Functional Modularity Methods and Applications for Human Diseases

Rahiminejad, Sara

2023

UNIVERSITY OF CALIFORNIA SAN DIEGO

Functional Modularity Methods and Applications for Human Diseases

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Engineering Sciences (Mechanical Engineering)

by

Sara Rahiminejad

Committee in charge:

Professor Shankar Subramaniam, Chair
Professor Ratneshwar Lal, Co-Chair
Professor Marcos Intaglietta
Professor Padmini Rangamani
Professor Sutanu Sarkar

2023

The Dissertation of Sara Rahiminejad is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my beloved parents and my sister,

for all their love, encouragement, and dedicated support throughout my life.

&

To my beloved husband,

for all his love and support that gave me the power to continue my journey,

with all its ups and downs.

TABLE OF CONTENTS

LIST OF SUPPLEMENTAL FILES

Rahiminejad_SupplementaryTables.zip

Rahiminejad_Tables_S2.xlsx

Rahiminejad_Tables_S3.xlsx

Rahiminejad_Tables_S4.xlsx

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ARI	Adjusted Rand Index
CP	Cellular Processes
CRC	Colorectal Cancer
DAVID	Database for Annotation, Visualization, and Integrated Discovery
DEG	Differentially Expressed Gene
DFS	Disease-Free Survival
EO	Extremal Optimization
ERW	Edge Random Walk
FDA	Food and Drug Administration
FDR	False Discovery Rate
FE	Fold Enrichment
GEO	Gene Expression Omnibus
GEPIA	Gene Expression Profiling Interactive Analysis
GO	Gene Ontology
GIP	Genetic Information Processing
HD	Human Diseases
JI	Jaccard Index
KEGG	Kyoto Encyclopedia of Genes and Genomes
LFR	Lancichinetti, Fortunato, Radicchi
M	Metabolism
mCRC	Metastatic Colorectal Cancer
MI	Mutual Information

MMRd	Mismatch Repair-deficient
MMRp	Mismatch Repair-proficient
NCI	National Cancer Institute
NMI	Normalized Mutual Information
OS	Organismal System
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PPI	Protein-Protein Interaction
RI	Rand Index
RMA	Robust Multi-array Average
scRNA-seq	Single-cell RNA Sequencing
STEM	Short Time-series Expression Miner
TCGA	The Cancer Genome Atlas
t-SNE	t-Distributed Stochastic Embedding
UMAP	Uniform Manifold Approximation and Projection
VEGFR	Vascular Endothelial Growth Factor Receptor
WGCNA	Weighted Gene Co-expression Network Analysis

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Professor Shankar Subramaniam for his support, guidance, and mentorship as the chair of my committee. Through many thought-provoking research discussions and multiple iterations of publication drafts, his guidance and feedback have been invaluable. I would also like to thank my doctoral committee members: Professor Ratneshwar Lal, Professor Marcos Intaglietta, Professor Padmini Rangamani, and Professor Sutanu Sarkar for their support and guidance.

I would like to extend my gratitude to Dr. Mano R. Maurya and Dr. Kavitha Mukund for their mentorship and valuable feedbacks on my publications, and all the past and present lab members who made the lab an enjoyable place during my Ph.D.

Finally, I would like to thank my family and friends, for their love and support during my time in school. I want to specially thank my mother, Fatemeh Rahiminejad, and my father, Hossein Rahiminejad, for simply everything. They instilled the value of education in me when I was a child and without their hard work and inspiration it would not have been possible for me to come this far. I would also like to thank my younger sister, Nazli Rahiminejad, for her constant love and encouragement. She filled my place for my parents through all these years that I was far from home. And last but not least, I want to thank my best friend, soulmate, and husband, Dr. Milad Mortazavi, from the bottom of my heart, for his unconditional love, support, and patience.

Chapter 2, in full, is a reprint of the material as it appears in “Topological and functional comparison of community detection algorithms in biological networks” by Rahiminejad, Sara; Maurya, Mano R.; and Subramaniam, Shankar., BMC Bioinformatics, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in “Modular and mechanistic changes across stages of colorectal cancer” by Rahiminejad, Sara; Maurya, Mano R.; Mukund, Kavitha; and Subramaniam, Shankar., BMC Cancer, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Rahiminejad, Sara; Mukund, Kavitha; Maurya, Mano R.; and Subramaniam, Shankar. The dissertation author was the primary researcher and author of this material.

VITA

2010 Bachelor of Science in Chemical Engineering, Sharif University of Technology, Iran

2016 Master of Science in Mechanical Engineering, University of California Irvine

2023 Doctor of Philosophy in Engineering Sciences (Mechanical Engineering), University of California San Diego

TEACHING EXPERIENCE

University of California San Diego

- Teaching Assistant, Department of Bioengineering,
  "Thermodynamics, Statistical Mechanics, Interfacial Phenomena in Living Systems", BENG 223, Dr. Shankar Subramaniam, Winter 2018.
- Teaching Assistant, Department of Mechanical and Aerospace Engineering,
  "Advanced Fluid Mechanics", MAE 101B, Dr. Keiko Nomura, Spring 2017.

PUBLICATIONS

Peer-Reviewed Research Articles

1. [Under preparation] **S. Rahiminejad**, K. Mukund, M. R. Maurya, and S. Subramaniam, "Analysis of gene modules across stages of colorectal cancer from signle cell transcriptomics", 2023.
2. [Under preparation] **S. Rahiminejad**, B. De Sanctis, R. Durbin, P. Pevzner, and A. Mushegian, "Synthetic lethality and the minimal genome size paradox", 2023.
3. **S. Rahiminejad**, M. R. Maurya, K. Mukund, and S. Subramaniam, "Modular and mechanistic changes across stages of colorectal cancer," *BMC cancer*, vol. 22, no 1, pp. 1-14, 2022.
4. **S. Rahiminejad**, M. R. Maurya, and S. Subramaniam, "Topological and functional comparison of community detection algorithms in biological networks," *BMC bioinformatics,* vol. 20, no. 1, pp. 1-25, 2019.

Conference Proceedings

5. **S. Rahiminejad**, M. R. Maurya, K. Mukund, and S. Subramaniam, "Stage-Specific Modulear and Molecular Network Analyses in Colorectal Cancer," *RECOMB/ISCB Conference on Regulatory and Systems Genomics with DREAM Challenges 2022*.
6. **S. Rahiminejad**, M. R. Maurya, and S. Subramaniam, "Comparison of Community Detection Algorithms in Biological Networks from a Topological and Functional Perspective," *2019 AIChеE Annual Meeting*.

ABSTRACT OF THE DISSERTATION

Functional Modularity Methods and Applications for Human Diseases

by

Sara Rahiminejad

Doctor of Philosophy in Engineering Sciences (Mechanical Engineering)

University of California San Diego, 2023

Professor Shankar Subramaniam, Chair
Professor Ratneshwar Lal, Co-Chair

Community detection in complex networks (graphs) has been the subject of investigation in numerous domains. In biological networks, communities are functionally contextual and often provide insights into mechanisms. Detecting communities and analyzing their biological functions is an important aspect of studying biological networks. Communities (aka modules) can yield useful insights into the structure of networks and serve as a basis for analyzing them at topological and functional levels. The work presented in this dissertation is aimed at community detection in human disease (specifically colorectal cancer) networks using different approaches, with the focus on analyzing their biological functions.

The study begins with the exploration of existing community detection algorithms and evaluation of their findings on two important Protein-Protein Interaction (PPI) networks, namely, Saccharomyces cerevisiae (Yeast) and Homo sapiens (Human) at both topological and functional levels. The main criteria to assess the performance of each method are 1) appropriate community size (neither too small nor too large), 2) representation of only one or two broad biological functions within a community, 3) most genes from the network belonging to a pathway should also belong to only one or two communities, and 4) performance speed. These criteria enable us to select one of the best methods for detecting communities in biological networks.

Next, a gene expression microarray dataset of colorectal cancer (CRC) is analyzed to detect stage-specific biomarkers as well as modular mechanisms, potentially causal for the progression of CRC from normal to stages I to IV. Constructing unweighted and weighted correlation networks for each stage, communities are identified and compared topologically and functionally across stages. Short Time-series Expression Miner (STEM) algorithm is also used to detect potential biomarkers having a role in CRC. Constructing a drug-target-PPI network provides insight, in the light of analyzed data, into understanding the functional mechanisms for some of the current drugs used in CRC treatment.

Lastly, gene modules across stages of CRC are analyzed from a single cell transcriptomic dataset to decipher mechanistic changes likely contributing to tumor growth and cancer progression. Cell down-sampling process is firstly performed at stages pT2 to pT4 (as well as right colon) to make cell count equal across different stages and tissue sites. Functional modules at early stage (pT1) and also right colon are identified utilizing Weighted Gene Co-expression Network Analysis (WGCNA). In particular, WGCNA's preservation statistics are used to detect gene modules that exhibit weak/strong preservation of network topology in late stages (pT234) vs. early stage (pT1) as well as left colon vs. right colon. Functional enrichment analysis of the non-preserved modules reveals mechanisms related to the initiation, progression and metastasis of CRC.

Chapter 1 INTRODUCTION

Network science plays a significant role in modeling and understanding complex systems in many domains including but not limited to physics, sociology, computer science, economics, biology, and neuroscience [1, 2, 3]. A network consists of nodes (vertices) and connections or edges between the nodes. The topology of networks is by itself complex. However, it is possible to identify groups of nodes that are relatively densely connected to each other but sparsely connected to other groups of the network. These interconnected groups of nodes are often called communities (clusters or modules) and occur in a wide variety of networks [4, 5, 6, 7]. Communities mark groups of nodes which could share common properties or have similar functions within the network of interest.

Community detection is important for many reasons; first, to classify the functions of nodes in accordance with their structural positions in their communities [8, 9], second, to help better understand the properties of dynamic processes taking place in a network [10], and third, to improve the performance and efficiency of processing, analyzing, and storing networked data [11].

Communities also have concrete applications. In social networks, communities represent groups of individuals with mutual interest and background [12]. In citation networks, they represent groups of related papers in one research area and identify scholars sharing research interests [13]. In brain networks, they represent groups of nodes that are intricately interconnected and that could perform local computations and give insights into structural units of the brain [14]. In biological networks, they represent groups of nodes that enable functional annotation of constituent biomolecules, detection of regulatory elements associated with disease phenotypes, or discovery of targets for therapeutic interventions [15, 16].

Considering the importance of community detection in networks, it is not surprising that many algorithms and methods have been developed for community detection algorithms during the past decade [17, 18, 19, 20]. The goal of all these methods is to identify meaningful communities, while keeping the computational complexity as low as possible.

Community detection methods can be broadly categorized into two types, *agglomerative* (bottom-up) methods and *divisive* (top-down) methods. In agglomerative methods, starting from the set of all nodes and no edges, links are iteratively added between pairs of nodes in the order of decreasing weight. Nodes are then grouped into larger and larger communities until the whole network is constructed. Louvain algorithm is one of the famous agglomerative algorithms [20]. Divisive methods start from the whole network and iteratively cut the edges, which results in the division of the network into smaller and smaller disconnected subnetworks (aka communities). The crucial point in a divisive algorithm is the selection of the edges to be cut, which have to be those connecting the communities and not those within communities. The Edge-Betweenness (or Girvan-Newman) [15, 17] and Leading Eigen [21] are a few examples of the divisive algorithms.

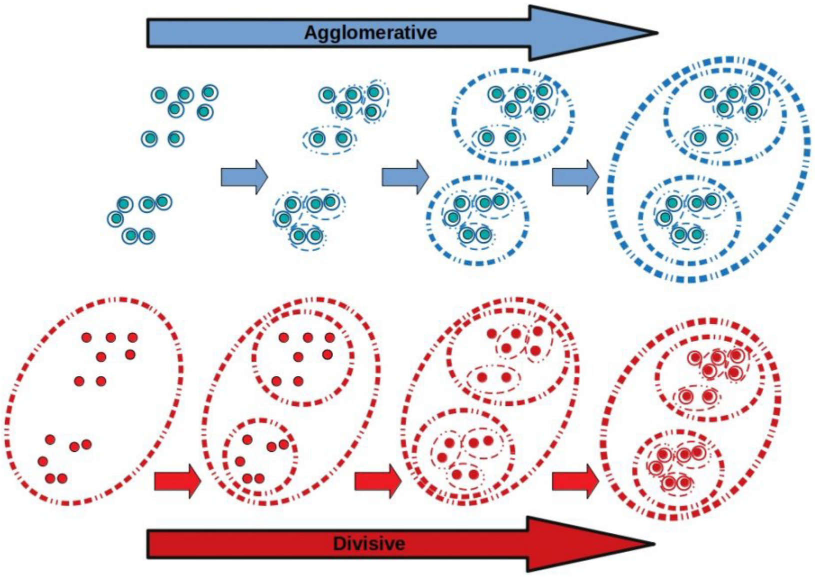


Figure 1.1 Agglomerative vs. divisive clustering.

1.1 Chapter Outline

Chapter 2, in full, is the material of the manuscript published in BMC Bioinformatics [22] and reviews different algorithms for detecting communities in biological networks with application to two important Protein-Protein Interaction (PPI) network, namely, Yeast and Homo Sapiens and compares them topologically and functionally. This study provides the first documented example for benchmarking community detection algorithms which could be used in biological networks. Overall, one of the divisive algorithms (Louvain) is found to be the best method for biological networks to find reasonably sized communities in a reasonable time.

Chapter 3, in full, is the material of the manuscript published in BMC Cancer [23], focuses on modeling each stage of Colorectal Cancer (CRC) as a molecular network and identifying their communities separately and then analyzing their progression for a better mechanistic interpretation of how CRC progresses. A few approaches are also used to detect candidate biomarkers with substantial/monotonic changes across stages. Additionally, a drug-target-PPI network is generated to provide insights into understanding stage-specific functional mechanisms associated with some of the current drugs used in CRC treatment.

Chapter 4 is a modified presentation of the manuscript being prepared for submission, analyzing gene modules (communities) across stages of CRC from a single-cell transcriptomics dataset. Due to the unequal number of cells at different stages and tissue sites, a cell down-sampling procedure is first applied to make the number of cells equal. Then, Weighted Gene Co-expression Network Analysis (WGCNA) is performed on the first stage as well as the right colon to detect their communities and also find the low/non-preserved modules out of them.

Chapter 5 represents the major findings of my research and conclusions of the dissertation.

Chapter 2 TOPOLOGICAL AND FUNCTIONAL COMPARISON OF COMMUNITY DETECTION ALGORITHMS IN BIOLOGICAL NETWORKS

2.1 Abstract

Background: Community detection algorithms are fundamental tools to uncover important features in networks. There are several studies focused on social networks but only a few deals with biological networks. Directly or indirectly, most of the methods maximize modularity, a measure of the density of links within communities as compared to links between communities. Results: Here, we analyze six different community detection algorithms, namely, Combo, Conclude, Fast Greedy, Leading Eigen, Louvain, and Spinglass, on two important biological networks to find their communities and evaluate the results in terms of topological and functional features through Kyoto Encyclopedia of Genes and Genomes pathway and Gene Ontology term enrichment analysis. At a high level, the main assessment criteria are 1) appropriate community size (neither too small nor too large), 2) representation within the community of only one or two broad biological functions, 3) most genes from the network belonging to a pathway should also belong to only one or two communities, and 4) performance speed. The first network in this study is a network of Protein-Protein Interactions (PPI) in *Saccharomyces cerevisiae* (Yeast) with 6,532 nodes and 229,696 edges and the second is a network of PPI in *Homo sapiens* (Human) with 20,644 nodes and 241,008 edges. All six methods perform well, i.e., find reasonably sized and biologically interpretable communities, for the Yeast PPI network but the Conclude method does not find reasonably sized communities for the Human PPI network. Louvain method maximizes modularity by using an agglomerative approach, and is the fastest method for community detection. For the Yeast PPI network, the results of Spinglass method are most similar to the results of Louvain method with regard to the size of communities and core pathways they identify, whereas for the Human PPI

network, Combo and Spinglass methods yield the most similar results, with Louvain being the next closest. Conclusions: For Yeast and Human PPI networks, Louvain method is likely the best method to find communities in terms of detecting known core pathways in a reasonable time.

2.2 Background

The use of networks to study complex interacting systems has been applied to many domains during the last two decades, including sociology, physics, computer science, and biology. An important task in the analysis of networks lies in the identification of communities or modules whose membership share one or more common features of the system. The problem that community detection attempts to solve is the identification of groups of nodes with more and/or better interactions amongst its members than between its members and the remainder of the network [17, 24]. For example, in social networks, a community may correspond to groups of friends who attend the same school or live in the same neighborhood; while in a biological network, communities may represent functional modules of interacting proteins.

Edges in a biological network may represent various types of direct interactions and indirect effects. Examples of direct interactions include protein-protein interactions as part of signaling pathways or as part of protein complexes and substrate-enzyme interactions. Indirect effects may include transport processes and regulatory effects, which, in most cases, can be substituted with a subnetwork of several direct interactions when modeled at a finer granularity. Examples of the latter are cholesterol and ion transport across the plasma membrane and protein-DNA interactions in gene-regulatory networks. Thus, in the context of a cell or tissue, subnetworks or communities may correspond to various cellular processes, pathways and functions, in which its components (nodes) exhibit a higher-degree of interaction as compared to those from outside the pathway.

Majority of the methods for community detection in networks are based on maximization of modularity. While the modularity metric Q, of a network, is defined in the Methods section, intuitively, given a network, if it can be partitioned in such a way that only a few connections exist between the nodes of different partitions and most connections are among the nodes within the partitions, then the modularity will be high. It is interesting to note that the modularity of a sparse network of fully connected subnetworks is higher than that of a fully connected network, which is zero. Any partition of a fully connected network results in $Q < 0$. Brandes et al. have carried out extensive theoretical analysis of properties of modularity and complexity of its maximization [25].

One of the most important objectives of any large-scale omics study is to identify mechanisms for specific functions and phenotypes in a chosen context. Biological networks derived from genome-scale experimental data and/or legacy knowledge are generally large and complex with thousands of nodes and many thousands of connections. Associating meaningful biological functions and interpretations to such networks is impossible. However, these large networks can be broken down into smaller (sub) networks (also called as modules or communities) which are more amenable to biological interpretation. Such communities are expected to represent one or a few biological functions and they may facilitate discovery of mechanisms relating the causes or perturbations to the observed phenotypes. Thus, community detection can provide valuable biological insights.

Several methods have been developed to find communities in networks using tools and techniques from different disciplines such as applied mathematics or statistical physics [26]. All these methods try to identify meaningful communities, while keeping the computational complexity of the underlying algorithm low [27]. Although these methods have proven to be successful in some cases, there is no guarantee that the resulting communities provide the best functional description of the system. Hence, selecting a suitable method to detect communities in a network is challenging. While there have been some studies comparing different methods for community detection [27], their focus has been on Lancichinetti, Fortunato, Radicchi (LFR) benchmark networks (artificial networks that have heterogeneity in the distributions of degree of nodes and the size of communities) [28]; comparisons with respect to biological networks are lacking.

Classical community detection algorithms initially divide networks into communities according to some network features such as edge betweenness. One of the most popular and prominent algorithms that uses edge betweenness is the Girvan-Newman algorithm [15, 17]. In this method edges are progressively removed from the original network till the modularity reaches its maximum value, making it an optimization problem. The connected nodes of the remaining network are the communities. The Girvan-Newman algorithm has been successfully applied to a variety of networks, including networks of email messages. However, its computational complexity, $O(m^2n)$ for a network with n nodes and m edges, practically restricts its use to networks of at most a few thousand nodes. There are other optimization-based algorithms with different objective functions that provide different approaches to solve the community detection problem. For example, Leading Eigen [21] algorithm also tries to maximize modularity but the modularity is expressed in the form of the eigenvalues and eigenvectors of a matrix called the modularity matrix. Spinglass method minimizes the Hamiltonian of the network [29].

Since the early 2000s, several methods have been developed that divide networks into communities based on the modularity [18, 20, 30, 31, 32, 33]. The modularity criterion was revisited in 2005 when Duch and Arenas proposed a divisive algorithm [34] that optimizes the modularity using a heuristic search based on the Extremal Optimization (EO) algorithm proposed by Boettcher and Percus [35, 36]. Pizzuti has suggested an algorithm named GA-net that uses a special assessment function described as the community score in addition to the modularity function [37]. There are also other approaches to the community detection problem in which the use of multiple objectives is preferred over the use of a single objective for complex networks. Since the objectives are usually directly related to the network properties, one advantage of using multi-objective optimization is that it balances among the multiple (important) properties of the network. The benefits of using multi-objective approach have been explained by Shi et al. [38].

In this manuscript, we briefly review eight algorithms for finding communities in biological networks such as Protein-Protein Interaction (PPI) networks (discussed in the Methods section). In such networks, each node represents a protein (or gene) and each edge represents an interaction between two proteins. In particular, we will apply six algorithms to the Yeast PPI network with 6,532 nodes and 229,696 edges and the Human PPI network with 20,644 nodes and 241,008 edges. Using several topological metrics, we assess which methods provide similar (or dissimilar) results. We evaluate the biological interpretation of the communities identified and compare the results in terms of their functional features. At a high level, the main criteria for assessment of the methods are 1) appropriate community size (neither too small nor too large), 2) representation within the community of only one or two broad biological functions, 3) most genes from the network belonging to a pathway should also belong to only a few communities, and 4) performance speed.

This paper is organized as follows: in the next section we will present the results of applying six methods on the Yeast and Human PPI networks and compare the communities based on their topological and functional features. In the last part of this section, we will describe an orthology analysis between the communities detected for the Yeast PPI network and the communities detected for the Human PPI network. In the following section, we will present discussion on the results providing insights into the algorithmic similarities and robustness of some of the methods. In the section after that, we will provide the conclusion of our paper. In the Methods section, we will describe eight different methods for finding communities in networks. We will also introduce three metrics to compare the communities identified by the algorithms.

2.3 Results

Six community detection methods, namely, Combo, Conclude, Fast Greedy, Leading Eigen, Louvain, and Spinglass, have been applied to the Yeast PPI network with 6,532 nodes and 229,696 edges and the Human PPI network with 20,644 nodes and 241,008 edges. A detailed description of the methods is included in the Methods section. We used the BioGRID database [39, 40] for the PPI networks for Yeast and Human. Since our focus in this paper is on undirected and unweighted networks, we removed repeated edges and self-loops from our data set.

In the first part of this section, we will present the results for the Yeast PPI network. In the second part, the results for the Human PPI network will be presented. In the third part, an orthology comparison will be provided between the Yeast and Human PPI networks.

2.3.1 Yeast PPI Network

Among the methods tested to find communities of the Yeast PPI network, Combo, Conclude, Fast Greedy, Leading Eigen, Louvain, and Spinglass give good partitioning results, i.e., the size of communities detected are not too small or too large compared to the size of the original network. Since the Yeast PPI network has 6,532 nodes, Girvan-Newman algorithm is not an appropriate method to detect communities. It takes about 44 minutes (on a PC with 4 GB RAM and 4 2.4 GHz processors) for Rattus PPI network with 3,379 nodes and 4,580 edges. Its computational complexity is proportional to m^2n (n: number of nodes and m: number of edges), so, it will take

~ 148 days to find communities in the Yeast PPI network (using the computational resource mentioned above). Infomap is also not a good method based on the size of communities it detects; the largest community has 6,195 nodes and the smallest one has just 2 nodes. Since very small communities (e.g., those with less than 100 nodes) are not expected to yield significant biological insights, we will not consider them in our analysis. We note that there may be some exceptions.

In the next subsection, first we will compare the methods from a topological perspective of the communities identified. Then we will provide a functional comparison. To begin with, the results for all these methods applied on the Yeast PPI network are described in Table 2.1.

Table 2.1: Number of nodes and edges for communities detected using different methods for the Yeast PPI network with 6,532 nodes and 229,696 edges.

Combo (8)	# nodes	2,231	1,514	1,337	1,284						
Q = 0.2654	# edges	25,137	23,690	38,523	30,585						
Conclude (66)	# nodes	788	744	602	468	423	359	288	271	252	199
Q = 0.2468	# edges	14,585	10,506	3,272	988	5,826	4,123	1,404	3,486	940	1,703
F. Greedy (10)	# nodes	2,608	2,410	1,466							
Q = 0.2112	# edges	61,665	72,998	7,180							
L. Eigen (4)	# nodes	2,661	1,910	984	977						
Q = 0.1686	# edges	75,812	28,664	7,373	7,203						
Louvain (9)	# nodes	1,538	1,472	1,190	1,151	993	131				
Q = 0.2643	# edges	16,015	23,394	13,247	31,202	22,553	676				
Spinglass (9)	# nodes	1,607	1,473	1,194	1,148	1,076					
Q = 0.2681	# edges	16,616	23,876	12,282	32,641	23,854					

The number in parenthesis after the name of each method represents the number of communities detected by that method. Modularity scores are also provided for different methods. For each method, we only consider the communities with 100 or more nodes and list up to 10 communities.

2.3.1.1 Comparison based on topological features of communities

Using three different metrics, namely, Rand Index (RI), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI) (described in the Methods section), we are able to compare different pair of methods. Table 2.2 represents the results of comparing six methods (Combo, Conclude, Fast Greedy, Leading Eigen, Louvain, and Spinglass) with respect to three topological metrics (RI, ARI and NMI).

Table 2.2: Comparison of different methods with respect to three topological metrics for the Yeast PPI network.

		Combo	Conclude	F. Greedy	L. Eigen	Louvain	Spinglass
RI	**Combo**	1	0.7608	0.7157	0.6788	0.8319	0.8409
ARI	**Combo**	1	0.1466	0.3125	0.1942	0.5163	0.5479
NMI	**Combo**	1	0.2905	0.4024	0.2447	0.5413	0.5723
RI	**Conclude**		1	0.6815	0.7061	0.8083	0.8012
ARI	**Conclude**		1	0.0818	0.0825	0.1659	0.1637
NMI	**Conclude**		1	0.1956	0.1472	0.3016	0.2924
RI	**F. Greedy**			1	0.6334	0.7098	0.7129
ARI	**F. Greedy**			1	0.146	0.2629	0.2764
NMI	**F. Greedy**			1	0.1918	0.3545	0.3652
RI	**L. Eigen**				1	0.6952	0.6936
ARI	**L. Eigen**				1	0.188	0.1914
NMI	**L. Eigen**				1	0.2231	0.2285
RI	**Louvain**					1	0.9021
ARI	**Louvain**					1	0.6922
NMI	**Louvain**					1	0.6644
RI	**Spinglass**						1
ARI	**Spinglass**						1
NMI	**Spinglass**						1

When a method is compared with itself, RI, ARI and NMI are 1 (diagonal elements). Larger (smaller) the value of RI, ARI and NMI, the more (less) similar are the two methods being compared.

Based on the results of Table 2.2, Louvain and Spinglass are most similar to each other amongst all pairs of comparisons. To maintain consistency in finding dissimilar methods, we selected a method which is dissimilar to Louvain, e.g., Conclude or Leading Eigen. Since Conclude finds 66 communities with sizes (number of nodes) ranging from 3 to 788, we compare Louvain with Leading Eigen here. We present the results from comparing Louvain and Conclude in the Supplementary Tables.

Table 2.3 provides Jaccard index (as a percentage) between communities identified by Louvain and Spinglass. We used the *intersect* function in R to find common genes between two communities and then divided the number of common genes by the total number of unique genes between the two communities (*union* function in R) to get the Jaccard index. Table 2.4 uses the same approach to calculate Jaccard index for communities detected by dissimilar methods, in particular, Louvain and Leading Eigen. The rest of Jaccard index matrices amongst all pairs of communities for all methods can be found in Table S2.1.

Table 2.3: Jaccard index (%) between the communities identified by two similar methods (Louvain and Spinglass) for the Yeast PPI network.

	L1 (1,538)	**L2 (1,472)**	**L3 (1,190)**	**L4 (1,151)**	**L5 (993)**
S1 (1,607)	**75.7**	2.35	1.79	0.15	2.43
S2 (1,473)	4.66	**70.53**	2.03	1.16	1.07
S3 (1,194)	2.01	1.1	**74.27**	2.95	1.03
S4 (1,148)	1.32	1.08	2.18	**81.02**	0.36
S5 (1,076)	0.74	0.33	0.46	0.09	**85.56**

L1 to L5: communities detected by Louvain; S1 to S5: communities detected by Spinglass. The numbers in parenthesis represent the number of genes in each community. Community pairs with maximum overlap are indicated in bold text.

Table 2.4: Jaccard index (%) between the communities identified by two dissimilar methods (Louvain and Leading Eigen) for the Yeast PPI network.

	L1 (1,538)	**L2 (1,472)**	**L3 (1,190)**	**L4 (1,151)**	**L5 (993)**
LE1 (2,661)	15.29	**40.67**	15.16	4.64	2.78
LE2 (1,910)	3.7	2.98	12.08	**44.87**	15.43
LE3 (984)	10.13	1.4	14.18	0.85	**27.47**
LE4 (977)	**33.14**	6.29	3.78	0.76	4.23

L1 to L5: communities detected by Louvain; LE1 to LE4: communities detected by Leading Eigen. The numbers in parenthesis represent the number of genes in each community. Community pairs with maximum overlap are indicated in bold text.

2.3.1.2 Comparison based on biological/functional features of communities

As described in the previous subsection, Louvain and Spinglass are most similar to each other and Louvain and Leading Eigen are most dissimilar. In order to know which communities of similar and dissimilar methods have to be compared to each other, we analyzed Tables 2.3 and 2.4 (Jaccard index) for all pairs of communities between similar and dissimilar methods. After selecting pairs of communities with highest value of Jaccard index for each column, we used Database for Annotation, Visualization and Integrated Discovery (DAVID) version 6.8 [41, 42] to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) [43] pathway and Gene Ontology (GO) term (GOTERM_BP_3) enrichment analysis for each community. In the following Tables (Table 2.5 through Table 2.7), we have considered pathways with more than 10 genes and with *p*-values less than or equal to 0.01. The number in parenthesis (e.g., after L1 or S1) is the number of genes that DAVID could annotate for that specific community for the Yeast PPI network. For example, the first community of Louvain (L1) has 1,538 genes (Table 2.1) and of those, DAVID is able to annotate 1,481 genes. In Tables 2.5, 2.6, and 2.7, the first column lists the broad category of pathways (M: Metabolism, CP: Cellular Processes, GIP: Genetic Information Processing, HD: Human Diseases, and OS: Organismal Systems). The second column lists the different pathways enriched. Columns 3 and 7 (#) represent number of genes enriched in the pathways, columns 4 and 8 (*p*-val) represent *p*-values for those pathways in the communities compared, and columns 5 and 9 (FE) represent Fold Enrichment for the pathways. FE is defined as $(s/b)/(k/N)$ where b is the total number of genes in a chosen pathway; s, the number of genes from the community in this pathway; N, the total number of genes for the species; and k, the number of genes in the community; all the four numbers are based on intersection/overlap with the respective DAVID database (e.g., KEGG). Essentially, FE represents the relative increase or decrease of the fraction of genes from the set of interest belonging to a pathway as compared to the genes from a background set (generally covering the whole-genome) belonging to the same pathway. The values in columns 5 and 9 are shaded from light orange to dark orange with increasing FE. Column 6 (Com) is the number of genes common to both communities for the different pathways.

2.3.1.2.1 Comparing similar methods

As the first column of Table 2.3 shows, the first community of Louvain (L1) and the first community of Spinglass (S1) have the maximum overlap and the results of comparing KEGG pathway enrichment analysis between L1 and S1 are presented in Table 2.5. Table S2.2 shows the results of comparing GO term enrichment analysis between these two communities. L2 and S2 are 71% similar to each other based on Table 2.3 and they are compared in Table 2.6 for KEGG pathway enrichment analysis and in Table S2.3 for GO term enrichment analysis. The rest of the comparison Tables (KEGG pathway enrichments analysis for L3 vs. S3, L4 vs. S4, and L5 vs. S5) are in the Tables S2.4-S2.6. Since DAVID did not find any pathways for small communities, such as L6 which has 131 nodes, those communities are not considered in the comparison Tables.

Based on Table 2.3, L1 and S1 are 76% similar to each other in terms of Jaccard index. In Table 2.5, KEGG pathway enrichment results of these two communities reveal that majority of these genes are related to various metabolic pathways such as carbohydrate metabolism, energy metabolism, amino acid metabolism, and metabolism of cofactors and vitamins. The top four pathways represent broad metabolism pathways. There are 13 pathways categorized as amino acid metabolism such as cysteine and methionine metabolism, or glycine, serine and threonine metabolism. Among pathways that are categorized as energy metabolism, oxidative phosphorylation is the one with the lowest *p*-value.

Table 2.5: A Comparison of KEGG pathway enrichment results between the first community of Louvain (L1) and the first community of Spinglass (S1) for the Yeast PPI network.

			L1 (1,481)				**S1 (1,556)**	
	Term (Pathway/function)	#	*p*-val	FE	Com	#	*p*-val	FE
M	Metabolic pathways	346	6.62E-86	2.16	337	357	2.64E-86	2.13
M	Biosynthesis of secondary metabolites	164	5.41E-38	2.37	156	167	5.61E-37	2.31
M	Biosynthesis of antibiotics	128	6.73E-32	2.5	123	131	1.04E-31	2.44
M	Carbon metabolism	79	5.22E-26	2.97	78	84	3.85E-29	3.01
M	Oxidative phosphorylation	55	3.83E-21	3.27	55	55	4.44E-20	3.12
M	Citrate cycle (TCA cycle)	30	5.55E-16	4.02	29	30	2.32E-15	3.83
M	Biosynthesis of amino acids	68	1.20E-14	2.37	67	74	1.39E-17	2.46
CP	Peroxisome	30	4.38E-12	3.38	30	31	1.45E-12	3.33
M	2-Oxocarboxylic acid metabolism	27	1.70E-10	3.31	27	28	5.42E-11	3.27
M	Pyruvate metabolism	27	7.41E-09	2.97	27	30	4.83E-11	3.14
M	Glyoxylate and dicarboxylate metabolism	20	8.84E-08	3.3	20	22	1.89E-09	3.46
M	Cysteine and methionine metabolism	24	1.87E-07	2.86	24	27	1.49E-09	3.06
M	Glycine, serine and threonine metabolism	21	2.05E-06	2.81	21	23	1.22E-07	2.94
M	Arginine and proline metabolism	16	3.54E-06	3.26	14	15	4.72E-05	2.92
M	Pentose phosphate pathway	19	3.79E-06	2.91	19	19	7.91E-06	2.77
M	Tryptophan metabolism	14	4.85E-06	3.53	12	12	4.79E-04	2.88
M	beta-Alanine metabolism	12	4.90E-06	3.96	12	12	8.11E-06	3.77
M	Pentose and glucuronate interconversions	13	5.16E-06	3.71	13	13	8.84E-06	3.54
M	Glycolysis / Gluconeogenesis	30	6.30E-06	2.22	30	32	1.39E-06	2.25
M	Starch and sucrose metabolism	23	1.22E-05	2.46	22	22	1.06E-04	2.25
M	Sulfur metabolism	12	1.81E-05	3.67	12	12	2.95E-05	3.5
M	Selenocompound metabolism	11	1.82E-05	3.93	11	11	2.86E-05	3.75
M	Methane metabolism	18	2.16E-05	2.75	18	21	1.75E-07	3.06
M	Tyrosine metabolism	11	1.73E-04	3.37	11	13	2.27E-06	3.79
M	Vitamin B6 metabolism	10	2.10E-04	3.57	10	10	3.12E-04	3.4
M	Glutathione metabolism	15	2.18E-04	2.68	15	16	7.62E-05	2.72
M	Fatty acid degradation	13	2.23E-04	2.93	12	12	1.79E-03	2.58
M	Porphyrin and chlorophyll metabolism	12	3.05E-04	3.02	12	12	4.79E-04	2.88
M	Alanine, aspartate and glutamate metabolism	17	3.18E-04	2.43	17	19	3.14E-05	2.59
M	Galactose metabolism	14	9.73E-04	2.5	14	14	1.57E-03	2.38
	Unique genes for all pathways	**406**			**394**	**417**		

The numbers inside parenthesis after L1 and S1 represent the number of genes that DAVID could annotate, which is generally less than the number of genes in those communities. The first column lists the broad category of pathways (M: Metabolism, CP: Cellular Processes). Many pathways enriched in L1 and S1 have good overlap (a large number of genes are common). FE: Fold Enrichment. False Discovery Rate (FDR) values for all pathways and both communities are approximately 1.10E+3 times p-value (the factor 1.10E+3 is related to the size of the community).

In terms of enzyme commission annotation, there are 1,738 enzyme-coding genes in the entire network. L1 and S1 have 571 and 605 enzyme-coding genes, respectively. Of these, 529 enzyme-coding genes are common between the two communities, showing a significant overlap. There are a few enzyme-coding genes which are present in L1 but not in S1 such as aminoacyl-tRNA hydrolase (PTH1) or glutamate 5-kinase (PRO1). Similarly, for genes that are present in S1 but not in L1, sulfuric ester hydrolase (BDS1) is an example. Since both Louvain and Spinglass find 9 non-overlapping communities, all enzyme-coding genes are part of one of the communities.

Table 2.6 shows KEGG pathway enrichment results for the communities L2 and S2. All pathways are related to genetic information processing with approximately similar genes enriched in the two methods. The first pathway with the lowest *p*-value is ribosome which is a complex molecule made of ribosomal RNA molecules and proteins. There are 151 genes enriched in L2 and 141 genes enriched in S2 for this pathway. Of these, 139 genes are common (a 92% overlap). Similar trend is observed for other pathways as well, e.g., there is a 95% overlap between L2 and S2 for Spliceosome and 95% overlap for RNA transport.

Table 2.6: A comparison of KEGG pathway enrichment results between the second community of Louvain (L2) and the second community of Spinglass (S2) for the Yeast PPI network.

		L2 (1,378)				**S2 (1,344)**		
	Term (Pathway/function)	#	*p*-val	FE	Com	#	*p*-val	FE
GIP	Ribosome	151	7.06E-65	3.25	139	141	1.96E-49	2.87
GIP	Spliceosome	74	9.60E-38	3.65	70	70	3.19E-30	3.26
GIP	RNA transport	82	2.37E-37	3.44	78	79	1.46E-31	3.13
GIP	Ribosome biogenesis in eukaryotes	75	2.86E-31	3.28	72	74	2.69E-28	3.06
GIP	RNA degradation	40	1.78E-10	2.56	38	42	3.90E-11	2.53
	Unique genes for all pathways	**380**			**356**	**365**		

The numbers inside parenthesis after L2 and S2 represent the number of genes that DAVID could annotate, which is generally less than the number of genes in those communities. The first column lists the broad category of pathways (GIP: Genetic Information Processing). Many pathways enriched in L2 and S2 have good overlap (a large number of genes are common). FE: Fold Enrichment. False Discovery Rate (FDR) values for all pathways and both communities are approximately 1.05E+3 times p-value.

The GO term enrichment results shown in Tables S2.2 and S2.3 also verify the similarity between L1 and S1, and L2 and S2, respectively. Counting all genes for all pathways in Table S2.2 yields 1,062 unique genes for L1 and 1,103 unique genes for S1 and of these, 957 genes are common between the two communities, which is an 87% overlap. This similarity value is 84% between L2 and S2 (Table S2.3).

Table S2.4 provides a comparison between the communities L3 and S3. The pathways enriched are classified into four different groups (metabolic processes, environmental information processing, cellular processes, and human diseases) as opposed to just one or two. Still, the overlap between L3 and S3 communities for each of the pathways is more than 80%.

L4 and S4 are 81% similar to each other based on Table 2.3 and the results of their comparison are shown in Table S2.5. Most pathways for these two communities are related to genetic information processing category. There are two pathways related to cellular processes and two pathways related to metabolic processes. Based on KEGG pathway enrichment results, there is a good overlap between genes enriched in different pathways for these two communities. For example, 71 genes of L4 are enriched in cell cycle pathway and 77 genes of S4 are also enriched in this pathway. Among genes enriched in cell cycle pathway, 70 genes are common between L4 and S4, giving a 91% overlap.

Table S2.6 compares L5 and S5. Based on Table 2.3, they are 86% similar to each other. KEGG pathway enrichment results of these two communities show that most pathways are related to metabolic processes and there are also other pathways related to other categories such as endocytosis, which is in the cellular processes category. The results of KEGG pathway also verify the similarity of Table 2.3. As seen from the enriched pathways, almost all of them have the same genes enriched in both communities. For example, there are 29 genes for L5 enriched in N-Glycan biosynthesis and the same genes are found in S5 in the same pathway.

2.3.1.2.2 Comparing dissimilar methods

In this subsection, we will compare the methods that are most dissimilar to each other, namely, Louvain and Leading Eigen. As Table 2.4 shows, the first community of Louvain has the maximum overlap with the fourth community of Leading Eigen (LE4). The results of this comparison based on KEGG and GO term enrichment analysis are shown in Tables 2.7 and Table S2.7, respectively. The rest of the comparisons can be found in Table S2.8 for L2 vs. LE1, Table S2.9 for L4 vs. LE2, and Table S2.10 for L5 vs. LE3.

The first pathway of Table 2.7 with the lowest *p*-value is metabolic pathways with 346 genes enriched in L1 and 253 genes enriched in LE4. Of these genes, 224 genes are common between the two communities, which is a 65% overlap. In contrast, the first pathway of Table 2.5 shows that 337 genes are common between L1 and S1, which is a 94% overlap. There are some pathways in Table 2.7 that are blank for LE4 such as biosynthesis of amino acids. For these pathways, although there are some genes enriched in L1, there are no genes enriched in LE4 or if there are any, the *p*-value for the pathway is higher than the defined cut-off of 0.01. Counting all genes for all pathways yields 406 unique genes for L1 and 273 unique genes for LE4 and of these, 239 genes are common between the two communities, which is a 59% overlap. Based on GO term enrichment analysis shown in Table S2.7, there are 474 genes common between L1 and LE4 out of 1,062 unique genes for L1 and 662 unique genes for LE4, which is a 45% overlap. These relatively low best-overlaps also confirm that these two methods are dissimilar to each other.

Table 2.7: A comparison of KEGG pathway enrichment results between the first community of Louvain (L1) and the fourth community of Leading Eigen (LE4) for the Yeast PPI network.

		L1 (1,481)				**LE4 (902)**		
	Term (Pathway/function)	#	*p*-val	FE	Com	#	*p*-val	FE
M	Metabolic pathways	346	6.62E-86	2.16	224	253	6.65E-56	2.14
M	Biosynthesis of secondary metabolites	164	5.41E-38	2.37	107	124	2.69E-27	2.43
M	Biosynthesis of antibiotics	128	6.73E-32	2.5	84	96	3.08E-22	2.54
M	Carbon metabolism	79	5.22E-26	2.97	56	62	7.38E-20	3.15
M	Oxidative phosphorylation	55	3.83E-21	3.27	49	50	2.07E-22	4.02
M	Citrate cycle (TCA cycle)	30	5.55E-16	4.02	25	25	6.78E-13	4.52
M	Biosynthesis of amino acids	68	1.20E-14	2.37	51	57	4.16E-14	2.68
CP	Peroxisome	30	4.38E-12	3.38				
M	2-Oxocarboxylic acid metabolism	27	1.70E-10	3.31	21	21	8.73E-08	3.47
M	Pyruvate metabolism	27	7.41E-09	2.97	19	21	9.37E-07	3.12
M	Glyoxylate and dicarboxylate metabolism	20	8.84E-08	3.3	14	15	2.11E-05	3.34
M	Cysteine and methionine metabolism	24	1.87E-07	2.86	18	21	1.66E-07	3.38
M	Glycine, serine and threonine metabolism	21	2.05E-06	2.81	13	14	1.42E-03	2.53
M	Arginine and proline metabolism	16	3.54E-06	3.26	10	10	5.12E-03	2.76
M	Pentose phosphate pathway	19	3.79E-06	2.91	15	16	1.15E-05	3.31
M	Tryptophan metabolism	14	4.85E-06	3.53	9	9	4.14E-03	3.06
M	beta-Alanine metabolism	12	4.90E-06	3.96	8	8	2.83E-03	3.56
M	Pentose and glucuronate interconversions	13	5.16E-06	3.71	8	8	7.75E-03	3.09
M	Glycolysis / Gluconeogenesis	30	6.30E-06	2.22	19	25	9.78E-06	2.5
M	Starch and sucrose metabolism	23	1.22E-05	2.46				
M	Sulfur metabolism	12	1.81E-05	3.67				
M	Selenocompound metabolism	11	1.82E-05	3.93				
M	Methane metabolism	18	2.16E-05	2.75	11	13	1.24E-03	2.69
M	Tyrosine metabolism	11	1.73E-04	3.37	7	8	4.84E-03	3.31
M	Vitamin B6 metabolism	10	2.10E-04	3.57				
M	Glutathione metabolism	15	2.18E-04	2.68	12	14	3.88E-05	3.38
M	Fatty acid degradation	13	2.23E-04	2.93				
M	Porphyrin and chlorophyll metabolism	12	3.05E-04	3.02				
M	Alanine, aspartate and glutamate metabolism	17	3.18E-04	2.43				
M	Galactose metabolism	14	9.73E-04	2.5				
	Unique genes for all pathways	**406**			**239**	**273**		

The numbers inside parenthesis after L1 and LE4 represent the number of genes that DAVID could annotate, which is generally less than the number of genes in those communities. The first column lists the broad category of pathways (M: Metabolism, and CP: Cellular Processes). FE: Fold Enrichment. False Discovery Rate (FDR) values for all pathways and both communities are approximately 1.10E+3 times p-value.

Based on Table 2.2, Louvain and Conclude methods are also dissimilar to each other. We compared the communities obtained from these two methods. As Table S2.1 shows, the first community of Louvain (L1) has the maximum overlap with the third community of Conclude (CL3). The results of this comparison based on KEGG pathway enrichment analysis are shown in Table S2.11. The metabolic pathway is the most enriched pathway with 346 genes enriched in L1, 230 genes enriched in CL3, and 173 genes common between the two communities, which is a 50% overlap. Counting all unique genes for all pathways yields 406 genes for L1 and 241 genes for CL3 and of these, 181 genes are common between the two communities (45% overlap). This similarity value is close to what we calculated for Louvain vs. Leading Eigen, using KEGG pathway and GO term enrichment analysis (Table 2.7 and Table S2.7). The rest of the comparisons can be found in Table S2.12 for L2 vs. CL2, Table S2.13 for L4 vs. CL1, and Table S2.14 for L5 vs. CL5. Overall, dissimilarity at the topological level translates into dissimilarity at the functional level as well.

2.3.2 Human PPI Network

Six methods, namely, Combo, Conclude, Fast Greedy, Leading Eigen, Louvain, and Spinglass, have been applied to the Human PPI network with 20,644 nodes and 241,008 edges. Although all of them were able to find communities, we will not consider the results of Conclude because it finds 495 communities, many of which are very small communities with less than 50 nodes. For Combo and Spinglass, since they use a random number generator in the procedure of finding communities, we ran them 10 times with 10 different seeds between 0 and 10,000 and used the results from the run with the largest modularity. Modularity scores and the number of communities detected in each run for Combo and Spinglass are summarized in Table 2.8 and Table 2.9, respectively.

Table 2.8: Modularity scores and number of communities detected by Combo for the Human PPI network.

Modularity	0.3735	0.3734	0.3734	0.3729	0.3723
# of communities detected	11	13	11	15	11
Modularity	0.3723	0.3718	0.3715	0.3711	0.3704
# of communities detected	13	12	10	12	13

Table 2.9: Modularity scores and number of communities detected by Spinglass for the Human PPI network.

Modularity	0.3729	0.3727	0.3725	0.3725	0.3724
# of communities detected	21	22	22	24	21
Modularity	0.3721	0.3716	0.3716	0.3711	0.3688
# of communities detected	22	23	21	23	23

After finding modularity scores and communities for 10 runs, we selected communities corresponding to the largest modularity score which is 0.3735 for Combo (11 communities) and 0.3729 (21 communities) for Spinglass.

The results of comparing all methods excluding Conclude are presented in Table 2.10. As seen from the table, Louvain and Spinglass are more similar to each other as compared to all other pairs of methods except Combo and Spinglass. Hence, we will compare Combo and Spinglass as well here. Since they are more similar to each other than Louvain and Spinglass, they will be compared first. The first community of Combo (C1) and the first community of Spinglass (S1) have been compared to each other using KEGG pathway enrichment analysis and the results for top 10 pathways are presented in Table 2.11. Table 2.12 presents the results of comparing top 10 pathways for the first community of Louvain (L1) and the first community of Spinglass (S1). Organization of these two tables is the same as that for Tables 2.5, 2.6, and 2.7 in the previous subsection. The complete versions of Tables 2.11 and 2.12 are in the Tables S2.15 and S2.16, respectively. The results of comparing all pathways for the communities C1 and S1 and L1 and S1 (with *p*-values less than 0.01) are illustrated in Figure 2.1 and Figure 2.2, respectively. The pie charts in Figures 2.1 and 2.2 show the broad functional categories. Essentially, pathways belonging to a broad category are selected and the genes of these pathways combined together and the number of unique genes is expressed as a percentage of total unique genes in all pathways with *p*-values less than 0.01. As an example, there are three different pathways belong to cellular processes in C1: lysosome, peroxisome and phagosome. There are 61 genes enriched in lysosome, 46 genes enriched in peroxisome, and 63 genes enriched in phagosome. Together, they have 159 unique genes, which is about 14% of the total unique genes for all pathways with *p*-value less than 0.01. We performed these calculations for all six broad categories of pathways for two community-pairs of Tables S2.15 and S2.16 and the corresponding results are shown in Figures 2.1 and 2.2, respectively.

Table 2.10: Comparison of different methods with respect to three topological metrics, for the Human PPI network with 20,644 nodes and 241,008 edges.

		Combo	**F. Greedy**	**L. Eigen**	**Louvain**	**Spinglass**
RI	**Combo**	1	0.7314	0.3606	0.8805	0.8948
ARI	**Combo**	1	0.1806	0.0315	0.416	0.4998
NMI	**Combo**	1	0.3025	0.0936	0.4601	0.5551
RI	**F. Greedy**		1	0.444	0.7243	0.7258
ARI	**F. Greedy**		1	0.0624	0.1609	0.1739
NMI	**F. Greedy**		1	0.0787	0.2682	0.3063
RI	**L. Eigen**			1	0.3531	0.3649
ARI	**L. Eigen**			1	0.0191	0.0326
NMI	**L. Eigen**			1	0.0711	0.0951
RI	**Louvain**				1	0.8832
ARI	**Louvain**				1	0.4479
NMI	**Louvain**				1	0.4679
RI	**Spinglass**					1
ARI	**Spinglass**					1
NMI	**Spinglass**					1

When a method is compared with itself, RI, ARI and NMI are 1 (diagonal elements). Larger (smaller) the value of RI, ARI and NMI, the more (less) similar are the two methods being compared.

Table 2.11: Top 10 pathways for a comparison of KEGG pathway enrichment results between C1 with and S1 for the Human PPI network.

		C1 **(3,028)**				**S1 (2,921)**		
	Term (Pathway/function)	#	*p*-val	FE	Com	#	*p*-val	FE
M	Oxidative phosphorylation	102	2.04E-53	4.69	98	99	4.47E-49	4.48
M	Metabolic pathways	385	1.62E-48	1.92	350	370	1.13E-39	1.81
HD	Parkinson's disease	87	1.73E-33	3.75	85	86	5.17E-32	3.65
HD	Alzheimer's disease	84	5.23E-24	3.06	80	82	4.30E-22	2.94
HD	Huntington's disease	83	1.02E-18	2.65	81	84	7.34E-19	2.63
CP	Lysosome	61	9.08E-18	3.09	55	56	6.40E-14	2.79
GIP	SNARE interactions in vesicular transport	29	3.24E-17	5.22	29	31	6.64E-20	5.49
CP	Peroxisome	46	1.96E-15	3.39	46	49	1.07E-17	3.55
HD	Non-alcoholic fatty liver disease (NAFLD)	66	3.63E-15	2.68	65	65	3.30E-14	2.59
M	N-Glycan biosynthesis	32	1.58E-13	4	31	31	2.14E-12	3.81

The numbers inside parenthesis after C1 and S1 represent the number of genes that DAVID could annotate. The first column lists the broad category of pathways (M: Metabolism, HD: Human Diseases, CP: Cellular Processes, and GIP: Genetic Information Processing). FE: Fold Enrichment.

As seen in Table 2.11, there is a good overlap between enriched genes in C1 and S1 communities for different pathways. The first pathway is oxidative phosphorylation with the lowest *p*-value. This pathway has 102 genes enriched in C1 and 99 genes enriched in S1. Of these genes, there are 98 genes common between the two communities, which is a 96% overlap. Counting all genes for all pathways yields 778 unique genes for C1 and 756 unique genes for S1. Of these genes, there are 696 genes common between the two communities, which is an 89% overlap. The results of GO term enrichment analysis between these two communities are presented in Table S2.17, where a similarity of 91% is observed.

As seen in Figure 2.1, communities C1 and S1 represent six different broad categories of functions and they are similar to each other in terms of the percentage of enriched genes in each category.

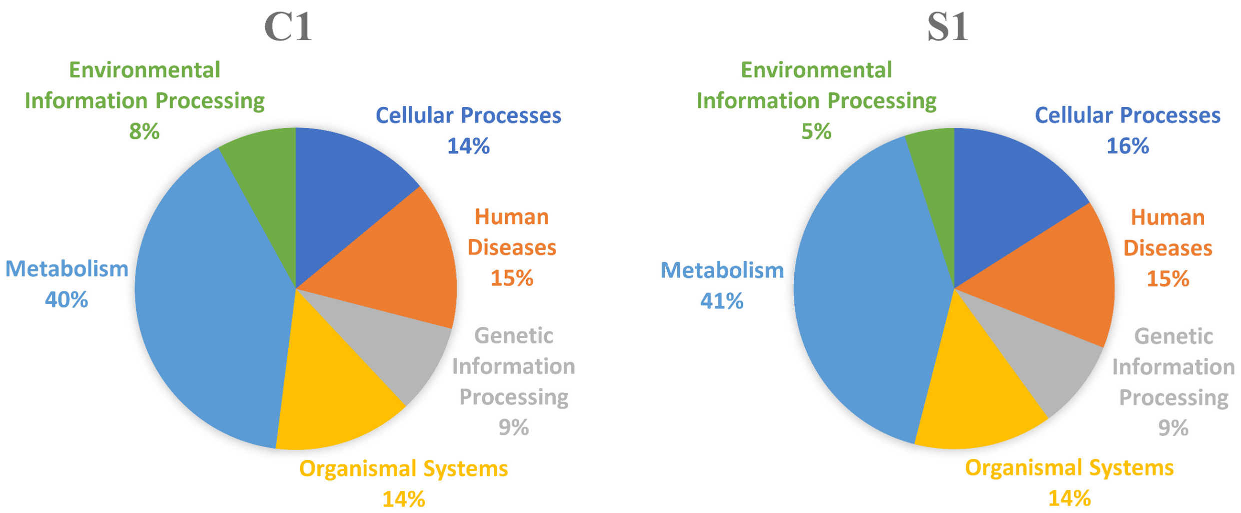


Figure 2.1: Pie charts for KEGG pathway enrichment results of C1 with 3,252 and S1 with 3,206 genes for the Human PPI network. Left chart shows the results for C1 and right chart shows the results for S1.

The results of comparing the top 10 pathways for L1 and S1 are summarized in Table 2.12 (Table S2.16 for the full list). Figure 2.2 shows the broad functional categories for comparing all pathways with *p*-values less than 0.01 for L1 and S1. Comparison of Figures 2.1 and 2.2 reveals that L1 and S1 are less similar as compared to C1 and S1. However, it is appropriate to say that Combo, Louvain, and Spinglass broadly yield similar and reasonably sized communities.

Table 2.12: Top 10 pathways for a comparison of KEGG pathway enrichment results between L1 and S1 for the Human PPI network.

		L1 (3,217)				**S1 (2,921)**		
	Term (Pathway/function)	#	*p*-val	FE	Com	#	*p*-val	FE
M	Oxidative phosphorylation	103	5.51E-48	4.06	97	99	4.47E-49	4.48
M	Metabolic pathways	398	1.08E-35	1.7	341	370	1.13E-39	1.81
HD	Parkinson's disease	91	8.68E-32	3.36	83	86	5.17E-32	3.65
HD	Alzheimer's disease	88	3.96E-22	2.75	80	82	4.30E-22	2.94
HD	Huntington's disease	87	8.64E-17	2.38	81	84	7.34E-19	2.63
GIP	SNARE interactions in vesicular transport	30	9.97E-17	4.63	30	31	6.64E-20	5.49
HD	Non-alcoholic fatty liver disease (NAFLD)	70	4.06E-14	2.43	64	65	3.30E-14	2.59
M	N-Glycan biosynthesis	34	1.69E-13	3.64	31	31	2.14E-12	3.81
CP	Phagosome	69	3.29E-13	2.37	61	66	1.80E-14	2.6
CP	Peroxisome	44	1.65E-11	2.78	40	49	1.07E-17	3.55

This table is arranged similar to Table 2.11.

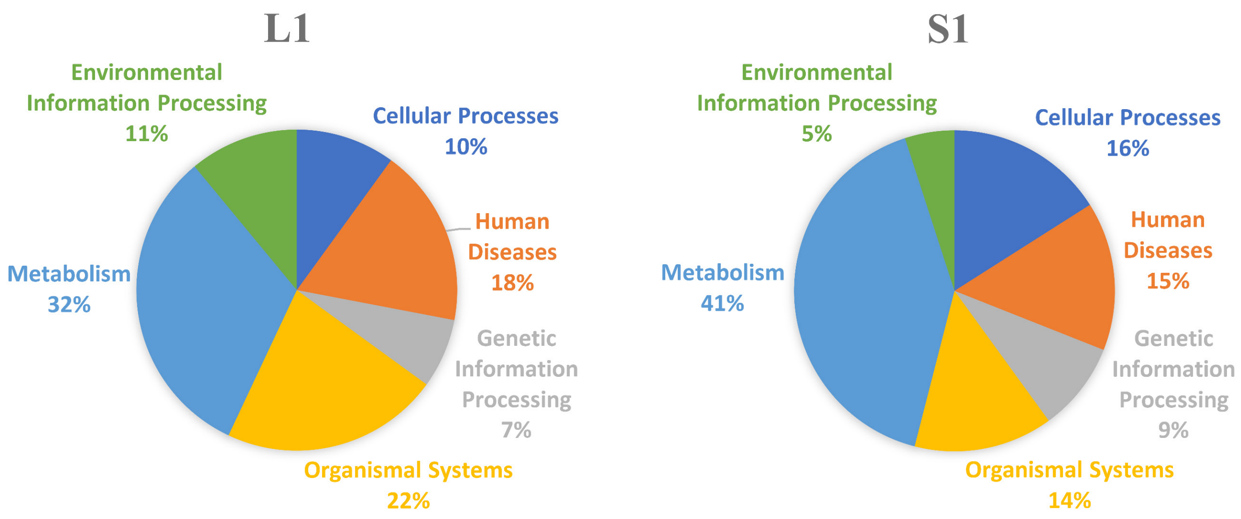


Figure 2.2: Pie charts for KEGG pathway enrichment results of L1 with 3,585 and S1 with 3,206 genes for the Human PPI network. Left chart shows the results for L1 and right chart shows the results for S1.

2.3.3 Orthology Comparison of Communities from Yeast and Human PPI Networks Using Louvain Method

In this sub-section, we will compare communities detected by Louvain for the Yeast PPI network and communities detected by the same method for the Human PPI network. Louvain could find 9 communities with sizes ranging from 4 to 1,538 for the Yeast PPI network (named SC1 for the biggest and SC9 for the smallest community) and 14 communities with sizes ranging from 3 to 3,585 for the Human PPI network. Using *biomaRt* package of R [44], we were able to find orthologous genes between Yeast and Human. Since the sizes of communities (the number of genes in the community) detected for the Human PPI network are larger than the size of communities detected for the Yeast PPI network, we found orthologous genes of the communities detected for the Human PPI network in Yeast (denoted HS ➔ SC) and then used DAVID to perform KEGG pathway enrichment for those genes. KEGG pathway enrichment results for the HS ➔ SC genes were compared to that for the communities of the Yeast PPI network. Table 2.13 shows the Jaccard index (as a percentage) between different pairs of communities and guided us on which community pairs should be compared with each other. For example, SC2 should be compared with HS3 ➔ SC. The results of comparing SC2 and HS3 ➔ SC are presented in Table 2.14. Tables for other comparisons of this sub-section are in the supplementary section (SC4 vs. HS1 ➔ SC in Table S2.18 and SC5 vs. HS2 ➔ SC in Table S2.19).

Table 2.13: Jaccard index (%) between the communities detected by Louvain for the Yeast PPI network and orthologous genes of the communities detected by the same method for the Human PPI network in Yeast.

	SC1 (1,538)	**SC2 (1,472)**	**SC3 (1,190)**	**SC4 (1,151)**	**SC5 (993)**
HS2 (3,511) → SC (702)	**12.11**	3.52	4.53	2.04	**18.36**
HS3 (2,660) → SC (621)	3	**20.84**	4.92	2.9	3.46
HS1 (3,585) → SC (531)	2.58	3.46	5.65	**19.97**	2.21
HS6 (1,561) → SC (408)	8.29	4.56	4.58	2.63	4.86
HS7 (1,308) → SC (367)	3.87	5.75	5.85	3.19	4.61
HS4 (2,547) → SC (335)	3.54	2.84	6.27	3.12	6.67
HS5 (1,639) → SC (244)	2.29	2.32	5.83	2.27	4.56

Community pairs with maximum overlap ≥ 10% (e.g., SC2 vs. HS3 → SC) are indicated in bold text.

Table 2.14: A comparison of KEGG pathway enrichment results between the second community detected by Louvain for the Yeast PPI network (SC2) and orthologous genes of the third community of the Human PPI network in Yeast (HS3 → SC).

			SC2 (1,378)				**HS3 (2,660)→SC (621)**	
	Term (Pathway/function)	#	*p*-val	FE	Com	#	*p*-val	FE
GIP	Ribosome	151	7.06E-65	3.25	104	112	6.72E-49	3.86
GIP	Spliceosome	74	9.60E-38	3.65	48	52	2.71E-23	4.1
GIP	RNA transport	82	2.37E-37	3.44	26	29	3.60E-04	1.94
GIP	Ribosome biogenesis in eukaryotes	75	2.86E-31	3.28	37	37	1.33E-08	2.59
GIP	RNA degradation	40	1.78E-10	2.56	14	17	2.60E-02	1.74

The first column lists the broad category of pathways (GIP: Genetic Information Processing). FE: Fold Enrichment.

As seen in Table 2.14, the most enriched pathway is the ribosome pathway with 151 genes enriched in SC2 and 112 genes enriched in HS3 ➔ SC. Of these, 104 genes are common between the two communities, which is a 69% overlap. Counting all genes for all pathways yields 380 unique genes for SC2 and 233 unique genes for HS3 ➔ SC. Of these genes, there are 218 genes common between the two communities, which is a 57% overlap. Although this similarity level is not impressive by itself, we did not expect much overlap between the two communities since Table 2.13 represents only a 21% similarity between them.

2.4 Discussion

As mentioned in the Results section, Louvain and Spinglass are most similar to each other for the Yeast PPI network (Table 2.2). Louvain tries to maximize the modularity (Q) whereas Spinglass tries to minimize the Hamiltonian ($\mathscr{H}$). However, it has been shown that there is a relation between Q and $\mathscr{H}$ as $Q = -\frac{\mathcal{H}}{2M}$ (Equation 2.16 in the Methods section). Thus, minimizing $\mathscr{H}$ is equivalent to maximizing Q. Still, since they use different algorithms to optimize their objective functions, the results are not exactly the same. Combo also tries to maximize modularity but in a different way than that in Louvain, thus resulting in slightly different communities as compared to those obtained by the Louvain method.

Table 2.2 suggests that Louvain and Spinglass are most similar to each other while Louvain and Leading Eigen are most dissimilar for the Yeast PPI network. Figure 2.3 illustrates the differences (as a percentage, i.e., 100*(#genes different between L1 and S1 (or L1 and LE4))/(max(L1,S1,LE4)) for each pathway) between the number of genes enriched in different pathways (with more than 10 genes and p-values less than 0.01) for L1 and S1 (blue columns), and for L1 and LE4 (orange columns). As seen from Figure 2.3, there is more difference between the number of genes enriched in L1 and LE4 compared to the difference between L1 and S1. This also verifies our results of topological comparison between L1 and S1, and L1 and LE4 (see also Table 2.2).

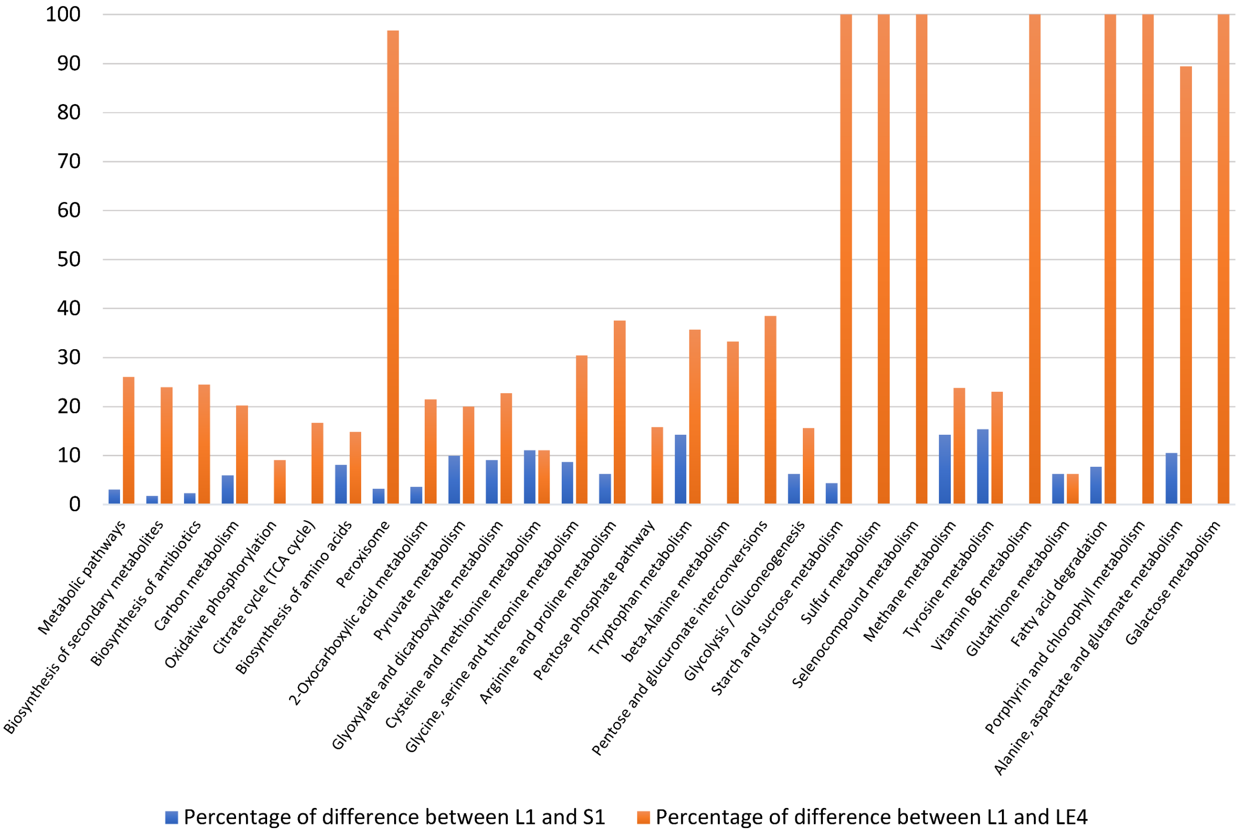


Figure 2.3: Comparing number of genes enriched in different pathways for the first community detected by Louvain (L1), the first community detected by Spinglass (S1), and the fourth community detected by Leading Eigen (LE4) for the Yeast PPI network.

KEGG pathway enrichment results for communities detected for the Yeast PPI network show that almost all pathways of each community belong to one broad function. For example, the first community of Louvain mostly includes pathways related to metabolic processes, the second community consists of pathways related to genetic information processing. On the other hand, the functions/pathways represented by communities detected for the Human PPI network are somewhat mixed and include several broad biological functions. Vis-a-vis the functional similarity of the methods, for the Human PPI network, Combo and Spinglass are similar, e.g., Figure 2.1 shows that in the first community of Combo (C1), 15% of the genes belong to pathways related to human diseases and the first community of Spinglass (S1) also has 15% of the genes related to the same broad category.

In order to confirm that the size of communities detected by different methods are reliable, we found sub-communities of the communities detected by Combo, Louvain, and Spinglass for the Yeast PPI network and analyzed the results to see if detected sub-communities of one community include pathways related to different biological functions or to the same one as the main community. As an example, most pathways of L1 belong to metabolic processes and the pathways for its sub-communities also belong to metabolic processes. Due to the result of this comparison and other comparisons for all communities and sub-communities detected by Combo, Louvain, and Spinglass, we can conclude that the size of communities detected are reliable.

We were also curious to know if all genes enriched in each pathway belong to one community or to different communities. To find this, we compared genes enriched in different pathways for all communities detected by Combo, Louvain, and Spinglass for the Yeast PPI network. The results of this comparison are shown in Tables 2.15, 2.16, and 2.17. The first column lists the broad category of pathways (M: Metabolism and GIP: Genetic Information Processing). The second column lists the different pathways enriched, columns 3 through 7 represent the number of enriched genes for each community, column 8 is the summation of all enriched genes, and the last column specifies the total number of genes in each pathway in the DAVID KEGG database. Some numbers in these tables are colored in gray, meaning their *p*-values are larger than the cut-off of 0.01. As seen in these tables, most enriched genes in each pathway belong to one community and the corresponding pathway is also significantly enriched. There are only a few exceptions such as metabolic pathways (which has a total of 685 genes and they are mainly distributed into two communities while still maintaining *p*-value less than 0.01 for the pathway (Tables 2.15-2.17, for the Yeast PPI network)), biosynthesis of amino acids (Table 2.16), and ribosome (Table 2.15). For the other pathways, most of the enriched genes belong to one community and if another community has some enriched genes, the *p*-value is greater than the cutoff of 0.01 (colored gray).

Table 2.15: Number of genes enriched in each pathway for different communities detected by Combo for the Yeast PPI network.

		C1	C2	C3	C4	C5		
	Term (Pathway/function)	Count	Count	Count	Count	Count	Sum	Pop Hits
M	Metabolic pathways	384	83	73	140		680	685
M	Biosynthesis of secondary metabolites	186	60	10	38		294	296
M	Biosynthesis of antibiotics	146	40	7	25		218	219
M	Carbon metabolism	89	18	5			112	114
M	Biosynthesis of amino acids	81	33	6	2		122	123
M	Glycolysis / Gluconeogenesis	34	19	3	2		58	58
GIP	Ribosome	3	145		2	24	174	181
GIP	RNA transport	5	79	8			92	93

The first column lists the broad category of pathways (M: Metabolism, and GIP: Genetic Information Processing). Column 2 lists the different pathways enriched. Columns 3 through 7 represent number of enriched genes for different communities. Column 8 lists the total of all enriched genes of all communities together and the last column represents the maximum number of genes in that pathway. For example, in DAVID database, "metabolic pathways" contains 685 genes and of these, 384 were found in C1 and 140 were found in C4.

Table 2.16: Number of genes enriched in each pathway for different communities detected by Louvain for the Yeast PPI network.

		L1	L2	L3	L4	L5		
	Term (Pathway/function)	Count	Count	Count	Count	Count	Sum	Pop Hits
M	Metabolic pathways	346	93	34	71	138	682	685
M	Biosynthesis of secondary metabolites	164	63	18	10	40	295	296
M	Biosynthesis of antibiotics	128	46	14	6	25	219	219
M	Carbon metabolism	79	19	9	5	2	114	114
M	Biosynthesis of amino acids	68	40	7	6	2	123	123
M	Glycolysis / Gluconeogenesis	30	19	5	3		57	58
GIP	Ribosome	21	151		2		174	181
GIP	RNA transport		82	4	5		91	93

This table is arranged similar to Table 2.15.

Table 2.17: Number of genes enriched in each pathway for different communities detected by Spinglass for the Yeast PPI network.

		S1	S2	S3	S4	S5		
	Term (Pathway/function)	Count	Count	Count	Count	Count	Sum	Pop Hits
M	Metabolic pathways	357	85	30	69	140	681	685
M	Biosynthesis of secondary metabolites	167	61	17	8	41	294	296
M	Biosynthesis of antibiotics	131	43	14	5	25	218	219
M	Carbon metabolism	84	18	6	4	2	114	114
M	Biosynthesis of amino acids	74	35	6	5	2	122	123
M	Glycolysis / Gluconeogenesis	32	17	5	3		57	58
GIP	Ribosome	32	141				173	181
GIP	RNA transport	5	79	2	6		92	93

This table is arranged similar to Table 2.15.

We note that the analysis provided above does not fully address the issue of selecting the best method. It is likely to be subjective and network specific. Hence, we recommend applying a few different methods, such as Louvain (from the group of similar methods) and one or two other (dissimilar) methods, and compare and interpret the results to obtain a consensus best method.

2.4.1 Robustness of Communities Obtained by the Louvain Method

We also analyzed the robustness of communities obtained by Louvain method for small perturbations in the network. Essentially, the network is randomly perturbed by deleting some nodes (and edges involving them) and the communities are identified. This is carried out 100 times to assess the robustness of the communities. First, the communities of the original Yeast PPI network are identified using Louvain method. Then, the following steps are repeated 100 times:

1. Remove 1% of nodes randomly (e.g., 65 nodes out of 6532 nodes for the Yeast PPI network).

2. Find the communities of the new network using Louvain method.

3. From the communities of the original network, delete the random nodes of step 1.

4. Calculate the Jaccard index matrix between the communities obtained in steps 2 and 3. We considered all communities with more than 100 nodes.

5. Compute the maximum value for each column of the Jaccard index matrix.

6. Compute the average of the resulting row-vector.

After performing the above steps 100 times, we have the vector of average of column-wise-maximum (avg-max) of Jaccard index values. Then, we generate scatter plots of avg-max. The scatter plots of avg-max for the Yeast and the Human PPI networks are shown in Figure 2.4 (left panel: Yeast, right panel: Human).

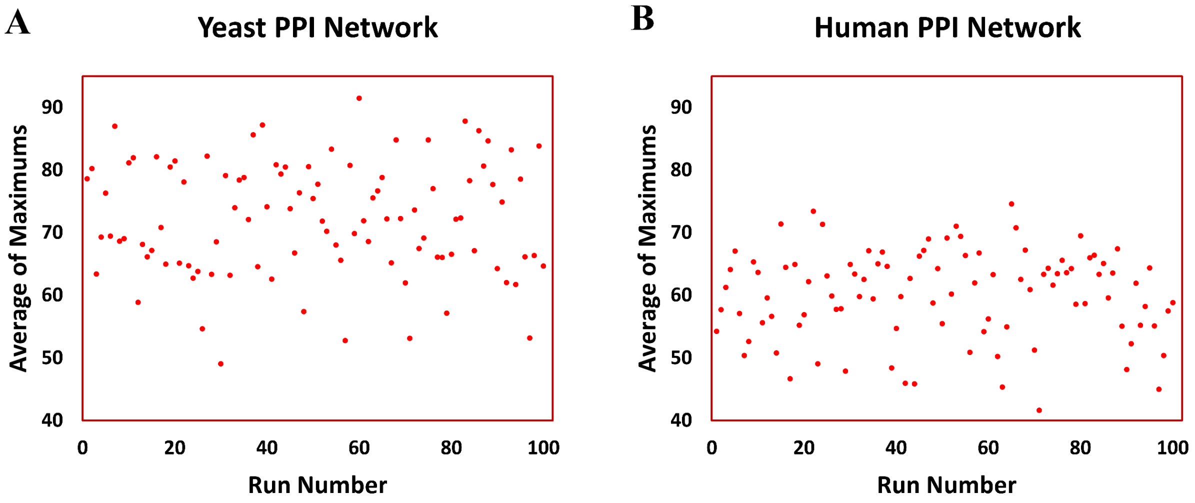


Figure 2.4: Scatter plots of avg-max (average of column-wise-maximum of Jaccard index matrix) values vs. run number. Panel **A** shows the results for the Yeast PPI network and panel **B** shows the results for the Human PPI network.

The mean and standard deviation of the avg-max vector are 72.21 and 9.02%, respectively for the Yeast PPI network, whereas, for the Human PPI network, they are 60.11 and 7.14%, respectively. These relatively large numbers for the mean suggest that the communities identified by Louvain method are robust to small perturbations. We repeated the process for Leading Eigen.

The mean and standard deviation of the avg-max vector for the Yeast network is 53.4 and 5.63%, respectively. Thus, Louvain is a better method, at least for the Yeast PPI network; for the Human PPI network, Leading Eigen finds only two or three very large communities, making avg-max artificially high (98%).

2.4.2 Generality of the Overall Results

The PPI networks that we used in our analysis were from BioGRID and included both physical and genetic interactions (combined network). Hence, we have also applied the various approaches for finding communities to a Yeast PPI network comprising of only physical interactions, which has 6,298 nodes and 83,788 edges, to find out if our broad conclusions based on the combined network are still valid for the physical interaction-only network. Table S2.20 shows the modularity scores and the number of communities detected by different methods for the physical interaction-only network. While Q is smaller for the combined network as compared to that for the physical interaction-only network for all methods, their relative trend for the different methods remain almost the same. Interestingly, the Q values for Louvain, Combo, and Spinglass are similar and among the largest. Thus, these three methods are one of the best methods in terms of Q value as well.

We have compared the various methods using three topological metrics (namely Rand Index, Adjusted Rand Index, and Normalized Mutual Information) with respect to the physical interaction-only network. The results for these comparisons are given in Table S2.21. Based on results of Table S2.21, Combo, Louvain, and Spinglass are similar to each other in terms of the topological metrics. We also compared the KEGG pathway enrichment results for the first 5 communities of Louvain and Spinglass. To find which communities from Louvain and Spinglass methods are similar, we generated the Jaccard index matrix for communities with more than 100 nodes for both methods (Table S2.22). After selecting pairs of communities with highest value of Jaccard index for each column, we used DAVID version 6.8, to perform KEGG pathway enrichment analysis. The results of those comparisons are presented in Tables S2.23-S2.25 for the L1 vs. S1, L2 vs. S2, and L3 vs. S4 comparisons, respectively. These tables are arranged similar to Table S2.4.

Two main results from Tables S2.23-S2.25 are: 1) good functional similarity between the communities from Louvain and Spinglass methods (e.g., an overlap of 73.06% between L1 and S1, 79.41% between L2 and S2, and 85.24% between L3 and S4) and 2) segregation of biological functions represented by the communities, e.g., communities L2, L3, and L4 represent mostly metabolism related pathways. L1 shows a mixed enrichment, akin to the mixed pathways represented by the L3 and L5 communities of the combined network (Tables S2.4 and S2.6). Some differences in the nature of the broad results for the combined network vs. the physical interaction-only network are likely due to the fact that the physical interaction-only network is much sparser (just 1/3rd of the edges are retained) as compared to the combined network. Using Cytoscape [45], we also analyzed the properties (Table S2.26) and node-degree distribution for the combined network and the physical interaction-only network (Figures 2.5A and 2.5B). Figure 2.5C shows a comparison of count of nodes for a given degree between the combined and the physical interaction-only networks. As can be seen from Figure 2.5C, good R^2-value suggests good agreement between the two degree distributions.

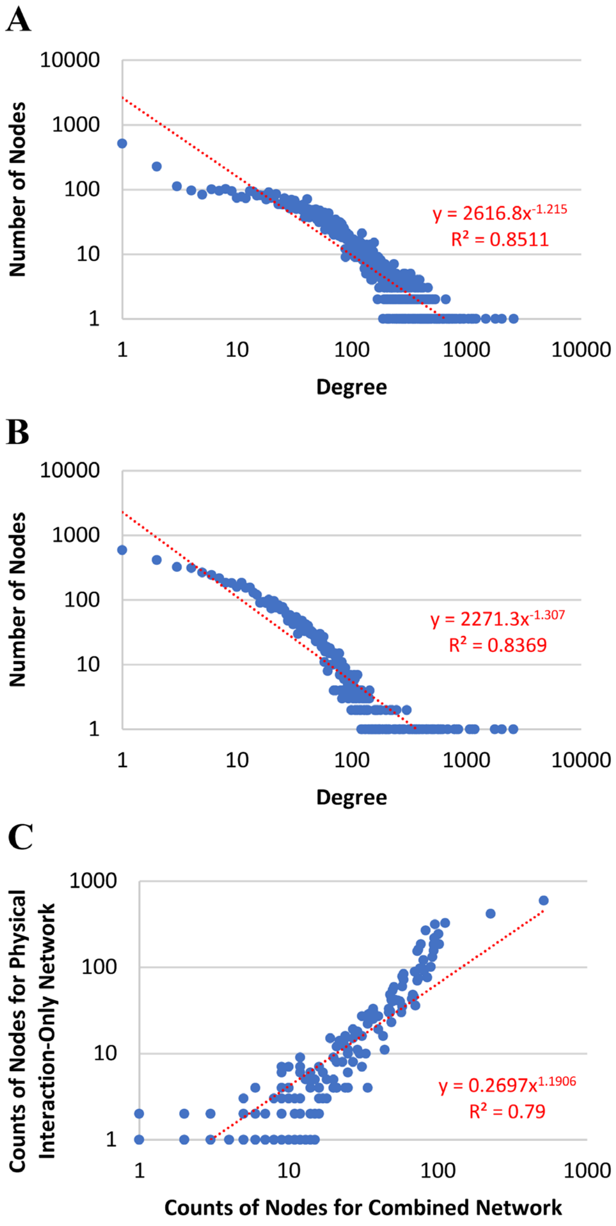


Figure 2.5: **A** Node-degree distribution for the combined network. **B** Node-degree distribution for the physical interaction-only network. **C** Comparison of counts of nodes between the combined network and the physical interaction-only network.

2.4.3 Optimization of Method-specific Parameters

We wanted to find out if the *Q* value for the different methods could be improved by optimizing their parameters, if any, or if their results varied across different runs. Of six methods, three of them, namely, Combo, Conclude, and Spinglass used a random number generator in the procedure for finding communities, although they do not have any parameters to be optimized. We carried out 10 repeats of Combo, Conclude, and Spinglass on the combined Yeast PPI network to assess the variation in *Q* across the runs. Table S2.27 shows the results of 10 runs for these methods (similar to those for the Combo and Spinglass methods for the Human PPI network in Tables 2.8 and 2.9, respectively). The standard deviations of *Q* across the 10 runs are much larger for Conclude (5.11%, std./mean) as compared to those for Combo (0.05%) and Spinglass (0.36%). Thus, for a small number of allowed runs, the results from Combo and Spinglass are more reliable.

We selected the run with the highest modularity for Conclude method and performed KEGG pathway enrichment analysis. Then we compared these enrichment results with the enrichment results for the communities from the Conclude method reported earlier. Although the size of communities is a little different, the broad categories of pathways for majority of the communities are still the same. For example, there is one community in both runs which has pathways mainly related to metabolic processes. Table S2.28 shows the results for comparison between the two communities with pathways related to metabolic processes (CL3) in the two runs.

Spinglass method uses simulated annealing in its procedure. Hence, we ran it 10 times with 10 different start and stop temperatures and cooling factor, but the default values in the *igraph* package in R yielded the best result in terms of the largest modularity. Overall, at a broad level, we found that in our case studies, the methods providing well-interpretable communities also resulted in near-optimal (largest) *Q* values. The fact that the *Q* value for Conclude in one of the runs is slightly higher (0.29) than those for Combo and Spinglass (0.265) does not violate this conclusion because some of the communities from Conclude are also well-interpretable. However, we note that the Q values across different runs vary substantially for Conclude as compared to those for Combo and Spinglass, suggesting that Combo and Spinglass are likely more robust than Conclude.

2.5 Conclusion

In this paper, we tested six methods to detect communities within the Yeast and Human PPI networks. An in-depth comparison of communities detected by these different methods has led us to conclude that Louvain and Spinglass are most similar for the Yeast PPI network whereas Combo and Spinglass are most similar for the Human PPI network. In terms of finding communities that include core pathways based on KEGG pathway and GO term enrichment results, Combo, Louvain, and Spinglass were able to find similar communities in which important biological functions and pathways were enriched. For the Yeast PPI network, all genes from the network belonging to a pathway also generally belong to only one or two communities. In terms of running time or computational complexity, for the Yeast PPI network, Louvain was the fastest method and Combo was faster than Spinglass. For the Human PPI network, Louvain was much faster than Spinglass, and Spinglass was faster than Combo. Overall, Combo, Louvain, and Spinglass provide reasonable results for community detection in biological networks. Their corresponding modularity values were also among the highest, except that Conclude yielded slightly better modularity for the Yeast PPI network in some runs; variation in modularity for Conclude was much larger than that for Combo and Spinglass across multiple runs. Overall, since Combo and Spinglass use stochastic search in their procedure and their running time is also more than that for Louvain, Louvain is likely the best method to find reasonably sized communities for biological networks in a reasonable time. While we applied these methods to PPI networks, we expect the broad results to be applicable to other types of networks such as gene-coexpression networks and hybrid/integrated networks.

2.6 Methods

The focus of this work is on undirected, unweighted, and connected graphs defined as $G(n,m)$ where $n = \{n_1, n_2, \ldots, n_{|n|}\}$ is an ordered set of nodes and m, a subset of $n*n$, is the set of edges; for convenience, here, n and m represent sets. In our case studies on the Yeast and Human PPI networks, each node represents an Entrez gene ID and each edge represents an interaction between two nodes. By finding communities, we imply the segregation of nodes into groups such that there is a higher density of edges within each group than between groups. Although there are many algorithms for detecting communities, we selected six algorithms that performed well in terms of finding core pathways in biological networks with several thousands of nodes and hundreds of thousands of edges within reasonable time (e.g., one day on a 24-processor machine).

Each of these algorithms tries to optimize an objective function which in most cases is modularity. However, two algorithms with the same objective function will generally yield slightly different communities using different amounts of run time if they use different optimization strategies. The algorithms that will be briefly described below are: Girvan-Newman [15, 17], Fast Greedy [18], Combo [46], Louvain [20], Conclude [47], Infomap [48], Leading Eigen [21], and Spinglass [29]. We used an R package named *igraph* [49] to run Fast Greedy, Louvain, Leading Eigen, and Spinglass. There is a Cytoscape plugin, named GLay, to visualize the network and the communities [50]. The GLay plugin utilizes *igraph* C library which provides the same methods as the *igraph* R package [49]. For the other two methods (Conclude and Combo), authors have made java [51] and C++ [52] codes of their algorithms available online. Girvan-Newman method did not provide any result for our networks in 24 hours on a 24-processor machine, so we did not consider its results in our comparison. Another method, called Infomap, did not find reasonably-sized communities (e.g., for the Yeast PPI network with 6,532 nodes, one community has 6,195 nodes, and the rest have less than 10 nodes). Due to these reasons, we did not consider Girvan-Newman and Infomap methods in our analysis. However, Fast Greedy and Louvain, which we have used in our comparison, are based on the original Girvan-Newman method. So, we will briefly describe Girvan-Newman algorithm in the next subsection.

2.6.1 Algorithms for Community Detection

Girvan-Newman

This is a divisive algorithm for identifying communities in networks where edges are iteratively removed based on the value of their betweenness. The steps of the algorithm are as follows [15, 17]:

1. Calculate betweenness score for all edges in the network. Betweenness is a measure that favors edges that lies between communities and disfavors those that lie inside communities. While there are many measures to find betweenness, Girvan-Newman algorithm uses the fast algorithm of Newman to find betweenness scores [53].

2. Find the edge that has the maximum value of betweenness and remove it from the network.

3. Recalculate betweenness for all remaining edges in the network and repeat from step 2 for the new scores until there are no edges remaining.

The output of this algorithm is a dendrogram that captures the possible divisions of the network into communities. In order to select the optimal division among all possible options, Modularity (Q) is used.

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{2.1}$$

where A_{ij} is the weight of the edge between node i and j (is equal to 1 when all edges have the same weight), k_i is the sum of the weights of the edges attached to node i or degree of node i, c_i is the community to which node i belongs to and the δ function is defined as $\delta(u,v) = 1$ if $u = v$ and 0 otherwise. Local peaks in the modularity during the process of community detection indicate good divisions of the network into communities. While using this algorithm gives good results in many cases, its computational complexity is $O(m^2n)$ or $O(n^3)$ on a sparse graph, where n is the number of nodes and m is the number of edges. This makes it impractical for very large graphs with several thousands of nodes and hundreds of thousands of edges.

Fast greedy by Clauset, Newman and Moore

This algorithm (namely, Fast Greedy) involves finding the changes in modularity that results from combining pair of communities, selecting the combination yielding largest gain in modularity, and implementing the combination of the corresponding pair [18]. At the beginning, each node is considered as a community. One way of performing the combination process is to consider the network as a multigraph where a whole community is represented by a node and the elements of the adjacency matrix are equal to the number of edges between the communities. Joining two communities (namely i and j) corresponds to replacing the i^{th} and j^{th} rows and columns of the adjacency matrix by a single row and column formed by their sum, respectively; the record of the list of nodes in the communities formed thus far is updated. In the algorithm proposed by Newman [31], this operation is carried out explicitly on the entire adjacency matrix. Calculating the change of modularity (ΔQ) and finding the pair i, j with the largest gain is time-consuming. Hence, here, instead of using the adjacency matrix and calculating ΔQ_{ij}, a matrix of ΔQ_{ij} values is initialized and updated directly. For sparse matrices, e.g., adjacency matrices for large networks, this results in substantial reduction in computation. The following parameters have to be set initially:

$$\Delta Q_{ij} = \begin{cases} {}^{1}\!/_{2m} - {}^{k_i k_j}\!/_{(2m)^2} & \text{if } i, j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

$$a_i = \frac{k_i}{2m} \tag{2.3}$$

for each *i*. The next steps are as follows:

1. Calculate the initial values of ΔQ_{ij} and a_i based on Equations 2.2 and 2.3, and form the max-heap, *H*, which contains the largest element of each row of the matrix ΔQ_{ij} along with the labels i, j of the corresponding pair of communities.

2. Select the largest ΔQ_{ij} from *H*, combine the corresponding communities, update the matrix ΔQ, *H*, and a_i (Equation 2.4) and increase *Q* by ΔQ_{ij}.

3. Repeat step 2 until only one community exists.

Due to the sparsity of the original adjacency matrix *A*, we will be able to perform updates in step 2 quickly and we need to only adjust a few elements of ΔQ. If communities *i* and *j* are joined together, labeling the combined community as *j*, we update the j^{th} row and column, and delete the i^{th} row and column altogether. The update rules are as follows:

If community *k* is connected to both *i* and *j*, then:

$$\Delta Q'_{jk} = \Delta Q_{ik} + \Delta Q_{jk} \tag{2.4a}$$

If k is connected to *i* but not to *j*, then:

$$\Delta Q'_{jk} = \Delta Q_{ik} - 2a_j a_k \tag{2.4b}$$

If k is connected to j but not to i, then:

$$\Delta Q'_{jk} = \Delta Q_{jk} - 2a_i a_k \tag{2.4c}$$

To update a: $a_j' = a_j + a_i$

Fast Greedy algorithm runs in time $O(m.d.logn)$ for a network with n nodes and m edges where $d \sim logn$ is the depth of the dendrogram. For sparse networks ($m \sim n$), the running time is $O(nlog^2n)$, which is essentially linear [18].

Combo

Most community detection algorithms use one of the following steps in the process of finding communities: they may join two communities, split a community into two, or move nodes between two distinct communities. Combo involves all three possibilities [46]. After selecting an initial partition made of a single community, the following steps are iterated until there is no gain in the objective function which may be modularity (Equation 2.1) or description code length:

1. For each source community, the best possible repartition of every source node into each destination community (either an existing community or a new community) is calculated. It would be possible that the source community totally joined the destination community in this step.

2. The best merger, split, or recombination is performed.

The basis of Combo is the selection of the best repartition of nodes between two communities. For each pair of sources and (maybe empty) destination communities, a shift of all the nodes using Kernighan-Lin algorithm [54] is performed. Particularly, Combo recombines the two communities starting from several initial configurations including: a) the original communities, b) the case in which the whole source community is moved to the destination community and c) a few intermediate mergers, where a random subset of the source community is shifted to the destination community.

For each starting configuration, a series of Kernighan-Lin shifts [54] is iterated until no further improvement is possible. Each configuration is carried out by initializing a list of available nodes to cover all the nodes from the original source community and then iterating the following steps until there are no more nodes in the list:

1. Find the node i in the list which when switched to another community results in the largest gain in modularity.

2. Switch i to the other community, remove it from the original list and save the intermediate result.

After performing a series of Kernighan-Lin iterations for each of the starting configurations, the intermediate result with the best score in terms of objective function (modularity) is selected.

Combo outperforms all other known algorithms when the objective function is modularity. However, it has limitations on the size of the network it could handle within a reasonable time, which is currently around 30,000 nodes and is not a serious limitation for most biological networks. When the objective function is description code length, Combo's results are similar to Infomap (which will be described later in this section) in most cases. Since the sequence of operations depends on the specific network, obtaining exact evaluations of the computational complexity of Combo is difficult, but the upper bound is $O(n^2 logc)$ where n is the number of nodes and c is the number of communities in the network [46].

Louvain

This algorithm detects communities in large networks by maximizing modularity and is much faster as compared to other methods [20, 55, 56]. The limiting factor for this method is the memory (RAM) requirement rather than the computation time, as is the case with Girvan-Newman and Spinglass algorithms. The algorithm is divided in two phases, which are repeated iteratively. First phase is to assign a different community to each node of the network. So, in the beginning, there are as many communities as there are nodes. Then, the gain of modularity (Equation 2.5) is calculated for removing node i from its community and placing it in one of its neighboring communities. The gain of modularity in moving node *i* into a community *C* can be computed by:

$$\Delta Q = \left[\frac{\Sigma_{in} + k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{2.5}$$

where Σ_{in} is the sum of the weights (or count for unweighted networks) of the edges inside *C*, Σ_{tot} is the sum of the weights of the edges incident to nodes in *C*, k_i is the sum of the weights of the edges incident to nodes *i* (degree of *i*), $k_{i,in}$ is the sum of the weights of the edges from *i* to nodes in *C* and *m* is the sum of the weights of all the edges in the network. If the gain is positive, the node *i* is placed in the community for which the gain is maximum. This process is applied repeatedly for all nodes until no further improvement can be achieved.

The second phase is to build a network whose nodes are now the communities detected during the first phase. In order to perform that, the weights of the edges between the new nodes are given by the sum of the edges between nodes in the corresponding two communities. Edges between nodes of the same community result in self-loops for this community in the new network. When this phase is completed, the first phase of the algorithm is reapplied to the new network. The combination of these two phases is referred to as a “pass”. The passes are iterated until a maximum of modularity is reached. This algorithm is extremely fast (*O*(*nlogn*)) and could be even faster by using some heuristics, e.g., stopping the first phase when the gain of modularity is less than a given threshold [20].

COmplex Network CLUster DEtection (CONCLUDE)

CONCLUDE is a fast community detection method. The algorithm takes a graph *G*, an integer κ, and an integer φ as inputs. The steps are:

1. Compute κ-path edge centralities using Edge Random Walk κ-path Centrality (ERW-Kpath) algorithm (described by De Meo, et al. [47]) on nodes of *G* by carrying out at most φ iteration. The output of ERW-Kpath algorithm is an array of weights.

2. Calculate the distance among all pairs of nodes by taking two inputs: graph *G* and the array of weights from the previous step. It uses the following equation (Equation 2.6) to find pairwise distances:

$$\sigma_{ij} = \sqrt{\frac{\sum_{k\epsilon N(i)-CN(i,j)}[L^{\kappa}(e_{ki})]^2}{|N(i)-CN(i,j)|}+\frac{\sum_{k\epsilon N(j)-CN(i,j)}\left[L^{\kappa}(e_{kj})\right]^2}{|N(j)-CN(i,j)|}+\frac{\sum_{k\epsilon CN(i,j)}\left[L^{\kappa}(e_{ki})-L^{\kappa}(e_{kj})\right]^2}{|CN(i,j)|}} \quad (2.6)$$

where L^{κ} is κ-path edge centrality and is calculated by:

$$L^k(e_{ij}) = \sum_{s\epsilon V} Pr(e,s) \quad (2.7)$$

Pr(*e*,*s*) is the probability of selecting the edge *e* in a random simple κ-path originating from an arbitrary source node *s*. The symbol *N*(*i*) is the set of neighbors of the node *i* and *CN*(*i*,*j*) indicates the subset of neighbors common to *i* and *j*.

The output of this step is a matrix Δ containing all pairs of distances between nodes.

3. Finally, apply Louvain method [20] on matrix Δ to find communities of *G*.

Maps of random walk (Infomap)

The Infomap approach closely follows the Louvain approach [20]; neighboring nodes are joined together to make small communities which subsequently are joined into bigger communities.

The difference between these two methods is that the objective function of Louvain is modularity while the objective function of Infomap is a lower bound on a quantity referred to as code-length (*M*), defined as:

$$L(M) = q_{\curvearrowright} H(Q) - \sum_{i=1}^{c} p^{i}{}_{\circlearrowright} H(P^{i}) \tag{2.8}$$

The aim of Infomap is to minimize the lower bound, *L*(*M*). The equation comprises of two terms: first is the entropy of the movement of nodes between communities and second is the entropy of movements of nodes within communities. Further details about this equation can be found elsewhere [48].

Each node is initially assigned to its own community. Then, in random sequential order, each node is placed into the neighboring community that results in the largest decrease in *L*(*M*) (Equation 2.8). If no move decreases *L*(*M*), the node will stay in its original community. This procedure is repeated in a new random sequential order each time until no move could decrease *L*(*M*). In each iteration, the nodes of the new network are the communities found at the last level and the process of joining nodes into communities is repeated on the new network until *L*(*M*) cannot be reduced further. The computational complexity of Infomap is a linear function of the number of edges, i.e., *O*(*m*) [48, 57].

Leading Eigen

Leading Eigen method [21] is also based on the modularity maximization but here, the modularity is expressed in terms of the eigenvalues and eigenvectors of a matrix called the modularity matrix, *B*:

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \tag{2.9}$$

where, A is the adjacency matrix, k_i is the sum of the weights of the edges attached to node i (or degree of node i), k_j is the degree of node j, and m is the total number of edges. The modularity matrix is a characteristic property of the network and is independent of any division of the network into communities. The procedure of finding communities of a network with this method consists of finding the eigenvector corresponding to the most positive eigenvalue of the modularity matrix, and then dividing the network into two groups based on the sign of the elements of the eigenvector. Defining an index vector, s, the signs of elements are:

$$s_i = \begin{cases} +1 & if\ u_i{}^{(1)} \geq 0 \\ -1 & if\ u_i{}^{(1)} < 0 \end{cases} \tag{2.10}$$

where, $u_i^{(1)}$ is the i^{th} element of u_1 (the normalized eigenvector of the modularity matrix). The nodes with positive sign form one community and the rest of the nodes form the other community. To avoid dividing the network into only two communities, an n (total number of nodes) by c (the number of non-overlapping communities) index matrix S has to be defined. Each column of this matrix is an index vector of (0,1) elements, such that:

$$S_{ij} = \begin{cases} 1 & if\ vertex\ i\ belongs\ to\ community\ j \\ 0 & otherwise \end{cases} \tag{2.11}$$

The modularity of this division of the network is then equal to:

$$Q = \sum_{i,j=1}^{n} \sum_{k=1}^{c} B_{ij} S_{ik} S_{jk} = Tr(S^T B S) \tag{2.12}$$

This form of modularity is different from other forms in a leading multiplicative constant 1/(2m) but since it has no effect on the position of the maximum of the modularity, it has been omitted from the equation. Writing $B=UDU^T$ where $U=(u_1|u_2|\ \ldots)$ is the matrix of eigenvectors of B, and D is the diagonal matrix of eigenvalues $D_{ii}=\beta_i$, Q can be written as:

$$Q=\sum_{i,j=1}^{n}\sum_{k=1}^{c}\beta_j(u_j^T s_k)^2 \tag{2.13}$$

The aim is still maximizing the modularity Q, but now, there is no constraint on the number of communities, c [21].

Spinglass

In this method, the community structure of a network is described as the spin configuration that minimizes the energy of the spin glass (Hamiltonian) with respect to the spin states (the community indices) [29]. Similar to any other quality function for an assignment of nodes into communities, Hamiltonian has to follow the principle of grouping together the nodes that are linked (there is an edge between them) and keep apart the ones that are not linked. From this, four requirements have to be satisfied: a) rewarding internal edges between nodes of the same community (in the same spin state), b) penalizing missing edges (non-links) between nodes in the same community, c) penalizing existing edges between different communities, and d) rewarding non-links between different communities. The following equation (Equation 2.14) satisfies these properties:

$$\begin{aligned}\mathcal{H}(\{\sigma\}) = &-\sum_{i\neq j} a_{ij}\,\underbrace{A_{ij}\delta(\sigma_i,\sigma_j)}_{internal\ links} + \sum_{i\neq j} b_{ij}\,\underbrace{(1-A_{ij})\delta(\sigma_i,\sigma_j)}_{internal\ non-links}\\ &+\sum_{i\neq j} c_{ij}\,\underbrace{A_{ij}(1-\delta(\sigma_i,\sigma_j))}_{external\ links} - \sum_{i\neq j} d_{ij}\,\underbrace{(1-A_{ij})(1-\delta(\sigma_i,\sigma_j))}_{external\ non-links}\end{aligned} \tag{2.14}$$

where, σ_i denotes the community index of node *i* in the graph, δ is the Kronecker delta function and a_{ij}, b_{ij}, c_{ij}, and d_{ij} represent the weights of the individual contributions, respectively. If links and non-links are each weighted equally, no matter they are external or internal, $a_{ij} = c_{ij}$ and $b_{ij} = d_{ij}$, then it would be enough to consider the internal links and non-links. Convenient choices of coefficients are $a_{ij} = 1 - \gamma p_{ij}$ and $b_{ij} = \gamma p_{ij}$ where p_{ij} denotes the probability that a link exists between node *i* and *j*, normalized, such that $\sum_{i \neq j} p_{ij} = 2m$, where *m* is the total number of edges in the network. When $\gamma = 1$, it leads to the natural situation in which the total amount of energy that can possibly be contributed by links and non-links is equal, i.e., $\sum_{i \neq j} A_{ij} a_{ij} = \sum_{i \neq j} (1\text{-}A_{ij}) b_{ij}$. As we are dealing with undirected, unweighted networks, our choice of weights allows us to simplify the Hamiltonian (Equation 2.14):

$$\mathcal{H}(\{\sigma\}) = -\sum_{i \neq j} (A_{ij} - \gamma p_{ij}) \delta(\sigma_i, \sigma_j) \tag{2.15}$$

where, p_{ij}, the probability, can be written as $p_{ij} = \frac{k_i k_j}{2m}$ and k_i and k_j represent degree of node *i* and degree of node *j,* respectively. Now, minimizing $\mathcal{H}$ gives the number of spin states (or communities) in a network. The minimization is carried out by using simulated annealing on the entire network. This method is rather fast and the computational complexity is approximately $O(n^{3.2})$. However, it cannot be used for disconnected networks as there is no guarantee that nodes from disconnected parts of the network also have different spin states and belong to different communities.

Substituting p_{ij} as $\frac{k_i k_j}{2m}$ and $\gamma = 1$ in Equation 2.15, we have:

$$\mathcal{H}(\{\sigma\}) = -\sum_{i \neq j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(\sigma_i, \sigma_j)$$

and comparing this equation with Equation 2.1 (modularity) yields:

$$Q = -\frac{\mathcal{H}(\{\sigma\})}{2m} \tag{2.16}$$

It is clear from Equation 2.16 that minimizing Hamiltonian is equivalent to maximizing modularity. Thus, we expect to get same results for Louvain (which maximizes modularity) and Spinglass (which minimizes Hamiltonian) when applied to our networks.

Table 2.18 summarizes the above methods described in this section.

Table 2.18: Summary of community detection methods.

Name of method	**Equation #**	**Complexity**	**Reference**
Girvan Newman	1	$O(m^2n)$	[15, 17]
Fast Greedy	2, 3, 4	$O(n\log^2 n)$	[18]
Combo	1	$O(n^2\log c)$	[46]
Louvain	1, 5	$O(n\log n)$	[20]
Conclude	6, 7	$O(m)$	[47]
Infomap	8	$O(m)$	[48]
Leading Eigen	9, 10, 12	$O(n^2)$	[21]
Spinglass	15	$O(n^{3.2})$	[29]

2.6.2 Metrics for Comparison of Different Algorithms

To compare different methods, we used three metrics, namely, Rand Index (RI), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI). We also used Jaccard Index for measuring similarity between different communities. These metrics are based on topological similarities of the communities identified and hence are relevant for biological networks; we expect that topologically similar communities will likely yield similar biological interpretations.

Rand index (RI)

Rand proposes a simple measure of agreement between the results of (i.e., the communities identified by) two methods *A* and *B* [58]. RI represents the fraction of node-pairs that are distributed to the communities obtained by the two methods in a similar manner. Let n_{11} be the number of pairs of nodes from a network *G* which are both in the same community detected by method *A* and are also in the same community detected by method *B*. Let n_{00} be the number of pairs of nodes from *G* which are in different communities in *A* and are also in different communities in *B*. n_{00} and n_{11} are interpreted as agreements in the classification of the nodes from a pair. Accordingly, two disagreement quantities n_{01} and n_{10} are also defined: n_{01} (n_{10}) is the number of pairs of nodes from *G* which are in the same community detected by method *A* (*B*) but they are in different communities detected by method *B* (*A*). Then, Rand Index (RI) is given by [59]:

$$RI(A,B) = \frac{n_{00} + n_{11}}{n_{00} + n_{11} + n_{01} + n_{10}} \tag{2.17}$$

As seen from Equation 2.17, RI has a probabilistic interpretation with respect to picking a pair of nodes at random, i.e., $\frac{n_{00}+n_{11}}{\binom{N}{2}}$, which is a probability of agreement (*N* is the total number of nodes). RI is not a normalized quantity, e.g., the upper bound is 1 but the lower bound is more than zero (network dependent). Due to this lack of normalization, Hubert and Arabie [60] suggested an improvement to RI as described below.

Adjusted Rand index (ARI)

ARI is equivalent to a normalized Rand Index. Consider a confusion matrix for methods *A* and *B* where rows correspond to the communities in *A* and columns correspond to the communities in *B*. N_{ij}, the $(i,j)^{th}$ entry in this matrix, is the number of nodes in both community *i* of method *A* and community *j* in method *B*. Denote by $N_{i.}$ the sum of all columns for row *i*; thus $N_{i.}$ is the number of nodes in community *i* of method *A*. Define $N_{.j}$ to be the sum of all rows for column *j*, i.e. $N_{.j}$ is the number of nodes in community *j* in method *B*. The Adjusted Rand Index (ARI), is calculated from the values N_{ij} of the confusion matrix for the two methods as follows [60]:

$$t_1 = \sum_{i=1}^{c_A} \binom{N_{i.}}{2}; \quad t_2 = \sum_{j=1}^{c_B} \binom{N_{.j}}{2}; \quad t_3 = \frac{2t_1t_2}{N(N-1)}$$

$$ARI(A,B) = \frac{\sum_{i=1}^{c_A} \sum_{j=1}^{c_B} \binom{N_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3} \tag{2.18}$$

where, c_A and c_B are the number of communities detected by methods *A* and *B*, respectively.

Normalized mutual information (NMI)

Another metric to calculate the similarity between two methods is Normalized Mutual Information (NMI). NMI is the normalized form of Mutual Information (MI). MI measures similarity between the results of two methods and is given by [59]:

$$MI(A,B) = \sum_{i=1}^{c_A} \sum_{j=1}^{c_B} \frac{N_{ij}}{N} log(\frac{N_{ij}N}{N_{i.}N_{.j}}) \tag{2.19}$$

where, *A* and *B* are the methods being compared. The terms used in this equation are the same as those used in the equation for ARI (Equation 2.18). Then, NMI between methods *A* and *B* is calculated as:

$$NMI(A,B) = \frac{-2\sum_{i=1}^{c_A} \sum_{j=1}^{c_B} \frac{N_{ij}}{N} log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\sum_{i=1}^{c_A} N_{i.} log(\frac{N_{i.}}{N}) + \sum_{j=1}^{c_B} N_{.j} log(\frac{N_{.j}}{N})} \tag{2.20}$$

Jaccard index

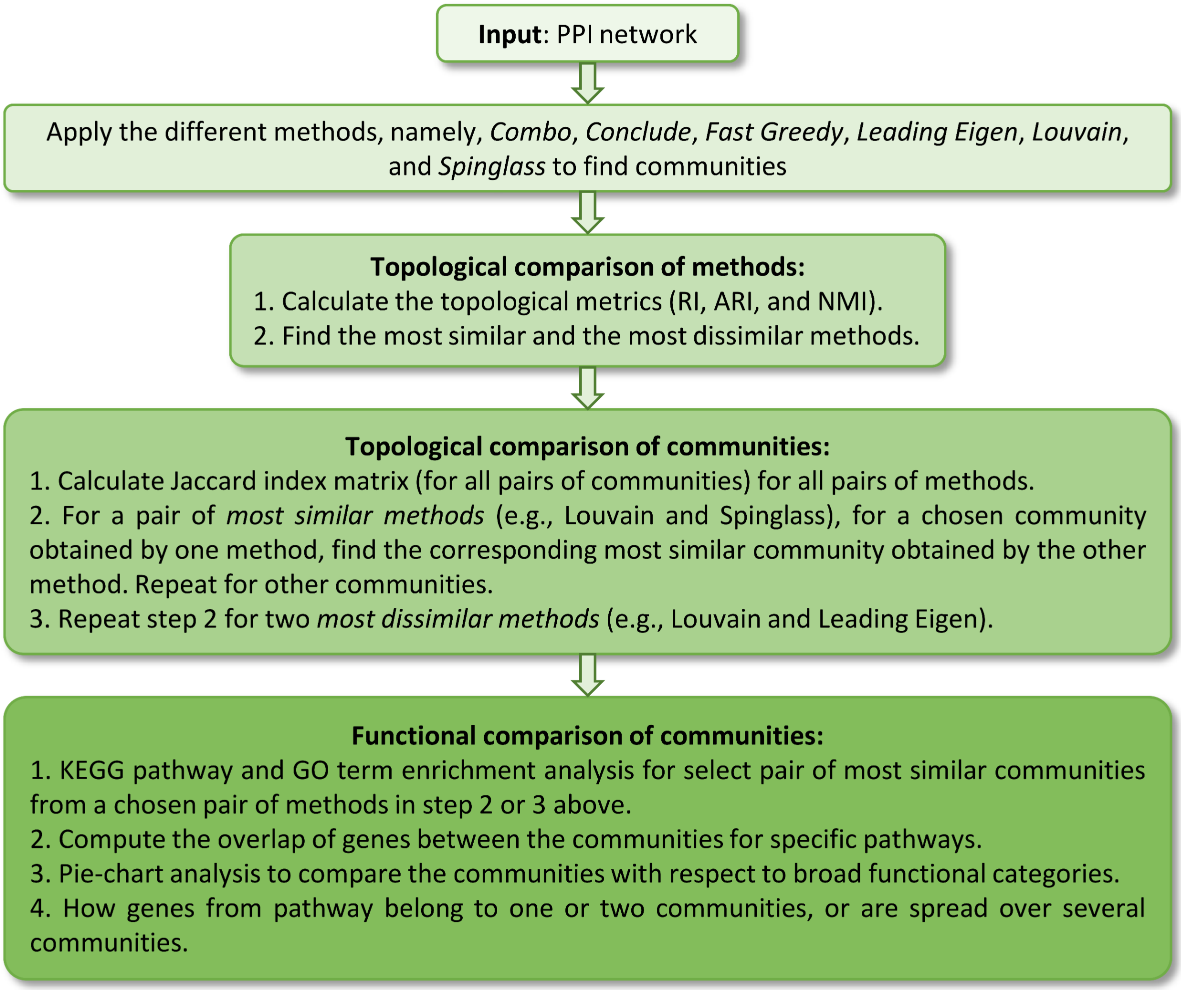
Jaccard index is a measure of similarity for two sets of nodes, with a range from 0 to 1 and is defined as the size of the intersection (overlap) divided by the size of the union of the sets:

$$J(A,B) = \frac{\#(A \cap B)}{\#(A \cup B)} \tag{2.21}$$

where, the numerator is the number of common elements between the sets A and B and the denominator is the number of all the elements in A and B combined.

2.6.3 Overall Approach for Topological and Functional Comparison of Communities Detected by Different Algorithms

Applying the above methods to PPI networks yields different number of communities with different number of nodes. The communities detected are compared in two ways: topological comparison and functional comparison. In topological comparison, methods are compared using different metrics (RI, ARI, and NMI). Based on the results of these metrics, we are able to figure out which methods are similar to each other and which are dissimilar. When a method is compared with itself, RI, ARI, and NMI are 1. Larger (smaller) the value of metrics, more (less) similar are the two methods being compared. After finding which methods are similar to each other from a topological perspective, functional comparisons (such as KEGG pathway enrichment analysis) have been used to further assess the functional similarity of the communities identified by these methods. Figure 2.6 shows a flow chart of our analysis pipeline.



Input: PPI network

Apply the different methods, namely, *Combo*, *Conclude*, *Fast Greedy*, *Leading Eigen*, *Louvain*, and *Spinglass* to find communities

Topological comparison of methods:
1. Calculate the topological metrics (RI, ARI, and NMI).
2. Find the most similar and the most dissimilar methods.

Topological comparison of communities:
1. Calculate Jaccard index matrix (for all pairs of communities) for all pairs of methods.
2. For a pair of *most similar methods* (e.g., Louvain and Spinglass), for a chosen community obtained by one method, find the corresponding most similar community obtained by the other method. Repeat for other communities.
3. Repeat step 2 for two *most dissimilar methods* (e.g., Louvain and Leading Eigen).

Functional comparison of communities:
1. KEGG pathway and GO term enrichment analysis for select pair of most similar communities from a chosen pair of methods in step 2 or 3 above.
2. Compute the overlap of genes between the communities for specific pathways.
3. Pie-chart analysis to compare the communities with respect to broad functional categories.
4. How genes from pathway belong to one or two communities, or are spread over several communities.

Figure 2.6: Flow chart of the steps used in our analysis.

2.7 Acknowledgement

Chapter 2, in full, is a reprint of the material as it appears in "Topological and functional comparison of community detection algorithms in biological networks" by Rahiminejad, Sara; Maurya, Mano R.; and Subramaniam, Shankar., BMC Bioinformatics, 2019. The dissertation author was the primary investigator and author of this paper.

Funding for this study was provided by the National Science Foundation grant STC0939370 and by the National Institutes of Health (NIH) Grants U01 DK097430, U01 CA200147, U01

CA198941, U19 AI090023, R01 HL106579, R01HL108735, R01 HD084633, R01 DK109365 (to SS), and R01LM012595 (to SS and MRM).

Availability of data and material

The data on Yeast (combined and physical interaction-only) and Human PPI networks are available from BioGRID (https://thebiogrid.org/; see also Tables S2.29-S2.31). All the source code used in this work is available from the respective publications and websites (R igraph: http://igraph.org/r/, Conclude method: http://www.emilio.ferrara.name/code/conclude/, and Combo method: http://senseable.mit.edu/community_detection/).

Chapter 3 MODULAR AND MECHANISTIC CHANGES ACROSS STAGES OF COLORECTAL CANCER

3.1 Abstract

Background: While mechanisms contributing to the progression and metastasis of colorectal cancer (CRC) are well studied, cancer stage-specific mechanisms have been less comprehensively explored. This is the focus of this manuscript. Methods: Using previously published data for CRC (Gene Expression Omnibus ID GSE21510), we identified differentially expressed genes (DEGs) across four stages of the disease. We then generated unweighted and weighted correlation networks for each of the stages. Communities within these networks were detected using the Louvain algorithm and topologically and functionally compared across stages using the normalized mutual information (NMI) metric and pathway enrichment analysis, respectively. We also used Short Time-series Expression Miner (STEM) algorithm to detect potential biomarkers having a role in CRC. Results: Sixteen Thousand and Sixty-Two DEGs were identified between various stages (p-value ≤ 0.05). Comparing communities of different stages revealed that neighboring stages were more similar to each other than non-neighboring stages, at both topological and functional levels. A functional analysis of 24 cancer-related pathways indicated that several signaling pathways were enriched across all stages. However, the stage-unique networks were distinctly enriched only for a subset of these 24 pathways (e.g., MAPK signaling pathway in stages I-III and Notch signaling pathway in stages III and IV). We identified potential biomarkers, including *HOXB8* and *WNT2* with increasing, and *MTUS1* and *SFRP2* with decreasing trends from stages I to IV. Extracting subnetworks of 10 cancer-relevant genes and their interacting first neighbors (162 genes in total) revealed that the connectivity patterns for these genes were different across stages. For example, *BRAF* and *CDK4*, members of the Ser/Thr kinase, up-regulated in cancer, displayed changing connectivity patterns from stages I to IV. Conclusions: Here, we report molecular and modular networks for various stages of CRC, providing a pseudo-temporal view of the mechanistic changes associated with the disease. Our analysis highlighted similarities at both functional and topological levels, across stages. We further identified stage-specific mechanisms and biomarkers potentially contributing to the progression of CRC.

3.2 Background

Colorectal cancer (CRC) refers to cancers affecting both colon and rectum. According to GLOBOCAN 2020 data, CRCs are the third most diagnosed and the second most deadly form of cancer worldwide, comprising 11% of all cancer diagnoses [61]. The survival is highly dependent upon the stage of disease at diagnosis and earlier detection portends higher chance of survival [62]. Two types of risk factors contribute to the incidence of CRC. The first type includes the ones that are beyond the control of the individual, such as age and hereditary factors. The second type is related to environmental and lifestyle risk factors such as diets high in fat, physical inactivity, smoking, and heavy alcohol consumption [63].

CRC is said to progress through five stages. The earliest stage, stage 0 represents the presence of abnormal cells in the mucosa of the colon wall. In stage I, tumor penetrates the submucosa of the colon or rectum wall, while at stage II the cancer has spread through the wall to the serosa, but not the nearby organs. Stage III represents cancer in the mucosa, submucosa, serosa and the spread into the nearby lymph nodes. Stage IV represents the most aggressive form of CRC, where the cancer metastasizes and spreads to other parts of the body [64]. Biomarkers, agnostic of stages, have been used for detection of CRC [65]. For example, p53, a key biomarker, is a tumor suppressor gene, mutated in 34% of the proximal colon tumors and in 45% of the distal colorectal tumors [66, 67]. Prior work from our group [68] and many others have identified potential causes and mechanisms of CRC, but a few have focused on identifying the stage-specific dysregulation, and biomarkers. Palaniappan et al. identified novel cancer genes that could underlie the stage-specific progression and metastasis of CRC [69]. Cai et al. performed a comprehensive untargeted metabolomics analysis on normal and tumor colon tissues from CRC patients and identified 28 highly discriminatory tumor tissue metabolite biomarkers [70].

In this study, we focused on modelling each stage as a molecular network and identifying subnetworks (communities) which enable better mechanistic interpretation [17, 24]. To this extent, we utilized a gene expression microarray dataset containing 104 human CRC samples (across stages I to IV) and 24 normal samples from Gene Expression Omnibus (GEO) to detect stage-specific biomarkers and modular mechanisms, potentially causal for the progression of CRC. We first constructed gene correlation networks for each of the stages, and detected communities using the Louvain algorithm [20, 22]. The communities are functionally interpreted in the context of CRC. We also developed stage-unique networks (by retaining edges unique to that specific stage) and functionally interpret them. Next, we utilized Short Time-series Expression Miner (STEM) approach to identify candidate biomarkers with substantial/monotonic changes across stages [71, 72]. A biologically driven analysis enabled characterization of the evolution of molecular subnetworks across stages. Lastly, a drug-target-PPI (Protein–Protein Interaction) network is generated which may provide insight into understanding stage-specific functional mechanisms for some of the current drugs used in CRC treatment. Figure 3.1 shows a flow chart for our analysis pipeline.

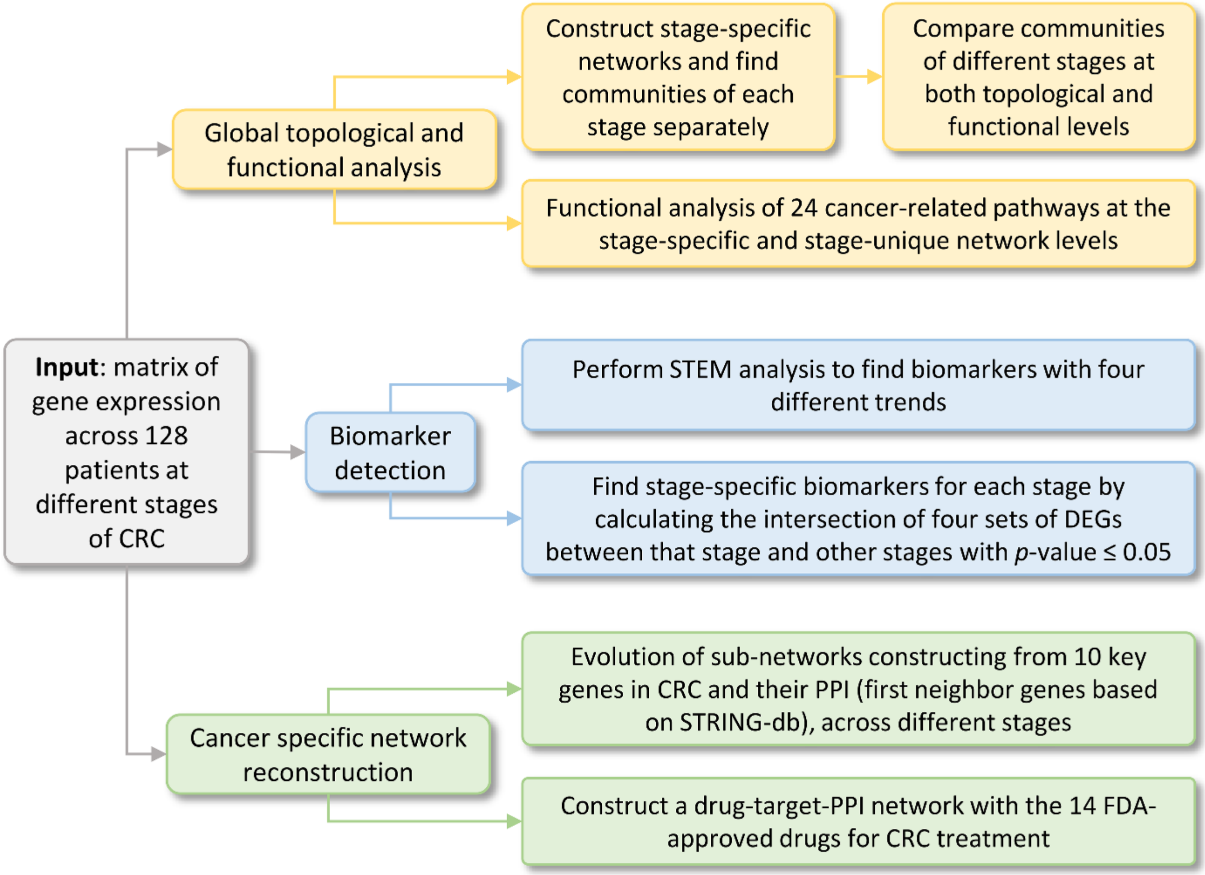

Construct stage-specific networks and find communities of each stage separately
Compare communities of different stages at both topological and functional levels
Global topological and functional analysis
Functional analysis of 24 cancer-related pathways at the stage-specific and stage-unique network levels
Input: matrix of gene expression across 128 patients at different stages of CRC
Perform STEM analysis to find biomarkers with four different trends
Biomarker detection
Find stage-specific biomarkers for each stage by calculating the intersection of four sets of DEGs between that stage and other stages with p-value ≤ 0.05
Evolution of sub-networks constructing from 10 key genes in CRC and their PPI (first neighbor genes based on STRING-db), across different stages
Cancer specific network reconstruction
Construct a drug-target-PPI network with the 14 FDA-approved drugs for CRC treatment

Figure 3.1: Flow chart of the approach used in our analysis.

3.3 Materials and Methods

Microarray data pre-processing

We used a CRC microarray dataset from the GEO (accession ID GSE21510) containing samples from 13 patients in stage I, 37 patients in stage II, 34 patients in stage III, and 20 patients in stage IV cancer along with 24 normal samples. There was only one sample associated with stage 0 and we excluded it from our analysis. More details about the clinical characteristics of the GSE21510 have been presented in the original publication by Tsukamoto et al. [73]. The raw dataset had 54,675 probe IDs across 128 samples/patients and it was re-normalized using Robust Multi-array Average (RMA) normalization [74]. Probe IDs with missing or multiple Entrez gene IDs (based on annotation file from GEO) were removed from the dataset. Both linear and non-linear dimensionality reduction algorithms (Principal Component Analysis (PCA) [75] and t-Distributed Stochastic Embedding (t-SNE) [76]) were used to detect outliers in the data. PCA was performed in R, using *prcomp* and *autoplot* functions (of *ggfortify* package). t-SNE was also performed in R, using *Rtsne* package.

Differentially expressed genes (DEGs)

We identified DEGs at the probe level using *limma* [77] between each pair of neighboring stages (i.e., stage I vs. normal, stage II vs. I, stage III vs. II and stage IV vs. III). For genes with multiple probe IDs, the geometric mean of the *p*-values of the multiple probes was used as the *p*-value for the gene. DEGs were then identified at p-value ≤ 0.05 for each comparison and their union across 4 comparisons (stage I vs. normal, stage II vs. I, stage III vs. II, and stage IV vs. III) was calculated as a master list of DEGs.

Network construction

Networks for each of the stages were constructed using correlation of gene expression values for the DEGs identified. Specifically, the Pearson Correlation Coefficient (PCC) [-1, 1], r, was calculated between all gene-pairs. Networks were then constructed using a cut-off for PCC at each stage (r_{th}) based on the degree of freedom (number of patients at that specific stage - 2) and a *p*-value threshold of 0.001. Edges with p-value ≤ 0.001 (i.e., $|r| \geq r_{th}$) were retained. The weight of edges was binary (0 or 1) for unweighted networks and non-binary ($0 \leq w \leq 1$) for weighted networks, with the absolute value of PCC being used as weights. We refer to these networks as stage-specific networks (whole networks for the normal, stage I, II, III, and IV). Stage-unique networks were also constructed for each of the stages by removing edges from stage-specific network of each stage that were common with any other stage-specific network.

Community detection

We used the Louvain algorithm to detect communities within each stage-specific network given its established status as the leading method for community detection [20, 22]. Louvain detects network communities by maximizing modularity (a measure of the density of links (edges) within communities compared to links between communities). Briefly, the search for communities using the algorithm proceeds in two phases. During the first phase, communities are detected by optimizing modularity locally. During the second phase, nodes of the same community are aggregated as pseudo-nodes to generate a new network. The combination of these two phases is iterated, until the modularity reaches a local maximum. The computational complexity of this algorithm is ($O(n\log n)$) which makes it extremely fast [20] (also see Supplementary Methods).

Topological and functional comparison of communities

Normalized Mutual Information (NMI) metric was utilized to compare communities of different stages at a topological level [59]. NMI is 1 when a network is compared with itself. Larger (smaller) the value of NMI, more (less) similar are the networks being compared (see Supplementary Methods). To assess the statistical significance of the NMI values, we needed to compute their *p*-value. Hence, we generated 1000 random networks with the same number of nodes, edges and degree distribution as the stage I-, II-, and III-specific networks. Communities of random networks were identified using the Louvain algorithm and compared between stage I- and II- and stage II- and III-specific networks using the NMI metric. *p*-values for comparing the stage-specific networks were then calculated from the histogram of the 1000 NMI values.

Jaccard index (JI), the ratio of the count of common genes to the count of union of genes in two groups, was used to identify pairs of communities which were similar to each other in terms of genes common between them. The most similar communities were then compared at a functional level. We used Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment available via DAVID version 6.8 [41, 42] for functional analysis [78].

Edge-based functional enrichment

p-values for edge-based enrichment were computed using a hypergeometric test for edges (gene pairs) [79] which accounted for the topology of the network. For d DEGs, the total number of edges, N, is calculated as $d(d - 1)/2$. Similarly, for a given KEGG pathway with d_{KEGG} enriched genes (from our master list of d DEGs), m_{KEGG} edges are calculated ($m_{KEGG} = d_{KEGG}(d_{KEGG} - 1)/2$). Suppose a network contains n edges, of which k edges are between d_{KEGG} genes of the given KEGG pathway, then a *p*-value for the edge-based enrichment of this pathway is calculated from a hypergeometric distribution as:

$$\begin{aligned} p(k|KEGG\ pathway) &= \sum_{i=k}^{m_{KEGG}} P(X = i|KEGG\ pathway) \\ &= \sum_{i=k}^{m_{KEGG}} \frac{\binom{m_{KEGG}}{i}\binom{N-m_{KEGG}}{n-i}}{\binom{N}{n}} \end{aligned} \tag{3.1}$$

Equation 3.1 provides an estimate for the probability of observing k or more edges between p_{KEGG} genes for the given KEGG pathway [79]. The R function *phyper* with 4 parameters was used to calculate the edge-based *p*-value using *phyper* (k - 1, n, N - n, m_{KEGG}, *lower.tail* = *FALSE*).

Biomarker identification

The STEM algorithm [71, 72] (see Supplementary Methods) was utilized to identify potential biomarkers. Since there were different number of patients at each stage, the median of gene-expression for patients at each stage was considered as the representative gene expression for that stage. STEM works by first selecting a set of potential profiles and then assigning genes to the profile that best captures their expression trend. We selected 60 model profiles and a maximum unit change of 1, which represents the change a gene could have between successive time points. Gene Expression Profiling Interactive Analysis (GEPIA2) [80] was next used to validate the biomarkers identified using STEM analysis within independent cohort (TCGA COAD-READ) at $|\log2FC| \geq 1$ and q-value (FDR adjusted *p*-value) ≤ 0.05.

Supervised analysis with key genes

We identified the interacting proteins of 10 key genes with known roles in CRC using STRING-db [81]. A subnetwork of the key genes and their first neighbors were extracted from each stage-specific network, separately. The analysis consisted of the following steps:

- Detection of the first neighbors of 10 key genes from STRING-db with two criteria: score threshold ≥ 0.4 and up to 20 connections between genes.
- Identification of unique genes from the union of 10 key genes and their first neighbors found in STRING-db.
- Extraction of subnetworks of the unique genes from each stage-specific network and their visualization using Cytoscape [45]. |PCC| between the genes were used as edge-weights.

Drug-target-PPI network

Approved drugs and their target genes for CRC were identified from National Cancer Institute (NCI) [82] and DrugBank databases [83]. We then projected the PPI information from STRING-db (score threshold ≥ 0.9) [81] and gene weights from the stage-specific networks onto the drug-target interactions detected above. We also identified important KEGG pathways related to these target genes. The constructed network was visualized using Cytoscape [45].

3.4 Results and Discussion

3.4.1 Identification of DEGs

The CRC dataset used here contained 41,834 probe IDs across 128 samples after pre-processing (see Materials and Methods). Outlier detection using PCA and t-SNE identified two normal samples as outliers which were eliminated, leaving 126 samples for our analysis. The first two PCs and the first two dimensions of t-SNE are shown in Figures S3.1A and S3.1B, respectively. In order to capture the most significant genes, we identified DEGs (see Materials and Methods) between neighboring stages with p-value ≤ 0.05 resulting in 15,634 DEGs between stage I and normal, 528 DEGs between stages II and I, 745 DEGs between stages III and II, and 503 DEGs between stages IV and III. The union of all DEGs (16,062 unique genes) was considered as the master list of DEGs for all downstream analysis.

3.4.2 Correlation-based Network Analysis

PCC was calculated for all pairs of DEGs to construct the networks (see Materials and Methods). For each stage, based on the number of patients and a fixed p-value, we identified the corresponding threshold for PCC. Unweighted, stage-specific and stage-unique networks were subsequently constructed using the 16,062 DEGs [49]. Table 3.1 lists some basic properties for different stage-specific networks. Properties for unweighted networks are listed in Table S3.1. Number of nodes and edges for all communities of stage-specific and unweighted networks can be found in Tables S3.2 and S3.3, respectively.

Table 3.1: Properties for stage-specific networks

Network	# of patients	PCC cut-off	# of edges	# of communities	Modularity
Normal	22	0.6523	1,809,792	18	0.43
Stage I	13	0.8009	507,603	17	0.51
Stage II	37	0.5186	1,063,390	9	0.44
Stage III	34	0.5392	1,214,109	9	0.45
Stage IV	20	0.6788	763,554	11	0.45

In the following section, we compare communities detected within stage-specific networks at the topological and functional levels. The NMI metric was used to compare networks at a topological level. Highly similar networks (at the topological level) were further analyzed at a functional level. KEGG pathway enrichment analysis was used to assess functional similarity of the communities detected.

3.4.2.1 Neighboring stages are functionally and topologically similar

Using the NMI metric, we evaluated the similarity between networks. Table 3.2 and S3.4 represent the results of comparing communities of stage-specific networks and unweighted networks using NMI. Based on the results of Table 3.2 (and Table S3.4), neighboring stages were found to be more similar to each other than non-neighboring stages. A permutation test was also performed to assess the statistical significance of the NMI values seen in Table 3.2 (see Materials and Methods). Figures 3.2A and 3.2B show histograms for the values of NMI between communities of the random networks of stages I and II, and II and III, respectively. Our analysis highlighted that the NMI calculated between stage-specific networks was highly significant (e.g., *p*-value of NMI between communities of the stages I- and II-specific networks was 0.001 and between communities of the stages II- and III-specific networks was 0.05).

Table 3.2: Comparing stage-specific networks using NMI

	Normal	Stage I	Stage II	Stage III	Stage IV
Normal	1	0.0282	0.0403	0.0444	0.0276
Stage I		1	0.0729	0.0689	0.045
Stage II			1	0.1501	0.0887
Stage III				1	0.0771
Stage IV					1

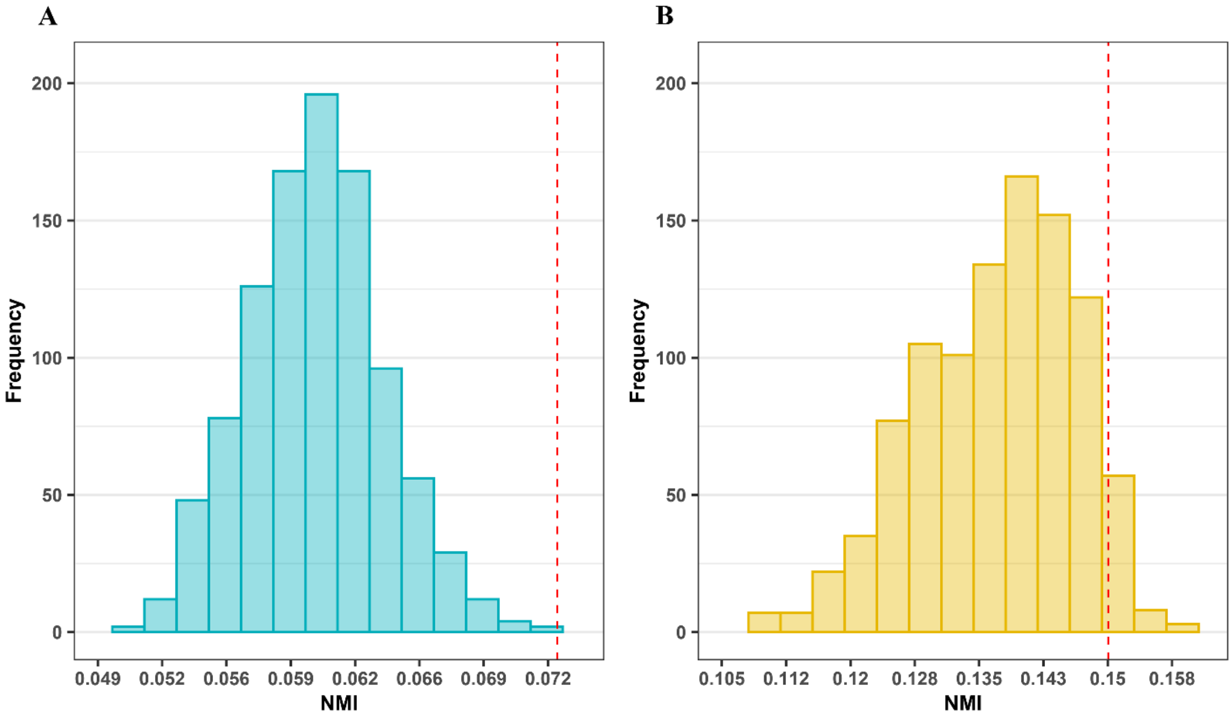


Figure 3.2: Histogram of a permutation test for comparing communities of different stages with degree preservation. **A** NMI metric between random networks with sizes equal to stage I- and II-specific networks. The actual value of NMI for comparing those stages is 0.0729 (vertical dotted line), corresponding to a p-value of 0.001 (significant for a *p*-value threshold of 0.05). **B** NMI metric between random networks with sizes equal to stage II- and III-specific networks. The actual value of NMI for comparing those stages is 0.1501 (vertical dotted line), (*p*-value of 0.05, significant).

We next used JI values for direct topological comparison of individual communities across stage-specific networks. The higher the value of JI, the more similar were the communities being compared. For example, JI values for comparing the communities of stage I-specific network with the communities of other stages is shown in Figure 3.3A (see also Table S3.5). Likewise, a functional comparison of the third community of stage I with the corresponding communities of other stages revealed that the third community of stage I was more similar to the first community of stage II (the neighboring stage) than the corresponding communities of other stages (based on JI) (see Figure 3B and Tables S3.6 through S3.9). Pathways indicated in this comparison were chosen at a *p*-value ≤ 0.01 and with more than 10 genes from the third community of stage I. For some pathways such as Ras signaling pathway, the number of enriched genes and the *p*-values were similar between the third community of stage I and the first community of stage II but did not meet the threshold for the corresponding communities of other stages.

Analyzing each community individually can leave out important functions due to the distribution of functionally related genes between them. Hence, we carried out an edge-based functional analysis (see Materials and Methods) at the whole network level (consisting of 16,062 genes). We constructed stage-unique networks and compared both types of networks (stage-specific and stage-unique) at a functional level.

3.4.2.2 Functional analysis at the whole network level

Some of the edges were common among two or more stage-specific networks. To identify edges unique to each stage, we constructed stage-unique networks (See Materials and Methods). We identified 1,668,692 edges for normal-, 430,446 edges for stage I-, 839,058 edges for stage II-

, 967,358 edges for stage III-, and 627,558 edges for stage IV-unique networks. The number of edges for the stage-specific networks are listed in Table 3.1.

To ascertain the functional relevance for the networks, we first selected 24 pathways associated with cancer progression (from initiation to metastasis) and carried out a supervised analysis. A list of all pathways enriched for the master list of genes can be found in Table S3.10. We calculated the edge-based *p*-values and performed an enrichment for the 24 pathways for the stage-specific (see Table 3.1) and stage-unique networks (see Figure 3.3C). The *p*-value cut-off was 0.05. The number of edges associated with genes enriched in the stage-unique networks were less than the stage-specific networks for all stages and for all 24 pathways. We noted that the number of edges for each stage-unique network was less than its value for the stage-specific network of that stage. For example, stage I-unique network had 430,446 edges as compared to the stage I-specific network with 507,603 edges.

Among the 24 cancer-related pathways, we observed that central carbon metabolism pathway was enriched across stages II and III and is known to play a role in cancer progression [84]. Cell cycle and DNA replication pathways were significantly enriched in almost all stages with more edges in the Cell cycle pathway. Several signaling pathways including PI3K-Akt, Ras, MAPK, TGF-beta, p53, and T cell receptor signaling pathway associated with cell growth were enriched across stages. PI3K-Akt signaling pathway plays an important role in the growth and progression of CRC. Both MAPK and PI3K-Akt serve as a molecular target for treatment of CRC [85, 86]. TGF-beta signaling pathway was particularly enriched only in stages II, and IV. TGF-beta is known to play a significant role in inflammation and tumorigenesis by modulating cell growth, differentiation, apoptosis, and homeostasis, contributing to tumor maintenance and cancer progression [87]. Besides changes in enrichment of specific pathways, changes in connectivity pattern of specific genes in key pathways were also observed across the CRC stages. For example, Figures 3.3D and 3.3E show the connectivity pattern of genes in the p53 signaling pathway across stages I-IV and normal, respectively. p53 signaling pathway has a critical role in the regulation of Cell cycle, DNA replication and apoptosis [88]. Comparing Figures 3.3D and 3.3E revealed that hub genes were different between cancer stages and normal. For example, *CCND1* and *CDK6* were two genes with high connectivity (degree) in normal only. *CCND1* is a proto-oncogene which is known to play a critical role in promoting the G1-to-S transition of the cell cycle in many cell types [89]. Likewise, *CCNE1*, also a proto-oncogene, displayed high degree of connectivity in stages II and III which was not present in normal. *CCNE1* serves as a positive regulator of cell cycle and promotes G1-to-S phase transition by activating *CDK2* [90, 91]. *CDK2* also showed high degree of connectivity in stage III, although it was not present in normal.

Focal adhesion pathway was more enriched in normal, stage II and stage III than in other stages. Focal adhesion kinase (*FAK* or *PTK2*) is a major integrin-dependent tyrosine phosphorylated protein in this pathway and known to contribute significantly to inflammatory signaling pathways. *PTK2* has been suggested to be a potential target for CRC therapies [92]. NF-kappa B signaling pathway was enriched in normal, and stages I, II and III, and is a regulator of immune response and inflammation and associated with carcinogenesis [93]. VEGF signaling associated genes, with known roles in angiogenesis and metastasis, were enriched in stages I and II [94], while Notch signaling pathway, a main pathway in metastasis and tumor angiogenesis processes, was enriched in stages III and IV.

Overall, most of the cancer related pathways were enriched across all stage-specific networks. However, the enrichment of those pathways was distinct across stage-unique networks.

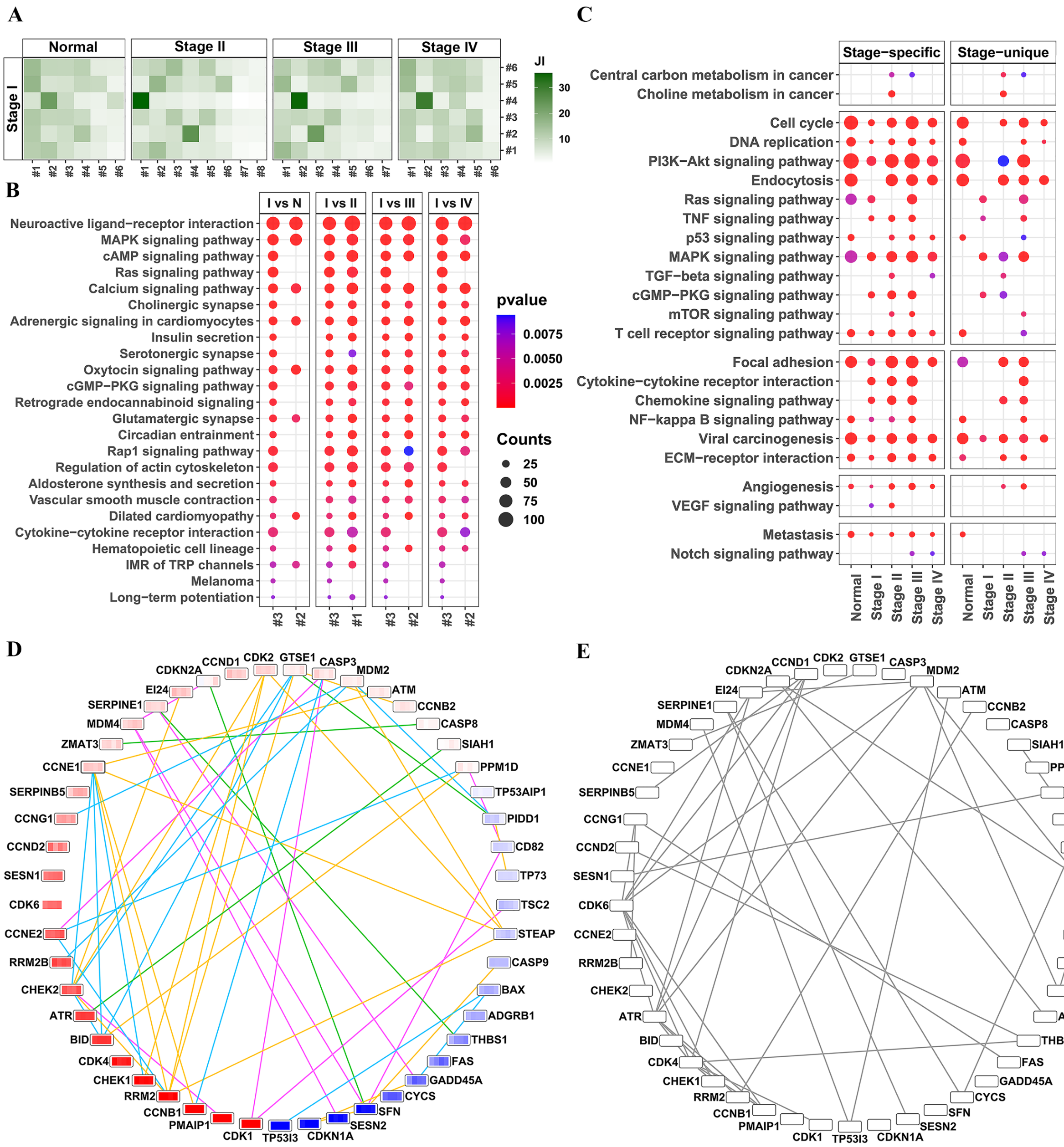


Figure 3.3: Topological and functional analysis of the weighted correlation networks. **A** Heat map for JI values for comparing the communities of stage I-specific network with the communities of other stages. The color-scale is from white for the minimum value of JI (0%) to green for the maximum value (36%). **B** Functional comparison of the KEGG pathways with p-values ≤ 0.01 and more than 10 genes for the third community of stage I-specific network with the corresponding communities of other stages. **C** Functional comparison (edge-based enrichment) of the stage-specific and stage-unique networks for 24 cancer-related pathways divided into five categories: cancer related, cell cycle/proliferation/growth, inflammation, angiogenesis, and metastasis. Number of edges related to the genes enriched in each pathway are indicated by the size of the dots. Color scale of the dots indicate p-value with a cut-off of 0.05. **D** Connectivity of p53 signaling pathway genes across different stage-unique networks. The nodes are colored based on the log2FC values (in a specific stage vs. normal) across the four stages I-IV (dark blue (log2FC of -2) to white (0) to dark red (2)). Each node represents four log2FC values, going from left to right. Edges are colored differently across stages as follows: green for edges in stage I, cyan in stage II, yellow in stage III, and purple in stage IV. **E** Connectivity of p53 signaling pathway genes in normal.

3.4.2.3 In-silico validation

To validate our result at the gene and pathway level, we analyzed the TCGA COAD-READ data available through GEPIA2 by identifying DEGs with q-value < 0.05 for COAD and READ cohorts, resulting in 16,438 DEGs common to both. Since GEPIA2 does not allow for stage-wise identification of DEGs, we calculated DEGs across all stages (104 samples) and normal (22 samples) at q-value < 0.05 within our dataset. A total of 16,641 DEGs were identified, of which 11,389 were common with the TCGA COAD-READ cohort. A hypergeometric test on the overlap indicated that the number of DEGs as common were statistically significant ($p = 0.05$). The total number of genes used for the hypergeometric test was 24,136. The log2FC of genes identified as common between the COAD, READ, and our dataset are also provided in Table S3.11. Of the 11,389 genes, ~ 65% of the genes showed expression trends in the same direction within COAD-READ as DEGs identified in our current study. Functional analysis of the 11,389 genes further revealed several signaling pathways enriched crucial to CRC consistent with our results including Ras, MAPK, PI3K-AKT, TGF-beta and WNT signaling (Figure 3.3C).

3.4.3 Biomarkers

We performed STEM analysis to identify potential biomarkers and validated them using TCGA COAD-READ cohort, available through GEPIA2 [80].

3.4.3.1 Four distinct biomarker trends identified in CRC via STEM analysis

We selected 60 model profiles and the maximum unit change of 1 for the STEM analysis (see Materials and Methods). Most of the genes were clustered in two main trends, (0,1,1,1,1) and (0,-1,-1,-1,-1), implying that the expression of genes changed extensively up or down from the normal condition but with little or no difference across stages I-IV (Figures 3.4A-D). The trends identified were consistent with TCGA COAD-READ cohort results from GEPIA2 (Figures S3.2A-D). Tables S3.12 through S3.17 list the genes belonging to each trend.

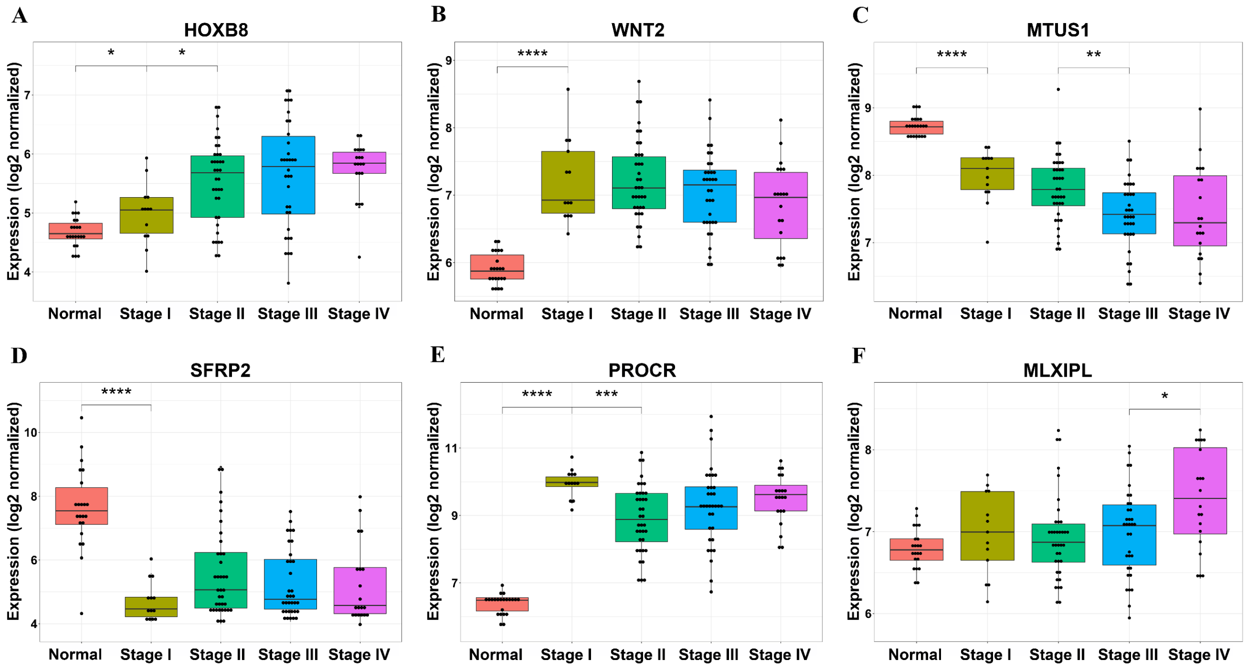


Figure 3.4: Biomarkers. **A-D** Boxplots for 4 biomarkers from STEM analysis and **E–F** boxplots for 2 stage-specific biomarkers, consistent with GEPIA2 COAD-READ cohort results. Each color indicates one stage and dots show the expressions of biomarker gene for patients in every stage. **A** HOXB8 with trend (0,1,2,3,4), **B** WNT2 with trend (0,1,1,1,1), **C** MTUS1 with trend (0,-1,-2,-3,-4), **D** SFRP2 with trend (0,-1,-1,-1,-1), **E** PROCR, stage I-specific biomarker, and **F** MLXIPL, stage IV-specific biomarker.

We highlighted some genes which exhibit the aforementioned trends including *HOXB8*, with monotonically increasing expression from normal through the cancer stages (Figure 3.4A). Studies have shown that knockdown of *HOXB8* inhibits cellular proliferation and invasion in vitro, as well as carcinogenesis and metastasis in vivo. *HOXB8* has been suggested as an independent prognostic factor in CRC [95, 96]. Likewise, *WNT2*, an oncogene, exhibited an increasing STEM trend and was over-expressed in CRC (Figure 3.4B), across stages, compared to normal tissues. *WNT2* is known to be involved in canonical Wnt signaling activation during CRC tumorigenesis, and has been suggested to enhance tumor growth and the invasion in a paracrine fashion [97, 98]. *WNT2* has been previously identified as a stool marker with a sensitivity of 74–78% and specificity of 88–89% [99, 100]. *MTUS1* expression (Figure 3.4C), was significantly down-regulated in human colon cancer tissues and has been documented in earlier studies [101]. It has been suggested to be involved in the loss of proliferative control in human colon cancer via its interference of *ERK2* pathway activation [102]. *SFRP2* gene, located upstream of the canonical Wnt signaling pathway, was also found to be suppressed across all stages [103]. *SFRP2* was the first reported DNA methylation marker in stool with a sensitivity of 32.1–94.2% and specificity of 54–100% [104]. DNA hypermethylation of *SFRP2* leads to the downregulation of the gene expression, inhibition of gene function and promotion of CRC [100]. GEPIA2 was additionally used to generate disease-free survival (DFS) plots of the four biomarkers identified by STEM analysis (Figures S3.3A-D). DFS plot of *HOXB8* confirmed that high expression of this genes was associated with poor disease-free survival of patients with CRC.

3.4.3.2 Stage-specific biomarkers

Biomarkers specific to each stage were identified as the intersection of four sets of DEGs between that stage and other stages with p-value ≤ 0.05. In total, 110 potential stage-specific biomarkers including 41 for stage I, 21 for stage II, 8 for stage III, and 40 for stage IV were identified (listed in Table S3.18). p-values for 10 comparisons (e.g. normal vs stage I) for all 110 potential biomarkers were listed in Table S3.19. Figures 3.4E and 3.4F show boxplots for *PROCR*, a stage I-specific biomarker and *MLXIPL* (*ChREBP*), a stage IV-specific biomarker, respectively. The trends for these two biomarkers identified were consistent with TCGA COAD-READ cohort results obtained through GEPIA2 (Figures S3.2E-F). High expression of *PROCR* and *MLXIPL* was associated with poor disease-free survival of CRC patients (Figures S3.3E-F). It has been shown that *PROCR* overexpressed in CRC epithelial tumor cells, through immunohistochemistry [105]. This upregulation is caused by gene amplification and DNA hypomethylation and occurs in concert with a cohort of neighboring genes on chromosome locus 20q [106]. Studies have shown that *ChREBP* mRNA and protein expression levels are significantly increased in colon cancer tissues compared to normal tissues [107]. Their expression positively correlated with colon malignancy and was suggested to contribute to cell proliferation. Given its functional roles in CRC, and its distinct expression with stage IV, we propose that *ChREBP* could serve as a clinically useful biomarker.

The results presented above were based on an unsupervised analysis at a global network level. We additionally carried out a more focused analysis, emphasizing key drivers of CRC.

3.4.4 Evolution of Subnetworks of Key Genes and First Neighbors across Different Stages

We performed a supervised analysis with 10 key genes with known roles in CRC (see Materials and Methods). The key genes were *TP53*, *APC*, *KRAS*, *BRAF*, *PIK3CA*, *EGFR*, *MLH1*, *TGFBR2*, *PTEN*, and *SMAD4*. The union of key genes and their first neighbors from STRING-db yielded 188 unique genes of which 162 were present within our master list of genes.

The subnetworks of 162 unique genes in stages I- and II-specific networks are shown in Figures 3.5A and 3.5B, respectively. The subnetworks from stages III- and IV-specific networks are shown in Figures S3.4A and S3.4B, respectively. The nodes were clustered based on the communities they belonged to in the stage-specific networks described in the earlier sections. The subnetwork of stage I was sparser but with stronger edge weights since the stage I-specific network had fewer and stronger edge weights (PCC $\geq$ 0.8009) than other stages. We observed these networks to be enriched for several drug targets including *BRAF*, *EGFR*, and *PDGFRB*, and several signaling pathways including Chemokine, PI3K-Akt, ErbB, Ras, TGF-beta, Wnt, p53, NF-kappa B, VEGF and MAPK (Figure 3.5). The subcommunities of both subnetworks included both up- and down-regulated genes. For instance, Figure 3.5A highlights a subcommunity in stage I enriched for several up-regulated genes associated with Chemokine and ErbB signaling pathways, both with known roles in cancer etiology [108, 109]. Likewise, there was a subcommunity in stage II, shown in Figure 3.5B, with genes mostly up-regulated and enriched for Ras signaling and mismatch repair pathways. We also detected a subcommunity within stage II with genes mostly down-regulated (Figure 3.5B) and enriched for pathways such as ErbB and VEGF signaling. VEGF family members play an essential role in tumor-associated angiogenesis, tissue infiltration, and metastasis formation [110].

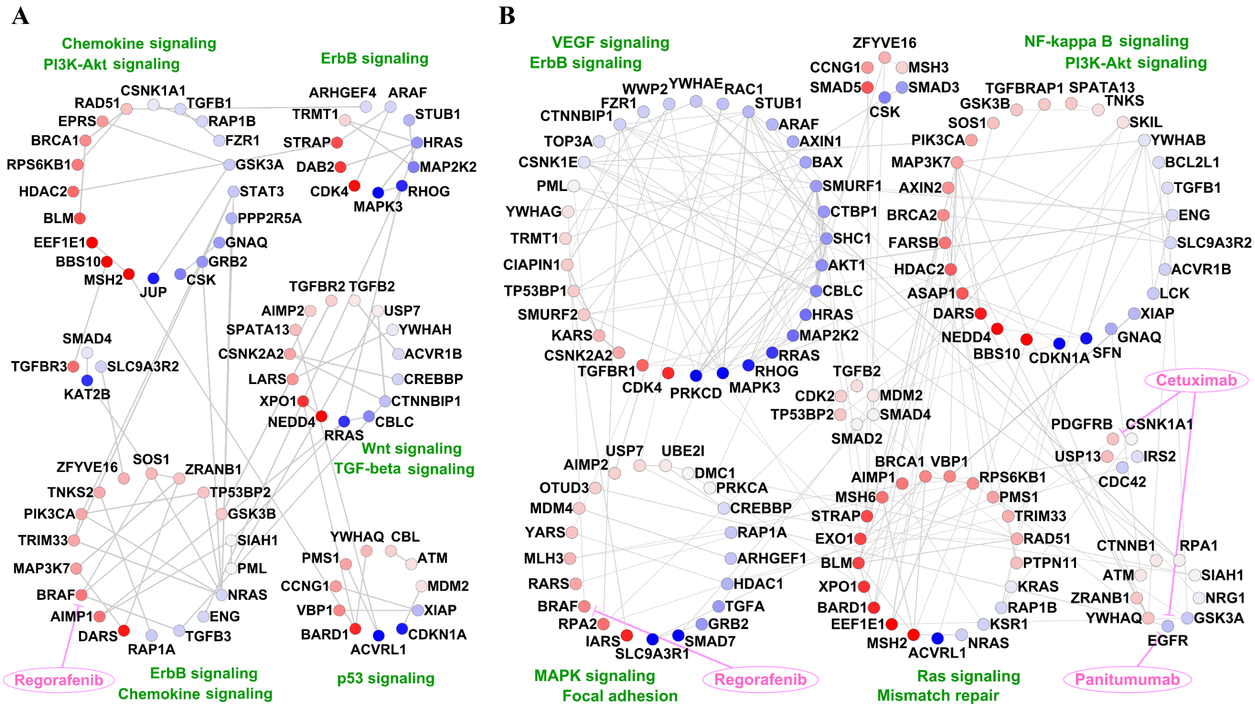


Figure 3.5: Subnetwork of 162 genes in stages I and II. **A** Stage I-specific network and **B** Stage II-specific network. Nodes of each subnetwork are grouped together based on the communities they belonged to in the stage-specific networks and colored based on the value of log2FC between that stage and normal: dark blue (log2FC of -2), to white (0) to dark red (2). The width of edges shows the strength of connections based on PCC between them. The thicker the edges are, the larger the PCC between the nodes is.

These subnetworks all showed differences in connectivity patterns for key genes. For example, *EGFR*, whose degree was zero in all subnetworks except for stage II, is known to play a critical role in oncogenesis, particularly in colon cancer development and is a potential target for therapy [111]. We identified that its expression was down-regulated in the subnetwork of stage II and was connected to *OTUD3* (a tumor promoter in lung cancer [112]). *EGFR* also serves as a drug target for Cetuximab and Panitumumab. *BRAF*, another key player in CRC was up-regulated across all cancer stages compared to normal, yet had distinct connectivity patterns across different stages. *BRAF* was connected to *TGFB3*, *TP53BP2*, and *SOS1* in the subnetwork of stage I. Although the stage II-specific network had more edges compared to stage I-specific network, *BRAF* was connected to only one gene, *YWHAG*, in the subnetwork of stage II. The chemotherapy drug for CRC, Regorafenib, targets *BRAF* and modulates the activity of its protein.

Finally, we sought to understand the functional mechanisms for some of the current drugs used in CRC treatment in the context of our current analysis and identify if any temporal variation in gene-expression of the drug-targets may indicate stage-specificity of the drugs.

3.4.5 Drug-target-PPI Network

We identified 14 FDA-approved drugs for CRC from the NCI website and 32 target genes (included in the master list of DEGs) for these 14 drugs from the DrugBank website [113]. There were 20 edges between the target genes based on STRING-db [81]. Figure 3.6 shows a drug-target-PPI network constructed with the approved drugs. Gene weight, the sum of the weights of edges connected to each gene, in each stage-specific network, are shown beside target genes. Some important pathways involving target genes, such as PI3K-Akt or Ras signaling, are also highlighted in the figure. We can see that the weight of different genes changes across the four stages extensively. log2FC (with respect to normal) for genes also changes albeit to a lesser degree.

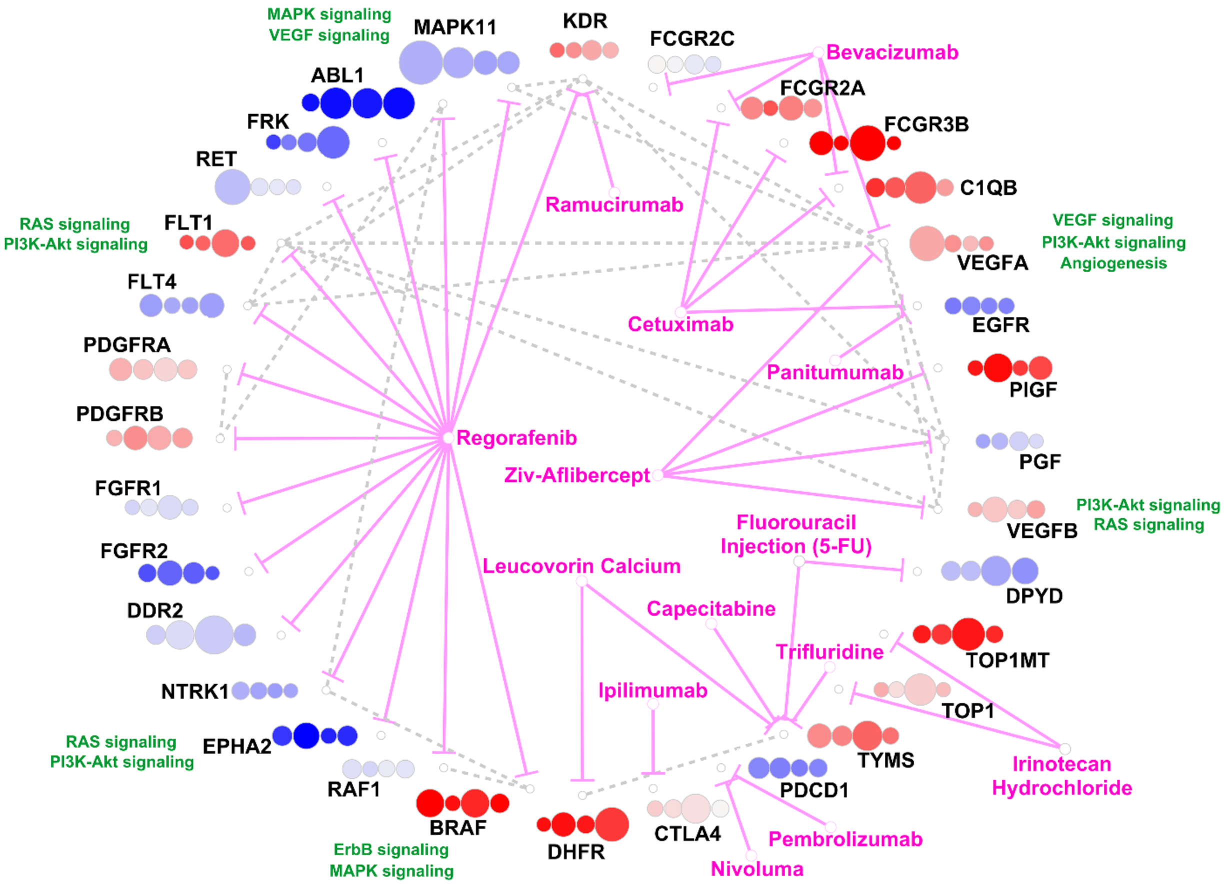


Figure 3.6: Drug-target-PPI network for CRC. Fourteen drugs approved by FDA for treating CRC (mainly when the cancer metastasizes) are used to construct this network; the drug nodes are shown in the center area. The target genes have been found from DrugBank. For each target gene node, four circles are associated with that gene corresponding to four stages I-IV, and are colored based on the log2FC values between a stage and normal (dark blue (log2FC of -1) to white (0) to dark red (log2 FC of 1)). The size of each circle represents the sum of the weights of edges connected to that gene (i.e., gene weights) in each stage. For example, *MAPK11* weight is greater in stage I as compared to that in other stages. PPI edges from STRING-db (score threshold ≥ 0.9) are also incorporated in this network by dashed grey-lines between the genes. For select functionally important genes, the related functions are listed.

Several targets of Regorafenib, a popular CRC drug, were found to be differentially regulated within our networks (Figure 3.6). Studies have shown that Regorafenib targets kinases involved in tumor angiogenesis (e.g. *VEGFR1/2/3*, *FGFR1/2*), proliferation (e.g. *MAPK11*, *RET*), tumor microenvironment, and metastasis [114, 115]. It can also disrupt tumor immunity through inhibition of *CSF-1R*, important for macrophage differentiation and survival [116]. Out of its targets, *MAPK11* and *RET* were both down-regulated and had greater weights in early stages. *MAPK11* is a member of protein kinases family involved in several cellular processes, including cell proliferation or differentiation. It was also enriched for MAPK and VEGF signaling pathways. *RET*, as a member of the tyrosine protein kinases family, has been identified as a novel tumor suppressor gene in the colon which can reduce apoptosis and is considered as a target for CRC treatment [117, 118]. There were also some targets for Regorafenib with larger weights in later stages, such as *FLT1* and *DDR2*. *FLT1*, a member of the vascular endothelial growth factor receptor (VEGFR) family, was up-regulated in CRC and strongly connected (PPI edge) to three ligands, namely, *VEGFA*, *VEGFB*, and *PGF* [119]. *DDR2*, down-regulated in CRC, is considered a critical regulator of cancer invasion and an attractive therapeutic target in metastatic CRC (mCRC) [120].

Two up-regulated and highly connected genes in this network, *VEGFA* and *VEGFB* are targets of the drug Ziv-Aflibercept, and participated in Ras and PI3K-Akt signaling pathways with known roles in CRC progression. *VEGFA* had larger weights in stage I whereas *VEGFB* had larger weights in stage II. *TYMS* (part of the Folate-mediated one-carbon metabolism pathway) is a crucial player of DNA methylation and repair and a critical target for Fluorouracil Injection (5-FU) drug, used in CRC treatment [121]. Studies have shown that *TYMS* is highly expressed in patients with CRC and might be used as a predictor for efficacy of chemotherapy [122]. Its weight was higher in stage III than in other stages. *TOP1* and *TOP1MT*, both up-regulated in CRC, had also greater weight in stage III and were targets for Irinotecan Hydrochloride which is one of the key drugs for the treatment of mCRC [123].

Besides Regorafenib, two other drugs, Cetuximab and Bevacizumab, commonly used in treating CRC also showed several targets enriched within our networks. *C1QB*, a target for both of those drugs, was up-regulated with greater weights in stage III. Cetuximab blocks ligand-induced receptor signaling and modulates tumor-cell growth by binding to the extracellular domain of *EGFR*. Studies have also shown that Cetuximab improves overall survival and progression-free survival and preserves quality-of-life measures in CRC patients in whom other treatments have failed [124]. Bevacizumab, which binds to and targets *VEGF*, also has demonstrated improved overall survival for patients with mCRC [125].

The pathogenesis of CRC is yet to be fully understood. In this study we detected a few potential biomarkers which were further validated in-silico, using a large cohort database (TCGA COAD-READ). However, further experimental validation is required to decipher their pathology-associated mechanisms. Additionally, we were limited by the unequal number of patient samples across stages and lacked sufficient clinical metadata to support downstream survival analysis. Nevertheless, the modular-network-based approach presented in this work will be useful for understanding mechanisms for disease progression and may contribute to identifying potential targets for disease intervention. In addition, while digital sequencing data are more robust, this microarray analog gene expression data set has been used extensively and our quest was to explore topological network analyses to demonstrate the ability to obtain stage-specific biomarkers and mechanisms. We demonstrate the validity of our conclusions through extant results and additional analyses.

3.5 Conclusion

In this study, we utilized a published transcriptomic data from 128 patients at various stages of CRC to find modular mechanisms potentially causal for progression of CRC from normal to stages I-IV and to find stage-specific biomarkers. We constructed stage-specific networks and identified their communities using the Louvain algorithm. Comparing communities of different networks at the topological and functional levels revealed that neighboring stages were more similar to each other than non-neighboring stages. We also carried out the functional analysis at the whole network level for the stage-specific and stage-unique networks by analyzing the enrichment of 24 cancer-related pathways across different stages. For the stage-specific networks, most of the pathways related to CRC such as PI3K-Akt and MAPK signaling pathways were enriched at all stages. However, stage-unique networks revealed functional differences across the stages. For example, MAPK signaling pathway was enriched across stages I-III and Notch signaling pathway (important for metastasis and tumor angiogenesis) was enriched in stages III and IV. We then identified key biomarkers to differentiate between CRC (any stage) and normal using STEM analysis. *WNT2* and *SFRP2* were two biomarkers validated by others in stool DNA and were over-expressed and under-expressed in CRC tissues, respectively. To incorporate legacy knowledge in our analysis, we performed a supervised analysis with 10 key genes related to CRC and their first neighbors based on STRING-db, across different stages. The subnetworks were analyzed to study the progression of cancer across stages. In particular, we identified that *BRAF*, a Ser/Thr kinase that activates MAP kinases, appeared in all subnetworks and was upregulated in stages I-IV as compared to normal. Its connectivity pattern changed across the subnetworks for normal and different stages of CRC. Finally, we constructed a drug-target-PPI network enabling us, in the light of present data, to understand the functional mechanisms for some of the current drugs for CRC treatment. We saw that the target gene weights changed across the four stages extensively. For example, *TYMS*, associated with folate-mediated one carbon metabolism and a target for some drugs such as Fluorouracil Injection (5-FU) and Capecitabine, was found to be upregulated in cancer stages with larger weights in stage III than in other stages.

3.6 Acknowledgement

Chapter 3, in full, is a reprint of the material as it appears in “Modular and mechanistic changes across stages of colorectal cancer” by Rahiminejad, Sara; Maurya, Mano R.; Mukund, Kavitha; and Subramaniam, Shankar., BMC Cancer, 2022. The dissertation author was the primary investigator and author of this paper.

Funding for this study was provided by the National Science Foundation grant CCF0939370 and by the National Institutes of Health (NIH) Grants U01 DK097430, U01 CA200147, U01 CA198941, U19 AI090023, R01 HL106579, R01HL108735, R01 HD084633, R01 DK109365, R01LM012595, and OTA OD030544.

3.7 Supplementary Methods

t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a nonlinear dimensionality reduction algorithm. It maps multi-dimensional data to a few (two or more) dimensions, which can be easily visualized [76].

The algorithm comprises three steps as follows:

1. In the 1st step, t-SNE measures similarities between points in the high dimensional space. For each data point, x_i, the algorithm centers a Gaussian distribution over that point and measures the density of all points under the distribution and renormalizes the densities. This gives a set of probabilities, p_{ij}, proportional to the similarity and is calculated by Equation S3.1 [76].

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma^2)}{\sum_{k\neq l}\exp(-\|x_k - x_l\|^2/2\sigma^2)} \tag{S3.1}$$

   where x’s are the points in the high-dimensional space and σ is the variance of the Gaussian distribution.

2. This step is similar to step 1 but here, the algorithm uses a student-t distribution rather than a Gaussian distribution to compute the similarities, q_{ij}, in the low-dimensional space (two or more). Mathematically, q_{ij} is given by.

$$q_{ij} = \frac{\exp(-\left\|y_i - y_j\right\|^2/2\sigma^2)}{\sum_{k\neq l}\exp(-\|y_k - y_l\|^2/2\sigma^2)} \tag{S3.2}$$

where y's are the points in the low-dimensional space.

3. In the last step, the algorithm matches the two probability distributions in the high and low dimensional spaces, by minimizing the Kullback-Leibler (KL) divergence. The KL divergence (cost function) is given by Equation S3.3 and the algorithm uses gradient descent to minimize it [76].

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{S3.3}$$

Normalized Mutual Information (NMI) metric

NMI is a metric to calculate the similarity between two groups. It is the normalized form of Mutual Information (MI). MI measures similarity between two methods (stages) and is given by [126]:

$$MI(A,B) = \sum_{i=1}^{c_A} \sum_{j=1}^{c_B} \frac{N_{ij}}{N} \log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right) \tag{S3.4}$$

where, *A* and *B* are the stages being compared. Denote by $N_{i.}$ the number of nodes (genes) in community *i* of stage *A* and $N_{.j}$ the number of nodes in community *j* in stage *B*. N_{ij} is the number of nodes in both community *i* of stage *A* and community *j* of stage *B*.

Then, NMI between stages *A* and *B* is calculated as [59]:

$$NMI(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}\frac{N_{ij}}{N}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{c_A}N_{i.}\log\left(\frac{N_{i.}}{N}\right)+\sum_{j=1}^{c_B}N_{.j}\log\left(\frac{N_{.j}}{N}\right)} \tag{S3.5}$$

Community detection algorithm, Louvain

As described in the main manuscript, *Louvain* algorithm detects communities in networks by maximizing modularity, calculated by Equation S3.6:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{S3.6}$$

where A_{ij} is the weight of the edge between node i and j (is equal to 1 when all edges have the same weight), k_i is the sum of the weights of the edges attached to node i or degree of node i, c_i is the community to which node i belongs to, and the δ function is defined as $\delta(u,v) = 1$ if $u = v$ and 0 otherwise. m is the total number of edges in an unweighted network and the sum of the weights of all edges in a weighted network.

The algorithm is divided in two phases, which are repeated iteratively. First phase is to assign a different community to each node of the network. So, in the beginning, there are as many communities as there are nodes. Then, the gain of modularity (Equation S3.6) is calculated for removing node i from its community and placing it in one of its neighboring communities. The gain of modularity in moving node i into a community C can be computed by:

$$\Delta Q = \left[\frac{\Sigma_{in} + k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{S3.7}$$

where Σ_{in} is the sum of the weights (or count for unweighted networks) of the edges inside C, Σ_{tot} is the sum of the weights of the edges incident to nodes in C, k_i is the sum of the weights of the edges incident to nodes i (degree of i), $k_{i,in}$ is the sum of the weights of the edges from i to nodes in C. If the gain is positive, the node i is placed in the community for which the gain is maximum. This process is applied repeatedly for all nodes until no further improvement can be achieved.

The second phase is to build a network whose nodes are now the communities detected in the first phase. In order to perform that, the weights of the edges between the new nodes are given by the sum of the weights of the edges between nodes in the corresponding two communities. Edges between nodes of the same community result in self-loops for this community in the new network. When this phase is completed, the first phase of the algorithm is reapplied to the new network. The combination of these two phases is referred to as a "pass". The passes are iterated until a maximum of modularity is reached [20].

STEM algorithm

Short Time-series Expression Miner (STEM) is an algorithm designed for clustering short time expression data. First, the algorithm selects a set of potential profiles and then genes are assigned to the profile that best represents their trend among the pre-selected profiles.

To define a set of model profiles, the user must specify a parameter c that controls the amount of maximum change a gene has between successive time points. For example, if c is 2, then a gene can go up either one or two units, stay the same or go down either one or two units between successive time points. So, for n time points, this strategy would give $(2c + 1)^{n-1}$ distinct profiles. Since most of these profiles are likely to be sparsely populated, a subset of them, R, has to be selected, such that the minimum distance between any two profiles in R, namely p_1 and p_2 is maximized. This can be formulized as:

$$\max_{R \subset P, |R|=m} \min_{p_1, p_2 \in R} d(p_1, p_2) \qquad \text{(S3.8)}$$

where P represents the total set of possible profiles, d is a distance metric, and m is the size of R (i.e., $|R| = m$). A greedy algorithm is used to calculate R. The algorithm starts with one of the two extreme profiles and in each iteration, selects the profile that is farthest from all profiles located in R so far. This process is repeated until m profiles have been selected.

In the next step, each gene $g \in G$ is assigned to a model profile m_i in the set of profiles M, such that $d(e_g, m_i)$ is the minimum over all m's, where e_g is the temporal expression profile for gene g. If the above equation is minimized by more than one profile ($h > 1$), then g is assigned to all those profiles, but the assignments is weighted as $1/h$. Further details about the algorithm can be found elsewhere [71, 72].

3.8 Supplementary Figures

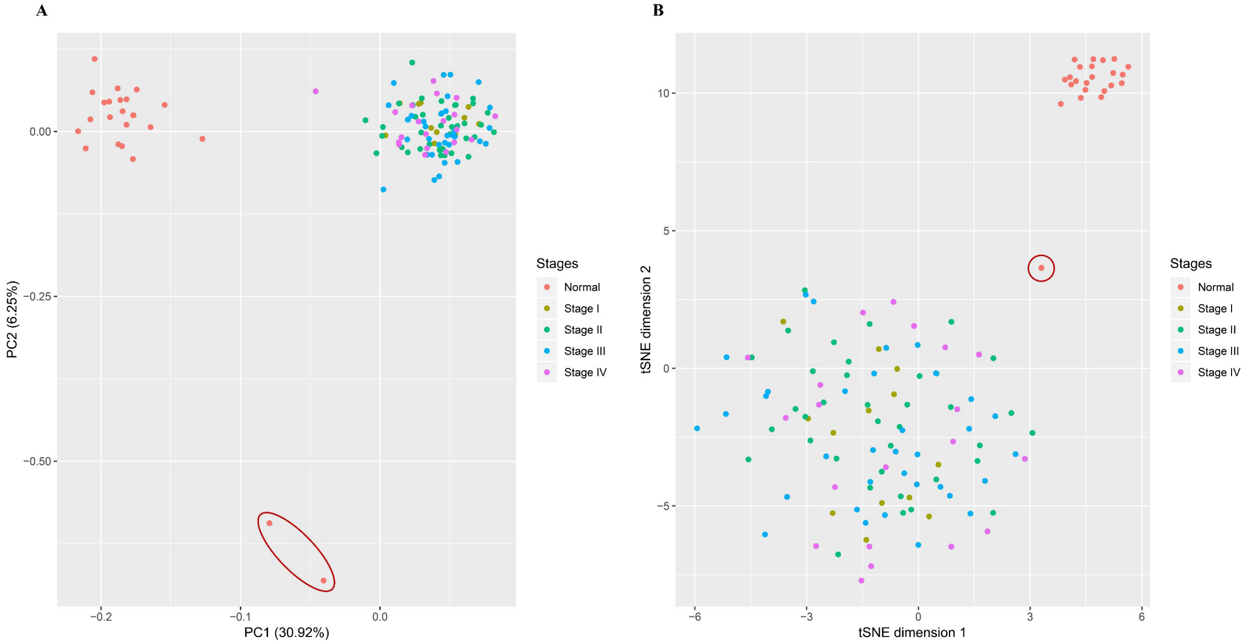


Figure S3.1: Dimensionality reduction techniques applied on 41834 probe IDs and 128 samples including 13 patients in stage I, 37 patients in stage II, 34 patients in stage III, 20 patients in stage IV, and 24 normal samples. **A** PC1 vs PC2. There are two outliers in normal, in a red circle. **B** First two dimensions of t-SNE method. There are two outliers in normal which are very close to each other, highlighted in a red circle. These two outliers are similar to the outliers found by PCA analysis.

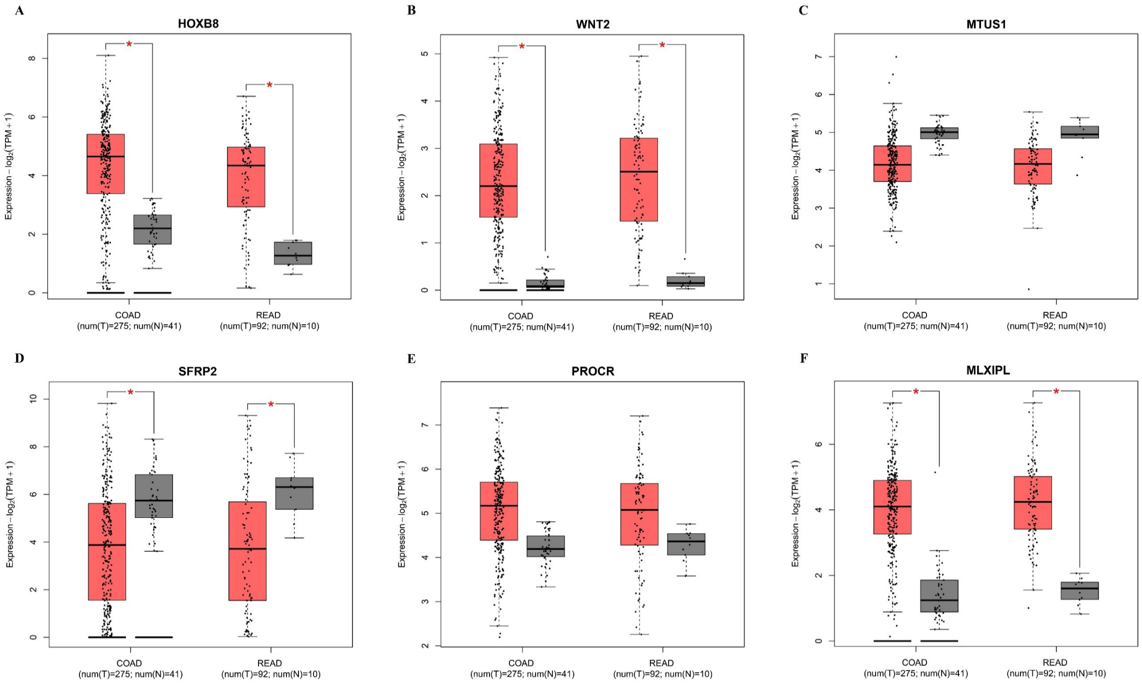


Figure S3.2: Gene expression boxplots of 6 biomarkers using TCGA COAD-READ cohort through GEPIA2. **A-D** Boxplots for 4 biomarkers from STEM analysis. **E-F** Boxplots for 2 stage-specific biomarkers. Red rectangles represent tumor and gray rectangles represent normal in each boxplot.

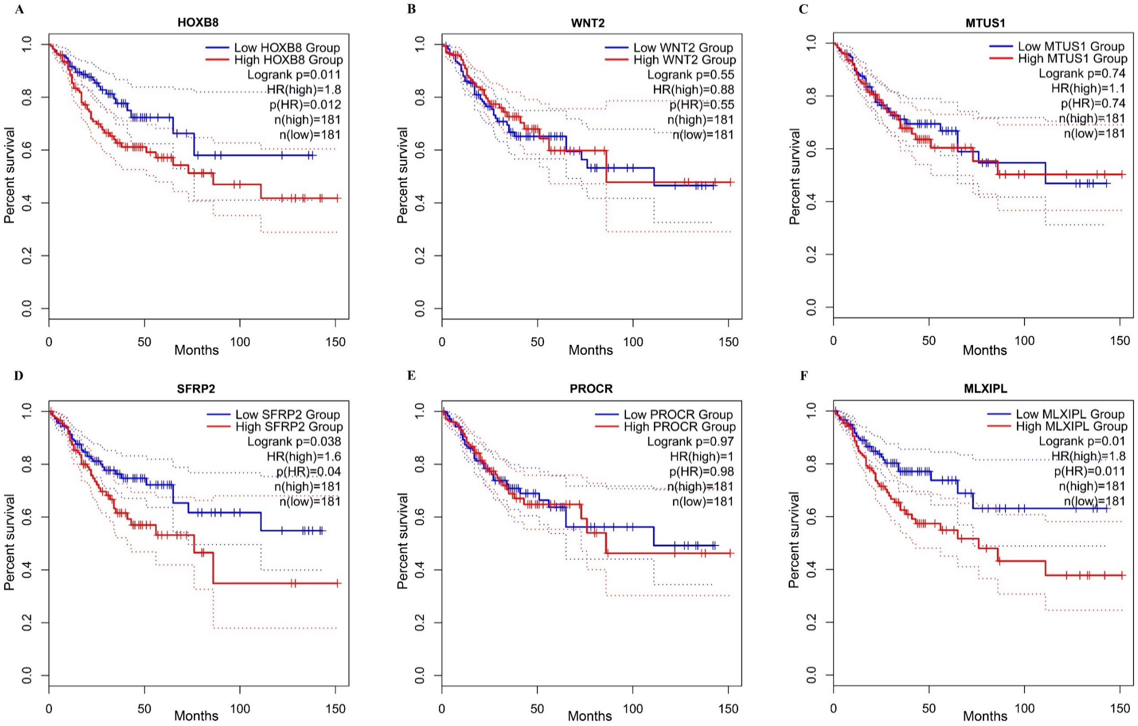


Figure S3.3: Kaplan-Meier curves for Disease-Free Survival (DFS) using TCGA COAD-READ cohort through GEPIA2. **A-D** DFS plots for 4 biomarkers from STEM analysis. **E-F** DFS plots for 2 stage-specific biomarkers. Red lines represent the samples with highly expressed genes and blue lines represent the samples with lowly expressed genes. Dotted line shows a 95% confidence interval. HR: hazard ratio.

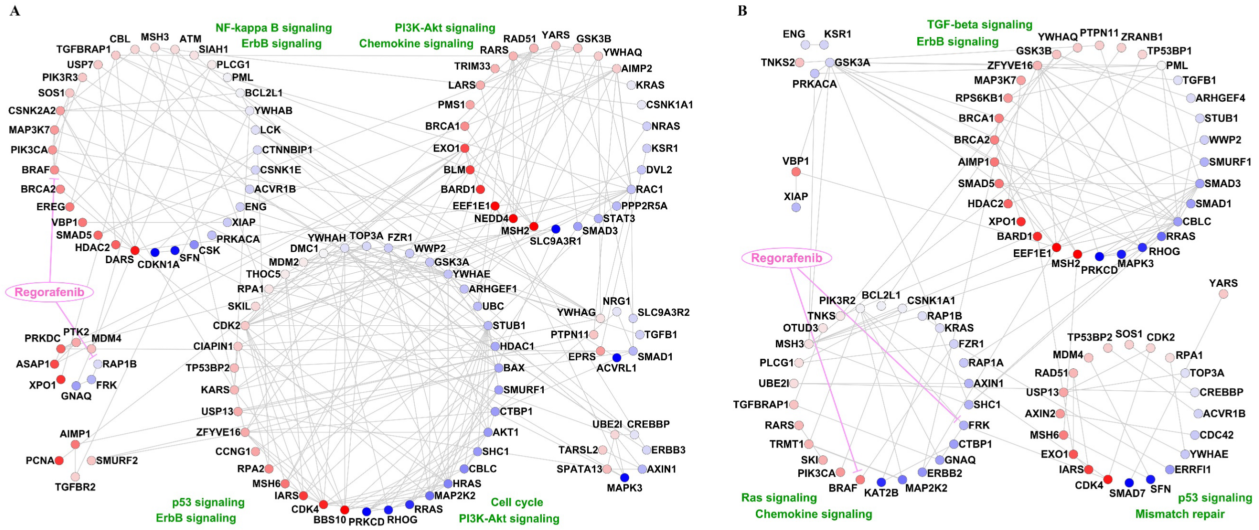


Figure S3.4: Subnetwork of 162 genes in **A** stage III-specific network and **B** stage IV-specific network. Nodes of each subnetwork are grouped together based on the communities they belonged to in the stage-specific networks and colored based on the value of log2FC between that stage and normal: dark blue (log2FC of -2), to white (0) to dark red (2). The width of edges shows the strength of connections based on PCC between them. The thicker the edges are, the larger the PCC between the nodes is.

Chapter 4 ANALYSIS OF GENE MODULES ACROSS STAGES OF COLORECTAL CANCER FROM SINGLE CELL TRANSCRIPTOMICS

4.1 Abstract

Background: The underlying mechanisms of how colorectal cancer (CRC) develops, progresses through stages pT1-pT4, and its manifestation in the left and right colons are poorly understood. The focus of this study is to analyze a single cell transcriptomic dataset of CRC from a modular perspective towards understanding mechanisms. Methods: Using previously published data (GEO-GSE178341), we parsed cells by stage and down-sampled cells at stages pT2-pT4 to make cell counts equal across different stages w.r.t pT1. We also down-sampled cells from the right colon to have the same cell count as the left colon. Using Weighted Gene Co-expression Network Analysis (WGCNA), we identified and functionally analyzed gene modules from the early stage (pT1) and the right colon that were not/lowly preserved in late stages (pT234) and the left colon, respectively. We also calculated Spearman's rank correlation (ρ) for gene degrees of non-preserved modules. Finally, we validated our results from this dataset with those from another scRNA-seq dataset on CRC. Results: All stages (pT1-pT4) had 7,540 cells after down-sampling. Modules tan and greenyellow of the early stage were found to be non-preserved in late stages. Both modules had different connectivity patterns between the early and late stages with low values for ρ. Module tan captured myeloid cells, the most abundant cell population in the tumor microenvironment, with genes enriched for cytokine-cytokine receptor interaction and signaling pathways. Module greenyellow was mostly enriched for transcriptional misregulation in cancer and bile secretion pathways. Both right and left colons had 46,379 cells after down-sampling. Modules greenyellow (ρ = 0.66) and purple (ρ = 0.69) of the right colon were non-preserved in the left colon. Module greenyellow was enriched for transcriptional misregulation in cancer and neuroactive ligand-receptor interaction pathways. Module magenta (mostly B cells) was enriched for B-cell receptor signaling and intestinal immune network for IgA production. Conclusions: We highlight topological and functional differences existing between the early and late stages as well as the right and left colons in CRC, using a co-expression network theoretic approach on scRNA-seq dataset. The non-preserved modules were enriched for important broad functions such as immune function and transcriptional misregulation in cancer.

4.2 Background

Despite the substantial progress in diagnosis and treatment of cancers, it continues to be one of the leading causes of death worldwide. Of all cancer types, colorectal cancer (CRC), as a term used to describe both colon and rectal cancers, is the second most common cause of cancer death in the US. It has been estimated that approximately 153,020 individuals will be diagnosed with CRC in 2023 and 52,550 will die from the disease [127]. CRC is a multifactorial disease involving genetic, environmental, and lifestyle risk factors. Although there are hereditary and non-hereditary types, the majority are non-hereditary and caused by somatic mutations in response to environmental factors [128].

CRC is staged using the American Joint Committee on Cancer’s TNM system in which T (tumor) refers to the size of the primary tumor, N (node) describes the involvement of lymph nodes near the primary tumor, and M (metastasis) indicates whether the cancer has spread to other organs or not. The letter ‘p’ is sometimes used before the letters TNM to indicate pathological stage. A number from 0 to 4 is assigned to each factor, with the higher number indicating increasing severity. Stage 0 is where the cancer cells are contained to the rectum's or colon's inner lining. In stage 1, the cancer cells are found in deeper layers of the colon or rectum wall, but have not spread beyond the wall. In stage 2, the cancer cells have not spread to the lymph nodes, but may have spread through and beyond the wall of the colon or rectum (sometimes into nearby tissues or organs too). In stage 3, the cancer cells have spread to one or more nearby lymph nodes and colon or rectum wall. In stage 4, which is also called metastatic CRC (mCRC), the cancer cells have spread beyond the lymph, colon or rectum to distant areas of the body, including other tissues and/or organs. According to the location of the tumor, CRCs are divided into the right-sided or proximal tumors (including the cecum, ascending colon, and the hepatic flexure), the left-sided or distal tumors (including the splenic flexure, descending colon, sigmoid colon, rectosigmoid, and rectum), and the transverse colon. Left-sided CRC accounts for two-thirds of CRCs as they are likely diagnosed earlier, and found to occur more commonly in men and in younger patients, whereas right-sided CRC occurs more frequently in women and older ages [129, 130].

The relationship between stage and outcome is evident: the higher the tumor stage, the shorter the survival time. Most patients with CRC are already in the advanced stage of diagnosis and tend to relapse after first-line treatment and chemotherapy. The 5-year survival rate for CRC has improved from 50% during the mid-1970s to 65% for patients diagnosed during 2011 through 2017, reflecting both earlier diagnoses and advances in surgical techniques and therapies. However, the 5-year overall survival rate of metastatic CRC (mCRC) is only ~4-12%. Tumor sidedness has been shown to affect clinical outcomes as well. Therefore, there are ongoing studies investigating more efficacious treatments for CRC [131].

In recent years, single-cell RNA sequencing (scRNA-seq) has contributed significantly to quantify the transcriptome status of tumor tissue at a single-cell resolution. scRNA-seq helps to gain better analysis and understanding of the different/rare cell types [132, 133], to detect the genetic information as well as the difference between gene expression in individual cells [134], and to describe the regulatory networks and developmental trajectories [135]. In CRC, single-cell omics analysis of the genome, transcriptome, and epigenome have elucidated the diversity within and between tumors. Dai et al. utilized scRNA-seq analysis to profile cells from cancer tissue of a CRC patient to provide more insight into the heterogeneity of the cell populations affected by the disease [136]. Willems et al. utilized scRNA-seq data to improve CRC survival prediction [137].

The current study aimed to analyze CRC tissues and investigate the events occurring at early stage (pT1) compared to late stages (stages pT2, pT3, and pT4 combined as pT234) and in the right-sided colon compared to the left-sided colon. To this end, we utilized a scRNA-seq dataset on CRC from Gene Expression Omnibus (GEO-GSE178341) containing 370,115 cells from 62 cancerous patients and adjacent normal tissues. Due to an unequal number of cells at different stages (and different sides), we first down-sampled each stage separately to make the number of cells equal. Then, we analyzed the data at the cell level using the *Seurat* package in R. Weighted Gene Co-expression Network Analysis (WGCNA) was employed to analyze the data at the gene level. We identified and functionally characterized gene modules that were not/lowly preserved using module preservation analysis between pT234 and pT1 [138, 139]. Similarly, we identified and compared gene modules from the right vs. left colon. Finally, we compared the broad results from this dataset with those from another scRNA-seq dataset on CRC. Figure 4.1 shows a flow chart for our analysis pipeline.

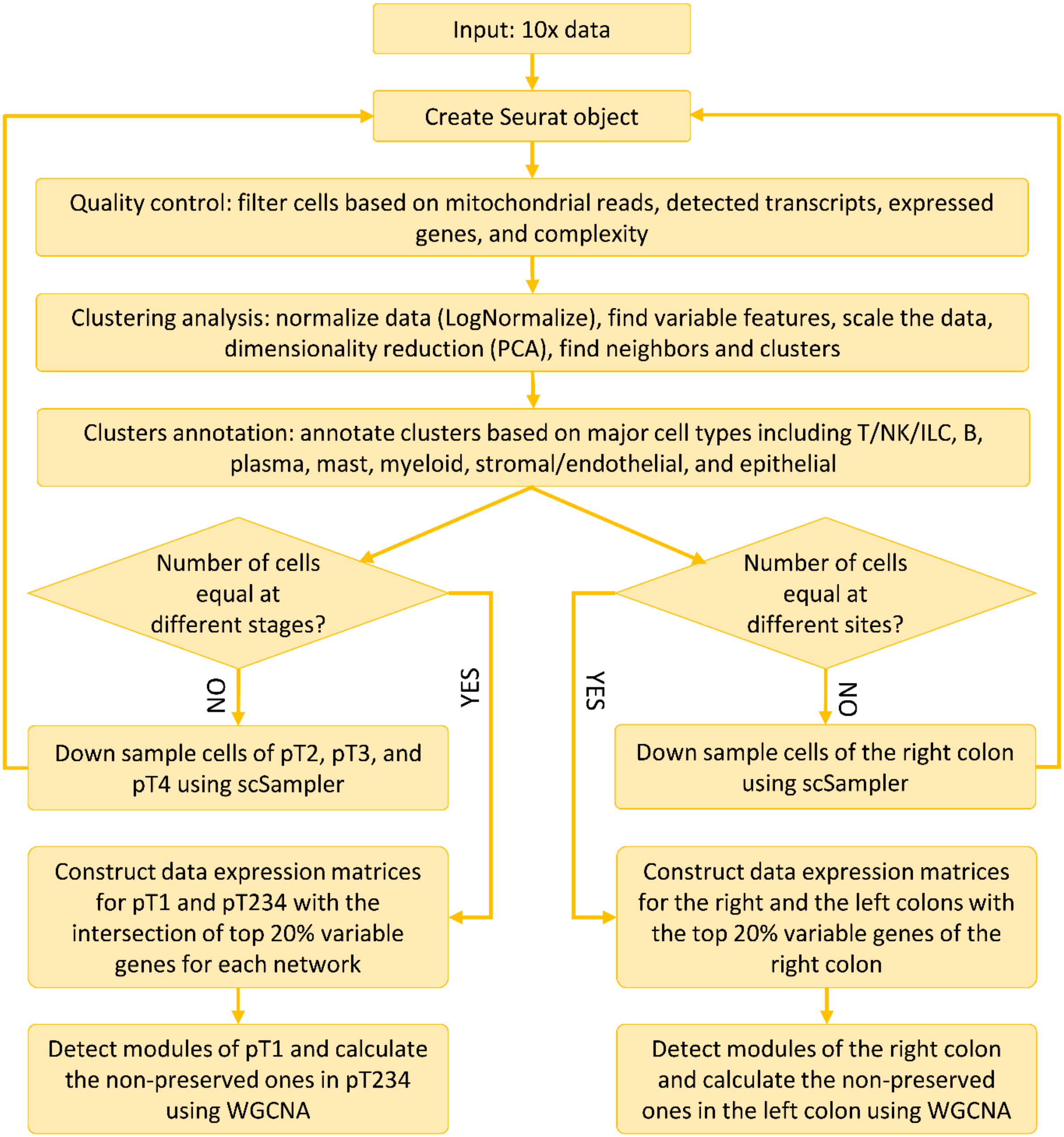

Input: 10x data
Create Seurat object
Quality control: filter cells based on mitochondrial reads, detected transcripts, expressed genes, and complexity
Clustering analysis: normalize data (LogNormalize), find variable features, scale the data, dimensionality reduction (PCA), find neighbors and clusters
Clusters annotation: annotate clusters based on major cell types including T/NK/ILC, B, plasma, mast, myeloid, stromal/endothelial, and epithelial
Number of cells equal at different stages?
Number of cells equal at different sites?
NO
YES
YES
NO
Down sample cells of pT2, pT3, and pT4 using scSampler
Down sample cells of the right colon using scSampler
Construct data expression matrices for pT1 and pT234 with the intersection of top 20% variable genes for each network
Construct data expression matrices for the right and the left colons with the top 20% variable genes of the right colon
Detect modules of pT1 and calculate the non-preserved ones in pT234 using WGCNA
Detect modules of the right colon and calculate the non-preserved ones in the left colon using WGCNA

Figure 4.1: Flow chart of the approach used in our analysis.

4.3 Materials and Methods

Single-cell RNA sequencing data

In this study, we analyzed publicly available scRNA-seq data from a study on single cell atlas of mismatch repair-deficient (MMRd) and mismatch repair-proficient (MMRp) colorectal cancer (GEO accession ID GES178341). The data included 370,115 cells from 62 patients at different pathological stages of cancer (pT1 for early invasive cancer through pT4 for growth of the tumor through the outer layer of the bowel wall) and 36 adjacent normal tissues. Of the total number of cancerous cells, there were 62,654 cells from the left colon and 194,597 cells from the right colon. We utilized the R *Seurat* package for processing the data [140, 141, 142, 143].

Data Filtering

The data was filtered at the cell level by excluding cells based on the following criteria: more than 15% mitochondrial reads, less than 300 detected transcripts, less than 500 expressed genes, and a complexity (log10 genes/UMI) of less than 70%. Gene level filtering was also performed by excluding genes expressed in less than 10 cells.

Data scaling and dimensionality reduction

The data was normalized using the *LogNormalize* method and the scale factor of 10,000. Top 20% variable genes were then calculated and the data was scaled. Principal Component Analysis (PCA) was used on the data for dimensionality reduction, with npcs = 40. The optimal dimension for clustering was selected using the following approach:

1. Determine the percentage of variance (PV) captured by each PC.

2. Calculate cumulative percentage (CM) for each PC.

3. Find PCs which have CM > 90% and PV < 5% variation.

4. Calculate the difference between variation of each PC and subsequent PC which had percentage change in variation < 0.1%.

5. Select the minimum value from steps 3 and 4 as the optimal number.

Uniform Manifold Approximation and Projection (UMAP) was performed on the optimal number of dimensions, with the cells clustered using the Louvain algorithm (*FindClusters* function in *Seurat* with resolution 2).

Cells down sampling using scSampler

Due to the different number of cells at different stages, stage 1 (pT1) with the lowest number of cells was selected as the reference and cells of other stages were down-sampled using *scSampler* [144], implemented in python, to have the same number of cells as in pT1. Clustered cells (detected in the previous step) were annotated based on major cell types, namely, T/natural killer (NK)/innate lymphoid cell (ILC), B, plasma, mast, myeloid, stroma/endothelial, and epithelial cells. Then, major cell types of each stage (pT2, pT3, pT4, and normal) were down-sampled in such a way that the sum of all cells in each stage was equal to the number of cells in pT1 and the (ratio) relative count of cells from different cell types was maintained for each stage as before down-sampling. After down-sampling, a subset of the original data with the down-sampled cells and all genes were selected and the previous steps (data scaling and dimensionality reduction) was applied on it.

Module detection and module preservation

We utilized the WGCNA R package [138] to find gene modules. In the first comparison, the early stage (pT1) was set as the reference network and late stages (pT2, pT3, and pT4 together referred to as pT234) were considered as the test network. To detect the most variable genes between the two networks, we selected the top 20% variable genes of the early and late stages separately and calculated their intersection as the list of variable genes. A ‘cell x gene’ expression matrix was then constructed for each network and the gene modules of the reference network were found using the *blockwiseModules* function of WGCNA. The preservation statistics of the detected modules were calculated w.r.t late stages. Non -preserved modules were selected based on two criteria; median rank ≥ 10 and module size ≤ 100.

In the second comparison, the right colon was selected as the reference network and the left colon as the test network. Other steps are similar to the first comparison except that here the top 20% variable genes of the reference network were used as the list of variable genes.

Change of non-preserved modules between early and late stages

Non-preserved modules were selected based on the median rank plot and visualized in a circle plot. A co-expression network was constructed for genes of non-preserved modules in reference (pT1) and test (pT234) networks separately. Edge weight in the circle plot represents the co-expression value between two genes and node (gene) degree is the sum of edge weights connected to that node. Nodes are colored based on the z-score between reference and test networks, changing from blue (-2 or lower) to white (0) and to red (2 or greater). Functional analysis on the modules or a subset of the nodes (genes) was carried out using Enrichr [145, 146, 147].

Spearman's rank correlation

To check the similarity/dissimilarity of non-preserved modules between two networks, we calculated Spearman's rank correlation coefficient (ρ) [148]. We first ranked the nodes of two networks based on their degrees. Then we used Equation 4.1 to calculate the correlation value, ρ. The lower (larger) the value, the less (more) similar the modules are between the two networks.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4.1}$$

where d_i is the difference between node degrees in two networks and n is the total number of genes.

Normalized two sample z-test for genes

z-score for a gene between the early and late stages was calculated using two sample z-test.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{4.2}$$

in which $\bar{x}_1$ and $\bar{x}_2$ are the sample mean of first sample and second sample respectively, μ_1 and μ_2 are the mean of first and second population, σ_1^2 and σ_2^2 are the population variance in first and second population, and n_1 and n_2 are the sample size of first and second groups respectively. Since the null hypothesis is $\mu_1 = \mu_2$, the equation will be:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

4.4 Results and Discussion

4.4.1 Down-sampling to Equalize the Number of Cells at Different Stages

The CRC dataset used in this study contained different numbers of cells at different stages (see Table 4.1). Since the number of cells in pT1 was less than the others, we selected pT1 as the reference and down-sampled cells of other stages to have the same number as in pT1. As described in the Materials and Methods, seven major cell types, i.e., T/natural killer (NK)/innate lymphoid cell (ILC), B, plasma, mast, myeloid, stroma/endothelial, and epithelial cells were selected for the down-sampling process (see Figures S4.1) [149]. Cell counts in each major cell type before and after down-sampling are listed in Table S4.1. The (ratio) relative count of cells from different cell types was maintained for each stage as before down-sampling. Figures 4.2A and 4.2B show UMAP plots for all cells before and after down-sampling, respectively. As seen from the figure, there was a good separation between major cell types.

Table 4.1: Properties of different stages

	# of samples	Tissue site	# of specimens	# of cells
Cancer	129	Left 62,654 cells	1 in pT1	4,502
			2 in pT2	10,956
			5 in pT3	24,790
			5 in pT4	22,406
		Right 194,597 cells	1 in pT1	4,677
			7 in pT2	33,571
			27 in pT3	90,138
			14 in pT4	66,211
Normal	52	Left	8	37,263
		Right	39	75,601

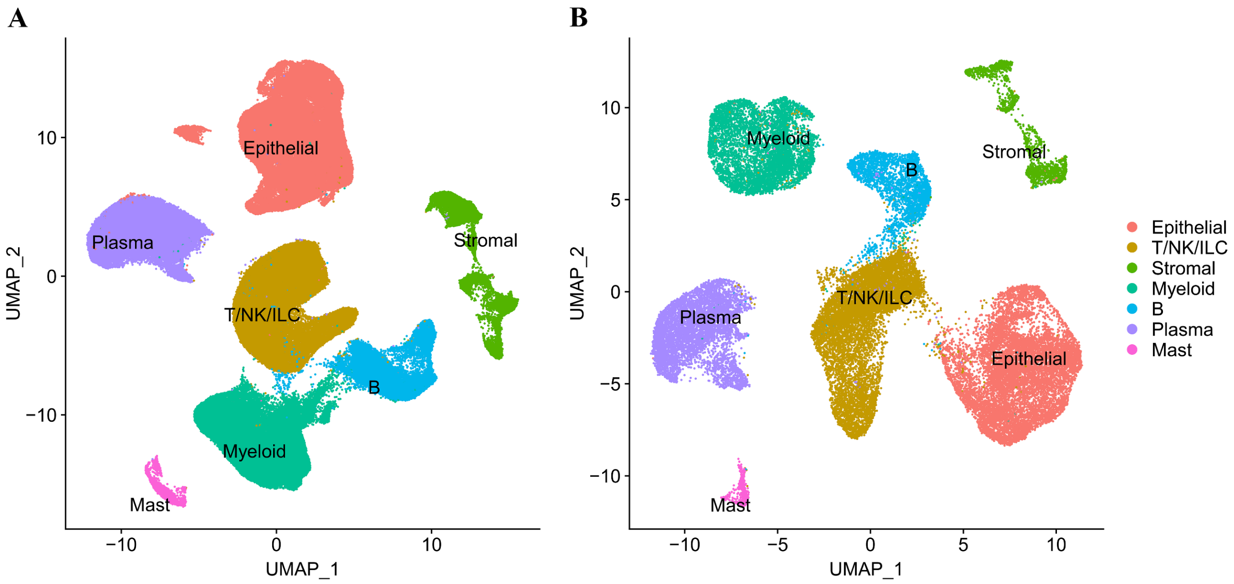


Figure 4.2: UMAP plot for all cells **A** before and **B** after down-sampling.

4.4.2 Early vs. Late Stages Comparison

After down-sampling, the early stage (pT1) had 7540 cells and late stages (pT234) had 22,620 cells. The intersection of top 20% of the most variable genes from the two networks yielded

3,107 unique genes. Using blockwiseModules function of WGCNA, we detected 13 modules for the early stage, out of which two modules, tan and greenyellow, were not preserved and magenta was the most preserved module in late stages based on the median rank plot (Figure 4.3). Figure 4.4 shows cell coverage of different modules in UMAP space. In the following section, we analyze non-preserved and preserved modules from a topological and functional perspective.

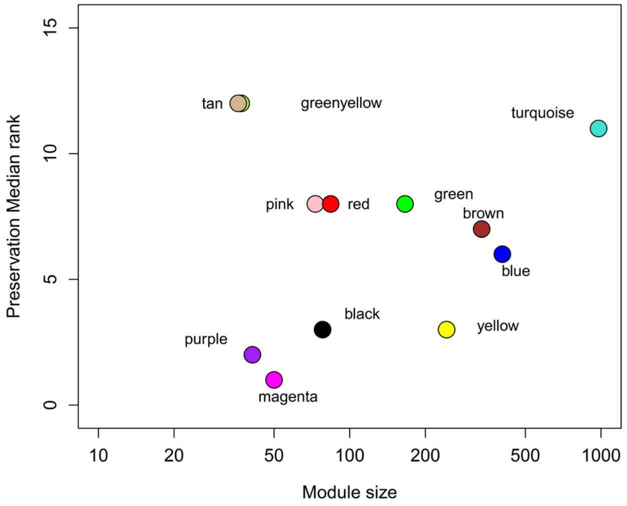


Figure 4.3: Median rank of module preservation between the early (reference) and late stages (test).

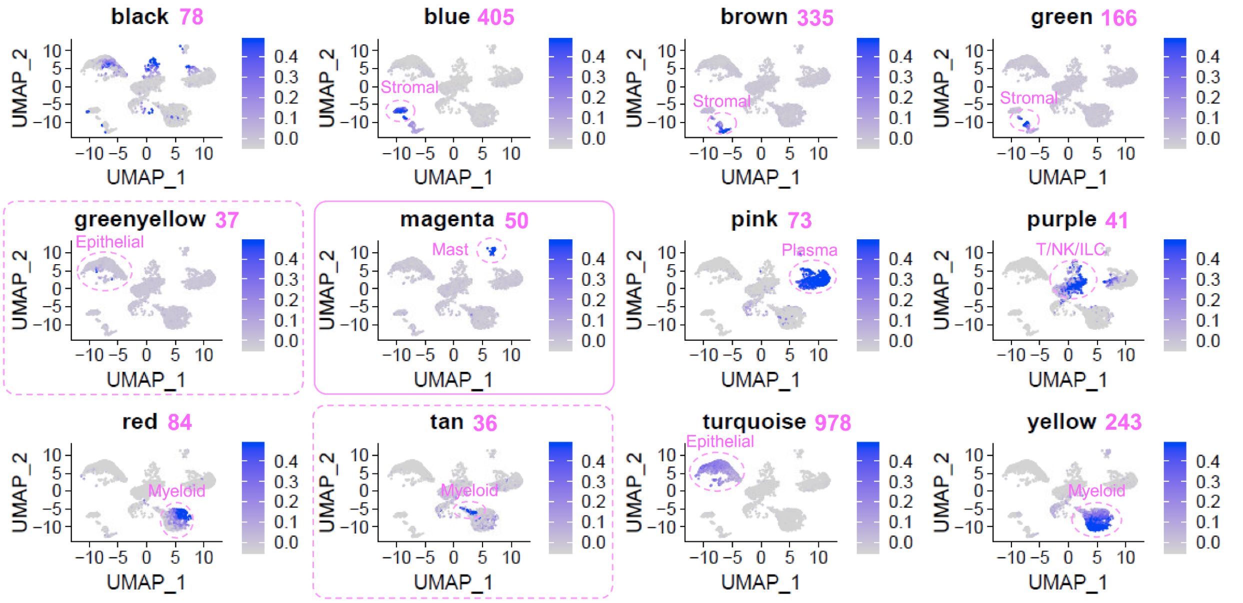


Figure 4.4: Cell coverage of different modules detected for the early stage (pT1). There is a number after each module's name representing the number of genes in that module. Non-preserved (or preserved) modules are highlighted by a dashed (or solid) purple rectangle around them.

4.4.2.1 Topological and functional differences of non-preserved modules between the early and late stages

As mentioned in the previous section, modules tan and greenyellow were non-preserved among all modules. Spearman's rank correlation (0 for the least similar and 1 for the most similar modules) also confirmed the non-preservation of gene degrees for these modules between the early and late stages (0.63 for module tan and 0.42 for module greenyellow). Figures 4.5A and 4.5B show top 10% of unique edges in the early and late stages for modules tan and greenyellow in circle plots, respectively. As seen from the figures, the connectivity pattern and the degree of genes changed from early to late stages. For both modules, the degree of genes and average weights of edges in the early stage was greater than late stages.

Module tan captured myeloid cells (see Figure 4.4) which are the most abundant cell population in the tumor microenvironment. Genes of module tan were mostly enriched for cytokine-cytokine receptor interaction, chemokine signaling, and NF-κB signaling pathways. It has been shown that inflammatory cytokines may promote tumor formation and enhance the progression of cancer from adenoma to invasive carcinoma [150]. Chemokines have been used as effective biomarkers in early diagnosis of CRC and they have contributed to making clinical decisions leading to improved survival [151]. NF-κB is hypothesized to promote tumorigenesis via pro-inflammatory factors, which can be activated by inflammatory cytokines [152]. The gene with the largest degree in the early stage of module tan (Figure 4.5A) was *CCL19* which is expressed abundantly in the T-cell zones, such as lymph nodes and thymus. It is also a vital regulator of immune responses, regulating the migration of DCs and T cells into secondary lymphatic tissues [153, 154]. The gene with the second largest degree was *FSCN1* which is an actin-bundling protein, oftentimes upregulated in different human cancers. In particular, *FSCN1* overexpression is known to promote cancer cell migration, invasion, and metastasis *in vitro* and *in vivo* [155]. *LAMP3*, the gene with the third largest degree, is known to participate in tumor metastasis and drug resistance with significant contribution to tumor cells proliferation, migration, and invasion [156].

Module greenyellow was mostly enriched for carbohydrate digestion and absorption, transcriptional misregulation in cancer, and bile secretion pathways. Carbohydrates lead to the proliferation of cancer cells through alterations in insulin levels and circulating glucose [157]. Transcriptional misregulation pathway is involved in the occurrence and development of CRC. *DEFA5* and *DEFA6*, two genes of the transcriptional misregulation pathway were down-regulated in the late stages as compared to the early stage and identified within this module (see Figure 4.5B). *DEFA5* was found to be closely related to colorectal adenocarcinoma and its high expression is also associated with better prognosis of CRC [158]. *DEFA6* was shown to have a promoting effect on the proliferation, migration, invasion, and colony invasion of CRC cell lines *in vitro* [159]. Bile acids are known to act as strong stimulators of the initiation of CRC by damaging colonic epithelial cells. They promote CRC progression through multiple mechanisms including apoptosis inhibition, enhancement of cancer cell proliferation, invasion, and angiogenesis [160]. Other markers that have been previously implicated with important roles in CRC were identified within this module. For example, *TTR*, whose degree in the early stage is higher than late stages, possesses cytokine functions to stimulate myeloid cell differentiation (known to play roles in the tumor environment) [161]. Likewise, *HEPACAM2* had a higher degree in the early stage compared to late stages. Studies have shown that it may be a diagnostic and prognostic biomarker for colon adenocarcinoma. *HEPACAM2* is also involved with immune response progress, chemokine signaling, Ras signaling and Hematopoietic cell lineage pathways [162]. *NEUROD1* is known to be highly expressed in CRC and its silencing induces the expression of p21, a master regulator of the cell cycle, leading to G2-M phase arrest and suppression of CRC cell proliferation and colony formation potential [163].

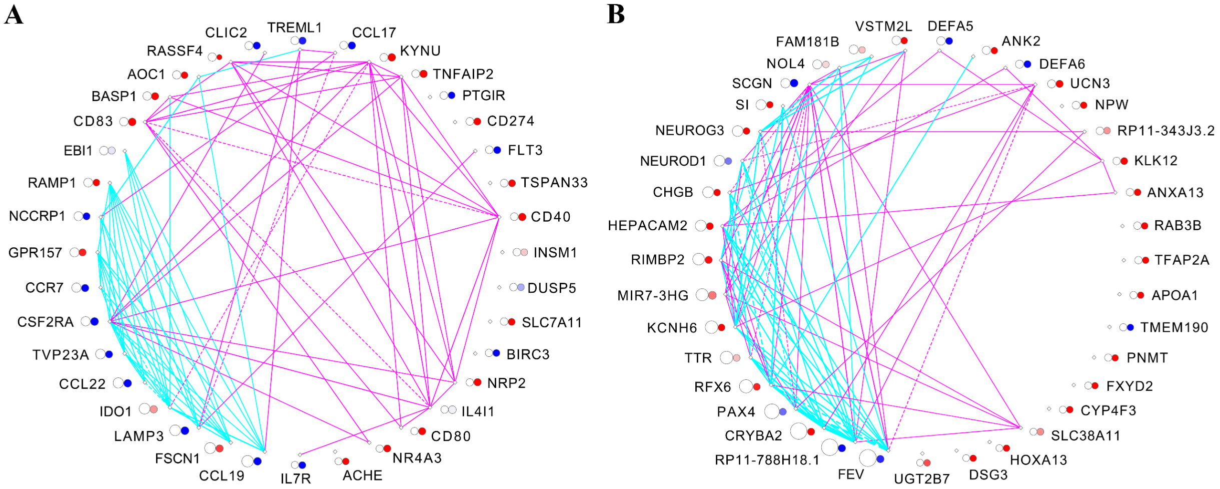


Figure 4.5: Circle plot for top 10% unique edges of module **A** tan and **B** greenyellow. Each cyan (or purple) edge represents the co-expression value between the two corresponding genes in the early stage (or late stages). The left (or right) circle near each gene represents the sum of weights of edges connected to that gene in the early stage (or late stages). Circles are colored based on the z-score between late stages and the early stage from blue (z-score ≤ -2) to white (0) to red (z-score ≥ 2). PPI edges are shown in dashed lines.

4.4.2.2 Mixed role of the preserved module in colorectal cancer

As seen from the median rank plot (Figure 4.3), module magenta was the most preserved module in the early vs. late stages comparison ($\rho = 0.88$). It was mostly enriched for renin-angiotensin metabolism, Fc epsilon RI signaling, and histidine metabolism pathways. It has been shown that the higher circulating levels of histidine, the lower the risk of colorectal cancer [164]. *MAOB* and *HDC*, two genes of the histidine metabolism pathway, had higher z-scores in late stages compared to the early stage. *IL5* and *IL13* were another two important genes enriched for Fc epsilon RI signaling pathway. Studies have shown that *IL5* has antitumor properties in CRC and it has been found elevated in mid CRC stages (pT2 and pT3) [165]. We also observed a higher z-score value for *IL5* in late stages compared to the early stage. Likewise, *IL13* was upregulated in late stages (higher z-score). It has been shown that *IL13* signaling could be involved early in intestinal stem cell self-renewal and homeostasis. It can also promote a tumorigenic microenvironment [166, 167]. Module magenta captured most of the mast cells. Although mast cells contribute to the transition from chronic inflammation to cancer, their exact role in tumor initiation and growth remains controversial. According to recent studies [168, 169] mast cell-derived mediators can either exert pro-tumorigenic functions and cause the progression and spread of the tumor or exert anti-tumorigenic functions limiting the tumor growth.

4.4.3 Right vs. Left Colon Comparison

Both right and left colons had 46,379 cells after down-sampling (see Table S4.2 for detailed cell counts). We constructed a data expression matrix with the top 20% variable genes (5,150 genes) of the right colon and found its modules using *blockwiseModules* of WGCNA. Of 12 modules found for the reference network (right colon), modules greenyellow and magenta were not preserved and module pink was preserved in the test network (left colon) (see Figure 4.6 for median rank plot). Spearman's rank correlation of gene degrees (0.66 for greenyellow and 0.69 for magenta) confirmed the non-preservation of these modules between the right and left colons as well. In the following section, we analyze non-preserved and preserved modules topologically and functionally.

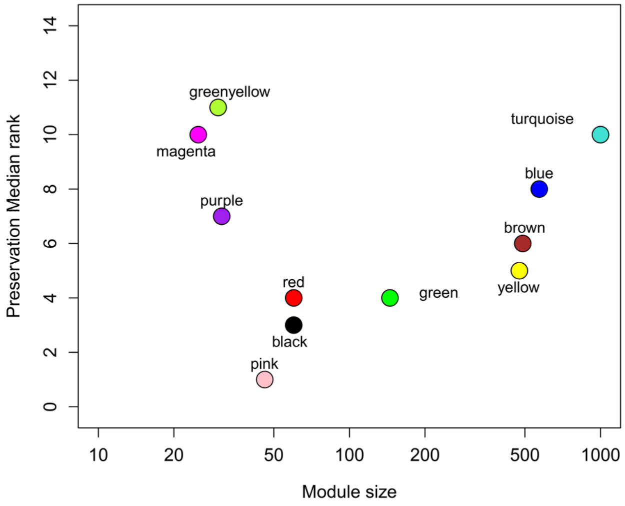


Figure 4.6: Median rank of module preservation between the right (reference) and left colon (test).

4.4.3.1 Different roles of non-preserved modules between the right and left colons

Module greenyellow was mostly enriched for transcriptional misregulation in cancer and neuroactive ligand-receptor interaction pathways based on KEGG and KRAS signaling down and epithelial mesenchymal transition based on MSigDB hallmark. Previous studies have shown that neuroactive ligand-receptor interaction pathways interact with the microenvironment cells and tumor cells in CRC and are significantly related to the development of the disease [170]. Epithelial mesenchymal transition (EMT), a collection of events during which cells lose their epithelial characteristics, is involved in the initial steps, progression and metastasis of CRC [171].

Module magenta captured B cells (see Figure S4.2) and it was enriched for B cell receptor signaling, hematopoietic cell lineage, and intestinal immune network for IgA production based on KEGG and notch signaling based on MSigDB Hallmark pathways. Hematopoietic cells play a central role in tumor growth and progression in the solid tumor microenvironment which could differentiate into hematopoietic cells [172]. Studies have shown that the intestinal immune network for IgA production was significantly enriched in the right-sided colon and the genes related to this pathway were expressed in CRC tumor tissues [173, 174].

In the following section, we analyze differences between the early and late stages for the non-preserved modules, i.e., modules greenyellow and magenta.

4.4.3.2 Changes in the connectivity pattern between the early and late stages for non-preserved modules

Figures 4.7 and 4.8 highlight the distinct differences in connectivity and edge weights for the two non-preserved modules, greenyellow and magenta, respectively. Degree of genes of module greenyellow did not change much between the early stage and late stages for the left colon (Figure 4.7A). However, in the right colon, genes had higher degrees and edge weights in the early stage compared to late stages (Figure 4.7B). Most genes of module greenyellow were up-regulated in late stages compared to the early stage in the left colon (Figure 4.7A), *i.e.*, higher z-scores, whereas in the right colon, some genes were up-regulated and some genes were down-regulated in late stages vs. the early stage (Figure 4.7B). Of 38 genes of module greenyellow, 17 genes were common with the module greenyellow of the previous analysis (the early stage vs. late stages). Some of these genes, such as *PAX4*, *HEPACAM2*, *TTR*, *DEFA5*, and *DEFA6,* have important roles in CRC. *PAX4*, the node with the largest degree in the right colon and enriched in the "KRAS signaling down" pathway (or gene set), is known to play an important role in the proliferation of CRC cells. Studies suggest a potential therapeutic role for *PAX4* inhibition in limiting cancer cell growth [175]. KRAS mutations are also associated with right-sided colon tumors [176]. *HEPACAM2*, as noted in our earlier comparison, was also present in this comparison (both in the early stage vs. late stages and the right vs. the left colon). It has been shown that the highly expressed *HEPACAM2* has a better prognosis and the reduced risk of death in patients with COAD based on different adjusted models. Therefore, *HEPACAM2* might be an independent diagnostic and prognostic biological indicator in patients with COAD [36]. *TTR* also had higher degree in the early stage in the right colon and down-regulated in late stages compared to the early stage. Studies have shown that its level in patients with CRC metastasis is significantly lower than that in patients without CRC metastasis [177]. *TTR* can be used as an indicator to evaluate the occurrence and prognosis of CRC. *DEFA5* and *DEFA6* are both down-regulated (in late stages vs. the early stage) in the right colon and up-regulated (in late stages vs. the early stage) in the left colon. Arijs et.al found a marked upregulation of *DEFA5* and *DEFA6* expression in the inflamed colon of patients with ulcerative colitis [178].

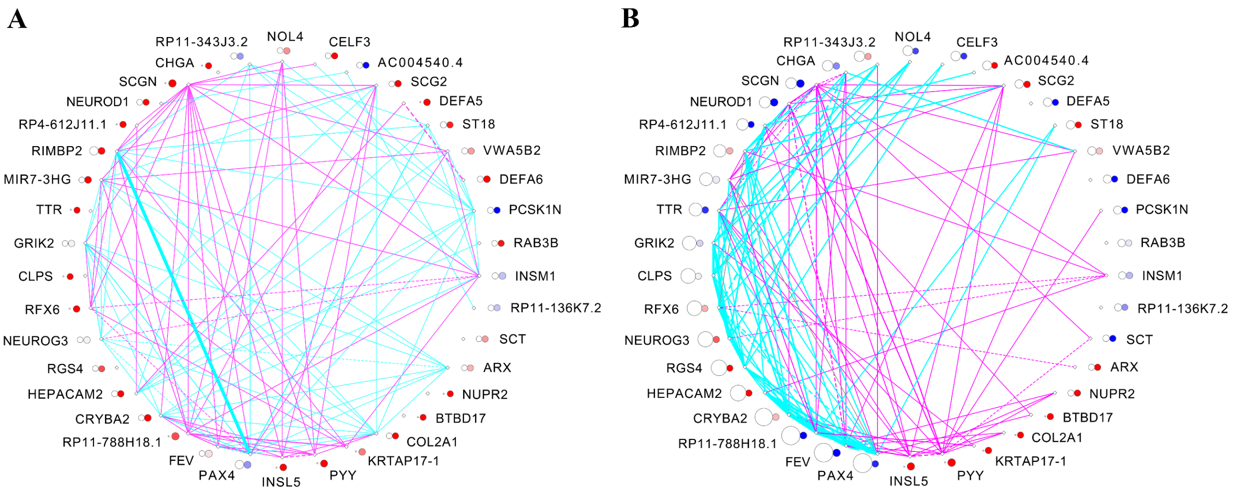


Figure 4.7: Circle plot for top 10% unique edges of module greenyellow in **A** the left colon and **B** the right colon. The plots are arranged similar to those in Figure 4.5.

Most genes of module magenta had larger degree in the early stage compared to late stages in the left colon and were mostly downregulated in late stages compared to the early stage (Figure 4.8A), whereas in the right colon the gene degrees in the early stage were lower than late stages and most genes were upregulated (Figure 4.8B). *MS4A1*, the gene with the largest degree in the circle plots, has been shown to be downregulated in colorectal carcinoma and its expression is positively correlated with CRC patient survival [179]. *CD19*, *CD22*, and *CD79B*, which were upregulated in late stages (vs. the early stage) in the right colon but downregulated in late stages of the left colon and enriched in B cell receptor signaling pathway, play a central role in colon cancer initiation and development [180]. *TCL1A*, was another high degree gene which was downregulated in late stages of the left colon but upregulated in late stages of the right colon. It has been shown to have higher expression in CRC cancer stages compared to normal which correlates with tumor differentiation and clinical stage [181]. *TCL1A* is also a useful biomarker for prognostic evaluation of patients at stages 2 and 3 of CRC [182]. *RGS13*, upregulated in late stages of both left and right colons, was connected to other genes in the early stage of the left colon (all *RGS13*'s connections are cyan in

Figure 4.8A) and connected to other genes in late stages of the right colon (all *RGS13*'s connections are purple in Figure 4.8B). Studies have shown that *RGS13* regulates mast and T cell migration and activation [183]. RGS proteins may also serve as a prognostic factor in CRC diagnosis [184].

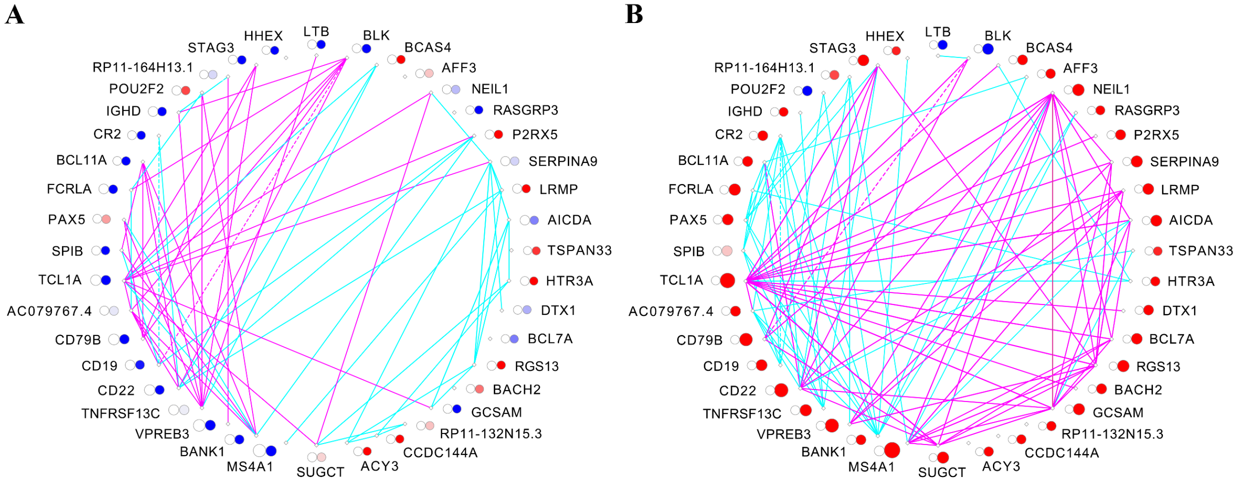


Figure 4.8: Circle plot for top 10% unique edges of module magenta in **A** the left colon and **B** the right colon. The plots are arranged similar to those in Figure 4.5.

4.4.3.3 Preserved module is mostly enriched for signaling pathways

As seen from Figure 4.6, module pink was the most preserved module enriched for renin angiotensin system, Fc epsilon RI signaling, and JAK-STAT signaling pathway based on KEGG and IL-2/STAT5 signaling, fatty acid metabolism, and angiogenesis based on MSigDB hallmark gene sets. Spearman's rank correlation for the gene degrees of module pink between the right and left colons was 0.96 which confirms the preservation of this module. Renin angiotensin system (RAS) components are known to be dysregulated in CRC, which indicates their potential role in CRC pathology with a pivotal role in metastasis [185]. JAK/STAT signal transduction is a common signaling pathway through which many growth factors and cytokines transmit signals in cells. Studies have shown that the JAK/STAT signal may be used as a novel tumor marker and prognostic factor for the diagnosis, assessment, and prognosis of CRC [186]. Fatty acid metabolism supports tumorigenesis and disease progression through a range of processes including energy storage and production [187]. Angiogenesis, the formation of new blood vessels, is a critical step in the cancer progression with an important role from the early stage of CRC to the late phase of metastasis [188].

4.4.3.4 In-silico validation

To validate our results of the comparison between right and left colon, we utilized another scRNA-seq data (henceforth referred to as validation dataset) from GEO (accession IDs are GSE132465 and GSE144735). The data contained 47,285 cells from 23 Korean and 17,678 cells from 6 Belgian patients at different stages of CRC [189]. We merged two datasets and followed the similar procedure (with similar values for the parameters) as the one for the original dataset to detect modules of the right colon and identify the non-preserved ones using *blockwiseModules* and *modulePreservation* functions of WGCNA, respectively (see Materials and Methods). The focus of this validation was on the right colons of the original and validation datasets. There was about 60% overlap between the top 20% variable genes of the right colon of original data and top 20% variable genes of the right colon of the validation data. We detected 12 modules for the right colon cells of the validation dataset, of which two of them were non-preserved in the left colon of the validation dataset. The size of modules is listed in Table S4.3.

Jaccard Index (JI) values were calculated between all modules of the original and validation datasets (Table S4.3). Although the maximum value of JI was only 48%, the non-preserved modules of the validation dataset represented biological functions similar to that for the non-preserved modules of the original dataset. They were enriched for pathways such as KRAS signaling down, notch signaling and angiogenesis. KRAS is one of the most frequently mutated oncogenes with a prevalence of about 40% in CRC and is involved in the occurrence, progression, treatment and recurrence of the disease [190]. Notch signaling, necessary to maintain intestinal homeostasis, is involved in regulating tumor progression and aggressiveness of CRC [191]. As mentioned in the previous section, angiogenesis plays a critical role in cancer progression because solid tumors need a blood supply to grow [188].

4.5 Conclusion

In this study, we utilized a scRNA-seq data containing 370,115 cells from 62 CRC patients and adjacent normal tissues of CRC to detect modular mechanisms different between early (pT1) and late stages (pT234) and between right and left colons. We first pre-processed the data and made the number of cells equal at all stages and right and left colons. Then, we constructed data expression matrices at early and late stages and detected modules at the early stage, of which two were non-preserved in late stages. Spearman's rank correlation of gene degrees also confirmed the non-preservation of these modules. The non-preserved modules captured myeloid cells and were enriched for cytokine-cytokine receptor interaction, NF-κB signaling, carbohydrate digestion and absorption, and bile secretion pathways. There were also important genes in these modules. For example, *FSCN1*, whose overexpression is known to promote cancer cell migration, invasion, and metastasis *in vitro* and *in vivo*, had higher z-score in late stages compared to the early stage. The most preserved module of this comparison captured mast cells with a mixed role in CRC.

The comparison between the right and the left colons revealed that two modules were non-preserved and enriched for neuroactive ligand-receptor interaction pathways, hematopoietic cell lineage, and intestinal immune network for IgA production based on KEGG, and KRAS signaling down and epithelial mesenchymal transition and notch signaling based on MSigDB hallmark gene sets. *TTR*, an important gene in these modules, can be used as an indicator to evaluate the occurrence and prognosis of CRC. *DEFA5* and *DEFA6* were two other genes and both were down-regulated (in late stages vs. the early stage) in the right colon and up-regulated (in late stages vs. the early stage) in the left colon. Studies have shown a marked upregulation of *DEFA5* and *DEFA6* expression in the inflamed colon of patients with ulcerative colitis. The preserved module of this comparison was mostly enriched for signaling pathways such as Fc epsilon RI and JAK-STAT.

Overall, we were able to detect topological and functional differences between the early stage (pT1) and late stages (pT234) of CRC as well as the right vs. left colon.

4.6 Acknowledgement

Chapter 4, in part, is currently being prepared for submission for publication of the material. Rahiminejad, Sara; Mukund, Kavitha; Maurya, Mano R.; and Subramaniam, Shankar. The dissertation author was the primary researcher and author of this material.

4.7 Supplementary Methods

scSampler

The input to the algorithm is a matrix $X \in R^{n\times \mathrm{p}}$ with columns corresponding to *p* features (top *p* PCs from a cell-by-gene log(count+1) matrix and scaled to [0,1]) and rows corresponding to *n* cells. Then, there is a set of cells $X = \{x_1, x_2, \dots x_\mathrm{n}\}$ where x_i € R^p is a row vector of X. The goal is to find a size n_s subset X_s subset of X which minimizes the Hausdorff distance:

$$d_H(X_s, X) = \max_{x_i \in X} \min_{x_j \in X_s} d(x_i, x_j) \qquad \text{(S4.1)}$$

where *d*(.,.) is the Euclidean distance. The Hausdorff distance measures the distance from X_s to *X*. A small value of the distance means that all data points in *X* are represented by at least one data point in X_s. Since the direct optimization of Equation S4.1 is computationally expensive, the authors proposed to approximate the optimality and search for X_s via the following optimization problem:

$$\min_{x_i, x_j \in X_s} \sum_{i=1}^{n_s - 1} \sum_{j=i+1}^{n_s} \frac{1}{[d(x_i, x_j)]^{\alpha}} \tag{S4.2}$$

for a sufficiently large alpha [ref]. Cells in X_s obtained from equation 2 have maximized distance and minimized similarity between each other and therefore can represent the diversity of X. The authors showed that $\alpha = 4p$ is big enough and keeps the algorithm numerically stable [144].

blockwiseModules function

The function first pre-clusters nodes (genes) into large clusters, named as blocks, using a variant of k-means clustering (*projectiveKMeans*) [138]. Then, hierarchical clustering is applied to each block and modules are defined as branches of the resulting dendrogram. Afterwards, an automatic module merging step (*mergeCloseModules*) is performed to merge modules whose eigengenes are highly correlated.

modulePreservation function

The function detects the conservation of gene pairs between two networks (early and late stages of cancer here). There are three types of network-based module preservation statistics: 1) Density based: determines if module nodes remain highly connected in the test network, 2) Separability based: determines if network modules remain distinct from one another in the test network, and 3) Connectivity based: determines if the connectivity pattern between nodes in the reference network is similar to that in the test network [192].

As those statistics measure different aspects of module preservation, two composite preservation statistics have been defined:

Z_{summary}: summarizes the individual Z statistic values resulting from permutation test and is calculated by:

$$Z_{summary} = \frac{Z_{density} + Z_{connectivity}}{2} \tag{S4.3}$$

medianRank: is a rank-based measure which relies on observed preservation statistics and less dependent on module sizes. It is calculated by the following equation:

$$medianRank = \frac{medianRank_{density} + medianRank_{connectivity}}{2} \tag{S4.4}$$

Permutation was performed 100 times due to the computational complexity of our network sizes. As prescribed in [192], modules with a $Z_{summary} > 10$ indicate strong evidence of preservation, 2 to 10 indicate low to moderate evidence of preservation and less than 2 indicate no preservation.

4.8 Supplementary Figures

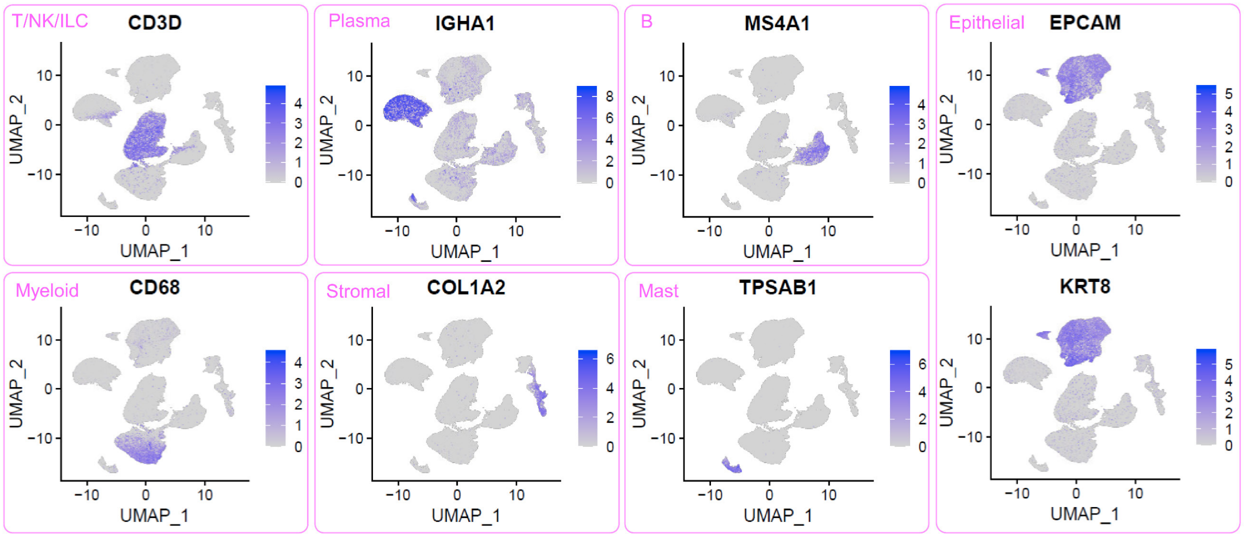


Figure S4.1: Feature plots for cell markers of major cell types before down-sampling.

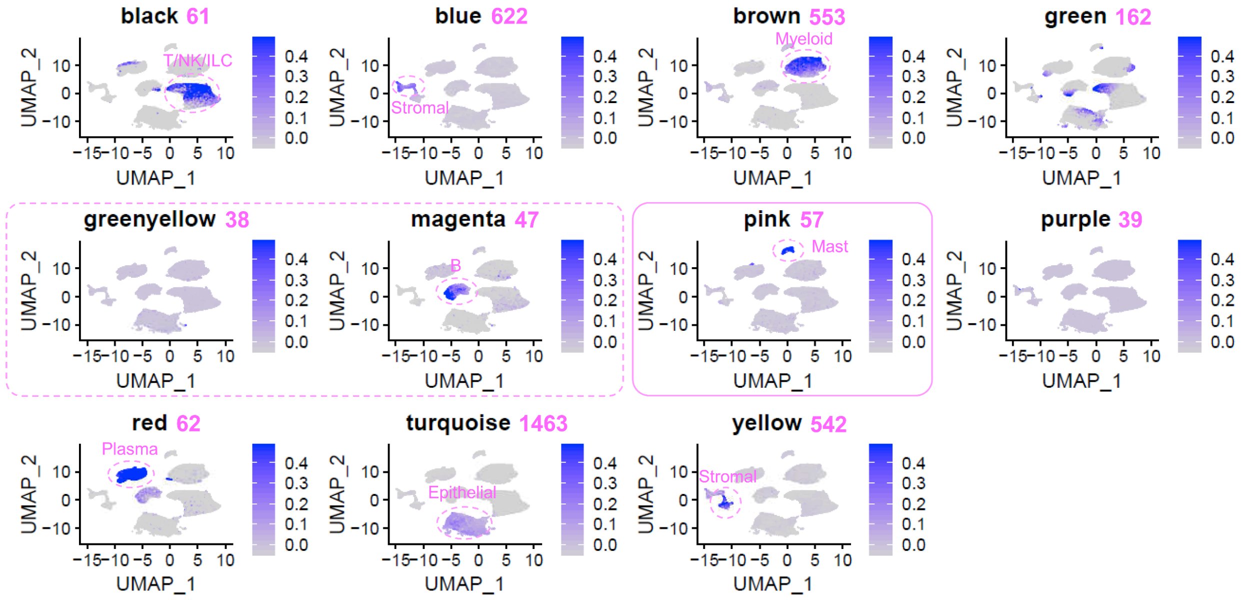


Figure S4.2: Cell coverage of different modules detected for the right colon. There is a number after each module's name representing the number of genes in that module. Non-preserved (or preserved) modules are highlighted by a dashed (or solid) purple rectangle around them.

Chapter 5 CONCLUSIONS

In this dissertation we explored community detection as an effective computational tool to analyze and understand biological networks and gain valuable insights about the networks from a topological and functional perspective. A variety of biological datasets have been utilized as case studies to illustrate the applicability of community detection in the field of biological networks.

In Chapter 2, we explored the existing community detection algorithms and evaluated findings of six algorithms, namely, Combo, Conclude, Fast Greedy, Leading Eigen, Louvain and Spinglass, on two important PPI networks, namely, Saccharomyces cerevisiae (Yeast) and Homo sapiens (Human) at both topological and functional levels. An in-depth comparison of communities detected by different methods has led us to conclude that Louvain and Spinglass are most similar for the Yeast PPI network whereas Combo and Spinglass are most similar for the Human PPI network. However, since Combo and Spinglass use stochastic search in their procedure and their running time is also more than that for Louvain, we concluded that Louvain is likely the best method to find reasonably sized communities for biological networks in a reasonable time.

In Chapter 3, we utilized a microarray gene expression data from 128 patients at various stages of CRC to detect modular mechanisms potentially causal for progression of disease from normal to stages I-IV and to find stage-specific biomarkers. Comparing communities of different stages, detected by Louvain algorithm, revealed that neighboring stages were more similar to each other than non-neighboring stages at both topological and functional levels. We carried out the functional analysis at the whole network level for the stage-specific and stage-unique networks by analyzing the enrichment of 24 cancer-related pathways. For the stage-specific networks, most of the CRC related pathways such as PI3K-Akt and MAPK signaling pathways were enriched at all stages. However, stage-unique networks revealed functional differences across the stages. We also identified key biomarkers to differentiate between any stages of CRC and normal using STEM analysis. *WNT2* and *SFRP2*, two biomarkers validated by other researchers in stool DNA, were over-expressed and under-expressed in CRC tissues, respectively. Finally, we constructed a drug-target-PPI network enabling us, in the light of the present data, to understand the functional mechanisms associated with some of the current drugs associated with CRC treatment. We found that the target gene weights changed across the stages extensively. For example, *TYMS*, a target for some drugs such as 5-FU, was upregulated in cancer stages with larger weights in stage III.

In Chapter 4, we analyzed a scRNA-seq dataset consisting of 370,115 cells from 62 CRC patients and adjacent normal tissues to detect modular mechanisms different between the early (pT1) and late stages (pT234) and between the right and left colons. After preprocessing the data, we constructed data expression matrices at the early and late stages as well as the right and left colons and detected modules using WGCNA. Of the modules detected for the early stage, two of them were non-preserved in late stages, capturing myeloid cells, and enriched for cytokine-cytokine receptor interaction, NF-κB signaling, carbohydrate digestion and absorption, and bile secretion pathways. *FSCN1*, an important gene of these modules, had a higher z-score in late stages compared to the early stage. Its overexpression has been known to promote cancer cell migration, invasion, and metastasis *in vitro* and *in vivo*. The comparison between the right and left colons also revealed that two modules were non-preserved and enriched for hematopoietic cell lineage, intestinal immune network for IgA production, KRAS signaling down, and notch signaling pathways. *DEFA5* and *DEFA6*, both down-regulated in the right colon and up-regulated in the left colon, have been shown to be upregulated in the inflamed colon of patients with ulcerative colitis.

The research presented in this dissertation paves the way for topological and functional network analyses in complex biological processes and diseases.

REFERENCES

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics reports,* vol. 424, no. 4-5, pp. 175-308, 2006.

[2] M. E. Newman, "The structure and function of complex networks," *SIAM review,* vol. 45, no. 2, pp. 167-256, 2003.

[3] S. Wasserman and K. Faust, Social network analysis: Methods and applications, Cambridge University Press, 1994.

[4] S. P. Borgatti, A. Mehra, D. J. Brass and G. Labianca, "Network analysis in the social sciences," *science,* vol. 323, no. 5916, pp. 892-895, 2009.

[5] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proceedings of the national Academy of sciences,* vol. 100, no. 3, pp. 1128-1133, 2003.

[6] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the national Academy of sciences,* vol. 100, no. 21, pp. 12123-12128, 2003.

[7] G. W. Flake, S. Lawrence, C. L. Giles and F. M. Coetzee, "Self-organization and identification of web communities," *Computer,* vol. 35, no. 3, pp. 66-70, 2002.

[8] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry,* pp. 35-41, 1977.

[9] R. S. Burt, "Positions in networks," *Social forces,* vol. 55, no. 1, pp. 93-122, 1976.

[10] A. Lancichinetti, M. Kivelä, J. Saramäki and S. Fortunato, "Characterizing the community structure of complex networks," *PloS one,* vol. 5, no. 8, p. e11976, 2010.

[11] B. Krishnamurthy and J. Wang, "On network-aware clustering of web clients," in *Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, Stockholm, 2000.

[12] J. Moody and D. R. White, "Structural cohesion and embeddedness: A hierarchical concept of social groups," *American sociological review,* pp. 103-127, 2003.

[13] S. Redner, "How popular is your paper? An empirical study of the citation distribution," *The European Physical Journal B-Condensed Matter and Complex Systems,* vol. 4, no. 2, pp. 131-134, 1998.

[14] A. E. Sizemore, C. Giusti, A. Kahn, J. M. Vettel, R. F. Betzel and D. S. Bassett, "Cliques and cavities in the human connectome," *Journal of computational neuroscience,* vol. 44, pp. 115-145, 2018.

[15] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences,* vol. 99, no. 12, pp. 7821-7826, 2002.

[16] G. Didier, C. Brun and A. Baudot, "Identifying communities from multiplex biological networks," *PeerJ,* vol. 3, p. e1525, 2015.

[17] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E,* vol. 69, no. 2, p. P026113, 2004.

[18] A. Clauset, M. E. Newman and C. Moore, "Finding community structure in very large networks," *Physical Review E,* vol. 70, no. 6, p. P066111, 2004.

[19] U. N. Raghavan, R. Albert and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E,* vol. 76, no. 3, p. 036106, 2007.

[20] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment,* vol. 2008, no. 10, p. 10008, 2008.

[21] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical review E,* vol. 74, no. 3, p. P036104, 2006.

[22] S. Rahiminejad, M. R. Maurya and S. Subramaniam, "Topological and functional comparison of community detection algorithms in biological networks," *BMC bioinformatics,* vol. 20, no. 1, pp. 1-25, 2019.

[23] S. Rahiminejad, M. R. Maurya, K. Mukund and S. Subramaniam, "Modular and mechanistic changes across stages of colorectal cancer," *BMC cancer,* vol. 22, no. 1, pp. 1-14, 2022.

[24] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the national academy of sciences,* vol. 101, no. 9, pp. 2658-2663, 2004.

[25] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski and D. Wagner, "On modularity clustering," *IEEE transactions on knowledge and data engineering,* vol. 20, no. 2, pp. 172-188, 2008.

[26] S. Fortunato, "Community detection in graphs," *Physics reports,* vol. 486, no. 3-5, pp. 75-174, 2010.

[27] Z. Yang, R. Algesheimer and C. J. Tessone, "A Comparative Analysis of Community Detection Algorithms on Artificial Networks," *Scientific reports,* vol. 6, no. 1, p. 30750, 2016.

[28] A. Lancichinetti, S. Fortunato and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical review E,* vol. 78, no. 4, p. P046110, 2008.

[29] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical review E,* vol. 74, no. 1, p. P016110, 2006.

[30] J. Li and Y. Song, "Community detection in complex networks using extended compact genetic algorithm," *Soft computing,* vol. 17, no. 6, pp. 925-937, 2013.

[31] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E,* vol. 69, no. 6, p. P066133, 2004.

[32] Z. M. Ibrahim and A. Ngom, "The relative vertex clustering value - a new criterion for the fast discovery of functional modules in protein interaction networks," *BMC bioinformatics,* vol. 16, pp. 1-14, 2015.

[33] P. Sah, L. O. Singh, A. Clauset and S. Bansal, "Exploring community structure in biological networks with random graphs," *BMC bioinformatics,* vol. 15, no. 1, pp. 1-14, 2014.

[34] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," *Physical review E,* vol. 72, no. 2, p. P027104, 2005.

[35] S. Boettcher and A. G. Percus, "Extremal optimization for graph partitioning," *Physical review E,* vol. 64, no. 2, p. P026114, 2001.

[36] S. Boettcher and A. G. Percus, "Optimization with extremal dynamics," *complexity,* vol. 8, no. 2, pp. 57-62, 2002.

[37] C. Pizzuti, "Ga-net: A genetic algorithm for community detection in social networks," in *International conference on parallel problem solving from nature*, Berlin, Heidelberg, 2008.

[38] C. Shi, Z. Yan, Y. Cai and B. Wu, "Multi-objective community detection in complex networks," *Applied Soft Computing,* vol. 12, no. 2, pp. 850-859, 2012.

[39] "The BioGRID," [Online]. Available: https://thebiogrid.org/. [Accessed 2018].

[40] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic acids research,* vol. 34, no. suppl_1, pp. D535-D539, 2006.

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

[42] D. W. Huang, B. T. Sherman and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research,* vol. 37, no. 1, pp. 1-13, 2009.

[43] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic acids research,* vol. 45, no. D1, pp. D353-D361, 2017.

[44] S. Durinck, P. T. Spellman, E. Birney and W. Huber, "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt," *Nature protocols,* vol. 4, no. 8, pp. 1184-1191, 2009.

[45] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research,* vol. 13, no. 11, pp. 2498-2504, 2003.

[46] S. Sobolevsky, R. Campari, A. Belyi and C. Ratti, "General optimization technique for high-quality community detection in complex networks," *Physical review E,* vol. 90, no. 1, p. P012811, 2014.

[47] P. De Meo, E. Ferrara, G. Fiumara and A. Provetti, "Mixing local and global information for community detection in large networks," *Journal of Computer and System Sciences,* vol. 80, no. 1, pp. 72-87, 2014.

[48] M. Rosvall, D. Axelsson and C. T. Bergstrom, "The map equation," *The European Physical Journal Special Topics,* vol. 178, no. 1, pp. 13-23, 2009.

[49] G. Cs´ardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal, complex systems,* vol. 1695, no. 5, pp. 1-9, 2006.

[50] G. Su, A. Kuchinsky, J. H. Morris, D. J. States and F. Meng, "GLay: community structure analysis of biological networks," *Bioinformatics,* vol. 26, no. 24, pp. 3135-3137, 2010.

[51] E. Ferrara, "CONCLUDE (COmplex Network CLUster DEtection) is a fast community detection algorithm.," 2014. [Online]. Available: http://www.emilio.ferrara.name/code/conclude/. [Accessed 2018].

[52] "Combo method," 2014. [Online]. Available: http://senseable.mit.edu/community_detection/. [Accessed 2018].

[53] M. E. Newman, "Scientific collaboration networks. I. Network construction and fundamental results," *Physical review E,* vol. 64, no. 1, p. P016131, 2001.

[54] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *The Bell system technical journal,* vol. 49, no. 2, pp. 291-307, 1970.

[55] G. K. Orman, V. Labatut and H. Cherifi, "Comparative evaluation of community detection algorithms: a topological approach," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2012, no. 08, p. P08001, 2012.

[56] L. Waltman and N. J. Van Eck, "A smart local moving algorithm for large-scale modularity-based community detection," *The European physical journal B,* vol. 86, pp. 1-14, 2013.

[57] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the national academy of sciences,* vol. 105, no. 4, pp. 1118-1123, 2008.

[58] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association,* vol. 66, no. 336, pp. 846-850, 1971.

[59] L. I. Kuncheva and S. T. Hadjitodorov, "Using diversity in cluster ensembles," *2004 IEEE international conference on systems, man and cybernetics,* vol. 2, no. 04CH37583, pp. 1214-1219, 2004.

[60] L. Hubert and P. Arabic, "Comparing partitions," *Journal of classification,* vol. 2, pp. 193-218, 1985.

[61] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians,* vol. 68, no. 6, pp. 394-424, 2018.

[62] N. Pawa, T. Arulampalam and J. D. Norton, "Screening for colorectal cancer: established and emerging modalities," *Nature reviews Gastroenterology & hepatology,* vol. 8, no. 12, pp. 711-722, 2011.

[63] P. Rawla, T. Sunkara and A. Barsouk, "Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors," *Gastroenterology Review/Przegląd Gastroenterologiczny,* vol. 14, no. 2, pp. 89-103, 2019.

[64] N. P. Brouwer, A. C. Bos, V. E. Lemmens, P. J. Tanis, N. Hugen, I. D. Nagtegaal, J. H. de Wilt and R. H. Verhoeven, "An overview of 25 years of incidence, treatment and outcome of colorectal cancer patients," *International journal of cancer,* vol. 143, no. 11, pp. 2758-2766, 2018.

[65] N. L. Henry and D. F. Hayes, "Cancer biomarkers," *Molecular oncology,* vol. 6, no. 2, pp. 140-146, 2012.

[66] K. M. Ryan, A. C. Phillips and K. H. Vousden, "Regulation and function of the p53 tumor suppressor protein," *Current opinion in cell biology,* vol. 13, no. 3, pp. 332-337, 2001.

[67] A. Russo, V. Bazan, B. Iacopetta, D. Kerr, T. Soussi and N. Gebbia, "The TP53 Colorectal Cancer International Collaborative Study on the Prognostic and Predictive Significance of p53 Mutation: Influence of Tumor Site, Type of Mutation, and Adjuvant Treatment," *Journal of clinical oncology,* vol. 23, no. 30, pp. 7518-7528, 2005.

[68] K. Mukund, N. Syulyukina, S. Ramamoorthy and S. Subramaniam, "Right and left-sided colon cancers - specificity of molecular mechanisms in tumorigenesis and progression," *BMC cancer,* vol. 20, no. 1, pp. 1-15, 2020.

[69] A. Palaniappan, K. Ramar and S. Ramalingam, "Computational Identification of Novel Stage-Specific Biomarkers in Colorectal Cancer Progression," *PloS one,* vol. 11, no. 5, p. e0156665, 2016.

[70] Y. Cai, N. J. Rattray, Q. Zhang, V. Mironova, A. Santos-Neto, E. Muca, A. K. Rosen Vollmar, K.-S. Hsu, Z. Rattray, J. R. Cross, Y. Zhang, P. B. Paty, S. A. Khan and C. H. Johnson, "Tumor tissue-specific biomarkers of colorectal cancer by anatomic location and stage," *Metabolites,* vol. 10, no. 6, p. 257, 2020.

[71] J. Ernst, G. J. Nau and Z. Bar-Joseph, "Clustering short time series gene expression data," *Bioinformatics,* vol. 21, no. suppl_1, pp. i159-i168, 2005.

[72] J. Ernst and Z. Bar-Joseph, "STEM: a tool for the analysis of short time series gene expression data," *BMC bioinformatics,* vol. 7, no. 1, pp. 1-11, 2006.

[73] S. Tsukamoto, T. Ishikawa, S. Iida, M. Ishiguro, K. Mogushi, H. Mizushima, H. Uetake, H. Tanaka and K. Sugihara, "Clinical significance of osteoprotegerin expression in human colorectal cancer," *Clinical cancer research,* vol. 17, no. 8, pp. 2444-2450, 2011.

[74] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic acids research,* vol. 31, no. 4, pp. e15-e15, 2003.

[75] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science,* vol. 2, no. 11, pp. 559-572, 1901.

[76] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *Journal of machine learning research,* vol. 9, no. 11, 2008.

[77] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic acids research,* vol. 43, no. 7, pp. e47-e47, 2015.

[78] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic acids research,* vol. 28, no. 1, pp. 27-30, 2000.

[79] F. Emmert-Streib, R. de Matos Simoes, P. Mullan, B. Haibe-Kains and M. Dehmer, "The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks," *Frontiers in genetics,* vol. 5, p. 15, 2014.

[80] Z. Tang, C. Li, B. Kang, G. Gao, C. Li and Z. Zhang, "GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses," *Nucleic acids research,* vol. 45, no. W1, pp. W98-W102, 2017.

[81] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen and C. von Mering, "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic acids research,* vol. 47, no. D1, pp. D607-D613, 2019.

[82] "Drugs Approved for Colon and Rectal Cancer," [Online]. Available: https://www.cancer.gov/about-cancer/treatment/drugs/colorectal. [Accessed 2021].

[83] D. S. Wishart, C. Knox, A. Chi Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research,* vol. 34, no. suppl_1, pp. D668-D672, 2006.

[84] A. D. Richardson, C. Yang, A. Osterman and J. W. Smith, "Central carbon metabolism in the progression of mammary carcinoma," *Breast cancer research and treatment,* vol. 110, pp. 297-307, 2008.

[85] J. Y. Fang and B. C. Richardson, "The MAPK signalling pathways and colorectal cancer," *The lancet oncology,* vol. 6, no. 5, pp. 322-327, 2005.

[86] L. W. Cheung and G. B. Mills, "Targeting therapeutic liabilities engendered by PIK3R1 mutations for cancer treatment," *Pharmacogenomics,* vol. 17, no. 3, pp. 297-307, 2016.

[87] Y. Itatani, K. Kawada and Y. Sakai, "Transforming growth factor-β signaling pathway in colorectal cancer and its tumor microenvironment," *International journal of molecular sciences,* vol. 20, no. 23, p. 5822, 2019.

[88] L. Gao, C. Ge, S. Wang, X. Xu, Y. Feng, X. Li, C. Wang, Y. Wang, F. Dai and S. Xie, "The role of p53-mediated signaling in the therapeutic response of colorectal cancer to 9F, a Spermine-modified Naphthalene Diimide derivative," *Cancers,* vol. 12, no. 3, p. 528, 2020.

[89] Q. Wang, G. He, M. Hou, L. Chen, S. Chen, A. Xu and Y. Fu, "Cell cycle regulation by alternative polyadenylation of CCND1," *Scientific reports,* vol. 8, no. 1, p. 6824, 2018.

[90] C. Zhang, Q. Zhu, J. Gu, S. Chen, Q. Li and L. Ying, "Down-regulation of CCNE1 expression suppresses cell proliferation and sensitizes gastric carcinoma cells to Cisplatin," *Bioscience Reports,* vol. 39, no. 6, p. BSR20190381, 2019.

[91] X. Shi, H. Li, H. Yao, X. Liu, L. Li, K. Leung, H. Kung and M. C. Lin, "Adapalene inhibits the activity of cyclin-dependent kinase 2 in colorectal carcinoma," *Molecular medicine reports,* vol. 12, no. 5, pp. 6501-6508, 2015.

[92] K.-Y. Jeong, "Inhibiting focal adhesion kinase: A potential target for enhancing therapeutic efficacy in colorectal cancer therapy," *World journal of gastrointestinal oncology,* vol. 10, no. 10, pp. 290-292, 2018.

[93] M. L. Slattery, L. E. Mullany, L. Sakoda, W. S. Samowitz, R. K. Wolff, J. R. Stevens and J. S. Herrick, "The NF-κB signalling pathway in colorectal cancer: associations between dysregulated gene and miRNA expression," *Journal of cancer research and clinical oncology,* vol. 144, no. 2, pp. 269-283, 2018.

[94] X. Liu, Q. Ji, Z. Fan and Q. Li, "Cellular signaling pathways implicated in metastasis of colorectal cancer and the associated targeted agents," *Future Oncology,* vol. 11, no. 21, pp. 2911-2922, 2015.

[95] T. Wang, F. Lin, X. Sun, L. Jiang, R. Mao, S. Zhou, W. Shang, R. Bi, F. Lu and S. Li, "HOXB8 enhances the proliferation and metastasis of colorectal cancer cells by promoting EMT via STAT3 activation," *Cancer Cell International,* vol. 19, no. 1, pp. 1-12, 2019.

[96] X. Li, H. Lin, F. Jiang, Y. Lou, L. Ji and S. Li, "Knock-Down of HOXB8 Prohibits Proliferation and Migration of Colorectal Cancer Cells via Wnt/β-Catenin Signaling Pathway," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research,* vol. 25, p. 711, 2019.

[97] Y.-S. Jung, S. Jun, S. H. Lee, A. Sharma and J.-I. Park, "Wnt2 complements Wnt/β-catenin signaling in colorectal cancer," *Oncotarget,* vol. 6, no. 35, pp. 37257-37268, 2015.

[98] N. Kramer, J. Schmöllerl, C. Unger, H. Nivarthi, A. Rudisch, D. Unterleuthner, M. Scherzer, A. Riedl, M. Artaker, I. Crncec, D. Lenhardt, T. Schwarz, B. Prieler, X. Han, M. Hengstschläger, J. Schüler, R. Eferl, R. Moriggl, W. Sommergruber and H. Dolznig, "Autocrine WNT2 signaling in fibroblasts promotes colorectal cancer progression," *Oncogene,* vol. 36, no. 39, pp. 5460-5472, 2017.

[99] F. J. Carmona, D. Azuara, A. Berenguer-Llergo, A. F. Fernández, S. Biondo, J. de Oca, F. Rodriguez-Moranta, R. Salazar, A. Villanueva, M. F. Fraga, J. Guardiola, G. Capellá, M. Esteller and V. Moreno, "DNA methylation biomarkers for noninvasive diagnosis of colorectal cancer," *Cancer prevention research,* vol. 6, no. 7, pp. 656-665, 2013.

[100] X. Liu, J. Fu, H. Bi, A. Ge, T. Xia, Y. Liu, H. Sun, D. Li and Y. Zhao, "DNA methylation of SFRP1, SFRP2, and WIF1 and prognosis of postoperative colorectal cancer patients," *BMC cancer,* vol. 19, pp. 1-14, 2019.

[101] O. Ozcan, M. Kara, O. Yumrutas, E. Bozgeyik, I. Bozgeyik and O. I. Celik, "MTUS1 and its targeting miRNAs in colorectal carcinoma: significant associations," *Tumor Biology,* vol. 37, no. 5, pp. 6637-6645, 2016.

[102] C. Zuern, J. Heimrich, R. Kaufmann, K. K. Richter, U. Settmacher, C. Wanner, J. Galle and S. Seibold, "Down-regulation of MTUS1 in human colon tumors," *Oncology reports,* vol. 23, no. 1, pp. 183-189, 2010.

[103] H. Hu, T. Wang, R. Pan, Y. Yang, B. Li, C. Zhou, J. Zhao, Y. Huang and S. Duan, "Hypermethylated promoters of secreted frizzled-related protein genes are associated with colorectal cancer," *Pathology & Oncology Research,* vol. 25, no. 2, pp. 567-575, 2019.

[104] A. Loktionov, "Biomarkers for detecting colorectal cancer non-invasively: DNA, RNA or proteins?," *World journal of gastrointestinal oncology,* vol. 12, no. 2, pp. 124-148, 2020.

[105] N. Tsuneyoshi, K. Fukudome, S.-i. Horiguchi, X. Ye, M. Matsuzaki, M. Toi, K. Suzuki and M. Kimoto, "Expression and anticoagulant function of the endothelial cell protein C receptor (EPCR) in cancer cell lines," *Thrombosis and haemostasis,* vol. 85, no. 2, pp. 356-361, 2001.

[106] N. Lal, C. R. Willcox, A. Beggs, P. Taniere, A. Shikotra, P. Bradding, R. Adams, D. Fisher, G. Middleton, C. Tselepis and B. E. Willcox, "Endothelial protein C receptor is overexpressed in colorectal cancer as a result of amplification and hypomethylation of chromosome 20q," *The Journal of Pathology: Clinical Research,* vol. 3, no. 3, pp. 155-170, 2017.

[107] Y. Lei, S. Zhou, Q. Hu, X. Chen and J. Gu, "Carbohydrate response element binding protein (ChREBP) correlates with colon cancer progression and contributes to cell proliferation," *Scientific reports,* vol. 10, no. 1, p. 4233, 2020.

[108] H.-K. Park, I.-H. Kim, J. Kim and T.-J. Nam, "Induction of apoptosis and the regulation of ErbB signaling by laminarin in HT-29 human colon cancer cells," *International Journal of Molecular Medicine,* vol. 32, no. 2, pp. 291-295, 2013.

[109] Y. Itatani, K. Kawada, S. Inamoto, T. Yamamoto, R. Ogawa, M. M. Taketo and Y. Sakai, "The role of chemokines in promoting colorectal cancer invasion/metastasis," *International journal of molecular sciences,* vol. 17, no. 5, p. 643, 2016.

[110] C. Ceci, M. G. Atzori, P. M. Lacal and G. Graziani, "Role of VEGFs/VEGFR-1 signaling and its inhibition in modulating tumor invasion: Experimental evidence in different metastatic cancer models," *International journal of molecular sciences,* vol. 21, no. 4, p. 1388, 2020.

[111] N. Yarom and D. J. Jonker, "The role of the epidermal growth factor receptor in the mechanism and treatment of colorectal cancer," *Discovery medicine,* vol. 11, no. 57, pp. 95-105, 2011.

[112] H. C. Rustamaji, Y. S. Suharini,, A. A. Permana, W. A. Kusuma, S. Nurdiati, I. Batubara and T. Djatna , "A network analysis to identify lung cancer comorbid diseases," *Applied Network Science,* vol. 7, no. 1, pp. 1-23, 2002.

[113] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox and M. Wilson, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic acids research,* vol. 46, no. D1, pp. D1074-D1082, 2018.

[114] L. Abou-Elkacem, S. Arns, G. Brix, F. Gremse, D. Zopf, F. Kiessling and W. Lederle, "Regorafenib inhibits growth, angiogenesis, and metastasis in a highly aggressive, orthotopic colon cancer model," *Molecular cancer therapeutics,* vol. 12, no. 7, pp. I322-I331, 2013.

[115] R. Schmieder, J. Hoffmann, M. Becker, A. Bhargava, T. Muller, N. Kahmann, P. Ellinghaus, R. Adams, A. Rosenthal, K.-H. Thierauch, A. Scholz, S. M. Wilhelm and D. Zopf, "Regorafenib (BAY 73-4506): Antitumor and antimetastatic activities in preclinical models of colorectal cancer," *International journal of cancer,* vol. 135, no. 6, pp. 1487-1496, 2014.

[116] A. Grothey, J.-Y. Blay, N. Pavlakis, T. Yoshino and J. Bruix, "Evolving role of regorafenib for the treatment of advanced cancers," *Cancer treatment reviews,* vol. 86, p. 101993, 2020.

[117] Y. Luo, K. D. Tsuchiya, D. I. Park, R. Fausel, S. Kanngurn, P. Welcsh, S. Dzieciatkowski, J. Wang and W. M. Grady, "RET is a potential tumor suppressor gene in colorectal cancer," *Oncogene,* vol. 32, no. 16, pp. 2037-2047, 2013.

[118] D. M. Oliveira, K. Grillone, C. Mignogna, V. De Falco, C. Laudanna, F. Biamonte, R. Locane, F. Corcione, M. Fabozzi, R. Sacco, G. Viglietto, D. Malanga and A. Rizzuto, "Correction to: Next-generation sequencing analysis of receptor-type tyrosine kinase genes in surgically resected colon cancer: identification of gain-of-function mutations in the RET proto-oncogene," *Journal of Experimental & Clinical Cancer Research,* vol. 37, no. 1, pp. 1-2, 2018.

[119] F. Mohammad Rezaei, S. Hashemzadeh, R. Ravanbakhsh Gavgani, M. Hosseinpour Feizi, N. Pouladi, H. Samadi Kafil, L. Rostamizadeh, V. Kholghi Oskooei, M. Taheri and E. Sakhinia, "Dysregulated KDR and FLT1 gene expression in colorectal cancer patients," *Reports of biochemistry & molecular biology,* vol. 8, no. 3, p. 244, 2019.

[120] M. Lafitte, A. Sirvent and S. Roche, "Collagen kinase receptors as potential therapeutic targets in metastatic colon cancer," *Frontiers in oncology,* vol. 10, p. 125, 2020.

[121] J. Ose, A. Botma, Y. Balavarca, K. Buck, D. Scherer, N. Habermann, J. Beyerle, K. Pfütze, P. Seibold, E. J. Kap, A. Benner, L. Jansen, K. Butterbach, M. Hoffmeister, H. Brenner, A. Ulrich, M. Schneider, J. Chang-Claude, B. Burwinkel and C. M. Ulrich, "Pathway analysis of genetic variants in folate-mediated one-carbon metabolism-related genes and survival in a prospectively followed cohort of colorectal cancer patients," *Cancer medicine,* vol. 7, no. 7, pp. 2797-2807, 2018.

[122] H. Jiang, B. Li, F. Wang, C. Ma and T. Hao, "Expression of ERCC1 and TYMS in colorectal cancer patients and the predictive value of chemotherapy efficacy," *Oncology letters,* vol. 18, no. 2, pp. 1157-1162, 2019.

[123] K.-i. Fujita, Y. Kubota, H. Ishida and Y. Sasaki, "Irinotecan, a key chemotherapeutic drug for metastatic colorectal cancer," *World journal of gastroenterology,* vol. 21, no. 43, p. 12234, 2015.

[124] D. J. Jonker, C. J. O'Callaghan, C. S. Karapetis, J. R. Zalcberg, D. Tu, H.-J. Au, S. R. Berry, M. Krahn, T. Price, R. J. Simes, N. C. Tebbutt, G. van Hazel, R. Wierzbicki, C. Langer and M. J. Moore, "Cetuximab for the treatment of colorectal cancer," *New England Journal of Medicine,* vol. 357, no. 20, pp. 2040-2048, 2007.

[125] P. L. McCormack and S. J. Keam, "Bevacizumab: a review of its use in metastatic colorectal cancer," *Drugs,* vol. 68, pp. 487-506, 2008.

[126] A. Strehl and J. Ghosh, "Cluster ensembles---a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research,* vol. 3, pp. 583-617, 2002.

[127] R. L. Siegel, K. D. Miller, N. S. Wagle and A. Jemal, "Cancer statistics, 2023," *Ca Cancer J Clin,* vol. 73, no. 1, pp. 17-48, 2023.

[128] H. Brenner, M. Kloor and C. P. Pox, "Pox cP. colorectal cancer," *Lancet,* vol. 383, no. 9927, pp. 1490-1502, 2014.

[129] M. Mik, M. Berut, L. Dziki, R. Trzcinski and A. Dziki, "Right- and left-sided colon cancer – clinical and pathological differences of the disease entity in one organ," *Archives of Medical Science,* vol. 13, no. 1, p. 157–162, 2017.

[130] D. L. Hanna and H.-J. Lenz, "How We Treat Left-Sided vs Right-Sided Colon Cancer," *Clinical Advances in Hematology & Oncology,* vol. 18, no. 5, pp. 253-257, 2020.

[131] N. P. Brouwer, A. C. Bos, V. E. Lemmens, P. J. Tanis, N. Hugen, I. D. Nagtegaal, J. H. de Wilt and R. H. Verhoeven, "An overview of 25 years of incidence, treatment and outcome of colorectal cancer patients," *International journal of cancer,* vol. 143, no. 11, pp. 2758-2766, 2018.

[132] H. M. Levitin, J. Yuan and P. A. Sims, "Single-cell transcriptomic analysis of tumor heterogeneity," *Trends in cancer,* vol. 4, no. 4, pp. 264-268, 2018.

[133] A. Nguyen, W. H. Khoo, I. Moran, P. I. Croucher and T. G. Phan, "Single cell RNA sequencing of rare immune cell populations," *Frontiers in immunology,* vol. 9, p. 1553, 2018.

[134] W. R. Becker, S. A. Nevins, D. C. Chen, R. Chiu, A. M. Horning, T. K. Guha, R. Laquindanum, M. Mills, H. Chaib, U. Ladabaum, T. Longacre, J. Shen, E. D. Esplin, A. Kundaje, J. M. Ford, C. Curtis, M. P. Snyder and W. J. Greenleaf, "Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer," *Nature Genetics,* vol. 54, no. 7, pp. 985-995, 2022.

[135] R.-Q. Wang, W. Zhao, H.-K. Yang, J.-M. Dong, W.-J. Lin, F.-Z. He, M. Cui and Z.-L. Zhou, "Single-Cell RNA sequencing analysis of the heterogeneity in gene regulatory networks in colorectal cancer," *Frontiers in Cell and Developmental Biology,* vol. 9, p. 765578, 2021.

[136] W. Dai, F. Zhou, D. Tang, L. Lin, C. Zou, W. Tan and Y. Dai, "Single-cell transcriptional profiling reveals the heterogenicity in colorectal cancer," *Medicine,* vol. 98, no. 34, 2019.

[137] A. Willems, N. Panchy and T. Hong, "Using Single-Cell RNA Sequencing and MicroRNA Targeting Data to Improve Colorectal Cancer Survival Prediction," *Cells,* vol. 12, no. 2, p. 228, 2023.

[138] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC bioinformatics,* vol. 9, no. 1, pp. 1-13, 2008.

[139] P. Langfelder and S. Horvath, "Fast R functions for robust correlations and hierarchical clustering," *Journal of statistical software,* vol. 46, no. 11, 2012.

[140] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nature biotechnology,* vol. 33, no. 5, pp. 495-502, 2015.

[141] A. Butler, P. Hoffman, P. Smibert, E. Papalexi and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature biotechnology,* vol. 36, no. 5, pp. 411-420, 2018.

[142] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert and R. Satija, "Comprehensive integration of single-cell data," *Cell,* vol. 177, no. 7, pp. 1888-1902, 2019.

[143] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert and R. Satija, "Integrated analysis of multimodal single-cell data," *Cell,* vol. 184, no. 13, pp. 3573-3587, 2021.

[144] D. Song, N. M. Xi, J. J. Li and L. Wang, "scSampler: fast diversity-preserving subsampling of large-scale single-cell transcriptomic data," *Bioinformatics,* vol. 38, no. 11, pp. 3126-3127, 2022.

[145] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. Vaz Meirelles, N. R. Clark and A. Ma'ayan, "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool," *BMC bioinformatics,* vol. 14, no. 1, pp. 1-14, 2013.

[146] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, M. D. Monteiro, G. W. Gundersen and A. Ma'ayan, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic acids research,* vol. 44, no. W1, pp. W90-W97, 2016.

[147] Z. Xie, A. Bailey, M. V. Kuleshov, D. J. Clarke, J. E. Evangelista, S. L. Jenkins, A. Lachmann, M. L. Wojciechowicz, E. Kropiwnicki, K. M. Jagodnik, M. Jeon and A. Ma'ayan, "Gene set knowledge discovery with Enrichr," *Current protocols,* vol. 1, no. 3, p. e90, 2021.

[148] J. H. Zar, Spearman rank correlation, vol. 7, Wiley Online Library, 2005.

[149] K. Pelka, M. Hofree, J. H. Chen, S. Sarkizova, J. D. Pirl, V. Jorgji, A. Bejnood, D. Dionne, W. H. Ge, K. H. Xu, S. X. Chao, D. R. Zollinger, D. J. Lieb, J. W. Reeves, C. A. Fuhrman, M. L. Hoang, T. Delorey, L. T. Nguyen, J. Waldman, M. Klapholz, I. Wakiro, O. Cohen, J.

Albers, C. S. Smillie, M. S. Cuoco, J. Wu, M.-j. Su, J. Yeung, B. Vijaykumar, A. M. Magnuson, N. Asinovski, T. Moll, M. N. Goder-Reiser, A. S. Applebaum, L. K. Brais, L. K. DelloStritto, S. L. Denning, S. T. Phillips, E. K. Hill, J. K. Meehan, D. T. Frederick, T. Sharova, A. Kanodia, E. Z. Todres, J. Jané-Valbuena, M. Biton, B. Izar, C. D. Lambden, T. E. Clancy, R. Bleday, N. Melnitchouk, J. Irani, H. Kunitake, D. L. Hiroko, A. Srivastava, J. L. Hornick, S. Ogino, A. Rotem, S. Vigneau, B. E. Johnson, R. B. Corcoran, A. H. Sharpe, V. K. Kuchroo, K. Ng, M. Giannakis, L. T. Nieman, G. M. Boland, A. J. Aguirre, A. C. Anderson, O. Rozenblatt-Rosen, A. Regev and N. Hacohen, "Spatially organized multicellular immune hubs in human colorectal cancer," *Cell,* vol. 184, no. 18, pp. 4734-4752, 2021.

[150] J. Borowczak, K. Szczerbowski, M. Maniewski, A. Kowalewski, M. Janiczek-Polewska, A. Szylberg, A. Marszałek and Ł. Szylberg, "The Role of Inflammatory Cytokines in the Pathogenesis of Colorectal Carcinoma—Recent Findings and Review," *Biomedicines,* vol. 10, no. 7, p. 1670, 2022.

[151] Z. Qian, L. Xue, A. Xu, Z. Li, Q. He, H. Xiujuan, G. Xu, F. Tian, Y. Ding and W. Zhu, "Chemokines in progression, chemoresistance, diagnosis, and prognosis of colorectal cancer," *Frontiers in Immunology,* vol. 13, p. 724139, 2022.

[152] H. Liu, G. Li, B. Zhang, D. Sun, J. Wu, F. Chen, F. Kong, Y. Luan, W. Jiang, R. Wang and X. Xue, "Suppression of the NF-κB signaling pathway in colon cancer cells by the natural compound Riccardin D from Dumortierahirsute," *Molecular medicine reports,* vol. 17, no. 4, pp. 5837-5843, 2018.

[153] T. R. Radstake, R. van der Voort, M. ten Brummelhuis, M. de Waal Malefijt, M. Looman, C. Figdor, W. van den Berg, P. Barrera and G. Adema, "Increased expression of CCL18, CCL19, and CCL17 by dendritic cells from patients with rheumatoid arthritis, and regulation by Fc gamma receptors," *Annals of the rheumatic diseases,* vol. 64, no. 3, pp. 359-367, 2005.

[154] S. K. Bromley, T. R. Mempel and A. D. Luster, "Orchestrating the orchestrators: chemokines in control of T cell traffic," *Nature immunology,* vol. 9, no. 9, pp. 970-980, 2008.

[155] H. Liu, Y. Zhang, L. Li, J. Cao, Y. Guo, Y. Wu and W. Gao, "Fascin actin-bundling protein 1 in human cancer: Promising biomarker or therapeutic target?," *Molecular Therapy-Oncolytics,* vol. 20, pp. 240-264, 2021.

[156] J. Zhu, T. Long, L. Gao, Y. Zhong, P. Wang, X. Wang, Z. Li and Z. Hu, "RPL21 interacts with LAMP3 to promote colorectal cancer invasion and metastasis by regulating focal adhesion formation," *Cellular & Molecular Biology Letters,* vol. 28, no. 1, pp. 1-21, 2023.

[157] E. K. Aglago, A.-L. Mayén, V. Knaze, H. Freisling, V. Fedirko, D. J. Hughes, L. Jiao, A. K. Eriksen, A. Tjønneland, M.-C. Boutron-Ruault, J. A. Rothwell, G. Severi, R. Kaaks, V. Katzke, M. B. Schulze, A. Birukov, D. Palli, S. Sieri, M. S. de Magistris, R. Tumino, F.

Ricceri, B. Bueno-de-Mesquita, J. W. Derksen, G. Skeie, I. T. Gram, T. Sandanger, J. R. Quirós, L. Luján-Barroso, M.-J. Sánchez, P. Amiano, M.-D. Chirlaque, A. B. Gurrea, I. Johansson, J. Manjer, A. Perez-Cornago, E. Weiderpass, M. J. Gunter, A. K. Heath, C. G. Schalkwijk and M. Jenab, "Dietary advanced glycation end-products and colorectal cancer risk in the European prospective investigation into cancer and nutrition (EPIC) study," *Nutrients,* vol. 13, no. 9, p. 3132, 2021.

[158] X. Zhao, M. Lu, Z. Liu, M. Zhang, H. Yuan, Z. Dan, D. Wang, B. Ma, Y. Yang, F. Yang, R. Sun, L. Li and C. Dang, "Comprehensive analysis of alfa defensin expression and prognosis in human colorectal cancer," *Frontiers in Oncology,* vol. 12, p. 974654, 2023.

[159] D. Jeong, H. Kim, D. Kim, S. Ban, S. Oh, S. Ji, D. Kang, H. Lee, T. S. Ahn, H. J. Kim, S. B. Bae, M. S. Lee, C.-J. Kim, H. Y. Kwon and M.-J. Baek, "Defensin alpha 6 (DEFA6) is a prognostic marker in colorectal cancer," *Cancer Biomarkers,* vol. 24, no. 4, pp. 485-495, 2019.

[160] T. T. Nguyen, T. T. Ung, N. H. Kim and Y. D. Jung, "Role of bile acids in colon carcinogenesis," *World journal of clinical cases,* vol. 6, no. 13, p. 577, 2018.

[161] C.-C. Lee, X. Ding, T. Zhao, L. Wu, S. Perkins, H. Du and C. Yan, "Transthyretin stimulates tumor growth through regulation of tumor, immune, and endothelial cells," *The Journal of Immunology,* vol. 202, no. 3, pp. 991-1002, 2019.

[162] S. Wang, G.-T. Ruan, S. Xu, Z. Zhu, Y. Qing, C.-G. He, F.-B. Kong, C.-C. Dong, J.-W. Lin, X.-C. Zhang, B.-C. Zhou, Y.-Z. Gong, C. Liao, L. Zhang and L.-M. Pang, "Identification and Validation of HEPACAM family member 2 is Correlated with Immune Infiltration as a Prognostic Biomarker in colon adenocarcinoma," 2022.

[163] K. Lei, W. Li, C. Huang, Y. Li, L. Alfason, H. Zhao, M. Miyagishi, S. Wu and V. Kasim, "Neurogenic differentiation factor 1 promotes colorectal cancer cell proliferation and tumorigenesis by suppressing the p53/p21 axis," *Cancer Science,* vol. 111, no. 1, pp. 175-185, 2020.

[164] J. A. Rothwell, J. Bešević, N. Dimou, M. Breeur, N. Murphy, M. Jenab, R. Wedekind, V. Viallon, P. Ferrari, D. Achaintre, A. Gicquiau, S. Rinaldi, A. Scalbert, I. Huybrechts, C. Prehn, J. Adamski, A. J. Cross, H. Keun, M. Chadeau-Hyam, M.-C. Boutron-Ruault, K. Overvad, C. C. Dahm, T. H. Nøst, T. M. Sandanger, G. Skeie, R. Zamora-Ros, K. K. Tsilidis, F. Eichelmann, M. B. Schulze, B. van Guelpen, L. Vidman, M.-J. Sánchez, P. Amiano, E. Ardanaz, K. Smith-Byrne, R. Travis, V. Katzke, R. Kaaks, J. W. Derksen, S. Colorado-Yohar, R. Tumino, B. Bueno-de-Mesquita, P. Vineis, D. Palli, F. Pasanisi, A. K. Eriksen, A. Tjønneland, G. Severi and M. J. Gunter, "Circulating amino acid levels and colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition and UK Biobank cohorts," *BMC medicine,* vol. 21, no. 1, pp. 1-13, 2023.

[165] P. Czajka-Francuz, S. Cisoń-Jurek, A. Czajka, M. Kozaczka, J. Wojnar, J. Chudek and T. Francuz, "Systemic Interleukins' profile in early and advanced colorectal cancer," *International journal of molecular sciences,* vol. 23, no. 1, p. 124, 2021.

[166] P. Zhu, X. Zhu, J. Wu, L. He, T. Lu, Y. Wang, B. Liu, B. Ye, L. Sun, D. Fan, J. Wang, L. Yang, X. Qin, Y. Du, C. Li, L. He, W. Ren, X. Wu, Y. Tian and Z. Fan, "IL-13 secreted by ILC2s promotes the self-renewal of intestinal stem cells through circular RNA circPan3," *Nature Immunology,* vol. 20, no. 2, pp. 183-194, 2019.

[167] M. Bruchard and F. Ghiringhelli, "Deciphering the roles of innate lymphoid cells in cancer," *Frontiers in immunology,* vol. 10, p. 656, 2019.

[168] R. Molfetta and R. Paolini, "The controversial role of intestinal mast cells in colon cancer," *Cells,* vol. 12, no. 3, p. 459, 2023.

[169] X. Liu, X. Li, H. Wei, Y. Liu and N. Li, "Mast cells in colorectal cancer tumour progression, angiogenesis, and lymphangiogenesis," *Frontiers in Immunology,* vol. 14, 2023.

[170] W. Lvu, X. Fei, C. Chen and B. Zhang, "In silico identification of the prognostic biomarkers and therapeutic targets associated with cancer stem cell characteristics of glioma," *Bioscience Reports,* vol. 40, no. 8, p. BSR20201037, 2020.

[171] T. Vu and P. Datta, "Regulation of EMT in colorectal cancer: a culprit in metastasis," *Cancers,* vol. 9, no. 12, p. 171, 2017.

[172] G. Hassan and M. Seno, "Blood and cancer: cancer stem cells as origin of hematopoietic cells in solid tumor microenvironments," *Cells,* vol. 9, no. 5, p. 1293, 2020.

[173] L. Liang, J.-h. Zeng, X.-g. Qin, J.-q. Chen, D.-z. Luo and G. Chen, "Distinguishable prognostic signatures of left-and right-sided colon cancer: a study based on sequencing data," *Cellular Physiology and Biochemistry,* vol. 48, no. 2, pp. 475-490, 2018.

[174] T. Shen, J.-L. Liu, C.-Y. Wang, Y. Rixiati, S. Li, L.-D. Cai, Y.-Y. Zhao and J.-M. Li, "Targeting Erbin in B cells for therapy of lung metastasis of colorectal cancer," *Signal Transduction and Targeted Therapy,* vol. 6, no. 1, p. 115, 2021.

[175] L. McDougall, B. Kueh, J. Ward, J. D. Tyndall, A. G. Woolley, S. Mehta, C. Stayner, D. S. Larsen and M. R. Eccles, "Chemical Synthesis of the PAX Protein Inhibitor EG1 and Its Ability to Slow the Growth of Human Colorectal Carcinoma Cells," *Frontiers in Oncology,* vol. 11, p. 709540, 2021.

[176] K. M. Haigis, "KRAS alleles: the devil is in the detail," *Trends in cancer,* vol. 3, no. 10, pp. 686-697, 2017.

[177] F. L. Ruberg, M. Grogan, M. Hanna, J. W. Kelly and M. S. Maurer, "Transthyretin Amyloid Cardiomyopathy: JACC State-of-the-Art Review," *Journal of the American College of Cardiology,* vol. 73, no. 22, pp. 2872-2891, 2019.

[178] I. Arijs, G. De Hertogh, K. Lemaire, R. Quintens, L. Van Lommel, K. Van Steen, P. Leemans, I. Cleynen, G. Van Assche, S. Vermeire, K. Geboes, F. Schuit and P. Rutgeerts, "Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment," *PloS one,* vol. 4, no. 11, p. e7984, 2009.

[179] T. W. Mudd Jr, C. Lu, J. D. Klement and K. Liu, "MS4A1 expression and function in T cells in the colorectal cancer tumor microenvironment," *Cellular immunology,* vol. 360, p. 104260, 2021.

[180] L. Zheng, Y. Yang and X. Cui, "Establishing and validating an aging-related prognostic four-gene signature in colon adenocarcinoma," *BioMed Research International,* pp. 1-17, 2021.

[181] J. Stachelscheid, Q. Jiang and M. Herling, "The modes of dysregulation of the proto-oncogene T-cell leukemia/lymphoma 1A," *Cancers,* vol. 13, no. 21, p. 5455, 2021.

[182] H. Li, X. Yan, L. Liu, L. Huang, M. Yin, C. Pan, P. Zhang and H. Qin, "T-cell leukemia/lymphoma-1A predicts the clinical outcome for patients with stage II/III colorectal cancer," *Biomedicine & Pharmacotherapy,* vol. 88, pp. 924-930, 2017.

[183] G. Bansal, J. A. DiVietro, H. S. Kuehn, S. Rao, K. H. Nocka, A. M. Gilfillan and K. M. Druey, "RGS13 controls g protein-coupled receptor-evoked responses of human mast cells," *The Journal of Immunology,* vol. 181, no. 11, pp. 7882-7890, 2008.

[184] M. Salaga, M. Storr, K. A. Martemyanov and J. Fichna, "RGS proteins as targets in the treatment of intestinal inflammation and visceral pain: New insights and future perspectives," *BioEssays,* vol. 38, no. 4, pp. 344-354, 2016.

[185] M. Almutlaq, A. A. Alamro, H. S. Alamri, A. A. Alghamdi and T. Barhoumi, "The effect of local renin angiotensin system in the common types of cancer," *Frontiers in Endocrinology,* vol. 12, p. 736361, 2021.

[186] S. Tang, X. Yuan, J. Song, Y. Chen, X. Tan and Q. Li, "Association analyses of the JAK/STAT signaling pathway with the progression and prognosis of colon cancer," *Oncology letters,* vol. 17, no. 1, pp. 159-164, 2019.

[187] S. R. Nagarajan, L. M. Butler and A. J. Hoy, "The diversity and breadth of cancer cell fatty acid metabolism," *Cancer & Metabolism,* vol. 9, pp. 1-28, 2021.

[188] Y. Tang, S. Zong, H. Zeng, X. Ruan, L. Yao, S. Han and F. Hou, "MicroRNAs and angiogenesis: A new era for the management of colorectal cancer," *Cancer cell international,* vol. 21, pp. 1-11, 2021.

[189] H.-O. Lee, Y. Hong, H. E. Etlioglu, Y. B. Cho, V. Pomella, B. Van den Bosch, J. Vanhecke, S. Verbandt, H. Hong, J.-W. Min, N. Kim, H. H. Eum, J. Qian, B. Boeckx, D. Lambrechts, P. Tsantoulis, G. De Hertogh, W. Chung, T. Lee, M. An, H.-T. Shin, J.-G. Joung, M.-H. Jung, G. Ko, P. Wirapati, S. H. Kim, H. C. Kim, S. H. Yun, I. B. H. Tan, B. Ranjan, W. Y. Lee, T.-Y. Kim, J. K. Choi, Y.-J. Kim, S. Prabhakar, S. Tejpar and W.-Y. Park, "Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer," *Nature genetics,* vol. 52, no. 6, pp. 594-603, 2020.

[190] G. Zhu, L. Pei, H. Xia, Q. Tang and F. Bi, "Role of oncogenic KRAS in the prognosis, diagnosis and treatment of colorectal cancer," *Molecular cancer,* vol. 20, no. 1, pp. 1-17, 2021.

[191] K. Huang, W. Luo, J. Fang, C. Yu, G. Liu, X. Yuan, Y. Liu and W. Wu, "Notch3 signaling promotes colorectal tumor growth by enhancing immunosuppressive cells infiltration in the microenvironment," *BMC cancer,* vol. 23, no. 1, p. 55, 2023.

[192] P. Langfelder, R. Luo, M. C. Oldham and S. Horvath, "Is my network module preserved and reproducible?," *PLoS computational biology,* vol. 7, no. 1, p. e1001057, 2011.