlation value. We computed the differences between CpG-exon correlations and the corresponding strongest CpG-isoform correlations, and visualized the differences in the histogram in Figure 3.4. It is interesting to observe that the histogram is skewed toward the positive side, with 70.20% of the differences greater than 0. When we set up

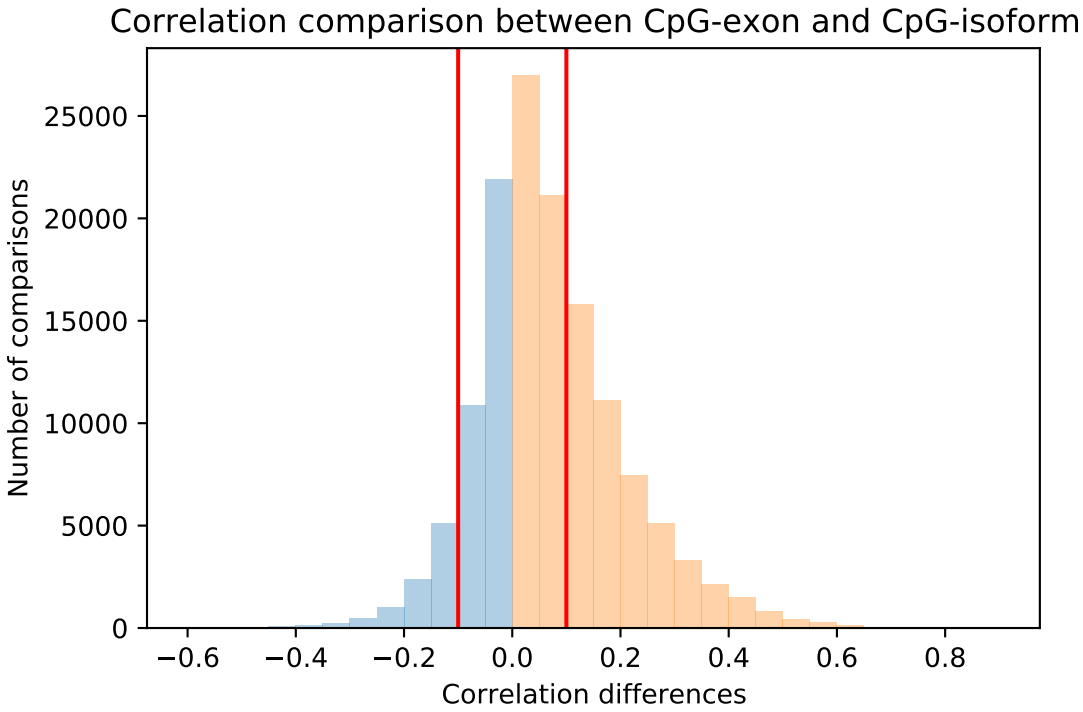

Figure 3.4: Comparison of correlation strength between CpG-exon and CpG-isoform. This histogram shows the differences between CpG-exon correlations and the corresponding strongest CpG-isoform correlations, which indicates that CpG-exon correlations tend to be stronger than the corresponding CpG-isoform correlations.

± 0.1 thresholds for the correlation differences, 34.76% of the differences were > 0.1, whereas only 6.81% of the differences were < -0.1, meaning that CpG-exon correlations tended to be stronger than the corresponding CpG-isoform correlations. This result may indicate that DNA methylation's impact on the inclusion or exclusion of exons could subsequently affect isoform usage.

3.4 Discussion

In this study, we examined the relationship between DNA methylation and alternative splicing by integrating multiple data types in TCGA. Our analyses successfully identified significant CpG-exon correlation patterns in various cancer contexts and explored the regulatory mechanism of CpG methylation on alternative splicing. Although most CpG-exon pairs showed a negative correlation in various cancer contexts, consistent with the concept of methylation-induced expression silencing, we also observed a substantial number of CpG-exon pairs exhibiting positive correlations. Remark-

ably, we showed that the majority of CpG-exon expression correlations had a consistent direction across multiple cancer types, indicating that the relations we have identified may share common molecular mechanisms across multiple cancers. Log-rank test was used to explore the connection between CpGs that correlated with exons and CpGs that correlated with survival outcomes. We showed that the CpG sites correlated with exon expression were enriched with CpG sites that correlated with survival outcomes, which may indicate that the CpGs correlated with exons have larger functional consequences than CpGs that do not correlate with exons. Therefore, the significant CpG-exon expression correlations we have identified in this study may provide useful candidates for functional studies in the future. Furthermore, we performed pairwise correlation analysis between CpG sites and isoforms for each cancer-gene combination, and compared the correlation strength between CpG-exon and CpG-isoform. Our analysis demonstrated stronger correlations between CpGs and exons compared to the correlations between CpGs and isoforms that contain the exons, indicating that CpG methylation may be associated with alternative splicing via regulating the inclusion or exclusion of exons, which subsequently impacts the relative usage of various isoforms.

This study is not without limitations. One limitation was the lack of independent validation datasets. It would be ideal to validate the identified CpG-exon correlations in independent patient cohorts. However, we were unable to find a multi-omics cancer dataset with exon-level gene expression data, CpG methylation data, and survival data. Therefore, our ability to perform independent validation was limited by data availability. Another limitation pertained to the literature search. Although a comprehensive literature review is a powerful tool for validating and evaluating our results based on existing knowledge, the large number of search results generated by simple searches with keywords made it difficult for us to distinguish useful information efficiently. In addition, experimental results on methylation analysis were rarely reported using CpG id in Table 3.1 and Table 3.2 as identifiers. Therefore, it is likely that we did not fully capture the relevant literature associated with the correlations observed in this study.

Despite these limitations, this study revealed the relationship between CpG methylation and alternative splicing in cancer and contributed toward the understanding of the role of methylation in alternative splicing during transcriptional processes and carcinogenesis. In addition, survival-

related CpG sites not only provided positive indications of the functional relevance of CpGs that correlate with exons, but also served as potential biomarkers predictive of clinical outcomes. In summary, the comprehensive survey of associations between methylation and alternative splicing will facilitate the exploration of the role of methylation regulation in transcriptional processes in cancers.

CHAPTER 4

PREDICTING POTENTIAL MIRNA TARGET GENES

4.1 Introduction

MicroRNAs (miRNAs) are small, non-coding RNAs that play a crucial role in the post-transcriptional regulation of gene expression [44, 115, 116]. Typically, miRNAs silence target genes by binding to complementary sequences in the 3' untranslated region of mRNAs, leading to their degradation and inhibiting their translation [117]. However, recent in vitro studies have suggested that miRNAs can also activate gene expression in certain circumstances through various mechanisms [102, 118]. Overall, miRNAs are versatile regulators of gene expression with complex roles in numerous biological processes, which has inspired many studies to explore the intricate relationship between miRNAs and their target genes [119, 120]. In recent years, extensive reviews have highlighted the pivotal role of miRNAs in human diseases. For example, miRNAs have been shown to regulate gene expression and are associated with a wide range of diseases, with computational models increasingly predicting miRNA-disease associations and integrating experimental data to better understand disease mechanisms [121]. Advances in bioinformatics tools and databases have further refined our knowledge of miRNA functions and their implications in human diseases. Systematic evaluations of these computational models have also demonstrated their potential to accurately predict miRNA-disease associations, enabling more comprehensive evaluations of miRNA functions in disease pathways [122]. Additionally, miRNAs are particularly studied for their regulatory roles in diseases such as cancer, where they significantly influence gene regulation and disease mechanisms [123]. Moreover, miRNAs modulate important signaling pathways involved in human diseases, such as PI3K-Akt and MAPK, making them attractive therapeutic targets [124]. Our study builds upon this foundation by predicting novel miRNA-gene interactions in cancer, contributing to the understanding of miRNA-mediated regulation in disease.

In the literature, researchers have determined many miRNA–gene relationships through experiments using high-throughput techniques on a transcriptome-wide scale, leading to a great number of

publications and repositories of experimentally validated miRNA target genes [125, 126, 127]. For example, miRBase [128], DIANA-TarBase [129], and miRTarBase [130] are databases that provide comprehensive information about miRNAs and their targets. These databases include information on experimental methods and supporting literature to ensure the accuracy and reliability of the miRNA-gene interactions cataloged. In parallel to the experimental efforts, several computational approaches have been developed to predict miRNA-gene relationships [131, 132, 133, 134, 135]. TargetScan is one of the widely used methods that predict miRNA target sites conserved among orthologous 3' UTRs of vertebrates based on the degree of sequence complementarity between the miRNA and its target mRNA [136]. While the recent developments in miRNA-gene databases have facilitated the exploration of miRNAs and provided new insights into their relationship with genes, these databases are not without limitations. Given the critical role of miRNA-gene regulation, the existing databases likely only represent a sparse coverage of miRNA-gene relationships, which provides an opportunity and motivation for this study, aiming to identify the yet unraveled miRNA-gene relationships.

In this study, we employed a machine learning approach to predict unknown miRNA-gene relationships by leveraging publicly available multi-omics gene expression data and several existing miRNA target databases. Using the miRNA and gene expression data from The Cancer Genome Atlas (TCGA) [137], we performed a comprehensive correlation analysis of each miRNA and each gene across 32 cancer types. We also compiled the known miRNA-gene relationships from the existing databases to label the miRNA-gene pairs. The total number of known miRNA-gene relationships in those databases is 197,877. After that, eXtreme Gradient Boosting (XGBoost) was used to build machine learning models that predict potential miRNA-gene relationships that have not been reported in those miRNA target databases. In addition, our findings were further validated by held-out miRNA target databases and a literature survey, which showed the ability of the models in terms of predicting novel miRNA-gene relationships. Furthermore, by exploring independent datasets that contain gene expression before and after specific miRNA interventions, we found consistent alignment between the correlation direction of our predicted miRNA-gene pairs and their regulatory patterns shown in these independent datasets.

4.2 Methods

4.2.1 Data access

The miRNA expression data was downloaded from TCGA Genomic Data Commons (https://portal.gdc.cancer.gov/). A total of 1,881 miRNAs were included in this study. The miRNA expression data was subsequently log-transformed for analysis. The gene expression data were downloaded from ucsc xena (http://xena.ucsc.edu) (IlluminaHiSeq, level 3 expression), which comprised a total of 20,530 genes. The miRNA and gene expression data covered 10,004 patients across 32 cancer types (glioblastoma was excluded from this study due to the small sample size in miRNA data).

In addition, we collected established miRNA targets from five different databases (mir2Disease, miRecords, TarBase, miRTarBase, and TargetScan) to label each miRNA-gene pair during model training. In total, 197,877 known miRNA-gene relationships were included in this study.

For the purpose of validating the predicted miRNA-gene pairs, a total of 9 datasets were collected from the GEO database. These datasets were specifically selected to contain gene expression profiles both before and after specific miRNA perturbations. These datasets were GSE68742, GSE9586, GSE62951, GSE69990, GSE12091, GSE83690, GSE25215, GSE18510, and GSE136665.

To compare our method with other existing tools, such as, miRanda, TargetMiner, and PITA, we downloaded 924 human miRNA sequences and 15,705 $3'$ UTRs from Ensembl database (https://useast.ensembl.org/index.html) using BiomaRt tool.

4.2.2 Correlation analysis between miRNAs and genes

We performed a Pearson's correlation analysis between miRNA expression and gene expression for each cancer type. To mitigate the effect of small sample sizes, we calculated correlation coefficients only for miRNA-gene pairs with expression data from at least 10 patients per cancer type. Pairs that could not be calculated in more than three cancer types were excluded from the study. All statistical tests were performed using standard Python functions.

4.2.3 Construction of prediction model

We used (Python library version 2.1.1) to perform binary classification, predicting potential miRNA-gene relationships. The analysis involved over 22 million miRNA-gene pairs, each labeled as either positive or negative based on existing miRNA target databases. Only 0.12% of the pairs were labeled as positive, representing confirmed miRNA-gene relationships, while the remaining 99.88% were labeled as negative since they were not documented in the databases. To address this significant class imbalance, we downsampled the negative class at various rates (ranging from 0.1% to 10%). For each downsampled version of the data, samples (miRNA-gene pairs) in the data were split, with 80% used for training and 20% for testing. Model prediction performance was evaluated using the Area Under the Curve (AUC) metric. At each downsampling level, we trained 1000 models, which were applied only to the negative class. MiRNA-gene pairs consistently predicted by these models were identified as potential, previously undiscovered miRNA target relationships.

4.2.4 Literature survey

A literature search was performed using PubMed. We programmatically searched the PubMed database using custom Python scripts. We searched through PubMed for all keywords in all fields, including title, abstract, and main texts of the articles.

4.2.5 Differential expression analysis

In each of the independent datasets, we performed a t-test to identify genes that expressed differently before and after the specific miRNA perturbations. Genes were considered as DEGs if the significant p-value <0.05.

4.3 Results

4.3.1 Construction of miRNA-gene prediction models

To build models for predicting potential miRNA-gene relationships, we used the correlations between miRNA and gene expression data to serve as features for describing miRNA-gene pairs. TCGA provided expression data of a total of 1,881 miRNAs and 20,530 genes, for 10,004 patients

across 32 cancer types. For each miRNA-gene pair, we computed their correlation in each of the 32 cancer types separately. Therefore, each miRNA-gene pair is described by a vector containing 32 correlation values. In order to avoid bias resulting from small sample sizes, for a miRNA-gene pair, their correlation in one cancer type was computed only when the number of patients expressing both the miRNA and the gene exceeded 10. We excluded miRNA-gene pairs whose correlations could not be calculated in more than three cancer types due to the limited sample sizes. As a result, we obtained the correlation values in the 32 cancer types for a total of 22,580,364 miRNA-gene pairs, involving 1,277 unique miRNAs and 17,822 unique genes.

The primary task of this study was to solve a binary classification problem to identify miRNA-gene association relationships. Binary class labels were needed to train the machine learning models. We gathered known miRNA-gene relationships from five existing databases of miRNA targets, including both experimentally validated and computationally predicted miRNA-gene interactions. These databases are miR2Disease [76], miRecords [77], TarBase [78], miRTarBase[130], and TargetScan[136]. We used mir2Disease, miRecords, and miRTarBase as training databases. The curated miRNA target genes in these training databases were used to define positive miRNA-gene pairs. The remaining two databases, TarBase and TargetScan served as validation databases. The miRNA target genes in the validation databases were used to evaluate the performance of our prediction. It is noteworthy that in evaluating model performance with the validation databases, we only considered those known miRNA-gene relationships that were only reported in the validation databases and not in the training databases.

For the 22,580,364 miRNA-gene pairs with correlation features computed, we labeled them as positives and negatives, depending on whether the miRNA-gene pair was reported in the training databases. Out of the 22,580,364 miRNA-gene pairs, 26,867 (0.12%) were labeled as positives because they were confirmed relationships in the training databases. The remaining 22,580,364 (99.88%) pairs were not present in the training databases, and thus, were labeled as negatives. Note that, due to the limited coverage of miRNA-gene interactions in the existing databases, the negative class included both unrelated pairs and true miRNA-gene relationships that were yet to be discovered.

To address the challenge of this extremely imbalanced binary classification task dominated by

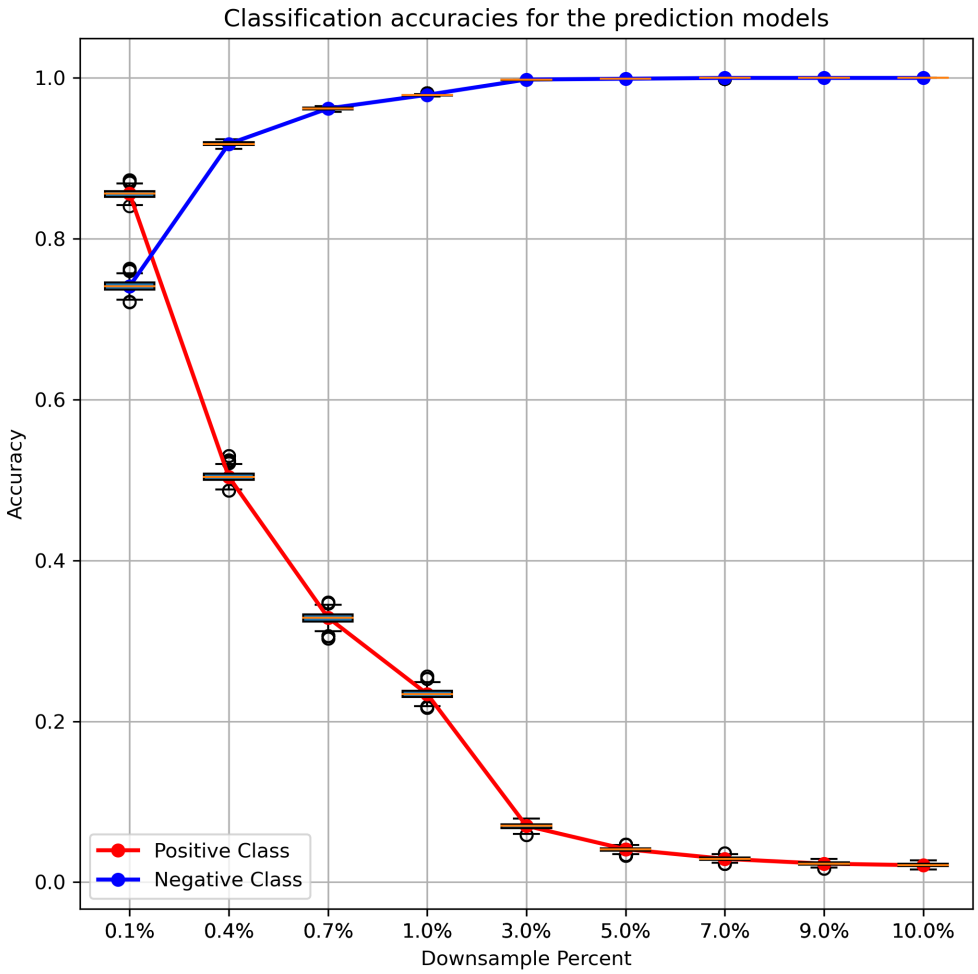

Figure 4.1: Classification accuracies for the prediction models on the testing datasets across 1000 experiments.

the negative class, we applied XGBoost (Extreme Gradient Boosting) to train machine learning models based on downsampled versions of the data. XGBoost is a powerful algorithm that excels in supervised classification problems [138]. Its ability to handle label-imbalanced data has been reported in many studies [139, 140], making it well-suited for identifying miRNA-gene relationships in our study. In addition, in order to reduce the imbalance, we generated many downsampled versions of the negative class to train XGBoost models. The size of the downsampled negative class ranged from 0.1% to 10% of the original negative class, so that the ratio between the positive and

negative class after downsampling ranged from 1-to-1 to 1-to-100. At each downsampling level, we generated 1000 versions of downsampled negative class. We combined each of the downsampled negative class and the positive class, split the data with an 80/20 ratio of training and testing, trained a classifier using XGBoost, and evaluated testing performance for the positive and negative classes separately. Figure 4.1 shows the performance across the 1000 models trained based on the 1000 versions of downsampled data. As detailed in Figure 4.1, incorporating more negatives into the training data led to a decreased prediction accuracy in the positive class (recall), while the prediction accuracy of the negative class (1-false positive) increased. This result was expected and suggested that a higher proportion of negatives might be more appropriate for identifying and hypothesizing miRNA-gene relationships for validation, because of the reduced false positives enabled by a higher proportion of negatives used in model training.

4.3.2 Validation of the identified miRNA-gene pairs in the validation databases and literature

At each specified level of downsampling, we applied our suite of 1000 trained predictive models exclusively on the miRNA-gene pairs in the negative class. Our objective was to identify miRNA-gene pairs that were repeated identified as positive by these models, which may highlight potential miRNA-gene interactions previously undocumented in the literature. To ascertain the reliability of our findings, we concentrated on miRNA-gene pairs that were consistently classified as positive in at least 700 out of the 1000 applications across each downsampling threshold. We designated these frequently predicted pairs as 'significant miRNA-gene pairs', thereby implying a higher confidence in their potential biological relevance. We evaluated the significant miRNA-gene pairs at different frequency requirements, including 700, 800, 900, and 950. Table 4.1 shows the number of identified miRNA-genes pairs at each downsampling level with various frequency requirements. As expected, with either increasing frequency requirement or increasing percentage of negatives, the number of significant miRNA-gene pairs decreased.

To evaluate the identified miRNA-gene pairs, we used two held-out miRNA target databases, which are the validation databases not included in the training and prediction process. The evaluation specifically focused on known miRNA-gene pairs that were reported in these validation databases but not in the training databases. For each downsample percentage and each frequency

Table 4.1: Number of miRNA-gene pairs consistently identified as positives by predictive models with different frequency thresholds, and the proportion of these pairs reported in the validation database

DOWNSAMPLE PERCENT	700 TIMES	800 TIMES	900 TIMES	950 TIMES
0.1%	5036798/0.94%	4397542/0.97%	3457560/1.02%	2653818/1.08%
0.4%	1049382/1.29%	736636/1.34%	408392/1.41%	227812/1.49%
0.7%	338755/1.46%	214150/1.53%	104061/1.64%	53939/1.70%
1.0%	143100/1.62%	86201/1.66%	40217/1.77%	20554/1.82%
3.0%	8431/2.19%	5115/2.25%	2680/2.46%	1719/2.39%
5.0%	2857/2.52%	1948/2.72%	1220/2.79%	871/3.21%
7.0%	1683/2.91%	1217/2.96%	829/3.02%	630/3.02%
9.0%	1212/3.14%	913/3.07%	656/3.20%	496/3.02%
10.0%	1062/3.11%	816/2.82%	581/2.75%	434/2.76%

requirement, we computed the percentage of the identified miRNA-gene pairs that were reported in the validation databases but not reported in the training databases. As shown in Table 1, at downsample percentage of 5% and frequency requirement of 950, 871 miRNA-gene pairs were consistently predicted, which achieved the highest validation rate of 3.21%. Interestingly, we observed a similar validation rate for the 656 miRNA-gene pairs identified at downsample percentage of 9% and frequency requirement of 900. After comparing these two sets of the identified miRNA-gene pairs, we noticed that the 656 miRNA-gene pairs were a subset of the 871. Therefore, we decided to focus on the 871 miRNA-gene pairs in the following analyses.

Beyond the 28 (3.21% out of the 871) miRNA-gene pairs confirmed by the validation databases, there may exist more miRNA-gene pairs that have been discovered in the previous studies but were not included in the validation databases. Hence, we performed a literature survey for each of the remaining 843 miRNA-gene pairs. The PubMed database was used to search for previous publications that implicated the identified miRNA-gene pairs, and we found supportive literature for 20 miRNA-gene pairs, which were related to 5 miRNAs, including let-7b, mir-192, mir-26b, mir-335, and mir-93. For example, our predictions suggested a potential relationship between let-7b and EZH2, which aligns with a previous study demonstrating that the loss of let-7b leads to the upregu-

lation of EZH2 expression, consequently promoting ovarian cancer growth both in vitro and in vivo [141]. Another example is miR-335 and TXNIP. A previous study has found that miR-335 targets several differential expressed genes, including TXNIP, and their interaction was associated with oral mucosal wound healing [142]. This previous study supported our prediction of the relationship between miR-335 and TXNIP. As a third example, mir-93 has been reported to control growth and proliferation through FOXO1, which served as a central regulator of endothelial activity[143]. Our analysis also indicates the potential association between mir-93 and FOXO1.

Table 4.2: Comparison of predicted miRNA targets of different methods and known miRNA targets based on differentially expressed genes in independent datasets.

miRNAs	mir-335	mir-335	mir-192	mir-192	mir-26b	mir-193b	mir-193b	mir-193b	mir-21
Dataset	GSE68742	GSE9586	GSE62951	GSE69990	GSE12091	GSE83690	GSE25215	GSE18510	GSE136665
DEGs	1162	1384	1377	1558	968	2108	1162	432	4677
Predicted targets	569	569	50	50	49	42	42	42	12
Known targets	3411	3411	1417	1417	1967	854	854	854	856
Overlap with DEGs:									
Our prediction	31 (5%)	26 (5%)	6 (12%)	4 (8%)	1 (2%)	7 (18%)	8 (20%)	2 (5%)	3 (25%)
PicTar	15 (11%)	17 (13%)	5 (10%)	4 (8%)	2 (4%)	9 (21%)	3 (7%)	0 (0%)	0 (0%)
RNA22	7 (6%)	6 (5%)	3.23 (6%)	3.23 (6%)	2.89 (6%)	4.25 (10%)	2.60 (6%)	0.89 (2%)	3.24 (27%)
miRanda	10 (2%)	2 (0%)	5 (10%)	3 (6%)	4 (8%)	0 (0%)	1 (2%)	0 (0%)	2 (17%)
TargetMiner	40 (7%)	32 (6%)	8 (16%)	2 (4%)	2 (4%)	5 (12%)	4 (10%)	1 (2%)	1 (8%)
PITA	37 (7%)	50 (9%)	6 (12%)	1 (2%)	5 (10%)	10 (24%)	5 (12%)	2 (5%)	3 (25%)
Known targets	130 (4%)	375 (11%)	87 (6%)	186 (13%)	154 (8%)	149 (17%)	98 (11%)	96 (11%)	259 (30%)

4.3.3 Validation of the identified miRNA-gene pairs in independent datasets

To more rigorously confirm the link between the predicted miRNA-gene pairs, we conducted a thorough search for independent datasets regarding each of the miRNAs associated with the predicted miRNA-gene pairs that have not been previously documented in the miRNA targets databases. We searched for datasets that provided gene expression data both before and after perturbation of a specific miRNA, either overexpression or knockdown. As a result, we retrieved 9 GEO datasets, each designed to investigate the consequences after overexpressing a particular miRNA.

A total of 88 miRNA-gene pairs in different cancer contexts were validated in the independent datasets, which involved five miRNAs: mir-335, mir-192, mir-26b, mir-193b, and mir-21. Utilizing the gene expression data before and after a specific miRNA perturbation in each independent dataset,

we performed differential expression analyses to identify the differential expressed genes (DEGs) that showed significant changes due to the miRNA perturbation. These genes may potentially be target genes for the miRNA. Each column of Table 4.2 corresponded to one independent dataset that investigated a particular miRNA. In each column of Table 4.2, we cataloged the number of identified DEGs from one independent dataset, the number of predicted miRNA-gene pairs we predicted for the miRNA, and the number of known target of the miRNA as documented in the five miRNA target databases. Additionally, we listed the intersection between the predicted targets and DEGs, along with the intersection between the known targets and DEGs. For example, in the existing literature, mir-335 has been associated with 3,411 target genes. Of these, 130 (4%) were detected as DEGs in the GSE68742 dataset. Among our predicted 569 target genes for mir-335, 31 (5%) were identified as DEGs. While a vast number of genes have been previously reported in the literature as targets of mir-335, only 4% of them manifested as differentially expressed genes (DEGs) in the GSE68742 dataset. Interestingly, our analysis, which predicted a smaller number of target genes, showed a slightly higher overlap of 5% with the DEGs. This observation was not an isolated occurrence. Upon extending this analysis for all 9 independent miRNA overexpression datasets, we found that our predicted miRNA target genes consistently demonstrated similar level of overlap with DEGs compared to the known target genes in the five miRNA target databases. This close alignment of overlap proportions indicated that our predictive method is potentially as robust as the broader literature in identifying miRNA targets.

We also compared our model's predictive performance against 5 established tools: PicTar [144], RNA22 [145], miRanda [146], TargetMiner [147], and PITA [148] using the 9 independent GEO datasets. For PicTar and RNA22, we assessed their performance using the precomputed predictions from the software platforms, while for miRanda, TargetMiner, and PITA, we computed scores for each miRNA-gene pair based on miRNA and 3' UTR sequencing data from Ensembl [149] and evaluated their performance. Since our model's predicted miRNA target genes consistently exhibited a similar overlap with DEGs as known targets, we extended this analysis to include predictions from existing tools to evaluate whether their predicted targets also showed enrichment in DEGs within the GEO datasets. Specifically, we evaluated the overlap between DEGs from the GEO datasets and the predicted targets from our model and the five other methods. As shown in the last

section of Table 4.2, all six methods achieved comparable levels of DEG enrichment across the 9 GEO datasets, aligning closely with the known targets. In addition, we compared the predicted targets among different methods and found that each method identified distinct sets of targets, with only partial overlap. For the 5 miRNAs listed in Table 4.2, the shared targets between any two methods accounted for an average of 3.8% to 13.4% of the total predicted targets for each method. This pattern of low overlap underscores the variability in prediction outcomes across methods, and suggests that our approach complements existing tools by providing additional insights that may not be captured by sequence-based prediction methods alone.

Table 4.3: Examination of the relationship between the correlation direction of the predicted miRNA-gene pairs and their regulatory direction

miRNA	Datasets	Cancer	Regulatory direction	Sign of correlation	Gene
mir-335	GSE68742	SARC	down	negative	CFLAR, FIGN, MSRB3, SOD3, RGS2, SGK1, PMP22
			down	positive	CCDC3, LDB2, DZIP1, KLF2, CXCL12, NTRK3, PABPC5, RBMS3, RCBTB2, RUNX1T1
			up	negative	A2M, ACTA2, DUSP3, TNFSF10, SGCB, SPARCL1, SSFA2, FAS
			up	positive	EGR2, JDP2, PRICKLE1, LOC728392, MATN2, LMO2
mir-335	GSE9586	LUAD	down	negative	CD247, DPYD, F3, LHFP, PTX3, MNDA, MT1E, MYL9, SCD5, SLFN11
			down	positive	OSR2
			up	negative	ADAM23, CALD1, CALHM2, CTTNBP2NL, DIXDC1, DUSP3, FBXL7, GCC2, GLI3, HVCN1, JAK1, MCTP1, SGCB, RHOJ, PCDHGA9
mir-192	GSE62951	LIHC	down	negative	E2F7, KIF2C, MTBP, NCS1, RAD54L
			up	positive	KIF4A
mir-192	GSE69990	OV	down	negative	SERTAD2
			down	positive	BIRC5, EZH2, PLK1
mir-26b	GSE12091	CESC	down	positive	C1S
mir-193b	GSE83690	SARC	down	positive	C9orf140, CCNB2, CENPA, KIF23, KIF2C, NUSAP1, TROAP
mir-193b	GSE25215	PAAD	up	positive	CCNB2, CDC25C, CENPO, KIF23, KIF2C, PTTG1, TROAP, TTK
mir-193b	GSE18510	SKCM	up	positive	BIRC5, KIF2C
mir-21	GSE136665	COAD	up	negative	ACADSB, NAP1L5
			up	positive	AKAP11

Moreover, the GSE68742 dataset explored the influence of mir-355 overexpression on gene expression within gastrointestinal stromal tumors (GISTs)[150], which are neoplasms closely related to SARC as delineated in the TCGA. As listed in the first column of Table 4.3, the overlap between the DEGs in GSE68742 and our potential mir-355 targets from TCGA is 31 genes. Among these 31 genes, 14 were up-regulated and 17 were down-regulated in the GSE68742 dataset. Since this dataset focused on overexpression of mir-355, the up-regulated genes are expected to have positive correlations with mir-355, while the down-regulated genes should exhibit negative correlations, especially within the context of SARC. We examined the observed correlation direction between the miRNA and these target genes in our TCGA analysis. Of the the 14 up-regulated DEGs, 6 exhibited the expected positive correlations with mir-355. Of the 17 down-regulated DEGs, 7 showed the expected negative correlations, as listed in the first section in Table 4.3. Such analysis of differential expression and correlation directions was conducted for each of the independent validation datasets. Overall, among the total of 88 miRNA-gene pairs (first overlap row in Table 4.2 and last column in Table 4.3) that showed differential expression in the validation datasets, we observed 40 whose directions of differential expression in the validation datasets agreed with their directions of miRNA-gene correlation in our TCGA analysis. These numbers translated to 45% of alignment between the differential expression after miRNA perturbation and the correlation in TCGA. Given the fact that TCGA data was generated from primary tumor samples while most of the independent datasets were generated from in vitro experiments, the 45% of alignment represented interesting miRNA target genes that exhibited robust regulatory patterns in different experimental designs, and can serve as hypotheses for further experimental validation.

4.4 Discussion

In this study, we aimed to identify potential miRNA-gene associations that have not been previously reported in the literature. Specifically, we used multi-omics data of cancer patients to compute correlate features for miRNA-gene pairs, leveraged existing miRNA target databases to define class labels, and applied XGBoost to train machine learning models to predict potential miRNA-gene relationships. Among the 871 predicted significant miRNA-gene pairs, 5.5% were validated using independent held-out miRNA target databases and literature survey, while the remaining could

serve as promising hypotheses for future experimental validation. In addition, we further validated the identified miRNAs by analyzing gene expression profiles before and after specific miRNA perturbations in multiple independent datasets, and we consistently observed an alignment between the correlation direction of the predicted miRNA-gene pairs and their regulatory patterns presented in these datasets. Although there were instances where the correlation direction diverged from their regulatory patterns in the independent datasets, such discrepancies can be attributed to the fact that these validation datasets were derived from cell lines, while the correlations we calculated were based on primary tumor samples profiled by TCGA.

This study is not without limitations. A key issue is that the existing literature provides only a limited number of confirmed miRNA-gene interactions. The lack of data makes it difficult for machine learning models to identify all potential relationships due to the imbalance between confirmed interactions and unreported ones. This imbalance has significant impact on our machine learning model. When working with such imbalanced data that was dominated by the negative class, the model is biased toward predicting negatives. To address this, we performed data downsampling to reduce the prevalence of the negative class and trained our model on various downsampled versions of the data. By doing so, we ensured that the model could focus more on potential miRNA-target relationships rather than being overwhelmed by the prevalence of negative pairs. The imbalance also means that our approach cannot be expected to capture all possible miRNA-target interactions without generating a large number of false positives. Instead, our goal was to focus on identifying high-confidence predictions, which may serve as hypotheses for further experimental validation. As we continue to refine our models, incorporating additional datasets and experimental techniques may help improve the predictive power of our approach.

Nevertheless, our study introduced a new methodology to discover previously unreported miRNA-gene associations, advancing a comprehensive understanding of miRNA-gene interactions. The findings from this investigation offer insight into miRNA regulation and hypotheses for experimental explorations.

CHAPTER 5

CLUSTERING ANALYSIS OF MULTIPLE OMICS DATA TYPES IDENTIFIES CANCER PATIENTS WITH CONSISTENT SURVIVAL OUTCOMES

5.1 Introduction

Cancer is a complex and diverse disease that can develop in various tissues and cell types. Even within a single type of cancer, multiple subtypes may be present, each characterized by distinct molecular profiles, histology, and genetic mutations, leading to varied clinical outcomes and treatment responses. Therefore, significant efforts have been directed toward identifying phenotypic variations as distinct cancer subtypes through molecular characterizations, aiming to categorize patients into distinct groups that exhibit similar clinical phenotypes, prognoses, and responses to treatments [151, 152]. Such classification is intended to facilitate the provision of precise clinical management and customized treatment for each specific subtype, ultimately improving the overall efficacy of cancer therapies [153, 154].

Clustering analysis is a fundamental computational method widely used to group patients into different clusters based on their similarity in genomic data. Patients within the same cluster may exhibit similar clinical outcomes or treatment responses, while those in different clusters may show greater dissimilarities. Several clustering methods, such as Cluster of Clusters (CoC) [155], consensus clustering (CC) [156], and consensus non-negative matrix factorization (CNMF) [157], have been frequently applied in cancer subtype discovery and have demonstrated success in multiple cancer contexts using different omics data types [158, 159, 160]. Typically, in the analysis of multi-omics data for cancer subtyping, these clustering methods aim to produce one clustering result that is supported by the multiple omics datasets. This is often achieved by either finding the consensus of clustering results of individual omics data types, or one integrative analysis that considers all omics data types jointly. In this study, we take a slightly different approach. Instead of finding one unified cluster result supported by multiple individual omics data types, we examine the patient clusters separately defined by different omics data types, and explore their consistency as well as

their correlation to survival outcomes.

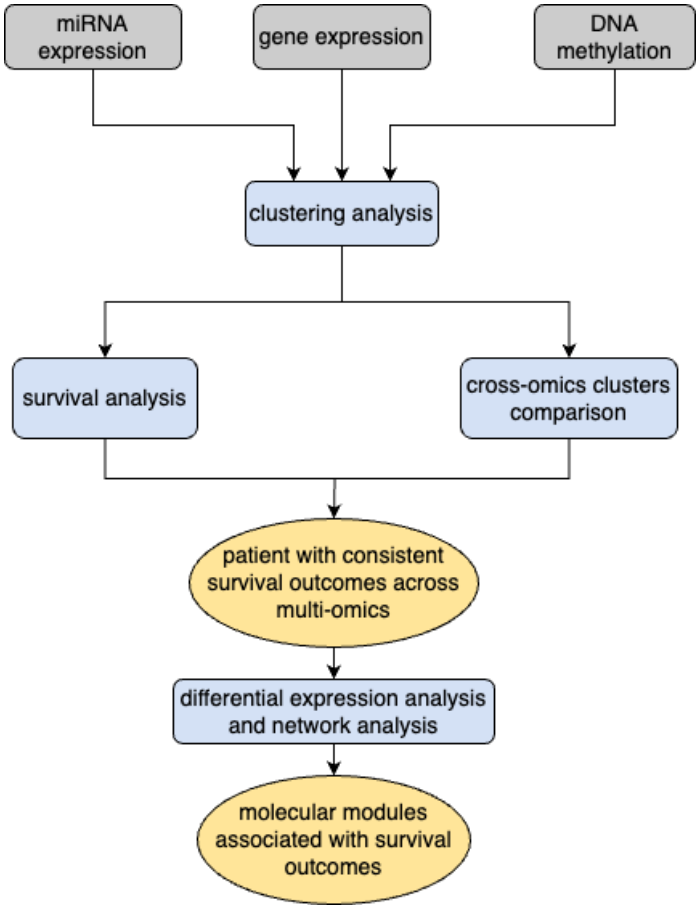

Figure 5.1: Overview of the analysis workflow in this study. The steps include the collection of three omics data types (miRNA expression, gene expression, and DNA methylation) from TCGA, followed by clustering analysis for each omics type. Subsequent analyses involve survival analysis of patient clusters and cross-omics cluster comparison to identify patients with consistent survival outcomes across multiple omics data for each cancer type. Finally, differential expression analysis and network analysis are performed on the identified patient groups to discover molecular modules associated with survival outcomes.

This workflow of this analysis is shown in Figure 5.1, we downloaded multiple omics data types, including miRNA expression, gene expression, and DNA methylation profiles, for all cancer patients available in The Cancer Genome Atlas (TCGA) [161]. We used the Seurat clustering pipeline to define patient clusters based on each omics data types separately. Seurat is an R package developed for scRNA-seq analysis, and offers a comprehensive suite of tools for data preprocessing, normalization, clustering, and visualization [162]. Although primarily tailored for scRNA-seq, Seurat can be adapted for bulk RNA-seq data. For each cancer type and each omics data type, we

used Seurat clustering pipeline to cluster the patients. We examined the similarities among patient clusters defined by different omics data types, and conducted survival analyses on the resulting patient clusters to assess their differences in survival outcomes. Interestingly, our analysis revealed groups of patients consistently captured by clusters from different omics data types, and such patients consistently exhibited either the most favorable or the most unfavorable survival outcomes. Additionally, to understand the molecular mechanisms associated with these patients, we examined differentially expressed molecular features and their interactions.

5.2 Methods

5.2.1 Data acquisition and preprocessing

miRNA, gene expression, CpG methylation expression, and survival data were downloaded from TCGA Genomic Data Commons (GDC) using the GDC Data Transfer Tool. Our analyses included 1,881 miRNAs, 60,498 genes and 485,577 CpG sites. Among the 33 cancer types in TCGA, we focused on the 20 cancers that contained at least 100 patients with miRNA, gene, and methylation data available. In total, there were 5,453 patients across the 20 cancer types used for clustering analyses in this study.

5.2.2 Clustering analysis

Clustering analysis was performed by using the 'Seurat' R package on each omics data type in each cancer type. The Seurat pipeline included identifying highly variable features, computing principle components, and constructing a shared neighborhood graph, and clustering by finding densely connected communities in the graph. We used the default parameters in the Seurat pipeline, except for the ones that related to the identification of highly variable features. Specifically, we selected 3000 top variable features for the analysis of gene expression and methylation data. For the miRNA data, we used all available features as highly variable features, given the limited number of miRNA features.

5.2.3 Survival analysis

For each cancer type and each omics data type, we performed survival analysis among the patient clusters to evaluate survival outcome differences among these clusters. Log-rank test was used to assess the statistical significance of the survival difference, with a threshold of $p < 0.05$ to define significance. Kaplan-Meier analysis and log-rank test were performed using the R package 'survival'. All statistical tests were performed using standard R functions.

5.2.4 Literature survey

A literature search was performed using PubMed, accessed via the National Library of Medicine PubMed interface (http://www.ncbi.nlm.nih.gov/pubmed). We programmatically searched the PubMed database using custom Python scripts. We searched through PubMed for all keywords in all fileds, including the title, abstract and main texts of the articles.

5.2.5 Construction of DE features network and analysis

To examine the relationships among the DE features that are associated with the consistently clustered patients with the most prominent survival outcomes, we obtained miRNA-target genes and gene-gene interactions from the existing databases. Human protein-protein interaction network (PPI) data were obtained from STRING database. We downloaded experimentally verified miRNA-target gene relationships from four databases: mir2Disease, miRecords, TarBase, and miRTarBase. We constructed networks based on relationships among the DE features by using Cytoscape 3.9.1, and applied Cytoscape plug-in MCODE to identify DE modules that are related to the survival outcomes. The default parameters were used throughout the network analyses.

5.3 Results

5.3.1 Clustering analysis of individual omics data types for each cancer type

In the literature, several cancer types have been divided into clinically meaningful subtypes based on the molecular profiles. For example, previous studies have recognized five molecular subtypes

Table 5.1: Number of patient clusters produced by different omics data types for each cancer type

Cancer Type	miRNA	Gene	Methylation
BLCA	4	3	5
BRCA	6	6	6
CESC	2	3	3
COAD	3	4	4
ESCA	2	2	2
HNSC	3	5	5
KIRC	3	4	4
KIRP	2	3	4
LGG	5	5	5
LIHC	3	4	4
LUAD	4	4	4
LUSC	3	4	3
PAAD	2	2	2
PRAD	4	4	4
SARC	3	4	4
SKCM	4	5	3
STAD	4	3	4
TGCT	3	3	4
THCA	4	4	4
UESC	2	3	3

of breast cancer based on the gene expression patterns, and validated the robustness of such classification in various clinical studies of breast cancer [163, 164]. Similarly, skin cutaneous melanomas has been classified into four genomic subtypes based on the mutation patterns of frequently mutated genes [165]. Moreover, De Cecco et al conducted a comprehensive meta-analysis to identify six distinct molecular subtypes of head and neck squamous cell carcinoma, each associated with distinct survival outcomes [166]. These examples highlight the proven effectiveness of using molecular profiling to identify significant cancer subtypes, supporting the validity of using omics-specific clustering to uncover meaningful subtypes.

We performed clustering analyses separately on miRNA expression data, gene expression data and CpG methylation data of each cancer type from TCGA. We selected cancer types with at least 100 patients whose miRNA expression, gene expression and DNA methylation data were available. Overall, 20 cancer types were included in this study, involving a total of 5,453 cancer patients. In terms of omics features, our analysis included 1,881 miRNAs, 60,498 genes and 485,577 CpG sites. For each cancer type and each omics data type, the Seurat clustering pipeline was applied to cluster the patients with that cancer into several clusters. It is important to note that patient clusters may differ depending on the omics data type which was used for clustering. Table 5.1 shows the number of patient clusters identified by each omics data type for each cancer type.

5.3.2 Significant survival differences were observed among patient clusters defined by multiple omics data types in various cancers

To examine the clinical relevance of the patient clusters produced by each omics data type, we applied the log-rank test to assess the survival differences among the corresponding patient clusters. The p-values of the log-rank tests were summarized in Figure 5.2. If the patient clusters exhibited distinct survival outcomes with the log-rank test yielding a p-value of less than 0.05, we considered the clustering result of the omics data type of the cancer type to be survival related. In total, significant survival differences among patient clusters were observed in at least one omics data type in 11 cancers.

Furthermore, for some of the cancer types, we observed significant survival differences among patient clusters in more than one omics data type. Such observation was present in six cancer types, including Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Head and Neck Squamous Cell Carcinoma (HNSC), Kidney Renal Clear Cell Carcinoma (KIRC), Kidney Renal Papillary Cell Carcinoma (KIRP), and Lower Grade Glioma (LGG). Notably, within the contexts of BRCA, KIRC, and LGG, we observed significant survival differences among patient clusters derived from all three omics data types.

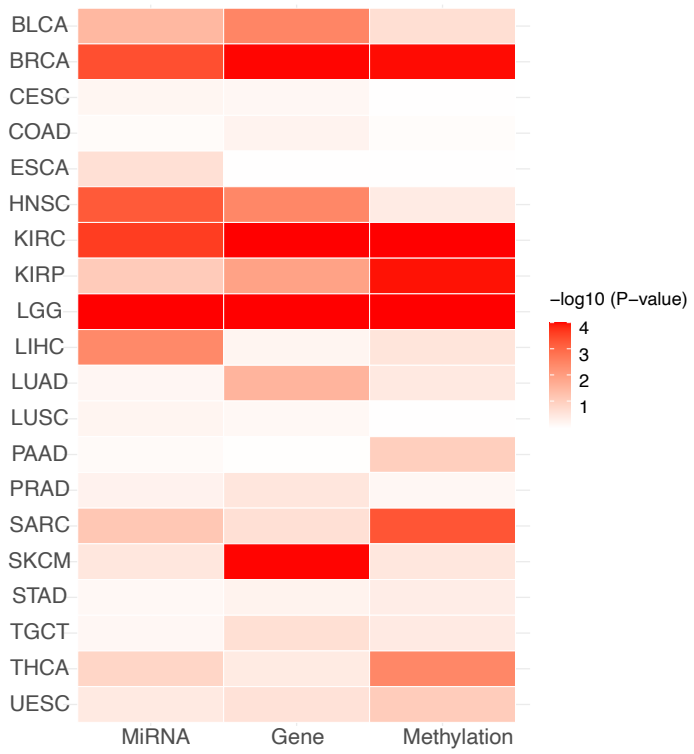

Figure 5.2: Log-rank analysis of patient clusters across 20 cancer types in TCGA. Summary of log-rank analysis to assess survival differences among patient clusters defined by different omics data types for each cancer type. Colors indicate statistical significance, highlighting clinically relevant clusters.

5.3.3 Patient clusters with the most prominent survival outcomes showed a high degree of consistency across omics data types

For each of the six cancer types where at least two omics data type produced patient clusters that exhibited significant survival differences, we further generated Kaplan-Meier (KM) plots to visualize the survival outcomes of those patient clusters. In Figure 5.3, the rows corresponded to the six cancer types, and the columns corresponded to the three omics data types. Each panel shows the KM curves of patient clusters in one cancer type defined by one omics data type.

In the KM plots for BLCA (first row in Figure 5.3), the p-values indicated that significant survival differences were observed in patient clustering analyses of both miRNA and gene expression data. Interestingly, the KM curves for the most favorable patient cluster in miRNA analysis (i.e., labeled as $miRNA_3$) and the most favorable patient cluster in gene expression analysis (i.e., labeled as $gene_2$) appeared to be similar, which was an indication that the $miRNA_3$ patient cluster and the

$gene_2$ patient cluster were likely composed of a highly overlapping set of bladder cancer patients. Indeed, the $miRNA_3$ cluster contained 66 patients, the $gene_2$ cluster contained 74 patients, and the overlap between these two clusters is 56. Such overlap can be quantified by the Jaccard Index (i.e., the ratio of intersection over union), which ranges from 0 to 1. The overlap between patient clusters $miRNA_3$ and $gene_2$ produced a high Jaccard index of 0.67. A previous study classified BLCA into five expression subtypes: luminal-papillary, luminal-infiltrated, luminal, basal-squamous, and neuronal, where luminal-papillary accounted for 33.2% of BLCA patients and exhibited best survival outcomes among the five subtypes [160]. Notably, all 56 overlapping patients identified in our analysis belonged to the luminal-papillary subtype. This strong representation of the luminal-papillary subtype aligns with our observations.

The KM plots for BRCA (second row of Figure 5.3) showed a similar observation that the KM curve for the least favorable patient cluster appeared to be similar across all three omics data types. These least favorable patient clusters were labeled as $miRNA_4$ (33 patients), $gene_5$ (31 patients), and $methyl_5$ (32 patients). The intersection among these three patient clusters was 28, and the pairwise Jaccard index among these three clusters were 0.88, 0.76, and 0.8. Therefore, these three breast cancer patient clusters defined essentially the same set of patients. While comparing this overlap with the well-established PAM50 intrinsic breast cancer subtypes [167] did not reveal significant correlations, the strong consistency across the three clusters suggests that this overlap may still warrant further investigation.

The fourth row of Figure 5.3 showed the KM plots for KIRC. Survival outcome of the least favorable patient clusters in the three data types appeared to be similar. These clusters were labeled as $miRNA_1$ (97 patients), $gene_0$ (101 patients), and $methyl_1$ (76 patients). The intersection among these three patient clusters was 54. The pairwise Jaccard index among these three clusters were 0.55, 0.49 and 0.62. A previous study of renal cell carcinoma has identified a subtype that has 9.9% prevalence and has significantly poor survival outcomes [158] Among the 54 consistently clustered patients in our analysis, 42.5% belonged to this previously established poor survival subtype, which presents significant enrichment that aligns with our observations.

The last row of Figure 5.3 showed the KM plots for LGG. The least favorable patient clusters in the three data types showed similar survival outcome. These clusters were labeled as $miRNA_3$ (47

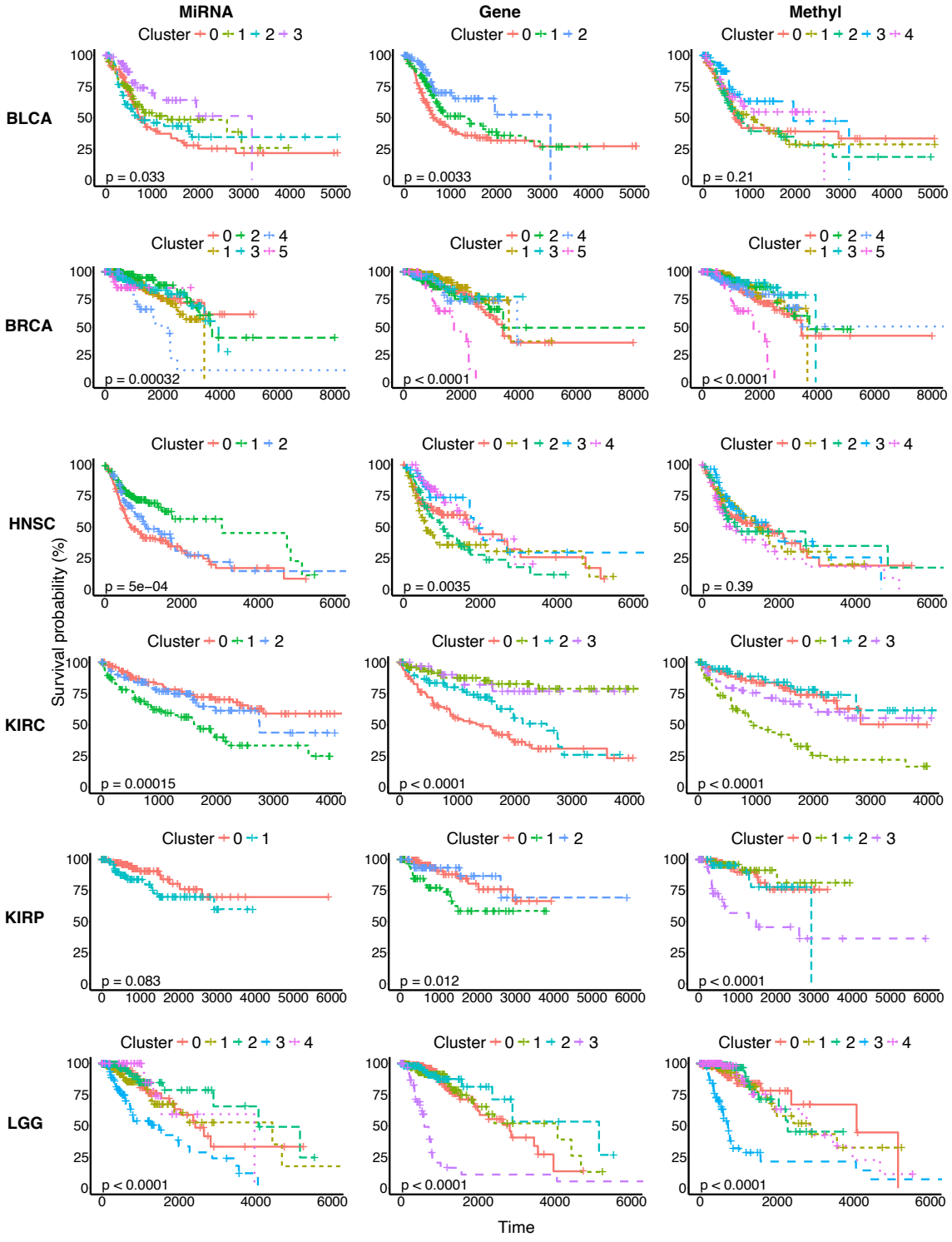

Figure 5.3: Kaplan-Meier survival curves depicting the survival outcomes of patient clusters derived from miRNA, gene expression, and DNA methylation data for each of the six cancer types where significant survival differences were observed in multiple omics. The rows represent different cancer types, and the columns represent different omics data types.

patients), $gene_3$ (63 patients), and $methyl_4$ (80 patients). The intersection among these three patient clusters was 42. The pairwise Jaccard index among these three clusters were 0.62, 0.53 and 0.70. Previous research on diffuse glioma has identified seven glioma subtypes, with the mesenchymal-like and classic-like subtypes being specifically linked to poor prognosis [168]. In our analysis, 57.1% of the 42 overlapping patients were of the mesenchymal-like subtype, and 26.1% were of the classic-like subtype. In comparison, only 8.5% of the broader LGG cohort had the mesenchymal-like subtype, and 4.3% had the classic-like subtype. This notable enrichment of poor-survival subtypes among overlapping patients reinforces the relevance of our observations. Although the overlap and Jaccard index were not as high compared to the cases of BLCA and BRCA, the overlap among these patient clusters did lead to visually appreciable similarity in the survival outcomes visualized by KM curves.

Collectively, these observations showed that in each of these four cancer types, there existed a set of patients that consistently clustered together in different omics data types, and these patients tended to exhibit either the most favorable or least favorable survival outcomes. We conjecture that the sets of patients with the most prominent survival outcomes likely showed distinct expression patterns in multiple genomics aspects, which enabled the clustering analysis to consistently capture them using different omics data types.

Table 5.2: The number of DE features identified in each of the 4 cancer types

Cancer Type	**DE miRNAs**	**DE genes**	**DE CpGs**
BLCA	260	6864	-
BRCA	343	10007	2083
KIRC	161	6120	1438
LGG	178	8172	2109

5.3.4 Differentially expressed features associated with the most prominent survival outcomes

We further focused on BLCA, BRCA, KIRC and LGG, the four cancer types where we observed that different omics data type led to consistently clustered patients with prominent survival outcomes. To explore the biological differences between these consistently clustered patients and the rest of the patients in the same cancer type, we performed Differential expression (DE) analyses to identify

miRNAs, genes, and CpG sites that showed significant differential expression. Specifically, for each cancer type and each omics data type, we applied t-tests to examine the molecular features. The resulting p-values were converted to false discovery rates (FDR), and an FDR threshold of 0.05 was used to identify significant DE features. Table 5.2 listed the number of DE features identified for each omics data type in each cancer type.

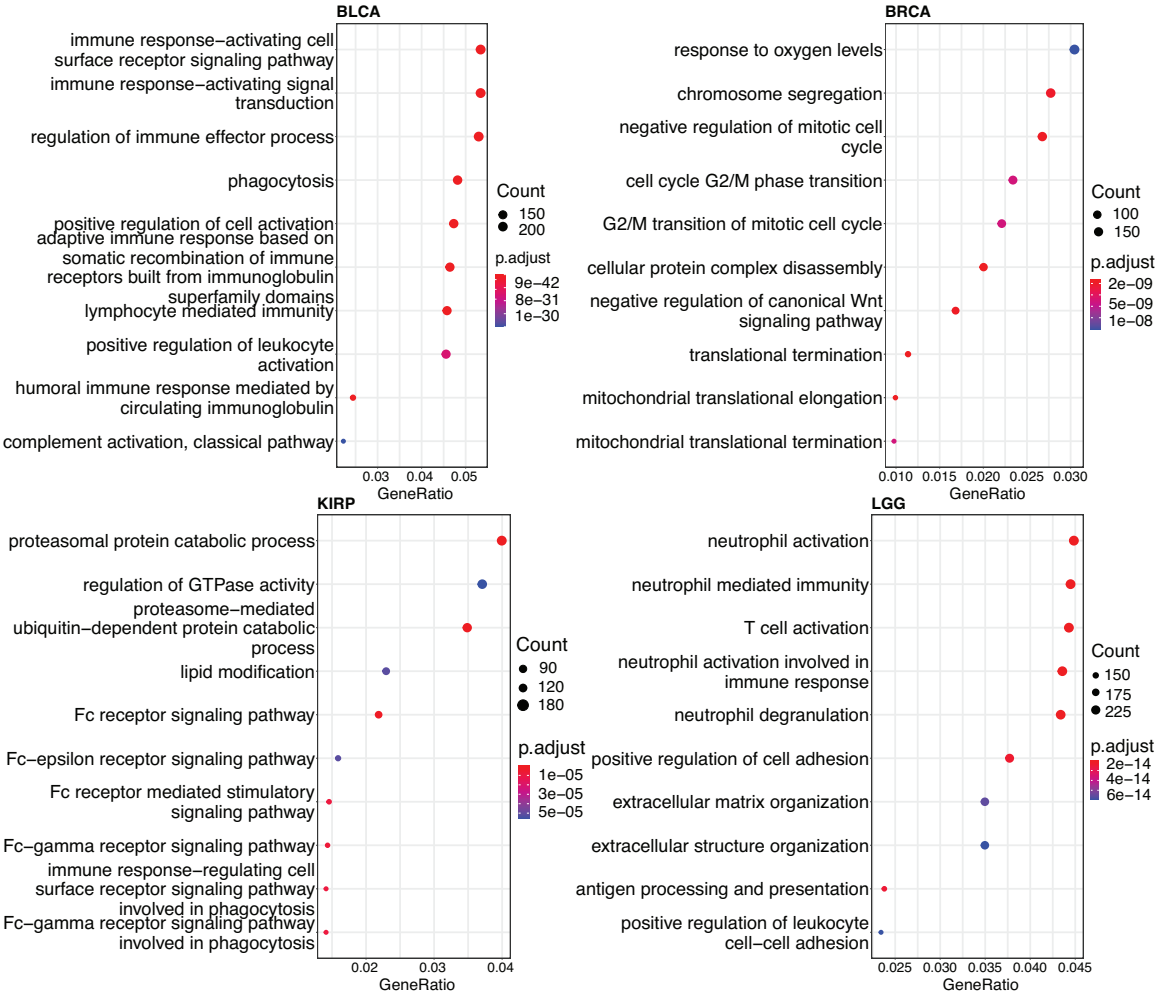

Figure 5.4: Gene Ontology (GO) enrichment analysis results for differentially expressed genes in BLCA, BRCA, KIRC, and LGG. The enriched biological processes are shown for each cancer type, with color intensity representing the significance level of enrichment. The analysis reveals key biological processes linked to cancer development and progression, underlining the molecular characteristics associated with different patient clusters.

We performed functional enrichment analysis to examine the biological roles of the DE genes. Gene Ontology (GO) enrichment analysis was conducted by using R package 'clusterProfiler'[169], searching against the entire Homo sapiens genome. We adopted a p-value threshold of < 0.05 to

identify the top enriched biological processes shown in Figure 5.4. This analysis revealed that the DE genes identified in these four cancer types were significantly enriched in several key biological processes, including cell development and immune response, which are critical for cancer development and progression. Moreover, a literature survey was conducted to examine the functional relevance of identified DE miRNAs for each cancer type. This survey yielded considerable evidence supporting the associations between the DE miRNAs and cancer survival outcomes. For example, in our analysis, we observed that the expression level of let-7 was associated with favorable survival outcomes in BLCA. Several studies have also highlighted the important role of let-7 as a cancer biomarker, and shown that decreased let-7 expression correlated with high-grade tumors and poor prognosis, while high let-7 levels are associated with better prognosis and prolonged patient survival [170, 171, 172, 173]. Another example is miR-155. Iorio et al. showed that miR-155 is significantly up-regulated in human breast cancer compared with normal breast tissue, aligning with our observation that miR-155 is highly expressed in patient clusters with unfavorable survival outcomes in BRCA [174]. Additionally, it has been discovered that high miR-21 expression is linked to poor survival outcome in squamous cell lung carcinoma and low-grade gliomas [162, 175]. Our analysis also revealed that the expression of miR-21 is correlated with unfavorable outcomes in LGG. These findings from the literature, which reported on the relationships between DE miRNAs and cancers, aligned well with our observations, providing support for our results in linking the DE miRNAs to cancer survival outcomes.

5.3.5 Network analyses identified molecular modules associated with patient survival outcomes in cancers

To investigate the underlying biological mechanisms associated with the consistently clustered patients with prominent survival outcomes, we collected and integrated relevant data from several sources, including miRNA-target gene relationships in the mir2Disease [176], miRecords [177], TarBase [178] and miRTarBase [179] databases, as well as protein-protein interaction data from the STRING database [180]. Additionally, we established the relationship between CpGs and genes based on the genomic locations of CpG sites relative to genes. We constructed one miRNA-gene-CpG network for each cancer type, using the identified DE features in the cancer type as nodes and

the relationships among miRNAs, genes and CpG sites in those databases as edges. The miRNA-gene-CpG networks were visualized using Cytoscape 3.9.1 [181]. We further employed the Cytoscape plug-in MCODE [182] to identify densely interconnected network modules, which may indicate molecular complexes relevant to the prominent survival outcomes of the consistently clustered patients. Figure 5.5 shows the identified network modules of DE features in each of the four cancer types.

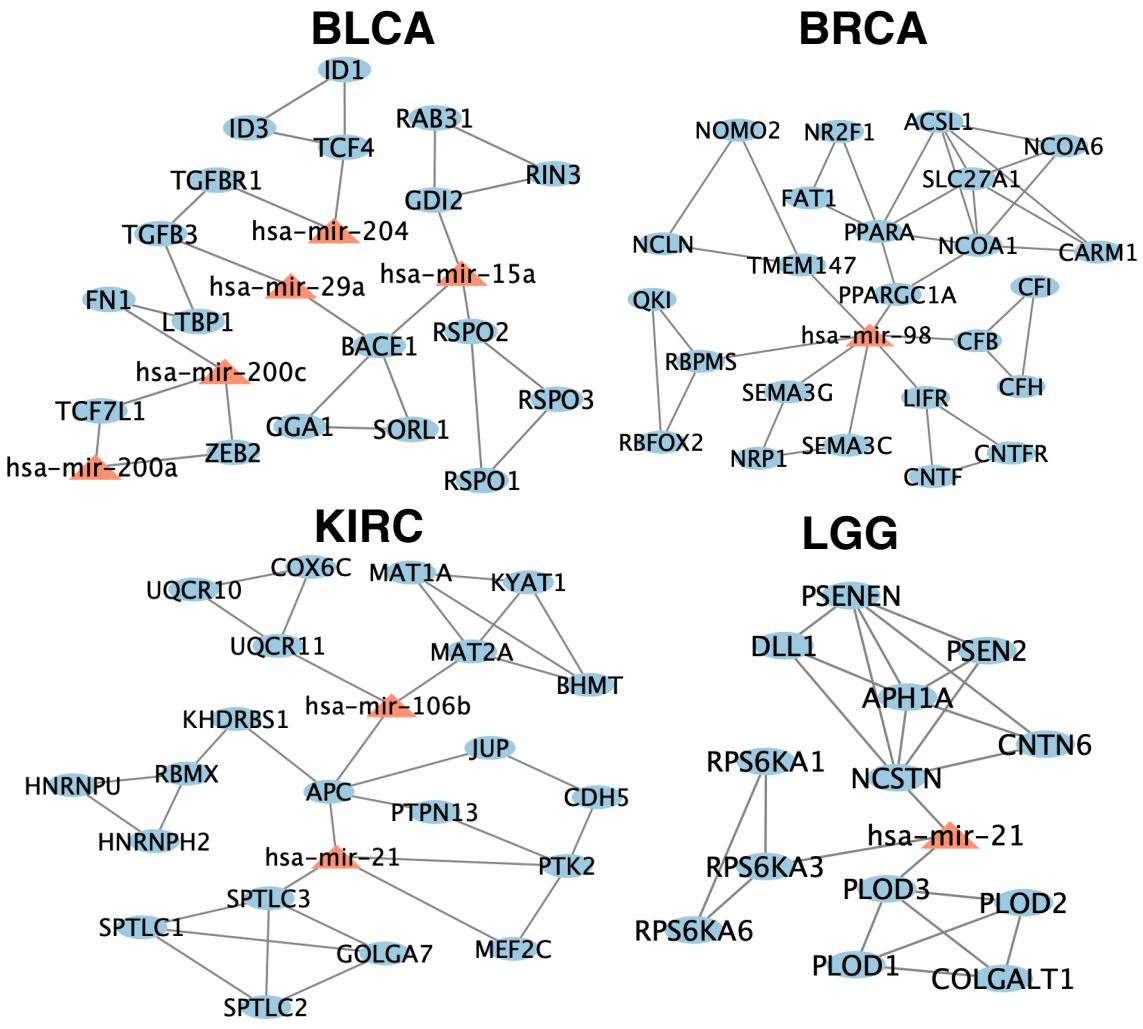

Figure 5.5: Densely connected network modules of DE features associated with consistently clustered patients in BLCA, BRCA, KIRC, and LGG. Nodes are color-coded to represent the type of feature, and edges indicate known interactions among the DE features based on miRNA-gene and gene-gene relationships. These network modules suggest potential molecular interactions that may influence patient survival outcomes.

In our network analysis of BLCA, we observed the association between miR-200c and its target gene ZEB2 in BLCA. This finding is supported by a previous study examining the tumor suppressor role of miR-200c-3p in prostate carcinoma, which suggested that targeting the miR-200c-3p/ZEB2 axis could be a promising design strategy for the treatment of prostate carcinoma [183]. In addition, we found that miR-98 emerged as a pivotal element, connecting target genes within the BRCA module. This is particularly notable as miR-98 is known to promote apoptosis in breast cancer cells [184, 185]. Previous studies have revealed the potential of miR-98 to affect the immune system in endometrial epithelial cells by down-regulating its target gene CFB [186]. Besides CFB, several target genes of miR-98 in the network, such as PPARGC1A, SEMA3C, and NRP1, have been implicated in key processes like metabolism, immune response, cell migration, and angiogenesis, all of which are critical in cancer progression [187, 188, 189]. These results suggest that miR-98 might play a crucial role in breast cancer. Moreover, we identified the regulation between miR-21 and APC gene in the network analysis of KIRC. In the literature, it has been reported that down-regulation of the APC gene in CRC correlated with up-regulation of miR-21, and patients with APC mutation and high miR-21 expression had poor overall survival rates in colorectal cancer [162]. Overall, our analysis underscores the relationships among DE features identified from the consistently clustered patients with prominent survival outcomes. These DE features and their associations, extensively documented in the literature, offer valuable biological insights into the factors correlated with prominent survival outcomes in cancer patients. It is worth noting that the DE CpG features were absent in the densely interconnected network modules for all four cancer types. This absence is likely because the miRNA-gene-CpG networks were constructed using several databases that describe miRNA-gene relationships and gene-gene relationships, while the CpGs were connected to genes only based on genomic location. Therefore, our miRNA-gene-CpG networks contained more connections among miRNAs and genes, which biased the densely connected modules.

5.4 Discussion

In this study, we explored cancer subtyping through multi-omics data types by conducting clustering analyses on miRNA expression, gene expression, and DNA methylation data of 20 cancer types in

TCGA. In 11 cancer types, we observed significant survival differences among patient clusters derived in at least one omics data type. In 6 of the 11 cancer types, significant survival differences among patient clusters were observed in more than one omics data type. We examined the survival outcome of the patient clusters derived from different omics data types. Interestingly, in 4 of the 6 cancer types, we noticed one set of patients who consistently clustered together irrespective of the omics data type, and these patients exhibited either the most favorable or the most unfavorable survival outcomes. This finding suggests that patients with the most prominent survival outcomes may display distinct expression patterns across multiple genomic dimensions, which can be detected by clustering analyses in different omics data types. Furthermore, we identified miRNAs, genes, and CpGs associated with these patient clusters in each cancer type. Based on the relationships among the differentially expressed molecular features, we constructed networks to illustrate the biological mechanisms that may be associated with the prominent survival outcomes of these patient clusters.

However, this study is not without its limitations. One of the limitations is the parameter tuning for Seurat clustering. The Seurat clustering process involves several components, including data filtering and normalization, identification of highly variable features and principal components, as well as construction of shared neighborhood graphs and patient clusters. Each component contains user-defined parameters that may affect the clustering results [190]. In our analysis, to ensure uniformity in comparing clustering analyses across various data types and cancer types, we adhered to the default parameter settings in Seurat. Although the default parameter settings in Seurat may not be optimal, they have been proven useful in many scRNA-seq analyses. In addition, it was encouraging that the default parameter enabled our clustering analyses to stratifying cancer patients into clusters with distinct survival outcomes, suggesting that these parameters are generally well-suited for our analysis. Another limitation of this study was the lack of independent validation datasets. Our analysis in BLCA, BRCA, KIRC and LGG showed that there exist a set of cancer patients with prominent survival outcomes that were consistently clustered across different data types. It would be ideal to validate these findings in an independent dataset. However, such an independent validation requires a multi-omics cancer dataset with miRNA expression, gene expression and DNA methylation data, as well as survival outcome data, for at least one of the four cancer types. However, we were unable to find such a validation dataset. Therefore, our ability to perform independent

validation was limited by data availability. Despite these limitations, it was striking to observe the prominent survival outcomes of the consistently clustered patients in multiple cancer types, and the molecular features associated to these consistently clustered patients pointed to targets and pathways with literature support. Future research could aim to include additional omics data types, such as proteomics and metabolomics, to provide a more comprehensive understanding of cancer biology and cancer subtyping.

CHAPTER 6

CONCLUSIONS AND FUTURE WORKS

This dissertation presents a comprehensive effort to address critical challenges in cancer bioinformatics by developing informatics methods to identify drug-specific molecular markers and decipher genetic regulatory mechanisms. By leveraging multi-omics data and computational tools, this work advances the understanding of cancer biology, uncovers novel biomarkers, and provides meaningful insights to guide cancer treatment strategies.

In Chapter 2, we focused on identifying drug-specific biomarkers predictive of survival outcomes. By integrating transcriptomic profiling, drug treatment information, and survival data, we identified molecular markers capable of stratifying patients based on their response to specific therapies. This analysis not only revealed novel protein and miRNA biomarkers but also revealed their regulatory relationships, including their influence on target and coding gene expression, providing insights into underlying biological mechanisms. Validation through external datasets and the existing literature also underscored the robustness of these findings, demonstrating their potential to inform personalized therapeutic decisions and improve treatment efficacy.

In Chapter 3, we explored the relationship between DNA methylation and alternative splicing. Through the integration of methylation and exon expression data from multiple cancer types, we identified CpG sites significantly associated with exon expression in cancers. These CpG-exon correlations associations were often consistent across different cancer types and were more strongly linked to survival outcomes than those without. In addition, CpG sites showed stronger correlations with exon expression than with isoform expression, suggesting that a major effect of CpG methylation on alternative splicing may be related to the inclusion or exclusion of exons, which subsequently impacts the relative usage of various isoforms. This analysis provides new insights into the role of methylation in the transcriptional process.

Chapter 4 delved into the prediction of novel miRNA-gene interactions using machine learning approaches. By analyzing miRNA and gene expression data across cancer types, we developed predictive models that leveraged known miRNA target databases as training data to predict previously

unreported miRNA-gene interactions. These predicted interactions were validated against independent datasets and through literature review, revealing biologically meaningful regulatory relationships. Furthermore, we validated a subset of these miRNA-gene pairs using external datasets from miRNA perturbation experiments, highlighting their potential role in miRNA-mediated regulation mechanism.

In Chapter 5, we applied clustering analyses on multi-omics data, including miRNA expression, gene expression, and DNA methylation, to identify cancer patient subgroups with consistent survival outcomes. By employing clustering pipelines similar to those used in scRNA-seq analysis, we observed significant survival differences among patient clusters, with some clusters displaying consistent patterns across multiple omics layers. Differentially expressed molecular features associated with these clusters were used to construct cancer-specific networks, revealing key molecular modules linked to survival outcomes. These findings collectively highlight that the clustering approach captures robust patterns across different omics layers, highlighting its potential to guide personalized cancer treatment strategies.

Future efforts will focus on advancing integrative computational approaches to improve patient stratification and uncover predictive biomarkers for therapeutic response. A key challenge remains the identification of robust patterns across diverse genomic and molecular datasets. To address this, I will first focus on enhancing the integration and analysis of multi-omics data to refine cancer subtyping. The Similarity Network Fusion (SNF) algorithm has emerged as a promising method for integrating heterogeneous datasets by constructing a unified similarity network that captures shared and unique biological signals. This approach has shown potential to cluster patients into biologically meaningful subgroups, but its broader application remains limited due to unresolved challenges, including sensitivity to hyperparameters, convergence behavior, and scalability for large datasets. To address these gaps, I propose a systematic evaluation of SNF's parameter space using TCGA datasets. By optimizing parameters such as the number of neighbors, iteration count, and the balance between local and global similarities, this work aims to enhance the algorithm's performance in terms of stability, computational efficiency, and biological relevance. These improvements are expected to enable the identification of robust patient clusters with similar clinical outcomes, such as survival outcomes and treatment responses. Improved parameter settings could reveal common

biological or molecular characteristics across cancers with similar tissue origins, such as liver and bile duct cancers, or previously unrecognized subtypes within complex cancers like breast and lung cancer. Beyond bulk multi-omics data, these insights could extend to single-cell RNA sequencing, where integrating multi-modal profiles may uncover cellular heterogeneity. These outcomes could refine clinical decision-making by enabling more precise patient stratification and uncovering new therapeutic targets.

Complementing this direction, I will also investigate the potential of miRNA expression patterns as predictive markers of drug sensitivity. Integrating miRNA expression profiles from DepMap with drug response data from GDSC, I aim to identify combinations of miRNAs that are predictive of drug sensitivity across cancer cell lines. For each drug, predictive models will be trained using miRNA expression as input features and drug sensitivity as the output, and repeated evaluations will be conducted to ensure robustness. miRNAs consistently ranked as highly important across 1,000 iterations will be selected as predictive markers. These findings will be further validated using independent datasets, such as TCGA, to assess their potential relevance in clinical settings. Together, these future efforts aim to improve our ability to stratify patients, predict treatment outcomes, and identify actionable biomarkers, contributing to the development of more effective and personalized cancer therapies.

In summary, this dissertation lays the groundwork for addressing key challenges in cancer treatment through the integration of multi-omics data and the development of computational tools. The findings contribute to a deeper understanding of cancer biology, providing insights into regulatory mechanisms and survival biomarkers that can guide personalized treatments. Together, these efforts pave the way for innovative approaches that have the potential to improve cancer care.

REFERENCES

[1] S. Chen *et al.*, "Estimates and projections of the global economic cost of 29 cancers in 204 countries and territories from 2020 to 2050," *JAMA oncology*, vol. 9, no. 4, pp. 465–472, 2023.

[2] J. S. Brown, S. R. Amend, R. H. Austin, R. A. Gatenby, E. U. Hammarlund, and K. J. Pienta, "Updating the definition of cancer," *Molecular Cancer Research*, vol. 21, no. 11, pp. 1142–1147, 2023.

[3] I. Dagogo-Jack and A. T. Shaw, "Tumour heterogeneity and resistance to cancer therapies," *Nature reviews Clinical oncology*, vol. 15, no. 2, pp. 81–94, 2018.

[4] P. L. Bedard, A. R. Hansen, M. J. Ratain, and L. L. Siu, "Tumour heterogeneity in the clinic," *Nature*, vol. 501, no. 7467, pp. 355–364, 2013.

[5] F. Lüönd, S. Tiede, and G. Christofori, "Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression," *British journal of cancer*, vol. 125, no. 2, pp. 164–175, 2021.

[6] R. Kaur, A. Bhardwaj, and S. Gupta, "Cancer treatment therapies: Traditional to modern approaches to combat cancers," *Molecular biology reports*, vol. 50, no. 11, pp. 9663–9676, 2023.

[7] I. I. Wistuba, J. G. Gelovani, J. J. Jacoby, S. E. Davis, and R. S. Herbst, "Methodological and practical challenges for personalized cancer therapies," *Nature reviews Clinical oncology*, vol. 8, no. 3, pp. 135–141, 2011.

[8] F. S. Ong *et al.*, "Personalized medicine and pharmacogenetic biomarkers: Progress in molecular oncology testing," *Expert review of molecular diagnostics*, vol. 12, no. 6, pp. 593–602, 2012.

[9] B. Heidecker and J. M. Hare, "The use of transcriptomic biomarkers for personalized medicine," *Heart failure reviews*, vol. 12, pp. 1–11, 2007.

[10] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-omics data integration, interpretation, and its application," *Bioinformatics and biology insights*, vol. 14, p. 1 177 932 219 899 051, 2020.

[11] L. Wyld, R. A. Audisio, and G. J. Poston, "The evolution of cancer surgery and future perspectives," *Nature reviews Clinical oncology*, vol. 12, no. 2, pp. 115–124, 2015.

[12] R. Sullivan *et al.*, "Global cancer surgery: Delivering safe, affordable, and timely cancer surgery," *The lancet oncology*, vol. 16, no. 11, pp. 1193–1224, 2015.

[13] D. L. Gustafson and R. L. Page, "Cancer chemotherapy," *Withrow and MacEwen's small animal clinical oncology*, pp. 157–179, 2013.

[14] R. Baskar, K. A. Lee, R. Yeo, and K.-W. Yeoh, "Cancer and radiation therapy: Current advances and future directions," *International journal of medical sciences*, vol. 9, no. 3, p. 193, 2012.

[15] M. Verma, "Personalized medicine and cancer," *Journal of personalized medicine*, vol. 2, no. 1, pp. 1–14, 2012.

[16] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," *New England Journal of Medicine*, vol. 363, no. 4, pp. 301–304, 2010.

[17] A.-M. Tsimberidou, "Targeted therapy in cancer," *Cancer chemotherapy and pharmacology*, vol. 76, pp. 1113–1132, 2015.

[18] F. Stegmeier, M. Warmuth, W. Sellers, and M. Dorsch, "Targeted cancer therapies in the twenty-first century: Lessons from imatinib," *Clinical Pharmacology & Therapeutics*, vol. 87, no. 5, pp. 543–552, 2010.

[19] S. Di Cosimo and J. Baselga, "Targeted therapies in breast cancer: Where are we now?" *European Journal of Cancer*, vol. 44, no. 18, pp. 2781–2790, 2008.

[20] K. Esfahani, L. Roudaia, N. a. Buhlaiga, S. Del Rincon, N. Papneja, and W. Miller, "A review of cancer immunotherapy: From the past, to the present, to the future," *Current Oncology*, vol. 27, no. s2, pp. 87–97, 2020.

[21] H. Mamdani, S. Matosevic, A. B. Khalid, G. Durm, and S. I. Jalal, "Immunotherapy in lung cancer: Current landscape and future directions," *Frontiers in immunology*, vol. 13, p. 823 618, 2022.

[22] A. Lahiri *et al.*, "Lung cancer immunotherapy: Progress, pitfalls, and promises," *Molecular cancer*, vol. 22, no. 1, p. 40, 2023.

[23] I. N. Ivashko and J. M. Kolesar, "Pembrolizumab and nivolumab: Pd-1 inhibitors for advanced melanoma," *American journal of health-system pharmacy*, vol. 73, no. 4, pp. 193–201, 2016.

[24] S. Méndez-Ferrer *et al.*, "Bone marrow niches in haematological malignancies," *Nature Reviews Cancer*, vol. 20, no. 5, pp. 285–298, 2020.

[25] G. Schochetman, C.-Y. Ou, and W. K. Jones, "Polymerase chain reaction," *The Journal of infectious diseases*, vol. 158, no. 6, pp. 1154–1157, 1988.

[26] A. Butte, "The use and analysis of microarray data," *Nature reviews drug discovery*, vol. 1, no. 12, pp. 951–960, 2002.

[27] S. Behjati and P. S. Tarpey, "What is next generation sequencing?" *Archives of Disease in Childhood-Education and Practice*, vol. 98, no. 6, pp. 236–238, 2013.

[28] K. R. Kukurba and S. B. Montgomery, "Rna sequencing and analysis," *Cold Spring Harbor Protocols*, vol. 2015, no. 11, pdb–top084970, 2015.

[29] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann, "The technology and biology of single-cell rna sequencing," *Molecular cell*, vol. 58, no. 4, pp. 610–620, 2015.

[30] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[31] F. A. Sinicrope, K. Okamoto, P. M. Kasi, and H. Kawakami, "Molecular biomarkers in the personalized treatment of colorectal cancer," *Clinical Gastroenterology and Hepatology*, vol. 14, no. 5, pp. 651–658, 2016.

[32] S. M. Dhanasekaran *et al.*, "Delineation of prognostic biomarkers in prostate cancer," *Nature*, vol. 412, no. 6849, pp. 822–826, 2001.

[33] C. O. Madu and Y. Lu, "Novel diagnostic biomarkers for prostate cancer," *Journal of Cancer*, vol. 1, p. 150, 2010.

[34] E. Kobayashi *et al.*, "Biomarkers for screening, diagnosis, and monitoring of ovarian cancer," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 21, no. 11, pp. 1902–1912, 2012.

[35] B. Burington, B. Barlogie, F. Zhan, J. Crowley, and J. D. Shaughnessy, "Tumor cell gene expression changes following short-term in vivo exposure to single agent chemotherapeutics are related to survival in multiple myeloma," *Clinical Cancer Research*, vol. 14, no. 15, pp. 4821–4829, 2008.

[36] J. C. G. Spainhour and P. Qiu, "Identification of gene-drug interactions that impact patient survival in tcga," *BMC bioinformatics*, vol. 17, no. 1, pp. 1–8, 2016.

[37] C. Garnis, T. P. Buys, and W. L. Lam, "Genetic alteration and gene expression modulation during cancer progression," *Molecular Cancer*, vol. 3, pp. 1–23, 2004.

[38] Y. Audic and R. S. Hartley, "Post-transcriptional regulation in cancer," *Biology of the Cell*, vol. 96, no. 7, pp. 479–498, 2004.

[39] I. M. Dykes and C. Emanueli, "Transcriptional and post-transcriptional gene regulation by long non-coding rna," *Genomics, Proteomics and Bioinformatics*, vol. 15, no. 3, pp. 177–186, 2017.

[40] G. A. Garinis, G. P. Patrinos, N. E. Spanakis, and P. G. Menounos, "Dna hypermethylation: When tumour suppressor genes go silent," *Human genetics*, vol. 111, no. 2, pp. 115–127, 2002.

[41] M. D. Long, D. J. Smiraglia, and M. J. Campbell, "The genomic impact of dna cpg methylation on gene expression; relationships in prostate cancer," *Biomolecules*, vol. 7, no. 1, p. 15, 2017.

[42] R. Shayevitch, D. Askayo, I. Keydar, and G. Ast, "The importance of dna methylation of exons on alternative splicing," *Rna*, vol. 24, no. 10, pp. 1351–1362, 2018.

[43] A. K. Maunakea, I. Chepelev, K. Cui, and K. Zhao, "Intragenic dna methylation modulates alternative splicing by recruiting mecp2 to promote exon recognition," *Cell research*, vol. 23, no. 11, pp. 1256–1269, 2013.

[44] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg, "Mechanisms of post-transcriptional regulation by microrna: Are the answers in sight?" *Nature reviews genetics*, vol. 9, no. 2, pp. 102–114, 2008.

[45] A. Shenoy and R. H. Blelloch, "Regulation of microrna function in somatic stem cell proliferation and differentiation," *Nature reviews Molecular cell biology*, vol. 15, no. 9, pp. 565–576, 2014.

[46] R. Visone and C. M. Croce, "Mirnas and cancer," *The American journal of pathology*, vol. 174, no. 4, pp. 1131–1138, 2009.

[47] O. Menyhárt and B. Győrffy, "Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis," *Computational and structural biotechnology journal*, vol. 19, pp. 949–960, 2021.

[48] J. C. G. Spainhour, J. Lim, and P. Qiu, "Gdisc: A web portal for integrative analysis of gene–drug interaction for survival in cancer," *Bioinformatics*, vol. 33, no. 9, pp. 1426–1428, 2017.

[49] N. Lynam-Lennon, S. G. Maher, and J. V. Reynolds, "The roles of microrna in cancer and apoptosis," *Biological Reviews*, vol. 84, no. 1, pp. 55–71, 2009.

[50] M. Garofalo, G. Condorelli, and C. M. Croce, "Microrna in diseases and drug response," *Current opinion in pharmacology*, vol. 8, no. 5, pp. 661–667, 2008.

[51] X. Liu *et al.*, "Construction of a potential breast cancer-related mirna-mrna regulatory network," *BioMed research international*, vol. 2020, 2020.

[52] R. Ma, C. Wang, J. Wang, D. Wang, and J. Xu, "Mirna–mrna interaction network in non-small cell lung cancer," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 8, no. 3, pp. 209–219, 2016.

[53] E. Andrés-León, I. Cases, S. Alonso, and A. M. Rojas, "Novel mirna-mrna interactions conserved in essential cancer pathways," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[54] T. Mazza *et al.*, "Multifaceted enrichment analysis of rna–rna crosstalk reveals cooperating micro-societies in human colorectal cancer," *Nucleic acids research*, vol. 44, no. 9, pp. 4025–4036, 2016.

[55] A. Piepoli *et al.*, "Mirna expression profiles identify drivers in colorectal and pancreatic cancers," *PloS one*, vol. 7, no. 3, e33663, 2012.

[56] A. Latini, P. Borgiani, G. Novelli, and C. Ciccacci, "Mirnas in drug response variability: Potential utility as biomarkers for personalized medicine," *Pharmacogenomics*, vol. 20, no. 14, pp. 1049–1059, 2019.

[57] Y. Ma *et al.*, "Predicting cancer drug response by proteomic profiling," *Clinical cancer research*, vol. 12, no. 15, pp. 4583–4589, 2006.

[58] M. Munoz, M. Henderson, M. Haber, and M. Norris, "Role of the mrp1/abcc1 multidrug transporter protein in cancer," *IUBMB Life*, vol. 59, no. 12, pp. 752–757, 2007.

[59] L. C. Young, B. G. Campling, S. P. Cole, R. G. Deeley, and J. H. Gerlach, "Multidrug resistance proteins mrp3, mrp1, and mrp2 in lung cancer: Correlation of protein levels with drug response and messenger rna levels," *Clinical Cancer Research*, vol. 7, no. 6, pp. 1798–1804, 2001.

[60] G. S. Ginsburg and H. F. Willard, *Essentials of Genomic and Personalized Medicine*. Academic Press, 2009.

[61] M.-A. Dillies *et al.*, "A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis," *Briefings in bioinformatics*, vol. 14, no. 6, pp. 671–683, 2013.

[62] D. Sahoo, D. L. Dill, R. Tibshirani, and S. K. Plevritis, "Extracting binary signals from microarray time-course data," *Nucleic acids research*, vol. 35, no. 11, pp. 3705–3712, 2007.

[63] D. Sahoo, D. L. Dill, R. Tibshirani, and S. K. Plevritis, "Extracting binary signals from microarray time-course data," *Nucleic acids research*, vol. 35, no. 11, pp. 3705–3712, 2007.

[64] Y. Liu *et al.*, "Circhipk3 promotes gemcitabine (gem) resistance in pancreatic cancer cells by sponging mir-330-5p and targets rassf1," *Cancer Management and Research*, vol. 12, p. 921, 2020.

[65] J. Hodzic, E. Giovannetti, B. D. Calvo, A. Adema, and G. Peters, "Regulation of deoxycytidine kinase expression and sensitivity to gemcitabine by micro-rna 330 and promoter methylation in cancer cells," *Nucleosides, Nucleotides and Nucleic Acids*, vol. 30, no. 12, pp. 1214–1222, 2011.

[66] L. Zhu, H. Deng, J. Hu, S. Huang, J. Xiong, and J. Deng, "The promising role of mir-296 in human cancer," *Pathology-Research and Practice*, vol. 214, no. 12, pp. 1915–1922, 2018.

[67] Y. Dong *et al.*, "Circ_0076305 regulates cisplatin resistance of non-small cell lung cancer via positively modulating stat3 by sponging mir-296–5p," *Life sciences*, vol. 239, p. 116 984, 2019.

[68] Y. Bai *et al.*, "Mir-296-3p regulates cell growth and multi-drug resistance of human glioblastoma by targeting ether-a-go-go (eag1)," *European journal of cancer*, vol. 49, no. 3, pp. 710–724, 2013.

[69] G. Pan, Y. Liu, L. Shang, F. Zhou, and S. Yang, "Emt-associated micrornas and their roles in cancer stemness and drug resistance," *Cancer Communications*, vol. 41, no. 3, pp. 199–217, 2021.

[70] Q. Fu *et al.*, "Miromics and proteomics reveal a mir-296-3p/prkca/fak/ras/c-myc feedback loop modulated by hdgf/ddx5/β-catenin complex in lung adenocarcinoma," *Clinical Cancer Research*, vol. 23, no. 20, pp. 6336–6350, 2017.

[71] J. Okazaki *et al.*, "Microrna-296-5p promotes cell invasion and drug resistance by targeting bcl2-related ovarian killer, leading to a poor prognosis in pancreatic cancer," *Digestion*, pp. 1–13, 2019.

[72] S. Fan *et al.*, "Mir-483-5p determines mitochondrial fission and cisplatin sensitivity in tongue squamous cell carcinoma by targeting fis1," *Cancer Letters*, vol. 362, no. 2, pp. 183–191, 2015.

[73] O. F. Karatas, M. Oner, A. Abay, and A. Diyapoglu, "Micrornas in human tongue squamous cell carcinoma: From pathogenesis to therapeutic implications," *Oral oncology*, vol. 67, pp. 124–130, 2017.

[74] N. Arrighetti *et al.*, "Pkc-alpha modulation by mir-483-3p in platinum-resistant ovarian carcinoma cells," *Toxicology and applied pharmacology*, vol. 310, pp. 9–19, 2016.

[75] H. Zhao *et al.*, "Mir-483-5p aggravates cisplatin-induced premature ovarian insufficiency in rats by targeting fkbp4," *Nan Fang yi ke da xue xue bao= Journal of Southern Medical University*, vol. 41, no. 6, pp. 801–810, 2021.

[76] Q. Jiang *et al.*, "Mir2disease: A manually curated database for microrna deregulation in human disease," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D98–D104, 2009.

[77] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, "Mirecords: An integrated resource for microrna–target interactions," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D105–D110, 2009.

[78] P. Sethupathy, B. Corda, and A. G. Hatzigeorgiou, "Tarbase: A comprehensive database of experimentally supported animal microrna targets," *Rna*, vol. 12, no. 2, pp. 192–197, 2006.

[79] S.-D. Hsu *et al.*, "Mirtarbase: A database curates experimentally validated microrna–target interactions," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D163–D169, 2011.

[80] M. Ashburner *et al.*, "Gene ontology: Tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[81] M. Kanehisa and S. Goto, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[82] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nature protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[83] B. Neary, J. Zhou, and P. Qiu, "Identifying gene expression patterns associated with drug-specific survival in cancer patients," *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.

[84] H. Ozawa *et al.*, "Smad4 loss is associated with cetuximab resistance and induction of mapk/jnk activation in head and neck cancer cells," *Clinical Cancer Research*, vol. 23, no. 17, pp. 5162–5175, 2017.

[85] D. Sahoo, D. L. Dill, R. Tibshirani, and S. K. Plevritis, "Extracting binary signals from microarray time-course data," *Nucleic Acids Research*, vol. 35, no. 11, pp. 3705–3712, 2007.

[86] M. A. Khan, H. Zubair, S. K. Srivastava, S. Singh, and A. P. Singh, "Insights into the role of micrornas in pancreatic cancer pathogenesis: Potential for diagnosis, prognosis, and therapy," in *microRNA: Cancer*, Springer, Cham, 2015, pp. 71–87.

[87] X. Chen *et al.*, "Therapeutic effects of argyrin f in pancreatic adenocarcinoma," *Cancer Letters*, vol. 399, pp. 20–28, 2017.

[88] S. Ni *et al.*, "Cbx7 suppresses cell proliferation, migration, and invasion through the inhibition of pten/akt signaling in pancreatic cancer," *Oncotarget*, vol. 8, no. 5, p. 8010, 2017.

[89] Y. Tian *et al.*, "Metformin mediates resensitivity to 5-fluorouracil in hepatocellular carcinoma via the suppression of yap," *Oncotarget*, vol. 7, no. 29, p. 46 230, 2016.

[90] T. E. Hutson *et al.*, "Randomized phase iii trial of temsirolimus versus sorafenib as second-line therapy after sunitinib in patients with metastatic renal cell carcinoma," *Journal of Clinical Oncology*, vol. 32, no. 8, p. 760, 2014.

[91] S. L. Abrams *et al.*, "Introduction of wt-tp53 into pancreatic cancer cells alters sensitivity to chemotherapeutic drugs, targeted therapeutics and nutraceuticals," *Advances in Biological Regulation*, vol. 69, pp. 16–34, 2018.

[92] N. Aissat *et al.*, "Antiproliferative effects of rapamycin as a single agent and in combination with carboplatin and paclitaxel in head and neck cancer cell lines," *Cancer Chemotherapy and Pharmacology*, vol. 62, no. 2, pp. 305–313, 2008.

[93] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation between protein and mrna abundance in yeast," *Molecular and Cellular Biology*, vol. 19, no. 3, pp. 1720–1730, 1999.

[94] G. Chen *et al.*, "Discordant protein and mrna expression in lung adenocarcinomas," *Molecular & Cellular Proteomics*, vol. 1, no. 4, pp. 304–313, 2002.

[95] R. A. Rippe and B. Stefanovic, "Methods for assessing the molecular mechanisms controlling gene regulation," in *Fibrosis Research*, Springer, 2005, pp. 141–160.

[96] C. Yin, Q. Mou, X. Pan, G. Zhang, H. Li, and Y. Sun, "Mir-577 suppresses epithelial-mesenchymal transition and metastasis of breast cancer by targeting rab25," *Thoracic cancer*, vol. 9, no. 4, pp. 472–479, 2018.

[97] S. Sharma, T. K. Kelly, and P. A. Jones, "Epigenetics in cancer," *Carcinogenesis*, vol. 31, no. 1, pp. 27–36, 2010.

[98] P. A. Jones, "Functions of dna methylation: Islands, start sites, gene bodies and beyond," *Nature Reviews Genetics*, vol. 13, no. 7, pp. 484–492, 2012.

[99] D. Schübeler, "Function and information content of dna methylation," *Nature*, vol. 517, no. 7534, pp. 321–326, 2015.

[100] M. Klutstein, D. Nejman, R. Greenfield, and H. Cedar, "Dna methylation in cancer and aging," *Cancer research*, vol. 76, no. 12, pp. 3446–3450, 2016.

[101] A. Razin and H. Cedar, "Dna methylation and gene expression," *Microbiological reviews*, vol. 55, no. 3, pp. 451–458, 1991.

[102] J. C. Spainhour, H. S. Lim, S. V. Yi, and P. Qiu, "Correlation patterns between dna methylation and gene expression in the cancer genome atlas," *Cancer informatics*, vol. 18, p. 1 176 935 119 828 776, 2019.

[103] S. B. Baylin, "Dna methylation and gene silencing in cancer," *Nature clinical practice Oncology*, vol. 2, no. 1, S4–S11, 2005.

[104] J. Attwood, R. Yung, and B. Richardson, "Dna methylation and the regulation of gene transcription," *Cellular and Molecular Life Sciences CMLS*, vol. 59, no. 2, pp. 241–257, 2002.

[105] K. B. Halpern, T. Vana, and M. D. Walker, "Paradoxical role of dna methylation in activation of foxa2 gene expression during endoderm development," *Journal of Biological Chemistry*, vol. 289, no. 34, pp. 23 882–23 892, 2014.

[106] S. C. Bonnal, I. López-Oreja, and J. Valcárcel, "Roles and mechanisms of alternative splicing in cancer—implications for care," *Nature Reviews Clinical Oncology*, vol. 17, no. 8, pp. 457–474, 2020.

[107] Z. Siegfried and R. Karni, "The role of alternative splicing in cancer drug resistance," *Current opinion in genetics & development*, vol. 48, pp. 16–21, 2018.

[108] X. Sun, Y. Tian, J. Wang, Z. Sun, and Y. Zhu, "Genome-wide analysis reveals the association between alternative splicing and dna methylation across human solid tumors," *BMC medical genomics*, vol. 13, no. 1, pp. 1–12, 2020.

[109] D. Kim, S. Han, and Y. Lee, *Crosstalk between methylation and alternative splicing in cancer*, 2018.

[110] N. Giannoukakis, C. Deal, J. Paquette, C. G. Goodyer, and C. Polychronakos, "Parental genomic imprinting of the human igf2 gene," *Nature genetics*, vol. 4, no. 1, pp. 98–101, 1993.

[111] J. H. Shin *et al.*, "Butyrate induced igf2 activation correlated with distinct chromatin signatures due to histone modification," *Gene regulation and systems biology*, vol. 7, GRSB–S11243, 2013.

[112] D. Monk *et al.*, "Imprinting of igf2 p0 transcript and novel alternatively spliced ins-igf2 isoforms show differences between mouse and human," *Human molecular genetics*, vol. 15, no. 8, pp. 1259–1269, 2006.

[113] V. K. Radhakrishnan *et al.*, "Methylation of a newly identified region of the ins-igf2 gene determines igf2 expression in breast cancer tumors and in breast cancer cells," *Oncotarget*, vol. 11, no. 44, p. 3904, 2020.

[114] C. S. Rosenfeld *et al.*, "Hypothalamic transcriptome of tame and aggressive silver foxes (vulpes vulpes) identifies gene expression differences shared across brain regions," *Genes, Brain and Behavior*, vol. 19, no. 1, e12614, 2020.

[115] R. J. Taft, K. C. Pang, T. R. Mercer, M. Dinger, and J. S. Mattick, "Non-coding rnas: Regulators of disease," *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, vol. 220, no. 2, pp. 126–139, 2010.

[116] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, "Human microrna targets," *PLoS biology*, vol. 2, no. 11, e363, 2004.

[117] D. P. Bartel, "Micrornas: Genomics, biogenesis, mechanism, and function," *cell*, vol. 116, no. 2, pp. 281–297, 2004.

[118] H. Tan, S. Huang, Z. Zhang, X. Qian, P. Sun, and X. Zhou, "Pan-cancer analysis on microrna-associated gene activation," *EBioMedicine*, vol. 43, pp. 82–97, 2019.

[119] L. He and G. J. Hannon, "Micrornas: Small rnas with a big role in gene regulation," *Nature reviews genetics*, vol. 5, no. 7, pp. 522–531, 2004.

[120] W. P. Kloosterman and R. H. Plasterk, "The diverse functions of micrornas in animal development and disease," *Developmental cell*, vol. 11, no. 4, pp. 441–450, 2006.

[121] L. Huang, L. Zhang, and X. Chen, "Updated review of advances in micrornas and complex diseases: Taxonomy, trends and challenges of computational models," *Briefings in bioinformatics*, vol. 23, no. 5, bbac358, 2022.

[122] L. Huang, L. Zhang, and X. Chen, "Updated review of advances in micrornas and complex diseases: Experimental results, databases, webservers and data fusion," *Briefings in bioinformatics*, vol. 23, no. 6, bbac397, 2022.

[123] L. Huang, L. Zhang, and X. Chen, "Updated review of advances in micrornas and complex diseases: Towards systematic evaluation of computational models," *Briefings in bioinformatics*, vol. 23, no. 6, bbac407, 2022.

[124] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, "Micrornas and complex diseases: From experimental results to computational models," *Briefings in bioinformatics*, vol. 20, no. 2, pp. 515–539, 2019.

[125] L. P. Lim *et al.*, "Microarray analysis shows that some micrornas downregulate large numbers of target mrnas," *Nature*, vol. 433, no. 7027, pp. 769–773, 2005.

[126] I. Bentwich, "Prediction and validation of micrornas and their targets," *FEBS letters*, vol. 579, no. 26, pp. 5904–5910, 2005.

[127] J. R. Brown and P. Sanseau, "A computational view of micrornas and their targets," *Drug Discovery Today*, vol. 10, no. 8, pp. 595–601, 2005.

[128] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "Mirbase: From microrna sequences to function," *Nucleic acids research*, vol. 47, no. D1, pp. D155–D162, 2019.

[129] D. Karagkouni *et al.*, "Diana-tarbase v8: A decade-long collection of experimentally supported mirna–gene interactions," *Nucleic acids research*, vol. 46, no. D1, pp. D239–D245, 2018.

[130] H.-Y. Huang *et al.*, "Mirtarbase 2020: Updates to the experimentally validated microrna–target interaction database," *Nucleic acids research*, vol. 48, no. D1, pp. D148–D154, 2020.

[131] J. Krüger and M. Rehmsmeier, "Rnahybrid: Microrna target prediction easy, fast and flexible," *Nucleic acids research*, vol. 34, no. suppl_2, W451–W454, 2006.

[132] Y. Chen and X. Wang, "Mirdb: An online database for prediction of functional microrna targets," *Nucleic acids research*, vol. 48, no. D1, pp. D127–D131, 2020.

[133] M. Kiriakidou *et al.*, "A combined computational-experimental approach predicts human microrna targets," *Genes & development*, vol. 18, no. 10, pp. 1165–1178, 2004.

[134] N. Rajewsky, "Microrna target predictions in animals," *Nature genetics*, vol. 38, no. Suppl 6, S8–S13, 2006.

[135] T. M Witkos, E. Koscianska, and W. J Krzyzosiak, "Practical aspects of microrna target prediction," *Current molecular medicine*, vol. 11, no. 2, pp. 93–109, 2011.

[136] 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*, vol. 120, no. 1, pp. 15–20, 2005.

[137] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "Review the cancer genome atlas (tcga): An immeasurable source of knowledge," *Contemporary Oncology/Współczesna Onkologia*, vol. 2015, no. 1, pp. 68–77, 2015.

[138] T. Chen *et al.*, "Xgboost: Extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[139] Z. Zhao, H. Peng, C. Lan, Y. Zheng, L. Fang, and J. Li, "Imbalance learning for the prediction of n 6-methylation sites in mrnas," *BMC genomics*, vol. 19, pp. 1–10, 2018.

[140] W. Li, Y. Yin, X. Quan, and H. Zhang, "Gene expression value prediction based on xgboost algorithm," *Frontiers in genetics*, vol. 10, p. 1077, 2019.

[141] Y. Kuang *et al.*, "Inhibition of microrna let-7b expression by kdm2b promotes cancer progression by targeting ezh2 in ovarian cancer," *Cancer Science*, vol. 112, no. 1, pp. 231–242, 2021.

[142] W. Ning *et al.*, "Identification of the potential biomarkers involved in the human oral mucosal wound healing: A bioinformatic study," *BioMed Research International*, vol. 2021, pp. 1–16, 2021.

[143] Z. Feng, R. Chen, N. Huang, and C. Luo, "Long non-coding rna asmtl-as1 inhibits tumor growth and glycolysis by regulating the mir-93-3p/mir-660/foxo1 axis in papillary thyroid carcinoma," *Life sciences*, vol. 244, p. 117 298, 2020.

[144] A. Krek *et al.*, "Combinatorial microrna target predictions," *Nature genetics*, vol. 37, no. 5, pp. 495–500, 2005.

[145] K. C. Miranda *et al.*, "A pattern-based method for the identification of microrna binding sites and their corresponding heteroduplexes," *Cell*, vol. 126, no. 6, pp. 1203–1217, 2006.

[146] A. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. Marks, "Microrna targets in drosophila," *Genome biology*, vol. 4, pp. 1–27, 2003.

[147] S. Bandyopadhyay and R. Mitra, "Targetminer: Microrna target prediction with systematic identification of tissue-specific negative examples," *Bioinformatics*, vol. 25, no. 20, pp. 2625–2631, 2009.

[148] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, "The role of site accessibility in microrna target recognition," *Nature genetics*, vol. 39, no. 10, pp. 1278–1284, 2007.

[149] F. J. Martin *et al.*, "Ensembl 2023," *Nucleic acids research*, vol. 51, no. D1, pp. D933–D941, 2023.

[150] M. Isosaka *et al.*, "A screen for epigenetically silenced microrna genes in gastrointestinal stromal tumors," *PloS one*, vol. 10, no. 7, e0133754, 2015.

[151] A. Marusyk and K. Polyak, "Tumor heterogeneity: Causes and consequences," *Biochim Biophys Acta*, vol. 1805, no. 1, pp. 105–17, 2010.

[152] P. L. Bedard, A. R. Hansen, M. J. Ratain, and L. L. Siu, "Tumour heterogeneity in the clinic," *Nature*, vol. 501, no. 7467, pp. 355–64, 2013.

[153] L. Zhao, V. H. F. Lee, M. K. Ng, H. Yan, and M. F. Bijlsma, "Molecular subtyping of cancer: Current status and moving toward clinical applications," *Brief Bioinform*, vol. 20, no. 2, pp. 572–584, 2019.

[154] X. Dai, H. Cheng, Z. Bai, and J. Li, "Breast cancer cell line classification and its relevance with breast tumor subtyping," *J Cancer*, vol. 8, no. 16, pp. 3131–3141, 2017.

[155] A. G. Robertson *et al.*, "Comprehensive molecular characterization of muscle-invasive bladder cancer," *Cell*, vol. 171, no. 3, 540–556. e25, 2017.

[156] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning*, vol. 52, no. 1, pp. 91–118, 2003.

[157] K. Devarajan, "Nonnegative matrix factorization: An analytical and interpretive tool in computational biology," *PLoS Comput Biol*, vol. 4, no. 7, e1000029, 2008.

[158] C. J. Ricketts *et al.*, "The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma," *Cell reports*, vol. 23, no. 1, 313–326. e5, 2018.

[159] N. Cancer Genome Atlas Research, "Comprehensive molecular characterization of gastric adenocarcinoma," *Nature*, vol. 513, no. 7517, pp. 202–9, 2014.

[160] A. G. Robertson *et al.*, "Comprehensive molecular characterization of muscle-invasive bladder cancer," *Cell*, vol. 174, no. 4, p. 1033, 2018.

[161] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[162] T.-H. Chen *et al.*, "The prognostic significance of apc gene mutation and mir-21 expression in advanced-stage colorectal cancer," *Colorectal disease*, vol. 15, no. 11, pp. 1367–1374, 2013.

[163] F. Bertucci, P. Finetti, and D. Birnbaum, "Basal breast cancer: A complex and deadly molecular subtype," *Curr Mol Med*, vol. 12, no. 1, pp. 96–110, 2012.

[164] T. Sørlie, "Molecular portraits of breast cancer: Tumour subtypes as distinct disease entities," *European journal of cancer*, vol. 40, no. 18, pp. 2667–2675, 2004.

[165] N. Cancer Genome Atlas, "Genomic classification of cutaneous melanoma," *Cell*, vol. 161, no. 7, pp. 1681–96, 2015.

[166] L. De Cecco *et al.*, "Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data," *Oncotarget*, vol. 6, no. 11, pp. 9627–42, 2015.

[167] A. C. Berger *et al.*, "A comprehensive pan-cancer molecular study of gynecologic and breast cancers," *Cancer cell*, vol. 33, no. 4, 690–705. e9, 2018.

[168] M. Ceccarelli *et al.*, "Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma," *Cell*, vol. 164, no. 3, pp. 550–563, 2016.

[169] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "Clusterprofiler: An r package for comparing biological themes among gene clusters," *Omics: a journal of integrative biology*, vol. 16, no. 5, pp. 284–287, 2012.

[170] E. Chirshev, K. C. Oberg, Y. J. Ioffe, and J. J. Unternaehrer, "Let-7 as biomarker, prognostic indicator, and therapy for precision medicine in cancer," *Clinical and translational medicine*, vol. 8, no. 1, p. 24, 2019.

[171] S. Ali, K. Almhanna, W. Chen, P. A. Philip, and F. H. Sarkar, "Differentially expressed mirnas in the plasma may provide a molecular signature for aggressive pancreatic cancer," *American journal of translational research*, vol. 3, no. 1, p. 28, 2011.

[172] R. Ghanbari *et al.*, "Simultaneous underexpression of let-7a-5p and let-7f-5p micrornas in plasma and stool samples from early stage colorectal carcinoma: Sup-

plementary issue: Biomarkers for colon cancer," *Biomarkers in cancer*, vol. 7, BIC. S25252, 2015.

[173] N. H. Heegaard, A. J. Schetter, J. A. Welsh, M. Yoneda, E. D. Bowman, and C. C. Harris, "Circulating micro-rna expression profiles in early stage nonsmall cell lung cancer," *International journal of cancer*, vol. 130, no. 6, pp. 1378–1386, 2012.

[174] M. V. Iorio *et al.*, "Microrna gene expression deregulation in human breast cancer," *Cancer research*, vol. 65, no. 16, pp. 7065–7070, 2005.

[175] W. Gao, H. Shen, L. Liu, J. Xu, J. Xu, and Y. Shu, "Mir-21 overexpression in human primary squamous cell lung carcinoma is associated with poor patient prognosis," *Journal of cancer research and clinical oncology*, vol. 137, pp. 557–566, 2011.

[176] Q. Jiang *et al.*, "Mir2disease: A manually curated database for microrna deregulation in human disease," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D98–D104, 2009.

[177] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, "Mirecords: An integrated resource for microrna–target interactions," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D105–D110, 2009.

[178] T. Vergoulis *et al.*, "Tarbase 6.0: Capturing the exponential growth of mirna targets with experimental support," *Nucleic acids research*, vol. 40, no. D1, pp. D222–D229, 2012.

[179] H.-Y. Huang *et al.*, "Mirtarbase 2020: Updates to the experimentally validated microrna–target interaction database," *Nucleic acids research*, vol. 48, no. D1, pp. D148–D154, 2020.

[180] D. Szklarczyk *et al.*, "The string database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic acids research*, vol. 49, no. D1, pp. D605–D612, 2021.

[181] P. Shannon *et al.*, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res*, vol. 13, no. 11, pp. 2498–504, 2003.

[182] M. Kohl, S. Wiese, and B. Warscheid, "Cytoscape: Software for visualization and analysis of biological networks," *Methods Mol Biol*, vol. 696, pp. 291–303, 2011.

[183] J. Zhang *et al.*, "Microrna-200c-3p/zeb2 loop plays a crucial role in the tumor progression of prostate carcinoma," *Annals of Translational Medicine*, vol. 7, no. 7, 2019.

[184] "Retraction: Microrna-98 inhibits the proliferation, invasion, migration and promotes apoptosis of breast cancer cells by binding to hmga2," *Biosci Rep*, vol. 41, no. 6, 2021.

[185] D. Sun, X. Luo, L. Ma, Y. Wang, and F. Zhang, "Identifying of mir-98-5p/igf1 axis contributes breast cancer progression using comprehensive bioinformatic analyses methods and experiments validation," *Life Sci*, vol. 261, p. 118 435, 2020.

[186] K. Nakamura, K. Kusama, A. Ideta, K. Kimura, M. Hori, and K. Imakawa, "Effects of mir-98 in intrauterine extracellular vesicles on maternal immune regulation during the peri-implantation period in cattle," *Sci Rep*, vol. 9, no. 1, p. 20 330, 2019.

[187] M. Hernandez-Quiles, M. F. Broekema, and E. Kalkhoven, "Ppargamma in metabolism, immunity, and cancer: Unified and diverse mechanisms of action," *Frontiers in endocrinology*, vol. 12, p. 624 112, 2021.

[188] D. Zhang *et al.*, "Sema3c supports pancreatic cancer progression by regulating the autophagy process and tumor immune microenvironment," *Frontiers in oncology*, vol. 12, p. 890 154, 2022.

[189] B. Herzog, C. Pellet-Many, G. Britton, B. Hartzoulakis, and I. C. Zachary, "Vegf binding to nrp1 is essential for vegf stimulation of endothelial cell migration, complex formation between nrp1 and vegfr2, and signaling via fak tyr407 phosphorylation," *Molecular biology of the cell*, vol. 22, no. 15, pp. 2766–2776, 2011.

[190] I. Schneider *et al.*, "Use of "default" parameter settings when analyzing single cell rna sequencing data using seurat: A biologist's perspective," *Journal of Translational Genetics and Genomics*, vol. 5, no. 1, pp. 37–49, 2021.