INTEGRATIVE NETWORK APPROACHES TO UNDERSTANDING MICRORNA AND GENE EXPRESSIONS IN CANCER

by

Benika Chenae Hall

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics & Computational Biology

Charlotte

2017

Approved by:

Dr. Xinghua Shi

Dr. Jun-tao Guo

Dr. Anthony Fodor

Dr. Jessica Schlueter

Dr. Erik Saule

ABSTRACT

BENIKA CHENAE HALL. Integrative Network Approaches to Understanding MicroRNA and Gene Expressions in Cancer. (Under the direction of DR. XINGHUA SHI)

Cancer is among the leading cause of deaths in men and women today. Regulation of gene expression by microRNAs (miRNA) has been linked to cancer progression in recent years. Therefore there has been a growing interest understanding how miRNAs mediate gene expression in cancer. In this dissertation we aimed to construct an integrative miRNA-gene network to understand how miRNAs affect gene expression and their downstream genetic neighborhoods in ovarian cancer and colon cancer. To conduct this research, we applied different techniques to determine how miRNAs impact downstream genetic neighborhoods and identify cancer enriched neighborhoods in ovarian and colon cancer. First, we developed a community based method utilizing the spin-glass model to construct an integrative miRNA-gene network. Second, we developed a label propagation framework to construct a multi-layer miRNA-gene network to exploit the downstream effects of miRNAs throughout integrating multiple networks. We identified multiple communities enriched in cancer-driven pathways across multiple networks in ovarian cancer. We also uncovered enriched genetic neighborhoods and identified key network signatures in patients with differ-

ent pathological stages in colon cancer. The methods developed provide a better outlook on how miRNAs affect gene expression and their downstream genetic neighborhoods, which will improve our understanding of their role in tumorigenesis and cancer progression.

ACKNOWLEDGMENTS

First and foremost, I would like to thank God for all of his tremendous glory he has bestowed upon me and my family during this journey.

I want to thank my advisor for her guidance, support and encouragement throughout the years. She has been an excellent advisor, a great leader, an honorable mentor and without her support and knowledge, this dissertation would not have been feasible. I'm truly grateful for the researcher she has helped me become.

I sincerely want to thank my committee members - Dr. Jun-tao Guo, Dr. Jessica Schlueter, Dr. Anthony Fodor and Dr. Erik Saule for committing to this task and providing invaluable feedback and suggestions. I greatly appreciate their collective knowledge and support. I also want to thank Dr. Shannon Schlueter who convinced me to pursue the Ph.D program. A special thank you to Dr. Dennis Livesay for all of his support, kind words of encouragement and advisement.

I really want to give a special thank you to our members in Shi Lab. They make a great team and are hard to replace. Special thank you to Jia Wen and Andrew Quitadamo.

I want to thank my parents and siblings. Although they are far away, they have also played an intricate part in this accomplishment. My parents have been supportive

of all of my academic achievements. To my sisters, their belief in me has always kept me focused. The numerous texts of encouraging words have lifted me during many long nights. I'd like to thank my aunt Linda for encouraging me every chance she had. Special thank you to my mother-in-law for her support. Special thank you to my dear friends, Shetocqua McKoy, Jessye Bemley and Vivian Lee. For all of their love and support, I am always grateful.

I express my deepest gratitude to my husband Chris and our son Desmond. This dissertation would not be possible if not for the patience, love and support they have shown. My husband has been the source of my confidence and strength in pursuing this PhD. Our son has made this journey worthy.

Finally, I want to thank the Department of Bioinformatics and Genomics for the opportunity to pursue a Ph.D. in Bioinformatics. I also want to thank the Graduate Assistant Support Plan (GASP) for the financial support. Special thank you to all of the faculty and staff.

DEDICATION

This dissertation is dedicated to my husband and son.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Introduction

Human genomics is the study of all of a person's genes, in which a gene is a set of deoxyribonucleic acid(DNA). Each person has a genome comprised of the same set of genes. However, what makes each person different are the minor variations in their genes. There are approximately 30,000 genes in the human genome [1]. Each of these genes carry information which is used in a process called gene expression to synthesize a gene product. Often, the created products are proteins, which dictate the functions and actions of a cell. Therefore, the thousands of genes that are expressed in a single cell can dictate the role of the cell [2].

Gene expression is a complex trait or phenotype. There are several factors that influence gene expression such as environmental, genetic and epigenetic factors [3]. Many complex diseases are caused by the combination of these factors rather than the expression of a single gene. In fact, most human diseases are caused by mutations in multiple genes, meaning they are multigenic [4]. The regulation of gene expression is important in complex diseases such as diabetes, Alzheimer's disease and

cancers. There are many mechanisms that can regulate gene expression including but not limited to, chromatin modification [4], alternative splicing, post-translational modifications and protein degradation [3].

In genomics research, we seek to understand the complexity of gene regulation in human disease. MicroRNAs (miRNAs) have been reported to regulate gene expression [5–7]. MiRNAs are small non-coding ribonucleic acid (RNA) molecules approximately 22 base-pairs in length [5, 8, 9]. MicroRNAs are also involved several biological processes including, cell proliferation, differentiation, and apoptosis [5, 10, 11]. However, it remains unclear how these small molecules regulate gene expression in complex diseases.

1.2 Problem Statement

There is evidence that suggests that miRNAs play a pivotal role in many biological processes. In fact, exhaustive research has shown that miRNAs are key regulators of gene expression [12–18]. However, it remains unclear exactly how miRNAs affect gene expression. Because of this gap, there has been a growing trend of using network analysis to study the complex relationships between miRNAs and gene expression [19–23]. Based on the existing body of research on constructing miRNA-gene networks, there is a gap in capturing the complexity of true genetic networks.

Although there are methods that have proposed an integrative approach to con-

struct miRNA-gene networks with miRNA expression and gene expression [22], these methods do not consider the correlations between miRNAs and the correlations between genes. These correlations in addition to protein-protein interaction networks add to complexity of genetic networks. Furthermore, these methods do not consider the community structure and how miRNAs affect the downstream communities. That said, Zhang et al. [19] proposed an alternative approach to include individual miRNA co-expression networks and gene co-expression networks. However, this kind approach is limited due to the fact that it does not capture the true heterogeneous nature of genetic networks.

It is worth noting that advancements have been made to address this limitation by integrating heterogeneous data such as miRNAs, miRNA-target genes, protein-protein interactions and gene expression [20,21,24]. This kind of integrative approach is important in understanding the biological functions of genetic networks in complex diseases. While the majority of these integrative methods have drawn attention to integrative network analysis, there exists a need to use such methods to expand our knowledge of disease progression in cancer subtypes.

Given the gap described in the sections above, this dissertation proposes new methodologies to construct an integrative miRNA-gene network in two cancer subtypes. There are three goals of this dissertation: 1. To construct an integrative

miRNA-gene network in ovarian serous cystadenocarcinoma (ovarian cancer) using a community-based network expansion method to explore the relationships between miRNAs and downstream genes in ovarian cancer networks. 2. To construct a multi-layer miRNA-gene network using a label propagation approach and capture the complexity of heterogeneous genetic networks. 3. Examine the relationships between miRNA-gene networks and the pathologic stages in colon adenocarcinoma (colon cancer). As explained, the work in this study aims to close the gap in miRNA-gene network construction and expand our knowledge of the functional mechanisms of miRNA-gene networks in ovarian and colon cancer.

1.3 Research Aims

As previously stated, our research aims are as follows:

* Our first research aim is to investigate the relationship between miRNAs and gene expression in ovarian cancer by constructing an integrative miRNA-gene network via community-based network expansion.

* Our second research aim is to improve our network expansion methodology by integrating multiple networks to identify downstream relationships between miRNA expression, gene expression, protein-protein networks, gene regulatory networks and the downstream effects on genetic networks in patients with ovarian cancer.

* Our third research aim is to utilize our methods to explore the relationship

between miRNA-gene networks and the pathologic stages of colon cancer and expand our knowledge by identifying prognostic network signatures in enriched communities and patients with stage II and stage III colon cancer.

1.4 Research Contributions

To support the advancement of genomics research in the area of integrative miRNA-gene network analysis, the proposed work will address the aforementioned challenges in the following ways:

• **A robust framework to improve integrative miRNA-gene network expansion.** This research provides an integrative framework capable of constructing an integrated miRNA-gene network via community-based network expansion. The method also extends the current knowledge of miRNA-gene relationships in downstream subnetworks in ovarian cancer.

•**A powerful framework for integrating multiple networks.** This research provides an efficient and scalable method for integrating more than 2 biological networks. This method also contributes a set of functional groups containing potential biomarkers and potential candidate genes in ovarian cancer for the research community.

•**Prognostic signatures in colon cancer.** The results generated in this research provide potentially new targets for cancer therapeutics in colon cancer. Results reveal

network signatures related to different stages of cancer progression to help improve diagnosis and early detection.

1.5 Significance

My interest in this area developed while I was conducting research on gene networks in pancreatic cancer. Years later, as we began to design the studies explained in this dissertation, we noticed the gaps in the methods for constructing integrative networks in cancer and that has driven this research. This work sheds new light on methods for constructing integrative miRNA-gene networks in cancer. Further, this work provides the opportunity to expand our knowledge of functional enrichment of complex miRNA-gene networks in cancer subtypes. Moreover, the findings should also make an important contribution to the field of prognostic markers in colon cancer. To the future researchers, this work can provide a baseline methodology for integrating heterogeneous genetic networks to explore the relationships between miRNAs and various regulatory elements.

CHAPTER 2: BACKGROUND

2.1 Introduction

The recent discovery of miRNAs has added to the complexity of gene regulation. In fact, deregulated microRNA(miRNA) expression has been found in many human diseases including colon adenocarcinoma(colon cancer) [25–27]. That said, it has become more challenging to understand how miRNAs affect downstream genes in genetic networks in cancer. Therefore, a great deal of previous research into genetic networks has focused on constructing miRNA-gene networks [19–23]. Using this approach, researchers have been able to identify relationships between miRNAs and gene expression. Much of the current literature on miRNA-gene networks pays particular attention to the direct relationship between miRNAs and genes [20, 28], but neglect to consider the miRNA-miRNA correlations and gene-gene correlations. To understand why this is neglected, we must examine the different methodologies developed in this topic. As mentioned in the previous chapter, many of the current methods do not capture the complexity of heterogeneous genetic networks. In addition, it is important to consider the community structure of genetic networks because

communities contain miRNAs or genes with similar functions. Therefore, in the next section, we review the literature on community structure in biological networks.

2.2 Community structure in biological networks

Over the years, we have learned that underlying mechanisms can be investigated by learning the community structure in biological networks [29–34]. This exhaustive research in community structure has increased interest in determining the community structure in biological networks as it may reveal insights into the functional components of a network system. The broad use of the term "community" is defined as a hub of densely connected nodes in which each node has more connections within the community than with the rest of the network [29, 35, 36]. In a community, the internal degree is the sum of the internal edges and vertices within the community. The external degree is the sum of the edges to the vertices out of the community. Community structure in biological systems can provide insight on how functional groups in complex networks affect the topology. In biological networks, communities represent functional groups such as genes or proteins that have similar functions and are involved in similar cellular processes [37]. Now that we know what the community structure is, let us consider how to detect communities by investigating the current methods of community detection.

2.3 Community detection methods

In this section, we will review various community detection methods as community detection is key to understanding the structure of complex networks. As indicated in the previous section, community detection methods are of vital in understanding the functional patterns in biological systems. Identifying communities in biological networks can be a challenging task. Most research on community detection has been carried out utilizing modularity-based based methods. While modularity-based methods have been the most popular [32,36,38,39], these methods have a key disadvantage in real-world networks as they are unable to handle noise perturbations in the network [40, 41]. This limitation is due the resolution limit of modularity which is an intrinsic scale that depends on the interconnectedness between node pairs of communities [42]. Fortunato et al. [40] found that the natural community structure is not detected by modularity-based optimization if the communities are smaller than the resolution limit of the graph. Furthermore, modularity optimization could miss important communities in real-world networks. With respect to this information, an alternative to these methods is to use a spin-glass model approach from statistical mechanics. The spin-glass model is a semi-supervised learning approach that is coupled with simulated annealing [41, 43, 44]. It is far less computationally intensive compared to other methods solely based on modularity [38]. Due to the large

size of our network, the computational efficiency makes this algorithm attractive for large-scale networks.

In the sections to follow, we review community detection methods. We will discuss the approaches used in each method. Regarding each community detection method, we also determine the advantages and disadvantages of each method as it pertains to the field and the work in this dissertation.

2.3.1 Edge-betweenness community detection

As indicated previously, many community detection methods measure the quality of a community using the modularity score. This notion was first introduced by Newman and Girvan [45]. This method is now one of the most popular methods in community detection. In this approach, Newman and Girvan introduced the notion of modularity which is a measure of the quality of a community. In terms of the algorithm, the edge-betweenness algorithm is a hierarchical decomposition method that removes edges based on the descending order of the edge-betweenness score [29, 34, 45]. To calculate the edge-betweenness score, one must calculate the number of shortest paths that interact with a given edge. The theory behind this methodology is that edges connecting different communities are likely to be involved in many shortest paths because they are often the only option connecting the groups. Having defined what is entailed in the algorithm, we will now move on to discuss the

advantages and disadvantages of this method.

We first examine the edge-betweenness method because it is highly effective in discovering community structure in real biological networks [31, 45]. Another advantage is that the accuracy of the edge-betweenness algorithm is not dependent on the size of the network [35], which makes it applicable to large-scale biological networks. Further review of literature showed that while this method is effective, it does have key limitations. Although this method is popular and it provides good results, it has a high cost of computational cost. This is because each time an edge-betweenness score is re-calculated after every edge removal step. One disadvantage of this method is that it is limited in the size of the input network. This limitation is a major drawback in community detection of large-scale biological networks. Another disadvantage is that it yields a full dendrogram, but without guidance on where to cut the dendrogram in order to obtain the final communities. One way to identify the final communities is to calculate the modularity score of each partition in the graph. Although the method would yield good results for a limited size network, it is not recommended for networks involving multiple layers and heterogeneous data types. With these limitations in mind, we review other methods that have a reasonable computation speed in the following sections.

2.3.2 Leading eigenvector community detection

Regarding the limitations discussed in the previous section, we now review methods with better computational performance. This leads us to another popular community detection method known as the leading eigenvector community detection. This method is a top-down hierarchical method that optimizes the modularity function [35]. To understand how this algorithm works, let us discuss the steps involved to divide the graph.

During the process, the graph is divided into two parts such that this division increases the modularity of the graph [45]. In order to partition the graph, each split is based evaluating the leading eigenvector of the modularity matrix, which is similar to the Laplacian matrix used in spectral partitioning. To complete the partitioning, the eigenvector corresponding to the most positive eigenvalues in the modularity matrix is chosen and then divided inn to two groups based on the sign of the eigenvalue. This method works well and is not constrained to finding communities of a specific size. The benefit of this method is that there is no loss of information from the remainder of the eigenvectors. A disadvantage of this method is the computational expense to split the network into communities. Further investigation into the leading eigenvector method revealed more limitations. Yang et al [35] showed that while this method has a more reasonable computation speed, it underestimates the number

of communities in large-scale networks. Moreover, the accuracy of this method decreases as the size of the network increases. In regards to the diminishing accuracy of the leading eigenvector method in large networks, we will now review methods that address this limitation.

2.3.3 Info-map community detection

As previously indicated, the leading eigenvector method loses accuracy in large networks. Therefore, we investigate the Info-map community detection method, which is based on the flow of information and minimizes the map equation [35, 46]. Unlike the previously mentioned methods, the Info-map method utilizes a random walk to analyze the flow of information in the network [47].To be more precise, this method divides the network into communities by determining the shortest description lengths for a random walk on the network. For each community detected, there is an association cost that describes the path of the random walk movements. The splits in the graph with the shortest description length (information flow) represents the community structure in the network. The map equation is minimized over all partitions in the network. Although there is a lacking presence of research using this method, a comparative study by Yang et al. [35] demonstrated that Info-map performs better on smaller networks. This is a disadvantage for analyzing large biological networks. Furthermore, similar to the leading eigenvector approach, it overestimates the num-

ber of communities in larger networks. Regarding the limitations mentioned in this section, we turn our attention to methods that focus on quality of the community structure in the following sections.

2.3.4 Walk-trap community detection

Most recently, studies have shown that with methods that excel in computational time, there is a trade-off in the quality of the community structure [48]. Therefore, we investigate the walk-trap community detection which is a bottom-up approach based on random walks. The concept of this method is that the walks are more likely to stay within the same community because there are a small number of edges leading outside of the community. It merges the communities of walks consisting of length 3-4-5. One disadvantage of this method is that it is much slower and therefore not ideal for large scale biological networks. The determine the final communities, the modularity score can be used to select where to partition the network. One of the major advantages of the walk-trap community detection method is that it has better accuracy on larger networks. On the other hand, a disadvantage is that the computation speed does not scale with large networks [35, 48]. While this method may seem suitable for mid-size networks, the work in this dissertation requires much larger networks and may be far more complex for this method. That said, we will discuss methods that have been proposed for more complex real-world networks in

the sections to follow.

2.3.5 Spin-glass community detection

The spin-glass model is a statistical mechanics model in which each spin state (of the spin-glass) exclusively interacts with the adjacent spin states based on an interaction energy depending on the similarity of the spins. Using the spin-glass model, identifying community indices is equivalent to identifying the spin states of a spin glass that lead to the ground state of the system. In community detection, the ground state would be the natural structure of a network partitioned into communities. The model is based on optimizing the Hamiltonian energy function in Equation 2.3.5 [41] such that it minimizes the energy of the system and maximizes the modularity in Equation 2.3.5 [40].

$$H(\sigma) = (A_{i,j}\gamma p_{i,j})\delta(\sigma_m, \sigma_n), \tag{1}$$

where $A_{i,j}$ is the adjacency matrix consisting of nodes i and j; γ represents the weights and $p_{i,j}$ represents the probability of an edge between nodes i and j. Thus, the adjacency matrix represents the relationships between nodes and edges based on the weights determined by γ. The Kronecker delta function formulates the energy; in this part of the equation, σ_m and σ_n represent the spin states or community indices in which nodes i and j belong to in the network. The Kronecker delta is a binary value

of 1 if the nodes represented by σ_m and σ_n are in the same community, otherwise it is 0. In reference to σ, m and n represent each community index. This function is updated and optimized throughout the community detection process.

Modularity can be defined in the following equation: 2.3.5.

$$Q = \frac{1}{m}\sum_{i,j}(A_{i,j} - \frac{k_i k_j}{m})\delta(c_i, c_j) \tag{2}$$

,

where m is the total weight of the graph, $A_{i,j}$ is the element in the adjacency matrix and C is the set of all communities. c_i, c_j is a community in C, k_i,k_j are the degrees of nodes i, j. The Kronecker $\delta(c_i,c_j)$ is 1 if the nodes i and j are in the same community and 0 if they are not in the same community. In order to calculate the Q, we iterate through each element in the adjacency matrix.

In optimizing the spin-glass model, the minimum number of edges required to partition a graph can be retrieved from the ground-state energy of the Hamiltonian. To minimize the Hamiltonian, we apply a simulated annealing for optimization [49]. Simulated annealing is a stochastic optimizing technique that is very efficient and easy to implement. The optimization procedure of simulated annealing is an iterative improvement completed at a finite interval of temperatures. Along with simulated annealing, the spin glass algorithm implements efficient update rules to improve

computation when adding a node to its proper community. This is applicable to our setting because we aim to assign each node to a unique community much like identifying spin states. Another advantage of this algorithm is that the spin-glass method can detect overlapping communities without the network being affected by the degeneracy of the Hamiltonian. One disadvantage of this method is the slow computational speed. Despite the flaw in computational speed, this method has many advantages that help meet the goals of this dissertation.

2.3.6 Bayesian network models

In this section, we will continue to review methods for constructing networks. A Bayesian network (BN) is a directed acyclic graph where $G = (X, E)$ based on a set of local conditional probabilities P. X is a set a nodes (x_i, x_j,.x_n) representing a set of variables [50, 51]. E is a directed edge that represents a conditional dependence between two variables. In a BN, the directed edge from X_i to X_j represents the relationship from parent node X_i to child node X_j. To construct the graph, BN employ belief propagation for network inference. Belief propagation works by updating and propagating beliefs about the network structure based on evidence in the observations. Bayesian networks have the advantage of predicting non-linear relationships between nodes. Another advantage is the ability to handle noisy or missing data [52].

There are many applications of Bayesian networks for graph construction. In Yu

et al. [53], a BN was used to develop the gene network inference software named Banjo. Banjo is based on the BN framework and implements Bayesian and Dynamic Bayesian networks. Dynamic Bayesian networks are an extension of standard Bayesian networks, but are much more capable of handling feedback loops. This is beneficial when handling time-series data. However, if the data is static, Banjo and regular BNs cannot infer cyclic relationships. Other applications of BNs were used to analyze expression data in [54] and to model gene regulatory networks in [55, 56].

2.3.7 Correlation-based methods

Correlation-based approaches are centered on correlation measures of mutual information, such rank correlation, Pearson correlation, Euclidean distance and the angle between the expression vectors [57]. For a given set nodes, the correlation measure is computed in order to compare the respective co-expression profiles. The most common correlation measure is the Pearson coefficient. A disadvantage of correlation based methods, is that one cannot infer the direction of an edge between two nodes. It is also impossible to determine if there is a direct or indirect interaction between two nodes based on co-expression profiles. This method has its limitations but has also been very successful in biological applications [58].

2.3.8 Graph-based learning methods

Turning now to a different approach, this section reviews graph-based learning approaches which have been widely used in the machine learning paradigm. Although there are many semi-supervised methods, graph-based learning approaches have shown to be the most promising. Similar to the standard graph representation, the labeled and unlabeled data points are vertices and the relationship between two data points are weighted edges.

Much of the current literature on graph-based learning [59–61] suggests that graph-based learning can be used to find strongly connected subnetworks within a larger network. In fact, Long et al. [60] proposed a graph approximation model learn a link-pattern based community structure. The link-pattern based communities are a group of nodes that have similar link patterns. For example, the nodes within a strongly intra-connected community have similar link patterns with other nodes in that community and the same is true for weakly intra-connected communities. A slight disadvantage of this method is that it requires some knowledge about the link-patterns in community structure. Also, the lack of application to the biological domain raises concerns if it can handle complex network structures. However, this study does reveal that this type of community detection method is capable of using labeled data to infer labels on unlabeled data. With this in mind, we will now explore

other suitable methods that incorporate labeled and unlabeled data.

There are two issues in graph-based learning. The first issue is handling noise perturbations in large-scale datasets. The other problem is the scalability of the large-scale datasets. Recent algorithms have been proposed to address these issues. In fact, a recent review by Liu et al. proposed a scalable graph-based learning approach based on anchor points in a graph [62, 63]. Anchor points are defined as points in the graph that allow the use of a non-parametric regression model to predict the labels of each data point as the weighted average of the labels of the anchor points [63]. Unlike other community detection methods, the anchor graph model infers label information onto the input data points instead of reconstructing the entire network. Thus, one major advantage of this method is the minimal impact on storage and computation. In the sections to follow, we will continue to review different types of graph-based learning methods.

2.3.9 Label propagation method

Semi-supervised methods have been used to predict unknown parts of network topology. The label propagation method is known to be fast and efficient. The label propagation algorithm (LPA) is a graph-based semi-supervised learning algorithm. In this algorithm, class labels are propagated from labeled nodes to unlabeled nodes [64–66]. The algorithm proposed in Zhu et al. [67] and Raghavan et al. [68] have

been the basis of many variations of the label propagation framework. The general idea of this method is that if a node x has a set of neighbors (x_1, x_2,.x_n), then each of its neighbors will have the same label. The label of x is determined by the label of the maximum number of neighbors in a community. Each node is initialized with a label and the labels propagate through the graph. As they propagate, a densely connected group of nodes is created, called a community. Each community created will have a unique label associated with it.

Label propagation is a robust method that uses the structure of the given network to guide the propagation process. [69]. Over the years, LPA has been applied to social networks [69, 70], computational linguistics [65, 71–73], neurological imaging [74], drug interaction discovery [75,76], disease-gene discovery [66,77] and protein function prediction and classification [37, 78, 79]. In many of these real-world applications research has shown it is possible to utilize data from various sources. This is a major advantage for research in constructing genetic networks. For example, Shahreza et al [75] included similarity networks for genes and targets, drug-target interactions and drug-disease associations in order to allow LPA to predict potential drug repositioning targets. Such implementations of this magnitude can attest to the robustness of LPA. We believe this algorithm can adapt to the noisy and complex nature of genetic networks including protein protein interaction networks and

regulatory networks as we investigate here.

2.3.10 Graph partitioning

In a broader aspect, the problem we are addressing is comparable to graph partitioning problem. The problem in graph partitioning is to divide the vertices of a graph into smaller equal sized hubs in a way that minimizes the number of edges connecting vertices of different hubs [80]. This problem is an NP-Complete problem, While the problem is the same, the mathematical description of the graph partition problem is as follows: data can be represented as a graph G = (V,E), with V vertices and E edges. This graph can be partitioned into smaller sub-graphs. A good partition is obtained when the number of edges between each partition is minimal and the edges within each partition is greater than the edges between the neighboring partitions. This problem has gained much traction due to its application in social network analysis, analyzing biological networks and clustering functionality.

There has been a multitude of methods developed to solve the graph partitioning problem. The field has gained much attention due to its applications in clustering for complex networks such as biological and social networks [81]. Moreover, these exists a great body of research on multi-level partitioning. *LaSalle* et al. [80] developed a multilevel partitioning algorithm built on k-way partitioning (KMetis) [82] and parallel partitioning (ParMetis) [83] methodologies. This is an optimization problem

where the goal is to compute the k-way partition the graph with the lowest edge-cut. Regarding the algorithm, we will now discuss the details of the partitioning algorithm.

The k-way partitioning of a graph G is breaking the graph into k disjoint subsets V. A partitioning vector is used to store the partition number of each vertex v. The edges connecting these vertices are cut based on the sum of the weight of the cut edges which is called the edge-cut, but it is constrained to the balance to be upper bounded by $1 + \epsilon$. The balance measures the weight distribution and evenly-weighted the partitions are in the graph. If the balance is close to 1, then the partitions are evenly weighted. If greater than one, then there are some partitions who have more heavily weighted vertices than average.

This method has a slight disadvantage due the serial nature of its greedy refinement of the graph. To address this concern, ParMetis, which is a parallelized implementation of KMetis, splits the refinement into two iterative passes. While this method is optimal for large graphs, its application to biological networks is limited in scope. The even partitioning of graphs is unlikely to provide much insight on the function similarities of vertices in the same partition. In many biological networks, densely connected groups often contain vertices with large weights.

In summary, it has been shown from this review section that there are many

methods available for community detection. As indicated in the previous sections, there is a common disadvantage of computational speed or inability to handle large scale networks. Therefore, this section examined alternative approaches that have been proposed to address these challenges such as graph based learning and graph partitioning. The remaining sections of this review will explore the literature related to miRNAs and miRNA-gene networks.

2.4 Introduction to miRNAs

MicroRNAs (miRNAs) are a class of non-protein-coding RNAs ranging from 21-25 base-pairs in length that regulate gene expression. They target complementary messengerRNA (mRNA), which triggers a translational blockade or degradation [84]. MiRNAs have also evolved and play a critical role in gene regulation and many biological functions and processes. Recent studies have shown that miRNAs play an important role in various cancers [15, 85–88]. Moreover, miRNAs have been reported to have specific involvement in many cellular functions in cancer such as cell differentiation, cell proliferation and apoptosis. The subsequent alterations have downstream regulatory effects in key driver pathways which cause various cancers including ovarian, colorectal, lung and breast cancer. Thus, it is important to understand how miRNAs effect gene expression and the downstream genetic neighborhood. This could lead to ways of identifying novel therapeutic targets and biomarkers to

improve prognosis procedures.

2.5 Role of miRNAs in cancer

In recent years, the role of miRNAs in cancers have be investigated extensively. Studies have proven that the dysregulation of miRNA is critical in cancer progression [89]. It is suggested that miRNA profile signatures could be used to classify tumor subtypes, improve diagnosis and prognosis. In a recent review, Peng et al [10] compiled evidence that shows there is a direct relationship with alterations in miRNA regions and ovarian cancer and breast cancer. More involved experiments show that numerous miRNA genes are located in cancer-related genomic regions. As previously mentioned, miRNAs play an important role in cellular functions such as cell proliferation. The dysregulation of miRNAs is suspected to aide in the evasion of growth suppressors and promote cell proliferation.

The potential of miRNAs as oncogenes has been studied exhaustively in previous years. Calin et al. [90] spearheaded this field by uncovering the effects of significant down-regulation of $miR-15$ and $miR-16$ in chronic lymphoid leukemia (CLL) by frequent deletions in chromosome 13q14 locus. This study also revealed 186 miRNAs that were located in various genomic regions that are frequently altered in many cancers. In addition, high expression levels of $mir-155$ in pancreatic cancer, $miR-155$ helps decrease apoptosis by targeting a common tumor suppressor

known as $TP53INP1$.There have been several studies highlighting the effects of the up-regulation and down-regulation of $miR-21$ in various cancers cancer [17, 91, 92]. It is well documented that $miR-21$ is often up-regulated and targets mostly tumor suppressor genes. These effects impact many functions such as promoting cell proliferation, invasion and metastasis, evading cell death and even abnormal metabolic functions [93].

2.6 Identifying miRNA target genes

In this section, we will investigate the current methods for identifying miRNA target genes. As previously mentioned in the first chapter, the goals of this dissertation involve identifying downstream miRNA target genes in genetic networks. Current research suggests that we must first understand the direct relationship between miRNAs and their target genes [94]. That said, we will first examine methods to identify miRNA target genes and then expand our review to methods focused on identifying miRNA-gene associations using gene expression profiles.

The current research on miRNA-gene relationships has established that identifying miRNAs, target genes and their functional regulatory networks are critical in understanding miRNAs and their roles in the onset and progression of disease [95–99]. This evidence has drawn great attention on the development of methods for the prediction of miRNA target genes. Therefore, in this section we will examine current

methodologies and discuss the advantages and disadvantages of each method.

In a recent study, Rabiee-Ghahfarrokhi et al. [94] used a machine learning approach to predict the the target genes of miRNAs. In particular, the method developed in this study used a genetic algorithm-based decision tree to identify miRNA target genes. In this approach, the researchers use a decision tree and a set of logical classification rules. Regarding the details of this method, the process in this method is as follows: 1. Split the dataset into training, testing and validation sets. 2. Using the decision tree, extracted the relationships based on the rule-based classification. 3. Apply 10-fold cross-validation 4. Apply the genetic algorithm to select the best set of rules. 5. Apply the best set of rules to the test dataset. The results of this method yield the number of times a set of rules has predicted a miRNA-gene relationship correctly or incorrectly. The benefit of this approach is the low computational cost and its high prediction accuracy.

To further examine the use of machine learning algorithms to identify miRNA target genes, we review a newer methodology proposed by Ovando-Vasquez [100]. This study used gene expression profiles to identify miRNA-gene relationships. Specifically, this method trains a support vector machine algorithm on a feature-ranked set of for each miRNA target gene in the dataset. A major advantage of this method is that by using the trained SVM models, this study showed that gene expression is

the most predictive feature to identify true target genes. Furthermore, it has a low computational cost. One limitation is that this method focuses on single miRNA target gene relationships. As previously stated, research has shown that multiple miRNAs regulate the same gene.

In a different study, Wang et al. [101] proposed a multiple linear regression method to identify miRNA-gene relationships in colorectal cancer. This method began with selecting candidate genes commonly involved in colorectal cancer pathways. The next step involved pruning a selected miRNA target prediction databases to identify miRNAs targeting the predefined set of candidate genes. With the candidate genes, expression data and the miRNAs targeting those genes, Wang et al. used a multiple linear regression model to identify the associations between each miRNA and the corresponding target gene. In this case, there were multiple hypotheses tested, thus requiring a multiple test correction. A common standard when correcting for multiple hypotheses is to calculate the false discovery rate to control the expected proportion of false positives. Results from this study confirmed that by integrating miRNA targets from prediction databases, one can decrease the number of false positives. One major advantage of this approach is that it can determine if more than one miRNA are targeting the same gene.

In terms of accessing information on miRNA target genes, we will discuss databases

available for systematic searches for identifying miRNA target genes. Research shows that there is a growing usage of using curated databases to retrieve miRNA target genes. Most recently, many target prediction databases are available. Among these comprehensive prediction databases, some of the prediction algorithms predict miRNA targets using conservation among various species, such as TargetScan [102]. Databases such as miRDB [97, 103] use machine learning to predict miRNA targets. Moreover, comprehensive databases such as TarBase [104] and miRTarBase [105] contain a catalogue of experimentally validated targets. In addition, there are databases curated for miRNA targets in human disease, such as miR2Disease [106] and miRCancer [107]. With many resources available and various methods to identify miRNA target genes, combining this information has become increasingly popular. That said, in the next section, we will discuss methods for constructing miRNA-gene networks.

2.7 Constructing miRNA-gene networks

The previous section has shown that there are several methods to identify miRNA-gene relationships. In this section, we will discuss methods for constructing miRNA-gene networks as this is directly related to the work in this dissertation. Furthermore, an earlier review by Drakaki and Iliopoulos [108] revealed the importance of miRNA-gene networks in the oncogenesis of cancer. Moreover, miRNA-gene networks contain diagnostic and prognostic signatures in cancer.

There are various ways of constructing miRNA-gene networks. One common approach is to build co-expression networks and identify miRNA-gene associations. Mamdani et al. constructed miRNA-gene networks using methodologies such as WGCNA [109] to build co-expression modules [110]. Such correlation based method relies on calculating the pair-wise Pearson correlation coefficients between differentially expressed genes. The correlations are then represented as a signed similarity matrix in which the positive correlations are selected for the co-expression networks. Using co-expression modules to construct miRNA-gene networks has a key advantage in identifying how gene subnetworks are targeted by multiple miRNAs. However, like many correlation based methods, it is difficult to determine the direction of the relationship.

The powerful capabilities of machine learning algorithms have allowed the feasible integration of discrete and continuous variables. This setting requires the use of a mixed graphical Markov model. In Tur et al. [111], the mixed graphical Markov model was used to construct miRNA-gene networks. In this study, the discrete variables are mapped to the genotype alleles from the corresponding target. The continuous variables are the values of gene expression. This model is assumed to follow a joint Gaussian distribution whereby the continuous variable follows a multivariate normal distribution with conditional dependence on the discrete variable.

Another key assumption in this study is that the miRNA-gene network is a homogeneous mixed graphical Markov model which, in theory, assumes that the miRNAs can only influence the mean expression levels rather than any correlations between the genes. In terms of constructing the network, miRNA-gene associations are depicted by discrete vertices and expression values are depicted by continuous vertices. It is critical to understand in this setting, miRNA-gene associations affect the gene expression and this behavior is non-interchangeable. This method is advantageous when generating directed and some undirected mixed graphs or networks.

Research on miRNA-gene networks has drawn much attention to integrating expression data with interaction data and various other types of biological data. A recent study by Freiesleben et al. [112] implemented a consensus approach combining differentially expressed genes, miRNA targets from target prediction databases and transcription factor-miRNA interactions. Using this consensus strategy, the constructed miRNA-gene network revealed that miRNAs are involved in the complex regulation system in gene networks. One major advantage of this consensus-based approach is that there is a smaller risk of finding false positives [112]. Let us now consider an integrative approach to construct miRNA-gene networks.

Quitadamo et al. [21], proposed an integrated network that included four types of miRNA and gene interactions. More specifically, this approach combined gene asso-

ciations, miRNA-gene associations and correlated miRNAs. This integrative method began with identifying miRNA-gene associations and then identifying miRNA target genes using target prediction databases. Using the miRNA-gene associations and miRNA target genes as seed nodes, the network was extended with a statistical model provided by Disease Associated Protein-Protein Link Evaluator (DAPPLE) [113]. Unlike the other miRNA-gene network methods previously mentioned, this method captured miRNA co-expression networks and gene co-expression networks using a two-graph guided approach to determine network associations. The networks generated in this method were integrated to construct a comprehensive network of the relationship between miRNA expression and gene expression. One major advantage of this method is that it captures different types of interactions between miRNA expression and gene expression [21]. Furthermore, it captured relationships between miRNAs and known cancer genes in ovarian cancer. Regarding the methods discussed in this section, integrative approaches yield more insight into the complex relationships between miRNA expression and gene expression.

2.8 Conclusion

Overall, these studies highlight the need for methods that can capture the true complexity of miRNA-gene networks. In addition, after carefully reviewing the current state of constructing miRNA-gene networks, there is a need to expand our

knowledge by investigating how miRNAs affect gene expression and other downstream genes in genetic networks. Moreover, integrative network approaches can shed light on the relationships between miRNAs, genes, protein-protein interactions and even phenotypic associations. Therefore, the aim of this dissertation is to understand how miRNAs affect gene expression and the downstream genetic networks.

In this work, we investigate the challenges in constructing integrative miRNA-gene networks with multiple layers of expression data and interaction data in efforts to understand how miRNAs effect gene expression. In our first study, we introduce a community detection approach to expand miRNA-gene networks using miRNAs, miRNA affected genes and miRNA target genes in ovarian cancer. This methodology is focused on constructing miRNA-gene networks from large-scale datasets and identifying significant genetic networks. Then, we use a robust label propagation algorithm to construct miRNA-gene networks with miRNAs, miRNA target genes, protein-protein interactions, gene regulatory network interactions and miRNA-gene associations in colon cancer. We compare this method with our previous method and other network construction approaches. Finally, we apply these methods to identify enriched communities and network patterns associated with the pathologic stages in a miRNA-gene network in colon cancer. Our results show that by using these methods, we can gain a better understanding of the relationships between miRNAs

and gene expression and the effects on downstream targets in genetic networks.

In the chapter that follows, we present our first study to examine the relationship between miRNAs and gene expression in ovarian cancer by constructing an integrative miRNA-gene network via community-based network expansion. We also explore how miRNAs affect downstream genes in genetic networks. The chapter that follows moves on to consider the challenges discussed in this review.

CHAPTER 3: CONSTRUCTING INTEGRATIVE MICRORNA-GENE NETWORKS IN OVARIAN CANCER

3.1 Introduction

Ovarian cancer is the fifth leading cause of cancer related deaths in women today [114]. In 2017 alone, there will be 22,440 expected new cases of ovarian cancer and approximately 14,080 deaths. If diagnosed in its early stages, the five year survival rate is over 90%, but is less than 20% if diagnosed in its final stage. It is evident that understanding the underlying functions that induce progressive behavior can lead to better prognosis and early detection. Recent studies have shown that miRNAs mediate cell growth [89] and act as tumor suppressors [11].

The influence of miRNAs on gene expression have become an important part of disease studies. For instance, $mir-200c$, which modulates metastasis, targets gene $PTPN6$, which is a protein tyrosine phosphatase, non-receptor-type, 6. Although the role of $PTPN6$ in cancer development is still unclear, we know that it is targeted by $mir-200c$, a common biomarker in epithelial ovarian cancer [115]. Using this knowledge and our integrated network, we can investigate other downstream targets, such as $PTPN6$ and their neighbors. With the advancement in high-throughput

sequencing technologies and the availability of multiple layers of data, we should design methods to integrate these layers together for a more systematic view of how miRNAs perturb gene expression.

Based on these challenges, we developed a framework based on the spin-glass model to construct an integrative miRNA-gene network. We then address the question of how do miRNAs affect the downstream genes in genetic networks in ovarian cancer. Also, we investigate the enrichment of these genetic networks. We hypothesized that integrated miRNA-gene networks can identify the relationships between miRNAs and the downstream genes in genetic networks in ovarian cancer. In doing this, we believe that the miRNA-gene networks provide insight on how miRNAs affect progression of ovarian cancer.

3.2 Methods

This study was designed to construct an integrative miRNA-gene network in ovarian cancer. The goal was to study the relationship between miRNAs and downstream genes in genetic networks. Our research design was the following: First, we used miRNA and gene expression data, we identified miRNA-target genes using target prediction databases. Using miRNA-gene association analysis, we identified relationships between miRNAs and genes [20, 21]. We also performed correlation analysis to identify correlations between miRNAs and correlations between genes [19, 116]. Net-

work expansion was performed via community detection to integrate miRNA-gene associations, miRNA-miRNA and protein-protein interactions. Pathway enrichment analysis was performed to investigate the function of candidate genes and communities. A work-flow of our study is shown in Figure 1.

The data used in this study was generated by The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/) [117] which is a collaboration between the National Human Genome Research Institute and the National Cancer Institute. The mission of TCGA is to build a comprehensive profile of the major types and subtypes of cancer [118]. In this study, we chose to focus on Ovarian Serous Cystadenocarcinoma because it is the fifth leading cause of cancer deaths in women. From TCGA, we collected miRNA expression profiles and gene expression profiles for 480 Ovarian Serous Cystadenocarcinoma samples.

Our design required a comprehensive network of protein-protein interactions to expand our miRNA-gene network. We chose to use the InWeb human interactome, which is a protein-protein interaction network (`https://github.com/BenikaH/spinglass/blob/master/data/InWeb29.txt`) [119]. The InWeb network containing over 500 protein complexes associated with human disease. Lage et al. constructed this comprehensive protein-protein network by pooling human protein interaction data from various databases such as MINT [120], BIND [121], Reactome [122] and KEGG an-

notated protein-protein interactions [123]. To construct the network, in this study a probabilistic score was developed to obtain a global interaction score. Next,an interaction confidences was measured by calculating the reliability of the probabilistic score and confirmed by a fitted calibration curve. Using these high-confidence interactions, the final InWeb network contains 343,000 unique protein-protein interactions. Having discussed how the InWeb network was constructed, the next section of this dissertation addresses our data preprocessing.

3.2.1 Data preprocessing

We preprocessed the data in a series of data cleansing and filtering steps. First, we removed the samples with missing data. We then extracted expression data for each sample containing 183 miRNAs and 13,536 genes. For each miRNA expression set, we created an RxC miRNA expression matrix where rows = miRNAs and columns = sample ID. Thus, each row contained miRNA expression values for each patient sample. This was completed in Python with the following script `https://github.com/BenikaH/spinglass/blob/master/scripts/miRMatrix.py`. We then repeated this process for the gene expression data `https://github.com/BenikaH/spinglass/blob/master/scripts/ExpressionMatrix-1.py`. To reduce the variation between the samples, we utilized sample-based quantile normalization to normalize the miRNA and gene expression data separately [124, 125]. This was im-

portant because we had to normalize the read counts to account for any technical effects that may have occurred during sequencing. This is a standard approach with high-throughput data. This global normalization technique relies on the assumption that any global changes observed across all samples are caused by unwanted technical variability [126]. To normalize the expression data, we utilized a function in R called "normalize.quantiles" from the "PreprocessCore" library (`https://github.com/BenikaH/ColonStudy/blob/master/scripts/stage2_eqtl.R`).

3.3 MiRNA-gene associations

To determine miRNA-gene associations, we identified the relationships between the miRNA expression and gene expression profiles. We performed this analysis using Matrix eQTL [125], an R package that uses matrix operations to performs mapping using large matrix operations. We performed this analysis using miRNA and gene pairs whose genomic locations were within 1MB of each other. To identify associations between the miRNAs and genes, we used a linear regression model provided by the Matrix eQTL package. Using Matrix eQTL, we identified 44 miRNA-gene associations with a FDR <0.01.

Because we were testing many associations, the traditional statistical significance of $\alpha = 0.05$ would have yielded many false positives. For this reason, we corrected for multiple comparisons and to do so, we utilized a multi-test correction based on

false discovery rate (FDR). We implemented false discovery rate by applying the Benjamini-Hochberg test [127]. Benjamini-Hochberg test assumes that the tests are independent or positively correlated. When using the FDR estimation, we tested if each miRNA-gene pair is statistically significant by estimating the corresponding q-value in Equation (3.3).

$$E(Q) \leq \frac{m_0}{m}\alpha \leq \alpha \tag{3}$$

where m_0 is the number of true null hypotheses, m is the total number of hypotheses tested, α is the FDR level threshold. Implementing this procedure, we chose the significant miRNA-gene associations with an FDR threshold at level α <0.01.

3.3.1 Network Expansion

Using the genes from the miRNA-gene associations obtained in the miRNA-gene association analysis, we queried two databases: the MicroRNA Target Prediction And Functional Study Database (miRDB) [97] and Tarbase [104], which is an experimentally validated miRNA target database. If the database query results revealed targets for the matching miRNAs, we added the targets as edges to our network. We then expanded the network with the edges of the matching genes in the InWeb network.

In order to identify downstream communities that interact directly with miR-

NAs and associated genes, we implemented the spin-glass community detection algorithm in R's igraph library [128] (https://github.com/BenikaH/spinglass/blob/master/scripts/Spinglass.R). We first, generated an edge-list (node-node pairs) for our extended network. We then converted the edge-list into a graph object. We used the "spinglass.community" function with the following parameters: (graph, weights = NULL, start.temp = 1, stop.temp = 0.3, update.rule = "config", gamma = 2). Here, the graph parameter is the network to be analyzed and the weights are the corresponding weights of the network. The start.temp and stop.temp parameters are the temperatures for the simulation. More specifically, when to start and stop the simulation based on the lowering of the temperature to optimize the system. We chose the default value of 1 for start.temp, but increased stop.temp to 0.3. The update.rule parameter corresponds to the null-model of the simulation which uses a random graph with the same vertex degrees as the input graph as opposed to a the same number of edges as the probability baseline of the input graph. We chose the "config" option in order to use the same number of degrees as our input network for the simulation. The gamma parameter specifies the importance of edges present in a community and edges not present in a community. Here, lower values make the edges inside the community more important than edges outside of the community. We chose a lower value of 2 to make the edges inside the community more important

than edges outside of the community. Next, we ran the model with these parameters. We then extracted the node membership for each community in order to visualize the communities in Cytoscape [129].

3.3.2 Edge recovery of spin-glass

Before investigating the communities detected by spin-glass, it was important to determine if these communities were defined each time we ran the spin-glass method. To do this, we chose to run 10 simulations of spin-glass and measure the edge similarity of the communities in each simulation. This allowed us to quantitatively measure the community divergence in our miRNA-gene network.

To calculate the edge similarity of each community, we aligned the communities using the "GraphAlignment" package in R (`https://github.com/BenikaH/spinglass/blob/master/scripts/node_recovery.R`) [130]. We imported communities derived from the spin-glass algorithm into R. For each community, we generated the adjacency matrix. Using the adjacency matrices, we aligned the matching communities from each run to determine the edge similarity between the community structures. This was done to measure how many nodes does spin-glass recover when it defines a community. We wanted to know if a community consisted of the same nodes each time it is ran. For each community linked to the miRNA-gene associations, we performed pathway enrichment analysis using DAVID [131].

3.3.3 Statistical Analysis

After identifying communities with spin-glass and determining the consistency of these communities through simulations, we conducted a statistical analysis to test if the enriched genes found in our pathway analysis results were significant. Moreover, we wanted to test if our pathway gene sets were independent of the published pathway gene sets. Therefore, for each functionally enriched pathway in the reference, we created a 2x2 contingency table to represent the number of genes present in each pathway found in the original published data and our pathway analysis. First, we calculated the number of genes in our functionally enriched pathways and in the functionally enriched reference pathways. Then we calculated the number of genes that were unique to our functionally enriched pathways.

To compare the genes found in our Notch pathway with the genes found in the reference Notch pathway, we performed the Fisher's exact test of independence. In this test the null hypothesis is that the gene set in our Notch pathway were independent of the gene set in the reference Notch pathway (Table 1) and similarly for the Basal pathway (Table 2). Assuming this null hypothesis, we applied the Fisher's exact test on the 2x2 contingency table to generate a p-value and odds ratio, given a 95% confidence interval (`https://github.com/BenikaH/ColonStudy/blob/master/scripts/fisher.test_spinglass.R`).

Table 1: A 2x2 contingency table representing the number of genes in our Notch pathway and the reference Notch pathway, the unique genes in the reference Notch pathway, the unique interactions found in our Notch pathway and the total number of genes.

	Genes in Reference Notch Pathway	Genes not in Reference Notch Pathway
Genes in our Notch Pathway	7	22
Genes not in our Notch Pathway	2	297

Table 2: A 2x2 contingency table representing the number of genes in our Basal pathway and the reference Basal pathway, the unique genes in the reference Basal pathway, the unique interactions found in our Basal pathway and the total number of genes.

	Genes in Reference Basal Pathway	Genes not in Reference Basal Pathway
Genes in our Basal Pathway	9	182
Genes not in our Basal Pathway	1	147

3.4 Results

We generated an integrated network consisting of miRNAs, their directly associated genes, protein-protein interactions and indirect targets of miRNAs and the genes from the communities derived from the spin-glass model. The edges of the constructed network include the miRNA-gene associations, their affected genes and the interactions among their downstream genes in a genetic network. Our integrated

network contained 183 miRNAs, 13,536 genes, 44 miRNA associated genes and 2361 miRNA targets. These miRNAs and genes are connected through 855 miRNA-gene association edges and 380,000 protein-protein edges. In terms of the community detection analysis, we observed that the spin-glass model detected a total of 25 communities, however only 13 communities were linked to miRNA-gene associations. Therefore, these 13 communities were used for our downstream analysis described in the Methods.

Our simulation results showed that out of the 15 consistent communities, spin-glass recovered more than 76% of nodes and edges in all 10 simulations (Table 3).

Table 3: Table showing the percentage of edge recovery of the most consistent communities in all 10 simulations of spin-glass. Only 15 communities were consistently detected in all 10 simulations while the others were not consistently defined in the spin-glass algorithm. This shows that spin-glass recovered 76% or more of the edges in each of these communities.

Community Number	**Percent of Edge Recovery**
Community 1	97.48596
Community 2	97.85322
Community 3	97.87851
Community 4	95.64697
Community 5	95.08038
Community 6	97.56603
Community 7	87.34146
Community 8	76.0274
Community 11	80.95071
Community 12	86.31008
Community 13	87.98856
Community 14	84.6952
Community 15	93.61404

To validate our findings, we compared our communities to the subnetworks in The Cancer Genome Atlas Network's published study of the ovarian cancer data [132]. In the initial analysis, there were four altered pathways discovered for potential therapeutic targets.

We found 59 miRNA associated genes with disease associations identified in the OMIM database and 2361 targets with disease associations in our extended network.

However, many had unclassified phenotypes and deserve further investigation to help understand the effects of genetic variation throughout the specified communities. We also observed oncogenes and disease interactions present in the downstream genes in genetic networks of these miRNAs identified in the OMIM database [133]. This information will be vital in identifying new biomarkers for early prognosis in ovarian cancer.

We also discovered 6 communities were enriched in cancer related pathways. We observed that there were several genes present in these these cancer related pathways such as, $PTPN6$, $CCND1$, $TP53$. At this point in our analysis, it is unclear whether the cancer genes in these communities are drivers of the cancer related pathways or are mediated by the involved miRNAs. This requires further investigation and validation on the roles of the participating genes in these pathways.

Results from our statistical analysis showed that there is a significance (p-value $<$ 6.534e-07) in the gene sets in our Notch pathway and the reference Notch pathway. These findings imply that we can reject the null hypothesis that these gene sets are independent. Moreover, we can interpret the odds ratio of 45.75 as the odds of the genes in our Notch pathway are 45.75 times that for the genes in the reference Notch pathway [132]. It is unlikely that the gene-sets in our enriched Notch pathway equally occur by chance in the reference Notch pathway. This was in line with our

hypothesis that the genes in these pathways are not independent and they contribute to the functional enrichment of the Notch pathway. Similarly, for the Basal pathway, our analysis yielded a p-value < 0.04728. While this p-value is close to .05, it still implies that the gene set found in our Basal pathway is independent of the gene set found in the reference Basal pathway. We can interpret the odds ratio of 7.24 as the odds of the genes in our Basal pathway are 7.24 times that of the the genes in the reference Basal pathway [132].

3.4.1 Comparison to Metis

We compared spin-glass community structure with the graph partitioning algorithm Metis [80, 82, 134, 135]. We used "gpmetis" application and the default parameters but chose the following options for number of partitions = 25 and partition type = kway. We then compared the community membership of the vertex sequences from spin-glass with the graph partition file generated by Metis. For example, we collected the vertex sequences for all graph partitions detected by Metis and similarly for spin-glass. Next, we compared the community membership of the miRNAs and genes generated by each method. By doing this, we were able to identify the number of matching genes in each community between both methods.

What we found here is that Metis partitioned the network into nearly equal number of genes (Figure 2), while spin-glass did not split them equally (Figure 3). Shown

in Table **??** is a summary of counts describing which communities had the most overlap. While there was not much overlap between the spin-glass communities and Metis communities, pathway enrichment analysis revealed similar cancer related genes embedded in the communities. In addition, although the community membership differs from that of spin-glass method, the Metis graph partitioning method detected communities that were enriched in the same cancer related pathways (Table **??**). We also identified common genes in both Metis communities and its comparable spin-glass community and performed pathway enrichment analysis on those common genes. These communities had the most overlap when comparing Metis and spin-glass communities. Shown in Table **??** are the enriched pathways of the common genes found in Metis communities and its respective spin-glass community. These findings showed that both methods are capable of detecting communities in an integrated network. However, Metis has the computational advantage over the spin-glass method.

3.5 Discussion

One goal in this chapter was to explore the relationship between miRNAs and gene expression in ovarian cancer by constructing an integrative miRNA-gene network via community-based network expansion. In addition, we wanted to understand the effects of miRNAs and downstream genes in genetic networks. Thus, we integrated

the InWeb protein-protein interaction network and downstream miRNA targets. We believe that by extending the miRNA-gene expression network, we were able to observe direct and indirect downstream genes in genetic networks in ovarian cancer.

The enriched communities were enriched with oncogenes such as $PTPN6$ and $TP53$ which are known to be pivotal in tumorigenesis of epithelial tissues in ovarian cancer. We also observed that some of the oncogenes were directly affected by miRNAs. For example, the relationship between $miR-200c$ and $PTPN6$ which modulates metastasis and tumor progression in epithelial cells. On the other hand, when evaluating the enriched pathways of these downstream genes in the network, we identified more enriched pathways and novel genes within those pathways. Although the reference pathways contained different genes, we do believe that the genes identified are correct as they were conserved in each of our spin-glass simulations.

Our work here suggests that spin-glass is able to capture the direct and indirect effects of miRNAs and their downstream genes in genetic networks. One advantage of this method was that it allowed use to explore enriched communities that were affected by miRNA mediated target genes. One of the major concerns with the spin-glass model is the size of input network. As the size of the network increases, the algorithm becomes more computationally expensive. This is a slight challenge as we are interested in integrating multiple networks and identifying downstream

genes in genetic networks across multiple layers. While the algorithm successfully integrated miRNA affected genes and the InWeb network, it was unable to initialize seed nodes. However, when we compared to Metis, we found that spin-glass is the more preferred method. Metis partitioned the graph into nearly equal partitions which may not be the true nature of a biological network. In terms of computation performance, Metis outperformed spin-glass by a factor of 60x (minutes). It is also worth noting that Metis found similar enriched pathways, but did not have many genes in common for the detected communities. This could be due to the equal partitioning and the manner in which it assigns its membership. it could also be due to the community sizes for spin-glass, the communities were very dense and its possible not all genes contribute to the functional enrichment of their respective communities. For the reasons mentioned above, we developed an alternative approach, based on a label propagation framework which is capable of integrating multiple networks in an efficient manner.

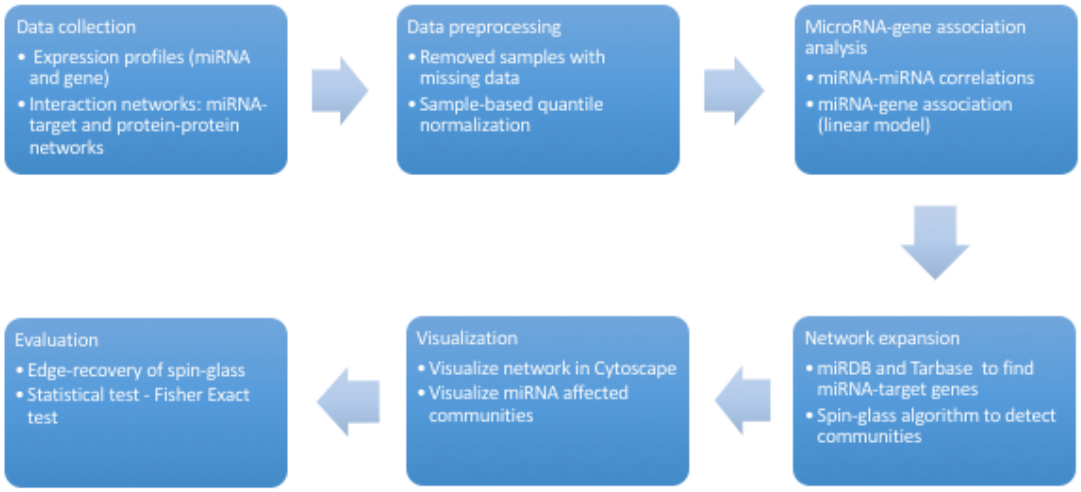

Figure 1: A work-flow diagram showing the detailed steps of the spin-glass method. First, we preprocessed the data. Then we identified miRNA-gene associations. We then performed spin-glass community to expand the network. Next, we visualized the integrated network in Cytoscape. Lastly, we evaluated the spin-glass model.

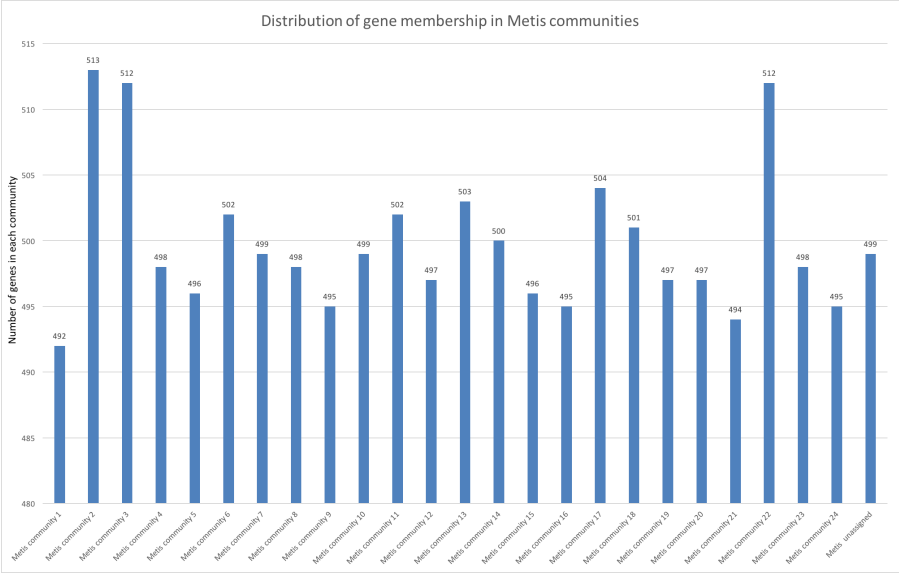

Figure 2: A histogram of the total number of vertices in each community detected by Metis.

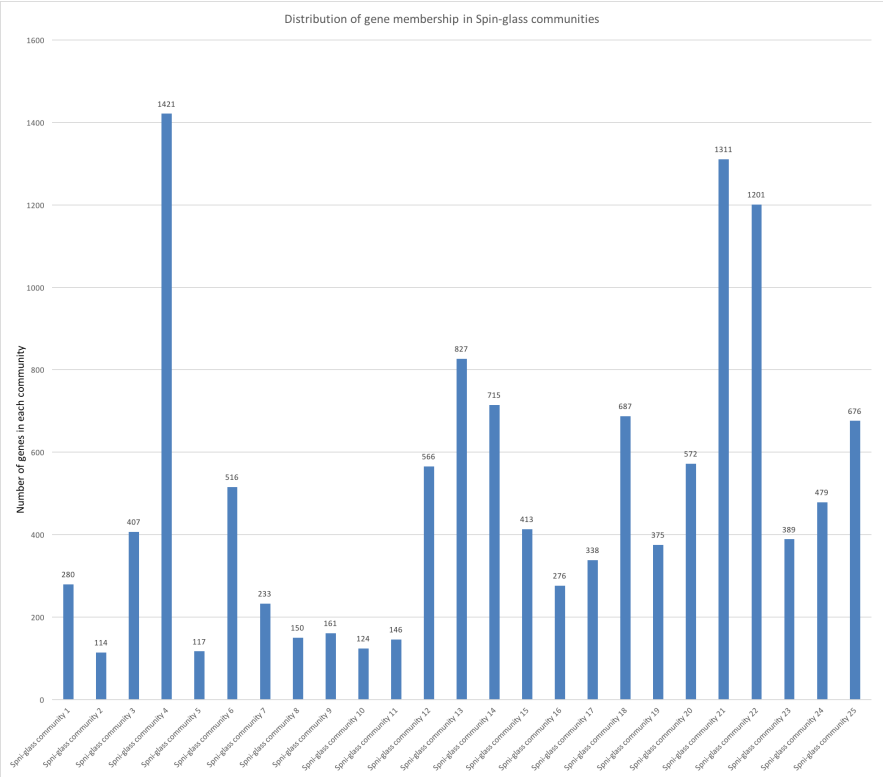

Figure 3: A histogram of the total number of vertices in each community detected by Spin-glass.

CHAPTER 4: CONSTRUCTING MULTI-LAYER MICRORNA-GENE NETWORKS USING A LABEL PROPAGATION APPROACH

4.1 Introduction

There has been great interest in determining how miRNA-gene relationships impact disease progression. These relationships play a critical role in various key cellular functions such as cell differentiation, proliferation, cell invasion and apoptosis. Interestingly enough, these relationships can be exploited via miRNA mediated gene regulatory networks. In order to understand the role of miRNAs and the regulation of gene expression in ovarian cancer, it is critical to learn how the miRNAs interact with targets in other networks such as gene regulatory networks and protein-protein networks.

Gene regulatory networks have been used to identify novel biomarkers in many cancers and various diseases [136–138]. While some genes in these networks are prone to miRNA regulation, it is unclear what the consequences are for those regulatory events [23]. The phenotypic outcome of a disease is not mediated by a single miRNA and a single target, therefore we must consider intermediate factors such as regulatory networks. By doing this, we are able to capture many different rela-

tionships between the miRNAs, protein-protein networks, regulatory networks and other genetic interactions. Generally, multiple genetic mutations occur in cancer progression, however it is unclear if the mutations are driven by miRNA regulation or in fact gene-gene regulation. For instance, the $TP53$ suppressor gene is reported to be the most frequently mutated gene in cancer and leads to the critical loss of the functional protein $p53$ [139]. This gene is also found to interact with or be mediated by $hsa-mir-877$, which is related to ovarian cancer [21].

Network-based integration of multi-omics data has been important in suggesting functional mechanisms for pathological phenotypes. Currently, majority of the methods involve only two layers of network interactions such as protein-protein interactions or correlation based gene interactions. Other methodologies have included DNA methylation data to identify gene regulatory modules [140, 141]. The inclusion of gene regulatory networks will expand the current knowledge of the complex relationship between miRNAs and the downstream genes in gene regulatory networks in cancer subtypes.

The task of integrating multiple networks in our setting was modeled as a network expansion problem, where a complete network topology was constructed based on known subnetworks or seed nodes. As opposed to combining bioinformatics analysis results from single layers, there is always a risk of loss of information [142]. To

address this issue, we implemented a robust semi-supervised learning approach called the label propagation algorithm, to model the complex relationships among miRNAs and their regulated genes. In this study, we integrated three heterogeneous networks to help understand how miRNAs interact with protein-protein interaction networks and gene regulatory networks in ovarian cancer via network expansion based on the label propagation algorithm.

4.2 Methods

The data used in this study was based upon data generated by the TCGA Research Network: `http://cancergenome.nih.gov/`. All scripts used to transform the data for our analysis can be found at `https://github.com/BenikaH/LPA/`. We utilized preprocessed data generated in the previous chapter.

The full work-flow of our integrative method is shown in Figure 4. Similar to the procedures described in Chapter 3, we collected miRNA and mRNA expression data from The Cancer Genome Atlas (TCGA) [117]. We also gathered networks from two respected databases. First, the protein-protein interaction network was obtained from Lage et al [119], which represented the human interactome. We retrieved a gene regulatory network from RegNetwork database [143] which consisted of experimentally verified targets. The second phase in our work-flow involved the data preprocessing steps required to conduct analysis on. We begin by filtering

the data to handle missing values. We then applied quantile-quantile normalization on both sets of expression data. Next, we filtered the networks to retain the most highly confident interactions in our data pool. The next phase of our approach included identifying associations between miRNAs and gene expression and discovering correlations between miRNAs as well as correlations between genes. To identify miRNA-gene associations, we used Matrix eQTL [125] with the following parameters :(useModel : modelLINEAR, cisDist : 1000000, pvOutputThreshold-cis : .01). Here, we identified 44 miRNA-gene associations with an FDR $\leq$ .01. MtLasso2G [116] was used to identify the correlations between miRNAs and correlations between genes in which we identified 48 associations. Next, we performed the network expansion phase in which we integrated multiple networks using the label propagation algorithm. Here, we used our miRNA affected genes as seed nodes and expanded our network with two additional networks, the InWeb protein-protein network and the RegNetwork regulatory network. To test the prediction accuracy of the model, we conducted three-fold cross validation to learn the hyper-parameters in the LPA algorithm. Lastly, we visualized the fully integrated network in Cytoscape [129].

4.2.1 Network expansion via label propagation framework

In general, the label propagation algorithm works as follows: Given a set of n nodes represented in an undirected graph G, $X = [x_1...., x_l]$ in which a portion of the nodes

are labeled l. LPA predicts the labels of the remaining unlabeled nodes $[y_1 ...., y_l]$. We assigned our miRNA-gene associations from our Matrix eQTL results generated in the previous chapter as seed nodes. Using these seed nodes, we expanded our network with the InWeb protein-protein network and the RegNetwork regulatory network using a robust label propagation algorithm. In this particular phase, we utilized the algorithmic implementation developed by Karasuyma [65], namely SMGI. This implementation used sparse integration of multiple graphs by combining Laplacian matrices linearly.

4.2.2 Label propagation algorithm

Algorithm 1 Label Propagation

1: **procedure** INPUT$L_{k,k=1}, \lambda_1, \lambda_2$
2: Output: f, u
3: *Initialize* u
4: Optimize f, the optimal f is reached using Equation (4.2.2), given the fixed u, then solved by Quadratic Problem solvers
5: Optimize u by using Quadratic Problem solvers to minimize u using Equation (2), the optimal u vector has m non-zero elements and the remainder of the elements are equal to 0
6: Repeat step 2 and step 3 and convergence.

• LPA Algorithm (1) requires a set of Laplacian graphs, the corresponding label vector and desired values for regularization parameters λ_1 and λ_2 as input. To generate the Laplacian matrices, we wrote an R script to convert the edge-lists to normalized Laplacians `https://github.com/BenikaH/LPA/blob/master/scripts/smgi_data_prep_experiments.R`.

• The resulting output is the minimized objective function and a resultant weight vector, u. The objective function is minimized by quadratic programming.

• Next, the objective function and the resultant vector are optimized and this process is repeated until convergence is reached. Convergence is reached when all nodes have been assigned a label.

• The optimization of f and u occurs in two separate phases. First, to optimize the score, f, Equation (4.2.2) is applied. Here, I is the identity matrix, u represent the

vector of weights and Z is the normalized constant for the Laplacians, represented by L.

$$f = \lambda_1(\lambda_1 \mathbf{I} + \sum_{k=1}^{K} u_k / Z_k L_k)^{-1} y \tag{4}$$

$$min \quad uv^T u + \frac{\lambda_2}{2}||u|| \tag{5}$$

Algorithm 2 Optimization of u

1: **procedure** INPUTv(has to be sorted), λ_2
2: Output: u
3: **for** $m \leftarrow$ *1 to* K **do**
4: $n \leftarrow (\lambda_2 + \sum_{k=1}^{m} v_k)/m$
5: **if** $m = k | n - v_k \geq 0, k = 1, ..., K$
6: *break*
7: **end if loop**
8: **end for loop**
9: $u_k \leftarrow (n - v_k)/\lambda_2 \ \ for \ \ all \ \ k = 1, ..., m.$
10: $u_k = 0$ for all k = m + 1,...,K.

The optimization of the resultant weighting coefficients, u, was applied in Algorithm (2). For all non-zero elements m from 1 to K (the total number of graphs) in the resultant vector u, the number of positive elements n was calculated in step 4 of the algorithm. In step 4, the number of positive elements, n is approximated by the regularization parameter λ_2 plus the summation of all data points, v_k in each graph

k, divided by the number of non-zero elements, m. The optimal resultant vector, u, is considered optimal if it has only m non-zero elements and the remaining elements are zero. Thus to find the optimal number of non-zero elements is key to finding the optimal resultant weight vector of the integrated graph. The algorithm then checked to see if u was equal to zero for each non-zero elements and counted the number of positive elements. If m is a set of positive integers for graphs $k = 1, ..., K$, then we have an optimal resultant weight vector, u_k. For all $k = 1, ..., m$, u_k is calculated by the number of positive elements minus the number of data points divided by the regularization parameter λ_2. Lastly, for all $k = m + 1, ..., K$, u_k equals 0. Thus, defining the optimal resultant weighting coefficients vector, u.

4.2.3 Cross-validation

In Matlab we wrote a function to performed three-fold cross validation to learn parameters λ_1 and λ_2 and to evaluate the model. We applied a grid search method to determine which parameter values would be most optimal for our networks. We chose to do three fold cross validation with our λ_1 values between $[0.01, 0.1]$ and our λ_2 values between $[0.1, 1]$. For each run, we increased λ_1 and λ_2 by intervals of .01 and 0.1 respectively until the max of each interval was met. Cross validation results yielded a minimum error rate of 2.2796e-04 with $\lambda_1 = .01$ and $\lambda_2 = 1$. Using these parameter values, we performed the integration with LPA.

4.2.4 Evaluation

To evaluate the LPA implementation, we tested the robustness of the model by initializing different seed nodes for 3 separate runs. The first implementation was run with only the significant miRNA-gene associations as seed nodes, the second and third implementations were ran with random seed initialization. We utilized the cosine similarity function to measure the similarity between each of the resultant vectors provided by the LPA algorithm. Cosine similarity is a measure of similarity between two non-zero vectors to determine the cosine angle between the two vectors. The resulting similarity ranges from -1 to 1, indicating if the similarity is 1, then the vectors are identical. If the resulting similarity is 0, then the vectors are orthogonal. Resulting similarity measures close to -1 or close to 1 represent dissimilarity or similarity between the two vectors respectively. The cosine similarity is shown in Table 4. `https://github.com/BenikaH/LPA/blob/master/scripts/lpa_cosine_similarity.R`

Table 4: Cosine similarity measure between LPA networks for 3 implementations with different seed initializations

Cosine Similarity measure between LPA networks

	LPA-1	LPA-2	LPA-3
LPA-1	1	0.2242562	0.2023935
LPA-2	0.2242562	1	0.9754067
LPA-3	0.2023935	0.9754067	1

4.2.5 Enrichment analysis

We performed functional enrichment analysis on the full integrated network gene set using DAVID [131]. After conducting pathway analysis on the full network, we expanded this analysis to investigate the community enrichment. To do this, we performed community detection on the integrated network. Then we performed community detection on each community. Using our previously developed method, spin glass community detection, we observed 25 relatively small communities. Of those 25 communities, we found 13 communities directly linked to the previously identified miRNA-gene associations.

4.2.6 Graph alignment of LPA integrated network and Spin-glass

We performed a graph alignment of the integrated networks from both methods. The graph alignment allowed us to measure the consistency of the edge interactions of the integrated networks constructed with spin-glass and the LPA method. We utilized the "GraphAlignment" package in R. First, we created an adjacency matrix from each of the edge-lists for both stages. Next, we calculated the node similarity between both networks by calculating the intersection of the two networks. Lastly, we extracted the weights for each interaction occurring in both networks. This was completed in the following R script `https://github.com/BenikaH/spinglass/blob/master/scripts/node_recovery.R`.

4.3 Results

We constructed a fully integrated multi-layered miRNA-gene network that includes miRNA-gene associations, correlations between miRNAs and genes, protein-protein interactions and a gene regulatory network. In the integrated network we have 174 miRNAs and 2,180 genes. These miRNAs and genes are connected through 803 regulatory edges, 1313 protein-protein edges, 9 correlated miRNAs, 18 correlated gene edges, 799 miRNA targets and a total of 855 miRNA-gene association edges (Figure 7).

We also identified a total of 251 genes that are somatic mutations or oncogenes in various cancers in the COSMIC database [144]. The integrated network revealed downstream target genes and subnetworks of several oncogenes including but not limited to, $AKT1$, $AKT2$, $TP53$, $FOXO3$ and $TGF-\beta 1$. As previously mentioned, these genes have critical roles in the onset and progression of ovarian cancer. Thus, exploiting the downstream targets across multiple layers of regulation.

4.3.1 Pathway Enrichment Analysis

Enrichment analysis revealed that over 50% (861) of our genes were identified in KEGG pathways and 120 of those genes were identified in cancer related pathways. We also found that many of the genes in our integrated network are involved in the $MAPK$ signaling pathway. The mitogen-activated protein kinase (MAPK)

pathway is reported to play an important role in cell migration, apoptosis, cell proliferation and cell differentiation [145]. These are critical functions that induce the progression of ovarian cancer. The community enrichment analysis revealed 9 out of 13 functionally enriched communities. Additional analysis revealed that 7 of the 9 enriched communities were enriched in cancer related pathways and various cell functions that impact tumor progression (Table 11). The integrated network highlighted downstream targets in subnetworks involved in various cancer related pathways. To investigate the interactions int he enriched pathways, we visualized the most enriched communities using Cytoscape [129]. Visualization of community 7 confirmed that it was enriched with multiple interactions found in cancer related pathways involved in bladder cancer, pancreatic cancer, prostate cancer and other critical signaling pathways(Figure 9). We found that it included interactions across all three layers in our integrated network which included the $E2F1$ oncogene.

4.3.2 Comparison with related methods

We compared these findings with GeNets [146] using the top 500 genes from our seed nodes and expanding with the InWeb protein-protein network. When only using the InWeb network, GeNets was able to identify 56 cancer related genes with only 3 of those specific to ovarian cancer. By adding the regulatory network in our analysis, the LPA algorithm was able to identify 4x more genes directly related to cancer

and nearly quadruple for ovarian cancer genes comparing with GeNets(Table 9). We also compared LPA to the spin-glass community detection method to extend the miRNA-gene network using the InWeb network as a template network [147]. The graph alignment we performed revealed that the integrated networks constructed by these two methods had an edge similarity of 58%. That means that LPA recovered 58% of the same interactions as the spin-glass method. We also found that many of the common interactions were affected by our miRNA-gene associations. Also, these conserved interactions were found in cancer related enriched pathways. When evaluating the robustness of LPA, out of our 3 LPA networks, we found that LPA-1 network and LPA-2 network had a similarity measure of .2242562, meaning these two networks had little similarity. LPA-1 network and LPA-3 network had a similarity measure of .2023935, meaning there was little similarity between these two networks. LPA-2 network and LPA-3 network had the most similarity with a cosine measure of .9754067. This means that the random seed initialization in our simulation had a strong influence on recovering the same network topology in each run and may impact the robustness of LPA. Let us consider other factors and discuss the properties of the networks integrated into the final LPA network. Given the density of the InWeb protein-protein network ($350\tilde{,}000$ interactions) and the addition of the gene regulatory network (RegNetwork), the LPA method recovered more interactions than

expected. The rationale of this expectation is that we used the InWeb network as a template in the spin-glass method and random seed node initialization. However, with LPA, we initialized our miRNA-gene associations as seed nodes and propagated the labels through the InWeb network and the RegNetwork using only the information of our seed genes. This suggests that with more labeled nodes in other networks, the percentage of recovery should increase drastically compared to using one subset of genes as seed nodes.

4.4 Discussion

Our results confirmed that there is collective action by multiple miRNAs, genes and other key interactions. We identified enriched communities across multiple networks of regulation, including miRNA-gene, protein-protein and regulatory networks. To exploit the large scope of regulation, we applied a network based learning approach to integrate multiple networks in ovarian cancer. In ovarian cancer and many other complex diseases, changes in miRNA expression and gene expression play a critical role in disease development and progression. Identifying more precise and detailed interactions on multiple levels of regulation can advance current prognosis and treatment procedures in ovarian cancer. Integrating multiple networks of miRNA-gene, protein-protein interactions, and regulatory networks is an effective way to explore these complex relationships. Such integrative methodology provides more insight on

understanding biological interactions in ovarian cancer and will lead to an advancement in current cancer therapies.

In comparison to the spin-glass model, LPA is more flexible allows for seed set network expansion via integrating multiple networks. While the spin-glass method was successful in network expansion, it produced a large number of communities and it was computationally expensive. We were unable to set seed nodes with the spin-glass model. Therefore, we chose to apply this method only to identify communities within our integrated network. Unlike the method proposed in the previous chapter, this method was able to predict the labels of unlabeled nodes using an LPA method. Thus, it is desirable to construct the integrated network utilizing multiple types of interactions while reducing the noise of irrelevant nodes and edges.

Also, our results showed that miRNAs regulate gene expression through multiple paths. The downstream genes in the subnetworks observed in this work are affected by miRNAs via protein-protein interactions and regulatory interactions. We used the miRNA affected genes as seed nodes to gain an understanding on the regulation of gene expression by miRNAs in these genetic networks. We observed that the miRNA-gene subnetworks consist of multiple interactions and are also enriched in multiple cancer related pathways. Some of the observed pathways had a high abundance of regulatory interactions affected by miRNAs. This shows how central regulatory

networks are in understanding the regulation in ovarian cancer phenotypes. Our results also showed that the integrated network constructed on ovarian cancer in this study, captures numerous cancer genes as a result of combining information from multiple genetic resources. Without this multi-layer approach, this information is not obtainable. This is beneficial to understanding the progression of ovarian cancer and could aid in improving early prognosis procedures.

This will expand the knowledge of candidate targets for therapeutic applications. In trying to understand how miRNAs affect disease progression through altered gene expression, we believe that we can exploit community enrichment at different stages of a disease using an integrated network such as the one generated in this study. To do so, we applied this theory and our developed methods to patients with stage II and stage III colon cancer to analyze the downstream genes in genetic networks for enriched network signatures. Additional results will show the benefits of integrating multiple networks to improve current prognostic knowledge in colon cancer.

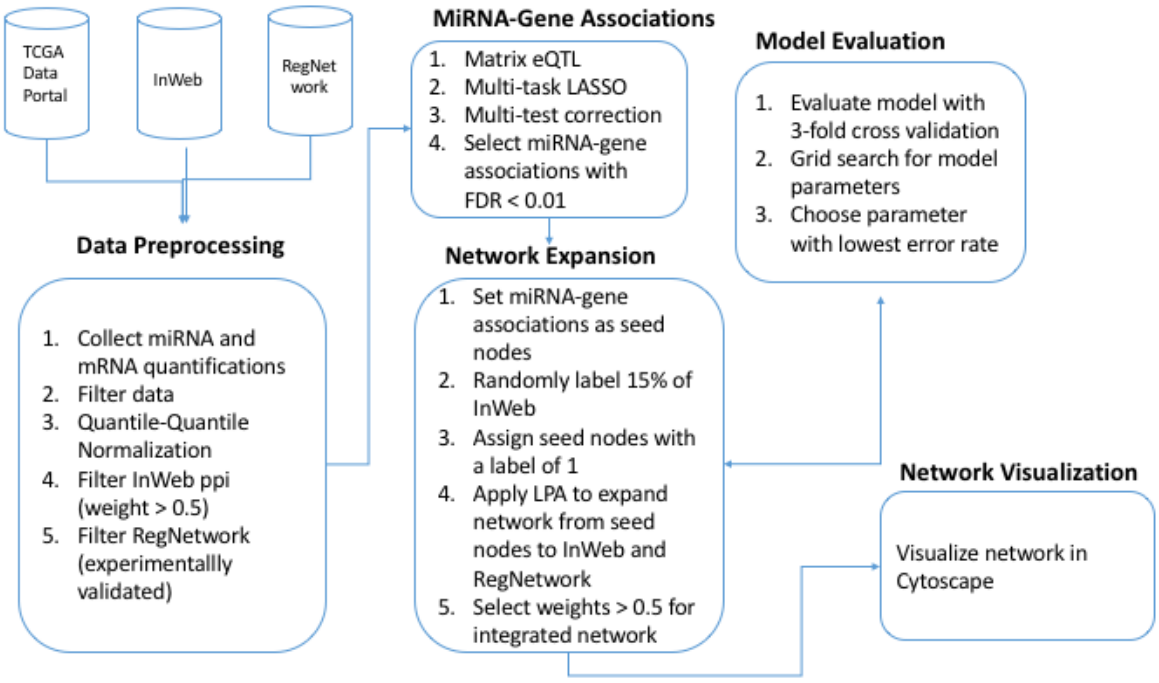

Figure 4: An overview of our integrative approach. We began with preprocessing the data. Then we identified miRNA-gene associations. We then expanded the network using LPA. We evaluated the model and tuned parameters. Next, we visualized the integrated network in Cytoscape.

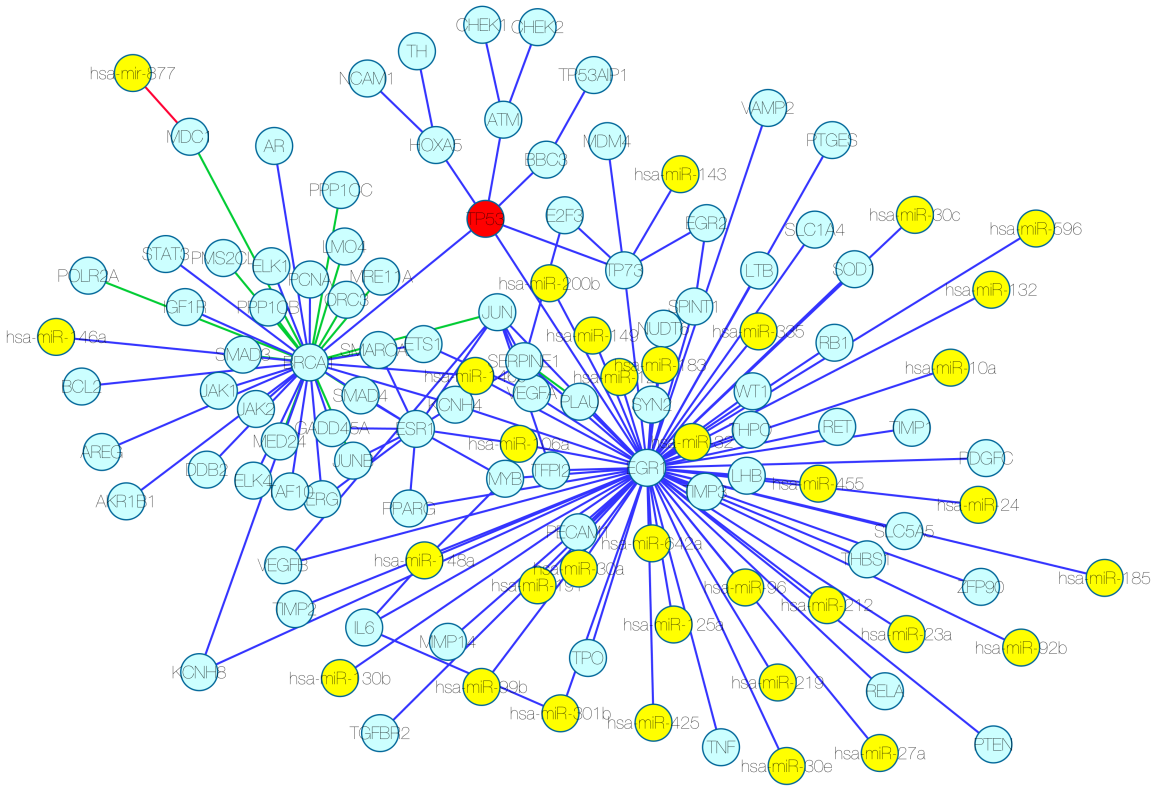

Figure 5: An example of our integrated network highlighting the $TP53$ suppressor gene. In this network, each yellow node represents a miRNA and light blue nodes represent a gene. The nodes connected with blue edges are from the RegNetwork, green edges are from InWeb network and the red edges are from our miRNA-gene network.

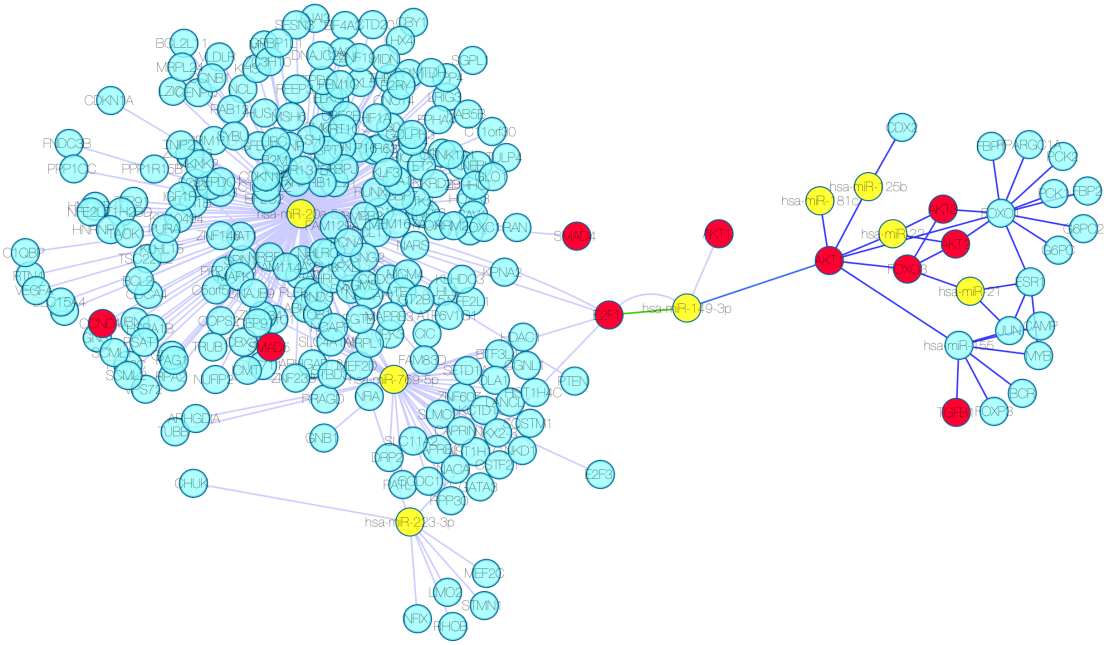

Figure 6: Another example of our integrated network highlighting the $AKT1$ genes responsible for activating the $PI3K/AKT1$ pathway. Cancer genes are depicted by red nodes. Each yellow node represents a miRNA and light blue nodes represent a gene. The nodes connected with blue edges are from the RegNetwork, miNRA targets are denoted by purple edges, green edges are from InWeb network and the red edges are from our miRNA-gene network.

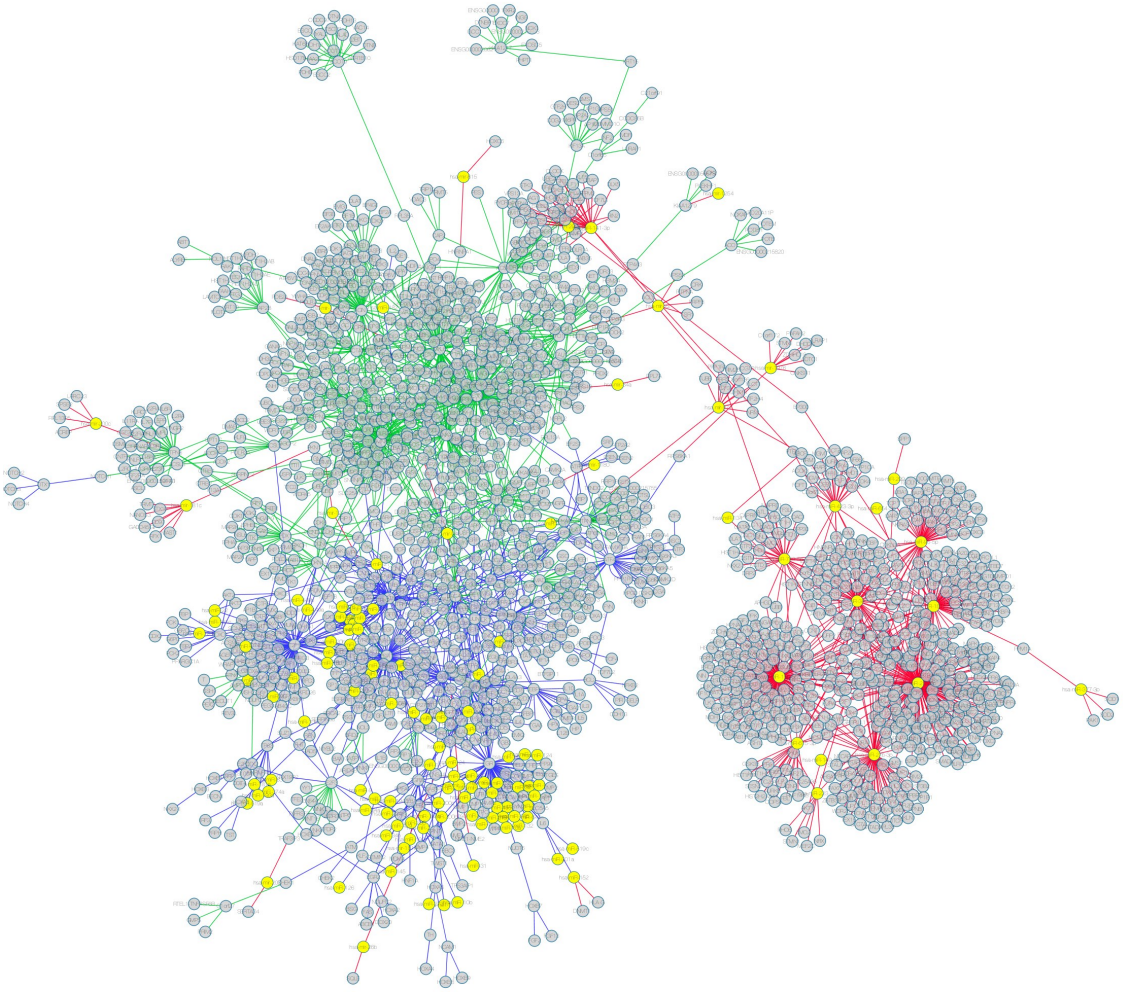

Figure 7: The fully integrated network after applying LPA . Each yellow node represents a miRNA and gray nodes represent interacting genes. The nodes connected with blue edges are from the RegNetwork, green edges are from InWeb network and the red edges are from our miRNA-gene network.

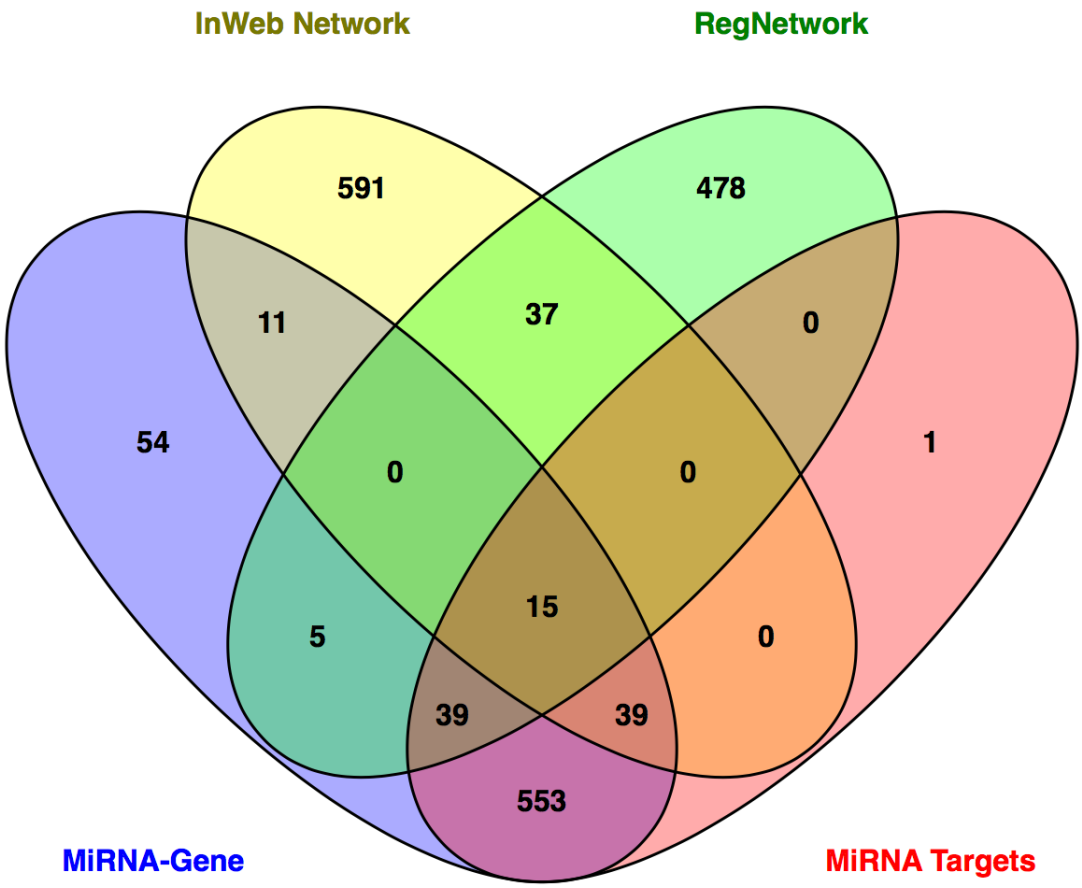

Figure 8: A Venn diagram showing the number of unique nodes in each network and the number of nodes in common between each network.

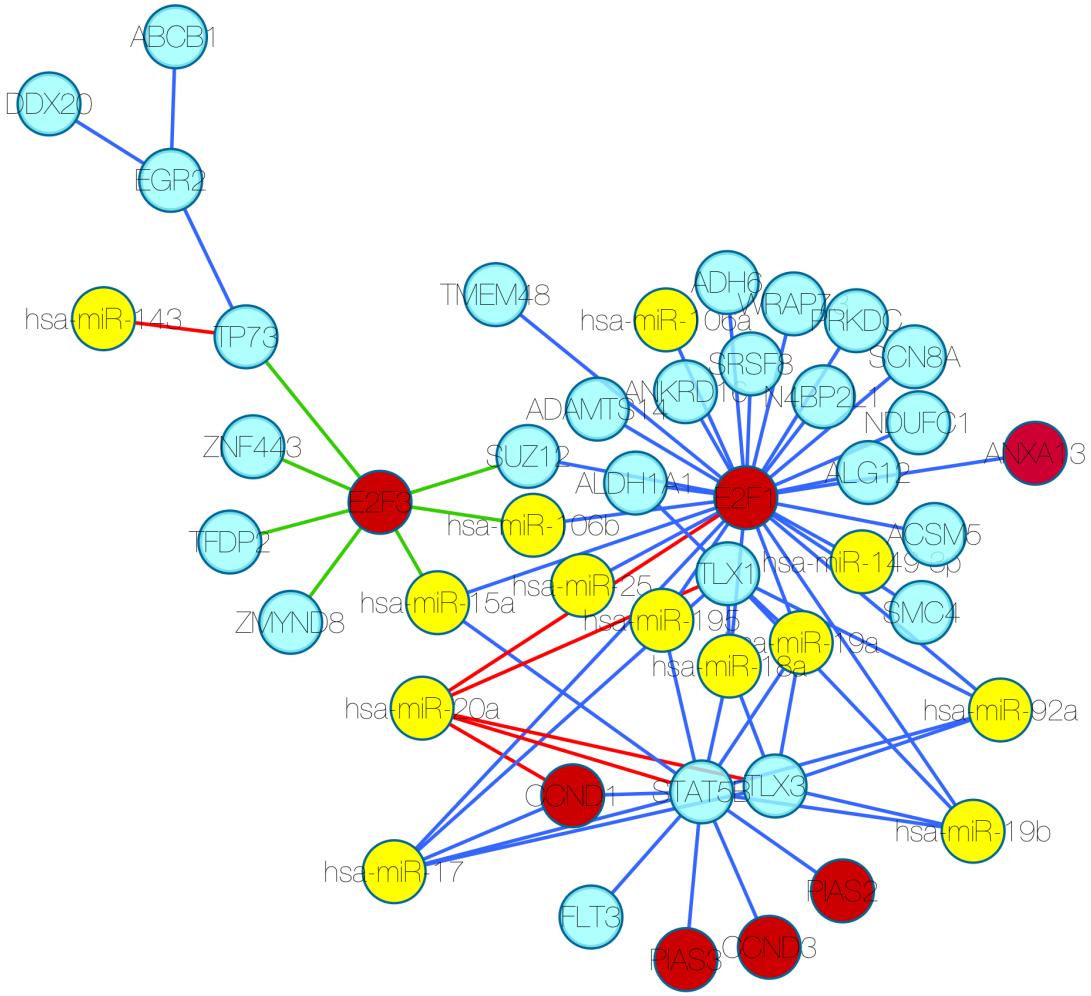

Figure 9: Enriched community 7 from our integrated network. Each yellow node represents a miRNA and light blue nodes represent a gene. The nodes connected with blue edges are from the RegNetwork, green edges are from InWeb network and the red edges are from our miRNA-gene network.

CHAPTER 5: IDENTIFYING MICRORNA-GENE NETWORKS SPECIFIC TO PATHOLOGIC STAGES IN COLON CANCER

5.1 Introduction

Colon cancer is one of the most common cancers in men and women today. Of the many cancers, it is reported to be one of the few curable cancers. However, the stage progression of colon cancer is very complex and makes for a poor prognosis. In stage I colon cancer, there is a 92% 5-year relative survival rate. For patients with stage II colon cancer, there are two stage subtypes: stage IIA and stage IIB colon cancers. There is an 87% 5-year relative survival rate for stage IIA and 63% for stage IIB. Similarly, for stage III colon cancer there are three subtypes: stage IIIA, IIIB and IIIC colon cancers. In patients with stage IIIA, the 5-year relative survival rate is 89%, for stage IIIB it is 69% and 53% for stage IIIC [114]. When the cancer has reached stage IV and metastasized to other parts of the body, the 5-year relative survival rate is decreased to approximately 12%. The drastic decrease in survival rate in colon cancer speaks to the need for better prognostic procedures.

The role of pathologic prognostic markers is important in the advancement of personalized medicine and can help reduce the risk of recurrence, especially in high-

risk patients with stage II colon cancer [148–150]. Due to the benefits of personalized medicine, these patients have an increased overall survival with therapies such as adjuvant chemotherapy. Gene expression signatures have shown much promise as prognostic markers [151]. For example, the progression of colon cancer is directly linked to the functional epithelial-mesenchymal transition (EMT) gene expression signatures [152]. Genes $ZEB1$ and $ZEB2$ are known repressors that regulate targets in the EMT pathway by changing the phenotype of normal cells to cancerous cells [153]. These genes are also known to be present in the beginning of metastasis.

Cell invasion and migration are also critical components in colon cancer progression. For instance, genes $PRKCQ$ and $PRKCZ$ are members of the protein kinase family and $PRKCZ$ is often involved in cell survival and cell migration in different cancers such as ovarian cancer [154]. It has also been reported that $ARID4B$ is a key player in pathogenesis and is classified as a metastasis modifier gene. Over-expression of this gene is thought to enhance the cell migration process as well as cell invasion. In contrast, the knockdown of $ARID4B$, causes metastasis of cancer cells to other regions of the body [155, 156]. More recently, miRNA expression profiles have been utilized as predictive markers for survival of colon cancer [157]. Studies have shown that specific miRNAs are directly related to poor prognosis in colon cancer as well as a poor therapeutic outcome [25,158,159]. For example, $miR-148$, $miR-26a-2$ and

$miR-130a$ were identified to be significantly associated with a poor clinical prognosis [160]. Exploiting the downstream neighborhoods of genes with such a critical role in the pathogenesis of colon cancer provide long-term benefits in personalized medicine and adjuvant therapies.

Recent studies have confirmed that there are genetic changes between stage II and stage III colon cancer [148, 161, 162]. However, it is unclear how these genetic changes drive the progression between stage II and stage III colon cancer. Therefore, in this study we focused on integrating genomics data and clinical phenotypes of colon cancer generated from the TCGA project. To demonstrate an application of our method, we applied this approach to study the relationships between miRNA expression, gene expression and clinical phenotypes to identify genetic and network patterns for stage II and stage III colon cancer. We aimed to identify pathologic prognostic network signatures such as enriched communities that are prevalent in stage II and stage III in colon cancer to understand how miRNAs affect gene expression in their downstream communities or neighborhoods. We hypothesized that as colon cancer progresses, there are unique network patterns present in stage II that are not present in stage III and network patterns specific in stage III that are not present in stage II. We believe that identifying these pathologic network signatures in enriched communities can lead to better prognosis and a better understanding in

the stage progression of colon cancer. The benefits of miRNA-gene networks will improve our understanding the complex biology of colon cancer.

5.2 Methods

The data used in this study generated by the TCGA Research Network: `http://cancergenome.nih.gov/`. Scripts used to transform the data for our analysis can be found at `https://github.com/BenikaH/ColonStudy/`. We retrieved miRNA expression data, gene expression data and clinical profiles for all colon cancer patients using the TCGA Data Portal located at `https://portal.gdc.cancer.gov`. We had a total of 196 patient samples with corresponding clinical profiles including information on pathologic stage, race and gender. We then chose to use stages II (82 samples) and stages III (59 samples) to run our miRNA-gene association analysis pipeline due to relatively small sample sizes for stage I (29 samples) and stage IV (26 samples) patient samples.

5.2.1 Data preprocessing

The data in this study was preprocessed using the same strategy implemented in Chapter 3. We removed samples with missing data. We retained all clinical profiles of patients whose pathologic stage was known. If the sample stage was unknown, we did not include that sample in the study. We also used the same scripts to perform the data transformation described in Chapter 3 to build miRNA

and gene expression matrices (https://github.com/BenikaH/ColonStudy/blob/master/scripts/miRMatrix.py, https://github.com/BenikaH/ColonStudy/blob/master/scripts/ExpressionMatrix-1.py. We then filtered the miRNA and gene expression matrices by removing rows with over 10% of missing values. The final expression matrix was used for our study.

5.2.2 MiRNA-gene associations

In this study we performed miRNA-gene association analysis between the miRNA expression and gene expression profiles in both stage II and stage III separately. The analysis was performed using Matrix eQTL [125], which is an R package that uses matrix operations to identify miRNA-gene associations. We performed *cis*-miRNA-gene association analysis, in which we only tested miRNA and gene pairs where miRNAs were within 1MB window from the genes. Prior to this, we extracted the gene positions from the human reference genome 37 using a command line tool called "gff2bed" https://github.com/BenikaH/ColonStudy/blob/master/scripts/Extract_positions_gff.txt. Next, similar to the process in Chapter 3, the miRNA-gene association analysis was completed with the following R code:(https://github.com/BenikaH/ColonStudy/blob/master/scripts/stage2_eqtl.R and https://github.com/BenikaH/ColonStudy/blob/master/scripts/stage3_eqtl.R).

5.2.3 Network expansion with the InWeb network

We ran the significant miRNA-gene associations from Matrix eQTL in the mirDB target prediction database [97] To expand our miRNA-gene association network for each stage, we used GeNets `https://apps.broadinstitute.org/genets` [146] to identify candidate genes in the InWeb protein-protein network. GeNets is a web-application developed by the Broad Institute for network analysis and visualization. We expanded our network with the InWeb protein-protein interaction network by identifying candidate genes in GeNets was able to import a gene list (max = 250) and expand to a single network, such as InWeb, BLAST, CLIME, ConsensusPathDB, GEO Expression, GeNets MetaNetwork and Achilles Cancer Co-dependency network to identify candidate genes. We chose our top ranked genes from the mirDB target prediction database. Using the pathway gene classifier within GeNets, we performed a pathway analysis and expanded our miRNA-gene network with the InWeb protein-protein interaction network. This was completed by determining the similarity of our test genes utilizing the similarity score within GeNets. The candidate genes were then identified based on the connectivity and the probability of being in the same pathways. Using this measure of connectivity, we expanded our network within GeNets by identifying candidate genes in the InWeb network. After expanding the network, we performed community detection analysis within GeNets for further analysis.

We performed a comparative analysis for all samples in stage II and samples in stage III by investigating miRNAs and miRNA-gene associations that were unique to each stage via enriched communities. First, we performed a pathway enrichment analysis on all genes in both stages separately within GeNets. By integrating miRNA-gene associations and miRNA targets, we exploited specific network signatures that were embedded in the community structures in different stages in colon cancer. In our stage II samples, we found that 4 of the 13 communities were enriched in pathways directly involved in various cancers, such as colorectal, bladder, thyroid, lung, prostate and pancreatic cancer. Similarly, we found 3 notable communities enriched in cancer related pathways in our stage III samples. We further investigated all communities to identify unique network signatures.

5.2.4 Network expansion via the Spin-glass model

In comparison with GeNets, we applied the spin-glass model to the stage II and stage III miRNA-gene associations generated with Matrix eQTL. From our previous evaluations of other methods such as Metis, GeNets and LPA, we decided to use spin-glass because of the following reasons: 1) Metis partitions networks into equal sized networks. Because we are interested in highly enriched communities, the equal-sized method may miss key interactions by dividing the graph into equal partitions. 2) GeNets limits the number of genes one can analyze to 250 genes. Due to the size of

our networks, this option was less effective. 3) Although, it did not allow seed initialization, the spin-glass model was consistent detecting interactions through multiple simulations. Also, the consistency was not impacted by the size of the networks. For these reasons, we found spin-glass was the most reliable for this comparison. We used the InWeb protein-protein interaction network as a template and expanded the network with miRNA-gene associations and predicted mirDB targets. For each stage we applied the spin-glass algorithm in the following scripts `https://github.com/BenikaH/ColonStudy/blob/master/scripts/stage2_spinglass.R` and `https://github.com/BenikaH/ColonStudy/blob/master/scripts/stage3_spinglass.R`.

5.2.5 Network alignment of Stage II and Stage III

We were also interested in seeing how the networks for stage II and stage III aligned with one another. Therefore, we performed a graph alignment to identify the commonality between the two networks. To perform the network alignment, we utilized the "GraphAlignment" package in R [130]. The "GraphAlignment" package uses the adjacency matrices of two networks to align them and take the intersection of the two graphs. Then it creates an adjacency matrix of the similar nodes found in the intersection resulting in a similarity score for the two graphs. Using this package, we first created an adjacency matrix from each of the edgelists for both stages. Next, we calculated the node similarity between both net-

works by calculating the intersection of the two networks. Lastly, we extracted the weights for each interaction occurring in both networks. This was completed in the following R script `https://github.com/BenikaH/ColonStudy/blob/master/scripts/stage2_stage3_graph_alignment.R`.

5.2.6 Statistical Analysis

We created a 2x2 contingency table (Table 5) to represent the number of interactions in each stage and in both stages. First, we calculated how many interactions were identified in both stage II and stage III (97). Next, we calculated the number of interactions that were in stage II but not in stage III (202). We then calculated the number of interactions that were in stage III but not in stage II (439). Lastly, we calculated all possible edge interactions that were not identified in either of the two stages (696,584). In order to compare the interactions found in the stage II network with the interactions found in the stage III network, we tested the null hypothesis that the interactions in the stage II network were equally likely to occur in the stage III network. Assuming this null hypothesis, we applied the Fisher's exact test on the 2x2 contingency table to generate a p-value (`https://github.com/BenikaH/ColonStudy/blob/master/scripts/fisher.test_stages.R`).

Table 5: A 2x2 contingency table representing the number of interactions found in stage II & III, the unique interactions found in stage II, the unique interactions found in stage III and the total possible interactions.

	Interactions in stage III	Interactions not in stage III
Interactions in stage II	97	202
Interactions not in stage II	439	696,584

5.3 Results

We identified 14 significant miRNA-gene associations (FDR ≤ 0.05) in stage II samples (Table 12) and 12 significant miRNA-gene associations in stage III samples) (Table 13). For stage II, there were 13 communities identified and 14 communities identified for stage III. In our stage II samples, we identified 27 enriched cancer pathways across 9 communities directly involved with various cancers (Table 14).

Results from our community enrichment analysis revealed that out of the 26 miRNA-gene associations identified, we observed two of these miRNA-gene in our stage II samples were directly associated with tumorigenesis, $miR-429$-$ZEB1$, $miR-429$-$ZEB2$ (Figure 10). We also found evidence that these miRNAs target multiple genes that are enriched in cancer-related pathways in colon and bladder cancer. Further investigation revealed downstream interactions in our study that are enriched in stage II, but are not present in stage III samples. While $miR-429$ was

significant in both samples, $miR-147b$ was specific only to stage II. In our network alignment, we discovered that there were only 97 ($\tilde{1}3\%$) shared interactions between the stage II and the stage III network. The similar interactions between stage II and stage III networks are shown in Table **??**. Pathway enrichment analysis revealed the pathways of the similar interactions between stage II and stage III. This evidence supports our hypothesis that the network patterns are significantly different in stage II and stage III colon cancer.

Additionally, we observed 16 enriched pathways directly involved in cancer across 5 communities (Table 15) in stage III samples. Results revealed a network signature containing $miR-429$, $PRKCQ$, $PRKCZ$, $ARID4$8. Further investigation is suggested to understand the downstream effects in the later stages of colon cancer. However, we believe this network signature and may be involved in the advanced stages of breast, lung, ovarian and colon cancer.

Results from our statistical analysis showed that there is a significant difference (p-value $<$ 2.2e-16) in the interactions identified in stage II and stage III. It is unlikely that the interactions identified in stage II samples equally occur in stage III samples. This was in line with our hypothesis that there are specific interactions unique to each stage in the progression of colon cancer.

5.4 Discussion

This study was an application of network expansion to understand how miRNAs affect gene expression and the downstream genetic neighborhoods of pathologic stages in colon cancer. We found that there is a significant difference in the presence of network edge interactions between the stages. We found that both network structures contained a similar number of communities. However, in our community enrichment analysis, we found more cancer enriched communities in stage II samples than in the stage III samples.

When evaluating the communities, we found that there were several interactions in stage III samples that were not present in stage II samples and vice versa, but they were not associated with our miRNAs. This could explain why we were only able to identify a small number of network signatures in each stage. Another explanation is that we had a limited sample size and many of the patients' prognosis was unknown. The size limitation of GeNets may have affected our candidate gene pool in the InWeb network. This was because GeNets only allowed a maximum of 250 genes as an input. This limited the number of possible candidate genes from the InWeb network within GeNets.

Lastly, there was strong evidence in our statistical analysis that there was a significant difference between the interactions identified in stage II and the interac-

tions identified in stage III. There was also evidence in provided literature that supported our hypothesis on the different roles of network signatures in the pathogenesis [163, 164], metastasis [165] and the tumorigenesis of colon cancer. Therefore, we do believe the interactions discovered and their cross-talk between the multiple layers in the integrated network.

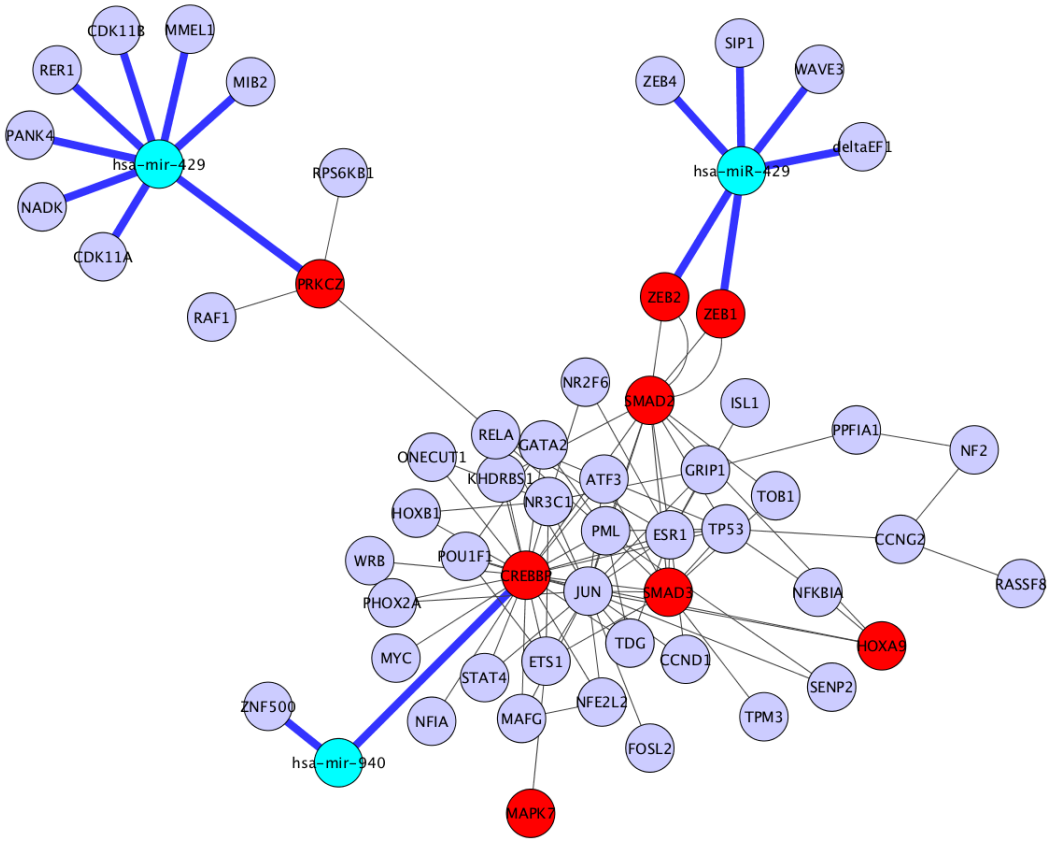

Figure 10: Enriched community 7 in our stage II samples. We visualized the network signature with $miR-429$, $ZEB1$, $ZEB2$ and $SMAD2$ which is only present in our samples with stage II prognosis. Red nodes are cancer genes, blue nodes are miRNAs associated with target genes and miRNA-gene associations are solid blue lines.

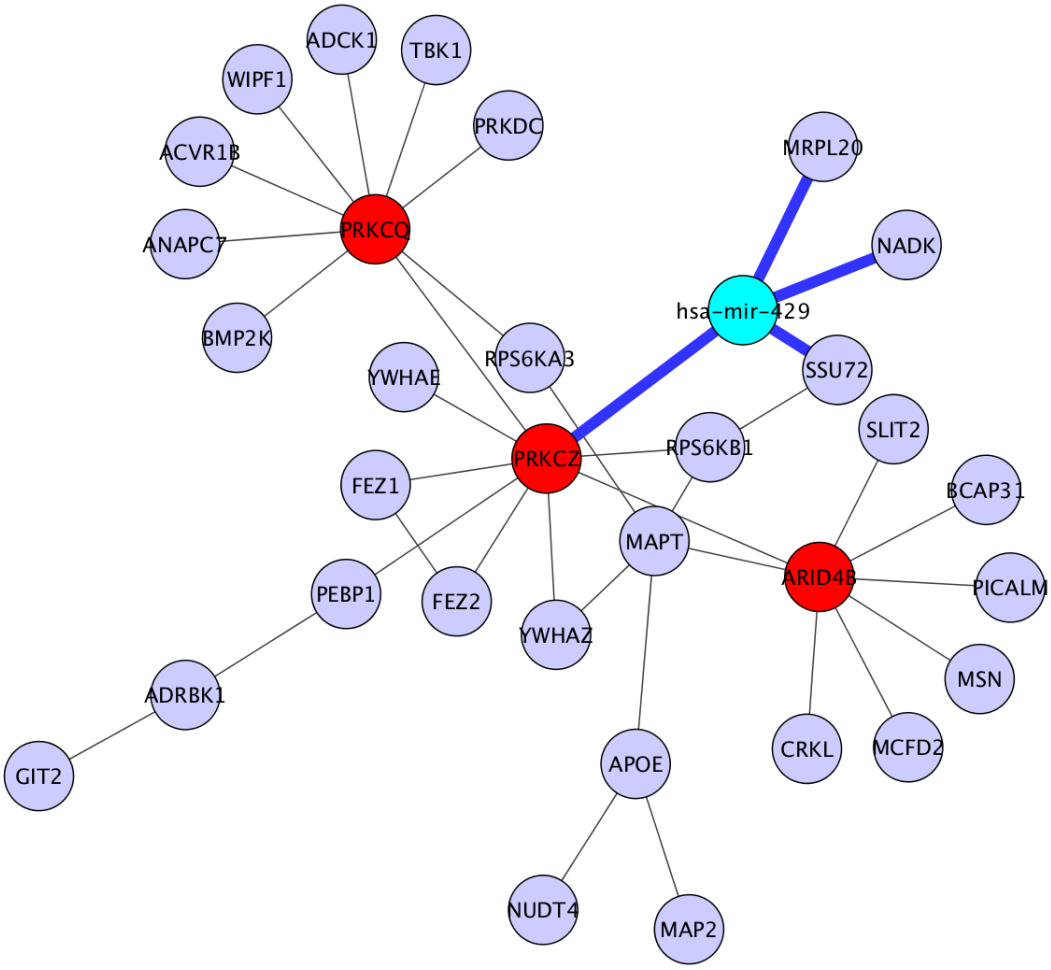

Figure 11: Enriched community in our stage III samples. The $miR-429$, $PRKCZ$ and $ARID4B$ interactions all have key roles in cell migration, invasion, pathogenesis and metastasis. Red nodes are cancer genes, blue nodes are miRNAs associated with target genes and miRNA-gene associations are solid blue lines.

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Conclusion

It has been shown that miRNAs are key regulators of gene expression. In fact, miRNAs have also been identified as tumor suppressors and oncogenes in various cancers. Moreover, miRNA-gene networks are embedded with miRNA-gene signatures that have diagnostic and prognostic significance. To better understand how miRNAs affect gene expression, we must first know more about the many complex relationships between miRNAs, genes and their downstream targets in genetic networks. Furthermore, we must understand the functional roles of miRNA-gene interactions and miRNA-gene subnetworks in cancer subtypes. There are numerous methods available to identify miRNA-gene relationships and construct miRNA-gene networks and this dissertation explores and compares these methods.

The first aim of the present research was to examine the relationship between miRNAs and gene expression in ovarian cancer by constructing an integrative miRNA-gene network via community-based network expansion. In this study, we addressed this by using the spin-glass community detection method to construct an integrated

miRNA-gene network and identify downstream communities in ovarian cancer. Our findings revealed that there is a complex relationship between miRNAs and their downstream targets and their neighbors embedded in smaller communities. Results also showed that these communities are functionally enriched in cancer related pathways that are. The evidence from this study suggests that miRNAs not only regulate expression levels of their immediate target genes but also downstream genes in genetic networks. The current findings add to a growing body of literature on the importance of miRNA-gene networks and their functional roles in ovarian cancer. A few limitations of this study is that it was computationally expensive and it produced extremely dense communities which were challenging to analyze as a whole.

The second aim of this research was to improve our network expansion methodology by integrating multiple networks to identify downstream relationships between miRNA expression, gene expression, protein-protein networks, gene regulatory networks and the downstream effects on genetic networks in patients with ovarian cancer. In this study we used the label propagation method to construct a multi-layer miRNA-gene network to investigate the miRNA-gene relationships in different levels of regulation in ovarian cancer. By expanding from miRNA-genes as seed nodes, the results of this investigation show that there are miRNAs that mediate multiple layers of regulation. This study has found that generally, miRNA-gene networks are also

embedded with smaller multi-layer subnetworks that are enriched in cancer related pathways. This research provides a framework for the exploration of integrating multiple networks and downstream analyses. Taken together, these findings enhance our understanding of how miRNAs affect gene expression on multiple levels in ovarian cancer. A limitation of this study is that we were unable to assess this study in other cancer subtypes.

The third aim of this study was to investigate the downstream effects of the relationship between miRNA-gene networks and the pathologic stages of colon cancer and expand our knowledge by identifying prognostic network signatures in enriched communities and patients with stage II and stage III colon cancer. In this study we used our spin-glass method to construct integrate miRNA-gene networks for stage II and stage III colon cancer samples. This study has identified novel prognostic network signatures in both stage II and stage III colon cancer. Moreover, the prognostic markers identified play pivotal roles in pathogenesis and metastasis in colon cancer. The results of this research support the idea that miRNA-gene network patterns are significantly different in stage II and stage III colon cancer. This study has raised important questions about the nature of prognosis of colon cancer and how miRNA-gene networks can improve the current knowledge. The present study also provides additional evidence with respect to the enrichment of cancer genes and prognostic

markers in miRNA-gene subnetworks in colon cancer. A limitation of this study is that it did not include additional layers such as gene regulatory networks, which we believe would yield more novel prognostic network signatures.

In summary, we developed frameworks to construct integrative miRNA-gene networks consisting of miRNA expression, miRNA correlations, gene correlation, gene expression, protein-protein networks, miRNA targets, clinical profiles and regulatory networks. This dissertation revealed methods to help to exploit the complexity of regulation, by constructing robust integrative miRNA-gene networks. We applied different approaches to attain this goal, specifically community based network approaches, and a semi-supervised learning approach to integrate multiple networks containing multiple regulatory elements in ovarian and colon cancer. Although there is much to learn about miRNA and gene expressions in cancer, the methods composed in this dissertation yield new ideas to inspect the relationships between miRNAs, genes, proteins and other regulatory elements in complex diseases. This will provide a better outlook on how these prognostic signatures can improve early detection techniques.

In terms of directions for future research, further work could involve several directions. For instance, one direction is to integrate cellular signaling networks with miRNA-gene networks. This direction can help researchers understand how miR-

NAs regulate cell migration via signaling networks. Because metastasis and tumor invasion are harmful consequences of cancerous cells migrating to other regions of the body, researchers can investigate this phenomenon using multi-layer networks.

Another direction for this research would be to integrate metabolic networks to explore how miRNA-gene networks respond to changes in expression and regulatory pathways involved in colon cancer. This direction will also allow researchers to study therapeutic strategies by investigating enriched metabolic pathways. By investigating the combination of miRNA-associated genes, proteins and metabolic pathways, one could understand metabolic responses in colon cancer and pancreatic cancer.

Further work needs to be done to establish whether the miRNA-gene relationships are drivers of their respective cancer related pathways. Moreover, it would also be interesting to compare the miRNA-gene networks and communities for multiple cancer subtypes to determine if there are conserved network patterns across various cancers. By investigating conserved miRNA-gene networks, researchers can identify enriched pathways across multiple cancer subtypes. In addition to this direction, these methods will be useful in studying prognostic network patterns within the different cancer subtypes. This will yield clinical benefits in current prognostic procedures.

In conclusion, the work in this dissertation was aimed to construct integrative miRNA-gene networks consisting of multiple data sources. The methods explained

in the previous chapters represent my intentions to explore the relationship between miRNAs and gene expression in ovarian and colon cancer. Moreover, in ovarian cancer and many other complex diseases, changes in miRNA expression and gene expression play a critical role in disease development and progression. In addition to cancer, the methods in this dissertation can also be applied to investigate other complex diseases. In fact, we have the means to provide more knowledge in miRNA-gene networks that can help establish better prognostic procedures and therapeutic strategies to improve human health.

REFERENCES

[1] Mark P Sawicki, Ghassan Samara, Michael Hurwitz, and Edward Passaro. Human genome project. *The American journal of surgery*, 165(2):258–264, 1993.

[2] Clare M O'Connor, Jill U Adams, and Jennifer Fairman. Essentials of cell biology. *Cambridge: NPG Education*, 2010.

[3] Barbara E Stranger and Emmanouil T Dermitzakis. The genetics of regulatory variation in the human genome. *Human genomics*, 2(2):126, 2005.

[4] A Ralston. Gene interaction and disease. *Nature education*, 1(1):16, 2008.

[5] Allan M Gurtan and Phillip A Sharp. The role of mirnas in regulating gene expression networks. *Journal of molecular biology*, 425(19):3582–3600, 2013.

[6] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

[7] Ramesh A Shivdasani. Micrornas: regulators of gene expression and cell differentiation. *Blood*, 108(12):3646–3653, 2006.

[8] Paul WC Hsu, Hsien-Da Huang, Sheng-Da Hsu, Li-Zen Lin, Ann-Ping Tsou, Ching-Ping Tseng, Peter F Stadler, Stefan Washietl, and Ivo L Hofacker. mirnamap: genomic maps of microrna genes and their target genes in mammalian genomes. *Nucleic acids research*, 34(suppl 1):D135–D139, 2006.

[9] Lutfi Tutar, Esen Tutar, and Yusuf Tutar. Micrornas and cancer; an overview. *Current pharmaceutical biotechnology*, 15(5):430–437, 2014.

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

[11] Aaron J Schetter, Hirokazu Okayama, and Curtis C Harris. The role of micrornas in colorectal cancer. *Cancer journal (Sudbury, Mass.)*, 18(3):244, 2012.

[12] Juan Manuel Moreno-Moya, Felipe Vilella, and Carlos Simón. Microrna: key gene expression regulators. *Fertility and sterility*, 101(6):1516–1523, 2014.

[13] Richard J Jackson and Nancy Standart. How do micrornas regulate gene expression. *Sci Stke*, 367(re1), 2007.

[14] M. Prokopi, C. A. Kousparou, and A. A. Epenetos. The secret role of micrornas in cancer stem cell development and potential therapy: A notch-pathway approach. *Frontiers in Oncology*, 4:389, 2014.

[15] Nicola Meola, Vincenzo Alessandro Gennarino, and Sandro Banfi. micrornas and genetic diseases. *Pathogenetics*, 2(1):1, 2009.

[16] Haiyan Zhai and Jingfang Ju. Implications of micrornas in colorectal cancer development, diagnosis, prognosis, and therapeutics. *Frontiers in genetics*, 2:78, 2011.

[17] Y Yang, H Meng, Q Peng, X Yang, R Gan, L Zhao, Z Chen, J Lu, and QH Meng. Downregulation of microrna-21 expression restrains non-small cell lung cancer cell proliferation and migration through upregulation of programmed cell death 4. *Cancer gene therapy*, 22(1):23–29, 2015.

[18] Jian Lu and Andrew G Clark. Impact of microrna regulation on variation in human gene expression. *Genome research*, 22(7):1243–1254, 2012.

[19] Shuqin Zhang and Michael K Ng. Gene-microrna network module analysis for ovarian cancer. *BMC systems biology*, 10(4):117, 2016.

[20] Min Ding, Jiang Li, Yong Yu, Hui Liu, Zi Yan, Jinghan Wang, and Qijun Qian. Integrated analysis of mirna, gene, and pathway regulatory networks in hepatic cancer stem cells. *Journal of translational medicine*, 13(1):259, 2015.

[21] Andrew Quitadamo, Lu Tian, Benika Hall, and Xinghua Shi. An integrated network of microrna and gene expression in ovarian cancer. *BMC bioinformatics*, 16(5):1, 2015.

[22] Yue Li, Cheng Liang, Ka-Chun Wong, Jiawei Luo, and Zhaolei Zhang. Mirsynergy: detecting synergistic mirna regulatory modules by overlapping neighbourhood expansion. *Bioinformatics*, 30(18):2627–2635, 2014.

[23] Xin Lai, Olaf Wolkenhauer, and Julio Vera. Understanding microrna-mediated gene regulatory networks through mathematical modelling. *Nucleic Acids Research*, page gkw550, 2016.

[24] Shihua Zhang, Qingjiao Li, Juan Liu, and Xianghong Jasmine Zhou. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics*, 27(13):i401–i409, 2011.

[25] Raquel Díaz, Javier Silva, José M García, Yolanda Lorenzo, Vanesa García, Cristina Peña, Rufo Rodríguez, Concepción Muñoz, Fernando García, Félix Bonilla, et al. Deregulated expression of mir-106a predicts survival in human colon cancer patients. *Genes, Chromosomes and Cancer*, 47(9):794–802, 2008.

[26] Rajesha Rupaimoole, George A Calin, Gabriel Lopez-Berestein, and Anil K Sood. mirna deregulation in cancer cells and the tumor microenvironment. *Cancer discovery*, 6(3):235–246, 2016.

[27] Edenir Inês Palmero, Silvana Gisele P de Campos, Marcelo Campos, Naiara C Souza, Ismael Dale C Guerreiro, Andre L Carvalho, and Marcia Maria C Marques. Mechanisms and role of microrna deregulation in cancer onset and progression. *Genetics and molecular biology*, 34(3):363–370, 2011.

[28] Da Yang, Yan Sun, Limei Hu, Hong Zheng, Ping Ji, Chad V Pecot, Yanrui Zhao, Sheila Reynolds, Hanyin Cheng, Rajesha Rupaimoole, et al. Integrated analyses identify a master microrna regulatory network for the mesenchymal subtype in serous ovarian cancer. *Cancer cell*, 23(2):186–199, 2013.

[29] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[30] Jianhua Ruan and Weixiong Zhang. An efficient spectral algorithm for network community discovery and its applications to biological and social networks. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 643–648. IEEE, 2007.

[31] Tejaswini Narayanan and Shankar Subramaniam. Community detection in biological networks using a variational bayes approach. In *BICoB*, pages 61–66, 2011.

[32] Tejaswini Narayanan, Merril Gersten, Shankar Subramaniam, and Ananth Grama. Modularity detection in protein-protein interaction networks. *BMC research notes*, 4(1):569, 2011.

[33] Franck Picard, Vincent Miele, Jean-Jacques Daudin, Ludovic Cottret, and Stéphane Robin. Deciphering the connectivity structure of biological networks using mixnet. *BMC bioinformatics*, 10(6):S17, 2009.

[34] Pratha Sah, Lisa O Singh, Aaron Clauset, and Shweta Bansal. Exploring community structure in biological networks with random graphs. *BMC bioinformatics*, 15(1):220, 2014.

[35] Zhao Yang, René Algesheimer, and Claudio J Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6, 2016.

[36] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[37] Koji Tsuda, Hyunjung Shin, and Bernhard Schölkopf. Fast protein classification with multiple networks. *Bioinformatics*, 21(suppl 2):ii59–ii65, 2005.

[38] Yen-Yi Ho et al. Modular network construction using eqtl data: an analysis of computational costs and benefits. *Frontiers in genetics*, 5:40–40, February 2014.

[39] Yen-Yi Ho, Leslie M Cope, and Giovanni Parmigiani. Modular network construction using eqtl data: an analysis of computational costs and benefits. 2014.

[40] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[41] Eric Eaton and Rachael Mansbach. *A Spin-Glass Model for Semi-Supervised Community Detection*. AAAI, 2012.

[42] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[43] Lei Pan, Chongjun Wang, and Junyuan Xie. A spin-glass model based local community detection method in social networks. *Tools with Artificial Intelligence (ICTAI)*, 2013, 2013.

[44] Jorg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74:1, 2006.

[45] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[46] Ludvig Bohlin, Daniel Edler, Andrea Lancichinetti, and Martin Rosvall. Community detection and visualization of networks with the map equation framework. In *Measuring Scholarly Impact*, pages 3–34. Springer, 2014.

[47] Leto Peel. Estimating network parameters for selecting community detection algorithms. In *Information Fusion (FUSION), 2010 13th Conference on*, pages 1–8. IEEE, 2010.

[48] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2):191–218, 2006.

[49] Scott Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of statistical physics*, 34(5-6):975–986, 1984.

[50] Lian En Chai, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, and Zalmiyah Zakaria. A review on the computational approaches for gene regulatory network construction. *Computers in biology and medicine*, 48:55–65, 2014.

[51] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[52] Jing Yu, V Anne Smith, Paul P Wang, Alexander J Hartemink, and Erich D Jarvis. Using bayesian network inference algorithms to recover molecular genetic regulatory networks. In *International Conference on Systems Biology*, volume 2002, 2002.

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

[54] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[55] Nguyen Xuan Vinh, Madhu Chetty, Ross Coppel, and Pramod P Wangikar. Gene regulatory network modeling via global optimization of high-order dynamic bayesian network. *BMC bioinformatics*, 13(1):131, 2012.

[56] William Chad Young, Adrian E Raftery, and Ka Yee Yeung. Fast bayesian inference for gene regulatory networks using scanbma. *BMC systems biology*, 8(1):47, 2014.

[57] Jia Wang, Huihui Yu, Xiaoyu Weng, Weibo Xie, Caiguo Xu, Xianghua Li, Jinghua Xiao, and Qifa Zhang. An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. *Journal of experimental botany*, page ert464, 2014.

[58] Jeffrey D Allen, Yang Xie, Min Chen, Luc Girard, and Guanghua Xiao. Comparing statistical methods for constructing large scale gene networks. *PloS one*, 7(1):e29348, 2012.

[59] Masayuki Karasuyama and Hiroshi Mamitsuka. Adaptive edge weighting for graph-based learning algorithms. *Machine Learning*, 106(2):307–335, 2017.

[60] Bo Long, Xiaoyun Xu, Zhongfei Zhang, and S Yu Philip. Community learning by graph approximation. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 232–241. IEEE, 2007.

[61] Amarnag Subramanya and Partha Pratim Talukdar. Graph-based semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(4):1–125, 2014.

[62] Wei Liu, Jun Wang, and Shih-Fu Chang. Robust and scalable graph-based semisupervised learning. *Proceedings of the IEEE*, 100(9):2624–2638, 2012.

[63] Wei Liu, Junfeng He, and Shih-Fu Chang. Large graph construction for scalable semi-supervised learning. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 679–686, 2010.

[64] Yasuhiro Fujiwara and Go Irie. Efficient label propagation. In *Proceedings of the 31st international conference on machine learning (ICML-14)*, pages 784–792, 2014.

[65] Masayuki Karasuyama and Hiroshi Mamitsuka. Multiple graph label propagation by sparse integration. *IEEE transactions on neural networks and learning systems*, 24(12):1999–2012, 2013.

[66] TaeHyun Hwang and Rui Kuang. A heterogeneous label propagation algorithm for disease gene discovery. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 583–594. SIAM, 2010.

[67] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.

[68] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):036106, 2007.

[69] Jierui Xie, Mingming Chen, and Boleslaw K Szymanski. Labelrankt: Incremental community detection in dynamic networks via label propagation. In *Proceedings of the Workshop on Dynamic Networks Management and Mining*, pages 25–32. ACM, 2013.

[70] Jierui Xie, Boleslaw K Szymanski, and Xiaoming Liu. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 344–349. IEEE, 2011.

[71] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[72] Zheng-Yu Niu, Dong-Hong Ji, and Chew Lim Tan. Word sense disambiguation using label propagation based semi-supervised learning. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 395–402. Association for Computational Linguistics, 2005.

[73] Zheng Lin, Songbo Tan, Yue Liu, Xueqi Cheng, and Xueke Xu. Cross-language opinion lexicon extraction using mutual-reinforcement label propagation. *PloS one*, 8(11):e79294, 2013.

[74] Rolf A Heckemann, Joseph V Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.

[75] Maryam Lotfi Shahreza, Nasser Ghadiri, Seyed Rasul Mossavi, Jaleh Varshosaz, and James Green. Heter-lp: A heterogeneous label propagation algorithm and its application in drug repositioning. *arXiv preprint arXiv:1611.02945*, 2016.

[76] Ping Zhang, Fei Wang, Jianying Hu, and Robert Sorrentino. Label propagation prediction of drug-drug interactions based on clinical side effects. *Scientific reports*, 5:12339, 2015.

[77] Pingzhao Hu, Gary Bader, Dennis A Wigle, and Andrew Emili. Computational prediction of cancer-gene function. *Nature Reviews Cancer*, 7(1):23–34, 2007.

[78] Marco Frasca and Nicolò Cesa Bianchi. Multitask protein function prediction through task dissimilarity. *arXiv preprint arXiv:1611.00962*, 2016.

[79] Tsuyoshi Kato, Hisahi Kashima, and Masashi Sugiyama. Robust label propagation on multiple networks. *IEEE Transactions on Neural Networks*, 20(1):35–44, 2009.

[80] Dominique LaSalle and George Karypis. Multi-threaded graph partitioning. In *Parallel & Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on*, pages 225–236. IEEE, 2013.

[81] Aydın Buluç, Henning Meyerhenke, Ilya Safro, Peter Sanders, and Christian Schulz. Recent advances in graph partitioning. In *Algorithm Engineering*, pages 117–158. Springer, 2016.

[82] George Karypis and Vipin Kumar. Multilevelk-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed computing*, 48(1):96–129, 1998.

[83] G Karypis and V Kumar. Parallel multilevel k-way partitioning scheme for irregular graphs, department of computer science tech. rep. 96-036. *University of Minnesota, Minneapolis, MN*, 1996.

[84] Hua Yang, William Kong, Lili He, Jian-Jun Zhao, Joshua D O'Donnell, Jiawang Wang, Robert M Wenham, Domenico Coppola, Patricia A Kruk, Santo V Nicosia, et al. Microrna expression profiling in human ovarian cancer: mir-214 induces cell survival and cisplatin resistance by targeting pten. *Cancer research*, 68(2):425–433, 2008.

[85] Francesca M Buffa, Carme Camps, Laura Winchester, Cameron E Snell, Harriet E Gee, Helen Sheldon, Marian Taylor, Adrian L Harris, and Jiannis Ragoussis. microrna associated progression pathways and potential therapeutic targets identified by integrated mrna and microrna expression profiling in breast cancer. *Cancer research*, pages canres–0489, 2011.

[86] Isaac Bentwich, Amir Avniel, Yael Karov, Ranit Aharonov, Shlomit Gilad, Omer Barad, Adi Barzilai, Paz Einat, Uri Einav, Eti Meiri, et al. Identification of hundreds of conserved and nonconserved human micrornas. *Nature genetics*, 37(7):766–770, 2005.

[87] Hai-Son Le and Ziv Bar-Joseph. Integrating sequence, expression and interaction data to determine condition-specific mirna regulation. *Bioinformatics*, 29(13):i89–i97, 2013.

[88] Gianpiero Di Leva, Michela Garofalo, and Carlo M Croce. Micrornas in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 9:287–314, 2014.

[89] Pai-Sheng Chen, Jen-Liang Su, and Mien-Chie Hung. Dysregulation of micrornas in cancer. *Journal of biomedical science*, 19(1):90, 2012.

[90] George Adrian Calin, Calin Dan Dumitru, Masayoshi Shimizu, Roberta Bichi, Simona Zupo, Evan Noch, Hansjuerg Aldler, Sashi Rattan, Michael Keating,

Kanti Rai, et al. Frequent deletions and down-regulation of micro-rna genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*, 99(24):15524–15529, 2002.

[91] Kamini Sundarbose, Reena V Kartha, and Subbaya Subramanian. Micrornas as biomarkers in cancer. *Diagnostics*, 3(1):84–104, 2013.

[92] Susan R Pfeffer, Chuan He Yang, and Lawrence M Pfeffer. The role of mir-21 in cancer. *Drug development research*, 76(6):270–277, 2015.

[93] Lindsey E Becker Buscaglia and Yong Li. Apoptosis and the target genes of mir-21. *Chinese journal of cancer*, 30(6):371, 2011.

[94] Behzad Rabiee-Ghahfarrokhi, Fariba Rafiei, Ali Akbar Niknafs, and Behzad Zamani. Prediction of microrna target genes using an efficient genetic algorithm-based decision tree. *FEBS open bio*, 5:877–884, 2015.

[95] Bing Liu, Jiuyong Li, and Murray J Cairns. Identifying mirnas, targets and functions. *Briefings in bioinformatics*, 15(1):1–19, 2012.

[96] Carlo M Croce. Causes and consequences of microrna dysregulation in cancer. *Nature reviews. Genetics*, 10(10):704, 2009.

[97] Xiaowei Wang. mirdb: a microrna target prediction and functional annotation database with a wiki interface. *Rna*, 14(6):1012–1017, 2008.

[98] Yukinori Okada, Tomoki Muramatsu, Naomasa Suita, Masahiro Kanai, Eiryo Kawakami, Valentina Iotchkova, Nicole Soranzo, Johji Inazawa, and Toshihiro Tanaka. Significant impact of mirna–target gene networks on genetics of human complex traits. *Scientific reports*, 6:22223, 2016.

[99] ND Mendes, Ana T Freitas, and M-F Sagot. Current tools for the identification of mirna genes and their targets. *Nucleic acids research*, 37(8):2419–2433, 2009.

[100] Cesaré Ovando-Vázquez, Daniel Lepe-Soltero, and Cei Abreu-Goodger. Improving microrna target prediction with gene expression profiles. *BMC genomics*, 17(1):364, 2016.

[101] Fengfeng Wang, SC Wong, Lawrence WC Chan, William Cho, SP Yip, and Benjamin YM Yung. Multiple regression analysis of mrna-mirna associations in colorectal cancer pathway. *BioMed research international*, 2014, 2014.

[102] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microrna target sites in mammalian mrnas. *elife*, 4:e05005, 2015.

[103] Nathan Wong and Xiaowei Wang. mirdb: an online resource for microrna target prediction and functional annotations. *Nucleic acids research*, 43(D1):D146–D152, 2014.

[104] Giorgos L Papadopoulos, Martin Reczko, Victor A Simossis, Praveen Sethupathy, and Artemis G Hatzigeorgiou. The database of experimentally supported targets: a functional update of tarbase. *Nucleic acids research*, 37(suppl 1):D155–D158, 2009.

[105] Sheng-Da Hsu, Feng-Mao Lin, Wei-Yun Wu, Chao Liang, Wei-Chih Huang, Wen-Ling Chan, Wen-Ting Tsai, Goun-Zhou Chen, Chia-Jung Lee, Chih-Min Chiu, et al. mirtarbase: a database curates experimentally validated microrna–target interactions. *Nucleic acids research*, 39(suppl_1):D163–D169, 2010.

[106] Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu. mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic acids research*, 37(suppl 1):D98–D104, 2009.

[107] Boya et al. Xie. mircancer: a microrna cancer association database constructed by text mining on literature. *Bioinformatics*, 14, 2013.

[108] Alexandra Drakaki and Dimitrios Iliopoulos. Microrna gene networks in oncogenesis. *Current genomics*, 10(1):35–41, 2009.

[109] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[110] Mohammed Mamdani, Vernell Williamson, Gowon O McMichael, Tana Blevins, Fazil Aliev, Amy Adkins, Laura Hack, Tim Bigdeli, Andrew D van der Vaart, Bradley Todd Web, et al. Integrating mrna and mirna weighted gene co-expression networks with eqtls in the nucleus accumbens of subjects with alcohol dependence. *PloS one*, 10(9):e0137671, 2015.

[111] Inma Tur Mongé. Mapping eqtl networks with mixed graphical models. 2014.

[112] Sherry Freiesleben, Michael Hecker, Uwe Klaus Zettl, Georg Fuellen, and Leila Taher. Analysis of microrna and gene expression profiles in multiple sclerosis: Integrating interaction data to uncover regulatory mechanisms. *Scientific reports*, 6, 2016.

[113] Elizabeth J Rossin, Kasper Lage, Soumya Raychaudhuri, Ramnik J Xavier, Diana Tatar, Yair Benita, Chris Cotsapas, Mark J Daly, International Inflammatory Bowel Disease Genetics Constortium, et al. Proteins encoded in ge-

nomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS genetics*, 7(1):e1001273, 2011.

[114] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67(1):7–30, 2017.

[115] S. et al. Prislei. Mir-200c and hur in ovarian cancer. *BMC Cancer*, 13:72, 2013.

[116] Chen et al. A two-graph guided multi-task lasso approach for eqtl mapping. In *Proceedings of the 15th International Conference of Artificial Intelligence and Statistics (AISTATS)*, pages 208–217, Journal of Machine Learning Research (JMLR) W&CP 22, 2012.

[117] Cancer Genome Atlas Research Network:. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609–615, 2011.

[118] The cancer genome atlas-genomic data commons.

[119] Kasper Lage, E Olof Karlberg, Zenia M Størling, Páll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, 25(3):309–316, 2007.

[120] Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni. Mint: a molecular interaction database. *FEBS letters*, 513(1):135–140, 2002.

[121] Gary D Bader, Doron Betel, and Christopher WV Hogue. Bind: the biomolecular interaction network database. *Nucleic acids research*, 31(1):248–250, 2003.

[122] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl_1):D428–D432, 2005.

[123] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[124] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

[125] Andrey a. Shabalin. Matrix eqtl: Ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10)::1353–1358, 05 2012.

[126] Stephanie C Hicks and Rafael A Irizarry. When to use quantile normalization? *bioRxiv*, page 012203, 2014.

[127] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[128] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, 1695, 2006.

[129] P. et al. Shannon. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:11, 2003.

[130] Johannes Berg and Michael Lässig. Cross-species analysis of biological networks by bayesian alignment. *Proceedings of the National Academy of Sciences*, 103(29):10967–10972, 2006.

[131] Xiaoli Jiao, Brad T Sherman, Da Wei Huang, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. David-ws: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13):1805–1806, 2012.

[132] Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011.

[133] Online Mendelian. *Inheritance in Man (OMIM)*. URL.

[134] George Karypis and V Kumar. Metis-serial graph partitioning and fill-reducing matrix ordering, 2012.

[135] George Karypis, Kirk Schloegel, and Vipin Kumar. Parmetis: Parallel graph partitioning and sparse matrix ordering library. *Version 1.0, Dept. of Computer Science, University of Minnesota*, 1997.

[136] Rotem Ben-Hamo and Sol Efroni. Gene expression and network-based analysis reveals a novel role for hsa-mir-9 and drug control over the p38 network in glioblastoma multiforme progression. *Genome medicine*, 3(11):77, 2011.

[137] Li Chen, Jianhua Xuan, Rebecca B Riggins, Robert Clarke, and Yue Wang. Identifying cancer biomarkers by network-constrained support vector machines. *BMC systems biology*, 5(1):161, 2011.

[138] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in cell and developmental biology*, 2:38, 2014.

[139] Marcus Q Bernardini, Tsukasa Baba, Paula S Lee, Jason C Barnett, Gregory P Sfakianos, Angeles Alvarez Secord, Susan K Murphy, Edwin Iversen, Jeffrey R

Marks, and Andrew Berchuck. Expression signatures of tp53 mutations in serous ovarian cancers. *BMC cancer*, 10(1):237, 2010.

[140] Wenyuan Li, Shihua Zhang, Chun-Chi Liu, and Xianghong Jasmine Zhou. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, 28(19):2458–2466, 2012.

[141] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17(Suppl 2):15, 2016.

[142] Claudio Angione, Max Conway, and Pietro Lió. Multiplex methods provide effective integration of multi-omic data in genome-scale models. *BMC bioinformatics*, 17(4):83, 2016.

[143] Zhi-Ping Liu, Canglin Wu, Hongyu Miao, and Hulin Wu. Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, 2015:bav095, 2015.

[144] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, 2004.

[145] Amardeep Singh Dhillon, Suzanne Hagan, O Rath, and W Kolch. Map kinase signalling pathways in cancer. *Oncogene*, 26(22):3279–3290, 2007.

[146] Mercer, j. d., et al (n.d.). an analysis of differential pathway topologies in diverse functional genomics networks. manuscript in preparation. retrieved from http://www.broadinstitute.org/genets.

[147] Benika Hall, Andrew Quitadamo, and Xinghua Shi. Identifying microrna and gene expression networks using graph communities. *Tsinghua Science and Technology*, 21(2):176–195, 2016.

[148] Efrat Dotan and Steven J Cohen. Challenges in the management of stage ii colon cancer. In *Seminars in oncology*, volume 38, pages 511–520. Elsevier, 2011.

[149] Saima Sharif and Michael J OâĂŹConnell. Gene signatures in stage ii colon cancer: a clinical review. *Current colorectal cancer reports*, 8(3):225–231, 2012.

[150] Rodrigo Dienstmann, Ramon Salazar, and Josep Tabernero. Personalizing colon cancer adjuvant therapy: selecting optimal treatments for individual patients. *Journal of Clinical Oncology*, 33(16):1787–1796, 2015.

[151] Piero Dalerba, Debashis Sahoo, Soonmyung Paik, Xiangqian Guo, Greg Yothers, Nan Song, Nate Wilcox-Fogel, Erna Forgó, Pradeep S Rajendran,

Stephen P Miranda, et al. Cdx2 as a prognostic biomarker in stage ii and stage iii colon cancer. *New England Journal of Medicine*, 374(3):211–222, 2016.

[152] Christine Woischke, Cristina Blaj, Eva Marina Schmidt, Sebastian Lamprecht, Jutta Engel, Heiko Hermeking, Thomas Kirchner, and David Horst. Cyb5r1 links epithelial-mesenchymal transition and poor prognosis in colorectal cancer. *Oncotarget*, 7(21):31350, 2016.

[153] JA Galván, Inti Zlobec, Martin Wartenberg, Alessandro Lugli, Beat Gloor, Aurel Perren, and Evanthia Karamitopoulou. Expression of e-cadherin repressors snail, zeb1 and zeb2 by tumour and stromal cells influences tumour-budding phenotype and suggests heterogeneity of stromal cells in pancreatic cancer. *British journal of cancer*, 112(12):1944–1950, 2015.

[154] Kelly KY Seto and Irene L Andrulis. Atypical protein kinase c zeta: Potential player in cell survival and cell migration of ovarian cancer. *PloS one*, 10(4):e0123528, 2015.

[155] Mei-Yi Wu, Karen W Eldin, and Arthur L Beaudet. Identification of chromatin remodeling genes arid4a and arid4b as leukemia suppressor genes. *Journal of the National Cancer Institute*, 100(17):1247–1259, 2008.

[156] Scott F Winter, Luanne Lukes, Renard C Walker, Danny R Welch, and Kent W Hunter. Allelic variation and differential expression of the msin3a histone deacetylase complex gene arid4b promote mammary tumor growth and metastasis. *PLoS Genet*, 8(5):e1002735, 2012.

[157] P Li, Q Ou, G Chen, and FS Oduncu. Microrna as predictive biomarker of survival for stage iiib colon cancer patients. *Annals of Oncology*, 27(suppl 6):532P, 2016.

[158] Jialu Li, Yang Liu, Cheng Wang, Ting Deng, Hongwei Liang, Yifei Wang, Dingzhi Huang, Qian Fan, Xia Wang, Tao Ning, et al. Serum mirna expression profile as a prognostic biomarker of stage ii/iii colorectal adenocarcinoma. *Scientific reports*, 5:12921, 2015.

[159] Aaron J Schetter, Suet Yi Leung, Jane J Sohn, Krista A Zanetti, Elise D Bowman, Nozomu Yanaihara, Siu Tsan Yuen, Tsun Leung Chan, Dora LW Kwong, Gordon KH Au, et al. Microrna expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. *Jama*, 299(4):425–436, 2008.

[160] Jian Xu, Jian Zhao, and Rui Zhang. Four micrornas signature for survival prognosis in colon cancer using tcga data. *Scientific Reports*, 6, 2016.

[161] Steven A Frank. Multistage progression. 2007.

[162] Elrasheid AH Kheirelseid, Nicola Miller, Kah Hoong Chang, Mary Nugent, and Michael J Kerin. Clinical applications of gene expression in colorectal cancer. *Journal of gastrointestinal oncology*, 4(2):144, 2013.

[163] Weihua Li, Shaoxin Cai, Le Wang, Changshun Yang, Biaohuan Zhou, and Huan Wang. Hint2 downregulation promotes colorectal carcinoma migration and metastasis. *Oncotarget*, 8(8):13521, 2017.

[164] Juliette Martin, Marie V St-Pierre, and Jean-François Dufour. Hit proteins, mitochondria and cancer. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1807(6):626–632, 2011.

[165] Martin Hart, Stefanie Rheinheimer, Petra Leidinger, Christina Backes, Jennifer Menegatti, Tobias Fehlmann, Friedrich Grässer, Andreas Keller, and Eckart Meese. Identification of mir-34a-target interactions by a combined network based and experimental approach. *Oncotarget*, 7(23):34288, 2016.

APPENDIX: SUPPLEMENTARY TABLES

Table 6: A summary table showing the genes affected by miRNAs that overlapped with disease genes according to the OMIM database.

Gene	Description	Disease/Phenotype
CDKN1C, KIP2, BWS, IMAGE	Cyclin-dependent kinase inhibitor 1C (p57, Kip2)	IMAGE syndrome
CDKN1C, KIP2, BWS, IMAGE	Cyclin-dependent kinase inhibitor 1C (p57, Kip2)	Beckwith-Wiedemann syndrome
DGCR2, DGS2	.l.o;-.?[;/ syndrome chromosome region-2	DiGeorge syndrome/velocardiofacial syndrome complex-2
HNRNPA1, IBMPFD3, ALS20	Heterogeneous nuclear ribonucleoprotein A1	Inclusion body myopathy wtih early-onset Paget disease without frontotemporal dementia 3
HNRNPA1, IBMPFD3, ALS20	Heterogeneous nuclear ribonucleoprotein A1	Amyotrophic lateral sclerosis 20
KIAA1279	KIAA1279 gene	Goldberg-Shprintzen megacolon syndrome
KIF1A, ATSV, UNC104, SPG30, HSN2C, MRD9	Kinesin family member 1A	Spastic paraplegia 30, autosomal recessive
KIF1A, ATSV, UNC104, SPG30, HSN2C, MRD9	Kinesin family member 1A	Neuropathy, hereditary sensory, type IIC
KIF1A, ATSV, UNC104, SPG30, HSN2C, MRD9	Kinesin family member 1A	Mental retardation, autosomal dominant 9
LDLRAP1, ARH, FHCB2, FHCB1	Low density lipoprotein receptor adaptor protein 1	Hypercholesterolemia, familial, autosomal recessive
NUMA1	Nuclear mitotic apparatus protein-1	Leukemia, acute promyelocytic, NUMA/RARA type
POMC	Proopiomelanocortin (adrenocorticotropin/beta-lipotropin)	Obesity, adrenal insufficiency, and red hair due to POMC deficiency
POMC	Proopiomelanocortin (adrenocortic'otropin/beta-lipotropin)	Obesity, early-onset, susceptibility to
RECQL4, RTS, RECQ4	DNA helicase, RecQ-like 4	Rothmund-Thomson syndrome
RECQL4, RTS, RECQ4	DNA helicase, RecQ-like 4	RAPADILINO syndrome
RECQL4, RTS, RECQ4	DNA helicase, RecQ-like 4	Baller-Gerold syndrome
SLC25A3, PHC	Solute carrier family 25 (mitochondrial carrier), member 3	Mitochondrial phosphate carrier deficiency
TUBB, TUBB5, M40, CDCBM6	Tubulin, beta polypeptide	Cortical dysplasia, complex, with other brain malformations 6

Table 7: List of cancer related pathways discovered for each community. Each community linked to our miRNA affected genes was investigated for enriched pathways.

Pathway	Pathway size	# of genes in pathway	Genes	Community
KEGG CHEMOKINE SIGNALING PATHWAY	190	18	ADRBK2 CCL11 CCL13 CCL14 CCL16 CCL2 CCL3 CCL3L1 CCL5 CCL7 CCL8 CCR1 CCR2 CCR3 CCR5 CXCL10 CXCL11 CXCR3	3
KEGG ENDOCYTOSIS	183	2	ADRBK2 CCR5	3
KEGG PARKINSONS DISEASE	133	4	COX4I1 COX5A COX5B COX6A1	3
KEGG ALZHEIMERS DISEASE	169	4	COX4I1 COX5A COX5B COX6A1	3
KEGG HUNTINGTONS DISEASE	185	4	COX4I1 COX5A COX5B COX6A1	3
KEGG MELANOGENESIS	102	2	ASIP POMC	6
KEGG INSULIN SIGNALING PATHWAY	137	2	RPS6KB1 TSC1	7
KEGG COLORECTAL CANCER	62	2	APC RALGDS	7
KEGG VASCULAR SMOOTH MUSCLE CONTRACTION	115	2	CALCRL RAMP1	7
KEGG LEUKOCYTE TRANSENDOTHELIAL MIGRATION	118	2	MLLT4 RAPGEF3	7
KEGG ENDOCYTOSIS	183	2	ARRB2 GRK5	7
KEGG CHEMOKINE SIGNALING PATHWAY	190	2	ARRB2 GRK5	7
KEGG INSULIN SIGNALING PATHWAY	137	5	CALML3 PHKA1 PHKA2 PHKB PHKG1	7
KEGG ALZHEIMERS DISEASE	169	3	CACNA1D CALML3 RYR3	7
KEGG VASCULAR SMOOTH MUSCLE CONTRACTION	115	2	CACNA1D CALML3	7
KEGG LEISHMANIA INFECTION	72	3	CYBA NCF1 NCF2	7
KEGG LEUKOCYTE TRANSENDOTHELIAL MIGRATION	118	3	CYBA NCF1 NCF2	7
KEGG NATURAL KILLER CELL MEDIATED CYTOTOXICITY	137	3	NFATC1 NFATC2 NFATC4	7
KEGG ENDOMETRIAL CANCER	52	2	CTNNA3 TCF7	7
KEGG COLORECTAL CANCER	62	2	BCL2 TCF7	7
KEGG SMALL CELL LUNG CANCER	84	2	BCL2 NOS2	7
KEGG PROSTATE CANCER	89	2	BCL2 TCF7	7
KEGG ACUTE MYELOID LEUKEMIA	60	2	RUNX1T1 ZBTB16	11
KEGG BASAL CELL CARCINOMA	55	2	SMO WNT16	11
KEGG PANCREATIC CANCER	70	2	BRCA2 RAD51	11
KEGG SMALL CELL LUNG CANCER	84	2	LAMA4 TRAF4	11
KEGG HUNTINGTONS DISEASE	185	2	AP2A2 CREB1	11
KEGG MATURITY ONSET DIABETES OF THE YOUNG	25	3	HES1 MAFA NKX2-2	11
KEGG PRIMARY IMMUNODEFICIENCY	35	3	CD3D RAG1 RAG2	11
KEGG NOTCH SIGNALING PATHWAY	47	3	DLL3 DTX3 HES1	11
KEGG PARKINSONS DISEASE	133	4	COX4I2 NDUFB7 SNCAIP UCHL1	11
KEGG VASOPRESSIN REGULATED WATER REABSORPTION	44	2	CREB3L1 CREB3L4	11
KEGG HUNTINGTONS DISEASE	185	4	COX4I2 CREB3L1 CREB3L4 NDUFB7	11
KEGG TGF BETA SIGNALING PATHWAY	86	2	ID4 SMAD2	11
KEGG ALZHEIMERS DISEASE	169	3	ATF6 COX4I2 NDUFB7	11
KEGG PROSTATE CANCER	89	2	CREB3L1 CREB3L4	11
KEGG MELANOGENESIS	102	2	CREB3L1 CREB3L4	11
KEGG ENDOCYTOSIS	183	2	ACAP3 STAM	11
KEGG P53 SIGNALING PATHWAY	69	6	CHEK1 MDM2 MDM4 PPM1D TP53 TP73	11
KEGG PROSTATE CANCER	89	6	CREB5 CREBBP EP300 LEF1 MDM2 TP53	11
KEGG WNT SIGNALING PATHWAY	151	7	CREBBP EP300 JUN LEF1 SMAD3 SMAD4 TP53	11
KEGG COLORECTAL CANCER	62	5	JUN LEF1 SMAD3 SMAD4 TP53	11
KEGG RENAL CELL CARCINOMA	70	5	CREBBP EP300 ETS1 HIF1A JUN	11
KEGG UBIQUITIN MEDIATED PROTEOLYSIS	138	6	BRCA1 MDM2 PIAS1 PIAS3 PIAS4 UBE2I	11
KEGG TGF BETA SIGNALING PATHWAY	86	5	CREBBP EP300 SMAD3 SMAD4 SP1	11
KEGG NOTCH SIGNALING PATHWAY	47	4	CREBBP EP300 NOTCH1 NUMB	11
KEGG HUNTINGTONS DISEASE	185	6	CREB5 CREBBP EP300 SP1 TGM2 TP53	11
KEGG CHRONIC MYELOID LEUKEMIA	73	4	MDM2 SMAD3 SMAD4 TP53	11
KEGG SMALL CELL LUNG CANCER	84	4	PIAS1 PIAS3 PIAS4 TP53	11
KEGG PANCREATIC CANCER	70	3	SMAD3 SMAD4 TP53	11
KEGG THYROID CANCER	29	2	LEF1 TP53	11
KEGG MELANOGENESIS	102	3	CREBBP EP300 LEF1	11
KEGG BLADDER CANCER	42	2	MDM2 TP53	11
KEGG ENDOMETRIAL CANCER	52	2	LEF1 TP53	11
KEGG BASAL CELL CARCINOMA	55	2	LEF1 TP53	11
KEGG GLIOMA	65	2	MDM2 TP53	11
KEGG MELANOMA	71	2	MDM2 TP53	11
KEGG NOTCH SIGNALING PATHWAY	47	2	CTBP1 CTBP2	11
KEGG HEDGEHOG SIGNALING PATHWAY	56	2	GLI2 ZIC2	11
KEGG CHRONIC MYELOID LEUKEMIA	73	2	CTBP1 CTBP2	11
KEGG WNT SIGNALING PATHWAY	151	2	CTBP1 CTBP2	11
KEGG TGF BETA SIGNALING PATHWAY	86	17	ACVR1C BMP2 BMP4 BMP6 BMP7 BMPR2 GDF5 SMAD1 SMAD5 SMAD6 SMAD7 SMAD9 SMURF1 SMURF2 TGFBR2 ZFYVE16 ZFYVE9	11
KEGG BASAL CELL CARCINOMA	55	2	BMP2 BMP4	11
KEGG P53 SIGNALING PATHWAY	69	13	CCNB1 CCND1 CCND2 CCND3 CCNE1 CCNE2 CDK4 CDK6 CDKN1A CDKN2A GADD45B GADD45G TSC2	11
KEGG SMALL CELL LUNG CANCER	84	11	CCND1 CCNE1 CCNE2 CDK4 CDK6 CDKN1B E2F1 E2F2 E2F3 RB1 SKP2	11
KEGG CHRONIC MYELOID LEUKEMIA	73	10	CCND1 CDK4 CDK6 CDKN1A CDKN1B CDKN2A E2F1 E2F2 E2F3 RB1	11
KEGG GLIOMA	65	9	CCND1 CDK4 CDK6 CDKN1A CDKN2A E2F1 E2F2 E2F3 RB1	11
KEGG BLADDER CANCER	42	8	CCND1 CDK4 CDKN1A CDKN2A E2F1 E2F2 E2F3 RB1	11
KEGG MELANOMA	71	9	CCND1 CDK4 CDK6 CDKN1A CDKN2A E2F1 E2F2 E2F3 RB1	11
KEGG NON SMALL CELL LUNG CANCER	54	8	CCND1 CDK4 CDK6 CDKN2A E2F1 E2F2 E2F3 RB1	11
KEGG PROSTATE CANCER	89	9	CCND1 CCNE1 CCNE2 CDKN1A CDKN1B E2F1 E2F2 E2F3 RB1	11
KEGG PANCREATIC CANCER	70	8	CCND1 CDK4 CDK6 CDKN2A E2F1 E2F2 E2F3 RB1	11
KEGG TGF BETA SIGNALING PATHWAY	86	5	E2F4 E2F5 RBL1 RBL2 TFDP1	11
KEGG WNT SIGNALING PATHWAY	151	3	CCND1 CCND2 CCND3	11
KEGG ACUTE MYELOID LEUKEMIA	60	2	CCNA1 CCND1	11
KEGG ERBB SIGNALING PATHWAY	87	2	CDKN1A CDKN1B	11
KEGG NOTCH SIGNALING PATHWAY	47	4	CIR1 MAML3 NOTCH4 RBPJ	11
KEGG WNT SIGNALING PATHWAY	151	2	TBL1X TBL1XR1	11
KEGG LONG TERM DEPRESSION	70	2	PPP1R17 PRKG1	14
KEGG THYROID CANCER	29	1	TPR	14
KEGG TYPE I DIABETES MELLITUS	44	1	PTPRN	20
KEGG ENDOMETRIAL CANCER	52	1	APC2	20
KEGG BASAL CELL CARCINOMA	55	1	APC2	20
KEGG COLORECTAL CANCER	62	1	APC2	20

Table 8: Summary of common vertex counts in each community detected by Metis and the number of matching vertices present in spin-glass communities. For example, Metis community 1 and spin-glass community 1 had 18 vertices in common. Spin-glass community 1 had a total of 280 vertices and Metis community 1 had a total of 492 unique vertices.

	Metis 1	Metis 2	Metis 3	Metis 4	Metis 5	Metis 6	Metis 7	Metis 8	Metis 9	Metis 10	Metis 11	Metis 12	Metis 13	Metis 14	Metis 15	Metis 16	Metis 17	Metis 18	Metis 19	Metis 20	Metis 21	Metis 22	Metis 23	Metis 24	Metis unassigned	Total unique vertices
Spin-glass 1	18	11	12	17	8	6	11	26	11	17	9	9	13	5	11	6	8	6	11	11	12	14	15	2	11	280
Spin-glass 2	6	2	3	6	5	8	3	3	4	7	7	3	5	8	3	3	5	4	1	4	1	5	5	5	8	114
Spin-glass 3	12	9	25	9	13	26	19	14	12	15	31	21	15	18	22	12	17	9	21	12	12	17	19	17	10	407
Spin-glass 4	59	53	56	59	53	61	66	60	55	50	56	59	63	62	56	49	63	65	57	58	58	53	44	56	50	1421
Spin-glass 5	6	6	7	7	3	3	2	3	4	5	7	1	4	3	8	6	6	7	3	5	6	6	0	1	8	117
Spin-glass 6	16	24	18	23	17	23	37	25	25	22	19	20	15	18	17	23	16	16	21	11	22	17	25	26	20	516
Spin-glass 7	10	10	9	13	12	11	4	7	7	13	7	8	5	9	7	11	12	10	11	14	7	10	6	11	9	233
Spin-glass 8	3	9	7	4	5	6	8	0	4	9	4	2	6	11	4	7	6	6	6	4	3	7	8	14	7	150
Spin-glass 9	2	7	6	4	4	6	10	13	6	5	7	7	6	7	14	10	7	7	5	7	0	4	3	6	8	161
Spin-glass 10	8	3	1	10	4	4	3	5	4	6	14	3	2	7	10	5	5	4	4	4	3	4	4	5	2	124
Spin-glass 11	3	5	8	6	13	7	3	3	9	4	9	6	6	10	7	5	6	3	6	8	4	2	5	3	5	146
Spin-glass 12	29	19	24	24	14	29	22	17	21	20	29	17	20	23	18	20	33	18	20	25	24	18	27	21	34	566
Spin-glass 13	30	22	24	36	40	35	36	36	31	31	32	30	37	33	24	42	35	29	31	31	40	36	39	36	31	827
Spin-glass 14	33	35	34	13	30	22	27	34	36	27	25	38	34	19	21	31	29	34	24	26	29	35	26	24	29	715
Spin-glass 15	13	17	11	17	20	19	9	15	12	12	12	13	12	14	22	34	23	18	20	13	21	17	17	16	16	413
Spin-glass 16	14	15	10	14	5	9	7	5	6	17	10	14	8	9	17	9	7	23	15	14	11	5	13	10	9	276
Spin-glass 17	7	15	10	6	15	19	19	11	15	15	5	9	16	22	8	11	14	18	8	22	12	13	19	21	8	338
Spin-glass 18	24	39	26	24	27	23	39	23	40	21	24	23	30	35	20	27	32	33	23	29	23	25	21	27	29	687
Spin-glass 19	7	15	24	13	16	14	13	14	15	14	22	25	12	9	14	13	16	12	10	14	24	14	14	12	19	375
Spin-glass 20	30	25	26	24	24	17	10	35	19	16	17	20	17	23	25	22	11	28	28	21	25	35	21	23	30	572
Spin-glass 21	52	63	53	53	52	55	55	54	60	65	53	45	58	51	58	50	52	42	54	46	51	45	49	45	50	1311
Spin-glass 22	53	48	50	53	50	42	47	46	46	41	45	46	54	55	44	42	44	49	44	56	49	52	42	49	54	1201
Spin-glass 23	6	14	19	15	16	13	14	9	13	8	13	17	17	13	18	15	17	20	17	18	12	21	31	22	11	389
Spin-glass 24	22	23	26	20	17	17	13	17	17	26	13	32	27	11	16	17	15	19	17	24	14	20	22	19	15	479
Spin-glass 25	29	24	23	28	33	27	22	23	23	33	32	29	21	25	32	25	25	21	40	20	31	37	23	24	26	676
Total unique vertices	492	513	512	498	496	502	499	498	495	499	502	497	503	500	496	495	504	501	497	497	494	512	498	495	499	12494

Table 9: Total number of genes in our integrated network that overlap with ovarian cancer and other cancer subtypes.

Table of Census Genes in Cancer

	MiRNA-gene Network	InWeb Network	RegNetwork	Integrated Network	GeNets
No. of Genes	776	895	637	2354	494
No. of Cancer Genes	7	111	133	251	56
No. of Ovarian Cancer Genes	1	5	5	11	3

Table 10: List of genes that are enriched in cancer related pathways from the KEGG pathway database. These genes are also found in our integrated network.

Gene Symbol	Gene Name	KEGG Pathways
AKT1	AKT serine/threonine kinase 1(AKT1)	MAPK signaling pathway
AKT2	AKT serine/threonine kinase 2(AKT2)	MAPK signaling pathway
ATF2	activating transcription factor 2(ATF2)	MAPK signaling pathway
ATF4	activating transcription factor 4(ATF4)	MAPK signaling pathway
BCR	BCR, RhoGEF and GTPase activating protein(BCR)	Pathways in cancer, Chronic myeloid leukemia
BMP2K	BMP2 inducible kinase(BMP2K)	Transcriptional misregulation in cancer
BRAF	B-Raf proto-oncogene, serine/threonine kinase(BRAF)	MAPK signaling pathway
CASP3	caspase 3(CASP3)	MAPK signaling pathway
CCNT1	cyclin T1(CCNT1)	Transcriptional misregulation in cancer
CDCA5	cell division cycle associated 5(CDCA5)	MicroRNAs in cancer
CDK9	cyclin dependent kinase 9(CDK9)	Transcriptional misregulation in cancer
CHUK	conserved helix-loop-helix ubiquitous kinase(CHUK)	MAPK signaling pathway
CRK	CRK proto-oncogene, adaptor protein(CRK)	MAPK signaling pathway
DAPK3	death associated protein kinase 3(DAPK3)	Pathways in cancer, Bladder cancer
EGFR	epidermal growth factor receptor(EGFR)	MAPK signaling pathway
ELK4	ELK4, ETS transcription factor(ELK4)	MAPK signaling pathway
FGFR2	fibroblast growth factor receptor 2(FGFR2)	MAPK signaling pathway
FLNA	filamin A(FLNA)	MAPK signaling pathway
FOS	Fos proto-oncogene, AP-1 transcription factor subunit(FOS)	MAPK signaling pathway
GOLPH3	golgi phosphoprotein 3(GOLPH3)	Transcriptional misregulation in cancer
GPC1	glypican 1(GPC1)	Proteoglycans in cancer
GRB2	growth factor receptor bound protein 2(GRB2)	MAPK signaling pathway
HOXD10	homeobox D10(HOXD10)	Proteoglycans in cancer, MicroRNAs in cancer
HPGD	hydroxyprostaglandin dehydrogenase 15-(NAD)(HPGD)	Transcriptional misregulation in cancer
HSPB1	heat shock protein family B (small) member 1(HSPB1)	MAPK signaling pathway
IKBKB	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta(IKBKB)	MAPK signaling pathway
IKBKG	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma(IKBKG)	MAPK signaling pathway
IL1A	interleukin 1 alpha(IL1A)	MAPK signaling pathway
IL1B	interleukin 1 beta(IL1B)	MAPK signaling pathway
JUN	Jun proto-oncogene, AP-1 transcription factor subunit(JUN)	MAPK signaling pathway
KIF23	kinesin family member 23(KIF23)	MicroRNAs in cancer
KLF3	Kruppel like factor 3(KLF3)	Transcriptional misregulation in cancer
KRAS	KRAS proto-oncogene, GTPase(KRAS)	MAPK signaling pathway
LMO2	LIM domain only 2(LMO2)	Transcriptional misregulation in cancer
MAP2K2	mitogen-activated protein kinase kinase 2(MAP2K2)	MAPK signaling pathway
MAPK1	mitogen-activated protein kinase 1(MAPK1)	MAPK signaling pathway
MAPKAPK2	mitogen-activated protein kinase-activated protein kinase 2(MAPKAPK2)	MAPK signaling pathway
MAX	MYC associated factor X(MAX)	MAPK signaling pathway
MEF2C	myocyte enhancer factor 2C(MEF2C)	MAPK signaling pathway
MKNK1	MAP kinase interacting serine/threonine kinase 1(MKNK1)	MAPK signaling pathway
MKNK2	MAP kinase interacting serine/threonine kinase 2(MKNK2)	MAPK signaling pathway
MLLT1	MLLT1, super elongation complex subunit(MLLT1)	Transcriptional misregulation in cancer

MYC	v-myc avian myelocytomatosis viral oncogene homolog(MYC)	MAPK signaling pathway
NRAS	neuroblastoma RAS viral oncogene homolog(NRAS)	MAPK signaling pathway
PAK1	p21 (RAC1) activated kinase 1(PAK1)	MAPK signaling pathway
PPM1A	protein phosphatase, Mg2+/Mn2+ dependent 1A(PPM1A)	MAPK signaling pathway
PPP3CA	protein phosphatase 3 catalytic subunit alpha(PPP3CA)	MAPK signaling pathway
PPP5C	protein phosphatase 5 catalytic subunit(PPP5C)	MAPK signaling pathway
PRKACA	protein kinase cAMP-activated catalytic subunit alpha(PRKACA)	MAPK signaling pathway
PTPN5	protein tyrosine phosphatase, non-receptor type 5(PTPN5)	MAPK signaling pathway
PTPN7	protein tyrosine phosphatase, non-receptor type 7(PTPN7)	MAPK signaling pathway
PTPRR	protein tyrosine phosphatase, receptor type R(PTPRR)	MAPK signaling pathway
RAC1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)(RAC1)	MAPK signaling pathway
RAF1	Raf-1 proto-oncogene, serine/threonine kinase(RAF1)	MAPK signaling pathway
RAP1A	RAP1A, member of RAS oncogene family(RAP1A)	MAPK signaling pathway
RAP1B	RAP1B, member of RAS oncogene family(RAP1B)	MAPK signaling pathway
RECK	reversion inducing cysteine rich protein with kazal motifs(RECK)	MicroRNAs in cancer
RPS6KA1	ribosomal protein S6 kinase A1(RPS6KA1)	MAPK signaling pathway
RUNX1	runt related transcription factor 1(RUNX1)	Pathways in cancer, Transcriptional misregulation in cancer, Chronic myeloid leukemia, Acute myeloid leukemia
ST14	suppression of tumorigenicity 14(ST14)	MicroRNAs in cancer
STK3	serine/threonine kinase 3(STK3)	MAPK signaling pathway
STMN1	stathmin 1(STMN1)	MAPK signaling pathway
TAB2	TGF-beta activated kinase 1/MAP3K7 binding protein 2(TAB2)	MAPK signaling pathway
TAOK1	TAO kinase 1(TAOK1)	MAPK signaling pathway
TAOK2	TAO kinase 2(TAOK2)	MAPK signaling pathway
TFE3	transcription factor binding to IGHM enhancer 3(TFE3)	Transcriptional misregulation in cancer
TFG	TRK-fused gene(TFG)	Pathways in cancer, Thyroid cancer
TGFB1	transforming growth factor beta 1(TGFB1)	MAPK signaling pathway
TGFB2	transforming growth factor beta 2(TGFB2)	MAPK signaling pathway
TGFBR2	transforming growth factor beta receptor 2(TGFBR2)	MAPK signaling pathway
TLX1	T-cell leukemia homeobox 1(TLX1)	Transcriptional misregulation in cancer
TLX3	T-cell leukemia homeobox 3(TLX3)	Transcriptional misregulation in cancer
TP63	tumor protein p63(TP63)	MicroRNAs in cancer
TRAF6	TNF receptor associated factor 6(TRAF6)	MAPK signaling pathway
TWIST1	twist family bHLH transcription factor 1(TWIST1)	Proteoglycans in cancer
ZEB1	zinc finger E-box binding homeobox 1(ZEB1)	MicroRNAs in cancer
ZEB2	zinc finger E-box binding homeobox 2(ZEB2)	MicroRNAs in cancer
ZFPM2	zinc finger protein, FOG family member 2(ZFPM2)	MicroRNAs in cancer

Table 11: List of genes that are enriched in cancer related pathways from the KEGG pathway database. These genes are also found in our LPA integrated network.

Pathway	# of genes in pathway	Genes	Community
KEGG BLADDER CANCER	2	E2F1 E2F3	7
KEGG NON SMALL CELL LUNG CANCER	2	E2F1 E2F3	7
KEGG GLIOMA	2	E2F1 E2F3	7
KEGG PANCREATIC CANCER	2	E2F1 E2F3	7
KEGG MELANOMA	2	E2F1 E2F3	7
KEGG CHRONIC MYELOID LEUKEMIA	2	E2F1 E2F3	7
KEGG SMALL CELL LUNG CANCER	2	E2F1 E2F3	7
KEGG PROSTATE CANCER	2	E2F1 E2F3	7
KEGG SMALL CELL LUNG CANCER	2	PIAS2 PIAS3	7
KEGG P53 SIGNALING PATHWAY	3	CCND1 CCND3 TP73	7
KEGG ACUTE MYELOID LEUKEMIA	2	CCND1 FLT3	7
KEGG WNT SIGNALING PATHWAY	2	CCND1 CCND3	7
KEGG ENDOMETRIAL CANCER	2	CTNNB1 GSK3B	11
KEGG BASAL CELL CARCINOMA	2	CTNNB1 GSK3B	11
KEGG COLORECTAL CANCER	2	CTNNB1 GSK3B	11
KEGG PROSTATE CANCER	2	CTNNB1 GSK3B	11
KEGG MELANOGENESIS	2	CTNNB1 GSK3B	11
KEGG WNT SIGNALING PATHWAY	2	CTNNB1 GSK3B	11
KEGG TGF BETA SIGNALING PATHWAY	3	BMP2 BMP4 TGFB1	12
KEGG BASAL CELL CARCINOMA	2	BMP2 BMP4	12
KEGG GLIOMA	2	PTEN SHC1	12
KEGG CHRONIC MYELOID LEUKEMIA	2	SHC1 TGFB1	12
KEGG P53 SIGNALING PATHWAY	3	ATM CHEK1 TP53	13
KEGG AMYOTROPHIC LATERAL SCLEROSIS ALS	2	SOD1 TP53	13
KEGG APOPTOSIS	2	ATM TP53	13
KEGG HUNTINGTONS DISEASE	2	SOD1 TP53	13
KEGG P53 SIGNALING PATHWAY	2	CHEK2 MDM2	13
KEGG COLORECTAL CANCER	2	BCL2 KRAS	16
KEGG PROSTATE CANCER	2	BCL2 KRAS	16
KEGG PROSTATE CANCER	4	AKT1 AKT2 AKT3 FOXO1	17
KEGG INSULIN SIGNALING PATHWAY	4	AKT1 AKT2 AKT3 FOXO1	17
KEGG ENDOMETRIAL CANCER	3	AKT1 AKT2 AKT3	17
KEGG NON SMALL CELL LUNG CANCER	3	AKT1 AKT2 AKT3	17
KEGG ACUTE MYELOID LEUKEMIA	3	AKT1 AKT2 AKT3	17
KEGG COLORECTAL CANCER	3	AKT1 AKT2 AKT3	17
KEGG GLIOMA	3	AKT1 AKT2 AKT3	17
KEGG PANCREATIC CANCER	3	AKT1 AKT2 AKT3	17
KEGG RENAL CELL CARCINOMA	3	AKT1 AKT2 AKT3	17
KEGG MELANOMA	3	AKT1 AKT2 AKT3	17
KEGG SMALL CELL LUNG CANCER	3	AKT1 AKT2 AKT3	17
KEGG APOPTOSIS	3	AKT1 AKT2 AKT3	17
KEGG JAK STAT SIGNALING PATHWAY	3	AKT1 AKT2 AKT3	17
KEGG PANCREATIC CANCER	3	RAC2 TGFBR2 VEGFA	23
KEGG COLORECTAL CANCER	2	RAC2 TGFBR2	23
KEGG P53 SIGNALING PATHWAY	2	SESN3 THBS1	25
KEGG TGF BETA SIGNALING PATHWAY	2	BMPR2 THBS1	25

Table 12: Table of miRNA-gene associations for stage II samples.

MiRNA	Gene	FDR
hsa-mir-429	CDK11B	0.00019764
hsa-mir-429	PRKCZ	0.000244602
hsa-mir-147b	MESP2	0.002152882
hsa-mir-429	MMEL1	0.012186697
hsa-mir-940	CREBBP	0.014576292
hsa-mir-429	MIB2	0.01953022
hsa-mir-940	ZNF500	0.022218129
hsa-mir-429	NADK	0.024168807
hsa-mir-429	RER1	0.024168807
hsa-mir-429	CDK11A	0.024168807
hsa-mir-147b	WDR93	0.031533831
hsa-mir-1291	CDK17	0.037043283
hsa-mir-429	PANK4	0.037968774
hsa-mir-3687	HSPA13	0.038841051

Table 13: Table of miRNA-gene associations for stage III samples.

MiRNA	**Gene**	**FDR**
hsa-mir-940	FAM86A	0.022834838
hsa-mir-940	TFAP4	0.031093644
hsa-mir-429	PRKCZ	0.031093644
hsa-mir-940	ALG1	0.031093644
hsa-mir-940	HMOX2	0.031093644
hsa-mir-940	DNAJA3	0.039209633
hsa-mir-429	SSU72	0.039209633
hsa-mir-484	C16orf93	0.039209633
hsa-mir-429	NADK	0.039497555
hsa-mir-484	PHKG2	0.039497555
hsa-mir-484	PRSS36	0.044826117
hsa-mir-429	MRPL20	0.046093757

Table 14: List of enriched pathways by community directly involved in various cancers. In stage II, there are four distinct communities enriched in cancer related pathways.

Pathway	Pathway size	# of genes in pathway	Gene list	Community
KEGG BLADDER CANCER	42	3	MYC RAF1 RASSF1	1
KEGG COLORECTAL CANCER	62	3	MYC RAF1 RALGDS	1
KEGG THYROID CANCER	29	2	MYC TPM3	1
KEGG ENDOMETRIAL CANCER	52	2	MYC RAF1	1
KEGG NON SMALL CELL LUNG CANCER	54	2	RAF1 RASSF1	1
KEGG ACUTE MYELOID LEUKEMIA	60	2	MYC RAF1	1
KEGG PANCREATIC CANCER	70	2	RAF1 RALGDS	1
KEGG CHRONIC MYELOID LEUKEMIA	73	2	MYC RAF1	1
KEGG ERBB SIGNALING PATHWAY	87	2	MYC RAF1	1
KEGG GAP JUNCTION	90	2	MAPK7 RAF1	1
KEGG GNRH SIGNALING PATHWAY	101	2	MAPK7 RAF1	1
KEGG NEUROTROPHIN SIGNALING PATHWAY	126	2	MAPK7 RAF1	1
KEGG SMALL CELL LUNG CANCER	84	2	COL4A6 FN1	2
KEGG HYPERTROPHIC CARDIOMYOPATHY HCM	85	2	ITGA1 ITGA11	2
KEGG DILATED CARDIOMYOPATHY	92	2	ITGA1 ITGA11	2
KEGG CELL ADHESION MOLECULES CAMS	134	2	ICAM2 ITGAL	2
KEGG NATURAL KILLER CELL MEDIATED CYTOTOXICITY	137	2	ICAM2 ITGAL	2
KEGG RNA DEGRADATION	59	2	ENO3 RQCD1	3
KEGG PARKINSONS DISEASE	133	2	NDUFA4 NDUFA5	4
KEGG OXIDATIVE PHOSPHORYLATION	135	2	NDUFA4 NDUFA5	4
KEGG ALZHEIMERS DISEASE	169	2	NDUFA4 NDUFA5	4
KEGG HUNTINGTONS DISEASE	185	2	NDUFA4 NDUFA5	4
KEGG INSULIN SIGNALING PATHWAY	137	2	PRKCZ RPS6KB1	6
KEGG WNT SIGNALING PATHWAY	151	7	CCND1 CREBBP JUN SENP2 SMAD2 SMAD3 TP53	7
KEGG COLORECTAL CANCER	62	5	CCND1 JUN SMAD2 SMAD3 TP53	7
KEGG PANCREATIC CANCER	70	5	CCND1 RELA SMAD2 SMAD3 TP53	7
KEGG CELL CYCLE	128	5	CCND1 CREBBP SMAD2 SMAD3 TP53	7
KEGG CHRONIC MYELOID LEUKEMIA	73	4	CCND1 RELA SMAD3 TP53	7
KEGG PROSTATE CANCER	89	4	CCND1 CREBBP RELA TP53	7
KEGG ACUTE MYELOID LEUKEMIA	60	3	CCND1 PML RELA	7
KEGG P53 SIGNALING PATHWAY	69	3	CCND1 CCNG2 TP53	7
KEGG RENAL CELL CARCINOMA	70	3	CREBBP ETS1 JUN	7
KEGG ADHERENS JUNCTION	75	3	CREBBP SMAD2 SMAD3	7
KEGG SMALL CELL LUNG CANCER	84	3	CCND1 RELA TP53	7
KEGG TGF BETA SIGNALING PATHWAY	86	3	CREBBP SMAD2 SMAD3	7
KEGG THYROID CANCER	29	2	CCND1 TP53	7
KEGG NEUROTROPHIN SIGNALING PATHWAY	126	3	JUN RELA TP53	7
KEGG JAK STAT SIGNALING PATHWAY	155	3	CCND1 CREBBP STAT4	7
KEGG BLADDER CANCER	42	2	CCND1 TP53	7
KEGG ENDOMETRIAL CANCER	52	2	CCND1 TP53	7
KEGG NON SMALL CELL LUNG CANCER	54	2	CCND1 TP53	7
KEGG GLIOMA	65	2	CCND1 TP53	7
KEGG EPITHELIAL CELL SIGNALING IN HELICOBACTER PYLORI INFECTION	68	2	JUN RELA	7
KEGG MELANOMA	71	2	CCND1 TP53	7
KEGG APOPTOSIS	88	2	RELA TP53	7
KEGG TOLL LIKE RECEPTOR SIGNALING PATHWAY	102	2	JUN RELA	7
KEGG T CELL RECEPTOR SIGNALING PATHWAY	108	2	JUN RELA	7
KEGG HUNTINGTONS DISEASE	185	2	CREBBP TP53	7
KEGG FOCAL ADHESION	201	2	CCND1 JUN	7
KEGG WNT SIGNALING PATHWAY	151	4	SFRP1 SFRP4 SFRP5 WNT16	9
KEGG CYTOSOLIC DNA SENSING PATHWAY	56	2	NFKBIA TBK1	10
KEGG RIG I LIKE RECEPTOR SIGNALING PATHWAY	71	2	NFKBIA TBK1	10
KEGG TOLL LIKE RECEPTOR SIGNALING PATHWAY	102	2	NFKBIA TBK1	10

Table 15: List of enriched pathways by community directly involved in various cancers. In stage III, there are 5 distinct communities enriched in cancer related pathways.

Pathway	Pathway size	# of genes in pathway	Gene list	Community
KEGG OOCYTE MEIOSIS	114	2	ANAPC7 RPS6KA3	8
KEGG ACUTE MYELOID LEUKEMIA	60	1	KIT	8
KEGG LONG TERM POTENTIATION	70	1	RPS6KA3	8
KEGG RENAL CELL CARCINOMA	70	1	CRKL	8
KEGG CHRONIC MYELOID LEUKEMIA	73	1	CRKL	8
KEGG ERBB SIGNALING PATHWAY	87	1	CRKL	8
KEGG MELANOGENESIS	102	1	KIT	8
KEGG CELL CYCLE	128	1	ANAPC7	8
KEGG ENDOCYTOSIS	183	1	KIT	8
KEGG HYPERTROPHIC CARDIOMYOPATHY HCM	85	4	ITGA1 ITGA2B ITGA3 ITGA5	6
KEGG SMALL CELL LUNG CANCER	84	3	FN1 ITGA2B ITGA3	6
KEGG PRION DISEASES	35	1	C1QA	6
KEGG SYSTEMIC LUPUS ERYTHEMATOSUS	140	1	C1QA	6
KEGG HUNTINGTONS DISEASE	185	1	TGM2	6
KEGG COLORECTAL CANCER	62	2	CCND1 JUN	4
KEGG CELL CYCLE	128	2	CCND1 PRKDC	4
KEGG WNT SIGNALING PATHWAY	151	2	CCND1 JUN	4
KEGG THYROID CANCER	29	1	CCND1	4
KEGG BLADDER CANCER	42	1	CCND1	4
KEGG ENDOMETRIAL CANCER	52	1	CCND1	4
KEGG NON SMALL CELL LUNG CANCER	54	1	CCND1	4
KEGG ACUTE MYELOID LEUKEMIA	60	1	CCND1	4
KEGG GLIOMA	65	1	CCND1	4
KEGG EPITHELIAL CELL SIGNALING IN HELICOBACTER PYLORI INFECTION	68	1	JUN	4
KEGG P53 SIGNALING PATHWAY	69	1	CCND1	4
KEGG PANCREATIC CANCER	70	1	CCND1	4
KEGG RENAL CELL CARCINOMA	70	1	JUN	4
KEGG MELANOMA	71	1	CCND1	4
KEGG LEISHMANIA INFECTION	72	1	JUN	4
KEGG CHRONIC MYELOID LEUKEMIA	73	1	CCND1	4
KEGG VIRAL MYOCARDITIS	73	1	CCND1	4
KEGG SMALL CELL LUNG CANCER	84	1	CCND1	4
KEGG ERBB SIGNALING PATHWAY	87	1	JUN	4
KEGG PROSTATE CANCER	89	1	CCND1	4
KEGG T CELL RECEPTOR SIGNALING PATHWAY	108	1	JUN	4
KEGG ENDOCYTOSIS	183	3	ADRBK1 GIT2 PRKCZ	2
KEGG ALZHEIMERS DISEASE	169	2	APOE MAPT	2
KEGG TYPE II DIABETES MELLITUS	47	1	PRKCZ	2
KEGG ACUTE MYELOID LEUKEMIA	60	1	RPS6KB1	2
KEGG ADIPOCYTOKINE SIGNALING PATHWAY	67	1	PRKCQ	2
KEGG TGF BETA SIGNALING PATHWAY	86	1	RPS6KB1	2
KEGG ERBB SIGNALING PATHWAY	87	1	RPS6KB1	2
KEGG T CELL RECEPTOR SIGNALING PATHWAY	108	1	PRKCQ	2
KEGG LEUKOCYTE TRANSENDOTHELIAL MIGRATION	118	1	MSN	2
KEGG CELL ADHESION MOLECULES CAMS	134	1	SELPLG	2
KEGG ADHERENS JUNCTION	75	3	ACTB SORBS1 WASF3	1
KEGG PATHOGENIC ESCHERICHIA COLI INFECTION	59	2	ACTB YWHAZ	1
KEGG OOCYTE MEIOSIS	114	2	YWHAE YWHAZ	1
KEGG CELL CYCLE	128	2	YWHAE YWHAZ	1
KEGG MATURITY ONSET DIABETES OF THE YOUNG	25	1	NR5A2	1
KEGG VIBRIO CHOLERAE INFECTION	56	1	ACTB	1
KEGG VIRAL MYOCARDITIS	73	1	ACTB	1
KEGG HYPERTROPHIC CARDIOMYOPATHY HCM	85	1	ACTB	1
KEGG LEUKOCYTE TRANSENDOTHELIAL MIGRATION	118	1	ACTB	1
KEGG WNT SIGNALING PATHWAY	151	1	RUVBL1	1

Table 16: List of similar nodes found in stage II and stage III network alignment.

Node	Node	Weight
ADRBK1	GIT2	1
AFF3	RQCD1	1
RQCD1	WIPF1	1
ANK3	CRKL	1
CRKL	ARID4B	1
CRKL	SLIT2	1
CRKL	WIPF1	1
CRKL	MAP4K5	1
ANXA6	CFL2	1
ANXA6	GMFB	1
ANXA6	MPRIP	1
ANXA6	TMOD3	1
ANXA6	UACA	1
CFL2	GMFB	1
CFL2	MPRIP	1
CFL2	TMOD3	1
CFL2	UACA	1
GMFB	MPRIP	1
GMFB	TMOD3	1
GMFB	UACA	1
MPRIP	TMOD3	1
MPRIP	UACA	1
MPRIP	CKAP4	1
MPRIP	LRP1B	1
MPRIP	RANBP9	1
MPRIP	SASH1	1
MPRIP	TFAP2A	1
TMOD3	UACA	1
TMOD3	EPS8	1
ARID4B	MCFD2	1
ARID4B	MSN	1
ARID4B	PICALM	1
ARID4B	PRKCZ	1
ARID4B	SLIT2	1
PRKCZ	FEZ1	1
PRKCZ	FEZ2	1
PRKCZ	RPS6KB1	1
SLIT2	GPC1	1
SLIT2	GTF3C4	1
ENO3	SEC23A	1
SEC23A	WIPF1	1

ATF3	JUN	1
JUN	CCND1	1
JUN	NR3C1	1
JUN	FOSL2	1
ATXN1	RBFOX1	1
BCL11A	KIF13B	1
BCL11A	PPAP2B	1
BCL11A	UTP6	1
KIF13B	PPAP2B	1
KIF13B	UTP6	1
PPAP2B	UTP6	1
RASSF8	CLASP1	1
RASSF8	NADK	1
RASSF8	RAB11FIP2	1
RASSF8	GOLGA1	1
CFD	FN1	1
FN1	COL2A1	1
FN1	PKD1	1
FN1	TGFBI	1
FN1	COL6A2	1
FN1	TCEB1	1
FN1	IGLL1	1
CKAP4	LRP1B	1
CKAP4	PLAT	1
CKAP4	RANBP9	1
LRP1B	PLAT	1
LRP1B	RANBP9	1
RANBP9	NR3C1	1
CLASP1	NADK	1
CLASP1	RAB11FIP2	1
NADK	RAB11FIP2	1
COL2A1	PKD1	1
COL2A1	TGFBI	1
TGFBI	RAP2C	1
CPA5	RECK	1
NR3C1	TCEB1	1
MAP4K5	CSNK1G3	1
MAP4K5	TBK1	1
CSNK1G3	TBK1	1
ELL2	PHACTR2	1
ERRFI1	SASH1	1
ISL1	RABIF	1

MAPK7	HIPK3	1
MAP2	MARCKS	1
MAP2	NIN	1
FEM1C	PPM1F	1
FEM1C	TCEB1	1
PPM1F	TCEB1	1
FEZ1	FEZ2	1
GATA4	ZFPM2	1
GTF3C4	SIP1	1
MARCKS	NIN	1
MARCKS	TOB1	1
MDK	PTPRZ1	1
PTPRZ1	TNR	1
PDS5B	WAPAL	1

Table 17 : List of common pathways between stage II and stage III network interactions. There were 97 interactions in common between stage II and stage III. Several pathways are involved in cancer and important signaling pathways that aid in cancer progression.

Gene	KEGG PATHWAY
CRKL	MAPK signaling pathway
GATA4	cGMP-PKG signaling pathway
ISL1	Signaling pathways regulating pluripotency of stem cells
JUN	MAPK signaling pathway
NADK	Nicotinate and nicotinamide metabolism
RAB11FIP2	Endocytosis
TBK1	Ras signaling pathway
WIPF1	Endocytosis
ATF3	HTLV-I infection
ANK3	Proteoglycans in cancer
CSNK1G3	Hedgehog signaling pathway
COL2A1	PI3K-Akt signaling pathway
COL6A2	PI3K-Akt signaling pathway
CCND1	FoxO signaling pathway
ENO3	Glycolysis / Gluconeogenesis
FN1	PI3K-Akt signaling pathway
GPC1	Proteoglycans in cancer
MAPK7	MAPK signaling pathway
MSN	Leukocyte transendothelial migration
MARCKS	MicroRNAs in cancer
PLAT	Transcriptional misregulation in cancer,
PRKCZ	Rap1 signaling pathway
PTPRZ1	Epithelial cell signaling in Helicobacter pylori infection
RECK	MicroRNAs in cancer
RPS6KB1	ErbB signaling pathway
TNR	PI3K-Akt signaling pathway
TCEB1	Pathways in cancer - Renal cell carcinoma
ZFPM2	MicroRNAs in cancer

PUBLICATIONS

[1] **Hall B**, Quitadamo A, and Shi X. “"A Label Propagation Approach to Constructing an Integrative MicroRNA and Gene Expression Network in Ovarian Cancer"“, in submission.

[2] **Hall B**, Quitadamo A, and Shi X. “Constructing an Integrative MicroRNA eQTL Network on Ovarian Cancer: A Label Propagation Approach Utilizing Multiple Networks“, in Proceedings of the 13th International Symposium on Bioinformatics Research and Applications (ISBRA2017, Track 2 paper), Honolulu, Hawaii, May 29 - June 1, 2017.

[3] **Hall B**, Quitadamo A, and Shi X. “Identifying microRNA and Gene Expression Networks Using Graph Communities", Tsinghua Science and Technology. 2016. 21(2): 176-95.

OTHER PUBLICATIONS

[4] Wen J, Quitadamo A, **Hall B**, and Shi X “An Empirical Bayesian Elastic Nets Method for Epistasis Analysis of microRNAs on Pathological Stages in Colon Cancer“, in press, BMC Genomics, 2017

[5] Wang Z, **Benika H**, Xu J, and Shi X. “A Sparse Learning Framework for Joint Effect Analysis of Copy Number Variants." IEEE/ACM Transactions on Computational Biology and Bioinformatics (Volume: PP, Issue: 99) 2015

[6] Quitadamo A, Tian L, **Benika H**, and Shi X. “An Integrated Network of microRNA and Gene Expression in Ovarian Cancer" *BMC Bioinformatics, 16, 2015.*

[7] Wen J, Quitadamo A, **Hall B**, and Shi X “"An Empirical Bayesian Elastic Nets Method for Epistasis Analysis of microRNAs on Pathological Stages in Colon Cancer" Conference

[8] Wen J, Quitadamo A, **Hall B**, and Shi X. “"Epistasis Analysis of microRNAs in Colon Cancer Using Empirical Bayesian Elastic Nets." In Proceedings of

ISBRA2016, Minsky, Belaruse, June 6-8, 2016.

[9] He A, **Hall B**, Wen J, Liang Y, and Shi X, “"Sequential and Parallel LASSO: Novel Scalable Methods for eQTL Mapping." In Proceedings of the 8th International Conference on Bioinformatics and Computational Biology (BICoB), Pages 533-534, April 4-6, 2016 Las Vegas, Nevada, USA

[10] Aniket, Robert Reid, **Benika Hall**, Ian Marriott and Ahmed El-Ghannam. “Early osteoblast responses to orthopedic implants: Synergy of surface roughness and chemistry of bioactive ceramic coating" *Journal of Biomedical Materials Research.* 2014.

CONTRIBUTIONS OF THE AUTHOR

Contributions to the related papers in this dissertation are:

Publication 1: In this publication, I contributed to establishing the methodology with my advisor Dr. Xinghua Shi, building a prototype network using the label propagation model. I also contributed to designing the experiment and extending the label propagation model. I contributed to filtering the data and analyzing the data. More contributions include evaluating the model, performing the downstream analysis, visualizing the results and interpreting the results. I also contributed to writing all sections in the paper and preparing the corresponding Track 2 paper for Publication 2.

Publication 3: In this publication, the methodology was developed by my advisor

Dr. Xinghua Shi and we designed the project based on her ideas. I contributed to designing the framework and analyzing the data. More contributions include simulating the experiments and evaluating the model. I also contributed to performing the downstream analysis and interpreting the results. Additional contributions include visualizing the results. I also contributed to writing all sections in the paper.