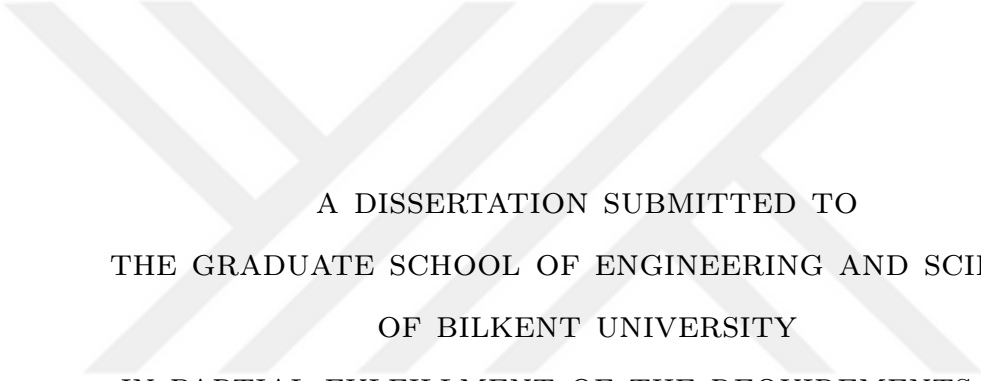


DISCOVERING REGULATORY NON-CODING RNA INTERACTIONS



A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

By
Gülden Olgun
September 2019

DISCOVERING REGULATORY NON-CODING RNA INTERAC-
TIONS

By Güliden Olgun

September 2019

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Abdullah Ercüment Çiçek(Advisor)

Öznur Taştan Okan(Co-Advisor)

Can Alkan

Özlen Konu Karakayalı

Aybar Can Acar

Tunca Doğan

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

DISCOVERING REGULATORY NON-CODING RNA INTERACTIONS

Gülden Olgun

Ph.D. in Computer Engineering

Advisor: Abdullah Ercüment Çiçek

Co-Advisor: Öznur Taştan Okan

September 2019

The vast majority of eukaryotic transcriptomes comprise noncoding RNAs (ncRNAs) which are not translated into proteins. Despite the accumulating evidence on the functional roles of ncRNAs, we are still far from understanding the whole spectrum of molecular functions ncRNAs can undertake and how they accomplish them. In this thesis we develop computational methods for discovering interactions among ncRNAs and tools to analyze them functionally. In the first part of the thesis, we present an integrative approach to discover long non-coding RNA (lncRNA) mediated sponge interactions where lncRNAs can indirectly regulate mRNAs expression levels by sequestering microRNAs (miRNAs), and act as sponges. We conduct partial correlation analysis and kernel independence tests on patient gene expression profiles and further refine the candidate interactions with miRNA target information. We use this approach to find sponge interactions specific to breast-cancer subtypes. We find that although there are sponges common to multiple subtypes, there are also distinct subtype-specific interactions with high prognostic potential. Secondly, we develop a method to identify synergistically acting miRNA pairs. These pairs have weak or no repression on the target mRNA when they act individually, but when together they induce strong repression of their target gene expression. We test the combinations of RNA triplets using non-parametric kernel-based interaction tests. In forming the triplets to test, we consider target predictions between the miRNAs and mRNA. We apply our approach on kidney tumor samples. The discovered triplets have several lines of biological evidence on a functional association among them or their relevance to kidney tumors. In the third part of the thesis, we focus on functional enrichment analysis of noncoding RNAs while some non-coding RNAs (ncRNAs) have been found to play critical regulatory roles in biological processes, most remain functionally uncharacterized. This presents a challenge whenever an interesting set of ncRNAs set needs to be analyzed in a functional context. We

develop a method that performs cis enrichment analysis for a given set of ncRNAs. Enrichment is carried out by using the functional annotations of the coding genes located proximally to the input ncRNAs. To demonstrate how this method could be used to gain insight into the functional importance of a list of interesting ncRNAs, we tackle different biological questions on datasets of cancer and psychiatric disorders. Particularly, we also analyze 28 different types of cancers in terms of molecular process perturbed and linked to altered lncRNA expression. We hope that the methods developed herein will help elucidate functional roles of ncRNAs and aid the development of therapies based on ncRNAs.

Keywords: noncoding RNA, miRNA, lncRNA, sponge interactions, synergistic miRNA, lncRNA functional enrichment.

ÖZET

DÜZENLEYİCİ KODLANMAYAN RNA ETKİLEŞİMLERİNİN KEŞFİ

Gülden Olgun

Bilgisayar Mühendisliği, Doktora

Tez Danışmanı: Abdullah Ercüment Çiçek

İkinci Tez Danışmanı: Öznur Taştan Okan

Eylül 2019

Ökaryotik transkriptomların büyük çoğunluğu, proteinlere çevrilmeyen kodlayıcı olmayan RNA'ları (ncRNA'lar) içerir. ncRNA'ların fonksiyonel rolleri hakkında biriken bilgilere rağmen, ncRNA'ların üstlenebileceği tüm moleküler fonksiyon spektrumunu ve bunları nasıl başarabileceklerini anlamaktan hala uzağız. Bu tez çalışmasında ncRNA'lar arasındaki etkileşimleri keşfetmek için hesaplamalı yöntemler ve bu etkileşimleri fonksiyonel olarak analiz etmek için araçlar geliştiriyoruz. Tezin ilk bölümünde, kodlayıcı olmayan uzun RNA (lncRNA) aracılı sünger etkileşimlerini keşfetmek için bütüncül bir yaklaşım sunuyoruz. lncRNA'lar, mikroRNA'lara (miRNA'lara) el koyarak mRNA'ların ekspresyon seviyelerini dolaylı olarak düzenleyebilirler ve miRNA süngeri gibi davranabilirler. Hasta gen ekspresyon profillerinde kısmi korelasyon analizi ve çekirdek bağımsızlık testleri yapıyoruz ve miRNA hedef bilgisi ile aday etkileşimlerini daha da geliştiriyoruz. Bu yaklaşımı meme kanseri alt tiplerine özgü sünger etkileşimlerini bulmak için kullanıyoruz. Birden fazla alt tipte ortak süngerleri bulmamıza rağmen, yüksek prognostik potansiyeli olan farklı alt tip-spesifik etkileşimler olduğunu da görüyoruz. İkincisi, sinerjik olarak etkili miRNA çiftlerini tanımlamak için bir yöntem geliştiriyoruz. Bu çiftler ayrı ayrı hareket ettiklerinde hedef mRNA üzerinde zayıf bir baskıya sahiptirler veya hiç baskı yapmazlar, ancak birlikte olduklarında hedef gen ifadelerinin kuvvetli bir şekilde baskılanmasına neden olurlar. Parametrik olmayan çekirdek bazlı etkileşim testleri kullanarak RNA üçüzlerinin kombinasyonlarını test ediyoruz. Test edilecek üçüzleri oluştururken, miRNA'lar ve mRNA arasındaki hedef tahminleri göz önünde bulunduruyoruz. Yaklaşımımızı böbrek tümör örneklerine uyguluyoruz. Keşfedilen üçüzlerin, aralarındaki fonksiyonel bir ilişki veya böbrek tümörleriyle olan ilgileri hakkında çeşitli biyolojik kanıtlar vardır. Tezin üçüncü bölümünde, biyolojik işlemlerde kritik düzenleyici roller oynadıkları halde hala pek çoğunun fonksiyonel

özellikleri bilinmeyen kodlayıcı olmayan RNA'ların fonksiyonel zenginleştirme analizine odaklandık. Bu problem, ilginç bir ncRNA setinin fonksiyonel bir bağlamda analiz edilmesi gerektiğinde bir meydan okuma sunar. Biz de bu amaçla, belirli bir ncRNA kümesi için gen yakınlığına bağlı zenginleştirme analizi yapan bir yöntem geliştiriyoruz. Zenginleştirme, girilen ncRNA'lara yakın olan mRNA genlerinin fonksiyonel açıklamaları kullanılarak gerçekleştirilir. Bu yöntemin ilginç ncRNA'ların bir listesinin işlevsel önemine dair bir içgörü kazanmada nasıl kullanılabileceğini göstermek için, kanser ve psikiyatrik bozuklukların veri setleri hakkında farklı biyolojik sorular ele alıyoruz. Ayrıca, 28 farklı tipteki kanseri, bozulan ve değiştirilmiş lncRNA ekspresyonuna bağlı moleküler işlemler açısından analiz ediyoruz. Burada geliştirilen yöntemlerin, ncRNA'ların fonksiyonel rollerini aydınlatmaya yardımcı olacağını ve ncRNA'lara dayanan tedavilerin geliştirilmesine yardımcı olacağını umuyoruz.

Anahtar sözcükler: kodlanmayan RNA, miRNA, lncRNA, sünger ilişkiler, sinerjistik miRNA, lncRNAların fonksiyonel zenginleştirilmesi.

Acknowledgement

I would like to thank and hereby express my gratitude to my supervisor Assist. Prof. Öznur Tastan for her continued guidance and support during my PhD studies. She took time to motivate me against countless hurdles that came my way. Her suggestions have always been precious for me. Also, I would like to thank Asist. Prof. Ercüment Çiçek for his generous help.

I would like to thank all committee members for their valuable time and critical feedback throughout this journey.

Additionally, I am deeply grateful for having my dear family on my side, especially my mom, with their support and patience. As a bonus, hugging my little boy, Gölge, helped me to stay motivated. I am delighted to acknowledge the valuable suggestions and the unwavering support of Assoc. Prof. Dr. İsmail Sengör Altingövde.

Of course, I would also like to thank my dearest friends Özlem Gür, Özlem Seven and Seda Dumlu for their support and patience. Last but not least, I would like to thank Gizem Orhan, Püren Güler, Nurulhude Baykal, Ömer Akpınar and all the Tastan Lab members.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Contributions	4
2	Identifying lncRNA Mediated Sponge Interactions in Breast Cancer Molecular Subtypes	6
2.1	Introduction	6
2.2	Related Work	8
2.2.1	Cefinder Method	8
2.2.2	Correlation Method	8
2.2.3	Sensitivity Correlation Method	9
2.2.4	Hermes Method	10
2.2.5	MuTaME Method	10
2.2.6	Other Methods	10
2.3	Methodology	11
2.3.1	Data Collection	11
2.3.2	Data Preprocessing	12
2.3.3	Statistical Analysis for Finding lncRNA Mediated ceRNA Interactions	12
2.3.4	Filtering ceRNAs Based on miRNA-Target Interactions	18
2.3.5	Identifying ceRNAs with Prognostic Value	19
2.3.6	Pathway and GO Enrichment Analysis	20
2.3.7	Clustering mRNAs	21
2.4	Results	22
2.4.1	Overview of Discovered ceRNA Interactions	22

2.4.2	Spatially Proximal ceRNAs Interactions	27
2.4.3	Functional Enrichment Analysis of mRNAs in ceRNAs . . .	29
2.4.4	Prognostic Sponge Interactions	31
2.5	Discussion	36
3	Identifying Cooperating Pairwise miRNAs	42
3.1	Introduction	42
3.2	Related Work	44
3.2.1	Cotargeting miRNAs	45
3.2.2	Coregulator miRNAs	45
3.2.3	Synergistic miRNA	46
3.3	Methods	46
3.3.1	miRCoop Step 1: Identifying Candidate miRNA–mRNA Target Interactions	47
3.3.2	miRCoop Step 2: Expression Filter	48
3.3.3	miRCoop Step 3: Testing Statistical Interactions of RNAs with Kernel-Based Interaction Tests	48
3.3.4	The Null Model for miRCoop Triplets	51
3.3.5	Data Collection and Processing	52
3.4	Results	53
3.4.1	Overview of the synergistic miRNAs and their targets in Kidney Cancer	53
3.4.2	Analysis of Identified Triplets	54
3.4.3	Comparison with Other Methods	55
3.4.4	Triplets Coded Close on the Genome	56
3.4.5	Common TF Regulating the RNAs of the Triplets	56
3.4.6	Functional Enrichment Analysis	58
3.5	Discussion	60
4	Functional Enrichment Analysis of the lncRNA Derived from Its Neighbors	62
4.1	Introduction	62
4.2	Related work	63
4.3	Methods	64

4.3.1	Data Sources	64
4.3.2	Differential Expression Analysis	66
4.3.3	Functional Enrichment Analysis of lncRNAs Based on Neighborhood Genes	67
4.3.4	GO Term Semantic Similarity Evaluation	68
4.3.5	Clustering Cancers	69
4.4	Results	69
4.4.1	Overview of Gene Ontology Enrichment Analysis of the lncRNAs	69
4.4.2	Identifying Functional Relevance between lncRNAs and mRNAs	71
4.4.3	Functional Similarity Analysis Between Cancers	71
4.4.4	Exploring the Metastasis in Cancers Based on a Functional Evidence	76
4.4.5	Inferring Functional mechanisms for the Super-enhancer Associated lncRNAs	78
4.4.6	Overview of the Pathway Analysis	79
4.4.7	Validation With Cancer Driver Genes	80
4.4.8	Comparison Between Our Proposed Method and the Other Methods	83
4.5	Conclusion	86
4.6	NoRCE R Package	87
4.7	Implementation and Features of the NoRCE	88
4.7.1	Enrichment Analysis of the Coding Genes	90
4.7.2	Visualization of the Results	91
4.8	Example Use Cases of NoRCE	92
4.8.1	Case Study 1: Functional enrichment analysis of ncRNAs differentially expressed in psychiatric disorders	92
4.8.2	Case Study 2: Functional enrichment analysis of variably expressed miRNAs in brain cancer	99
4.8.3	Data in NoRCE repository for the vignette and supplemen- tary results	105
4.8.4	Other features	106



List of Figures

2.1	Venn diagrams for lncRNAs that are found coding with CPAT [1] and/or CPC [2].	13
2.2	The distribution of the S_z values for all tested RNA triplets. 99 th percentile is illustrated with a red dashed line.	14
2.3	A) Overview of the methodology, each box represents a step in the methodology. Steps B-F are conducted for each breast cancer subtype separately. B) The number of ceRNAs remained after each main filtering step when $t = 0.2$ (Step C in Figure 1A). C) Venn diagram of ceRNA interactions discovered in each of the breast cancer molecular subtype.	15
2.4	A) Number of ceRNAs remained after each main filtering step when $t = 0.3$ threshold is used. B) Venn diagram of ceRNA interactions discovered in each of the breast cancer molecular subtype.	21
2.5	A-B) Heatmaps display the distribution of ceRNAs over the subtypes for A) $t = 0.3$ and B) $t = 0.2$. Blue and green cells indicate the ceRNA interaction is discovered in the given subtype, while white color indicates it is not. Number of ceRNA interactions discovered that per C) each lncRNA and per B) each miRNA in each breast cancer subtypes ($t = 0.3$).	24
2.6	Number of ceRNA interactions discovered that A) lncRNAs and B) miRNAs take part in each breast cancer subtype ($t = 0.2$). . .	27

2.7	lncRNA-mRNA network for each breast cancer subtypes. Green triangle nodes represent lncRNA and circle orange nodes represents mRNA. An edge between an mRNA and a lncRNA is drawn to represent a ceRNA interaction through a miRNA. Node size is in proportion to degree of the node. The network plot was generated with Cytoscape(v3.4.0)[3].	28
2.8	Number of mRNAs that each lncRNA-miRNA pair interacts with in each subtype ($t = 0.2$).	29
2.9	A) The network of sponge interactions between HOTAIR, hsa-miR-196a miRNAs and HOXC genes. The circles denote lncRNAs, and triangles denote mRNAs. An edge exists between a lncRNA and an mRNA if there is a sponge interaction between them; the edge label indicates the miRNA that regulates the interaction. B) The genomic locations of the sponge interactions on chromosome 12. Spatially proximal genes are underlined with the same color code of used in (A).	32
2.10	The dot plot for the most significant 27 enriched pathway which were filtered out by p -value cutoff 0.05 and FDR cutoff 1×10^{-4} . Dots in the plot are color coded depending upon the relevant FDR value. Color gradient changes from red(low FDR value, high enrichment) to blue(high FDR value, low enrichment). Dot size depends on the gene ratio, ratio of enriched genes to identified genes in the pathway. Number of identified genes in each subtypes were provided in parenthesis.	33
2.11	The dot plot for the most significant 27 enriched KEGG pathway which were filtered out by p -value cutoff 0.05 and FDR cutoff 1×10^{-4} . Dots in the plot are color coded depending upon the relevant FDR value. Color gradient changes from red(low FDR value, high enrichment) to blue(high FDR value, low enrichment). Dot size depends on the gene ratio, ratio of enriched genes to identified genes in the pathway. Number of identified genes in each subtypes were provided in parenthesis.	34

2.12	Venn diagram displaying the distribution of enriched pathways (p -value ≤ 0.05 and FDR $\leq 1 \times 10^{-4}$.) over the subtypes.	35
2.13	Kaplan-Meir survival plots when patients are divided based on individual expression patterns of the RNAs (the first three plots in each panel) and when patients are divided based on the sponge expression pattern (4th plot) for MEG3, hsa-miR-1245, COL12A1 sponge.	37
2.14	lncRNA:miRNA:mRNA network for all breast cancer subtypes. lncRNAs are represented by the green triangle symbol, mRNAs are represented by orange ellipse symbol and miRNAs are with the yellow rectangle. Each node size is scaled by its degree, the number of edges incident to the nodes and edge width is scaled by the number of occurrence of the node pair. The network was constructed using the Cytoscape (v3.4.0) [3].	38
2.15	The distribution of the f_{xyz} scores of the prognostic ceRNA interactions.	39
2.16	Distribution of the prognostic RNAs in breast cancer subtypes Colored cells indicate the RNA is discovered in the given subtype, while white color indicates it is not.	41
3.1	Overview of the methodology. Each box represents a step in the methodology.	47
3.2	The number of remaining candidate triplets at the end of each step of the pipeline. The number of available RNAs that participate in the analysis is provided at the beginning.	53
3.3	The network shows the miRNA synergistic pairs discovered for kidney cancer. Nodes are miRNAs and edges exist between them if they are participate in at least one miRCoop triplet.	54

3.4	Venn diagram for the number of synergistic miRNAs that are reported by different methods for human. In CancerNet [4] we use the kidney specific triplets. The TriplexRNA [5] database predicts the RNA triplexes constitute of two synergistic miRNAs and their mutual target mRNAs. In TriplexRNA, we utilized predictions that are called as TEC. In TEC, predicted RNA triplexes are closely within 13-35 nt and their predicted equilibrium concentration (TEC) > 50 nM.	55
3.5	The red marks on the genome axis indicate the miRNA positions of the synergistic miRNAs that belong to the miR-500 family. The figure was created with Gviz package in R [6].	56
3.6	Network for the synergistic triplets regulated by common TFs. Each node denotes a triplet or a TF while an edge represents an interaction between them. TFs are represented by the diamond-shaped yellow symbol, and triplets are represented by the circle-shaped blue symbol.	58
3.7	a) The the largest connected component of miRCoop triplet network. Blue nodes are miRNAs and pink colors are mRNAs of the triples. The gray nodes are dummy nodes to connect the triplet miRNAs to the target mRNA. b) Bar chart for the top 25 enriched biological process GO-terms found over the mRNA set of the largest connected component in part a)	60

4.1 A. Illustration of the gene intervals and gene sets that are covered in this analysis. Distal regions are scaled down in order to fit to the manuscript. B. The logic of the annotations for the set of lncRNAs based on neighborhood genes. A lncRNA that falls into any of the genomic intervals of mRNA is annotated with the corresponding mRNA. For example, when a set of lncRNAs consisting of lncRNA A to F is analyzed, lncRNA A is annotated with mRNA X in the exon and regulation based gene set, lncRNA F is annotated in exon based gene set with mRNA Y and in regulation gene set with mRNA Y and Z. lncRNA B is only annotated with mRNA X in regulation based gene set. LncRNA D, E and F is annotated with Y in regulation based whereas lncRNA C falls into the 3D/5D genomic intervals, as a result, it is not annotated with any mRNA in regulation and exon based gene set. Thus, to enrich the exon based gene set for the given lncRNA set, mRNA X and Y is utilized. Similarly, regulation based gene set contains mRNA X, Y and Z. 65

4.2 Overview of the methodology. Dysregulated lncRNAs are determined by a differential expression analysis (Step 1). Within the 10 kb range, both from upstream and downstream of the lncRNA gene, lncRNAs annotated with their neighborhood coding genes according to the genomic intervals of mRNA. The color codes of genes represent the functionality. The red color-coded genes cannot be annotated with any lncRNA since they are at least 10 kb far away from lncRNAs. On the neighborhood coding gene set, we perform enrichment analysis (Step 3). The same methodology is applied for 28 different cancers. 66

4.3 Dot plot for the number of tumor patients and the number of differentially expressed lncRNAs in each cancer. Color code of the dots represents the number of differentially expressed lncRNA genes and the size of dots is the number of tumor patients for the cancer. 70

4.4 Functional similarity scores between each cancer pairs are represented in the heatmap. Biological processes (A-B) are demonstrated in the first row and molecular functions (C-D) are in the second row. Colors are scaled between blue (low) to red (high) and white color represents the median value. Color key for regulation based heatmap is [0.2-1] and for exon based is [0.4-1]. Important clusters are highlighted with black square. Row and column labels describe the primary site of the cancers. 73

4.5 Representation of metastasis from the primary site to the metastatic site. Each node corresponds to a site. The size of a node depends on the out-degree, and the number of outward directed edges from the node. Weights of edges are occurrences of the related metastasis for the nodes, and the directions of the edges are from the primary sites to metastatic sites. 77

4.6 The NoRCE workflow 89

4.7 GO biological process enrichment results pertaining to differentially expressed brain ncRNAs. A) Top 35 GO terms are shown in the dot plot. The x-axis represents the p -values; the y-axis represents GO terms. Dot area is proportional to the size of the overlapping gene set, and the color signifies the p -value of the enrichment test for the corresponding GO term. B) The top 7 enriched GO terms and the coding genes which are associated with them are depicted as a network. The size of the node represents the number of the coding genes which are associated with the GO term, and each color represents a different GO term. 94

4.8 The enrichment results obtained by performing TAD based analysis on ncRNAs obtained from [7]. (A) The dot plot where the Y-axis denotes p -values and the x-axis denotes GO terms. The size and the color of dots signify the size of the enriched gene set and the p -value for the corresponding GO term, respectively. B) The network of GO term:protein coding gene. Size of the nodes is proportional to the degree of the node. Different colors are used to differentiate the cluster of proteins. 96

4.9 Directed acyclic graph of the enriched top 5 GO terms based on a gene set that are co-expressed and targeted by the brain cancer miRNA. GO terms are colored based on their p -values 103

4.10 Reactome pathway map for the enriched pathway, R-HSA-199992(trans-Golgi Network Vesicle Budding). Genes that are enriched for this pathway are marked with purple. 105



List of Tables

1.1	Information for some noncoding RNAs	3
2.1	Computationally and experimentally validated miRNA-target databases used for mRNA and lncRNA. Plus signs denote databases that are used for the miRNA interactions of the RNA type. ‘P’ denotes predicted target information while ‘E’ denotes experimentally supported target information.	18
2.2	Features of the prediction miRNA-target databases used for mRNA and lncRNA. Plus signs denote features that are used for the databases.	18
2.3	Pathway data sources utilized for enrichment analysis and number of pathways in each data source.	21
2.4	Number of ceRNA interactions identified in each breast cancer subtypes at $t = 0.3$ and $t = 0.2$. Total number of all ceRNA interactions and number of subtype specific ceRNAs are provided.	23
2.5	List of lncRNAs & miRNAs that are found to participate in sponges of all four subtypes.	23
2.6	List of lncRNAs & miRNAs that are found in sponges of single subtype.	25
2.7	Number nodes and edges for bipartite lncRNA-mRNA networks for each breast cancer subtypes where each node denotes lncRNA or mRNA and each edge represents a lncRNA-mRNA interaction,miRNA.	26

2.8	List of subtype specific enriched pathways. Bonferroni corrected p values are provided and all listed pathways are above FDR cutoff 1×10^{-4}	31
3.1	The list of identified synergistic miRNA pairs in kidney cancer that are encoded close in the genome and their family information. . .	56
4.1	Enriched GO-term comparison for the differentially expressed mRNA and lncRNA gene set	72
4.2	List of the GO-terms that are only enriched in both Acute Myeloid Leukemia and Diffuse Large B-cell Lymphoma. Last column shows the maximum <i>FDR</i> value that AML and DLBL take for the corresponding GO-terms.	75
4.3	Comparison between driver genes and our results for the corresponding cancers. In this table, each field represents the GO-term similarity score between driver genes and our results. Null test columns show the ratio of the alternative hypothesis for the GO-term similarity scores of the random selected driver genes greater than the actual GO-term similarity score for the actual driver genes for the corresponding cancer.	81
4.4	Comparison between driver genes and our results for the corresponding cancer. In this table, each field represents the GO-term similarity score between driver genes and our results. Null test columns show the ratio of the alternative hypothesis for the GO-term similarity scores of the random selected driver genes greater than the actual GO-term similarity score for the actual driver genes for the corresponding cancer.	82
4.5	Regulation based BP enrichment results for gene set HOTAIR, MEG3, H19	83
4.6	Exon based BP enrichment results for gene set HOTAIR, MEG3, H19	83
4.7	Regulation based MF enrichment results for gene set HOTAIR, MEG3, H19	84

4.8	Exon based MF enrichment results for gene set HOTAIR, MEG3, H19	84
4.9	The supported assemblies for different species in NoRCE.	88
4.10	Topological associating domain data that included in the NoRCE repository	90
4.11	List of miRNA target prediction algorithms and the corresponding number of target pairs for each species. To the best of our knowledge, no miRNA target predictions have been reported for <i>Saccharomyces cerevisiae</i>	91
4.12	The size of gene ontologies available for each species in NoRCE	91
4.13	The number of pathways and the different pathway sources included in the in the March 2019 Bader Lab pathway data set.	93
4.14	Top 10 enrichment results of brain related biological process GO terms that are obtained from the neighbourhood coding genes of ncRNA set of the [7]. The GeneRatio is computed by dividing the number of genes that are overlapping with the coding ones to the number of all protein-coding genes within the input set neighbourhood. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of the background gene set. The EGNo refers to the size of the overlap between the corresponding GO term gene set and the neighboring coding gene set. GeneList columns contain these genes.	97
4.15	The Top 10 biological process GO term enrichment results. These are obtained by performing TAD based analysis using <i>brain_disorder_ncRNA</i> data. The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes within the neighbourhood of the input set. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set. GeneList columns contain these genes.	98

4.16 Disease related pathway enrichment results for the *brain_disorder_ncRNA* dataset using customized pathway databases. The The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes within the neighbourhood of the input set. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set. GeneList columns contain these genes. 100

4.17 Some brain related biological process functions that are enriched based on target and custom co-expression analysis in brain cancer specific miRNAs. The The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes that are within the neighbourhood of the input set. The BGRatio column represents the ratio of the number of genes within the enriched GO term set to the size of the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set while GeneList columns list these genes. 102

4.18 Pathway enrichment of miRNA gene set based on neighbor targets. The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes within the input set neighborhood. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of all of the genes in the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set. GeneList columns list these genes. 104

Chapter 1

Introduction

1.1 Motivation

Completion on the transcription profiling studies reveal that only 1.5% of the human genome consists of the coding sequence [8]. Remained regions are mostly considered as junk DNA because they contain approximately 59% repeat sequences [9]. However; those regions among coding genes transcribed into various different length non-coding RNAs (ncRNAs). ncRNAs participate in biologically important functions including control of chromosome dynamics, splicing, RNA editing, translational inhibition and messenger RNA (mRNA) destruction and deregulations are implicated with diseases including cancer [10].

Based on their functional relevance, ncRNAs can be divided into two categories, housekeeping and regulatory. Housekeeping ncRNA category encompasses transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), RNase P RNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), and telomerase RNA. They are constantly expressed and needed for regular functions and viability of the cell. Additionally, they take part in transportation and insertion of the proteins into membranes and telomeric sequence addition [11].

The regulatory ncRNAs are expressed in development of specific tissues or disease. They may alter the expression of other genes at transcription or translation levels [12]. Regulatory ncRNAs include long non-coding RNAs (lncRNA) and small non-coding RNAs (less than 200 nt)[13] such as microRNAs (miRNAs), endogenous small interfering RNAs (endo-siRNAs), repeat-associated small interfering RNA (rasiRNA) and PIWI-interacting RNAs (piRNAs). Further details of each RNA is provided in Table 1.1.

miRNAs are small RNA, 20 - 24 nucleotides long that regulate gene expression at post-transcriptional level, intercede gene silencing by guiding Argonaute (AGO) proteins to bind complementary sequences of the specific target sites in the 3' untranslated region (UTR) of the mRNA [19, 13]. miRNA plays key part in animal development, cell differentiation, homeostasis and participate every cellular process [19]. Aberrant miRNA functioning has been associated with diseases including cancer[20]. miRBase release 22 [21] contains 1,917 human precursors and 2,656 mature miRNAs and it is reported that miRNAs can bind more than 60% of the known protein coding genes [22]. Recently, it has been experimentally provided that multiple miRNAs can cooperatively work to exert function on the target mRNA by repression the expression of mRNA. Yet, complete set of synergistic miRNAs are unknown.

Long non-coding RNAs are < 200 nt have tissue specific expression patterns. They have significant roles in regulating gene expression at transcriptional and post-transcriptional levels[23]. They have an essential role in tumorigenesis and lncRNA-mediated networks settle a significant place for the cancer progression. Moreover, lncRNAs play important roles in biological processes, such as cell proliferation, differentiation, chromatin remodeling, epigenetic regulation, genomic splicing, transcription, translation[23]. Due to lncRNAs' characteristics, it is difficult to assess the lncRNA functionality. Even some methods [24, 25, 26, 27] reveal to identify the lncRNA functionality and their functional impact on the genome and in disease state, their functional mechanisms are still missing. However, it is known that lncRNAs have cis and trans regulatory activity. Cis-regulator lncRNAs regulate neighbour genes to exert function whereas trans-regulators inspect distal genes from their transcription site [28]. Moreover, lncRNAs can act as

Table 1.1: Information for some noncoding RNAs

RNA	Length	Function	Ref.
miRNAs	20 – 24 bp	Responsible from regulation of the transcription and translation of protein-coding genes.	[14]
lncRNAs	> 200 bp	Key player in gene expression control during processes. Construct regulator networks with mRNAs and miRNAs. By reducing available number of miRNAs for target mRNA, they act as miRNA sponge.	[15]
endo-siRNAs	~ 20 bp	Derived from endogenous double stranded RNAs including transposable elements, cis-natural antisense transcripts, trans-natural antisense transcripts and hairpin RNA transcripts. AGO2 binds to endo-siRNAs, and responsible from the repression of the transposon transcripts or endogenous mRNAs.	[14]
piRNAs	21 – 35 bp	Create RNA-protein interactions with PIWI proteins. Generate the largest class of the small non-coding RNAs and take parts in germline transposon silencing. 24-29 nt long rasiRNAs are also considered as piRNAs.	[14, 16]
snoRNAs	60 – 300 bp	The role of the snoRNAs is modification of the rRNAs, tRNA and snRNA.	[14]
tRNAs	< 0.1 kb	t-shaped molecule that decode mRNA to produce protein. It is categorized into two groups, tiRNA and tRF. They act as microRNA in RNA interference, inhibit protein synthesis. They are responsible for the formation of many diseases including tumors.	[17]
circRNAs	500 bp	Product of alternative splicing know as backsplicing and covalently linked 3' and 5' ends. As in lncRNA, they act as miRNA sponge.	[18]

”miRNA sponge” and sequester miRNAs to prevent them interacting with their natural target mRNAs to play critical roles in various biological and pathological processes. Those RNAs that compete for the shared miRNAs are called as competing endogenous RNAs (ceRNA). However, our knowledge on the ceRNA regulation still incomplete.

Due to their various regulatory roles and disease causing deregulations, lncRNAs and miRNAs attract researchers to reveal the interactions while considering the mRNA. In this thesis, we shall be focusing on the regulatory interactions of the noncoding RNAs, lncRNAs and/or miRNAs, as well as mRNAs are one of the participants of these interactions. Moreover, we shall be analyzing those regulatory interactions for cancer state.

1.2 Contributions

In this thesis, we address three important biological questions for noncoding RNAs that our knowledge is limited: 1) Identifying miRNA sponges in lncRNA-mediated networks, 2) Exploring lncRNA functionality from a set of lncRNAs based on their regulatory effects on their neighborhood, 3) Discovering cooperative miRNAs that individually have weak influence on mRNA but their combine effect has an high impact by repression the expression of the mRNA. We develop new methodologies for the questions that we state above. Our contribution is approaching the first and third questions as a graphical model problem and presenting new methodologies accordingly. Moreover, for exploring the lncRNA functionality from a set of lncRNA, we present the methodology based on the lncRNAs ability to regulate neighborhood genes that involve transcription or splicing of RNA but the lncRNA may not be a functional. Also, we contribute to the literature by considering set of lncRNAs where set size is larger than 3. Our approaches can fit to any disease that have gene expression data but we want to contribute to the cancer research.

In chapter 2 (based on [29]), we consider competing endogenous RNAs

(ceRNA) hypothesis where lncRNAs allocate miRNAs by reducing available miRNAs to inhibit potential miRNA target interactions. We present methodology based on correlation analysis and kernel-based conditional independence test. Furthermore, proposed interactions take into account the biological evidence on miRNA target potential. We identify subtype specific potential lncRNA:miRNA:mRNA interactions for breast cancer since breast cancer subtypes have distinct biological and histopathological features, therapeutic resistance. We introduce the ceRNAs that can be tested with experimental analyses and helps to expose the unknown biological mechanisms or therapeutic decision of the breast subtypes.

In chapter 3 (based on [30]), we propose new method to explore synergistic miRNAs that have weak effect singly but induce stronger effect on the mRNA when bound together. To this end, we propose new method based on three variable independence test that make use of the kernel trick. We also test against the potential miRNA:mRNA interactions with two variable interaction test. Since no experimental data available, we validate our potential triplets with biological evidences. We conclude that our method will lead to identify more complex synergistic interactions.

In chapter 4 (based on [31]), we devise a methodology for the functional analysis of lncRNAs derived from their neighbors. We annotate each lncRNA with mRNA according to their genomic intervals and each set is enriched with GO and KEGG. Also, we develop an R package, NoRCE, for this algorithm. We enhance NoRCE with other biological evidences such as the topologically associating domain (TAD) regions, co-expression patterns, and miRNA target information. The results show that our method illuminates the functional mechanisms of ncRNA genes.

In chapter 5, finally we conclude our findings and point out some future works.

Chapter 2

Identifying lncRNA Mediated Sponge Interactions in Breast Cancer Molecular Subtypes

2.1 Introduction

Advances in sequencing technologies have revealed that there is a large number of RNAs that do not encode proteins [32]. One class of non-coding RNAs (ncRNAs) comprises microRNAs (miRNAs) that repress gene expression by preferentially binding the complementary sequence of their target mRNAs [33]. miRNAs play crucial roles in regulating gene expression programs in the normal cell, and their aberrant expression contributes to pathogenesis in several diseases, including cancer. To date, a large number of miRNAs have been shown to be associated with cancer progression, drug resistance or metastasis [34, 35, 36, 37, 38].

Another major class of non-coding RNAs is long non-coding RNAs (lncRNAs) that are longer than 200 nucleotides. Although the function of the vast majority of lncRNAs remains to be identified, accumulating evidence suggests that they are highly involved in regulating cellular and pathological processes [39, 40].

Deregulations of several lncRNAs have also been associated with cancer [41, 42].

Recent work has provided evidence for an emerging regulatory role of lncRNAs. According to the ceRNA hypothesis [43], lncRNAs can act as ceRNAs. By sequestering miRNAs, lncRNAs can reduce the number of miRNAs available for the target mRNA [43]; in this way, they indirectly prevent the target gene repression, acting like a sponge [44, 45, 46]. lncRNA-mediated sponge interactions and their protein-coding targets have been investigated in gastric cancer [47], glioblastoma multiforme [48], pancreatic cancer [49], ovarian cancer [50] and breast cancer [51].

Breast cancer subtypes differ significantly in their molecular profiles and response to therapy. Because miRNAs and mRNAs exhibit different molecular activity patterns in breast cancer subtypes [52, 53, 54], it is expected that there will be subtype-specific lncRNA-mediated sponge interactions. Identifying these miRNA sponges can both shed light on the uncharacterized mechanisms of the breast cancer subtypes and potentially help in making better therapeutic decisions. In this work, we use an integrative approach to identify subtype-specific lncRNA:miRNA:mRNA interactions through which lncRNAs compete for binding to shared miRNAs in breast cancer.

To find breast cancer subtype-specific interactions, we systemically analyze lncRNA, miRNA, and mRNA expression profiles of breast cancer patients made available through the Cancer Genome Atlas Project (TCGA) [54]. We first identify statistically related lncRNA:miRNA:mRNA interactions through correlation and partial correlation analysis as in Paci et al. [51] and further refine these candidate interactions using a kernel-based conditional independence test (KCI) [55]. KCI does not assume any parametric form for the random variables that are being tested. Also, for the first time, it is used for finding regulatory interactions. The potential candidate interactions are further filtered in the light of available evidence regarding the miRNA-target interactions. We examine the functional enrichment of mRNAs that participate in sponges, the genomic spatial organization and finally, through the survival analysis of patients, we discover lncRNA-mediated ceRNA interactions with prognostic value.

2.2 Related Work

Identifying sponges by experimental means is a challenge, and the experimental datasets are not available in different contexts such as cancer. We can categorize the methods that identify lncRNA sponges computationally as follows:

2.2.1 Cefinder Method

This method is the simplest method to identify the ceRNAs. The method only considers the mRNAs. It is based on condition independence scoring and ranking. It enumerates the number of miRNA response elements (MRE) of the shared miRNAs using target information. In the literature, this method is utilized with ceRDB [56].

2.2.2 Correlation Method

The baseline method to identify the miRNA sponge interactions is correlation coefficient methods. It is experimentally proven that there is a high positive correlation between competing RNAs and miRNAs [57]. Thus, in this method, each pairs of RNA is evaluated and coefficient between RNA pairs should be significantly positive (*pvalue* < 0.05 and $\rho > 0$), [58, 59]. The formula for the Pearson correlation coefficient is provided in Eq 2.1:

$$\rho_{lncRNA,mRNA} = \frac{cov(lncRNA,mRNA)}{\sigma_{lncRNA}\sigma_{mRNA}} \quad (2.1)$$

, where cov represent the covariance and σ represents the standard deviation.

2.2.3 Sensitivity Correlation Method

Partial correlation method measures the relationship between lncRNA and mRNA while controlling miRNA. This method reveals the role of miRNA in mediating lncRNA-mRNA network and calculates the remaining coefficient value when miRNA is removed from the interaction.

$$\rho_{lncRNA,mRNA|miRNA} = \frac{\rho_{lncRNA,mRNA} - \rho_{lncRNA,miRNA}\rho_{miRNA,mRNA}}{\sqrt{1 - \rho_{lncRNA,miRNA}^2}\sqrt{1 - \rho_{miRNA,mRNA}^2}} \quad (2.2)$$

It assumes that if there is no correlation between lncRNA or mRNA with miRNA ($\rho_{lncRNA,miRNA} = 0$ or $\rho_{mRNA,miRNA} = 0$), the partial correlation must be equal to the their correlation coefficient value. That is, correlation coefficient infers the direct relation between lncRNA-mRNA whereas partial correlation infers the indirect relation.

To deduce the role of the miRNA in the lncRNA-mRNA-miRNA network sensitivity correlation, S , is calculated, Eq 2.3. From this formula, it can be seen that if correlation between mRNA and lncRNA is weaker or none ($S \approx 0$), in lncRNA-mRNA relationship, miRNA has an effect.

$$S_{miRNA} = \rho_{lncRNA,mRNA} - \rho_{lncRNA,mRNA|miRNA} \quad (2.3)$$

To explore the sponges, several analysis use the sensitivity correlation method, [51, 60, 29, 61, 62]. It is experimentally proved that correlation between RNAs that compete for miRNAs is very high [57]. Thus, this method assumes there is a high correlation between lncRNA and mRNA.

2.2.4 Hermes Method

Hermes method [63] captures non-linear interactions whereas correlation analysis identifies linear ones. To identify the sponge interactions, the method makes use of the mutual information and conditional mutual information which calculates the expected value of mutual information of two variable given third variable. The method reveals the interaction between a modulator (M), regulator (R) and a regulated target (T) by extracting the mutual information of the regulator and target expression from the conditional mutual information for the regulator and target given modulator, $\Delta I = I[R; T|M] - I[R; T]$. P -value of the triplet is estimated with the permutation test. Significance of the ΔI is determined by converting the individual p -values to X^2 , Chi-squared, test statistics.

2.2.5 MuTaME Method

The mutually targetted MRE enrichment (MuTaME) [64] method measures the interaction metric between two ceRNAs based on these rules: 1) the frequency of the shared miRNA 2) The fraction of the MREs and their maximum distance between two MREs of each shared miRNAs. 3) Similarity of the MREs 4) The rate between total number of MREs for one miRNAs compared to the total number of miRNAs that bind to these MREs. Logarithm of these four scores are added and normalized scores of the RNA pairs that falls into the top 10% are announced as potential miRNA sponge interactions. This method is one of the major methods to identify potential miRNA sponges [65, 66, 62]

2.2.6 Other Methods

SpongeScan uses a sequence-based algorithm to detect potential sponge interactions [67]. The algorithm is applicable whenever sequence data is available but does not account for the expression of the RNAs, which provides evidence on the interaction of the RNA species within a given context. Tian et al. [47] find

sponge interactions in gastric cancer using micro array expression data. In their approach, they find up- and down-regulated lncRNAs and combine them with miRNA prediction algorithms to construct a lncRNA:miRNA network, but the approach does not consider the correlation structure of RNA expressions.

We use the baseline method, sensitivity correlation method, and we enhance the method with kernel based conditional independence.

2.3 Methodology

2.3.1 Data Collection

Illumina HTSeq level 3 RNA-seq gene level expression and miRNA-seq expression data for human breast cancer were retrieved from the The Cancer Genome Atlas[54] on August 9th 2014. As lncRNAs are not presented in TCGA, we curated the list of lncRNAs from the RNA-Seq using the GENCODE v24 [68]. TCGA breast cancer survival data were downloaded from the UCSC Cancer Genomics Browser on June 31st 2016. Patients that contains mRNA, lncRNA, and miRNA gene expression data simultaneously were considered in this analysis.

Patients were divided to the subtypes according to the PAM (Prediction Analysis of Microarray) 50 classifier. The PAM 50 classifier is commonly used method to divide the patients based on their gene expression into different subtypes of the breast cancer. There are four accepted intrinsic subtypes according to the PAM: Luminal A, Luminal B, HER2-enriched (HER2) and basal-like (Basal). In this analysis, 211 Luminal A, 112 Luminal B, 85 Basal and 54 HER2 patients are remained.

2.3.2 Data Preprocessing

Reads Per Kilobase Million Reads (RPKM) values are utilized for RNA-Seq and miRNA-Seq data. In order to eliminate lowly expressed genes, we consider the RPKM values below 0.05 as missing value and RNAs that are missing in more than 20% of the samples in each subtypes are filtered out. Remaining expressions are added to the constant 0.25 value to deal with the 0 values and transformed to log2. We remove the genes that do not vary across samples. To detect the variance across samples, we used median absolute deviation (MAD), Eq. 2.4 and genes with MAD value below 0.5 are filtered.

$$\text{MAD}(r) = \text{median}(|r_i - \text{median}(\mathbf{r})|) \quad (2.4)$$

2.3.2.1 lncRNA curation

As lncRNAs are not annotated in TCGA, we curated a list of lncRNAs using GENCODE v24 [68]. Based on GENCODE v24 annotation, 598 of the RNAs present in RNA-Seq expression data are designated as lncRNAs. To minimize erroneous annotations, we further examined each lncRNA’s coding potential with alignment-free method Coding-Potential Assessment Tool (CPAT) [1] and alignment-based method Coding Potential Calculator (CPC) [2]. LncRNAs whose all transcripts are predicted to have high coding potential by both tools are eliminated. The number of lncRNAs that are predicted to have high coding potential by each tool is provided in Figure 2.1.

2.3.3 Statistical Analysis for Finding lncRNA Mediated ceRNA Interactions

To identify ceRNA interactions between lncRNA:miRNA:mRNA, we performed correlation analysis and kernel-based conditional independence test on expression

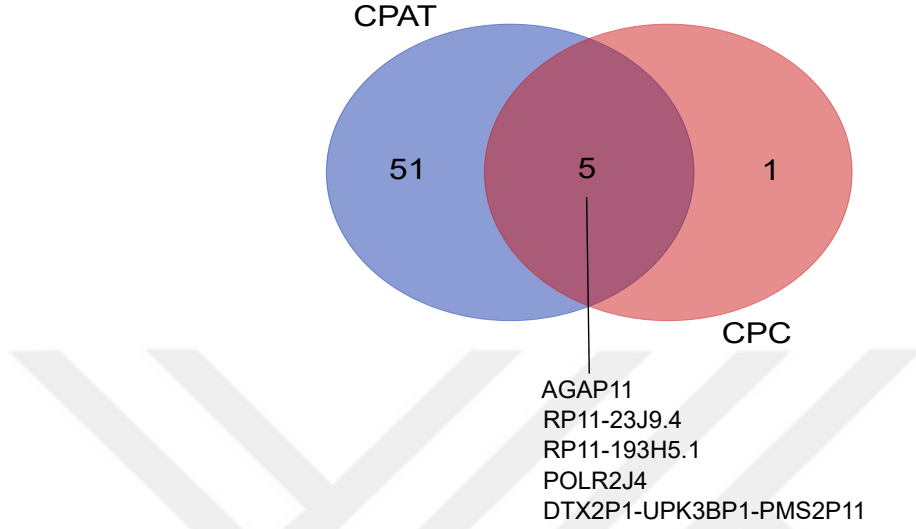


Figure 2.1: Venn diagrams for lncRNAs that are found coding with CPAT [1] and/or CPC [2].

data. Below, random variable X denotes the expression level of a lncRNA, random variable Y indicates the expression level of a mRNA, and finally random variable Z denotes the expression level of a miRNA.

2.3.3.1 Correlation and partial correlation analysis

For a given ceRNA interaction, we expect expression values of the lncRNA and mRNA to be positively correlated, and if this correlation relies on miRNA expression, the correlation between mRNA, and lncRNA should weaken when miRNA expression is taken into account. To quantify this, first Spearman rank order correlation was calculated between lncRNA and mRNAs, which we denote with $\rho_{lncRNA,mRNA}$. Next, we calculated the Spearman partial rank order correlation between lncRNA and mRNA, this time controlling for miRNA expression, $\rho_{lncRNA,mRNA|miRNA}$, as in Eq 2.2.

The difference between the correlation and the partial correlation for a miRNA measures the extent the miRNA Z is effective in the statistical correlation of X and Y . This value is calculated as in Eq 2.3.

As we look for strongly positively correlated lncRNA and mRNA pairs, only those with correlation $\rho_{X,Y} > 0.5$ ($p\text{-value} < 0.05$) were considered. Among those, RNA triplets where S_Z is larger than a threshold value, t , were retained. As it is hard to determine what cutoff is meaningful, we conducted our analysis at two different thresholds, $t = 0.2$ and $t = 0.3$. $t = 0.2$ corresponds to approximately to the 99th percentile of the distribution of the S , Figure 2.2 and $t = 0.3$ is chosen to obtain a even more stringent list.

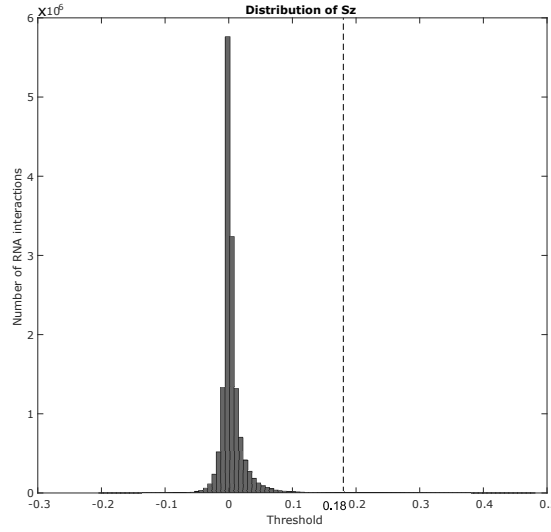


Figure 2.2: The distribution of the S_z values for all tested RNA triplets. 99th percentile is illustrated with a red dashed line.

2.3.3.2 Kernel based conditional independence test

To find lncRNA interactions we also test directly for conditional independence. In a ceRNA interaction, if the interaction of a particular pair of lncRNA (X) and mRNA (Y) were through a shared miRNA (Z), we would expect that lncRNA and mRNA expressions to be conditionally independent given the miRNA expression level. Conditional independence is denoted by $X \perp\!\!\!\perp Y \mid Z$. X and Y are conditionally independent given Z if and only if the $\mathbf{P}(X \mid Y, Z) = \mathbf{P}(X \mid Z)$ (or equivalently $\mathbf{P}(Y \mid X, Z) = \mathbf{P}(Y \mid Z)$ or $\mathbf{P}(X, Y \mid Z) = \mathbf{P}(X \mid Z) \mathbf{P}(Y \mid Z)$). That is if X and Y are conditionally independent given Z , further knowing

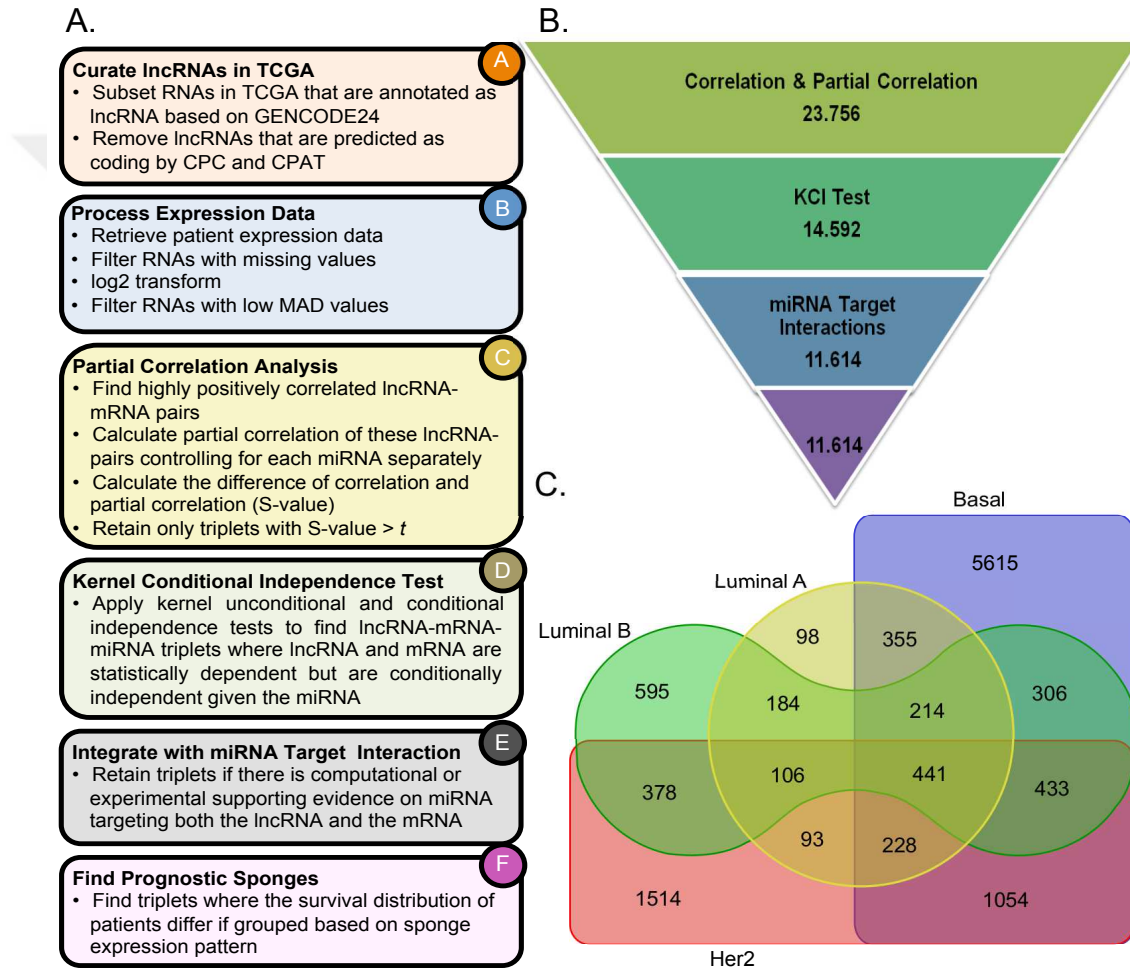


Figure 2.3: A) Overview of the methodology, each box represents a step in the methodology. Steps B-F are conducted for each breast cancer subtype separately. B) The number of ceRNAs remained after each main filtering step when $t = 0.2$ (Step C in Figure 1A). C) Venn diagram of ceRNA interactions discovered in each of the breast cancer molecular subtype.

the values of X (or Y) does not provide any additional evidence about Y (or X).

There are conditional independence tests available for continuous random variables [69, 70, 71, 55]. In our work we employ, kernel-based conditional independence (KCI) test proposed by Zhang et al. [55] as it does not make any distributional assumptions on the variables tested. Furthermore, KCI-test does not require explicit estimation of the joint or conditional probability densities and avoids discretization of the continuous random variables, both of which require large sample sizes for an accurate test performance. Below we describe the KCI-test briefly, details of which can be found in [55].

KCI-test defines a test statistic which is calculated from the kernel matrices associated with X , Y and Z random variables. A kernel function takes two input vectors and returns the dot product of the input vectors in a transformed feature space, $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. The feature transformation is denoted by $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ [72], $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \rangle$. In this work we use the Gaussian kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_x^2})$, where $\sigma > 0$ is the kernel width. CI and KCI are based on kernel matrices of X , Y and Z , which are calculated by evaluating the kernel function for all pairs of samples, i.e. the (i,j)th entry of $\mathbf{K}_X$ is $k(\mathbf{x}_i, \mathbf{x}_j)$. The corresponding centralized kernel matrix is $\tilde{\mathbf{K}}_X \triangleq \mathbf{H}\mathbf{K}_X\mathbf{H}$ where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{1}$ is a vector 1's. $\tilde{\mathbf{K}}_Y$ and $\tilde{\mathbf{K}}_Z$ are similarly calculated for Y and Z variables.

Given the i.i.d. samples $\mathbf{x} \triangleq (x_1, x_2, \dots, x_n)$ and $\mathbf{y} \triangleq (y_1, y_2, \dots, y_n)$, the unconditional kernel test first calculates the centralized kernel matrices, $\tilde{\mathbf{K}}_X$ and $\tilde{\mathbf{K}}_Y$ from the samples $\mathbf{x}$ and $\mathbf{y}$ and then eigenvalues of the centralized matrices. The eigenvalue decompositions of centralized kernel matrices $\tilde{\mathbf{K}}_X$ and $\tilde{\mathbf{K}}_Y$ are $\tilde{\mathbf{K}}_X = \mathbf{V}_x \Lambda_x \mathbf{V}_x^T$ and $\tilde{\mathbf{K}}_Y = \mathbf{V}_y \Lambda_y \mathbf{V}_y^T$. Here Λ_x and Λ_y are the diagonal matrices containing the non-negative eigenvalues $\lambda_{\mathbf{x},i}$ and $\lambda_{\mathbf{y},i}$ in descending order, respectively. $\mathbf{V}_x$ and $\mathbf{V}_y$ matrices contain the corresponding eigenvectors. Zhang et al. [55] show that under the null hypothesis that X and Y are independent, the following test statistic:

$$T_{UI} \triangleq \frac{1}{n} \text{Tr}(\tilde{\mathbf{K}}_X \tilde{\mathbf{K}}_Y) \quad (2.5)$$

has the same asymptotic distribution ($n \rightarrow \infty$) as

$$\tilde{T}_{UI} \triangleq \frac{1}{n^2} \sum_{i,j=1}^n \lambda_{\mathbf{x},i} \lambda_{\mathbf{y},i} z_{i,j}^2, \quad (2.6)$$

Here $z_{i,j}$ are i.i.d. standard Gaussian variables, thus $z_{i,j}^2$ are i.i.d χ_1^2 - distributed. The unconditional independence test procedure involves calculating T_{UI} according to Eq (2.5). Empirical null distribution of $\tilde{T}_{UI}$ is simulated by drawing i.i.d random samples for $z_{i,j}^2$ variable from $\tilde{\chi}^2$. Finally, the p -value is calculated by locating T_{UI} in the empirical null distribution.

The kernel conditional independence test also makes use of the centralized kernel matrices. Under the null hypothesis that X and Y are conditionally independent given Z , the following test statistic is calculated:

$$\tilde{T}_{CI} \triangleq \frac{1}{n} \text{Tr}(\tilde{\mathbf{K}}_{\ddot{X}|Z} \tilde{\mathbf{K}}_{Y|Z}), \quad (2.7)$$

where $\ddot{X} \triangleq (X, Z)$ and $\mathbf{K}_{\ddot{X}}$ is the centralized kernel matrix for $\ddot{X}$. As Zhang et al. [55] report has the same asymptotic distribution as

$$\tilde{T}_{CI} \triangleq \frac{1}{n} \sum_{k=1}^{n^2} \lambda_k^\circ \cdot z_k^2 \quad (2.8)$$

The details of the definition of λ_k° and z_k^2 can be found in [55]. The procedure involves calculating the empirical p -value based on the test statistic as defined in Eq (2.5) and simulating the null distribution based on Eq (2.8).

Using the unconditional kernel independence test, we first test the null hypothesis that a lncRNA and mRNA pair is independent. We consider pairs for whom the null hypothesis is rejected at the 0.01 significance level. For each of the lncRNA:mRNA pair, we test their conditional independence given each miRNA separately using KCI. The triplet where the lncRNA:mRNA pair given an miRNA is found to be independent at significance level 0.01 are considered as potential lncRNA-mediated ceRNAs.

2.3.4 Filtering ceRNAs Based on miRNA-Target Interactions

To identify interactions that are biologically meaningful, we filtered the potential ceRNA interactions that were not supported by miRNA target information. The miRNA:mRNA and miRNA:lncRNA interactions are retrieved from multiple databases as listed in Table 2.1. The candidate sponges are retained if both mRNA and lncRNA have support for being targeted by the miRNA of the sponge.

Table 2.1: Computationally and experimentally validated miRNA-target databases used for mRNA and lncRNA. Plus signs denote databases that are used for the miRNA interactions of the RNA type. ‘P’ denotes predicted target information while ‘E’ denotes experimentally supported target information.

miRNA-Target Databases	mRNA	lncRNA	P/E	Reference
TargetScan	+	+	P	[73]
miRcode	+	+	P	[74]
mirSVR	+	+	P	[75]
PITA	+	+	P	[76]
RNA22	+	+	P	[77]
lncCeDB		+	P	[78]
mirTarBase	+	+	E	[79]
Diana LncBase		+	E	[80]

Table 2.2: Features of the prediction miRNA-target databases used for mRNA and lncRNA. Plus signs denote features that are used for the databases.

miRNA-Target Databases	Features of the Prediction Algorithms		
	Seed Match	Free Energy	Conservation
TargetScan	+	+	+
miRcode	+		+
mirSVR		+	+
PITA	+	+	
RNA22	+	+	
lncCeDB	+		+

2.3.5 Identifying ceRNAs with Prognostic Value

To evaluate the ceRNA interactions in terms of their prognostic potential, we analyzed the survival of the patients based on the expression patterns of each sponge interaction. In a sponge interaction, we expect the lncRNA and mRNA to be regulated in the same direction and miRNA to be in the opposite direction. For each ceRNA found in a subtype, the patients are divided into two groups based on the regulation patterns of the RNAs that participate in the ceRNA. For the up-down-up pattern, the first group comprises patients whose sponge lncRNA and mRNA are up-regulated, and miRNA is down-regulated; the second group includes all patients that do not fit in this pattern. Similarly, we divide the patients based on the down-up-down pattern: if both lncRNA and mRNA are down-regulated whereas miRNA is unregulated, such patients constitute one group, and the rest of the patients constitute the second group.

Based on this grouping, we tested whether the ceRNA expression pattern can divide the patients into two groups, where the survival distribution of the groups are different using log-rank test [81] (p -value < 0.05). Note that since the log-rank test is not reliable when one of the group sizes is small, we only consider the cases where after dividing the patients based on the expression pattern, the group sizes are larger than 10. Since the observed significant difference could be due to a single RNA molecule prognostic value, we only considered ceRNAs as prognostic if none of the RNAs can by itself divide the patients into groups that differ in terms of their survival distributions significantly. In each subtype, we split the patients as up-regulated and down-regulated for each of the RNA participating in the ceRNA interaction separately. If at least one of the molecules leads to groups with significant survival difference (log-rank test, p -value < 0.05), we disregard this ceRNA from the list of prognostic ceRNAs. This last step ensures that the prognostic difference is due to the interactions between the RNAs but not stem from the expression of the single RNA's expression patterns. We also tested whether RNAs were prognostic in other subtypes by conducting the log-rank test on expression data of the RNAs in other subtypes.

The identified prognostic interactions' are further summarized with f -score that reflects the interaction's prognostic value compared to the most prognostic RNA of the interaction.

$$f_{xyz} = -\log \frac{p_{xyz}}{\min(p_x, p_y, p_z)} \quad (2.9)$$

Here, p_{xyz} is the p -value attained in testing whether patient survivals differ based on the log-rank test of the triplet whereas p_x, p_y, p_z indicate the p -values obtained by testing patient survival distribution differences due to lncRNA, mRNA, and miRNA expression patterns, respectively. Thus f -score is the log-fold decrease in the p -value when the patients are divided based on the interaction.

In the above analysis, RNAs that have expression levels above (or below) a particular threshold value are considered up-regulated (or down-regulated). This threshold value is selected among the candidate cut-off values of expression as the one that results in the lowest p -value in the log-rank test when patients are divided based on this cut-off. The candidate cut-off values are the 10th and 90th percentiles, mean, median or the lower and upper quartiles of the expression values of the patients in each subtype.

2.3.6 Pathway and GO Enrichment Analysis

We conduct gene ontologies (GO) enrichment analysis of mRNAs that participate in subtype-specific sponges. Enrichment tests are conducted with clusterProfiler [82] with Bonferroni multiple hypothesis test correction. In deciding enriched pathways and GO terms, a p -value cut-off of 0.05 and FDR cut-off of 1×10^{-4} are used. In both pathway and GO enrichment analyses, the background genes were the union of mRNAs that remained after the MAD filtering step (Step B in Figure 2.3(A)). For pathway enrichment analysis, different pathway data sources were downloaded from BaderLab GeneSets Collection [83]. List of all pathways that are employed in this analysis is provided in Table 2.3. Redundant pathways are eliminated when different sources are combined. Additionally, a

Table 2.3: Pathway data sources utilized for enrichment analysis and number of pathways in each data source.

Source	# of Pathways	Version/Frozen Date
HumanCyc	240	v20.5
Institute of Bioinformatics (IOB)	33	July 2011
MSigdb	520	v5.1
NCI	223	Feb 2016
NetPath	25	Jun 2016
Panther	175	Jul 2016
Reactome	18889	Dec 2016

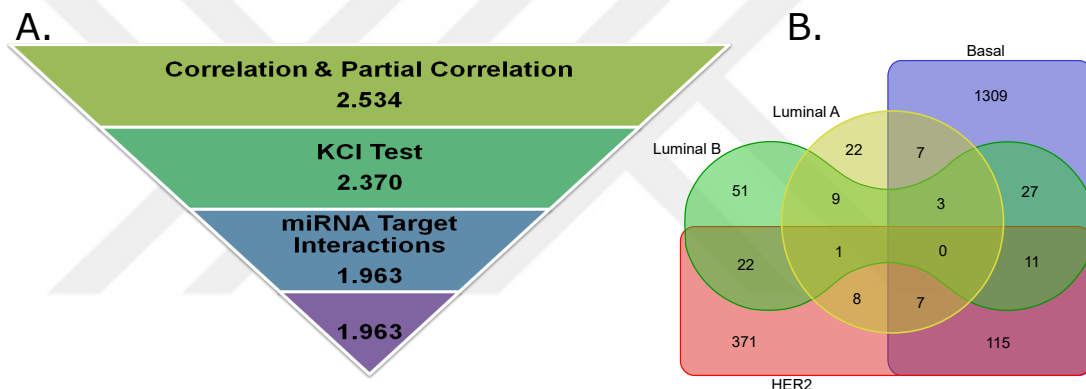


Figure 2.4: A) Number of ceRNAs remained after each main filtering step when $t = 0.3$ threshold is used. B) Venn diagram of ceRNA interactions discovered in each of the breast cancer molecular subtype.

pathway enrichment analysis is conducted with KEGG pathways (downloaded on February 28th 2017).

2.3.7 Clustering mRNAs

If mRNAs are highly correlated with each other, we often find that correlated mRNAs participate in ceRNA interactions with the lncRNA and miRNA pair.

We consider the mRNAs that participate in a ceRNA interaction with the same pair of lncRNA and miRNA. If all mRNAs are strongly correlated among each other, where all the pairwise correlations are above 0.7, all mRNAs are assigned into the same cluster. Otherwise, we apply Ward hierarchical clustering

method to find groups of correlated mRNAs [84]. We determine the optimal number of clusters with Mojena’s stopping rule [85] using Milligan and Cooper’s [86] correction.

2.4 Results

2.4.1 Overview of Discovered ceRNA Interactions

The ceRNA hypothesis states that transcripts with shared miRNA binding sites compete for post-transcriptional control [43, 46]. Based on this hypothesis, we set out to discover subtype-specific breast cancer ceRNA interactions where lncRNAs can act as miRNA sponges to reduce the amount of miRNAs available to target mRNAs. We employ the methodology summarized in Figure 2.3(A) and identify ceRNAs specific to four molecular subtypes of breast cancer: Luminal A, Luminal B, HER2, and Basal. The number of candidate ceRNA interactions that remain after each main step when in the partial correlation analysis step S value threshold $t = 0.2$ is employed, is provided in Figure 2.3(B) (see Figure 2.4 for $t = 0.3$). The total number of ceRNA interactions found in all subtypes is 11,614. Figure 2.3(C) shows the Venn diagram of number of ceRNA interactions discovered for the four subtypes (see Figure 2.4 for $t = 0.3$).

Although there are sponges that are detected in multiple subtypes, there are also a large number of sponges that are only specific to a single subtype (Table 2.4 and Figure 2.5).

We analyze the specificity of the individual RNAs that participate in each of the subtypes. Figures 2.6A and B display the number of sponges per lncRNA and miRNA for $t = 0.2$ (Figure 2.5 for $t = 0.3$). Some lncRNAs and miRNAs participate in sponges of all the subtypes (Table 2.5); i.e., KIAA0125 (FAM30A) participates in a large number of sponges across the four subtypes. KIAA0125 has been reported to act as an oncogene in bladder cancer related to cell migration and invasion [87]; however, no functional relevance to breast cancer has been reported

Table 2.4: Number of ceRNA interactions identified in each breast cancer subtypes at $t = 0.3$ and $t = 0.2$. Total number of all ceRNA interactions and number of subtype specific ceRNAs are provided.

Subtypes	# of ceRNA Interaction			
	0.3 Threshold		0.2 Threshold	
	Found	All	Found	All
Luminal A	57	22	1719	98
Luminal B	124	51	2657	595
Basal	1479	1309	8646	5615
HER2	535	371	4247	1514

Table 2.5: List of lncRNAs & miRNAs that are found to participate in sponges of all four subtypes.

miRNA	lncRNA
hsa-miR-142	LOC100188949
hsa-miR-196a-1	C5orf58
hsa-miR-127	LOC100233209
hsa-miR-155	HCP5
hsa-miR-150	KIAA0125
hsa-miR-196a-2	C21orf34
hsa-miR-125b-2	MIR155HG

to date. HOTAIR, which is one of the lncRNAs that has been associated with metastasis [88], is found to participate in sponges of all the subtypes except HER2. Similarly, miRNAs hsa-miR-142, hsa-miR-150, and hsa-miR-155 participate in ceRNA interactions of all subtypes.

There are also RNAs that take part in sponges exclusively in a single subtype, Table 2.6. For example, the lncRNA C17orf44 (LINC00324) is specific to HER2 (Figure 2.6(A)) while hsa-miR-342 is only found in Basal ceRNA interactions (Figure 2.6(B)). Several studies indicated that miR-342 is linked to BRCA1 mutated breast cancer, most of which are the Basal subtype [89, 90, 91]. Similarly, some mRNAs are involved in ceRNA interactions only in a single subtype. These subtype-specific RNAs are of great value for understanding the dysregulated cellular mechanisms in each subtype.

The lncRNA:mRNA networks for each subtype are shown in Figure 2.7. In

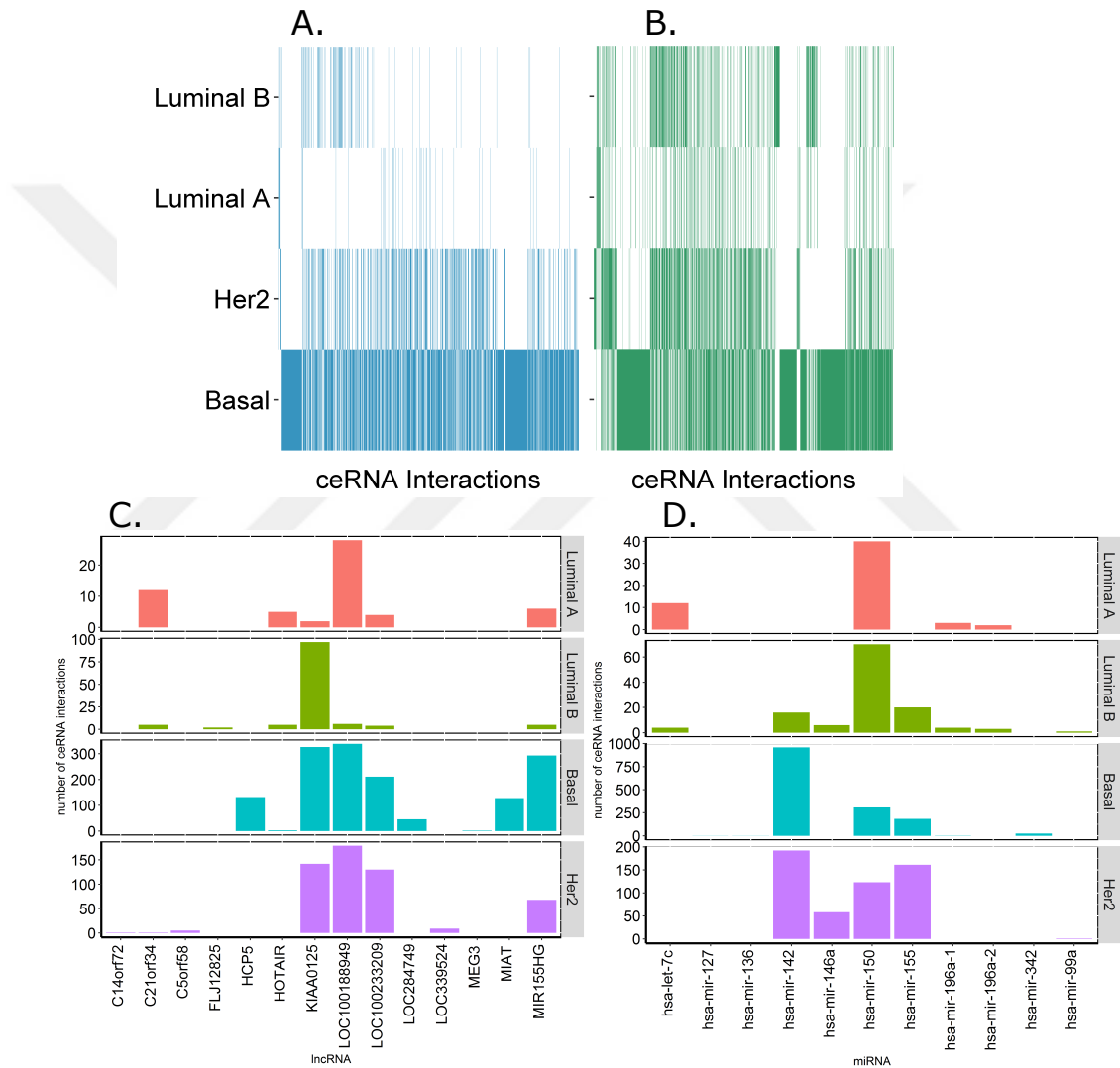


Figure 2.5: A-B) Heatmaps display the distribution of ceRNAs over the subtypes for A) $t = 0.3$ and B) $t = 0.2$. Blue and green cells indicate the ceRNA interaction is discovered in the given subtype, while white color indicates it is not. Number of ceRNA interactions discovered that per C) each lncRNA and per B) each miRNA in each breast cancer subtypes ($t = 0.3$).

Table 2.6: List of lncRNAs & miRNAs that are found in sponges of single subtype.

Subtype	miRNA	lncRNA
Luminal A	hsa-miR-381	
Luminal B	hsa-miR-431 hsa-miR-758 hsa-miR-708 hsa-miR-214 hsa-miR-370	
HER2	hsa-miR-29b-1 hsa-miR-140 hsa-miR-149 hsa-miR-23b hsa-miR-9-1 hsa-miR-379 hsa-miR-675 hsa-miR-101-1 hsa-miR-502 hsa-miR-30b hsa-miR-223 hsa-miR-34a hsa-miR-26a-2 hsa-miR-9-2 hsa-miR-511-1 hsa-miR-1270-1 hsa-miR-148a hsa-miR-146b hsa-miR-18a hsa-miR-29b-2 hsa-miR-301a	FLJ37453 MIR17HG C17orf44 LOC254559 C8orf51 PP14571 H19 SNHG3 HESRG
Basal	hsa-miR-10b hsa-miR-1245 hsa-miR-493 hsa-miR-342 hsa-miR-17 hsa-miR-20a hsa-miR-577 hsa-miR-337 hsa-miR-3614 hsa-miR-200c	LOC284749 C17orf91 LOC388692 KIAA1529 LOC678655

these networks, each node denotes a lncRNA or an mRNA while an edge represents an interaction through a shared miRNA. The number of nodes and edges are provided in Table 2.7. In Luminal A, lncRNA LOC100188949 (LINC00426) regulates the majority of the sponge interactions, while C21orf34 (MIRHG99AHG) also form a smaller connected component of its own. In Luminal B, KIAA0125 is at the center of the many interactions while a few other lncRNAs among them are HOTAIR and C21orf34 mediates a small number of interactions. Basal and HER2 subtypes include a large number of interactions. In Basal subtype, among others HCP5, MIR155HG, MIAT are the hubs of the network. In HER2 subtype KIAA0125, LOC100188949 and LOC100233209 (PCED1B-AS1) are the top 3 largest hubs.

Table 2.7: Number nodes and edges for bipartite lncRNA-mRNA networks for each breast cancer subtypes where each node denotes lncRNA or mRNA and each edge represents a lncRNA-mRNA interaction,miRNA.

Subtypes	# of Nodes	# of Edges
Luminal A	54	57
Luminal B	106	124
Basal	574	1479
HER2	272	535

We find that the same lncRNA:miRNA pair participates in multiple sponge interaction with different mRNAs. As an example, HER2 subtype-specific C14orf72:hsa-miR-150 lncRNA:miRNA pair interacts with 45 different mRNAs, the same is not true for lncRNA:mRNA pairs. The number of ceRNA interactions per lncRNA:miRNA pairs is provided in Figure 2.8. We also analyze the data by clustering mRNAs that participate in a sponge with the same lncRNA:miRNA pair based on mRNA expression correlation.

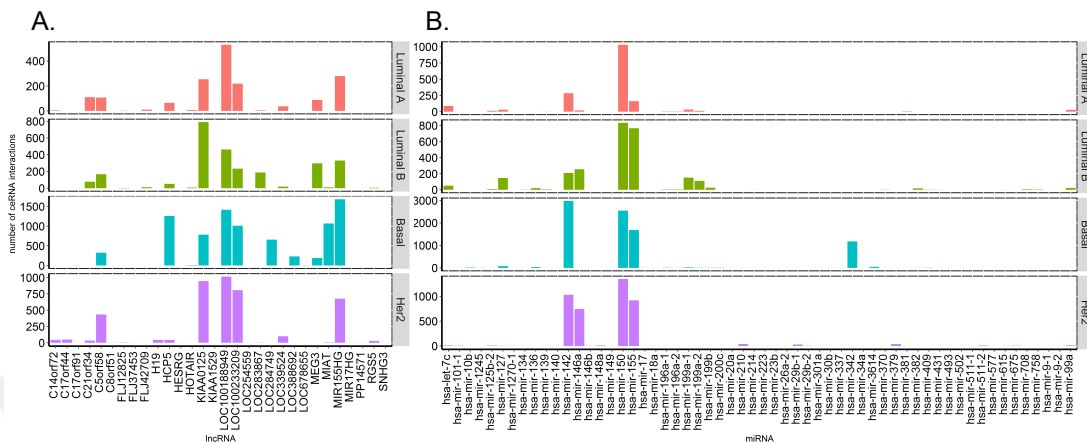


Figure 2.6: Number of ceRNA interactions discovered that A) lncRNAs and B) miRNAs take part in each breast cancer subtype ($t = 0.2$).

2.4.2 Spatially Proximal ceRNAs Interactions

In the prior section, we analyzed all possible ceRNA interactions including both distal and spatially proximal ceRNA interactions. Although the regulatory interactions can take place between molecules encoded in different chromosomes, spatial proximity often hints at a tight regulatory coordination. Also there are several studies highlighting the functional relevance of spatially proximal RNA interactions (not necessarily to be a ceRNA interaction), and we reasoned that chromosomal proximity of RNAs involved in a ceRNA interaction could also be functional. Therefore, we analyzed all the ceRNA participating RNAs within 100 Kb distance of each other, and identified several potentially important proximal ceRNA interactions. For example, we found that on chromosome 12, there is a potential sponge interaction that takes place between HOTAIR, hsa-miR-196a and HOXC genes (Figure 2.9(A)) which could be an important ceRNA interaction to contribute to the HOTAIR's known oncogenic functions as reported previously.

HOX genes are highly conserved transcription factors that take master regulatory roles in numerous cellular processes including development, apoptosis, receptor signaling, differentiation, motility, and angiogenesis. Their aberrant expression has been reported in multiple cancer types [92]. HOXA is reported to have altered expression in breast and ovarian cancers; other HOX genes are also

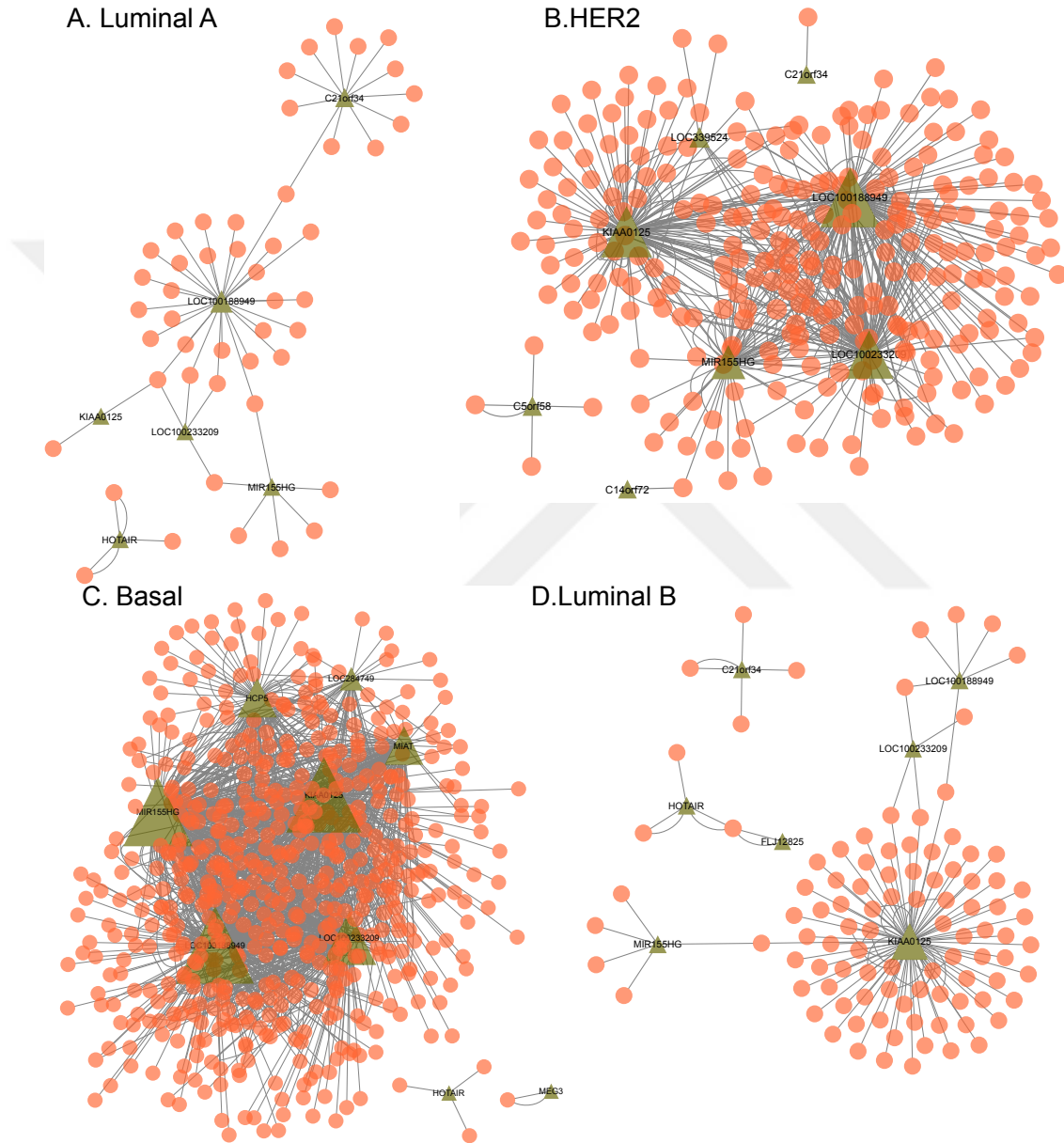


Figure 2.7: lncRNA-mRNA network for each breast cancer subtypes. Green triangle nodes represent lncRNA and circle orange nodes represents mRNA. An edge between an mRNA and a lncRNA is drawn to represent a ceRNA interaction through a miRNA. Node size is in proportion to degree of the node. The network plot was generated with Cytoscape(v3.4.0)[3].

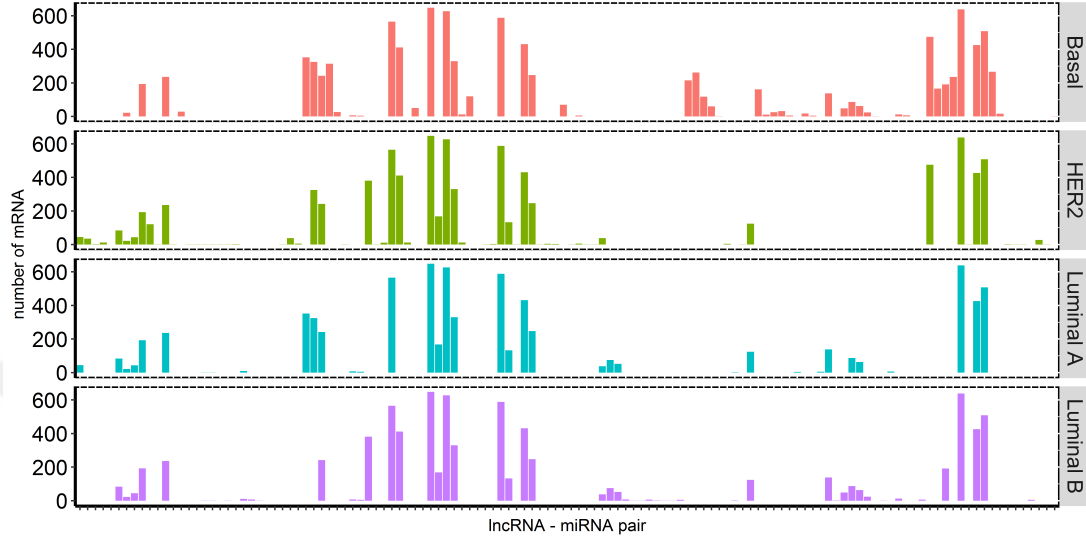


Figure 2.8: Number of mRNAs that each lncRNA-miRNA pair interacts with in each subtype ($t = 0.2$).

associated with multiple tumor types, including colon, lung, and prostate cancer. The lncRNA partner of this sponge interaction is HOTAIR. Up-regulation of HOTAIR is associated with metastatic progression and low survival rates in breast, colon, and liver cancer patients [45, 93, 87, 47, 48, 50, 94, 95, 96]. Please note that, whether it is spatially proximal or distal, all the ceRNA interactions are expected to occur in the cytoplasm as the RISC complex necessary for miRNA binding is localized in the cytoplasm.

2.4.3 Functional Enrichment Analysis of mRNAs in ceRNAs

To understand the patterns of pathways related to identified sponges, we conducted pathway enrichment analysis of mRNAs that participate in the sponges separately. The top enriched pathways are found to be common across subtypes, Figure 2.10 and Figure 2.11, and these pathways are mostly related to the immune system and signaling pathways, which are essential modulators of cancer progression and therapy response [97]. Interferon alpha/beta signaling pathway is among the top pathways for Basal subtype (p -value 7.20×10^{-23}) while it is not

found enriched in other subtypes (p -value cut-off 0.05 and FDR cutoff 1×10^{-4}).

Considering the key role of interferon signaling in the immune system, and the positive correlation between immune cell infiltration and aggressiveness of Basal subtype of breast cancer, our results suggest that mRNAs involved in ceRNA interactions might contribute to the different immune profile of the Basal subtype [98, 99]. Complement cascade induces cell proliferation which causes carcinogenesis including invasion, cell death, and metastasis [100], which are Basal subtype characteristics. We detected C2, C3, C3AR1, C4A, C7 complement genes in Basal ceRNA interactions. Consequently, complement cascade pathway may be significant for the Basal subtype.

The overlap between the enriched pathways in different subtypes is shown on a Venn diagram (Figure 2.12). The list of pathways that are found enriched only in a single subtype is listed in Table 2.8 with p -value cut-off 0.05 and FDR cutoff 1×10^{-4} . Interestingly, the PI3K pathway is found to be enriched specifically in Luminal A. This is interesting as the most frequently mutated gene in Luminal A is PIK3CA (45% of the patients in TCGA), and there are PIK3CA mutations that are specific to this subtype [54]. This suggests that ceRNA interactions might be key regulators of the PI3K pathway, especially in this subtype of tumors which comprises of more than 60% of breast cancers.

Integrin signaling is widely studied in breast cancer literature since integrins incorporate breast cancer progression [101]. Moreover, integrins play key roles in migration, invasion, and metastasis of cancer cells. Enrichment of integrating signaling in mRNAs involved in ceRNA networks might suggest that HER2-specific ceRNA interactions might contribute the aggressive progression of HER2 subtype, similar to the Basal subtype of breast cancer. Thus, they drive tumor cell to metastasis [101]. HER2 subtype-specific enriched pathways contain integrin signaling pathways (Table 2.8).

Table 2.8: List of subtype specific enriched pathways. Bonferroni corrected p values are provided and all listed pathways are above FDR cutoff 1×10^{-4} .

Subtypes	List of Subtype Specific Pathways	p-value
Luminal A	• Constitutive Signaling By Aberrant PI3K In Cancer	2.84×10^{-6}
	• Biocarta il17 Pathway	5.22×10^{-6}
	• Pid il2 1 Pathway	5.33×10^{-6}
	• Pid TxA2 Pathway	7.90×10^{-6}
Luminal B	• Naba Core Matrisome	3.58×10^{-20}
	• Extracellular Matrix Organization	2.30×10^{-13}
	• ECM Glycoproteins	2.66×10^{-10}
	• Naba Proteoglycans	2.51×10^{-8}
	• Formation	6.86×10^{-8}
	• Collagen Biosynthesis And Modifying Enzymes	1.19×10^{-7}
	• Integrin Signalling Pathway	5.37×10^{-7}
	• Assembly Of Collagen Fibrils And Other Multimeric Structures	1.89×10^{-6}
	• Pid Integrin1 Pathway	2.99×10^{-6}
	• $\beta 1$ Integrin Cell Surface Interactions	2.99×10^{-6}
Basal	• Pid $\alpha v \beta 3$ Integrin Pathway	3.42×10^{-6}
	• Pid Syndecan 1 Pathway	6.38×10^{-6}
	• Interferon Alpha Beta Signaling	7.20×10^{-23}
	• Antigen Presentation: Folding, Assembly And Peptide Loading Of Class I MHC	1.05×10^{-9}
	• ER-phagosome Pathway	2.52×10^{-8}
	• BCR Signaling Pathway	2.84×10^{-8}
	• Biocarta Complement Pathway	8.14×10^{-8}
HER2	• Pertussis	1.17×10^{-6}
	• Complement Cascade	7.16×10^{-6}
	• Integrin Cell Surface Interactions	3.68×10^{-9}
	• SA MMP Cytokine Connection	2.62×10^{-8}
	• Pid $\alpha m \beta 2$ Neutrophils Pathway	2.98×10^{-7}
	• Class I Pi3k Signaling Events	7.47×10^{-7}
	• Il27-Mediated Signaling Events	9.33×10^{-7}
	• Signaling Events Mediated By Stem Cell Factor Receptor (c-kit)	9.86×10^{-7}
	• Calcineurin-Regulated Nfat-Dependent Transcription In Lymphocytes	2.24×10^{-6}
	• CCR1	4.27×10^{-6}
	• Pid GMCSF Pathway	5.01×10^{-6}
	• DAP12 Interactions	7.27×10^{-6}
	• $\alpha m \beta 2$ Integrin Signaling	8.24×10^{-6}
	• JAK/STAT Signaling Pathway	9.25×10^{-6}
	• AGE-RAGE Signaling Pathway in Diabetic Complications	1.64×10^{-5}

2.4.4 Prognostic Sponge Interactions

To identify ceRNA interactions with prognostic value in each subtype, for each of the identified sponge we checked whether the sponge expression pattern divides the patients into groups that differ in their survival probability. To this end, for each potential ceRNA based on the participants up or down-regulation pattern, we divided the patients into two groups and checked if the survival of these groups differs significantly using log-rank test (details provided in the Methods section). For the cases where we observe a significant difference, we further checked if the observed difference could be attributed to the prognostic

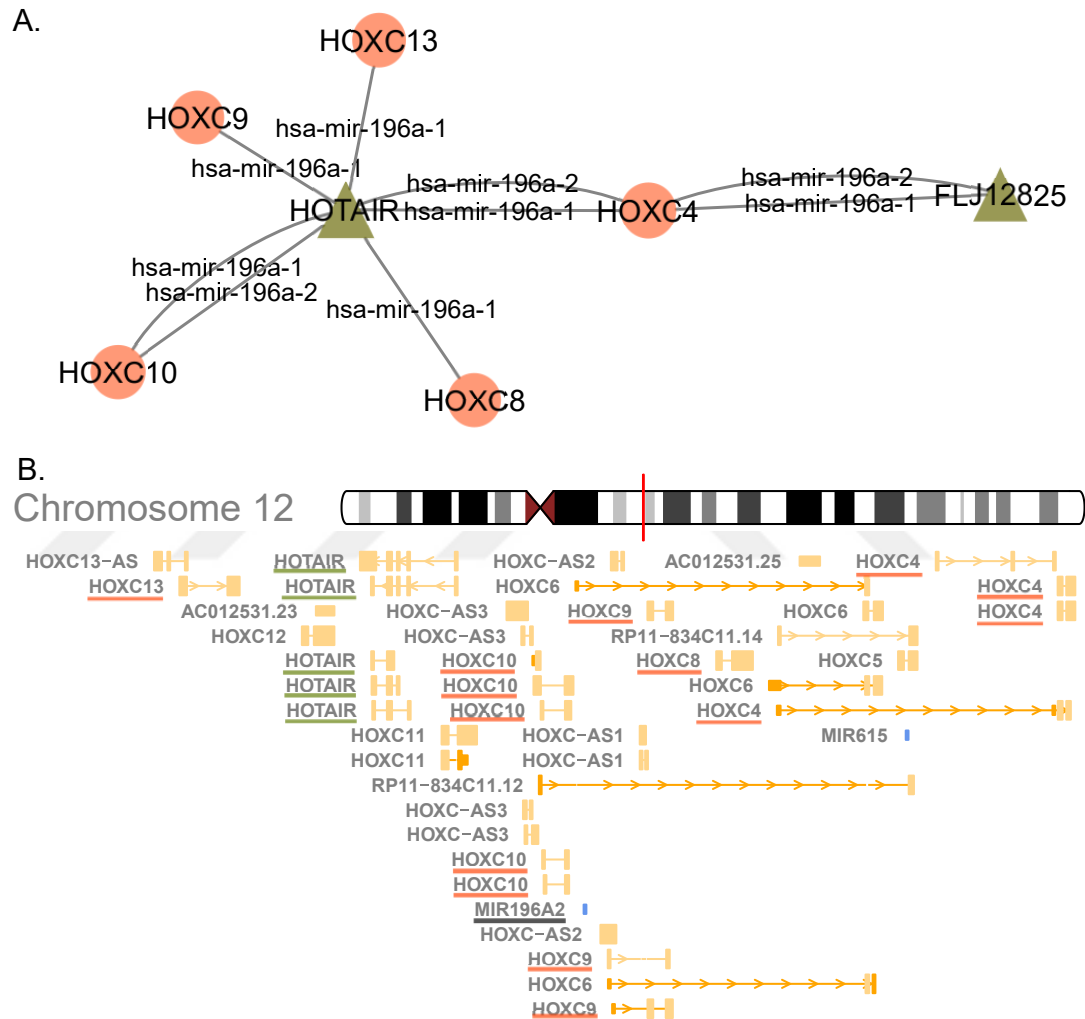


Figure 2.9: A) The network of sponge interactions between HOTAIR, hsa-miR-196a miRNAs and HOXC genes. The circles denote lncRNAs, and triangles denote mRNAs. An edge exists between a lncRNA and an mRNA if there is a sponge interaction between them; the edge label indicates the miRNA that regulates the interaction. B) The genomic locations of the sponge interactions on chromosome 12. Spatially proximal genes are underlined with the same color code of used in (A).

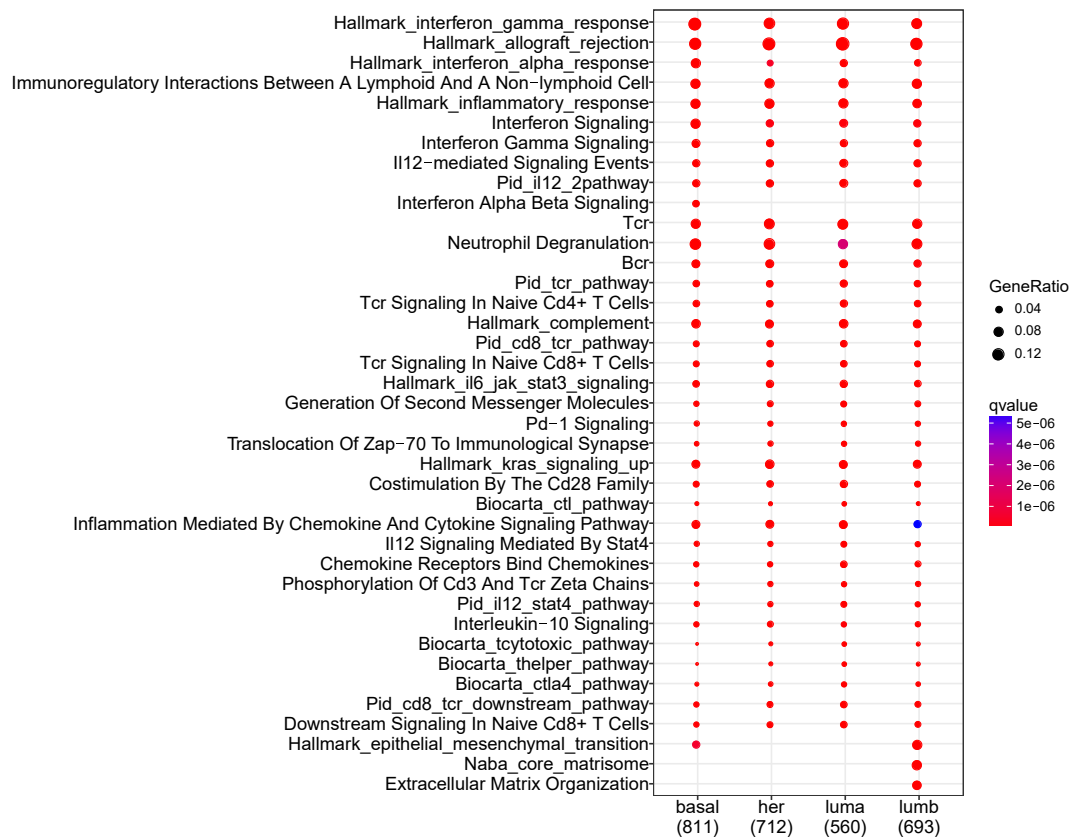


Figure 2.10: The dot plot for the most significant 27 enriched pathway which were filtered out by p -value cutoff 0.05 and FDR cutoff 1×10^{-4} . Dots in the plot are color coded depending upon the relevant FDR value. Color gradient changes from red(low FDR value, high enrichment) to blue(high FDR value, low enrichment). Dot size depends on the gene ratio, ratio of enriched genes to identified genes in the pathway. Number of identified genes in each subtypes were provided in parenthesis.

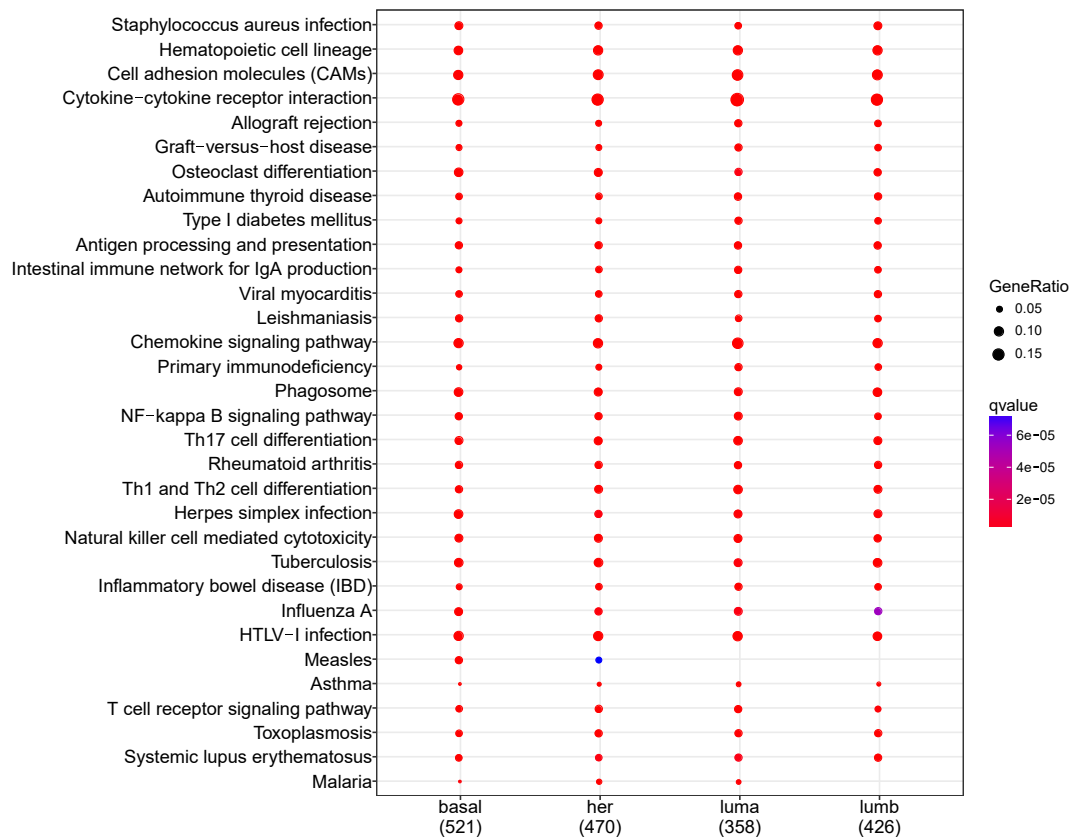


Figure 2.11: The dot plot for the most significant 27 enriched KEGG pathway which were filtered out by p -value cutoff 0.05 and FDR cutoff 1×10^{-4} . Dots in the plot are color coded depending upon the relevant FDR value. Color gradient changes from red(low FDR value, high enrichment) to blue(high FDR value, low enrichment). Dot size depends on the gene ratio, ratio of enriched genes to identified genes in the pathway. Number of identified genes in each subtypes were provided in parenthesis.

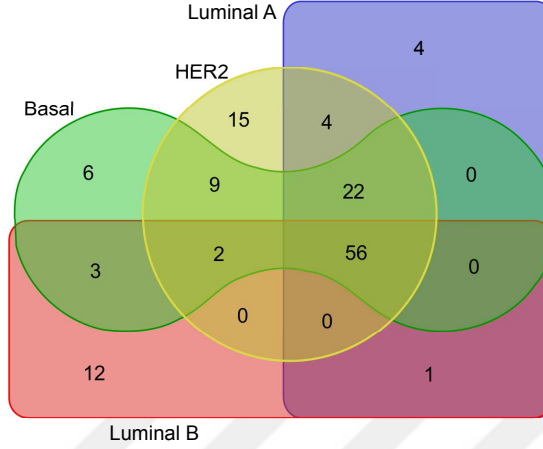


Figure 2.12: Venn diagram displaying the distribution of enriched pathways (p -value ≤ 0.05 and $FDR \leq 1 \times 10^{-4}$.) over the subtypes.

power of a single RNA molecule in the interaction by performing a log-rank test on each of the constituents' expression pattern. We only considered the ceRNA interactions as prognostic for this subtype if there was a significant difference in survival when patients were grouped based on ceRNA expression pattern but there was no significant difference if the patients were grouped based on a single RNA molecules' expression pattern. An example prognostic ceRNA interaction is shown in Figure 2.13; patients with a sponge pattern where lncRNA MEG3 and mRNA COL12A1 are high miRNA miR-1245 are low have better survival than other patients (Figure 2.13 d) while none of the three RNA molecules can separate the patients into groups that differ in survival probabilities individually (Figures 2.13 a, b and c). This result suggests that examining sponge patterns might have a better prognostic value than that of the individual genes. The networks of prognostic sponges in each subtype highlight that some of the lncRNAs and miRNAs are central in these interactions (Figure 2.14).

We summarized the prognostic ceRNA interactions' prognostic values with f -score (details in Methods), which is the log-fold decrease in the p -value of the ceRNA interaction compared to the best RNA molecule in the ceRNA interaction. The overall distribution of f -scores are provided in Figure 2.15. The presence of each RNA as a participant of a prognostic RNA across each subtype are provided in Figure 2.16 a, b and c.

2.5 Discussion

To find the triplets of lncRNA, mRNA, and miRNAs that are likely to form sponge interactions, we develop a method that uses statistical tests on patient RNA expression profiles. We start this analysis with likely physically interacting list of miRNA interactions. To this end, we retrieve experimentally validated and computationally predicted miRNA:lncRNA and miRNA:mRNA information from multiple databases. Due to the limited knowledge on experimentally confirmed interactions, we also choose to include predicted targets in our analysis. While the inclusion of predicted RNA interactions allows us to investigate a wider list of candidates it is also likely to increase the false positive rates. If more experimental target information becomes available, the framework can be altered to exclusively use the experimentally identified miRNA target information and the statistical tests can be performed only on this set of candidate triplets.

A statistical interaction does not automatically imply a physical interaction in the cell, because just like any other intracellular interaction, the RNA-RNA interactions are context dependent. Our predicted interaction set, therefore, constitutes a candidate list that can be probed and tested experimentally. To assist in prioritizing candidate interactions, we curate additional information such as the number of databases supporting a predicted interaction and the number of miRNA binding sites in the lncRNA partner. The latter is important because for the lncRNA to be tittered down by the miRNA, the presence of multiple miRNA binding sites might be required. We also provide subsets of interactions through computational means such as the cis-RNAs interactions and the interactions with prognostic potential. This additional information and the subsets can be used to prioritize interactions for experimental validation and can help explore different aspects of RNA regularity mechanisms in breast cancer subtypes.

One challenge is the unavailability of experimentally validated and lncRNA mediated sponge interactions for breast cancer subtypes. This limits the efforts to assess the statistical power and the false positive rate of our method and complicates the choice of cut-off values used in the compilation of the final candidate

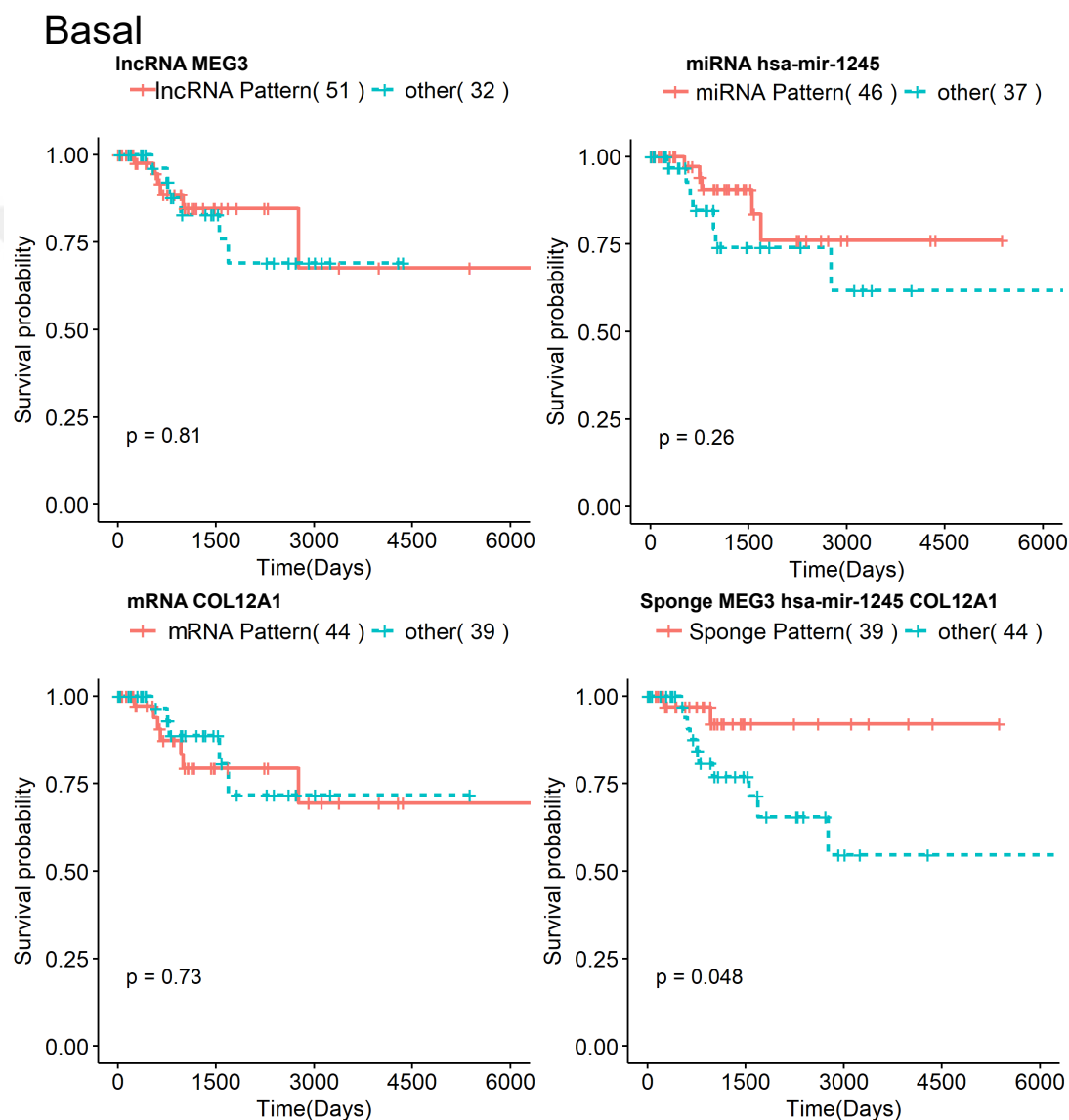


Figure 2.13: Kaplan-Meier survival plots when patients are divided based on individual expression patterns of the RNAs (the first three plots in each panel) and when patients are divided based on the sponge expression pattern (4th plot) for MEG3, hsa-miR-1245, COL12A1 sponge.

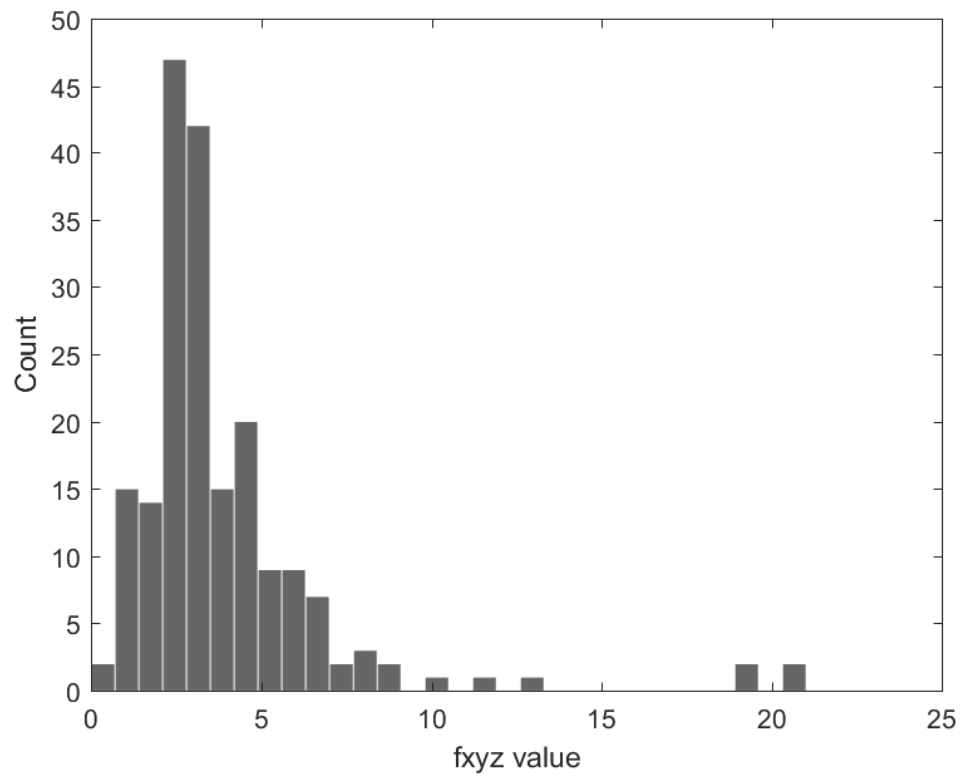


Figure 2.15: The distribution of the f_{xyz} scores of the prognostic ceRNA interactions.

list. For this reason, we report our results at two cut-off values that differ in their stringency. To assist with these analyses, we also perform several additional analyses to investigate the relevance of the discovered potential interactions. This way we can validate the interaction list with indirect supporting evidence. Firstly, a functional enrichment analysis of mRNAs that involve subtype-specific sponges is done. This analysis reveals subtype specific mRNA partners that are enriched with pathways/processes known to be specific to some of the subtypes. We consider this as an indirect validation. Secondly, the spatial organization of the RNA triplets that participate in the genome reveals that some of the sponges are positioned in close proximity of each other on the genome, hinting a regulatory relationship between these RNAs. Thirdly, a subset of the interactions is found to have prognostic value. Based on the sponge expression patterns, patients can be divided into two groups that differ in terms of their survivals.

Chapter 3

Identifying Cooperating Pairwise miRNAs

3.1 Introduction

MicroRNAs (miRNAs) are small non-coding RNAs that play extensive regulatory roles across different biological processes and pathologies such as cancer [22, 102]. By binding the complementary sequence of the target messenger RNAs (mRNAs), miRNAs repress their target mRNA gene expression [33]. Despite the widespread changes they have on the protein levels, single miRNAs typically induce modest repression on their targets [103]. Cooperation among miRNAs is one strategy to boost their activity. This cooperativity can be realized in the form of multiple miRNAs collectively co-regulating a cohort of genes that are part of a functional module, or multiple miRNAs can jointly co-target an individual gene to achieve cooperative control of a target gene (reviewed in [104]). In this work, we focus on the latter case.

Despite the accumulating evidence, our understanding of synergistic miRNAs is still rudimentary. Determining which pairs of miRNAs can synergize and which gene is the target of this cooperativity is critical for understanding how miRNAs

function and how they regulate various gene expression programs. Furthermore, synergism can be exploited in designing miRNA based therapeutics as pointed in a recent survey by [105]. Single miRNA treatments require high amounts of miRNAs, which can provoke off-target effects while therapeutics based on synergistic miRNAs have the potential to overcome these unintended effects by enabling lower levels of each of the miRNAs to be administered [105].

In this study, we develop miRCoop for detecting synergistic miRNAs and their mutual targets. We focus on a specific type of synergistic relationship: the miRNA pairs which have weak or no repression on the target mRNA but when together induce strong repression of the expression. Note that this does not cover all the synergistic relationships, i.e., the case where miRNAs have moderate expression themselves can also be in a synergistic relationship. We focus on this specific case as it is statistically more realistic to find. Additionally, it can uncover more novel RNAs that are important for a disease or condition which cannot be captured by merely checking the miRNAs individual effect.

We cast the problem of finding such triplets as a statistical dependence discovery problem. Using matched miRNA and mRNA expression data and considering the expression level of each miRNA and mRNA as a random variable, miRCoop finds a specific three-way statistical interaction among these random variables. Here, the pairs of miRNAs have weak or no statistical dependence with the mRNAs, yet when taken jointly they are dependent with the mRNA. This case is similar to the example wherein adding sugar to coffee or stirring the coffee alone have weak or no effect on the coffee’s sweetness, but the joint presence of the two has a strong effect [106]. To test whether the triplets of RNAs have such relationships, we resort to the kernelized Lancaster three-variable interaction developed by [106]. This test relies on mapping the probability distributions into a reproducing kernel Hilbert space (RKHS), which is a powerful approach in machine learning (reviewed in [107]). The kernel trick helps us to deal with small samples sizes for the expression profiles available, and due to its non-parametric nature, it also avoids any distributional assumptions. In addition to the expression profiles, miRCoop make use of the putative target information among the miRNAs and mRNAs.

We apply miRCoop to kidney cancer using tumor samples from The Cancer Genome Atlas (TCGA) project. This analysis discovers 66 putative triplets. We use several validation strategies to check the existence of a regulatory relationship among the triplets’ participants. Investigating the triplets’ RNAs reveal that most of them are individually associated with kidney cancer, as well. Furthermore, we observe that some miRNA pairs are encoded very close on the genome hinting a tight regulation at the genome level and supporting a functional link among these miRNAs. To check other coordinated regulatory evidence, we investigate whether a shared transcription factor potentially coordinates the miRNAs and RNAs, and interestingly, we observe that this holds for almost all identified triplets. When we explore the relevance of the identified triplets to kidney cancer, our functional enrichment analysis on the largest connected component of miRNA–miRNA–mRNA network highlights critical processes related to kidney cancer, part of which are key driver genes. Matlab code for identification of the synergistic triplets are provided in <https://github.com/guldenolgun/miRCoop>

3.2 Related Work

There are several experimental evidence on the cooperativity of miRNAs to repress a common target. Enright et al. [108] are the first to experimentally show the cooperativity of lin-4 and let-7 in *Drosophila*. Krek et al. [109] find that in a murine pancreatic cell line that miR-375, miR-124 and let-7b cooperate to repress the MTPN gene. Similarly, Lai et al. [110] reports synergistic interaction of miR-93 and miR-572 in repressing the p21 gene in human SK-Mel-147 melanoma cells. Others also show similar lines of evidence on this synergistic interaction of miRNAs to regulate a single gene ([111, 112]). Despite these accumulating evidence mentioned above, our understanding of cooperative synergistic miRNAs is still rudimentary. Given the complexity between miRNAs and target gene, determining which pairs of miRNA can synergize and which mRNA is the target of this cooperativity is critical for understanding miRNA functions and how they regulate various gene expression programmes that underlie the cellular processes.

Computational studies reveal the interaction between multiple miRNAs and their target mRNA. Studies in the literature can be categorized into 3 subgroups, cotargeting, coregulator and synergistic miRNAs. The relationship between these methods are cotargeting miRNAs $\subseteq$ coregulator miRNA $\subseteq$ synergistic miRNAs.

3.2.1 Cotargeting miRNAs

Group of different miRNAs or same miRNAs can bind the same mRNA. These miRNAs called as cotargeting miRNAs. Cotargeting miRNAs can bind simultaneously or at different time. Thus, cotargeting miRNAs may or may not coexpress the target mRNA [113]. However, if they coexpress the target mRNA, their combined effect is stronger than the individual miRNA effect and they additively repress the expression of the target mRNAs [114].

In the literature, there are several studies that ensemble target predictions of multiple miRNAs to improve the prediction potential or to reveal the miRNA regulation [115, 116]. Also, C2Analyzer tool [117] identify the cotarget miRNA pairs and reveal whether the pair is co-functionally enriched. Tsang et al. [113] predict the miRNA functions from cotarget relationships since cotarget relationships facilitate to uncover functions of the miRNA from the function of its cotargeting miRNAs. Moreover, some studies that discover the synergistic and coregulator miRNAs are also find out the cotargeting miRNAs. GeneSet2miRNA [118] defines their miRNA-regulatory model using miRNAs cotargeting activity.

3.2.2 Coregulator miRNAs

Coregulator miRNAs are a set of miRNAs that regulates their mutual target mRNA or a set of mRNAs in the same pathway. Coregulation does not imply simultaneous regulation effect. Davis et al [119] propose a combinatorial method to detect the coregulator miRNAs. Na et al. [120] detect coregulating miRNAs that regulate mRNAs in the network simultaneously.

3.2.3 Synergistic miRNA

Synergistic miRNAs coregulate target mRNA or a set of mRNA or a pathway. They have weak or no repression on the mRNA but when they bound together, they induce a stronger effect.

There are several methods that inspect the synergism of miRNAs in regulating a set of genes that are part of a functional module ([121, 122, 123, 124, 125, 126, 127, 128, 4, 129], there are only few computational methods for discovering synergistic miRNAs that target the same mRNA. Among these methods, [130] analyze multiple miRNAs targeting a single gene, but their analysis solely relied on the miRNA-mRNA interactions derived from target prediction algorithms. TriplexRNA ([5]) is a computational framework that integrates miRNA target prediction, structure prediction, molecular dynamics simulations and mathematical modeling methods to find synergistic miRNAs and their mutual targets.

3.3 Methods

miRCoop relies on miRNA-mRNA target predictions and kernel statistical tests on the expression data. The overall pipeline of miRCoop is summarized in Figure 3.1. The first step curates a list of possible triplets based on miRNA target predictions and their predicted mRNAs. In the second step, the candidate triplets are pruned based on expression data. In the final stage, we test the remaining triplets for the statistical dependence relationships. Below, we detail these three steps. Following the miRCoop descriptions, we provide information on the null model used for triplets and the data collection and pre-processing steps when miRCoop is applied to kidney cancer.

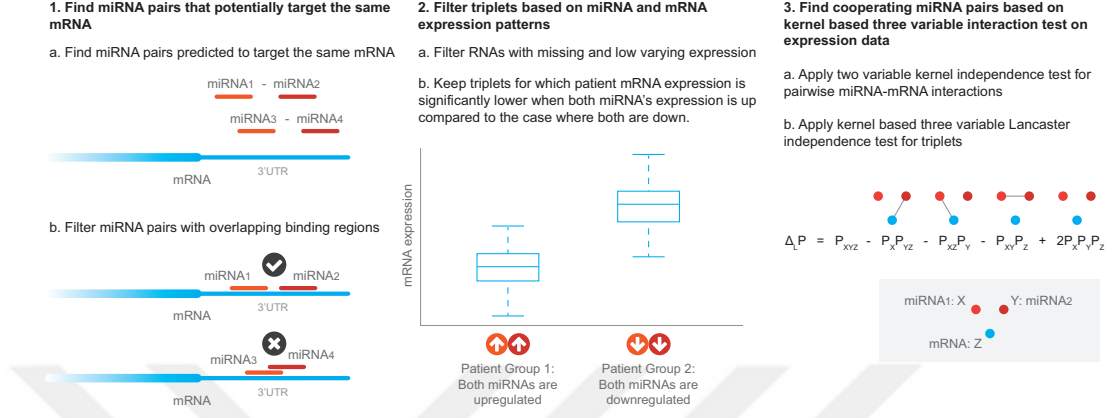


Figure 3.1: Overview of the methodology. Each box represents a step in the methodology.

3.3.1 miRCoop Step 1: Identifying Candidate miRNA–mRNA Target Interactions

miRCoop first identifies pairs of miRNA that can bind to a common target mRNA (Figure 3.1 Step 1). A pair of miRNAs and their mutual target mRNA constitute a candidate triplet. As the interactions of newly cataloged miRNAs are not included in experimentally validated miRNA target interaction databases, we resort to available target prediction algorithms. For this, we use TargetScan [73] and miRanda [75]. When using TargetScan, the 6–mer sites’ predictions are filtered since they are classified as poorly conserved [73]. Similarly, when running miRanda, to reduce false-positive predictions, we apply stringent parameter settings (a pairing score of > 155 and an energy score of < -20). We consider only the miRNA–mRNA pairs that are predicted by both algorithms. The triplets where miRNA pairs have an overlapping binding site on the target mRNA are also excluded.

Although the target information prunes many combinations, at the end of this step, there is still an extensive list of candidate triplets. For example, when we apply to kidney cancer(see Results), there are 31,184 potentially interacting miRNA–mRNA pairs, and there are 163,550 candidate miRNA–miRNA–mRNA triplets.

3.3.2 miRCoop Step 2: Expression Filter

We further narrow down the candidate list based on the expression profiles of the miRNAs and mRNAs measured over a set of samples, Figure 3.1 (Step 2). miRNAs repress the expression of their target mRNA genes. Therefore, we expect that if both miRNAs are upregulated, the mRNA expression will be lower compared to the case where both miRNAs are down-regulated. To filter for triplets where this expression pattern holds, we divide the samples into different expression subgroups based on the expression states of the potential synergistic miRNA pairs. The first group comprises the samples where both miRNAs are upregulated, and the second group is the subset of samples where both miRNAs are downregulated. In this analysis, miRNA expression in a sample is considered upregulated if it is in the fourth quartile of all samples, and it is deemed to be downregulated if it is in the first quartile. Having grouped the samples, for every triplet in the candidate list, we check whether the mRNA expression levels are significantly lower than the mRNA expression levels compared to the second group. To test for significance, we apply the one-sided Wilcoxon rank-sum test [131]. This test is used to determine whether two groups of expression values are selected from populations having the same mRNA expression distribution. We apply this test to the all possible triplets formed at the end of Step 1 and retain only those for which the null hypothesis is rejected at 0.05 significance level.

3.3.3 miRCoop Step 3: Testing Statistical Interactions of RNAs with Kernel-Based Interaction Tests

For every triplet candidate that passed through Step 2, we perform statistical interactions tests on RNA expression data to find putative miRCoop triplets, Figure 3.1 (Step 3). We are looking for a high order interaction between the three RNA entities that is while each miRNA expression levels have weak or no interaction on the mRNA expression, the simultaneous interaction of the two miRNAs with the mRNA expression levels is strong.

We denote the two miRNAs' expression levels as two random variables: X and Y and define the expression level of the mRNA to be tested as a third random variable denoted as Z . We are interested in cases where the miRNAs are pairwise independent with mRNA: $X \perp\!\!\!\perp Y$ and $Y \perp\!\!\!\perp Z$ but form a mutually dependent triplet because of the following holds: $(\neg((X, Y) \perp\!\!\!\perp Z))$. Note that this also corresponds to checking for a V-structure, a directed graphical model with two independent parents pointing to a single child, where the miRNAs' expression levels (X and Y) are the parent variables, and the child variable is the mRNA expression level (Z) [106].

To ensure that the miRNAs have weak influence individually, we first test for pairwise independence among X and Z and Y and Z . In each of these tests, the null hypothesis that the two variables are independent. We continue with triplets, where the null hypothesis is accepted. On these triplets, we apply three-variable kernel interaction tests proposed by [106] and considers only triplets which reject the null hypothesis. Here, we provide a brief description of these kernel-based tests and later detail on how we use this method for discovering miRCoop triplets.

3.3.3.1 Kernel Three-Variable Lancaster Interaction Test

An interaction measure associated to a multidimensional probability distribution P is a signed measure ΔP that vanishes whenever P can be factorized as a product of its marginal distributions [132]. Lancaster interaction [132, 106] is such a measure which is the difference between the joint probability distribution and the product of the marginal distributions. When there are two variables, X and Y , the Lancaster interaction is defined as:

$$\Delta_L P = P_{XY} - P_X P_Y. \quad (3.1)$$

For three random variables, the Lancaster interaction is defined as:

$$\Delta_L P = P_{XYZ} - P_{XY}P_Z - P_{YZ}P_X - P_{XZ}P_Y + 2P_XP_YP_Z. \quad (3.2)$$

A statistical dependency test can be constructed where the Lancaster interaction is the test statistic [132, 106]. The null hypothesis, in this case, is that the

Lancaster interaction is zero. The null distribution of the test statistic can be estimated using permutation-based approach where the indices of the variables are shuffled [106]. Sejdinovic et al. [106] formulates a kernel based three-variable Lancaster interaction test by embedding the probability distributions into reproducing kernel Hilbert spaces (RHKS). In their general formulation, the three variables to be tested form a random vector of triplets $(X, Y, Z) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p$, where p is the dimension. In our case $p = 1$.

The effectiveness of embedding probability distributions are demonstrated in different machine learning and inference task (reviewed in [133] and [107]). The key idea of is to implicitly map probability distributions into potentially infinite dimensional feature spaces using kernel functions [134]. Kernel embedding of the probability distributions avoids the restrictive assumptions about the form of the distributions such as, variables to be discrete-valued, or assumptions about the form of the relationships among the variables such as linear relationship between variables [133]. Using the kernel trick, instead of constructing the new feature space, computations on the probability distributions can be calculated implicitly via feature space operations, that is as Gram matrix operations. The probabilities can be expressed as certain expectations of kernel functions. Sejdinovic et al. [106] arrives at the following test statistic for the two variable and three-variable dependency test:

$$\| \Delta_L \hat{P} \|_{k \otimes l}^2 = \| \hat{P}_{XY} - \hat{P}_X \hat{P}_Y \|_{k \otimes l}^2 = \frac{1}{n^2} (\tilde{K} \circ \tilde{L})_{++} \quad (3.3)$$

$$\| \Delta_L \hat{P} \|_{k \otimes l \otimes m}^2 = \frac{1}{n^2} (\tilde{K} \circ \tilde{L} \circ \tilde{M})_{++} \quad (3.4)$$

Here, k denotes the kernel function defined in domain $\mathcal{X}$ that maps P_X to RHKS. K is a symmetric $n \times n$ Gram matrix for the kernel k computed over the observations' X dimension, n is the number of samples. Similarly, l and m are kernel functions defined in domains of $\mathcal{Y}$ and $\mathcal{Z}$, respectively, while L and M are the Gram matrices computed over observations sampled from domains $\mathcal{Y}$ and $\mathcal{Z}$, respectively. $\circ$ denotes an Hadamard product of matrices. $\otimes$ indicates the Kronecker product. $\mathbf{1}$ denotes an $n \times 1$ vector and $K_{++} = \sum_{i=1}^n \sum_{j=1}^n K_{ij}$. $\tilde{K}$ denotes the corresponding centered matrix of K , where $\tilde{K} = HKH$ and $H = I - \frac{1}{n} \mathbf{1}\mathbf{1}^T$ and I is the identity matrix. The same rules apply to the other gram matrices,

L and M . Given an i.i.d. sample $(x_i, y_i, z_i)_{i=1}^n$, the centered kernel matrices can be computed over the samples. Using equations 3.4, the norm of the embeddings can be computed using these centered matrices.

3.3.3.2 Applying Kernel-based Interaction Tests

We form a random vector of triplets $(X, Y, Z) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}$. In the pairwise interaction tests, where we test for independence of miRNA expression levels with the mRNA (X - Y and X - Z), we consider triplets where the pairwise independence test fails to be rejected ($\alpha = 0.01$). For the three-variable independence interaction test, we only consider triplets, where the null hypothesis of the three variable interaction is rejected ($\alpha = 0.01$). We use the Gaussian kernel as defined:

$$k(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|_2^2}{2\sigma^2}\right) \quad (3.5)$$

, where σ is a non-negative real number that determines the width of the kernel. We set σ using the median heuristic, [135]. Median heuristic is a popular method to set the Gaussian kernel width, σ . When $\sigma \rightarrow 0$, Gram matrix K becomes identity matrix and when $\sigma \rightarrow \infty$ all elements of the Gram matrix becomes 1. Therefore, important information will be lost in those boundaries [136]. Thus, the most reasonable value to set the width is the middle of the $\|x_i - x_j\|$.

In both tests, the null distribution is estimated by repeatedly permuting the indices of the mRNA expressions of the samples (Z variable), where we set the number of permutations to 2,000.

3.3.4 The Null Model for miRCoop Triplets

Having identified the triplets, to check the validity of different comparisons with the biological data, we use a random sampling-based approach to obtain a random set of triplets that constitutes the null model. We use these randomized triplets to obtain an empirical null distribution of the test statistic of interest in each

case. We construct the randomized triplets as follows: let $S = t_1, t_2, \dots, t_n$ be the set of miRCoop triplets identified. Let the i -th triplet member of S denoted by $t_i = (x_i, y_i, z_i)$, which include the miRNA pairs, x_i, y_i and the mRNA z_i . We sample a collection of B random triplet sets of each with size m . Let $S^b = \{t_1^b, t_2^b, \dots, t_n^b\}$ denote the b -th random triplet set. For the b -th sampling, t_i^b is generated as follows: the mRNA z_i is retrieved, and one miRNA pair of the triplet (x_i or y_i) is selected at random and retained, the paired miRNA triplet is though selected uniformly at random from the set of miRNAs that can potentially target z_i , R_z , as identified at the end of Step 2 of the miRCoop pipeline excluding the miRNA triplets (Figure 3.1). By repeating this procedure for all samplings $\forall b \in \{1, \dots, B\}$, we obtain a collection of B many triplet sets, $S_{b=1}^B$ the random counterparts of S . We must note that this procedure constructs a rather stringent null model. Note that if we were to select the mRNAs and miRNAs all at random, the miRNAs and mRNAs in these random triplets would likely be unrelated to each other. With such a null model, attaining significance would be essentially trivial.

We calculate the p -value of the observed statistic empirically based on this null distribution. To assess the statistical significance of the statistic of interest (such as the number of TFs (See sections 3.4.5, 3.4.4, 3.4.3), we obtain an empirical p -value by calculating the fraction of the cases in B random samplings where the test statistic is as extreme or more extreme than the observed statistic. The null model is constructed with 1,000 samplings.

3.3.5 Data Collection and Processing

We apply miRCoop on patient-derived kidney cancer samples. We retrieve RNASeq gene and miRNA expression data from GDC Data Portal [137] on Sept 2nd 2017. We only use primary tumor solid samples that concurrently include mRNA and miRNA expression data, which contain 995 tumor patient samples. To filter out the non-coding RNA genes from the RNASeq expression data and create an mRNA list, we used GENCODE v27 annotations [68]. We eliminate

mRNAs with 3'UTRs shorter than 12 nucleotides since two miRNAs cannot bind on such a short sequence at the same time. 3'UTR sequences of mRNAs are obtained using ENSEMBL BioMart [138].

We utilize the available Reads Per Kilobase Million Reads (RPKM) values and Fragments per Kilobase of transcript per Million mapped reads upper quartile normalization (FPKM-UQ) for miRNA and mRNA expression data, respectively. Genes with expression values lower than 0.05 in more than 20% of the samples are filtered. Expression values are added with a constant 0.25 to deal with the 0 gene expression values, and data are $\log_2$ transformed. We eliminate RNAs that do not vary across samples. For this purpose, we filter RNAs with median absolute deviation (MAD) below 0.5. In the final set, we have 12,113 mRNAs and 349 miRNAs in kidney tumor samples.



Figure 3.2: The number of remaining candidate triplets at the end of each step of the pipeline. The number of available RNAs that participate in the analysis is provided at the beginning.

3.4 Results

3.4.1 Overview of the synergistic miRNAs and their targets in Kidney Cancer

We applied the methodology summarized in Figure 3.1 to find synergistic miRNAs and their targeting mRNAs. The number of candidate triplets that remain after applying each of the filtering steps is provided in Figure 3.2. The target prediction step produce 31,184 potentially interacting miRNA–mRNA pairs and 163,550 miRNA–miRNA–mRNA interaction combinations. Following the

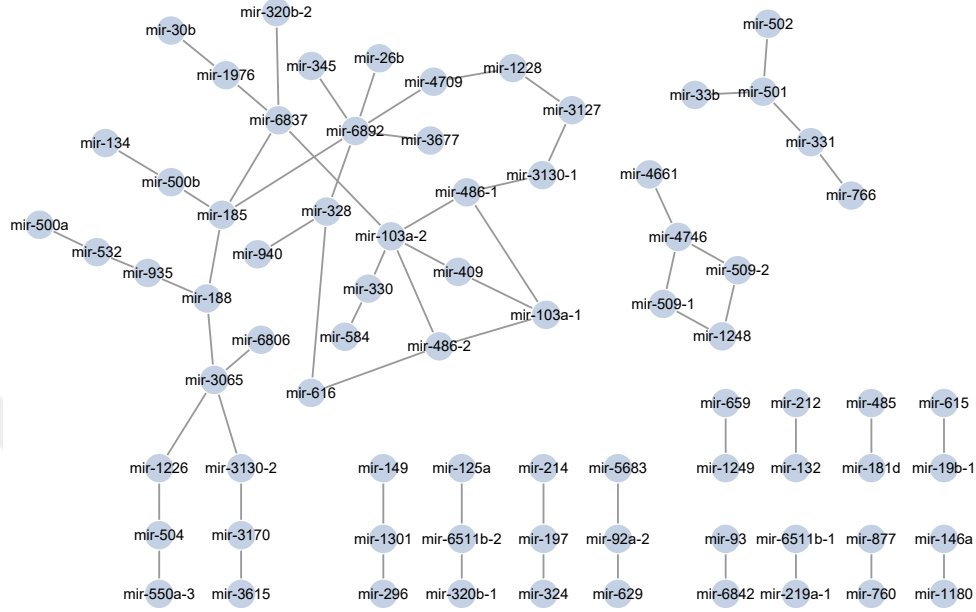


Figure 3.3: The network shows the miRNA synergistic pairs discovered for kidney cancer. Nodes are miRNAs and edges exist between them if they participate in at least one miRCoop triplet.

remaining steps, we arrive at potential 66 synergistic triplets, which also contains 66 different synergistic miRNA pairs. These triplets include 41 different mRNAs, 27 of which are reported to be cancer driver genes for kidney cancer as reported in IntOGen [139]. The complete list of the identified triplets is listed in Supplementary File 1 and the network for the potential miRNA-miRNA pairs are provided in Figure 3.3.

3.4.2 Analysis of Identified Triplets

There is no available experimentally identified set of synergistic miRNAs to validate the triplets we identified. Thus, we devise several strategies to investigate the miRCoop triplets in light of external biological information. To check the statistical significance of each of these findings, we use a null model by generating random synergistic triplets, which we detail in Methods section 3.3.4. Below we describe the biological evidence relevant to the identified triplet set.

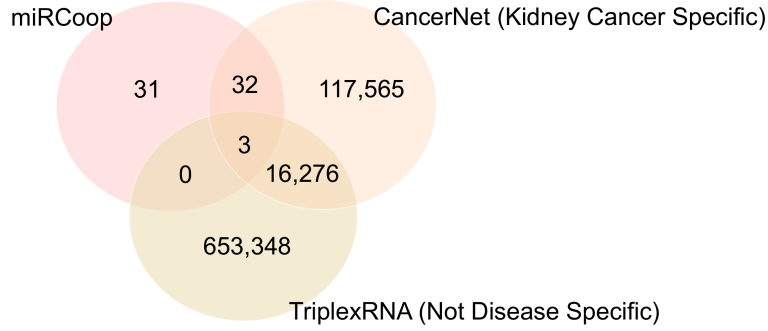


Figure 3.4: Venn diagram for the number of synergistic miRNAs that are reported by different methods for human. In CancerNet [4] we use the kidney specific triplets. The TriplexRNA [5] database predicts the RNA triplexes constitute of two synergistic miRNAs and their mutual target mRNAs. In TriplexRNA, we utilized predictions that are called as TEC. In TEC, predicted RNA triplexes are closely within 13-35 nt and their predicted equilibrium concentration (TEC) > 50 nM.

3.4.3 Comparison with Other Methods

We compare our proposed methodology with two other computational methods: CancerNet [4] and TriplexRNA [5]. CancerNet provides cancer-specific pre-computed synergistic pairs of miRNAs. Of the 66 synergistic miRNA pairs, miRCoop identifies, 35 of them are also suggested by CancerNet. This overlap is highly significant based on the sampling-based null model with a sampling size of 1,000 (p -value ≤ 0.001). When we consider the triplets that are only identified by our methodology, we observed that kidney related miRNA pairs such as miR-185 & miR-500b are discovered [140, 141].

We also check the overlap with the synergistic pairs provided by TriplexRNA; however, we do not find a significant overlap, Figure 3.4. One reason for the low overlap could be that TriplexRNA triplets are not cancer-specific while the triplets we identified are.

miRNA	miRNA	Distance	miRNA Family
miR-132	miR-212	262	miR-132
miR-501	miR-502	4792	miR-500
miR-500a	miR-532	5192	miR-500

Table 3.1: The list of identified synergistic miRNA pairs in kidney cancer that are encoded close in the genome and their family information.

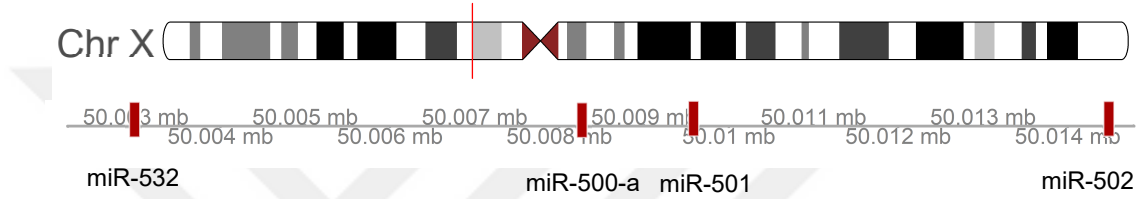


Figure 3.5: The red marks on the genome axis indicate the miRNA positions of the synergistic miRNAs that belong to the miR-500 family. The figure was created with Gviz package in R [6].

3.4.4 Triplets Coded Close on the Genome

Recent studies reveal that miRNAs, which regulate similar process, are located nearby on the genome [122]. Others report that proximal pairs of miRNAs are generally co-expressed [142] indicating a functional link. Thus, we investigate whether the identified synergistic pairs are close on the genome. Specifically, 10 kb of their transcription start and end sites within the genome are considered spatially close. We observe that three of such pairs are very close, and these pairs form two different clusters (Table 3.1). These clustered miRNAs are part of the miR-500 and miR-132 miRNA family. There is evidence reporting that miR-500 family take part in kidney cancer [143, 144, 145]. Genomic locations of these synergistic miR-500 family miRNAs are illustrated in Figure 3.5.

3.4.5 Common TF Regulating the RNAs of the Triplets

Coordinated regulation of genes by a shared transcription factor often indicates a functional link among the genes. To this end, we checked if the identified triplets are potentially co-regulated by a common TF. To be able to conduct this

analysis, we first curated a set of miRNA–TF and a set of mRNA–TF regulatory interactions.

Since none of the TF–miRNA databases are comprehensive, we use two strategies to curate the set of TF–miRNA regulatory interactions. We first use data from available databases; this includes TransmiR [146], PuTmiR [147] and miRTrans [148] databases. miRTrans database provides TF–miRNA interactions for different tissues and cell types; we only consider the renal cell lines HRE, human Renal Epithelial Cells, and RPTEC, hTERT immortalized human renal proximal tubular epithelial cell lines. Secondly, we compile the list by analyzing ChIP-Seq data made available by the ENCODE project [149] and TF motifs. We retrieve the 10,000 upstream regions of transcription start of the gene in question using the R package GenomicRanges [150] and check if this region overlaps with the ChIP-Seq peaks in three kidney-related cell lines (HRE, HEK293B, HEK293) using the GLANET tool [151]. We compile the TF–miRNA interactions from the databases and discovered through analyzing ChIP-Seq data. For the mRNA–TF regulatory interaction set, we conducted the same analysis on the ChIP-Seq data. Furthermore, we scan the promoter regions with the TFs’ binding motifs with JASPAR [152] and Hocomoco v.11 [153] using the FIMO [154] tool with default parameter settings. If the TF has a reported overlapping peak with the promoter and there is a matching motif inside the peak, we consider this as a TF–mRNA interaction.

Finally, we check if there is at least one common TF that regulate all RNAs (the miRNAs and the mRNA) that participate in a triplet interaction. In checking this, we disregard POL2 as it will overlap with promoter regions without any relevance to the functional cooperation. We observe that for 64 of the 66 triplets, all triplets participants are co-regulated by at least one common TF (p -value < 0.04). The list of these TF–triplet relations is provided in Supplementary File 2.

We inspect a subset of the TF–triplet regulations more in detail. Recent studies report that the transcription factors, HIF1A [155], TFEB, TFE3 and MITF [156] are related to the kidney cancer. We check if there are any miRCoop

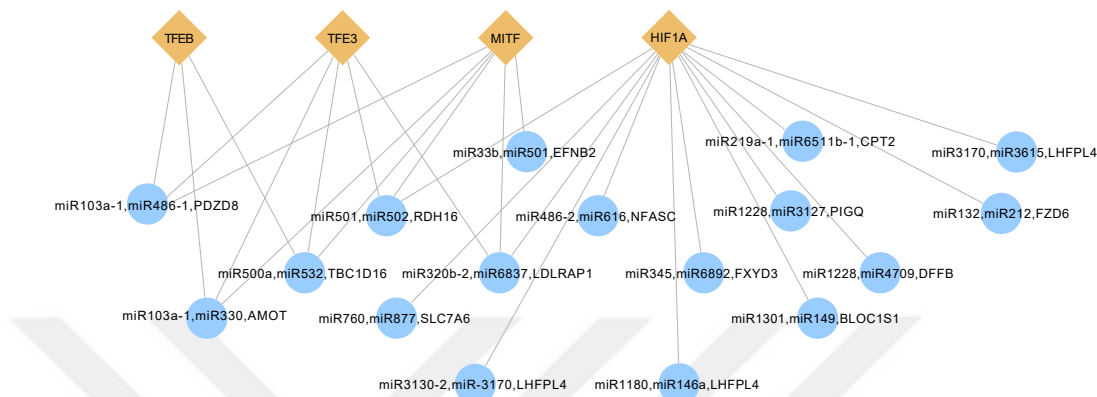


Figure 3.6: Network for the synergistic triplets regulated by common TFs. Each node denotes a triplet or a TF while an edge represents an interaction between them. TFs are represented by the diamond-shaped yellow symbol, and triplets are represented by the circle-shaped blue symbol.

triplets regulated by these transcription factors. These four TFs are linked to a total of 27 triplets (out of 66) and the sub-network includes triplets that are coordinated by multiple of these TFs Figure 3.6. Additionally, we observe that the mRNA pairs that are encoded very closely on the genome (see Section 3.4.4) are among them and have shared TFs. For example, TFE3 potentially regulates both miR-501, miR-502, miR-500a, and miR-532. Apart from this, interestingly, TFs related to the Xp11 translocation renal cell carcinoma disease (TFEB, TFE3 and MITF) [156] regulate the same triplets: (miR103a.1, miR486.1, PDZD8) and (miR103a.1, miR330, AMOT).

3.4.6 Functional Enrichment Analysis

Synergistic miRNAs are likely to regulate genes with the same or similar functions [125]. To investigate the potential biological process related to the synergistic miRNAs, we conduct a functional enrichment analysis on them. To this end, we first build a miRNA-miRNA-mRNA network. In this network, nodes are miRNAs and mRNAs, and an edge exists between two RNAs whenever they are

connected through at least one miRCoop triplet. A connected components analysis reveals that the network includes 13 connected components. Interestingly, the largest connected component covers a very large fraction of the all miRNAs: of the 75 miRNAs that is part of the triplets, 41 of them were in this largest connected component ($p - value < 0.032$) (Figure 3.7 (a)) indicating that the cooperative miRNAs are directly or indirectly linked to each other through their targets.

A gene ontology (GO) enrichment analysis reveals that the set of mRNAs are enriched in processes that are associated with kidney cancer. For example, metabolic networks such as the receptor metabolic processes, the negative regulation of hydrolase activity, and the regulation of metal ion transport are all found to be enriched (Figure 3.7 (b)). Among these, sodium and guanylate related processes are particularly reported as important in kidney cancer [157, 158, 159]. Also, the genes that are already known as the drivers of renal cancer are annotated with these processes. An example is Angiomotin (AMOT). Lv et al., [160], reports that AMOT is an oncogene in renal cancer. AMOT family members is also reported to promote cancer initiation in other cancers and act as a tumor suppressor in some others [161, 162].

Interestingly, the brain related biological processes such as layer formation in the cerebral cortex, synapse organization, embryonic brain development are found to be enriched (Figure 3.7 (b)). It is reported that brain metastasis is common for renal cancer and in 15 % of cases, the tumor metastasizes to the brain [163]. Additionally, brain metastasis is the main cause of morbidity and mortality in kidney cancer [163]. These findings suggest that mRNAs regulated by the synergistic triplets may participate in the mechanisms that promote brain metastasis.

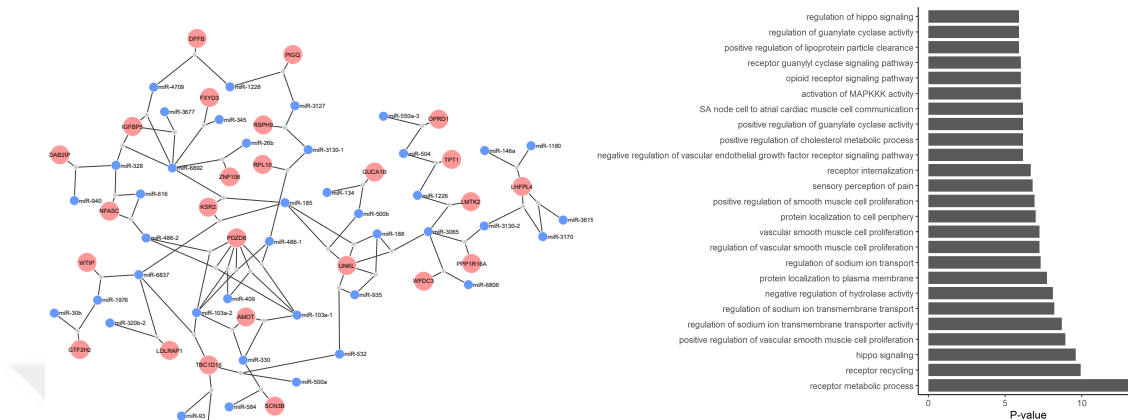


Figure 3.7: a) The the largest connected component of miRCoop triplet network. Blue nodes are miRNAs and pink colors are mRNAs of the triples. The gray nodes are dummy nodes to connect the triplet miRNAs to the target mRNA. b) Bar chart for the top 25 enriched biological process GO-terms found over the mRNA set of the largest connected component in part a)

3.5 Discussion

In this work, we focus on miRNAs that regulate a target gene directly that is those that simultaneously bind to the target. There are other ways in which pairs of miRNAs regulate a common target. For example, while one miRNA targets the gene of interest directly, another miRNA can target the transcription factor that regulates this gene [104, 164]. In this case, one miRNA is acting directly on the target, and the partner miRNA is acting indirectly through the transcription factor. An additional line of future work would be to extend miRCoop to discover these multi-level regulatory interactions.

Since there are no synergistic interaction datasets which are readily available, we devise various strategies to check if the identified triplets and the individual RNAs are implicated in kidney cancer and have any biological evidence supporting a functional theme or regulatory interaction among themselves. When experimentally validated synergistic miRNA pairs become available, the analysis can be further improved in the light of the ground truth interactions. For example, we set the kernel width using the median heuristic, but the width could be

tuned when an extensive set of ground truth interactions become available.

In the present study, to establish target relations between the miRNA and the mRNAs, we use two target prediction algorithms. We resort to prediction algorithms because newly cataloged miRNAs – which is valid for the TCGA profile miRNAs – are not incorporated in the experimentally validated miRNA target interaction datasets. To reduce the number of false positives, we only consider target interactions that are supported by both algorithms concurrently. The availability of experimentally validated miRNA–mRNA interactions should strengthen the outlined methodology. The number of synergistic triplets can be adjusted by changing the significance levels applied in the statistical tests. For example, in a scenario where a large number of interactions can be verified experimentally at scale, these settings can be adjusted to output more triplets.

Although we present results for kidney cancer, the method can be applied to other diseases as well as healthy cells. The method can be extended to work with other species other than human whenever the expression profiles of matched miRNAs and mRNAs are available for over a set of samples, and miRNA target predictions are available.

Chapter 4

Functional Enrichment Analysis of the lncRNA Derived from Its Neighbors

4.1 Introduction

Noncoding RNAs (ncRNAs) that do not encode proteins constitute most of the human genome[68]. The largest class of the ncRNAs is the long non-coding RNA genes (lncRNAs) which denote ncRNAs with more than 200 nucleotides. LncRNAs play an important role in various biological mechanisms, including physiological, chromatin modification, transcriptional/post-transcriptional regulation. Their expression is often associated with the expression of the neighborhood genes and tissue specific expression patterns[165, 166, 167]. Dysregulation in their expression states is associated with various human complex diseases such as cancer[168, 15, 169]. Even though it is known that their expression levels depend on nearby genes lncRNAs' association with biological mechanisms in cancer based on nearby genes is still largely unknown.

To analyze the functional roles of the lncRNAs based on their neighborhood

genes, we systematically examine the set of lncRNAs for 28 different cancer types that are in the Genomic Data Commons (GDC) Data Portal. We first identify dysregulated lncRNAs and perform functional enrichment analysis on these sets by transferring functional annotation knowledge from their nearby coding mRNAs on the genome. We consider nearby genes in the proximal and distal regions a set of mRNAs. We consider the overlaps in the exon, intron, 5p1, 5p2, 3p1, and 3p2 regions as described as gene-centric regions in [170]. To find the functional similarity among different cancers, we measure the semantic similarity of the enriched GO terms set between cancer pairs and successfully identify functional patterns of the cancers. To validate our results, for each cancer, driver coding genes in IntOGen database and differentially expressed mRNAs in GDC are enriched (p -value < 0.05 and $FDR < 0.05$). Functional similarities between driver genes and the results from our proposed method are evaluated. We further explore the functional role of the differentially expressed cis-acting super-enhancer lncRNAs that are known to regulate the close-by mRNAs. Moreover, we analyze the known metastasis interactions between cancers that have similar functional patterns to expose the indirect relationship between metastasis and lncRNA genes.

4.2 Related work

Several computational studies investigate the lncRNA functions, and only a few of the available methods consider the disease state. Guttman et al.[171] predict several mouse lincRNA functions by using chromatin-state maps and Li et al.[172] propose a method to find the functional similarity between lncRNAs by joining chromatin-states and gene expression data. The LncRNAdb[24] curates lncRNA functions from the literature. Linc2Go provides functional annotation data for lincRNA and also, supplies microRNA-mRNA and microRNA-lincRNA interactions to generate functional annotations of lincRNAs according to the competing endogenous RNA hypothesis[173]. The LncRNA2Function[25] utilizes expression correlation between lncRNA and protein coding genes and performs a hypergeometric test to identify lncRNA functions with the significantly enriched functional terms among the co-expressed protein-coding genes for the given lncRNAs.

Similarly, the lncRNAator identifies lncRNA functions by enriching co-expression patterns of genes and lncRNAs[26], and it primarily focuses on cancer. FARNA infers the functions of a transcript from the promoter-associated transcription factors and transcription co-factors whose genes are co-expressed with the transcript controlled by these TFs and TcoFs[27]. All of these studies focus on one lncRNA genes to reveal the lncRNA functionality. Only Zhao et al.[174] measure the combinatorial effects of the lncRNAs in a specific biological function for cancer and discover co-expressed protein coding genes of lncRNAs. However, none of these methods account for the importance of neighboring gene regulation in lncRNA. Moreover, it is known that cancers change gene functionality, but our knowledge on the lncRNA functionality in cancer is limited, and the effects of the set of lncRNAs are still missing.

4.3 Methods

The proposed methodology uses the property of lncRNA genes that regulate nearby genes. Thus our methodology is based on lncRNA annotation with nearby coding genes according to their gene centric regions. Figure 4.1 provides a graphical representation of mRNA centric regions and how lncRNA genes are annotated. From the retrieved mRNA genes, functional enrichment analysis is performed. The pipeline for the methodology is provided in Figure 4.2.

4.3.1 Data Sources

RNA-Seq gene expression data for 28 different cancer sites are collected from GDC Data Portal (July 15th 2017). Normal tissue samples are excluded from the analysis. Using the list of lncRNAs in human transcriptome obtained from GENCODE v26 [68], we subset the lncRNAs in TCGA RNASeq data. In total, 15,513 lncRNA genes are overlapped with GENCODE v26.

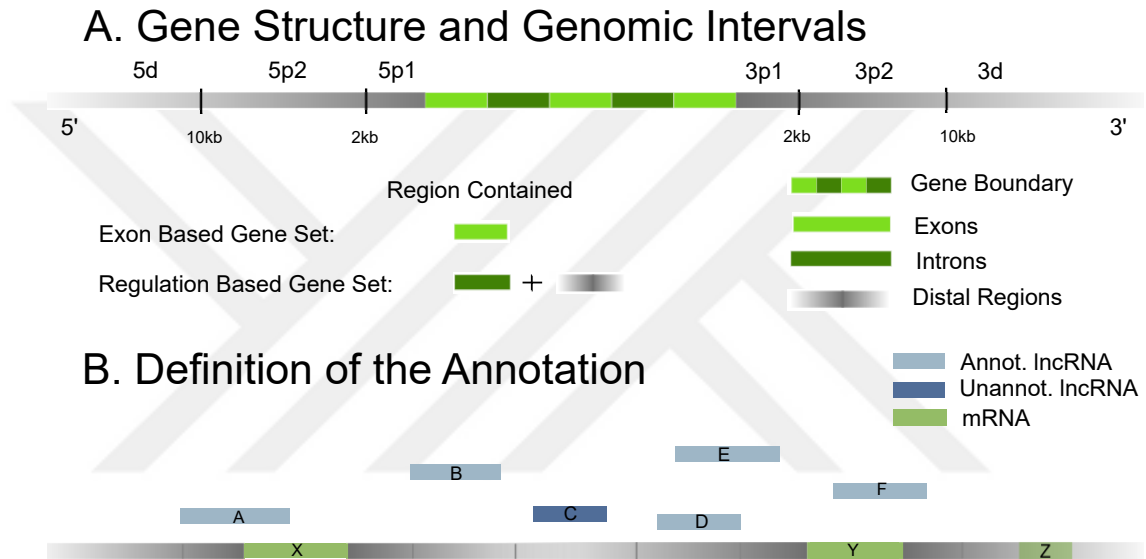


Figure 4.1: A. Illustration of the gene intervals and gene sets that are covered in this analysis. Distal regions are scaled down in order to fit to the manuscript. B. The logic of the annotations for the set of lncRNAs based on neighborhood genes. A lncRNA that falls into any of the genomic intervals of mRNA is annotated with the corresponding mRNA. For example, when a set of lncRNAs consisting of lncRNA A to F is analyzed, lncRNA A is annotated with mRNA X in the exon and regulation based gene set, lncRNA F is annotated in exon based gene set with mRNA Y and in regulation gene set with mRNA Y and Z. lncRNA B is only annotated with mRNA X in regulation based gene set. lncRNA D, E and F is annotated with Y in regulation based whereas lncRNA C falls into the 3D/5D genomic intervals, as a result, it is not annotated with any mRNA in regulation and exon based gene set. Thus, to enrich the exon based gene set for the given lncRNA set, mRNA X and Y is utilized. Similarly, regulation based gene set contains mRNA X, Y and Z.

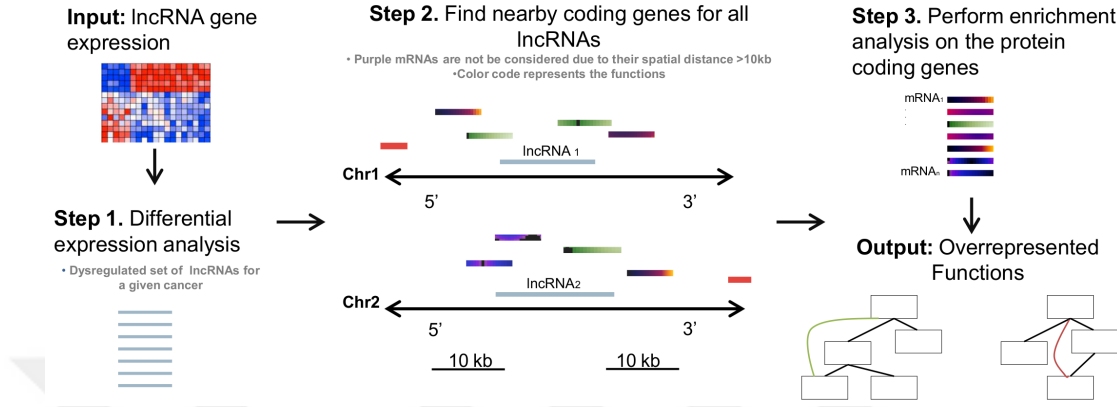


Figure 4.2: Overview of the methodology. Dysregulated lncRNAs are determined by a differential expression analysis (Step 1). Within the 10 kb range, both from upstream and downstream of the lncRNA gene, lncRNAs annotated with their neighborhood coding genes according to the genomic intervals of mRNA. The color codes of genes represent the functionality. The red color-coded genes cannot be annotated with any lncRNA since they are at least 10 kb far away from lncRNAs. On the neighborhood coding gene set, we perform enrichment analysis (Step 3). The same methodology is applied for 28 different cancers.

The metastasis information for each cancer type is collected from MSK-IMPACT Clinical Sequencing Cohort[175], (downloaded on April 10nd 2018), only metastatic samples are considered.

The cis-acting differentially expressed super-enhancer associated lncRNAs for different cancers are retrieved from the SELER database [176] (June 23th 2019).

4.3.2 Differential Expression Analysis

In our analyses, we use the fragments per kilo base of transcript per million mapped reads upper quartile normalization (FPKM-UQ) data. We filter out the lncRNA genes that have expression values less than 0.05 in more than %20 of the samples. Expression values are added with a constant 1 to deal with the 0 gene expression values and data are log 2 transformed. We eliminate RNAs that do not vary across samples which we filter out by using the median absolute deviation (MAD) Equation 4.1 leaving out those below 0.5.

$$\text{MAD}(r) = \text{median}(|r_i - \text{median}(\mathbf{r})|) \quad (4.1)$$

In order to identify the significantly dysregulated lncRNAs, we use linear model fitting followed by empirical Bayes t-test implemented in the Bioconductor limma package[177]. lncRNA genes are tested against the mean of expression of each gene across all samples to identify whether a given gene is statistically significantly expressed at $p\text{-value} < 0.05$.

4.3.3 Functional Enrichment Analysis of lncRNAs Based on Neighborhood Genes

To get a functional insight into each differentially expressed lncRNA set for the different cancer types, we conduct a functional enrichment analysis based on their neighborhood genes. For this, we use the GLANET method [170], which inputs the location of the genes of interest and performs sampling based on enrichment test to test for overlap with gene-sets. The gene-set is formed based on GO terms and KEGG pathways. In the case of GO-terms, a gene-set is formed for each GO term, and the set includes the genes annotated with the GO term. GLANET first forms genomic intervals from the input gene set and subsequently the overlap with the gene’s genomic intervals in these functional gene sets. The genomic regions can be defined over the exons of the genes in the functional set, over introns or based on proximal and distal regions as used in [178]. As gene-centric regulatory regions; exon, intron, 5p1, 5p2, 3p1, and 3p2 of mRNAs are employed. 5p1, 5p2 is defined as the regions 0 – 2 kb, 2 – 10 kb upstream of the first exon of the gene, respectively. Similarly, 3p1 and 3p2 as the regions 0 – 2 kb, 2 – 10 kb downstream of the last exon of the gene, respectively.

One or more common functional elements in the input gene set are enriched by measuring the statistical significance of overlaps with the genomic elements, which are lists of the annotated genomic regions from the literature. For quantifying overlaps, we use the number of overlapping bases which considered the actual

number of overlapping bases. The GO-terms which achieve significance with Benjamini-Hochberg correction with $FDR < 0.05$ are selected. The enrichment tests are conducted with 10,000 samplings. The random intervals are generated by matching the input's chromosome, length, mappability, and GC content of the input interval set.

4.3.4 GO Term Semantic Similarity Evaluation

Semantic similarity between pairwise cancers' GO-terms are evaluated to reveal how similar functions play their roles for the different cancer types. X is the cancer set and for any $x \in X$, genes in the x cancer are enriched in the set of GO-terms are represented with $GO_x = \{go_{1x}, go_{2x}, go_{3x}, \dots, go_{kx}\}$ and for any $y \in X$, genes in the y cancer are enriched in the set of GO-terms as with $GO_y = \{go_{1y}, go_{2y}, go_{3y}, \dots, go_{ly}\}$.

The similarity score between two GO-terms are calculated with graph-based Wang method[179], Equation 4.2, which depends on the position of these GO-terms in the DAG and their relations with their ancestor terms. Given a GO-term A , semantic value of GO-term A has a combinatorial semantic value of all terms in the DAG_A and GO-terms in the DAG_A closer to the A contribute the A semantics more.

$$sim_{Wang}(A, B) = \frac{\sum_{t \in T_A \cap T_B} S_A(t) + S_B(t)}{\sum_{t \in T_A} S_A(t) + \sum_{t \in T_B} S_B(t)} \quad (4.2)$$

where $S_A(t)$ is defined as contribution of a GO-term t to the semantic of GO-term A and $S_B(t)$ is defined as contribution of a GO-term t to the semantic of GO-term B .

Semantic similarity among the sets of GO-terms are combined with Best-Match average strategy (BMA) method which computes the average of all maximum

similarities on each row and column, and is defined as in Equation 4.3[179].

$$sim_{BMA}(GO_x, GO_y) = \frac{\sum_{i=1}^m \max_{1 \leq j \leq n} sim(go_{xi}, go_{yj})}{m+n} + \frac{\sum_{j=1}^n \max_{1 \leq i \leq m} sim(go_{xi}, go_{yj})}{m+n} \quad (4.3)$$

4.3.5 Clustering Cancers

Based on GO-term semantic similarities of the cancer pairs, we cluster the cancers using hierarchical clustering with average linkage method. Optimal cluster size is determined with Dunn index[180] which measures the compactness and separation of the clusters by calculating the ratio between the smallest distance between observations which are not in the same cluster to the largest intra-cluster distance.

4.4 Results

4.4.1 Overview of Gene Ontology Enrichment Analysis of the lncRNAs

In order to investigate a set of lncRNAs that may participate in common functions, we conduct the following analysis summarized in Figure 4.2 and we infer the functions of lncRNAs by utilizing the ability of lncRNA locus to regulate nearby genes for different cancer types. We carry out our analysis for exon based and regulation based gene sets where exon based gene set consists of genes that are in the exon region and where regulation based gene set contains introns and four different gene-centric regulatory regions, namely 5p1, 5p2, 3p1, and 3p2, of

genes. The purpose of using different gene set is to understand the spatial effects and variations imposed by different gene-centric regulatory regions.

We perform differential expression analysis to identify dysregulated lncRNAs in cancer. The numbers of lncRNAs for each cancer which are curated after differential expression analysis among 15.513 lncRNAs are provided in Figure 4.3 and Figure 4.3 illustrates the distribution of the differentially expressed lncRNAs in each cancer (p -value < 0.05). Although lncRNAs have tissue-specific expression compared to mRNAs [166], we find that differentially expressed 1.186 lncRNAs are common in all cancer types. This may indicate that some biological mechanisms may play roles in all cancers.

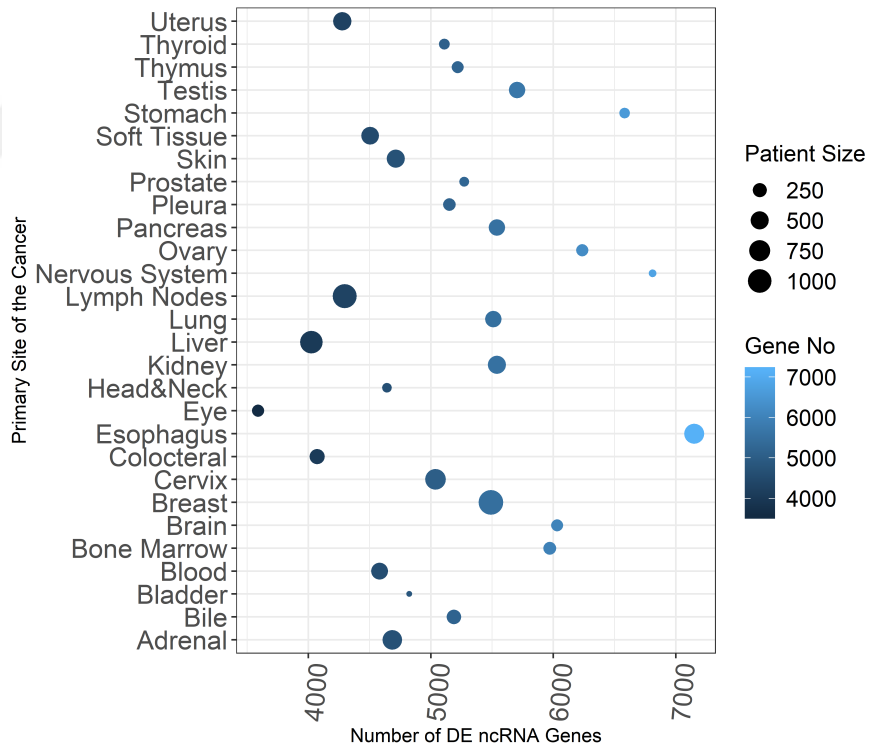


Figure 4.3: Dot plot for the number of tumor patients and the number of differentially expressed lncRNAs in each cancer. Color code of the dots represents the number of differentially expressed lncRNA genes and the size of dots is the number of tumor patients for the cancer.

We find that some enriched GO-terms are common in all cancers and among these GO-terms, some of them are related to the signaling, which is associated

with the cancer[181]. Moreover, lncRNAs that fall into the exon regions of the mRNAs are mostly enriched with the regulation, process, and biosynthetic biological process GO-terms. However, for lncRNAs that fall into the regulator regions of the mRNAs, most of the biological process GO-terms are related to regulation, morphogenesis, development, metanephric in all of the cancers except nervous, ovarian, pancreas, esophagus and stomach. Most of the GO-terms are enriched with long-chain fatty acid, secretion, and regulation related to ovarian cancer, secretion associated GO-terms for pancreas and esophagus cancer, cellular and acid associated GO-terms for nervous cancer.

4.4.2 Identifying Functional Relevance between lncRNAs and mRNAs

In this analysis, we aim to examine the functional relevance between a set of dysregulated lncRNAs and a set of dysregulated mRNAs for each cancer and state whether they participate in similar functions. For this purpose, we apply the same pipeline for the protein coding genes of each cancer. mRNAs are pre-processed and differential expression analysis is applied for $p\text{-value} < 0.05$ and $FDR < 0.05$. Curated mRNA set for each cancer is annotated with GO-terms, and GO-term enrichment analysis is performed. We compare enriched GO-terms of the differentially expressed mRNA gene set and lncRNA gene set[179]. The similarities between the functionality of the mRNA and lncRNA for each cancer are listed in Table 4.1 and Table 4.1 shows that lncRNAs promote cancer by regulating the neighborhood genes.

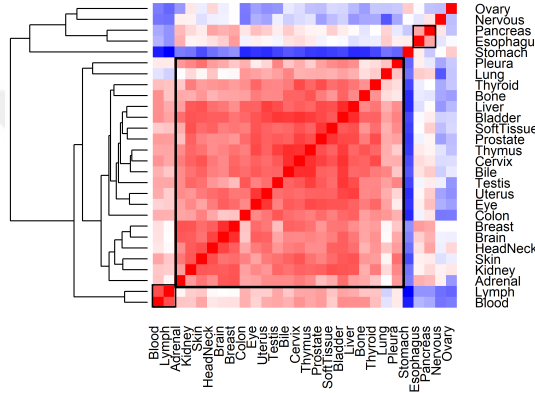
4.4.3 Functional Similarity Analysis Between Cancers

In order to explore functionally related cancers, we compute a semantic similarity score between cancer pairs[179] and determine similar functional patterns of the cancers with hierarchical clustering. We observe that similar mechanisms play roles in some cancers when functions of lncRNAs derived from its neighborhood,

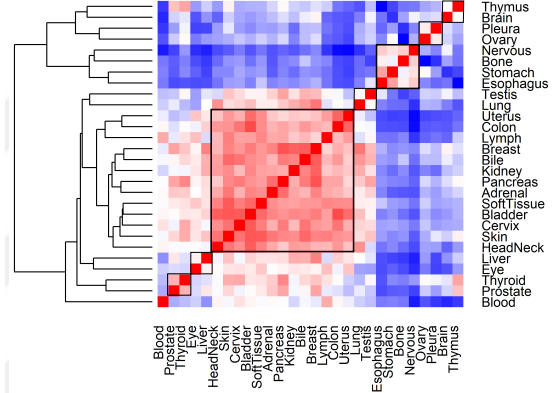
Table 4.1: Enriched GO-term comparison for the differentially expressed mRNA and lncRNA gene set

Primary Site	Comparison Score
Adrenal	0.94
Bile	0.93
Bladder	0.91
Bone	0.87
Brain	0.91
Breast	0.94
Cervix	0.91
Colon	0.96
Esophagus	0.93
Eye	0.89
HeadNeck	0.90
Kidney	0.89
Liver	0.96
Lung	0.90
Lymph	0.88
Nervous	0.93
Ovarian	0.90
Pancreas	0.92
Pleura	0.91
Prostate	0.91
Skin	0.92
SoftTissue	0.91
Stomach	0.90
Testis	0.87
Thymus	0.94
Thyroid	0.88
Uterus	0.92

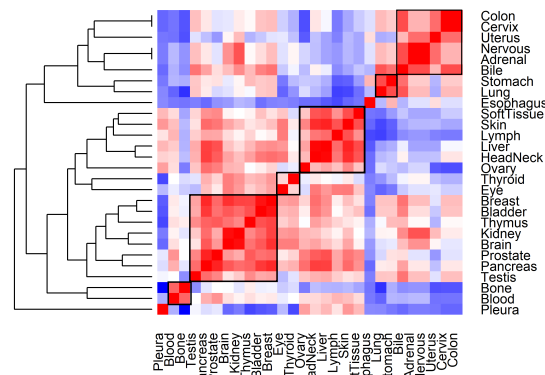
A. Regulation Based BP



B. Exon Based BP



C. Regulation Based MF



D. Exon Based MF

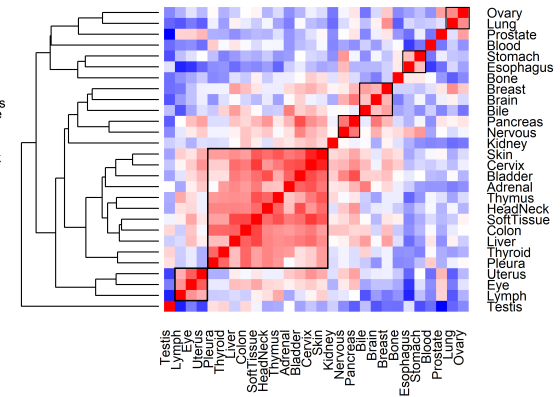


Figure 4.4: Functional similarity scores between each cancer pairs are represented in the heatmap. Biological processes (A-B) are demonstrated in the first row and molecular functions (C-D) are in the second row. Colors are scaled between blue (low) to red (high) and white color represents the median value. Color key for regulation based heatmap is $[0.2-1]$ and for exon based is $[0.4-1]$. Important clusters are highlighted with black square. Row and column labels describe the primary site of the cancers.

Figure 4.4. In Figure 4.4, only cancer cluster sizes greater than 2 are demonstrated with black squares, and the optimal cluster size is determined with Dunn index (see Section 4.3).

We observe that the similarity between two cancers is less in the regulation based gene set than exon based gene set, Figure 4.3 (A-C). This may be due to the observation that regulation based gene set contains distant gene centric regions such as 5p2 and 3p2, and lncRNA is correlated with the nearby genes[165]. This finding may show that lncRNAs have less effect on the regulation of distant genes.

4.4.3.1 Functional Similarity Between Acute Myeloid Leukemia and Diffuse Large B-cell Lymphoma in Regulation Based Biological Process

We detect that similar functions play a role in Diffuse Large B-cell Lymphoma (DLBL) and Acute Myeloid Leukemia (AML). Diffuse Large B-cell Lymphoma is the most common subtype of non-Hodgkin's lymphoma that affects B-cells, one type of white blood cell responsible for producing antibodies[182]. Acute Myeloid Leukemia is abnormally increased in a number of myeloid line of blood cells in the bone marrow.

We observe that differentially expressed lncRNAs of AML and DLBL that fall in the regulatory gene-centric regions of mRNAs are enriched with similar functions, Figure 4.4 (A). In Figure 4.4 (A), primary site blood represents AML, and primary site lymph represents DLBL.

Moreover, some GO-terms are only enriched in both AML and DLBL, which are given in Table 4.2. K63-linked ubiquitination is reported to be linked to the AML and DLBL[183]. Biosynthetic processes allow cells to develop macromolecules to cause abnormal cell division and tumor growth, which is essential for cancer[184]. It is known that biosynthetic processes inhibit AML and DLBL, but the specific mechanisms are not clear yet[185].

Table 4.2: List of the GO-terms that are only enriched in both Acute Myeloid Leukemia and Diffuse Large B-cell Lymphoma. Last column shows the maximum *FDR* value that AML and DLBL take for the corresponding GO-terms.

GO-ID	GO-term	FDR
GO:0046167	glycerol-3-phosphate biosynthetic process	0.0000
GO:0006071	glycerol metabolic process	0.0277
GO:0010870	positive regulation of receptor biosynthetic process	0.0325
GO:1902523	positive regulation of protein K63-linked ubiquitination	0.0330
GO:0032817	regulation of natural killer cell proliferation	0.0339
GO:1900165	negative regulation of interleukin-6 secretion	0.0427
GO:0035425	autocrine signaling	0.0448
GO:0034145	positive regulation of toll-like receptor 4 signaling pathway	0.0448
GO:0032481	positive regulation of type I interferon production	0.0487

4.4.3.2 Functional Similarity Between Stomach Adenocarcinoma and Esophageal Carcinoma in Exon Based Molecular Function

Stomach adenocarcinoma also called gastric cancer, is a general term of the tumor in the stomach. Esophageal carcinoma develops in the glands of the esophagus, the fibromuscular tube that connects the throat to the stomach.

We discover that differentially expressed lncRNAs are enriched with the GO-term acetyltransferase activator activity (GO:0010698) only for stomach and esophageal carcinoma. Previous studies[186, 187, 188] provide evidence that N-Acetyltransferase 1 and N-Acetyltransferase 2 gene activities increase the stomach and esophageal cancer risk. Thus, differentially expressed lncRNAs may activate the regulation of the neighbour genes and promote the stomach and esophageal cancer.

Moreover, fibroblast growth factor binding (GO:0017134) and C-X3-C chemokine binding (GO:0019960) GO-terms are detected which are important for both stomach adenocarcinoma and esophageal carcinoma [189, 190].

4.4.4 Exploring the Metastasis in Cancers Based on a Functional Evidence

We analyze the functional clusters to explore whether they promote metastasis when lncRNAs regulate close-by coding genes. Primary and metastatic sites of each patient are altered to the generic sites that are used to describe the primary sites of the cancers in TCGA, in Figure 4.4, such that, primary site “Transverse Colon” in MSK-IMPACT study is converted to the “Colon”. Figure 4.5 is a graphically represents which primary sites are causing which metastasis and the number of occurrences of each metastasis.

4.4.4.1 Metastasis from Uveal Melanoma and Colon Cancer: Liver

Uveal melanoma is the most common ocular malignancy which occurs in the iris, ciliary body or choroid[191]. Almost, 50% patients will produce metastasis within 6-12 months survival[191]. Most of the metastasis are formed in liver[192].

Our findings of biological process in exon based gene set, Figure 4.4 (B), shows that differentially expressed lncRNAs, whose primary sites are on the eye, overlaps with the exons of mRNAs may participate similar functional mechanisms and promote liver cancer. In MSK-IMPACT, 21 cancer patients whose primary site is eye have liver metastasis, Figure 4.5.

Moreover, our result is confirmed with Human Cancer Metastasis Database (HCMDB)[193]. HCMDB presents published cancer metastasis expression profiles from GEO, SRA, TCGA, and analyses mRNA-lncRNA co-expression profiles and miRNA regulatory networks. Experiment EXP00455 in HCMDB for uveal melanoma in TCGA, which is metastatic to liver shows that differentially expressed genes are functionally enriched (18953 mRNA; 3210 lncRNA; 2588 miRNA).

Colon cancer is the third leading cause of the cancer-related deaths in 2018[194]. The liver is the most common site of the colon cancer metastasis,

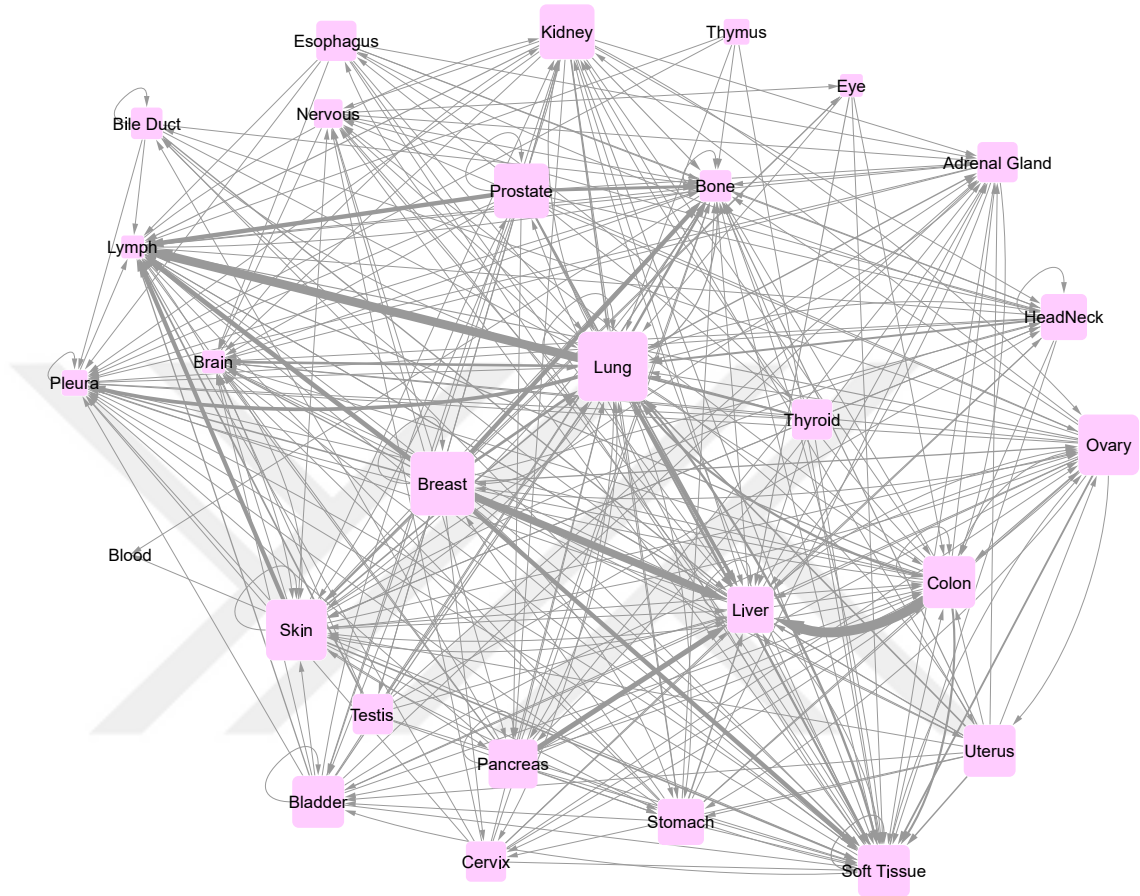


Figure 4.5: Representation of metastasis from the primary site to the metastatic site. Each node corresponds to a site. The size of a node depends on the out-degree, and the number of outward directed edges from the node. Weights of edges are occurrences of the related metastasis for the nodes, and the directions of the edges are from the primary sites to metastatic sites.

more than 50% of the colon cancers[195]. The most frequently seen metastatic couple in MSK-IMPACT is colon-liver and 297 patient whose primary site is colon cancer suffer from liver metastasis, Figure 4.5. We observe these metastatic pairs in regulation based biological processes and exon based molecular function.

4.4.4.2 Metastasis Analysis in the Big Clusters

Primary and metastasis cancer pairs in the same clusters are tested with MSK-IMPACT Clinical Sequencing Cohort data to assess whether metastasis in the

clusters are significant for the given biological mechanism. V is the cancer set, for any cancer $v_i \in V$, C is a set of cancers that are in a given cluster, $C = \{c_i \in C \wedge 1 \leq i \leq 28 \wedge C \subset V\}$, NC is a set of cancer that are not in a given cluster, $NC \subset V$, $NC = V \setminus C$. We applied Wilcoxon rank-sum test between $M = \{v_i v_j \mid v_i \in C \wedge v_j \in C \wedge 1 \leq i, j \leq 28\}$ and $NM = \{v_i v_k \mid v_i \in C \wedge v_k \in NC \wedge 1 \leq i, k \leq 28\}$ sets under the alternative hypothesis that M is greater than NM . Since the size of the two sets, M and NM , are not equal, we randomly pick $\min(|M|, |NM|)$ number of elements from the set that element size is $\max(|M|, |NM|)$ and permute the Wilcoxon rank-sum test 1.000 times.

It is observed that regulation based molecular function sets that contain $c_1 = \{\text{testis, pancreas, prostate, brain, kidney, thymus, bladder, breast}\}$ and $c_2 = \{\text{ovary, head-neck, liver, lymph, skin, soft tissue}\}$ make statistically significant metastatic couples. Moreover, cancers in the big cluster of the regulation based biological process also constitute significant metastatic pairs.

4.4.5 Inferring Functional mechanisms for the Super-enhancer Associated lncRNAs

Enhancers are cis-regulatory DNA elements that increase the transcription of a gene. Super-enhancers are groups of putative enhancers that are spatially proximal to the mediator binding sites[196]. They regulate gene expression by increasing the gene transcription[176]. Due to the disease related variations which occur in the super-enhancer regions for a disease specific cell-lines, super-enhancers are highly related to the diseases including cancer and can be used as biomarkers[197]. The recent studies show cis-acting super enhancer lncRNAs are transcribed from the super enhancers and regulate the nearby genes. Thus, understanding the functional roles of the super-enhancer associated lncRNAs is important.

In this analysis, we collect the differentially expressed cis-acting lncRNAs in SELER database, and we only consider the genes that more than %20 of the samples expressed higher than the 0.05 expression value and varied across samples

(See Methods). The final set contains 65 lncRNA for bladder cancer, 49 lncRNA for breast cancer, 179 lncRNA for kidney cancer, 97 lncRNA for lung cancer, 47 for prostate cancer, 42 and 50 lncRNAs for stomach and thyroid cancer, respectively.

We apply the same methodology for the cis-acting super-enhancer lncRNAs. We identify the set of coding genes that are proximal to the cis-acting super-enhancer lncRNAs, and we perform gene ontology and pathway enrichment analysis on the coding genes.

We observe that the enriched go terms in all cancers are regulation related GO-terms such as positive regulation of telomerase activity, negative regulation of transcription by competitive promoter binding. Interestingly, enriched functions for each cancer, excluding stomach cancer, comprise the cell death-related GO-terms such as negative/positive regulation of programmed cell death, autophagic cell death, engulfment of target by autophagosome, cell death, and other apoptosis-related GO-terms.

4.4.6 Overview of the Pathway Analysis

Pathway enrichment analysis for lncRNA is conducted with KEGG pathways under GLANET.

In regulation based gene set, only prostate cancer is enriched with pancreatic secretion (hsa04972, $FDR = 0.026$) pathway. Previous studies do not report links to this pathway with prostate cancer. However, serine peptidase inhibitor Kazal Type 1, SPINK1 gene, which is not in the pancreatic secretion pathway, is secreted from pancreatic acinar cells into pancreatic juice. Its role is to prevent trypsin-catalyzed premature activation of zymogens within the pancreas and the pancreatic duct[198], and it is associated with prostate cancer[199]. Thus, prostate cancer and pancreatic secretion (hsa04972) pathway are highly related, and lncRNAs of prostate cancer promote the transcription or splicing in tumorigenesis.

In the exon based gene set, phenylalanine metabolism (hsa00360, $FDR < 0.05$) pathway is enriched in prostate, liver, breast, brain, and bile duct cancer. Phenylalanine is the essential amino acid that is naturally found in breast milk and also cheese, eggs, and meat. It is utilized in the dopamine and norepinephrine neurotransmitters. Thus, it is intensively presented at the human brain and plasma. Phenylalanine is the source of the artificial sweetener aspartame[200] which is claimed to caused brain cancer[201, 202]. However, we do not find a relation between phenylalanine metabolism pathway and the other enriched cancers.

4.4.7 Validation With Cancer Driver Genes

To validate our enrichment analysis on lncRNAs, we employ cancer driver genes which are known to involve in tumorigenesis. Cancer driver genes are enriched, and we compare the enrichment results of cancer driver genes with our findings. The motivation of this analysis is to prove that the dysregulated lncRNAs for a given cancer are involved in the same functional mechanisms as the driver genes of the given cancer.

For this analysis, oncogenes of the IntOGen[139] database, which identifies cancer driver mutations, genes, and pathways across tumor types, are utilized. Driver genes of each cancer are collected on March 23rd 2018 and enrichment tests for each cancer are carried out with clusterProfiler[82] with Benjamini-Hochberg multiple hypothesis test corrections. In each test, p -value and FDR cut-offs are set to 0.05.

The similarity score is evaluated between GO-terms of the enrichment analysis of the driver genes and GO-terms of the enrichment analysis of the lncRNAs (see *Materials and Methods* section). For similarity score calculation, we used R package GOSemSim[179] with Wang method combined Best-Match Average strategy (BMA).

The permutation test with 1,000 permutations is used in the analysis, and

Table 4.3: Comparison between driver genes and our results for the corresponding cancers. In this table, each field represents the GO-term similarity score between driver genes and our results. Null test columns show the ratio of the alternative hypothesis for the GO-term similarity scores of the random selected driver genes greater than the actual GO-term similarity score for the actual driver genes for the corresponding cancer.

	Exon BP	Reg. BP	Null Exon BP	Null Reg. BP
Adrenal	0.392	0.391	0.007	0.007
Breast	0.410	0.353	0.060	0.104
Bladder	0.440	0.400	0.020	0.025
Blood	0.450	0.449	0.025	0.027
Bone	0.425	0.390	0.025	0.024
Brain	0.356	0.340	0.030	0.049
Colon	0.433	0.396	0.024	0.018
Esophagus	0.460	0.340	0.022	0.025
HeadNeck	0.440	0.345	0.035	0.082
Kidney	0.426	0.362	0.029	0.037
Liver	0.416	0.422	0.013	0.011
Lung	0.419	0.335	0.240	0.756
Ovary	0.363	0.231	0.026	0.028
Pancreas	0.363	0.291	0.011	0.012
Prostate	0.400	0.448	0.018	0.011
Skin	0.405	0.337	0.791	0.849
Stomach	0.444	0.188	0.050	0.137
Thyroid	0.451	0.418	0.011	0.009
Uterus	0.426	0.393	0.045	0.102

for each permutation, GO-terms enrichment analysis uses a threshold at p -value < 0.05 and $FDR < 0.05$ and similarity score is measured with Wang method combined with BMA. Permutation test result for each cancer is provided in Table 4.3.

Table 4.4: Comparison between driver genes and our results for the corresponding cancer. In this table, each field represents the GO-term similarity score between driver genes and our results. Null test columns show the ratio of the alternative hypothesis for the GO-term similarity scores of the random selected driver genes greater than the actual GO-term similarity score for the actual driver genes for the corresponding cancer.

	Exon MF	Reg. MF	Null Exon MF	Null Reg. MF
Adrenal	0.414	0.214	0.015	0.018
Breast	0.404	0.274	0.040	0.061
Bladder	0.432	0.284	0.034	0.020
Blood	0.401	0.485	0.025	0.022
Bone	0.380	0.370	0.020	0.021
Brain	0.444	0.238	0.044	0.053
Colon	0.369	0.177	0.017	0.042
Esophagus	0.360	0.306	0.026	0.034
HeadNeck	0.384	0.348	0.056	0.052
Kidney	0.378	0.254	0.028	0.040
Liver	0.416	0.335	0.022	0.020
Lung	0.408	0.167	0.047	0.249
Ovary	0.443	0.301	0.021	0.021
Pancreas	0.463	0.334	0.011	0.009
Prostate	0.526	0.442	0.018	0.019
Skin	0.404	0.353	0.117	0.376
Stomach	0.492	0.343	0.054	0.069
Thyroid	0.449	0.272	0.016	0.011
Uterus	0.442	0.203	0.041	0.081

Table 4.5: Regulation based BP enrichment results for gene set HOTAIR, MEG3, H19

GO ID	GO Term
GO:0032274	gonadotropin secretion
GO:0034337	RNA folding
GO:2000279	negative regulation of DNA biosynthetic process
GO:0008285	negative regulation of cell proliferation
GO:0045893	positive regulation of transcription, DNA-dependent
GO:0045892	negative regulation of transcription, DNA-dependent

Table 4.6: Exon based BP enrichment results for gene set HOTAIR, MEG3, H19

GO ID	GO Term
GO:0032274	gonadotropin secretion
GO:0034337	RNA folding
GO:2000279	negative regulation of DNA biosynthetic process
GO:0008285	negative regulation of cell proliferation
GO:0045892	negative regulation of transcription, DNA-dependent
GO:0045893	positive regulation of transcription, DNA-dependent

4.4.8 Comparison Between Our Proposed Method and the Other Methods

4.4.8.1 Comparison for the Gene Ontology Enrichment Results

We compare our results with the methods discussed previously. As a case study, we use HOTAIR, H19 and MEG3 lncRNAs which are experimentally proved to be associated with the tumorigenesis in breast cancer[203, 204]. For this analysis, we only enrich differentially expressed HOTAIR, H19, MEG3 lncRNAs with GLANET[170] and compare the results with other methods. The list of common GO-terms that enriched for the selected lncRNAs are given in Table 4.6, Table 4.8, Table 4.5, Table 4.7.

Table 4.7: Regulation based MF enrichment results for gene set HOTAIR, MEG3, H19

GO ID	GO Term
GO:0003713	transcription coactivator activity
GO:0001077	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription
GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding

Table 4.8: Exon based MF enrichment results for gene set HOTAIR, MEG3, H19

GO ID	GO Term
GO:0003713	transcription coactivator activity
GO:0001077	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription
GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding

We can only compare the results with the Co-LncRNA and Farna[174, 27] since Linc2GO[173] and LncRNA2Function[25] are no longer accessible. Moreover, lncRNAator[26] doesn't support breast cancer data.

FARNA allows to input only single lncRNA. Thus, the enrichment results are compared on an individual basis. Mutually negative regulation of cell proliferation (GO:0008285) GO-term is enriched for H19 and HOTAIR. However, mutual enrichment result for lncRNA MEG3 is not observed.

Co-LncRNA investigates combinatorial effects of the lncRNAs. Therefore, it permits to input maximum 3 lncRNAs for functional enrichment analysis. For enrichment analysis, we use Linear Regression and all p -value cut-offs set to 0.05. Negative regulation of cell proliferation (GO:0008285), positive regulation of transcription, DNA-dependent (GO:0045893), negative regulation of transcription, DNA-dependent (GO:0045892) biological process GO-terms are detected in both Co-LncRNA and our proposed method. Moreover, transcription co-activator activity and RNA polymerase II core promoter proximal region sequence-specific DNA binding GO-terms are found mutually for molecular function.

4.4.8.2 Comparison for the Pathway Enrichment Results

To show the robustness of our methodology, results are compared with LncRNAs2Pathways[205]. LncRNAs2Pathways is an R-based tool under a Lnc-Path package name. It identifies the pathways affected by a set of lncRNAs based on a Kolmogorov-Smirnov like statistical measure weighted by the network propagation score. Package considers the expression correlation among lncRNAs and coding RNA genes.

In order to perform LncRNAs2Pathways analysis, we build the coding-non-coding gene correlation (CNC) network where nodes are protein-coding and non-coding genes, and edges are interactions between co-expressed protein coding and non-coding gene pairs. To construct the network, we apply the same preprocessing step for expression of the lncRNA and mRNA genes explained in the Methods

section. Pearson correlation coefficient is calculated for each lncRNA-mRNA pair and co-expression patterns between lncRNA and mRNA is detected if their coefficient value ranked in the top or bottom of the %0.1 and for p -value < 0.01 . Differentially expressed lncRNAs which were collected priorly is mapped to the CNC.

In LncRNAs2Pathways analysis, we use brain cancer as a case study. Given the CNC network, LncRNAs2Pathways[205] finds that Ribosome pathway is enriched in brain cancer ($FDR < 0.05$). We do not find a direct relationship between the brain and ribosome even though our proposed method finds related pathways.

4.5 Conclusion

We present a new methodology which systematically analyzes different cancers to explore the functional mechanisms that lncRNAs participate in those by regulating the nearby genes. Our contribution to the literature is threefold. We utilize a set of interesting lncRNA to explore the substantial functions that lncRNAs may contribute to the formation by its ability to regulate nearby genes. Moreover, we also contribute to the knowledgebase in elucidating the functional relationship between cancers when the neighborhood of the lncRNAs are considered. Finally, we make use of the combinatorial lncRNAs to distinguish the important functions.

Our analysis bias to GO-term annotations. However, some genes are experimentally and/or computationally analyzed more and they have more robust GO-term annotations than other genes, as well. Furthermore, newly found genes have no recognized GO-term annotations. Thus, the developments in GO-term annotations will enhance our analysis and render it more robust.

As a validation, we use driver genes for each cancer and perform enrichment analysis to measure the strength of our method. However; statistical tests may contain false positives which limits our validation power. Yet, experimentally validated GO-term annotations for lncRNAs are not available and providing such

annotations will make a high impact on the current understanding of functional mechanisms in cancers.

4.6 NoRCE R Package

In this work, we introduce an R package, NoRCE, which performs functional enrichment of a set of ncRNAs by transferring functional annotation available from the coding genes nearby. Although not all functionally related genes are located nearby on the genome, previous work has shown that functionally related genes show substantial spatial concentration within chromosomes and across chromosomes [206]. There is also evidence suggesting that both ncRNAs and coding genes can influence the expression of neighboring genomics loci [165].

For each ncRNA in the input set, NoRCE first determines the set of coding genes located within a user-defined distance range (in base-pairs) of this ncRNA and use the union of annotations of all coding genes assigned to the ncRNA set to perform enrichment. NoRCE provides options to perform enrichment analysis by using Fisher’s exact, hypergeometric, binomial, or chi-squared tests.

Several useful options are available for the curation of the input ncRNA set. Users can filter ncRNAs with respect to their biotypes (sense, antisense, lincRNA, etc.). The input set’s neighboring coding gene set can be filtered or expanded with coding genes found to be co-expressed with the input ncRNAs. The co-expression analysis can be performed by using the data available in NoRCE repository, such as The Cancer Genome Atlas (TCGA) project expression data or by using custom expression data. Since the interruption in TAD boundaries affects the expression of neighboring genes, filtering option with TAD boundaries allows the user to limit the genes to be enriched to a set that falls into TAD regions[207]. For miRNA specific input, the neighbor coding set can be filtered or expanded with predicted targets of the input miRNAs or with co-expressed genes. Finally, the outcome of the above analysis can be summarized in a tabular format or visualized using relevant graphs, including KEGG pathway map, Reactome diagram, and directed

acyclic graph (DAG) of the enriched GO-term (See Figure 4.6 for the detailed workflow of the package). The analysis is available for several species including human, mouse, rat, zebrafish, fruit fly, worm and yeast (Details of the assemblies supported is provided in Table 4.9). NoRCE R package is platform independent, available at <https://github.com/guldenolgun/NoRCE>

Table 4.9: The supported assemblies for different species in NoRCE.

Species	Supported Assembly	
	UCSC	NCBI
<i>Homo sapiens</i>	hg19	GRCh37, GCA_000001405.1, Feb. 2009
<i>Homo sapiens</i>	hg38	GRCh38, GCA_000001405.15, Dec. 2013
<i>Mus musculus</i>	mm10	GRCm38.p6, INSDC Assembly GCA_000001635.8, Jan 2012
<i>Rattus norvegicus</i> (brown rat)	rn6	Rnor_6.0, INSDC Assembly GCA_000001895.4, Jul 2014
<i>Drosophila melanogaster</i> (fruit fly)	dm6	BDGP6, INSDC Assembly GCA_000001215.4, Jul 2014
<i>Danio rerio</i> (zebrafish)	danRer10	GRCz10, GCA_000002035.3, Sep. 2014
<i>Caenorhabditis elegans</i> (worm)	ce11	WBcel235, GCA_000002985.3, Feb. 2013
<i>Saccharomyces cerevisiae</i> (yeast)	sacCer3	UCSC version sacCer3, 2011

4.7 Implementation and Features of the NoRCE

The workflow for the NoRCE is provided in Figure 4.6. NoRCE accepts a set of ncRNAs, $S = \{r_1, \dots, r_n\}$. If the user wants to conduct a biotype specific analysis (e.g., lincRNAs), NoRCE can select the elements in S with the selected

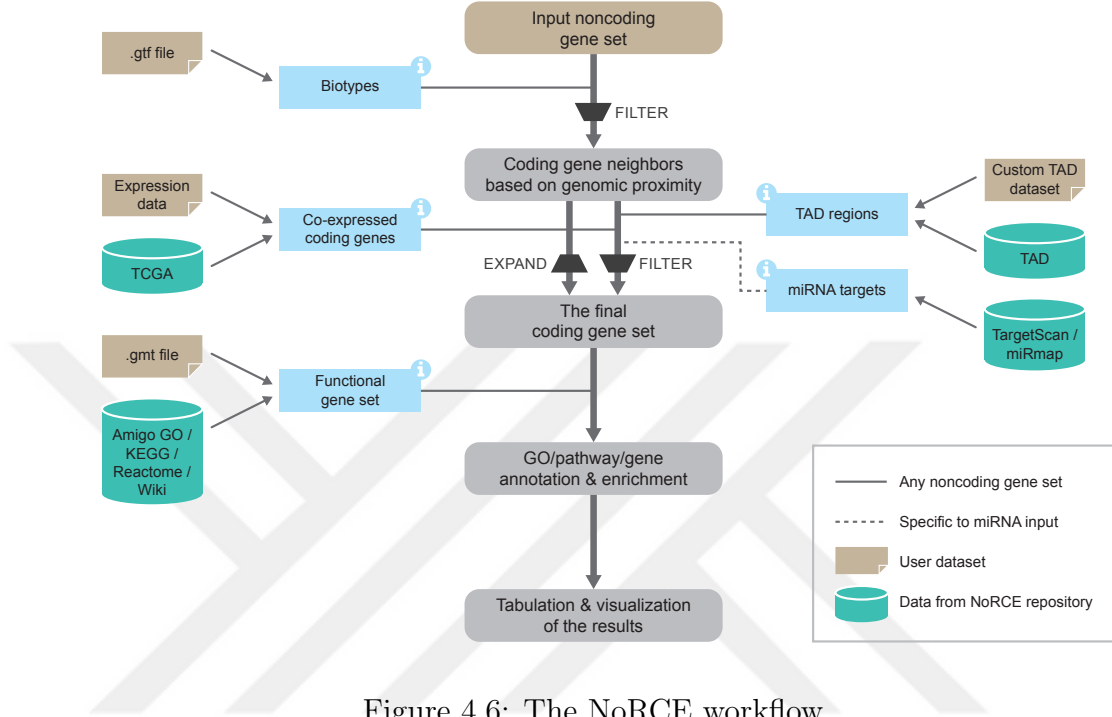


Figure 4.6: The NoRCE workflow

biotype and use this subset in the subsequent steps. For each ncRNA in the input list, NoRCE identifies the nearby protein-coding genes that are located upstream and downstream in a given user-specified base pair limit. Thus, for each $r_i \in S$, NoRCE creates a coding gene list of C_i and conducts the analysis with the union of these $C = \cup_{i=1}^n C_i$. This set of coding genes, C , can be further filtered or expanded based on additional biological evidence:

Co-expressed coding genes: If the filtering option is set, for every input ncRNA r_i , only the nearby genes in C_i that are co-expressed with r_i are placed into the final coding gene set C . If the expansion option is set, a coding gene is co-expressed with any of the $r_i \in S$, and each coding gene is included in C . For cancer analysis, we provide pre-computed results based on expression analysis of the patient data obtained from The Cancer Genome Atlas (TCGA) portal [208]. Users can also load custom expression data and perform correlation based expression analysis.

TAD regions: Chromosomes are organized into distinct, evolutionarily conserved, cell-type invariant secondary structures called TADs. TADs are known

to have functional importance since the subsets of genes within a TAD are co-regulated and some gene clusters, such as cytochrome genes and olfactory receptors are organized into individual TADs [209]. Since the interruption in TAD boundaries affects the expression of neighboring genes [207], we allow a filtering option with the TAD regions. If this option is on, enrichment is conducted only with ncRNAs that fall into a TAD region. Also, in the coding gene assignment, only those that are in the same TAD region are included in the neighborhood coding gene set. NoRCE provides several bed formatted cell-line specific TAD regions in its data repository, Table 4.10. Alternatively, the user can input a custom TAD file.

Table 4.10: Topological associating domain data that included in the NoRCE repository

Species	Name	Vers.	# of cell lines	# of TAD regions	Source
<i>Homo sapiens</i>	tad_hg19	hg19	37	74,424	3D Genome Browser[210]
<i>Homo sapiens</i>	tad_hg38	hg38	42	96,526	3D Genome Browser[210]
<i>Mus musculus</i>	tad_mm10	mm10	5	16,866	3D Genome Browser [210]
<i>D. melanogaster</i>	tad_dmel	dm6	1	2,846	HiCBrowser [211]

miRNA target information: If the user list is a set of miRNAs, the neighborhood of miRNA genes can be subsetting with the potential miRNA targets. For this step, NoRCE utilizes several available target prediction algorithms. We obtain computationally predicted conserved miRNA-target interactions from the TargetScan [212] for the species except rat. Rat data was obtained from the miRmap [213]. More information is provided in Table 4.11.

4.7.1 Enrichment Analysis of the Coding Genes

Once the coding gene list, C , is curated, NoRCE can conduct enrichment analysis based on gene sets derived GO-terms, KEGG, Reactome, WikiPathways, coding genes, or integrated pathway dataset. We retrieve the gene ontologies for each species from the EMBL-EBI GOA database(<https://www.ebi.ac.uk/GOA/>

Table 4.11: List of miRNA target prediction algorithms and the corresponding number of target pairs for each species. To the best of our knowledge, no miRNA target predictions have been reported for *Saccharomyces cerevisiae*.

Species	miRNA-Target Prediction Database	Version/Date	No of Pairs	Reference
<i>Homo sapiens</i>	TargetScan	v7.2	235,128	[212]
<i>Mus musculus</i>	TargetScan	v7.1	179,430	[214]
<i>Rattus norvegicus</i>	miRmap	2013	2,258,219	[213]
<i>Drosophila melanogaster</i>	TargetScan	v6.2	15,641	[215]
<i>Danio rerio</i>	TargetScan	v6.2	1,147,028	[214]
<i>Caenorhabditis elegans</i>	TargetScan	v6.2	208,146	[216]

downloads) and the GO slim annotations from the Gene Ontology (GO) Consortium [217]. Table 4.12 reports the size of gene ontologies available for each species in NoRCE. Users can input custom gene sets as well. Enrichment is performed according to the standard candidate versus background sets using Fisher’s exact, hypergeometric, binomial and chi-squared tests. Also, the available multi-test hypothesis correction methods are Holm, Hochberg, Hommel, Benjamini & Hochberg, and Benjamini & Yekutieli and Bonferroni correction.

Table 4.12: The size of gene ontologies available for each species in NoRCE

	Biological Process	Cellular Component	Molecular Function
<i>Homo sapiens</i>	137,575	80,584	66,106
<i>Mus musculus</i>	161,638	103,261	91,633
<i>Rattus norvegicus</i>	205,265	107,124	87,949
<i>Drosophila melanogaster</i>	50,525	27,156	27,894
<i>Danio rerio</i>	64,866	46,429	55,383
<i>Caenorhabditis elegans</i>	36,699	30,189	32,716
<i>Saccharomyces cerevisiae</i>	35,136	33,417	26,526

4.7.2 Visualization of the Results

Coding genes found in enriched pathways can be visualized within the context of KEGG or Reactome maps marked with color. Moreover, GO enrichment results

can be visualized with directed acyclic graph, dot plot, where p -value sorted list of enriched GO terms with gene count, and GO-term:coding or GO-term:noncoding gene network.

4.8 Example Use Cases of NoRCE

We showcase NoRCE with two use cases related to the brain disease to demonstrate how NoRCE could be used to gain insight into the functional importance of a list of interesting ncRNAs. As demonstrated in Case Study 1, a list of ncRNAs differentially expressed in three psychiatric disorders, including autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD) generated by Gandal et al. [7], are analyzed with NoRCE. We repeat this analysis with a TAD filter where the TAD regions of the dorsolateral prefrontal cortex are used. Thirdly, on this dataset, we show the ability of NoRCE to perform pathway enrichment with pathways in the data repository and also with user-defined gene sets. Next, to demonstrate miRNA specific functionalities, we conduct a functional enrichment analysis on a set of differentially expressed miRNAs obtained from [218]. In this study, we filter the neighboring coding genes with target and custom co-expression analyses. Also, we perform miRNA pathway enrichment based on non-coding gene neighbors.

4.8.1 Case Study 1: Functional enrichment analysis of ncRNAs differentially expressed in psychiatric disorders

To demonstrate how NoRCE could be used to gain functional insights given a set of ncRNAs, we retrieved a set of ncRNAs exhibiting gene- or isoform-level differential expression data on brain disorders. These ncRNAs are differentially expressed in at least one of the following disorders: autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD) and pertinent data were

generated by Gandal et al. [7]. In total, the ncRNA gene set contains 1,363 differentially expressed human ncRNAs. In this case study, we also showcase the ability to constrain protein-coding genes with the TAD regions. For this, we use the dorsolateral prefrontal cortex TAD regions data obtained from the same study [7]. Finally, to demonstrate the capability of NoRCE working with custom gene sets, we use the Bader lab pathway gene set [83]. This dataset contains 3,758 pathways compiled from various databases (see Table 4.13). This data set is accessible through the NoRCE data repository and can be loaded by the users, see Section 4.8.4 for a summary and data frame name.

Table 4.13: The number of pathways and the different pathway sources included in the in the March 2019 Bader Lab pathway data set.

Number of Pathways	Pathway Database
NetPath	25
IOB	33
Panther	173
NCI	223
Cyc	240
WikiPathways	490
MSigdb	520
Reactome	2,054

4.8.1.1 Functional enrichment results and example outputs

We use the default parameter setting in all the analyses. In NoRCE, by default, the neighborhood of each ncRNA is taken as the 10 kb extensions towards the transcription start site (TSS) and transcription end site (TES). The protein-coding genes that fall into this neighborhood region of the ncRNAs are input to the enrichment analysis. To increase the power of the statistical analysis, only the GO terms with at least 5 annotated protein-coding genes are considered as suggested by [25].

Table 4.14 lists a subset of the enriched GO terms and shows what the tabular format outputs should include. Interestingly, this list contains terms which are related to neurological functions as long-term memory, sensory perception of

pain, neuronal action potential, and positive regulation of dendritic spine morphogenesis. There are also more general terms that are related to the cell cycle arrest, RNA processing, and signaling. On the downside, we observe some irrelevant terms, which are associated with bone development. However, taken as a whole, relevant terms are more pervasive, suggesting that NoRCE can indeed provide insights into ncRNA functions.

Figure 4.7 shows an alternative representation of the results. The dot plot views the top 35 enriched BP GO terms, sorted based on the significance of enrichment. Figure 4.7 shows the network that can be generated by NoRCE. This network includes the top 7 BP GO terms - enriched gene pairs interaction. The node coloring highlights the clusters with optimal modularity based on a modularity measure maximized over all possible partitions [219]. NoRCE deploys igraph package [220] for this analysis.

4.8.1.2 Filtering with TAD regions

We repeat the analysis in the previous section; but this time, applying a filter based on TAD regions. When this filter is applied, the assignment of the neighboring coding genes is altered; only the protein-coding genes that are nearby to the ncRNA and at the same time reside within the same TAD regions are included. In this analysis, we use custom defined TAD regions for the adult dorsolateral prefrontal cortex that are provided by the [7] study.

Top 10 GO term enrichment results produced by this TAD based analysis are provided in Table 4.15. The enrichment results are visualized in Figure 4.8 (A). The networks of the enriched GO terms. Coding genes are provided in Figure 4.8 (B).

In this use case, we use a custom TAD region file. Additionally, NoRCE repository includes TAD regions for the *human*, *mouse* and *fruit fly* and can be readily used for analysis (See Section 4.8.3).

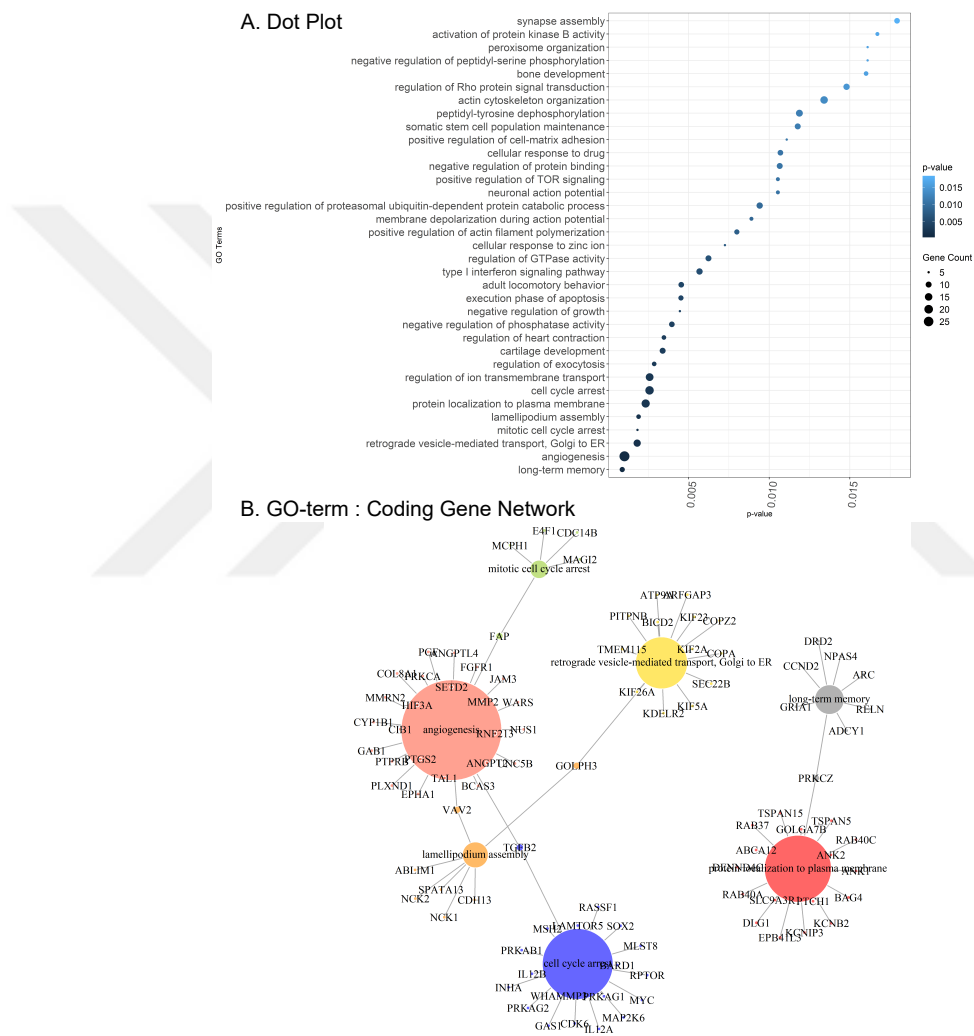


Figure 4.8: The enrichment results obtained by performing TAD based analysis on ncRNAs obtained from [7]. (A) The dot plot where the Y-axis denotes p -values and the x-axis denotes GO terms. The size and the color of dots signify the size of the enriched gene set and the p -value for the corresponding GO term, respectively. B) The network of GO term:protein coding gene. Size of the nodes is proportional to the degree of the node. Different colors are used to differentiate the cluster of proteins.

Table 4.14: Top 10 enrichment results of brain related biological process GO terms that are obtained from the neighbourhood coding genes of ncRNA set of the [7]. The GeneRatio is computed by dividing the number of genes that are overlapping with the coding ones to the number of all protein-coding genes within the input set neighbourhood. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of the background gene set. The EGNo refers to the size of the overlap between the corresponding GO term gene set and the neighboring coding gene set. GeneList columns contain these genes.

GO-ID	GO term	Pvalue	GeneRatio	BGRatio	EGNo	GeneList
GO:0007616	long-term memory	1.92e-4	8/1,260	33/17,523	8	PRKCZ GRIA1 ADCY1 RELN ARC DRD2 CCND2 NPAS4
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	2.02e-4	15/1,260	91/17,523	15	COPA TMEM115 GOLPH3 KIF2A KDELR2 RINT1 BICD2 KIF26A KIF23 COPZ2 ATP9A PITPNB ARFGAP3 SEC22B KIF5A
GO:0006349	regulation of gene expression by genetic imprinting	3.11e-4	5/1,260	15/17,523	5	IGF2 EED CTCF CTCFL ARID4A
GO:0071850	mitotic cell cycle arrest	3.11e-4	5/1,260	15/17,523	5	FAP MAGI2 MCPH1 CDC14B E4F1
GO:0034765	regulation of ion transmembrane transport	3.48e-4	17/1,260	115/17,523	17	CLCNKA CLCNKB KCNK12 SCN3A SCN2A SCN1A SCN9A CACNA2D2 KCNQ5 KCNB2 KCNV1 TMEM109 KCNK4 SCN8A ASIC2 CACNA1G KCNC3
GO:0072659	protein localization to plasma membrane	3.66e-4	19/1,260	135/17,523	19	PRKCZ KCNIP3 ABCA12 DLG1 TSPAN5 ANK2 BAG4 ANK1 KCNB2 DENND4C PTCH1 RAB40A TSPAN15 GOLGA7B RAB40C RAB37 SLC9A3R1 MRAP EPB41L3
GO:0030032	lamellipodium assembly	3.75e-4	7/1,260	29/17,523	7	NCK2 GOLPH3 VAV2 ABLIM1 SPATA13 CDH13 NCK1
GO:0007050	cell cycle arrest	4.32e-4	20/1,260	147/17,523	20	LAMTOR5 TGFB2 MSH2 BARD1 INHA RASSF1 IL12A SOX2 IL12B CDK6 PRKAG2 MYC GAS1 PRKAB1 WHAMMP3 MLST8 TP53 MAP2K6 RPTOR PRKAG1
GO:0032008	positive regulation of TOR signaling	4.59e-4	7/1,260	30/17,523	7	LAMTOR5 GOLPH3 MIOS RELN WDR24 MLST8 RPTOR

Table 4.15: The Top 10 biological process GO term enrichment results. These are obtained by performing TAD based analysis using *brain-disorder-ncRNA* data. The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes within the neighbourhood of the input set. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set. GeneList columns contain these genes.

GO-ID	GO term	Pvalue	GeneRatio	BGRatio	EGNo	GeneList
GO:0007616	long-term memory	8.86e-5	8/1,118	33/17,523	8	PRKCZ GRIA1 ADCY1 RELN ARC DRD2 CCND2 NPAS4
GO:0001525	angiogenesis	1.02e-4	27/1,118	223/17,523	27	TAL1 PTGS2 TGFB2 CYP1B1 FAP SETD2 COL8A1 PLXND1 GAB1 NUS1 EPHA1 ANGPT2 FGFR1 VAV2 UNC5B MMRN2 JAM3 PTPRB PGF WARS CIB1 MMP2 BCAS3 PRKCA RNF213 ANGPTL4 HIF3A
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	1.81e-4	14/1,118	91/17,523	14	COPA TMEM115 GOLPH3 KIF2A KDELR2 BICD2 KIF26A KIF23 COPZ2 ATP9A PITPNB ARFGAP3 SEC22B KIF5A
GO:0071850	mitotic cell cycle arrest	1.84e-4	5/1,118	15/17,523	5	FAP MAGI2 MCPH1 CDC14B E4F1
GO:0030032	lamellipodium assembly	1.90e-4	7/1,118	29/17,523	7	NCK2 GOLPH3 VAV2 ABLIM1 SPATA13 CDH13 NCK1
GO:0072659	protein localization to plasma membrane	2.34e-4	18/1,118	135/17,523	18	PRKCZ KCNIP3 ABCA12 DLG1 TSPAN5 ANK2 BAG4 ANK1 KCNB2 DENND4C PTCH1 RAB40A TSPAN15 GOLGA7B RAB40C RAB37 SLC9A3R1 EPB41L3
GO:0007050	cell cycle arrest	2.58e-4	19/1,118	147/17,523	19	LAMTOR5 TGFB2 MSH2 BARD1 INHA RASSF1 IL12A SOX2 IL12B CDK6 PRKAG2 MYC GAS1 PRKAB1 WHAMMP3 MLST8 MAP2K6 RPTOR PRKAG1
GO:0034765	regulation of ion transmembrane transport	2.59e-4	16/1,118	115/17,523	16	CLCNKA CLCNKB KCNK12 SCN3A SCN2A SCN1A SCN9A CACNA2D2 KCNQ5 KCNB2 KCNV1 TMEM109 SCN8A ASIC2 CACNA1G KCNC3
GO:0017157	regulation of exocytosis	2.897e-4	7/1,118	31/17,523	7	STXBP5L STXBP5 RALA RAB40A RAB40C RAB37 LILGL2
GO:0051216	cartilage development	3.39e-4	10/1,118	58/17,523	10	HOXD3 SATB2 HYAL2 TAPT1 SMAD1 ZEB1 ESRRA SOX9 GNAS COL11A2

4.8.1.3 Pathway enrichment using predefined pathway gene sets and custom user-defined gene sets

For pathway enrichment analysis, the latest versions of the Reactome (Reactome database in R [221], Wiki Pathways, and KEGG (KEGG database in R, [222]) databases are available in NoRCE. These databases will be updated as new versions are made available. Also, NoRCE supports pathway enrichment using custom pathway databases such as MSigDb, or other user curated data. To demonstrate this, we use the Bader Lab dataset as the user-defined pathway gene set analysis. In this analysis, given ncRNA set are expanded within 10kb, and the neighborhood coding genes are enriched with custom BaderLab gene sets.

Table 4.16 lists the pathways that are found to be enriched. Many gene sets that are involved in diseases are related to brain functions such as axon guidance. This observation suggests that it is possible to gain insights into the functions of ncRNAs by studying the functions of the protein-coding genes located proximally to the ncRNAs.

4.8.2 Case Study 2: Functional enrichment analysis of variably expressed miRNAs in brain cancer

Integrating biological information from different sources such as co-expression data is likely to enhance the quality of functional analysis for ncRNAs. Therefore, NoRCE also supports filtering of input ncRNA gene set based on co-expression analysis. The user can input an expression data that includes coding and non-coding RNA expression. NoRCE will compute the co-expressed pairs. When defining coding gene neighborhoods for an ncRNA, the user can choose the option to include a coding gene only if it is co-expressed. Or alternatively, the user can choose to augment the list of coding genes with the co-expressed coding genes.

Exclusively for the miRNA input, NoRCE also allows filtering with miRNA targets. User can choose to filter ncRNA neighbors so that only those that are

Table 4.16: Disease related pathway enrichment results for the *brain_disorder_ncRNA* dataset using customized pathway databases. The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes within the neighbourhood of the input set. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of the background gene set. The EGNNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set. GeneList columns contain these genes.

GO term	Database	GO ID	Pvalue	GeneRatio	BGRatio	EGNNo	GeneList
MEIOSIS	REACTOME	R-HSA-1500620.2	5.49e-5	3.25e-4	22/967	118/12,930	HIST2H4A HIST2H2BE ATR TOP3A POT1 TEX12 FKBP6 SUN1 HIST2H2AC
AXON GUIDANCE	REACTOME	R-HSA-422475.5	5.60e-5	3.25e-4	65/967	533/12,930	RELN DAB1 LAMB1 RPL3L TRPC1 NCK1 PSMD14 PSMD7 ABLM1 ABLIM2 ITGA10 GAB1 PRKCA ANK2 ANK1 ACTG1 SEMA4D DPYSL5 COL4A3 PLXND1 RPS2 TIAM1 RPS6 KALRN RPL11 EPHA1 RPL38 RPL14 RPS4Y1 VAV2 RPLP1 RPS28 RPS24 UNC5B DSCAM AGRN LDB1 DLG1 RASA1 PFN2 SCN9A SCN8A SCN3A SCN2A SCN1A MMP2 DOK6 NCK2 TRPC3 PSMB11 PSMA3 CXCL12 PSMD1 PSMB5 RBM8A NCAM1 CACNA1G MAGOH RNPS1 PRKCQ EPHA4 VASP ARHGEF28
Amyotrophic lateral sclerosis (ALS)	WikiPathways	WP2447	4.40e-5	1.77e-3	10/967	39/12,930	GRIA1 BAD NEFM PPP3CC CASP1 NEFH CASP12 TP53 MAP2K6 CASP9
Parkinsons Pathway	WikiPathways	WP2371	4.93e-5	1.88e-3	16/967	85/12,930	MIR17HG ATXN2 GPR37 UBE2L3 UBE2J2 CASP9 SNCAIP MIR136 MIR16-2 CCNE2

among the targets of the miRNA are included. To demonstrate these options, we use a set of miRNAs differentially expressed in brain cancer obtained from dbDEMC 2.0[218]. This set contains 407 miRNAs and is provided in NoRCE with the name *brain_miRNA*. For the co-expression filter, we use the brain cancer patient data obtained from TCGA. The data include expression levels for mRNA and miRNA for 527 tumor patients. We apply the same pre-processing step as in our previous method [29]. Genes and miRNAs that have very low expression, ($\text{RPKM} < 0.05$) in many patients (more than %20 of the samples) are filtered. The gene expression values are log2 transformed, and those that have high variability are retained for co-expression analysis. For this, only the genes with median absolute deviation (MAD) above 0.5 are used.

Final expression data set contains 444 miRNA and 12,643 mRNA genes on 527 tumor patients on which we perform Pearson correlation analysis. The mRNAs that have more than 0.3 correlation with a miRNA are retained. Additionally, we apply the miRNA target filter. NoRCE only allows the coding genes that are highly correlated with the mRNA and intersected with the nearby coding genes. Enrichment analysis is conducted on this list.

4.8.2.1 Functional annotation of miRNA based on custom expression

We perform enrichment analysis for the miRNAs that are differentially expressed in brain cancer. We apply default parameter settings for this analysis. The results obtained after performing enrichment analysis using a gene set that is both targeted by and co-expressed with the input miRNAs are shown in Table 4.16. We observe that most of the results are related to brain functionality, suggesting that miRNAs differentially expressed in brain cancer might be involved in these processes (4.17). Here, Figure 4.9 displays the hierarchical view of the enriched GO terms, which is an alternative visualization available under NoRCE.

Table 4.17: Some brain related biological process functions that are enriched based on target and custom co-expression analysis in brain cancer specific miRNAs. The The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes that are within the neighbourhood of the input set. The BGRatio column represents the ratio of the number of genes within the enriched GO term set to the size of the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set while GeneList columns list these genes.

GO-ID	GO term	Pval	PAdj	GeneRatio	BGRatio	GeneList
GO:0007411	axon guidance	4.66e-15	5.62e-12	88/3,617	193/17,523	ARX BDNF CDH4 GDNF GRI1 CHL1 CNTN2 CNTN4 CSF1R CXCL12 CXCR4 DAB1 DOK1 DOK4 DOK5 DOK6 DPYSL5 DRAXIN DRGX EFNA1 EFNA2 EFNA3 EFNA4 EFNA5 EFNB3 EFNB4 EFNA5 EPHA8 EPHB2 EPHB3 ETV1 EZR FAM129B FEZF1 FLRT2 FLRT3 FOXD1 GAP43 GATA3 GFRA2 GFRA3 GLI2 GLI3 GRB10 GRB7 KIF5A KIF5C LICAM LAMA2 LGH1 LHX2 MATN2 NFASC NGFR NR4A3 NRP1 NRXN1 NRXN3 PAX6 PDLIM7 PIK3CD PRKCQ RAPIGAP RELN ROBO2 RPS6KA5 SCN1B SEMA3A SEMA3C SEMA3F SEMA4F SEMA6C SHANK3 SHC3 SHH SLIT1 SLIT2 SLIT3 SPTB SPTBN2 SPTBN4 TNFRSF8 UNC5A UNC5C UNC5D VASP VAX1 VLDLR WNT5A
GO:0007399	nervous system development	1.39e-12	1.01e-9	121/3,617	322/17,523	ADORA1 APBA1 APOB ARHGAP26 BARX2 BDNF BZW2 CABLES1 CAMK2A CAMK2G CBLN1 CHRDL1 CHRM1 CHRM2 CHRM3 CNTRF CNTN3 CNTN4 CPLX2 CRIM1 CSGALNACT1 CXCL1 CYP46A1 DCLK1 DCX DLG2 DLG3 DLG4 DLX5 DLX6 DDC2A DPF1 DPYSL4 DPYSL5 EDNRB EFNA5 EFNB3 ELAVL3 ENC1 EPHB2 ERBB4 FGF12 FGF13 FGF14 FGF2 FGF5 FOS FUT9 GAS7 GDA GFRA3 GJB1 GLRB HES1 INTU JAG1 KALRN LICAM LGH1 MAFK MDK MEF2C MPPED2 MYT1 MYTIL N4BP3 NBL1 NDN NDP NEUROD2 NIN1 NOTCH2 NR2E1 NRG1 NRGN NRN1 NRSN1 NUMBL OLFM1 P2RX5 PCDH1 PCDH10 PCDH18 PCDH19 PCDH8 PCDHA11 PCDHA13 PCDHA5 PCDHA6 PCDHAC2 PCDHGA12 PCDHGA3 PCDHGA9 PCDHGC5 PCSK2 PTN PURA RAPGEF5 RAPGEF11 RBFOX1 RBFOX2 RBFOX3 SCN2B SCN3B SCN8A SEMA3E SEMA4F SEMA6B SHOX2 SLC4A10 SPOCK1 SRRM4 STSIA4 TMOD2 TRPC5 TLL7 VEGFA VLDLR ZC3H12A ZIC5
GO:0007409	axonogenesis	3.16e-9	1.13e-6	46/3,617	97/17,523	AMIGO1 ANK3 ATL1 ATP8A2 BCL2 BRSK1 BRSK2 CNP CNTN4 CTNNA2 DOCK7 DSCAML1 ECM2 FGFR2 FMOD LINGO1 LINGO2 LRNF2 LRNF5 LRRN3 LUM MAP2 NEFH NKX2-8 NRCAM NTNG1 NTNG2 NUMBL OGN OMD OMG PAK3 PAX2 POU4F1 PRELP RTN4R SEMA4A SLITRK1 SLITRK2 SLITRK3 SLITRK4 SLITRK5 SPTBN4 STMN1 VAX2 WNT7A
GO:0007612	learning	3.81e-8	7.87e-6	32/3,617	61/17,523	ARC ATP8A1 CHRN2 CNTN2 CNTNAP2 DLG4 DRD1 EPHB2 FGF13 FOSL1 GRM5 JPH4 JUN NLGN3 NPAS4 NRXN1 NRXN3 NTRK2 PRKAR2B PTGS2 PTN RGS1 SHANK2 SHANK3 SLC12A5 SLC24A2 SLC6A1 SLC8A2 SLC8A3 SORCS3 SYNJ1

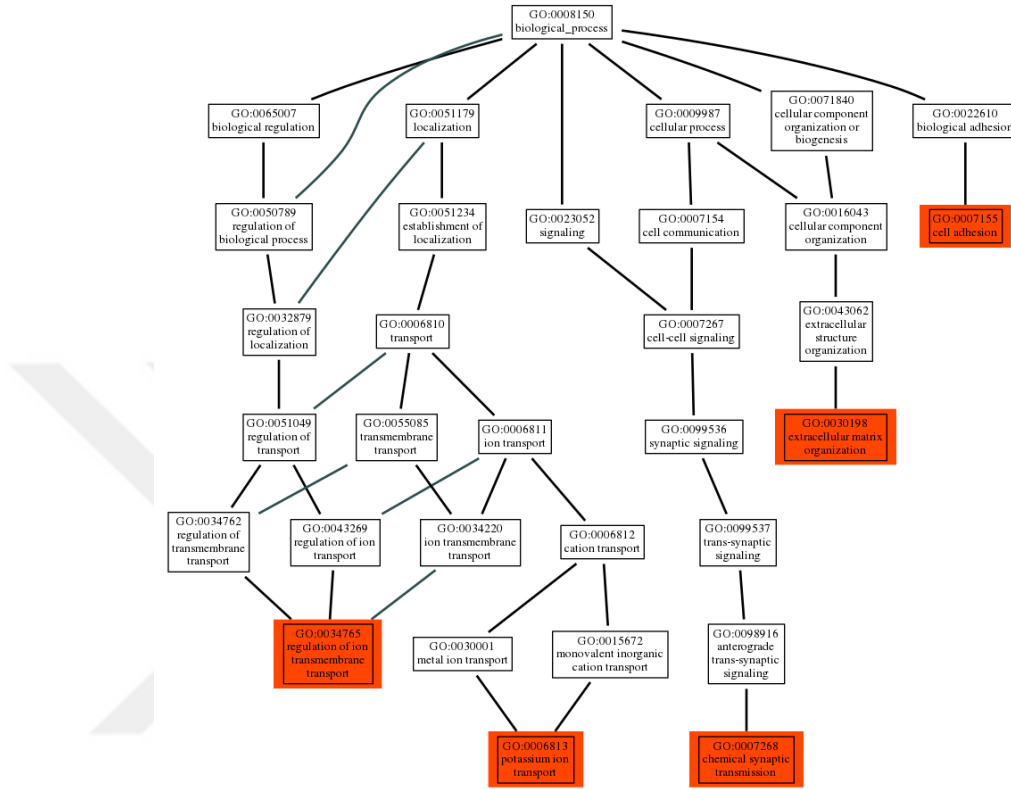


Figure 4.9: Directed acyclic graph of the enriched top 5 GO terms based on a gene set that are co-expressed and targeted by the brain cancer miRNA. GO terms are colored based on their p -values

4.8.2.2 Reactome pathway enrichment of miRNA set based on neighboring protein-coding genes and miRNA target analysis

In this section, the neighboring protein-coding genes targeted by the input miRNA list (used in the previous section) are used for Reactome pathway enrichment. Results show that cellular transport based processes are enriched processes. Table 4.18 lists a term which is found to be enriched in the previous analysis, as well. Moreover, previous studies have shown that transport mechanisms play a role in metastatic brain tumors [223]. Together these observations suggest that NoRCE can facilitate the identification of important pathways in which ncRNAs participate. 4.10 also shows an alternative visualization provided by NoRCE. The genes that overlap with the ncRNA's coding neighbors are highlighted in the pathway map.

Table 4.18: Pathway enrichment of miRNA gene set based on neighbor targets. The GeneRatio column represents the ratio of the number of coding genes overlapping with the input coding genes within the GO term set to the number of all protein-coding genes within the input set neighborhood. The BGRatio column represents the ratio of the number of genes found in the enriched GO term set to the size of all of the genes in the background gene set. The EGNo denotes the size of the overlap with the corresponding GO term gene set and the neighboring coding gene set. GeneList columns list these genes.

GO term	GO ID	Pvalue	PAdjust	GeneRatio	BGRatio	EGNo	GeneList
R-HSA-199992	trans-Golgi Network Vesicle Budding	7.16e−5	1.00e−4	5/64	72/10,554	5	AP1S1 CLVS1 DNM2 BLOC1S3 ARRB1
R-HSA-421837	Clathrin derived vesicle budding	7.16e−5	1.00e−4	5/64	72/10,554	5	AP1S1 CLVS1 DNM2 BLOC1S3 ARRB1

4.8.2.3 Brain miRNA gene list enrichment based on driver cancer gene set

We apply gene enrichment analysis to identify the regions that are statistically similar to the given gene set. In this case study, we use differentially expressed brain cancer miRNA named as *brain_mirna* in NoRCE to identify whether the neighborhood of the miRNA set is statistically similar to a brain cancer driver gene set obtained from Intogen [224]. In this analysis, only the neighborhood of the miRNAs is considered, and we apply the default parameter settings for non-coding gene enrichment. We find that brain miRNA gene list is statistically enriched for the driver gene list ($p\text{-value} = 3.33\text{e}^{-132}$ and $p\text{-Adjust} = 3.33\text{e}^{-132}$). This analysis showcases the use of a custom gene-set for functional enrichment analysis.

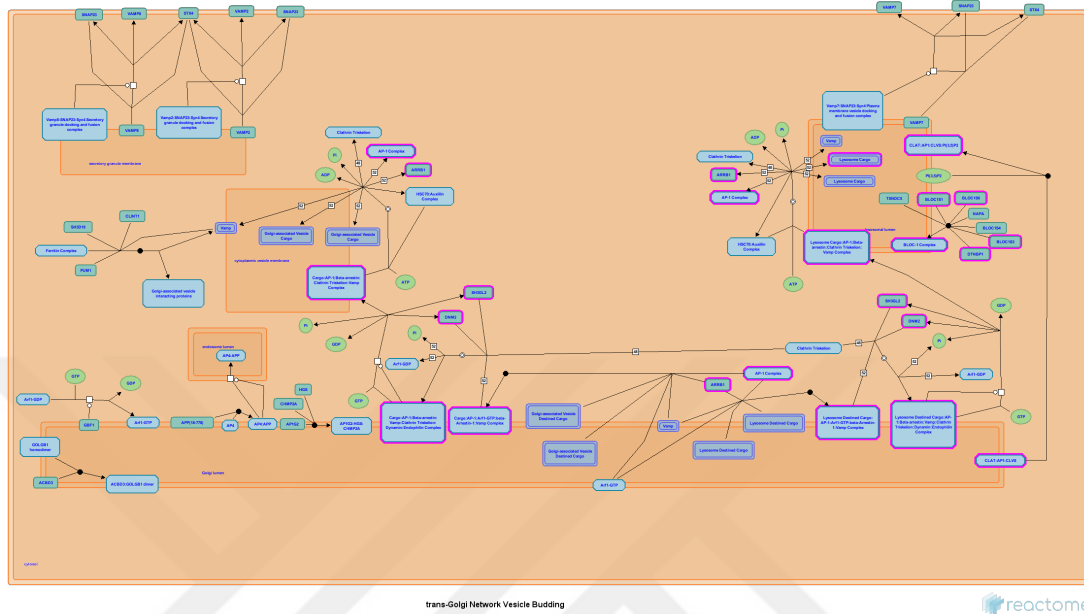


Figure 4.10: Reactome pathway map for the enriched pathway, R-HSA-199992(trans-Golgi Network Vesicle Budding). Genes that are enriched for this pathway are marked with purple.

4.8.3 Data in NoRCE repository for the vignette and supplementary results

- `brain_disorder_ncRNA` : A list of ncRNAs differentially expressed in three psychiatric disorders including autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD) generated by Gandal et al. [7]. The set comprises 1,363 ncRNAs. (Case study 1, Section 3.1).
- `tad_custom` : The TAD regions for dorsolateral prefrontal cortex obtained from [7]. The dataset contains 2,735 TAD regions. (Case study 1, Section 3.1).
- `miRCancerdb`: The pre-computed Pearson correlation values of the expressions of pairs of 18,069,409 miRNAs and genes for 34 cancer types. The data incorporates 1,015 miRNAs, which are curated based on patient expression profiles that are provided by the TCGA. miRCancerdb is available at https://figshare.com/articles/miRCancer_db_gz/5576329 [208].

(Case study 2, Section 3.3).

- `brain_mirna` : 407 differentially expressed miRNAs in human brain obtained from dbDEMC 2.0. (Case study 2, Section 3.4).
- `ncRegion` : The relevant regions of the differentially expressed human ncRNA genes (except pseudogenes) for three psychiatric disorders. The diseases include autism spectrum disorder (ASD), schizophrenia (SCZ), and bipolar disorder (BD), and the data were generated by Gandal et al. [7]. The final set contains 930 gene regions.

4.8.4 Other features

A list of features and the related questions that can be answered using NoRCE.

- Filter the ncRNAs based on biotypes and perform enrichment analysis on the curated ncRNA gene list
- Perform GO annotation and enrichment on the GO-slim dataset
- Neighbourhood gene set can be limited to the protein coding genes that are exons or introns and fall into the user-defined neighborhood region

Chapter 5

Conclusion

As transcriptome gets cataloged in greater depth, it becomes evident that the vast majority of the mammalian transcriptome is non-coding. One type of non-coding RNAs is lncRNA. The dysregulation of ncRNA expression is also associated with many diseases, including cancer. ncRNAs can not only serve for potential biomarkers but also stand out as promising therapeutic agents. Therefore, untangling their complex regulatory interactions is - essential for shedding light into gene expression programs regulated by them and for translating this knowledge into effective therapies. Although many different regulatory interactions may arise among ncRNAs and coding RNAs, we focused on two specific regulatory interactions in this thesis.

In chapter 2, we present our results on sponge interactions. The basic motivation follows from the accruing evidence that lncRNAs compete for binding to miRNAs, thus acting as sponges which regulate the gene activity. The three-way regulatory interaction between lncRNAs, miRNAs, and the mRNAs is observed in multiple cancer types, including breast cancer. In this study, we develop a computational framework that relies on kernel interaction tests on gene expression data. We apply this framework to identify potential subtype-specific lncRNA mediated sponge interactions in breast cancer. We recognize sponges that are highly subtype-specific. Some of these sponges also have high prognostic potential. The

list of sponges, which we discovered, can be probed and tested by experimental analyses and potentially help uncover unknown molecular mechanisms underlying breast cancer subtypes. These interactions can serve for biomarkers for the subtypes. The methodology we develop is not breast-cancer specific. It has broader applicability and relevance to studies on other cancers, diseases, and healthy cells. As future work, we are in the process of obtaining results for different cancer types and presenting them in a database that can be readily accessed by the community.

The second type of regulatory interactions we focus on within this work is the synergistic target regulation of miRNAs, in chapter 3. To identify synergistic miRNAs pairs and their target mRNA, we propose a novel method which we call miRCoop. MiRCoop uses kernel-based statistical interactions and finds triplets where miRNA expression levels have weak or no statistical dependencies with the mRNA levels but are jointly dependent. MiRCoop makes use of both RNA sequence information for target predictions. We demonstrate the use of miRCoop in kidney cancer and find triplets with supporting biological evidence on their relevance. The work presented here can be extended in several directions. Finding higher-order interactions is one possible direction. Secondly, we focus on miRNAs that regulate a target gene directly that is those that simultaneously bind to a target. There are other ways in which the pairs of miRNAs regulate a common target. An additional line of future work would be to extend miRCoop to discover these multi-level regulations, where one miRNA exert function on the mRNA directly while the other activities are mediated through an intermediary transcription factor. In future work, we will use miRCoop on other diseases as well as on healthy cells.

In chapter 4, we present a new methodology for the functional enrichment of ncRNA sets. This method transfers knowledge of coding gene annotations to the ncRNAs to gain insight into the functional importance of a list of interesting ncRNAs. We showcase this method with different scenarios in cancer and psychiatric diseases. We also explore the similarities of different cancer based on dysregulated lncRNAs. We present this methodology as an R package, NoRCE.

The ncRNAs open up a fascinating window into the workings of the cell. With

their great therapeutic potential, they will continue to write the rules of molecular biology. We hope that this thesis, along with the tools developed herein, will contribute to the continued endeavor in understanding the biology of ncRNAs and their interactions in different biological contexts.



Bibliography

- [1] L. Wang, H. J. Park, S. Dasari, S. Wang, J.-P. Kocher, and W. Li, “Cpat: Coding-potential assessment tool using an alignment-free logistic regression model,” *Nucleic acids research*, vol. 41, no. 6, pp. e74–e74, 2013.
- [2] L. Kong, Y. Zhang, Z.-Q. Ye, X.-Q. Liu, S.-Q. Zhao, L. Wei, and G. Gao, “Cpc: assess the protein-coding potential of transcripts using sequence features and support vector machine,” *Nucleic acids research*, vol. 35, no. suppl_2, pp. W345–W349, 2007.
- [3] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, “Cytoscape: a software environment for integrated models of biomolecular interaction networks,” *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [4] X. Meng, J. Wang, C. Yuan, X. Li, Y. Zhou, R. Hofestädt, and M. Chen, “Cancernet: a database for decoding multilevel molecular interactions across diverse cancer types,” *Oncogenesis*, vol. 4, no. 12, p. e177, 2015.
- [5] U. Schmitz, X. Lai, F. Winter, O. Wolkenhauer, J. Vera, and S. K. Gupta, “Cooperative gene regulation by microrna pairs and their identification using a computational workflow,” *Nucleic acids research*, vol. 42, no. 12, pp. 7539–7552, 2014.
- [6] F. Hahne and R. Ivanek, “Statistical genomics: methods and protocols,” *Chapter visualizing genomic data using Gviz and Bioconductor. New York: Springer New York*, pp. 335–51, 2016.

- [7] M. J. Gandal, P. Zhang, E. Hadjimichael, R. L. Walker, C. Chen, S. Liu, H. Won, H. van Bakel, M. Varghese, Y. Wang, *et al.*, “Transcriptome-wide isoform-level dysregulation in asd, schizophrenia, and bipolar disorder,” *Science*, vol. 362, no. 6420, p. eaat8127, 2018.
- [8] I. H. G. S. Consortium *et al.*, “Initial sequencing and analysis of the human genome,” *nature*, vol. 409, no. 6822, p. 860, 2001.
- [9] X.-D. Fu, “Non-coding rna: a new frontier in regulatory biology,” *National science review*, vol. 1, no. 2, pp. 190–204, 2014.
- [10] J. S. Mattick and I. V. Makunin, “Non-coding rna,” *Human molecular genetics*, vol. 15, no. suppl_1, pp. R17–R29, 2006.
- [11] S. R. Eddy, “Non-coding rna genes and the modern rna world,” *Nature Reviews Genetics*, vol. 2, no. 12, p. 919, 2001.
- [12] K. V. Prasanth and D. L. Spector, “Eukaryotic regulatory rnas: an answer to the ‘genome complexity’conundrum,” *Genes & development*, vol. 21, no. 1, pp. 11–42, 2007.
- [13] E. A. Gibb, C. J. Brown, and W. L. Lam, “The functional role of long non-coding rna in human carcinomas,” *Molecular cancer*, vol. 10, no. 1, p. 38, 2011.
- [14] H. Ling, M. Fabbri, and G. A. Calin, “Micrnas and other non-coding rnas as targets for anticancer drug development,” *Nature reviews Drug discovery*, vol. 12, no. 11, p. 847, 2013.
- [15] A. Fatica and I. Bozzoni, “Long non-coding rnas: new players in cell differentiation and development,” *Nature Reviews Genetics*, vol. 15, no. 1, p. 7, 2014.
- [16] E.-M. Weick and E. A. Miska, “pirnas: from biogenesis to function,” *Development*, vol. 141, no. 18, pp. 3458–3471, 2014.
- [17] Y. Shen, X. Yu, L. Zhu, T. Li, Z. Yan, and J. Guo, “Transfer rna-derived fragments and trna halves: biogenesis, biological functions and their roles

- in diseases,” *Journal of Molecular Medicine*, vol. 96, no. 11, pp. 1167–1176, 2018.
- [18] S. P. Barrett and J. Salzman, “Circular rnas: analysis, expression and potential functions,” *Development*, vol. 143, no. 11, pp. 1838–1847, 2016.
 - [19] V. Ambros, “The functions of animal micrnas,” *Nature*, vol. 431, no. 7006, p. 350, 2004.
 - [20] C. M. Croce, “Causes and consequences of micrna dysregulation in cancer,” *Nature reviews genetics*, vol. 10, no. 10, p. 704, 2009.
 - [21] A. Kozomara and S. Griffiths-Jones, “mirbase: integrating micrna annotation and deep-sequencing data,” *Nucleic acids research*, vol. 39, no. suppl_1, pp. D152–D157, 2010.
 - [22] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel, “Most mammalian mrnas are conserved targets of micrnas,” *Genome research*, vol. 19, no. 1, pp. 92–105, 2009.
 - [23] J. Cao, “The functional role of long non-coding rnas and epigenetics,” *Biological procedures online*, vol. 16, no. 1, p. 42, 2014.
 - [24] X. C. Quek, D. W. Thomson, J. L. Maag, N. Bartonicek, B. Signal, M. B. Clark, B. S. Gloss, and M. E. Dinger, “lncrnadb v2. 0: expanding the reference database for functional long noncoding rnas,” *Nucleic acids research*, vol. 43, no. D1, pp. D168–D173, 2014.
 - [25] Q. Jiang, R. Ma, J. Wang, X. Wu, S. Jin, J. Peng, R. Tan, T. Zhang, Y. Li, and Y. Wang, “Lncrna2function: a comprehensive resource for functional investigation of human lncrnas based on rna-seq data,” *BMC genomics*, vol. 16, no. 3, p. S2, 2015.
 - [26] C. Park, N. Yu, I. Choi, W. Kim, and S. Lee, “lncrnator: a comprehensive resource for functional investigation of long non-coding rnas,” *Bioinformatics*, vol. 30, no. 17, pp. 2480–2485, 2014.

- [27] T. Alam, M. Uludag, M. Essack, A. Salhi, H. Ashoor, J. B. Hanks, C. Kapfer, K. Mineta, T. Gojobori, and V. B. Bajic, “Farna: knowledgebase of inferred functions of non-coding rna transcripts,” *Nucleic acids research*, vol. 45, no. 5, pp. 2838–2848, 2017.
- [28] A. Zampetaki, A. Albrecht, and K. Steinhofel, “Long-noncoding rna structure and function: Is there a link?,” *Frontiers in physiology*, vol. 9, p. 1201, 2018.
- [29] G. Olgun, O. Sahin, and O. Tastan, “Discovering lncrna mediated sponge interactions in breast cancer molecular subtypes,” *BMC genomics*, vol. 19, no. 1, p. 650, 2018.
- [30] G. Olgun and O. Tastan, “mircoop: Identifying cooperating mirnas via kernel based interaction tests,” *bioRxiv*, p. 769307, 2019.
- [31] G. Olgun, A. Nabi, and O. Tastan, “Norce: Non-coding rna sets cis enrichment tool,” *bioRxiv*, p. 663765, 2019.
- [32] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. M. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, *et al.*, “Landscape of transcription in human cells,” *Nature*, vol. 489, no. 7414, p. 101, 2012.
- [33] D. P. Bartel, “MicroRNAs: target recognition and regulatory functions,” *cell*, vol. 136, no. 2, pp. 215–233, 2009.
- [34] S. Lin and R. I. Gregory, “MicroRNA biogenesis pathways in cancer,” *Nature reviews cancer*, vol. 15, no. 6, pp. 321–333, 2015.
- [35] G. A. Calin and C. M. Croce, “MicroRNA signatures in human cancers,” *Nature reviews. Cancer*, vol. 6, no. 11, p. 857, 2006.
- [36] L. Wang and J. Wang, “MicroRNA-mediated breast cancer metastasis: from primary site to distant organs,” *Oncogene*, vol. 31, no. 20, p. 2499, 2012.
- [37] U. Raza, Ö. Saatci, S. Uhlmann, S. A. Ansari, E. Eyüpoğlu, E. Yurdusev, M. Mutlu, P. G. Ersan, M. K. Altundağ, J. D. Zhang, *et al.*, “The mir-644a/ctbp1/p53 axis suppresses drug resistance by simultaneous inhibition

- of cell survival and epithelial-mesenchymal transition in breast cancer,” *Oncotarget*, vol. 7, no. 31, p. 49859, 2016.
- [38] M. Mutlu, Ö. Saatci, S. A. Ansari, E. Yurdusev, H. Shehwana, Ö. Konu, U. Raza, and Ö. Şahin, “mir-564 acts as a dual inhibitor of pi3k and mapk signaling networks and inhibits proliferation and invasion in breast cancer,” *Scientific reports*, vol. 6, p. 32541, 2016.
- [39] K. C. Wang and H. Y. Chang, “Molecular mechanisms of long noncoding rnas,” *Molecular cell*, vol. 43, no. 6, pp. 904–914, 2011.
- [40] J. R. Prensner and A. M. Chinnaiyan, “The emergence of lncrnas in cancer biology,” *Cancer discovery*, vol. 1, no. 5, pp. 391–407, 2011.
- [41] J.-h. Yuan, F. Yang, F. Wang, J.-z. Ma, Y.-j. Guo, Q.-f. Tao, F. Liu, W. Pan, T.-t. Wang, C.-c. Zhou, *et al.*, “A long noncoding rna activated by $\text{tgf-}\beta$ promotes the invasion-metastasis cascade in hepatocellular carcinoma,” *Cancer cell*, vol. 25, no. 5, pp. 666–681, 2014.
- [42] J. R. Prensner, M. K. Iyer, A. Sahu, I. A. Asangani, Q. Cao, L. Patel, I. A. Vergara, E. Davicioni, N. Erho, M. Ghadessi, *et al.*, “The long noncoding rna schlaf1 promotes aggressive prostate cancer and antagonizes the swi/snf complex,” *Nature genetics*, vol. 45, no. 11, pp. 1392–1398, 2013.
- [43] L. Salmena, L. Poliseno, Y. Tay, L. Kats, and P. P. Pandolfi, “A cerna hypothesis: the rosetta stone of a hidden rna language?,” *Cell*, vol. 146, no. 3, pp. 353–358, 2011.
- [44] M. Cesana, D. Cacchiarelli, I. Legnini, T. Santini, O. Sthandier, M. Chinappi, A. Tramontano, and I. Bozzoni, “A long noncoding rna controls muscle differentiation by functioning as a competing endogenous rna,” *Cell*, vol. 147, no. 2, pp. 358–369, 2011.
- [45] J. M. Franco-Zorrilla, A. Valli, M. Todesco, I. Mateos, M. I. Puga, I. Rubio-Somoza, A. Leyva, D. Weigel, J. A. García, and J. Paz-Ares, “Target mimicry provides a new mechanism for regulation of microrna activity,” *Nature genetics*, vol. 39, no. 8, p. 1033, 2007.

- [46] D. W. Thomson and M. E. Dinger, “Endogenous microRNA sponges: evidence and controversy,” *Nature Reviews Genetics*, vol. 17, no. 5, p. 272, 2016.
- [47] T. Xia, Q. Liao, X. Jiang, Y. Shao, B. Xiao, Y. Xi, and J. Guo, “Long noncoding rna associated-competing endogenous rnas in gastric cancer,” *Scientific reports*, vol. 4, 2014.
- [48] Y.-C. Chiu, T.-H. Hsiao, Y. Chen, and E. Y. Chuang, “Parameter optimization for constructing competing endogenous rna regulatory network in glioblastoma multiforme and other cancers,” *BMC genomics*, vol. 16, no. 4, p. S1, 2015.
- [49] S. Ye, L. Yang, X. Zhao, W. Song, W. Wang, and S. Zheng, “Bioinformatics method to predict two regulation mechanism: Tf-mirna-mrna and lncrna-mirna-mrna in pancreatic cancer,” *Cell biochemistry and biophysics*, vol. 70, no. 3, pp. 1849–1858, 2014.
- [50] M. Zhou, X. Wang, H. Shi, L. Cheng, Z. Wang, H. Zhao, L. Yang, and J. Sun, “Characterization of long non-coding rna-associated cerna network to reveal potential prognostic lncrna biomarkers in human ovarian cancer,” *Oncotarget*, vol. 7, no. 11, p. 12598, 2016.
- [51] P. Paci, T. Colombo, and L. Farina, “Computational analysis identifies a sponge interaction network between long non-coding rnas and messenger rnas in human breast cancer,” *BMC systems biology*, vol. 8, no. 1, p. 83, 2014.
- [52] S. Kurozumi, Y. Yamaguchi, M. Kurosumi, M. Ohira, H. Matsumoto, and J. Horiguchi, “Recent trends in microRNA research into breast cancer with particular focus on the associations between microRNAs and intrinsic subtypes,” *Journal of human genetics*, 2016.
- [53] C. Blenkiron, L. D. Goldstein, N. P. Thorne, I. Spiteri, S.-F. Chin, M. J. Dunning, N. L. Barbosa-Morais, A. E. Teschendorff, A. R. Green, I. O. Ellis, *et al.*, “MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype,” *Genome biology*, vol. 8, no. 10, p. R214, 2007.

- [54] C. G. A. Network *et al.*, “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, no. 7418, p. 61, 2012.
- [55] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, “Kernel-based conditional independence test and application in causal discovery,” *arXiv preprint arXiv:1202.3775*, 2012.
- [56] A. L. Sarver and S. Subramanian, “Competing endogenous rna database,” *Bioinformatics*, vol. 8, no. 15, p. 731, 2012.
- [57] L. Polisen, L. Salmena, J. Zhang, B. Carver, W. J. Haveman, and P. P. Pandolfi, “A coding-independent function of gene and pseudogene mrnas regulates tumour biology,” *Nature*, vol. 465, no. 7301, p. 1033, 2010.
- [58] X. Zhou, J. Liu, and W. Wang, “Construction and investigation of breast-cancer-specific cerna network based on the mrna and mirna expression data,” *IET systems biology*, vol. 8, no. 3, pp. 96–103, 2014.
- [59] J. Xu, Y. Li, J. Lu, T. Pan, N. Ding, Z. Wang, T. Shao, J. Zhang, L. Wang, and X. Li, “The mrna related cerna–cerna landscape and significance across 20 major cancer types,” *Nucleic acids research*, vol. 43, no. 17, pp. 8169–8182, 2015.
- [60] F. Conte, G. Fiscon, M. Chiara, T. Colombo, L. Farina, and P. Paci, “Role of the long non-coding rna pvt1 in the dysregulation of the cerna-cerna network in human breast cancer,” *PLoS One*, vol. 12, no. 2, p. e0171661, 2017.
- [61] J. Zhang, T. D. Le, L. Liu, and J. Li, “Identifying mirna sponge modules using biclustering and regulatory scores,” *BMC bioinformatics*, vol. 18, no. 3, p. 44, 2017.
- [62] J. Zhang, L. Liu, T. Xu, Y. Xie, C. Zhao, J. Li, and T. D. Le, “mirsponger: an r/bioconductor package for the identification and analysis of mirna sponge interaction networks and modules,” *BMC bioinformatics*, vol. 20, no. 1, p. 235, 2019.

- [63] P. Sumazin, X. Yang, H.-S. Chiu, W.-J. Chung, A. Iyer, D. Llobet-Navas, P. Rajbhandari, M. Bansal, P. Guarnieri, J. Silva, *et al.*, “An extensive microrna-mediated network of rna-rna interactions regulates established oncogenic pathways in glioblastoma,” *Cell*, vol. 147, no. 2, pp. 370–381, 2011.
- [64] Y. Tay, L. Kats, L. Salmena, D. Weiss, S. M. Tan, U. Ala, F. Karreth, L. Poliseno, P. Provero, F. Di Cunto, *et al.*, “Coding-independent regulation of the tumor suppressor pten by competing endogenous mrnas,” *Cell*, vol. 147, no. 2, pp. 344–357, 2011.
- [65] L.-J. Li, Z.-W. Zhu, W. Zhao, S.-S. Tao, B.-Z. Li, S.-Z. Xu, J.-B. Wang, M.-Y. Zhang, J. Wu, R.-X. Leng, *et al.*, “Circular rna expression profile and potential function of hsa_circ_0045272 in systemic lupus erythematosus,” *Immunology*, vol. 155, no. 1, pp. 137–149, 2018.
- [66] R. M. Piro and A. Marsico, “Network-based methods and other approaches for predicting lncrna functions and disease associations,” in *Computational Biology of Non-Coding RNA*, pp. 301–321, Springer, 2019.
- [67] P. Furió-Tarí, S. Tarazona, T. Gabaldón, A. J. Enright, and A. Conesa, “spongescan: A web for detecting microrna binding elements in lncrna sequences,” *Nucleic acids research*, vol. 44, no. W1, pp. W176–W180, 2016.
- [68] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, *et al.*, “Gen-code: the reference human genome annotation for the encode project,” *Genome research*, vol. 22, no. 9, pp. 1760–1774, 2012.
- [69] L. Su and H. White, “A nonparametric hellinger metric test for conditional independence,” *Econometric Theory*, vol. 24, no. 4, pp. 829–864, 2008.
- [70] T.-M. Huang *et al.*, “Testing conditional independence using maximal non-linear conditional correlation,” *The Annals of Statistics*, vol. 38, no. 4, pp. 2047–2091, 2010.
- [71] K. Song, “Testing conditional independence via rosenblatt transforms,” 2007.

- [72] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.
- [73] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, “Predicting effective microrna target sites in mammalian mrnas,” *elife*, vol. 4, p. e05005, 2015.
- [74] A. Jeggari, D. S. Marks, and E. Larsson, “mircode: a map of putative microrna target sites in the long non-coding transcriptome,” *Bioinformatics*, vol. 28, no. 15, pp. 2062–2063, 2012.
- [75] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie, “Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites,” *Genome biology*, vol. 11, no. 8, p. R90, 2010.
- [76] 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, p. 1278, 2007.
- [77] K. C. Miranda, T. Huynh, Y. Tay, Y.-S. Ang, W.-L. Tam, A. M. Thomson, B. Lim, and I. Rigoutsos, “A pattern-based method for the identification of microrna binding sites and their corresponding heteroduplexes,” *Cell*, vol. 126, no. 6, pp. 1203–1217, 2006.
- [78] S. Das, S. Ghosal, R. Sen, and J. Chakrabarti, “lncdb: database of human long noncoding rna acting as competing endogenous rna,” *PloS one*, vol. 9, no. 6, p. e98965, 2014.
- [79] C.-H. Chou, N.-W. Chang, S. Shrestha, S.-D. Hsu, Y.-L. Lin, W.-H. Lee, C.-D. Yang, H.-C. Hong, T.-Y. Wei, S.-J. Tu, *et al.*, “mirtarbase 2016: updates to the experimentally validated mirna-target interactions database,” *Nucleic acids research*, vol. 44, no. D1, pp. D239–D247, 2015.
- [80] M. D. Paraskevopoulou, I. S. Vlachos, D. Karagkouni, G. Georgakilas, I. Kanellos, T. Vergoulis, K. Zagganas, P. Tsanakas, E. Floros, T. Dalmagas, *et al.*, “Diana-lncbase v2: indexing microrna targets on non-coding transcripts,” *Nucleic acids research*, vol. 44, no. D1, pp. D231–D238, 2016.

- [81] D. P. Harrington and T. R. Fleming, “A class of rank test procedures for censored survival data,” *Biometrika*, vol. 69, no. 3, pp. 553–566, 1982.
- [82] 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.
- [83] D. Merico, R. Isserlin, O. Stueker, A. Emili, and G. D. Bader, “Enrichment map: a network-based method for gene-set enrichment visualization and interpretation,” *PloS one*, vol. 5, no. 11, p. e13984, 2010.
- [84] J. H. Ward Jr, “Hierarchical grouping to optimize an objective function,” *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.
- [85] R. Mojena, “Hierarchical grouping methods and stopping rules: An evaluation,” *The Computer Journal*, vol. 20, no. 4, pp. 359–363, 1977.
- [86] G. W. Milligan and M. C. Cooper, “An examination of procedures for determining the number of clusters in a data set,” *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.
- [87] W. Lv, L. Wang, J. Lu, J. Mu, Y. Liu, and P. Dong, “Long noncoding rna kiaa0125 potentiates cell migration and invasion in gallbladder cancer,” *BioMed research international*, vol. 2015, 2015.
- [88] R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M.-C. Tsai, T. Hung, P. Argani, J. L. Rinn, *et al.*, “Long noncoding rna hotair reprograms chromatin state to promote cancer metastasis,” *nature*, vol. 464, no. 7291, p. 1071, 2010.
- [89] E. Crippa, L. Lusa, L. De Cecco, E. Marchesi, G. A. Calin, P. Radice, S. Manoukian, B. Peissel, M. G. Daidone, M. Gariboldi, *et al.*, “mir-342 regulates brca1 expression through modulation of id4 in breast cancer,” *PloS one*, vol. 9, no. 1, p. e87039, 2014.
- [90] E. Crippa, M. Folini, M. Pennati, N. Zaffaroni, M. A. Pierotti, and M. Gariboldi, “mir-342 overexpression results in a synthetic lethal phenotype in

- brca1-mutant hcc1937 breast cancer cells,” *Oncotarget*, vol. 7, no. 14, p. 18594, 2016.
- [91] N. Petrovic, R. Davidovic, V. Bajic, M. Obradovic, and R. E. Isenovic, “MicroRNA in breast cancer: The association with brca1/2,” *Cancer Biomarkers*, vol. 19, no. 2, pp. 119–128, 2017.
- [92] S. Bhatlekar, J. Z. Fields, and B. M. Boman, “Hox genes and their role in the development of human cancers,” *Journal of molecular medicine*, vol. 92, no. 8, pp. 811–823, 2014.
- [93] F. A. Karreth and P. P. Pandolfi, “CERNA cross-talk in cancer: when ce-bling rivalries go awry,” *Cancer discovery*, vol. 3, no. 10, pp. 1113–1121, 2013.
- [94] R. Siegel, J. Ma, Z. Zou, and A. Jemal, “Cancer statistics, 2014,” *CA: a cancer journal for clinicians*, vol. 64, no. 1, pp. 9–29, 2014.
- [95] K. P. Sørensen, M. Thomassen, Q. Tan, M. Bak, S. Cold, M. Burton, M. J. Larsen, and T. A. Kruse, “Long non-coding rna hotair is an independent prognostic marker of metastasis in estrogen receptor-positive primary breast cancer,” *Breast cancer research and treatment*, vol. 142, no. 3, pp. 529–536, 2013.
- [96] J. Li, J. Wang, Y. Zhong, R. Guo, D. Chu, H. Qiu, and Z. Yuan, “Hotair: a key regulator in gynecologic cancers,” *Cancer cell international*, vol. 17, no. 1, p. 65, 2017.
- [97] P. Eroles, A. Bosch, J. A. Pérez-Fidalgo, and A. Lluch, “Molecular biology in breast cancer: intrinsic subtypes and signaling pathways,” *Cancer treatment reviews*, vol. 38, no. 6, pp. 698–707, 2012.
- [98] M. Miyan, J. Schmidt-Mende, R. Kiessling, I. Poschke, and J. Boniface, “Differential tumor infiltration by t-cells characterizes intrinsic molecular subtypes in breast cancer,” *Journal of translational medicine*, vol. 14, no. 1, p. 227, 2016.
- [99] I. Acerbi, L. Cassereau, I. Dean, Q. Shi, A. Au, C. Park, Y. Chen, J. Liphardt, E. Hwang, and V. Weaver, “Human breast cancer invasion

- and aggression correlates with ecm stiffening and immune cell infiltration,” *Integrative Biology*, vol. 7, no. 10, pp. 1120–1134, 2015.
- [100] M. J. Rutkowski, M. E. Sughrue, A. J. Kane, S. A. Mills, and A. T. Parsa, “Cancer and the complement cascade,” *Molecular Cancer Research*, vol. 8, no. 11, pp. 1453–1465, 2010.
 - [101] A. W. Lambert, S. Ozturk, and S. Thiagalingam, “Integrin signaling in mammary epithelial cells and breast cancer,” *ISRN oncology*, vol. 2012, 2012.
 - [102] Y. Peng and C. M. Croce, “The role of micrnas in human cancer,” *Signal transduction and targeted therapy*, vol. 1, p. 15004, 2016.
 - [103] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky, “Widespread changes in protein synthesis induced by micrnas,” *nature*, vol. 455, no. 7209, p. 58, 2008.
 - [104] C. P. Bracken, H. S. Scott, and G. J. Goodall, “A network-biology perspective of micrna function and dysfunction in cancer,” *Nature Reviews Genetics*, vol. 17, no. 12, p. 719, 2016.
 - [105] X. Lai, M. Eberhardt, U. Schmitz, and J. Vera, “Systems biology-based investigation of cooperating micrnas as monotherapy or adjuvant therapy in cancer,” *Nucleic acids research*, 2019.
 - [106] D. Sejdinovic, A. Gretton, and W. Bergsma, “A kernel test for three-variable interactions,” in *Advances in Neural Information Processing Systems*, pp. 1124–1132, 2013.
 - [107] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, *et al.*, “Kernel mean embedding of distributions: A review and beyond,” *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.
 - [108] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, “Micrna targets in drosophila,” *Genome biology*, vol. 5, no. 1, p. R1, 2003.

- [109] A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. Da Piedade, K. C. Gunsalus, M. Stoffel, *et al.*, “Combinatorial microrna target predictions,” *Nature genetics*, vol. 37, no. 5, pp. 495–500, 2005.
- [110] X. Lai, U. Schmitz, S. K. Gupta, A. Bhattacharya, M. Kunz, O. Wolkenhauer, and J. Vera, “Computational analysis of target hub gene repression regulated by multiple and cooperative mirnas,” *Nucleic acids research*, vol. 40, no. 18, pp. 8818–8834, 2012.
- [111] J. J. Gam, J. Babb, and R. Weiss, “A mixed antagonistic/synergistic mirna repression model enables accurate predictions of multi-input mirna sensor activity,” *Nature communications*, vol. 9, no. 1, p. 2430, 2018.
- [112] S. Wu, S. Huang, J. Ding, Y. Zhao, L. Liang, T. Liu, R. Zhan, and X. He, “Multiple micrnas modulate p21cip1/waf1 expression by directly targeting its 3 prime untranslated region,” *Oncogene*, vol. 29, no. 15, p. 2302, 2010.
- [113] J. S. Tsang, M. S. Ebert, and A. van Oudenaarden, “Genome-wide dissection of microrna functions and cotargeting networks using gene set signatures,” *Molecular cell*, vol. 38, no. 1, pp. 140–153, 2010.
- [114] M. S. Ebert and P. A. Sharp, “Roles for micrnas in conferring robustness to biological processes,” *Cell*, vol. 149, no. 3, pp. 515–524, 2012.
- [115] C. Coronello and P. V. Benos, “Comir: combinatorial microrna target prediction tool,” *Nucleic acids research*, vol. 41, no. W1, pp. W159–W164, 2013.
- [116] Y. Friedman, G. Naamati, and M. Linial, “Mirror: a combinatorial analysis web tool for ensembles of micrnas and their targets,” *Bioinformatics*, vol. 26, no. 15, pp. 1920–1921, 2010.
- [117] M. Aftabuddin, C. Mal, A. Deb, and S. Kundu, “C2analyzer: Co-target-co-function analyzer,” *Genomics, proteomics & bioinformatics*, vol. 12, no. 3, pp. 133–136, 2014.

- [118] A. V. Antonov, S. Dietmann, P. Wong, D. Lutter, and H. W. Mewes, “Gene-set2mirna: finding the signature of cooperative mirna activities in the gene lists,” *Nucleic acids research*, vol. 37, no. suppl_2, pp. W323–W328, 2009.
- [119] J. A. Davis, S. J. Saunders, M. Mann, and R. Backofen, “Combinatorial ensemble mirna target prediction of co-regulation networks with non-prediction data,” *Nucleic acids research*, vol. 45, no. 15, pp. 8745–8757, 2017.
- [120] Y.-J. Na and J. H. Kim, “Understanding cooperativity of micrnas via micrna association networks,” *BMC genomics*, vol. 14, no. 5, p. S17, 2013.
- [121] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel, “Global and local architecture of the mammalian micrna–transcription factor regulatory network,” *PLoS computational biology*, vol. 3, no. 7, p. e131, 2007.
- [122] X. Yuan, C. Liu, P. Yang, S. He, Q. Liao, S. Kang, and Y. Zhao, “Clustered micrnas’ coordination in regulating protein-protein interaction network,” *BMC systems biology*, vol. 3, no. 1, p. 65, 2009.
- [123] G. Boross, K. Orosz, and I. J. Farkas, “Human micrnas co-silence in well-separated groups and have different predicted essentialities,” *Bioinformatics*, vol. 25, no. 8, pp. 1063–1069, 2009.
- [124] J. An, K. P. Choi, C. A. Wells, and Y.-P. P. Chen, “Identifying co-regulating micrna groups,” *Journal of bioinformatics and computational biology*, vol. 8, no. 01, pp. 99–115, 2010.
- [125] J. Xu, C.-X. Li, Y.-S. Li, J.-Y. Lv, Y. Ma, T.-T. Shao, L.-D. Xu, Y.-Y. Wang, L. Du, Y.-P. Zhang, *et al.*, “Mirna–mirna synergistic network: construction via co-regulating functional modules and disease mirna topological features,” *Nucleic acids research*, vol. 39, no. 3, pp. 825–836, 2010.
- [126] Y. Li, C. Liang, K.-C. Wong, J. Luo, and Z. Zhang, “Mirsynergy: detecting synergistic mirna regulatory modules by overlapping neighbourhood expansion,” *Bioinformatics*, vol. 30, no. 18, pp. 2627–2635, 2014.

- [127] V. A. Gennarino, G. d'Angelo, G. Dharmalingam, S. Fernandez, G. Rusolillo, R. Sanges, M. Mutarelli, V. Belcastro, A. Ballabio, P. Verde, *et al.*, “Identification of microRNA-regulated gene networks by expression analysis of target genes,” *Genome research*, vol. 22, no. 6, pp. 1163–1172, 2012.
- [128] J. Ding, X. Li, and H. Hu, “MicroRNA modules prefer to bind weak and unconventional target sites,” *Bioinformatics*, vol. 31, no. 9, pp. 1366–1374, 2014.
- [129] J. Ding, X. Li, and H. Hu, “Ccmir: a computational approach for competitive and cooperative microRNA binding prediction,” *Bioinformatics*, vol. 34, no. 2, pp. 198–206, 2017.
- [130] L. S. Hon and Z. Zhang, “The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression,” *Genome biology*, vol. 8, no. 8, p. R166, 2007.
- [131] F. Wilcoxon, “Individual comparisons by ranking methods. biom bull 1: 80–83,” 1945.
- [132] H. O. Lancaster and E. Seneta, *Chi-square distribution*. Wiley Online Library, 1969.
- [133] L. Song, K. Fukumizu, and A. Gretton, “Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models,” *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 98–111, 2013.
- [134] A. Smola, A. Gretton, L. Song, and B. Schölkopf, “A hilbert space embedding for distributions,” in *International Conference on Algorithmic Learning Theory*, pp. 13–31, Springer, 2007.
- [135] K. Fukumizu, A. Gretton, G. R. Lanckriet, B. Schölkopf, and B. K. Sriperumbudur, “Kernel choice and classifiability for rkhs embeddings of probability distributions,” in *Advances in neural information processing systems*, pp. 1750–1758, 2009.

- [136] D. Garreau, W. Jitkrittum, and M. Kanagawa, “Large sample analysis of the median heuristic,” *arXiv preprint arXiv:1707.07269*, 2017.
- [137] C. G. A. R. Network *et al.*, “Comprehensive molecular characterization of clear cell renal cell carcinoma,” *Nature*, vol. 499, no. 7456, pp. 43–49, 2013.
- [138] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, “Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomaRt,” *Nature protocols*, vol. 4, no. 8, p. 1184, 2009.
- [139] A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M. P. Schroeder, A. Jene-Sanz, A. Santos, and N. Lopez-Bigas, “Intogen-mutations identifies cancer drivers across tumor types,” *Nature methods*, vol. 10, no. 11, p. 1081, 2013.
- [140] G. C. Tan, E. Chan, A. Molnar, R. Sarkar, D. Alexieva, I. M. Isa, S. Robinson, S. Zhang, P. Ellis, C. F. Langford, *et al.*, “5’ isomir variation is of functional and evolutionary importance,” *Nucleic acids research*, vol. 42, no. 14, pp. 9424–9435, 2014.
- [141] S. Li, Y. Ma, X. Hou, Y. Liu, K. Li, S. Xu, and J. Wang, “Mir-185 acts as a tumor suppressor by targeting akt1 in non-small cell lung cancer cells,” *International journal of clinical and experimental pathology*, vol. 8, no. 9, p. 11854, 2015.
- [142] S. Baskerville and D. P. Bartel, “Microarray profiling of micrnas reveals frequent coexpression with neighboring mirnas and host genes,” *Rna*, vol. 11, no. 3, pp. 241–247, 2005.
- [143] A. Mangolini, A. Bonon, S. Volinia, G. Lanza, R. Gambari, P. Pinton, G. R. Russo, L. del Senno, L. Dell’Atti, and G. Aguiari, “Differential expression of microrna501-5p affects the aggressiveness of clear cell renal carcinoma,” *FEBS Open Bio*, vol. 4, pp. 952–965, 2014.
- [144] X. Shu, M. Hildebrandt, J. Gu, N. M. Tannir, S. Matin, J. A. Karam, C. G. Wood, and X. Wu, “MicroRNA profiling in clear cell renal cell carcinoma tissues potentially links tumorigenesis and recurrence with obesity,” *British journal of cancer*, vol. 116, no. 1, p. 77, 2017.

- [145] M. Fedorko, D. Pacik, R. Wasserbauer, J. Juracek, G. Varga, M. Ghazal, and M. I. Nussir, “Micrnas in the pathogenesis of renal cell carcinoma and their diagnostic and prognostic utility as cancer biomarkers,” *The International journal of biological markers*, vol. 31, no. 1, pp. 26–37, 2016.
- [146] J. Wang, M. Lu, C. Qiu, and Q. Cui, “Transmir: a transcription factor–microRNA regulation database,” *Nucleic acids research*, vol. 38, no. suppl_1, pp. D119–D122, 2009.
- [147] S. Bandyopadhyay and M. Bhattacharyya, “Putmir: a database for extracting neighboring transcription factors of human micrnas,” *BMC bioinformatics*, vol. 11, no. 1, p. 190, 2010.
- [148] X. Hua, R. Tang, X. Xu, Z. Wang, Q. Xu, L. Chen, E. Wingender, J. Li, C. Zhang, and J. Wang, “mirtrans: a resource of transcriptional regulation on micrnas for human cell lines,” *Nucleic acids research*, vol. 46, no. D1, pp. D168–D174, 2017.
- [149] E. P. Consortium *et al.*, “An integrated encyclopedia of dna elements in the human genome,” *Nature*, vol. 489, no. 7414, p. 57, 2012.
- [150] M. Lawrence, W. Huber, H. Pages, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey, “Software for computing and annotating genomic ranges,” *PLoS computational biology*, vol. 9, no. 8, p. e1003118, 2013.
- [151] B. Otlu, C. Firtina, S. Keleş, and O. Tastan, “Glanet: genomic loci annotation and enrichment tool,” *Bioinformatics*, vol. 33, no. 18, pp. 2818–2828, 2017.
- [152] A. Khan, O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. van der Lee, A. Bessy, J. Cheneby, S. R. Kulkarni, G. Tan, *et al.*, “Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework,” *Nucleic acids research*, vol. 46, no. D1, pp. D260–D266, 2017.

- [153] I. V. Kulakovskiy, I. E. Vorontsov, I. S. Yevshin, R. N. Sharipov, A. D. Fedorova, E. I. Rumynskiy, Y. A. Medvedeva, A. Magana-Mora, V. B. Bajic, D. A. Papatsenko, *et al.*, “Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis,” *Nucleic acids research*, vol. 46, no. D1, pp. D252–D259, 2017.
- [154] C. E. Grant, T. L. Bailey, and W. S. Noble, “Fimo: scanning for occurrences of a given motif,” *Bioinformatics*, vol. 27, no. 7, pp. 1017–1018, 2011.
- [155] D. J. Sanchez and M. C. Simon, “Transcriptional control of kidney cancer,” *Science*, vol. 361, no. 6399, pp. 226–227, 2018.
- [156] S. L. Skala, H. Xiao, A. M. Udager, S. M. Dhanasekaran, S. Shukla, Y. Zhang, C. Landau, L. Shao, D. Roulston, L. Wang, *et al.*, “Detection of 6 tfeb-amplified renal cell carcinomas and 25 renal cell carcinomas with mitf translocations: systematic morphologic analysis of 85 cases evaluated by clinical tfe3 and tfeb fish assays,” *Modern Pathology*, vol. 31, no. 1, p. 179, 2018.
- [157] A. Jeppesen, H. Jensen, F. Donskov, N. Marcussen, and H. Von Der Maase, “Hyponatremia as a prognostic and predictive factor in metastatic renal cell carcinoma,” *British journal of cancer*, vol. 102, no. 5, p. 867, 2010.
- [158] E. Oosterwijk and R. Gillies, “Targeting ion transport in cancer,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1638, p. 20130107, 2014.
- [159] I. Z. Nagy, G. Lustyik, V. Z. Nagy, B. Zarandi, and C. Bertoni-Freddari, “Intracellular na⁺: K⁺ ratios in human cancer cells as revealed by energy dispersive x-ray microanalysis,” *The Journal of cell biology*, vol. 90, no. 3, pp. 769–777, 1981.
- [160] M. Lv, S. Li, C. Luo, X. Zhang, Y. Shen, Y. Sui, F. Wang, X. Wang, J. Yang, P. Liu, *et al.*, “Angiomotin promotes renal epithelial and carcinoma cell proliferation by retaining the nuclear yap,” *Oncotarget*, vol. 7, no. 11, p. 12393, 2016.

- [161] C. Yi, Z. Shen, A. Stemmer-Rachamimov, N. Dawany, S. Troutman, L. C. Showe, Q. Liu, A. Shimono, M. Sudol, L. Holmgren, *et al.*, “The p130 isoform of angiomin is required for yap-mediated hepatic epithelial cell proliferation and tumorigenesis,” *Sci. Signal.*, vol. 6, no. 291, pp. ra77–ra77, 2013.
- [162] M. Lv, Y. Shen, J. Yang, S. Li, B. Wang, Z. Chen, P. Li, P. Liu, and J. Yang, “Angiomin family members: oncogenes or tumor suppressors?,” *International journal of biological sciences*, vol. 13, no. 6, p. 772, 2017.
- [163] L. Wyler, C. Napoli, B. Ingold, T. Sulser, M. Heikenwlder, P. Schraml, and H. Moch, “Brain metastasis in renal cancer patients: metastatic pattern, tumour-associated macrophages and chemokine/chemoreceptor expression,” *British journal of cancer*, vol. 110, no. 3, p. 686, 2014.
- [164] J. Zhang, T. D. Le, L. Liu, J. He, and J. Li, “Identifying mirna synergistic regulatory networks in heterogeneous human data via network motifs,” *Molecular bioSystems*, vol. 12, no. 2, pp. 454–463, 2016.
- [165] J. M. Engreitz, J. E. Haines, E. M. Perez, G. Munson, *et al.*, “Local regulation of gene expression by lncrna promoters, transcription and splicing,” *Nature*, vol. 539, no. 7629, p. 452, 2016.
- [166] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn, “Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses,” *Genes & development*, vol. 25, no. 18, pp. 1915–1927, 2011.
- [167] S. Luo, J. Y. Lu, L. Liu, Y. Yin, C. Chen, X. Han, B. Wu, R. Xu, W. Liu, P. Yan, *et al.*, “Divergent lncrnas regulate gene expression and lineage differentiation in pluripotent cells,” *Cell Stem Cell*, vol. 18, no. 5, pp. 637–652, 2016.
- [168] P. J. Batista and H. Y. Chang, “Long noncoding rnas: cellular address codes in development and disease,” *Cell*, vol. 152, no. 6, pp. 1298–1307, 2013.

- [169] J. T. Lee and M. S. Bartolomei, “X-inactivation, imprinting, and long non-coding rnas in health and disease,” *Cell*, vol. 152, no. 6, pp. 1308–1323, 2013.
- [170] B. Otlu, C. Firtina, S. Keles, and O. Tastan, “Glanet: Genomic loci annotation and enrichment tool,” *Bioinformatics*, 2017.
- [171] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, *et al.*, “Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals,” *Nature*, vol. 458, no. 7235, pp. 223–227, 2009.
- [172] Y. Li, H. Chen, T. Pan, C. Jiang, Z. Zhao, Z. Wang, J. Zhang, J. Xu, and X. Li, “Lncrna ontology: inferring lncrna functions based on chromatin states and expression patterns,” *Oncotarget*, vol. 6, no. 37, p. 39793, 2015.
- [173] K. Liu, Z. Yan, Y. Li, and Z. Sun, “Linc2go: a human lincrna function annotation resource based on cerna hypothesis,” *Bioinformatics*, vol. 29, no. 17, pp. 2221–2222, 2013.
- [174] Z. Zhao, J. Bai, A. Wu, Y. Wang, *et al.*, “Co-lncrna: investigating the lncrna combinatorial effects in go annotations and kegg pathways based on human rna-seq data,” *Database*, vol. 2015, p. bav082, 2015.
- [175] A. Zehir, R. Benayed, R. H. Shah, A. Syed, S. Middha, H. R. Kim, P. Srinivasan, J. Gao, D. Chakravarty, S. M. Devlin, *et al.*, “Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients,” *Nature medicine*, vol. 23, no. 6, p. 703, 2017.
- [176] Z.-W. Guo, C. Xie, K. Li, X.-M. Zhai, G.-X. Cai, X.-X. Yang, and Y.-S. Wu, “Seler: a database of super-enhancer-associated lncrna-directed transcriptional regulation in human cancers,” *Database*, vol. 2019, 2019.
- [177] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, “limma powers differential expression analyses for rna-sequencing and microarray studies,” *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

- [178] K. R. Blahnik, L. Dou, H. O’Geen, T. McPhillips, X. Xu, A. R. Cao, S. Iyengar, C. M. Nicolet, B. Ludäscher, I. Korf, *et al.*, “Sole-search: an integrated analysis program for peak detection and functional annotation using chip-seq data,” *Nucleic acids research*, vol. 38, no. 3, pp. e13–e13, 2009.
- [179] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, “Gosemsim: an r package for measuring semantic similarity among go terms and gene products,” *Bioinformatics*, vol. 26, no. 7, pp. 976–978, 2010.
- [180] J. C. Dunn, “Well-separated clusters and optimal fuzzy partitions,” *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.
- [181] G. S. Martin, “Cell signaling and cancer,” *Cancer cell*, vol. 4, no. 3, pp. 167–174, 2003.
- [182] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, *et al.*, “Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling,” *Nature*, vol. 403, no. 6769, p. 503, 2000.
- [183] G. Wang, Y. Gao, L. Li, G. Jin, J.-I. Chao, H.-K. Lin, *et al.*, “K63-linked ubiquitination in kinase activation and cancer,” *Frontiers in oncology*, vol. 2, p. 5, 2012.
- [184] R. J. DeBerardinis and N. S. Chandel, “Fundamentals of cancer metabolism,” *Science advances*, vol. 2, no. 5, p. e1600200, 2016.
- [185] C. S. Ahn and C. M. Metallo, “Mitochondria as biosynthetic factories for cancer proliferation,” *Cancer & metabolism*, vol. 3, no. 1, p. 1, 2015.
- [186] R. J. Boissy, M. A. Watson, D. M. Umbach, M. Deakin, J. Elder, R. C. Strange, and D. A. Bell, “A pilot study investigating the role of nat1 and nat2 polymorphisms in gastric adenocarcinoma,” *International journal of cancer*, vol. 87, no. 4, pp. 507–511, 2000.

- [187] T. Katoh, R. Boissy, N. Nagata, K. Kitagawa, Y. Kuroda, H. Itoh, T. Kawamoto, and D. A. Bell, “Inherited polymorphism in the n-acetyltransferase 1 (nat1) and 2 (nat2) genes and susceptibility to gastric and colorectal adenocarcinoma,” *International journal of cancer*, vol. 85, no. 1, pp. 46–49, 2000.
- [188] K. Shibuta, T. Eto, A. Kataoka, H. Inoue, H. Ueo, T. Suzuki, G. F. Barnard, and M. Mori, “Genetic polymorphism of n-acetyltransferase 2 in patients with esophageal cancer,” *The American journal of gastroenterology*, vol. 96, no. 12, pp. 3419–3424, 2001.
- [189] H. Miyazaki, K. Takabe, and W. A. Yeudall, “Chemokines, chemokine receptors and the gastrointestinal system,” *World Journal of Gastroenterology: WJG*, vol. 19, no. 19, p. 2847, 2013.
- [190] H. Murase, M. Inokuchi, Y. Takagi, K. Kato, K. Kojima, and K. Sugihara, “Prognostic significance of the co-overexpression of fibroblast growth factor receptors 1, 2 and 4 in gastric cancer,” *Molecular and clinical oncology*, vol. 2, no. 4, pp. 509–517, 2014.
- [191] B. A. Krantz, N. Dave, K. M. Komatsubara, B. P. Marr, and R. D. Carvajal, “Uveal melanoma: epidemiology, etiology, and treatment of primary disease,” *Clinical ophthalmology (Auckland, NZ)*, vol. 11, p. 279, 2017.
- [192] R. D. Carvajal, G. K. Schwartz, T. Tezel, B. Marr, J. H. Francis, and P. D. Nathan, “Metastatic disease from uveal melanoma: treatment options and future prospects,” *British Journal of Ophthalmology*, pp. bjophthalmol–2016, 2016.
- [193] G. Zheng, Y. Ma, Y. Zou, A. Yin, W. Li, and D. Dong, “Hcmdb: the human cancer metastasis database,” *Nucleic acids research*, vol. 46, no. D1, pp. D950–D955, 2017.
- [194] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2018,” *CA: a cancer journal for clinicians*, vol. 68, no. 1, pp. 7–30, 2018.

- [195] L. R. Zarour, S. Anand, K. G. Billingsley, W. H. Bisson, A. Cercek, M. F. Clarke, L. M. Coussens, C. E. Gast, C. B. Geltzeiler, L. Hansen, *et al.*, “Colorectal cancer liver metastasis: evolving paradigms and future directions,” *Cellular and molecular gastroenterology and hepatology*, vol. 3, no. 2, pp. 163–173, 2017.
- [196] S. Pott and J. D. Lieb, “What are super-enhancers?,” *Nature genetics*, vol. 47, no. 1, p. 8, 2015.
- [197] A. Khan and X. Zhang, “dbsuper: a database of super-enhancers in mouse and human genome,” *Nucleic acids research*, vol. 44, no. D1, pp. D164–D171, 2015.
- [198] A. Horii, T. Kobayashi, N. Tomita, T. Yamamoto, S. Fukushige, T. Murotsu, M. Ogawa, T. Mori, and K. Matsubara, “Primary structure of human pancreatic secretory trypsin inhibitor (psti) gene,” *Biochemical and biophysical research communications*, vol. 149, no. 2, pp. 635–641, 1987.
- [199] S. A. Tomlins, D. R. Rhodes, J. Yu, S. Varambally, R. Mehra, S. Perner, F. Demichelis, B. E. Helgeson, B. Laxman, D. S. Morris, *et al.*, “The role of spink1 in ets rearrangement-negative prostate cancers,” *Cancer cell*, vol. 13, no. 6, pp. 519–528, 2008.
- [200] “National center for biotechnology information, pubchem compound database; cid=6140.”
- [201] J. G. Gurney, J. M. Pogoda, E. A. Holly, S. S. Hecht, and S. Preston-Martin, “Aspartame consumption in relation to childhood brain tumor risk: results from a case-control study,” *Journal of the National Cancer Institute*, vol. 89, no. 14, pp. 1072–1074, 1997.
- [202] J. W. Olney, N. B. Farber, E. Spitznagel, and L. N. Robins, “Increasing brain tumor rates: is there a link to aspartame?,” *Journal of Neuropathology & Experimental Neurology*, vol. 55, no. 11, pp. 1115–1123, 1996.
- [203] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, and Q. Cui, “Lncrnadisease: a database for long-non-coding rna-associated diseases,” *Nucleic acids research*, vol. 41, no. D1, pp. D983–D986, 2012.

- [204] S. Ning, J. Zhang, P. Wang, H. Zhi, J. Wang, Y. Liu, Y. Gao, M. Guo, M. Yue, L. Wang, *et al.*, “Lnc2cancer: a manually curated database of experimentally supported lncrnas associated with various human cancers,” *Nucleic acids research*, vol. 44, no. D1, pp. D980–D985, 2015.
- [205] J. Han, S. Liu, Z. Sun, Y. Zhang, F. Zhang, C. Zhang, D. Shang, H. Yang, F. Su, Y. Xu, *et al.*, “Lncrnas2pathways: Identifying the pathways influenced by a set of lncrnas of interest based on a global network propagation method,” *Scientific reports*, vol. 7, p. 46566, 2017.
- [206] A. Thévenin, L. Ein-Dor, M. Ozery-Flato, and R. Shamir, “Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome,” *Nucl. Acids Res.*, vol. 42, no. 15, pp. 9854–9861, 2014.
- [207] D. G. Lupiáñez, M. Spielmann, and S. Mundlos, “Breaking tads: how alterations of chromatin domains result in disease,” *Trends in Genet.*, vol. 32, no. 4, pp. 225–237, 2016.
- [208] M. Ahmed, H. Nguyen, T. Lai, and D. R. Kim, “mircancerdb: a database for correlation analysis between microrna and gene expression in cancer,” *BMC research notes*, vol. 11, no. 1, p. 103, 2018.
- [209] J. R. Dixon, D. U. Gorkin, and B. Ren, “Chromatin domains: the unit of chromosome organization,” *Molecular cell*, vol. 62, no. 5, pp. 668–680, 2016.
- [210] Y. Wang, F. Song, B. Zhang, L. Zhang, J. Xu, D. Kuang, D. Li, M. N. Choudhary, Y. Li, M. Hu, *et al.*, “The 3d genome browser: a web-based browser for visualizing 3d genome organization and long-range chromatin interactions,” *Genome biology*, vol. 19, no. 1, p. 151, 2018.
- [211] F. Ramírez, V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Grüning, J. Villaveces, B. Habermann, A. Akhtar, and T. Manke, “High-resolution tads reveal dna sequences underlying genome organization in flies,” *Nature communications*, vol. 9, no. 1, p. 189, 2018.

- [212] C. Shin, J.-W. Nam, K. K.-H. Farh, H. R. Chiang, A. Shkumatava, and D. P. Bartel, “Expanding the microRNA targeting code: functional sites with centered pairing,” *Molecular cell*, vol. 38, no. 6, pp. 789–802, 2010.
- [213] C. E. Vejnar and E. M. Zdobnov, “Mirmap: comprehensive prediction of microRNA target repression strength,” *Nucleic acids research*, vol. 40, no. 22, pp. 11673–11683, 2012.
- [214] 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.
- [215] J. G. Ruby, A. Stark, W. K. Johnston, M. Kellis, D. P. Bartel, and E. C. Lai, “Evolution, biogenesis, expression, and target predictions of a substantially expanded set of drosophila microRNAs,” *Genome research*, vol. 17, no. 12, pp. 000–000, 2007.
- [216] C. H. Jan, R. C. Friedman, J. G. Ruby, and D. P. Bartel, “Formation, regulation and evolution of caenorhabditis elegans 3 utrs,” *Nature*, vol. 469, no. 7328, p. 97, 2011.
- [217] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, “Gene ontology: tool for the unification of biology,” *Nature genetics*, vol. 25, no. 1, p. 25, 2000.
- [218] Z. Yang, L. Wu, A. Wang, W. Tang, Y. Zhao, H. Zhao, and A. E. Teschendorff, “dbdemc 2.0: updated database of differentially expressed mirnas in human cancers,” *Nucleic acids research*, vol. 45, no. D1, pp. D812–D818, 2016.
- [219] U. Brandes, D. Dellling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, “On modularity clustering,” *IEEE transactions on knowledge and data engineering*, vol. 20, no. 2, pp. 172–188, 2008.
- [220] G. Csardi and T. Nepusz, “The igraph software package for complex network research,” *InterJournal*, vol. Complex Systems, p. 1695, 2006.

- [221] W. Ligtenberg, “reactome.db,” 2019. reactome.db: A set of annotation maps for reactome. R package version 1.68.0.
- [222] M. Carlson, “kegg.db,” 2016. KEGG.db: A set of annotation maps for KEGG. R package version 3.2.3.
- [223] S. Tiwary, J. E. Morales, S. C. Kwiatkowski, F. F. Lang, G. Rao, and J. H. McCarty, “Metastatic brain tumors disrupt the blood-brain barrier and alter lipid metabolism by inhibiting expression of the endothelial cell fatty acid transporter mfsd2a,” *Scientific reports*, vol. 8, no. 1, p. 8267, 2018.
- [224] D. Tamborero, C. Rubio-Perez, J. Deu-Pons, M. P. Schroeder, A. Vivancos, A. Rovira, I. Tusquets, J. Albanell, J. Rodon, J. Tabernero, *et al.*, “Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations,” *Genome medicine*, vol. 10, no. 1, p. 25, 2018.