CANCER DRIVER SUBNETWORK IDENTIFICATION BY NETWORK DISMANTLING METHODS

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

BY

FATMA ÜLKEM KASAPOĞLU

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

SEPTEMBER 2021

Approval of the thesis:

CANCER DRIVER SUBNETWORK IDENTIFICATION BY NETWORK DISMANTLING METHODS

submitted by **FATMA ÜLKEM KASAPOĞLU** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, Graduate School of **Informatics Institute** _____

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

Assoc. Prof. Dr. Nurcan Tunçbağ
Supervisor, **Chemical and Biological Engineering, Koç University, Health Informatics, METU** _____

Examining Committee Members:

Prof. Dr. Tolga Can
Computer Engineering, METU _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Chemical and Biological Engineering, Koç University, Health Informatics, METU _____

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

Assoc. Prof. Dr. Özlen Konu
Molecular Biology and Genetics, Bilkent University _____

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

Date: 07.09.2021

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

Name, Surname: Fatma Ülkem Kasapoğlu

Signature :

ABSTRACT

CANCER DRIVER SUBNETWORK IDENTIFICATION BY NETWORK DISMANTLING METHODS

Kasapoğlu, Fatma Ülkem
M.Sc., Department of Bioinformatics
Supervisor: Assoc. Prof. Dr. Nurcan Tunçbağ

September 2021, 92 pages

Tissue-specific protein protein interaction networks(TSPPI) are important for studying tissue-based cellular processes and protein functions. As the functions of tissues are various, the proteins they contain and the functions of these proteins also different than each other. Thus, these networks may help clinical studies both in tissue-specific cancer prediction by identification of the important groups in TSPPI. In this study, we aim to detect driver subnetworks in various TSPPIs and to understand effectiveness of the driver subnetworks. We performed iterative centrality attacks, Generalized Network Dismantling(GND), GND with Reinsertion(GNDR) on breast, liver, lymph node, ovary, and peripheral nerve TSPPIs from TissueNet v2 database. After the comparison by using driver gene information of each TSPPI from the Cancer Genome Interpreter and cBioPortal databases, we applied Personalized PageRank to each resulting TSPPI subnetworks to get more compact and robust subnetworks. Finally, we performed enrichment analysis for the proposed driver subnetworks. We found that genes in final subnetworks were enriched in both common and tissue-based cancer related pathways. As a result, we found that there was not optimal attack strategy for all TSPPIs. However, it is possible to obtain promising results for the cancer research by the comparison of these strategies.

Keywords: Tissue-Specific Protein Protein Interaction Networks, Subnetwork Identification, Network Attacks, Network Dismantling

ÖZ

AĞ SÖKME YÖNTEMLERİNİ KULLANARAK KANSER SÜRÜCÜSÜ ALT AĞ TANIMLAMA

Kasapoğlu, Fatma Ülkem
Yüksek Lisans, Biyoenformatik Bölümü
Tez Yöneticisi: Doç. Dr. Nurcan Tunçbağ

Eylül 2021 , 92 sayfa

Dokuya özgü protein protein etkileşim ağları (TSPPI), doku bazlı hücresel süreçleri ve protein fonksiyonlarını incelemek için önemlidir. Dokuların işlevleri çeşitli olduğu için içerdikleri proteinler ve bu proteinlerin işlevleri de birbirinden farklıdır. Böylece bu ağlar, TSPPI'deki önemli grupları tanımlayarak hem dokuya özgü kanser tahmininde klinik çalışmalara yardımcı olabilir. Bu çalışmada, çeşitli TSPPI'lerdeki sürücü alt ağlarını tespit etmeyi ve sürücü alt ağlarının etkinliğini anlamayı amaçlıyoruz. TissueNet v2 veritabanından meme, karaciğer, lenf düğümü, yumurtalık ve periferik sinir TSPPI'lerinde yinelemeli merkezilik saldırıları, Genelleştirilmiş Ağ Dağıtma (GND), Yeniden Yerleştirmeli GND (GNDR) gerçekleştirdik. Cancer Genome Interpreter ve cBioPortal veritabanlarından alınan her TSPPI'nin sürücü gen bilgilerini kullanarak karşılaştırmadan sonra, daha kompakt ve sağlam alt ağlar elde etmek için ortaya çıkan her TSPPI alt ağına Kişiselleştirilmiş PageRank uyguladık. Son olarak, önerilen sürücü alt ağları için zenginleştirme analizi gerçekleştirdik. Nihai alt ağlardaki genlerin hem yaygın hem de doku bazlı kanserle ilgili yollarda zenginleştiğini bulduk. Sonuç olarak, tüm TSPPI'ler için optimal saldırı stratejisi olmadığını bulduk. Ancak bu stratejilerin karşılaştırılması ile kanser araştırmaları için umut verici sonuçlar elde etmek mümkündür.

Anahtar Kelimeler: Dokuya Özgü Protein Protein Etkileşim Ağları, Alt Ağ Tanımlama, Ağ Saldırıları, Ağ Sökülmesi

To the Civriz family

ACKNOWLEDGMENTS

Primarily, I would like to thank my dear supervisor Assoc. Prof. Dr. Nurcan Tunçbağ for conveying her valuable information to me, providing a solution to all my questions, making this process bearable and her understanding to me. It would not have been possible for me to come up with this thesis without her guidance. I am very lucky to have such an advisor. Her insightful feedback pushed me to sharpen my thinking.

I would also like to extend my deepest gratitude to my thesis committee members Prof. Dr. Tolga Can, Assoc. Prof. Dr. Yeşim Aydın Son, Assoc. Prof. Dr. Özlen Konu and Assist. Prof. Dr. Aybar Can Acar for taking time to attend my thesis dissertion and their informative comments.

In addition, I would like to acknowledge, my collegues from Digital Transformation Office of the Presidency of the Republic of Turkey, Dr. Umut Demirezen, Dr. Ramazan Terzi, Assoc. Prof Dr. Tolga Çakmak, Bilge Miraç Atıcı, Seda Özlü, Alican Aşan, Ahmet Enes Kılıç and Öyküm Demir for their patient support.

Special thanks to Hikmet Baş for always being by my side and never letting me down. I also had great pleasure of being friends with Eda Tosun, Burak Kızıl, Pınar Arıkoğlu, Baran Barış Kıvılcım, Merve Bozo and Mustafa Batuhan Çevirgen. They always supported and nurtured me.

Finally, and most importantly, I am grateful to my cousin, Sinem Civriz Bozdağ, for always being there for me whenever I need support and being a shoulder to cry. In short, I would like to thank her for acting as a sister and as well as a friend. I would like to also thank Gürkan Bozdağ for his patience to all my concerns and Yiğit Bozdağ for giving joy to our lives. I'm deeply indebted to Ayşe Civriz, Hüseyin Civriz, Şemsünnisa Civriz, and Fadime Civriz. I am grateful for their profound belief in my studies. I could not have completed this thesis without the support of them.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AP	Articulation Point
CGI	Cancer Genome Interpreter
CGP	Cancer Genome Project
COSMIC	Catalogue of Somatic Mutations in Cancer
CRM	Cis-Regulatory Modules
DAVID	The Database for Annotation, Visualization, and Integrated Discovery
IID	Integrated Interaction Database
GCC	Giant Connected Component
GEO	The Gene Expression Omnibus
GND	Generalized Network Dismantling
GNDR	Generalized Network Dismantling with Reinsertion
GO	Gene Ontology
GTEx	Genotype-Tissue Expression
GWAS	Genome-Wide Association Study
HGNC	HUGO Gene Nomenclature Committee
HPA	Human Proteome Atlas
ICGC	International Cancer Genome Consortium
IMEx	The International Molecular Exchange Consortium
KEGG	Kyoto Encyclopedia of Genes and Genomes
MEMo	Mutual Exclusivity Modules in cancer
ND	Network Dismantling
NetWAS	Network-Wide Association Study
PARADIGM	Pathway Recognition Algorithm using Data Integration on Genomic Models
PPI	Protein-Protein Interaction
PPIN	Protein-Protein Interaction Network
PRIDE	The Proteomics Identifications Database
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
TieDIE	Tied Diffusion Through Interacting Events
TIGER	Tissue-specific Gene Expression and Regulation
TSPPI	Tissue-Specific PPI Network
tsv	Tab-Seperated Values

CHAPTER 1

INTRODUCTION

Protein-protein interactions (PPIs) are to perform almost all molecular mechanisms in cell such signaling pathways, cell to cell communication, replication, transcription etc.; therefore, understanding PPIs is crucial to understanding cell dynamics. It may also be considered to predict and understand the possible effects of drugs on the cell (Mabonga & Kappo, 2019). Recent developments in biotechnology have resulted in PPI data created with different perspectives. Modeling PPI networks (PPINs) with simple graphs enables us to understand the basic components of cell physiology and from the organizational network level. Many public databases have been created to ensure that researchers can easily access this information. In addition to the standard guidelines to store PPI data, many primary and meta-databases have been created by The International Molecular Exchange consortium (IMEx) (Orchard et al., 2012).

Recently, it has gained popularity that biological networks can contribute to cancer studies by deciphering biological mechanisms. It has been shown to be very useful for the prediction therapeutic responses at both the molecular and systems levels. Biological networks are generally scale-free, meaning that a few nodes have huge number of links, called hub nodes, and the remaining nodes have only one or few links (Barabási & Bonabeau, 2003). Hub genes/proteins in biological networks tend to be essential (Goymer, 2008).

PPINs represent physical contacts between proteins. Tissue-specific PPI network (TSPPI) may be defined as a subnetwork of a PPIN that includes proteins expressed in a particular tissue (Bossi & Lehner, 2009). To generate a TSPPI, a PPIN is needed as a template and expression data for each protein. Since an expression dataset provides protein expression data for multiple tissues, multiple TSPPIs can be generated, one per tissue for the PPIN and the expression dataset.

Cancer types show behavioral changes from tissue to tissue. Likewise, TSPPIs are quite different in terms of both hub genes, also called 'driver genes', and topological features (Stratton, Campbell, & Futreal, 2009). Different structures produce different results as a result of network attacks e.g. mutations for a PPIN.

In network science, dismantling a network by attacking minimum number of nodes is a crucial problem. The adapted version of the problem for TSPPIs is that when a gene or a set of genes is mutated, a pathway that can lead to cancer may be disrupted. In this case, the problem evolves to find cancer driver genes in the TSPPI or genes that interact with them. This is called "driver subnetwork". It should be noted that

finding driver subnetworks is an NP-hard problem (Hormozdiari, Salari, Bafna, & Sahinalp, 2010). Thus, there is not any optimal solution for this problem. In this thesis, we compared several approaches to extract the best subnetworks for various TSPPIs. We tried to attack the nodes with the highest betweenness, closeness, degree, or eigenvector centralities first in each iteration. As another method, we implemented various ways of the weight function of Generalized Network Dismantling (GND) (Ren, Gleinig, Helbing, & Antulov-Fantulin, 2019), which was created using the node degrees of the network, by replacing it with these centrality measures. Then, we reinserted the nodes that are unnecessarily removed from the GND algorithm. Finally, we evaluated our results by using 5 different TSPPIs consisting of breast, liver, lymph node, peripheral nerve and ovary. We used driver node information from the Cancer Genome Interpreter (Tamborero et al., 2018) and cBioPortal (Gao et al., 2013) in the method evaluation phase. After determining the best attack strategy for each TSPPI seperately according to their resulting subnetworks and driver node counts, we applied Personalized PageRank (Jeh & Widom, 2003) to resulting subnetworks, produced by selected strategy, in order to extract a smaller, meaningful, and compact subnetwork for each tissue. We conducted functional enrichment analysis to show the resulting proposed subnetworks include both tissue-related and common cancer pathways.

In Chapter 2, we present a detailed literature review on the subject, starting with the knowledge of the data sources used. After introducing network centrality measures, we review network dismantling algorithms, biomarker identification and driver subnetwork construction algorithms.

In Chapter 3, we present our step-by-step methodology. First of all, we explain how we obtain and process data on TSPPI and driver gene information for each TSPPI. After that, we introduce details about the algorithms in our pipeline. We clarify how we conduct the functional enrichment analysis on these results.

In Chapter 4, we start by giving the dataset statistics. We show that the TSPPIs are quite different from each other. Then, we present all results seperately for each TSPPI. We present method evaluation results to find best subnetwork to continue with. We visualize Personalized PageRank result statistics and resulting networks. Finally, we infer final functional enrichment results.

As a conclusion, we interpret our results while giving an overview of our pipeline. Finally, we discuss our resulting networks. Through this study, we have shown that there is no single optimal attack strategy and that even though TSPPIs are generated from the same interactome, the impact of each of these strategies is quite different, as there are great topological differences between them.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we review available tissue-specific data sources, network dismantling algorithms, and biomarker and driver subnetwork identification studies in the literature. Also, we refer to the topological features of a network and databases used in this study.

2.1 Tissue-Specific PPI Data Sources

2.1.1 Tissue-specific Gene Expression and Regulation (TIGER) Database

TIGER (http://bioinfo.wilmer.jhu.edu/tiger/) is a publicly available relational database that accommodates broad information on tissue-specific gene regulation regarding both expression and regulatory data which include tissue-specific expression profiles, combinatorial editing for interacting transcription factor (TF) pairs, and cis-regulatory modules (CRM) for tissue-specific genes. (Liu, Yu, Zack, Zhu, & Qian, 2008)

2.1.2 SPECTRA

SPECTRA (http://alpha.dmi.unict.it/spectra) integrates 13 authoritative sources including Protein Atlas (Uhlén et al., 2005), ArrayExpress (Brazma et al., 2003), The Gene Expression Omnibus (GEO) (Barrett et al., 2007) and The Cancer Genome Atlas (TCGA)(Weinstein et al., 2013), then combines human PPIs with gene expressions. This resulting database contains protein-coding genes and gene interactions (GIs) and tissue-specific gene expression data entries.

2.1.3 Integrated Interaction Database (IID)

IID (http://ophid.utoronto.ca/iid) is a database that contains 6 living species including yeast, worm, fly, rat, mouse, and human. There is a maximum of 30 tissue per organism. It consists of experimentally and computationally predicted PPIs. Users query IID by giving a set of proteins or interactions, and tissues of interest from either of these organisms. (Kotlyar, Pastrello, Sheahan, & Jurisica, 2016)

2.1.4 TissueNet v.2

TissueNet was one of the first databases to enable tissue-sensitive views of PPIs. TissueNet provided comprehensive views of 16 major human tissues by integrating gene and protein expression profiles of human tissues into a combined expression dataset. Importantly, in the output network, TissueNet implemented tissue-specific or global proteins and presented with it an informative view of TSPPIs. The TissueNet database establishes the link between PPIs and tissues (Barshir et al., 2013).

TissueNet v.2 (http://netbio.bgu.ac.il/tissuenet) is an improved version of the first version. It includes 40 TSPPIs which are profiled through RNA sequencing or protein based assay. By the updates to TissueNet, users can dynamically change the resulting network and enhance the appearance of the tissue specificity of the proteins, by the flexibility to interactively select the source of expression for tissue associations and adjust the expression threshold(Basha et al., 2017).

2.2 Cancer Genome Interpreter(CGI)

The Cancer Genome Interpreter (http://www.cancergenomeinterpreter.org), is used to interpret the probability of a mutation being a driver mutation and its treatment automatically. The results are arranged at different levels of evidence based on the available information that may support a wide variety of oncology use cases (Tamborero et al., 2018).

2.3 cBioPortal

cBioPortal is a publicly available platform for exploratory analysis and interactive visualization of large-scale cancer genomics datasets. cBioPortal gives the advantage of integration of multiple analysis tools such as mutual exclusivity calculation, survival analysis, mutated gene identification from clinical data, sample comparison of different groups. It is freely available at https://www.cbioportal.org/ (Gao et al., 2013).

2.4 BioMart

BioMart provides an integrated searching on many biological data sources. BioMart is extended through integration with various commonly used tools such as BioConductor (Gentleman et al., 2004), Galaxy (Giardine et al., 2005), Cytoscape (Shannon et al., 2003), Taverna (Oinn et al., 2004). BioMart is an convenient and universal system and an integral part of big data sources such as Ensembl (Hubbard et al., 2002), UniProt (Apweiler et al., 2004), HapMap (Gibbs et al., 2003), Wormbase (Stein, Sternberg, Durbin, Thierry-Mieg, & Spieth, 2001), Gramene (Ware et al., 2002), Dictybase (Chisholm et al., 2006), the proteomics identifications database (PRIDE)

(Martens et al., 2005) and Reactome (Joshi-Tope et al., 2005). BioMart is freely accessible at http://www.biomart.org (Smedley et al., 2009).

2.5 Topological Features of a Network

The molecular causes of diseases are investigated by many different techniques such as the study of their driver genes, disruption of related pathways, and various other external influences. Disruption of certain genes in a biological process affects the organism more than other genes. Likewise, critical areas in a network mean a lot to network connectivity. The scientific field that tries to address this coupling is biological networks. For example, there are several applications on PPINs (Jeong, Mason, Barabási, & Oltvai, 2001) (Fraser, Hirsh, Steinmetz, Scharfe, & Feldman, 2002) (Eisenberg & Levanon, 2003) (Saeed & Deane, 2006) (Jordan, Wolf, & Koonin, 2003) (Wachi, Yoneda, & Wu, 2005) (Xu & Li, 2006).

Network topology is the element structure of a network. Network centrality assigns different node rankings within a network according to their localization. For example, identification of the most influential person(s) in a social network (Zhao, Liu, Wang, Li, et al., 2017), super-spreaders of diseases and brain networks(van den Heuvel & Sporns, 2013) (Saberi, Khosrowabadi, Khatibi, Misic, & Jafari, 2021) are some applications. There are various network centrality metrics in the literature.

Degree centrality (Bell, Atkinson, & Carlson, 1999) is the simplest network centrality measure that means ratio of number of edges connected to a node to the number of all edges in undirected graphs. Corresponding centralities are calculated by dividing them into the total number of edges.

Closeness centrality (Freeman, 1978) is the inverse of the average shortest path distance to that a node over all accessible nodes. Larger closeness centrality indicates a less central node while a smaller closeness centrality score indicates a more central node.

Betweenness centrality (Freeman, 1977) measures the influence of a node on the information flow in a graph. Thus, it is often used to find important nodes connecting two parts of a graph. It is calculated with respect to the number of times a node is located on the shortest paths between any pair of nodes using the Brandes algorithm (Brandes, 2001).

Percolation centrality (Piraveenan, Prokopenko, & Hossain, 2013) is the distribution of 'percolated paths' through a node. The percolation states are values in the range 0.0 to 1.0.

Eigenvector centrality (Ruhnau, 2000) is another measure of the impact of a node. According to this metric, connections to high scoring nodes give more to that node's score than connections with lower scores, relative scores depend on the impact of their neighbors. PageRank (Langville & Meyer, 2011) and the Katz centrality (Katz, 1953) are modifications of this measure.

2.6 Network Dismantling

Network Dismantling is a key challenge in complex networks to find a minimum collection of vertices where removal of this collection results in subnetworks which may be the most comprehensive and robust connected sub-components available. There is no solution in polynomial time because of its NP-complete complexity due to the presence of non-trivial connectivity patterns (Braunstein, Dall'Asta, Semerjian, & Zdeborová, 2016).

One of the most focused issues regarding the dismantling of the network has been the decision of its ability to continue with its expected function after a major collapse. Thus, the links between these vertices was affected by the removal of nodes, as a result of random or targeted processes. Then, the problem of identifying the most effective strategies to disrupt a network arose. Primarily, it makes move from evaluation to resilience building by predicting where a clever attacker might attack and thus determining where should be conserved at first glance. Second, there are situations where it is needed to disrupt a network, such as stopping the spread of a computer virus or transmission of a disease (Wandelt, Sun, Feng, Zanin, & Havlin, 2018).

A common practice in identifying the best attack is to use measures of degree, betweenness, closeness, and eigenvector node centrality or special network removal techniques such as Min-Sum (Braunstein et al., 2016), articulation points (Tian, Bashan, Shi, & Liu, 2017), and the Laplacian operator (Ren et al., 2019).

Network topology based network dismantling methods propose attacking nodes with respect to their topological features, i.e. different centrality measures and degree-based weight calculations. According to the Girvan-Newman method (Girvan & Newman, 2002), based on the edge betweenness, a network is gradually reduced by discarding edges having the highest betweenness centralities until the required number of subnetworks is reached. Another method based on centrality (Holme, Kim, Yoon, & Han, 2002) introduces a dynamic iterative calculation by taking into account the changes in both degree and betweenness centralities of nodes after each attack. Coreness (Kitsak et al., 2010) is a measure that can help identify strongly connected groups within a network. K-shell iteration factor (Wang, Zhao, Xi, & Du, 2016) is another topological measure that is based on the node coreness. The k-shell iteration algorithm integrates shell decomposition to iterative node removal to use variations in the neighborhood. CoreHD attacks (Zdeborová, Zhang, & Zhou, 2016) is a modified version of k-shell iteration algorithm to achieve a dismantled network (Altarelli, Braunstein, Dall'Asta, & Zecchina, 2013a) (Altarelli, Braunstein, Dall'Asta, & Zecchina, 2013b). The highest degree node with the 2-core graphs is iteratively removed until there are no 2-core graphs left. Then, the algorithm continues with processing the remaining part by tree breaking. Influence Maximization through optimal percolation (Morone & Makse, 2015), is a method using topological features, designed for determining the optimal structural node set onto optimal percolation in random networks. It considers the neighbors at a pre-determined distance.

The Articulation Points Targeted Attack (Tian et al., 2017) is one of the special network removal techniques established in the literature. An articulation point (AP) removal breaks down the network connectivity. All APs can be identified by performing

a linear-time algorithm based on depth-first search. From the community perspective, community-based attacks (Requião da Cunha, González-Avella, & Gonçalves, 2015) rely on identifying the communities that compose a network based on the concept of modularity. Louvain method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) is used to extract communities. After a community structure is extracted from the network, inter-community links are determined and removed iteratively by calculating the betweenness centrality of the nodes. For the resulting subnetworks, the nodes are ordered by their degree ranks. Another algorithm, specifically designed for dismantling problem, is Network Dismantling (ND) (Braunstein et al., 2016). There is an assumption in ND algorithm that the problem of dismantling a network is related to decycling. A three-stage Min-Sum algorithm consisting of decycling, tree breaking and loop closure is proposed. First, there is a variant of the Min-Sum message-passing algorithm for decycling, which is the most important part of the algorithm. After loops have been broken, some tree components may not satisfy the desired threshold. These subnetworks are subdivided into smaller subnetworks, eliminating some of the nodes lost. Finally, loops are greedily closed to increase the efficiency. Generalized network removal (GND) (Ren et al., 2019) is separated from ND via "node-specific costs". When cost for each node is calculated as unit cost, GND turns into the standard ND algorithm. GND uses a spectral approximation by a power laplacian operator. This process is called "node-weighted spectral cut" and it is implemented recursively until the network consists of smaller subnetworks.

2.7 Identification of Biomarkers and Driver Subnetworks in Biological Networks

Large-scale cancer catalogs of genomic alterations such as Cancer Genome Project (CGP) (Dickson, 1999), TCGA (Weinstein et al., 2013), International Cancer Genome Consortium (ICGC) (J. Zhang et al., 2011), and The Catalogue of Somatic Mutations in Cancer (COSMIC) database (Forbes et al., 2010) address molecular disruptions in different types of cancer. There are two important problems to be solved in these studies. The first is to identify mutations in driver genes, which are usually important biomarkers, and the second is the identification of biological pathways that are often disrupted in tumor cells and cause the acquisition of tumorigenic features (Weinberg & Hanahan, 2000).

The simplest method for predicting driver nodes in a biological network is to rank all genes according to a particular local topological feature (e.g affinity centrality, PageRank centrality, degree centrality). On the other hand, biological networks are very complex and therefore the application of the methods using only one or more local topological features is insufficient (Cámara, 2017).

To identify cancer-associated genes is the combination of different types of data about types of tumors (Cheng, Zhao, & Zhao, 2016). If there are consistent changes in gene expression, or DNA as a mutation occurs, it is likely that this mutation is associated with cancer. Several studies have used changes in copy number, expression, or methylation of the mutated gene to identify new cancer-associated mutations (Leary et al., 2008; Ding et al., 2008; Frattini et al., 2013). Recently, novel algorithms and frameworks have been developed by the data integration perspective to find op-

timal set of driver genes. DriverNet (Bashashati et al., 2012) is a framework created to identify possible driver mutations through their effects on mRNA expression networks. CONEXIC (Akavia et al., 2010) is a computational scheme that combines expression and chromosomal copy number data to detect abnormality that favor cancer. AMARETTO (Gevaert, Villalobos, Sikic, & Plevritis, 2013) is an algorithm that initially models the relationship between genomic data and disease-specific gene expression and then involves building a network of modules to link cancer drivers to downstream targets. OncoIMPACT (Bertrand et al., 2015) identifies patient-specific driver genes by integrating structural and copy number mutations and resulting disruptions in transcriptional process.

Networks may often be easily divided into modules or subnetworks by cutting them at their weakly-connected edges (Newman, 2006). Subnetwork construction usually starts with an exploration with one or more genes and expansion to their neighbors. This process continues until a specific threshold is reached. The resulting subnetworks may contribute in diseases or drug treatments by revealing disease effects or therapeutic responses. There are different perspectives to solve the driver subnetwork construction/identification problem.

First approach to identify driver subnetworks is to define a reduced number of candidate subnetworks and select the high-scoring subnetworks based on node scores. jActiveModules (Ideker, Ozier, Schwikowski, & Siegel, 2002) method proposes to combine a proper statistical measure to rank subnetworks with a search algorithm to identify high-scoring subnetworks. BioNet (Beisser, Klau, Dandekar, Müller, & Dittrich, 2010) is an R package that integrates multiple P-values from various experiments, assign rankings to nodes with a modular scoring function, calculates optimal and nonoptimal solutions, calculates high-scoring solutions, and finally visualizes these solutions in both 2 dimensions and 3 dimensions. Another example of functional module identification on PPINs (Dittrich, Klau, Rosenwald, Dandekar, & Müller, 2008) represents a method which is a combination of integer programming and prize-collecting Steiner tree with a novel scoring function. Omics Integrator (Tuncbag et al., 2013) uses omics data to reconstruct multiple paths changed under a given condition by solving a prize-collecting Steiner forest problem. MUFFINN (MUtations For Functional Impact on Network Neighbors) (Cho et al., 2016) combines gene mutations and the neighbors of these genes in functionally-associated networks.

In addition to the first approach that defines subnetworks via node scores, in the second approach, network topology is also included in the subnetwork scoring in the literature. For example, in a study, a tool is presented that can conduct protein function predictions by extracting the shortest paths calculated with the new metric between each protein pair, with Diffusion state distance, which is designed to grab subtle separations in proximity for the transmission of functional descriptions in PPINs (Cao et al., 2013). NetWalker (Komurov, Dursun, Erdin, & Ram, 2012) is an integrated platform designed for comparative interpretations of genomic data by using a number of comprehensive data integration methods. It presents a random walk-based novel method that gives analysis capabilities to evaluate data distributions and connectivity to compute priorities of subnetworks. HotNet (Vandin, Upfal, & Raphael, 2011) uses an undirected heat diffusion process. TieDIE (Paull et al., 2013) uses manually curated differentially expressed genes and mutated genes to find driver subnetworks by using a directed heat diffusion strategy. HotNet2 (Leiserson et al., 2015) is an algo-

rithm based on an isolated heat diffusion kernel algorithm to find altered subnetworks using both the heats of the genes and the PPIN topology.

By linking additional information to node scores and network topology information, another approach is applied to biological networks. PARADIGM (Vaske et al., 2010) proposes an algorithm by modelling each gene with a factor graph of expression variables and known activity of a gene and gene products, allowing for the inclusion of omics data. MEMo (Ciriello, Cerami, Sander, & Schultz, 2012) uses correlation and statistical analysis by determining genes that have mutually exclusive alterations across a set of tumor samples from the same biological process.

CHAPTER 3

MATERIALS AND METHODS

In this section, we explain the general structure, from the acquisition of TSPPI and driver node information data to the use of different methods, namely algorithms using topological features and the GND method.

3.1 Overview of the Pipeline

In this study, we first download TSPPIs from TissueNet v.2 (Basha et al., 2017), compare different algorithms and weight functions on TSPPIs and analyze them to determine the best method for each tissue. To analyze it, we first need to get the nodes of the network that may be drivers. For this reason, we extract tissue-specific biomarkers and common biomarkers from CGI (Tamborero et al., 2018). Then, in order to find out the occurrence frequency of the common biomarkers play the role of driver gene in this cancer, we remove the genes with 0% frequency but marked as biomarkers from the driver gene list by reaching the incidence in patients through cBioPortal (Gao et al., 2013). We use the biomarker information to determine the optimal subnetwork among the dismantled networks and to determine the driver subnetworks for each tissue in the initial large networks by applying the Personalized PageRank algorithm starting from the driver nodes on the selected subnetworks. Finally, we evaluated our results by functional enrichment analysis via DAVID (Dennis et al., 2003).

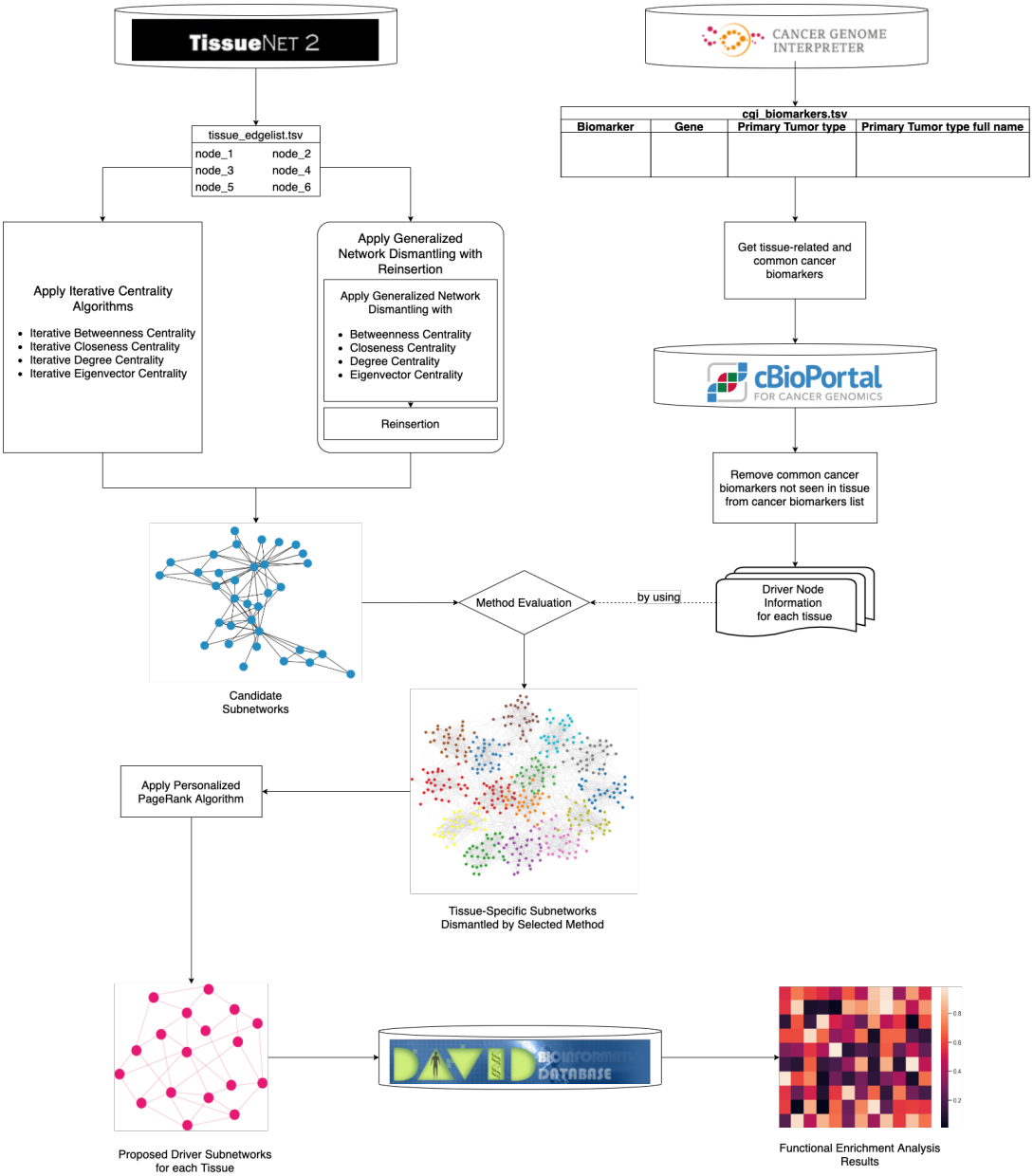


Figure 3.1: Summary of our methodology

3.2 Dataset

3.2.1 Data from TissueNet v2

The TSPPI data in TissueNet v2 are located in the "Download" Section of its website. There are tissues belongs to Homo sapiens profiled via RNA-sequencing or protein-based assays. Data from RNA-sequencing were obtained by Human Proteome Atlas (HPA) (Uhlén et al., 2005) or Genotype-Tissue Expression (GTEx) (Consortium et al., 2015). We downloaded TSPPI data generated by using HPA protein-based assays.

The downloaded zip folder contains 47 tab-seperated values (tsv) files. Each of these files is the edgelist of the tissue written in the name of the file. TSPPI information representation consists of Ensembl Gene IDs.

3.2.2 Data from CGI

The cancer biomarkers in CGI are manually-curated. This database is provided through two tsv files. In the cgi_biomarkers.tsv, variants that constitute the biomarkers of the same type of response to the same drug in a certain cancer type are grouped in the same order. In the cgi_biomarkers_per_variant.tsv, each variant is written in a separate row. In addition to the first file, genomic and protein coordinates of variants are included in Strand, Region, Info, cDNA, and gDNA columns.

Firstly, we downloaded these two files from https://www.cancergenomeinterpreter.org/biomarkers. The materials in this page is shown in the Figure 3.2.

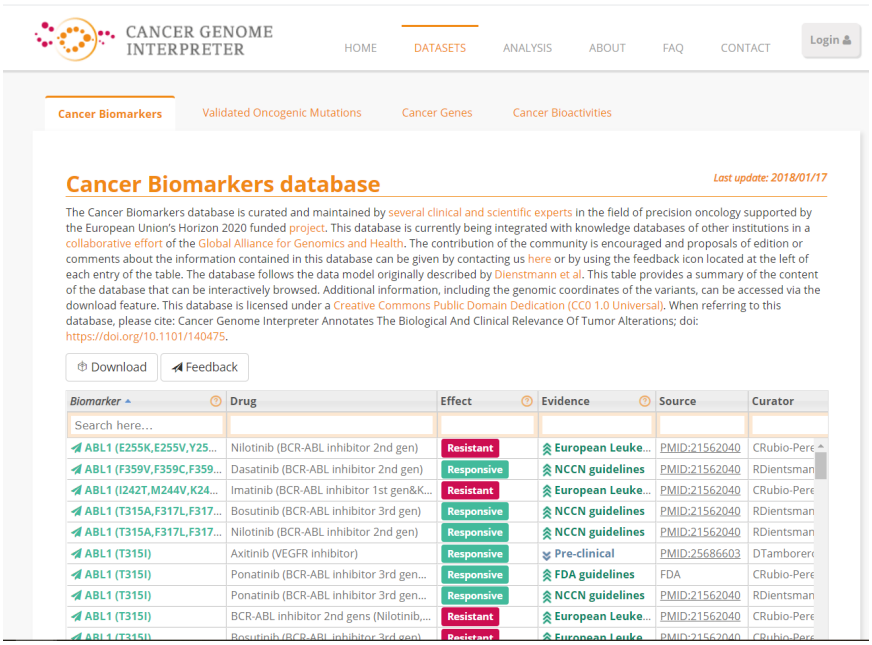


Figure 3.2: The dataset page of the Cancer Biomarkers database.

Since cgi_biomarkers.tsv file is enough to extract driver nodes for selected TSPPI networks, we get only the data from "Gene" and "Primary Tumor type full name" columns from this file. Keywords used to find driver genes for the related TSPPI are shown in Table 3.1. Values in the 'Gene' column is represented in HUGO Gene Nomenclature Committee (HGNC) symbol.

Table 3.1: Tissue Related Tumor Type Search Keywords in CGI

Tissue	Primary Tumor Type
Breast	Breast Adenocarcinoma & Any Cancer Type
Liver	Hepatic Carcinoma & Any Cancer Type
Lymph Node	Lymphoma & Any Cancer Type
Ovary	Ovary & Any Cancer Type
Peripheral Nerve	Malignant Peripheral Nerve Sheat & Any Cancer Type

3.2.3 Data from cBioPortal

cBioPortal provides information on each gene frequency of the given gene list appears in the samples in the selected study. In this study, we determined how often genes associated with tissue-related cancer type and any cancer type from CGI were mutated in clinical datasets in cBioPortal. Then, we select the final driver gene set, accordingly.

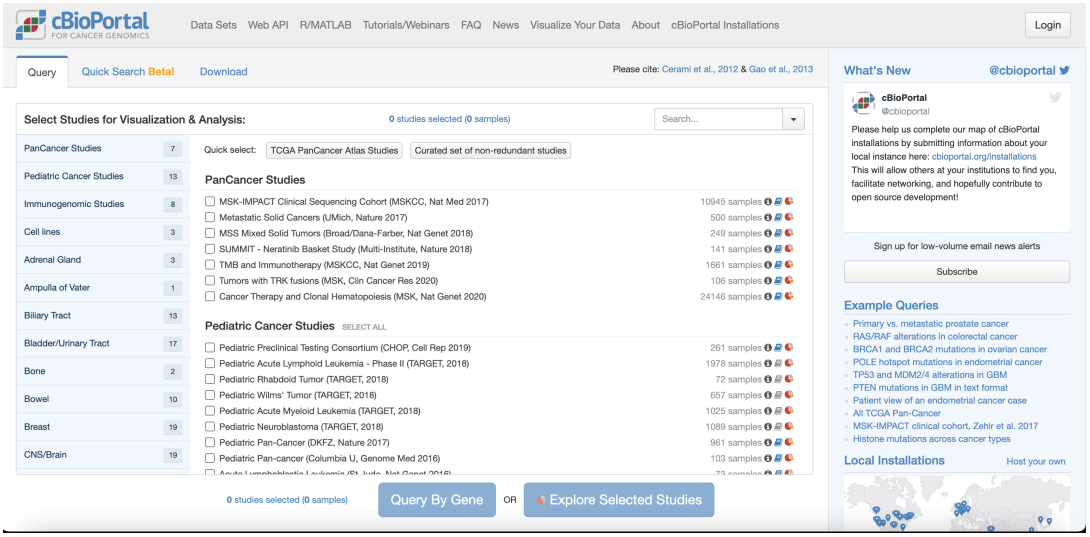


Figure 3.3: Search Query Page of cBioPortal Web Interface.

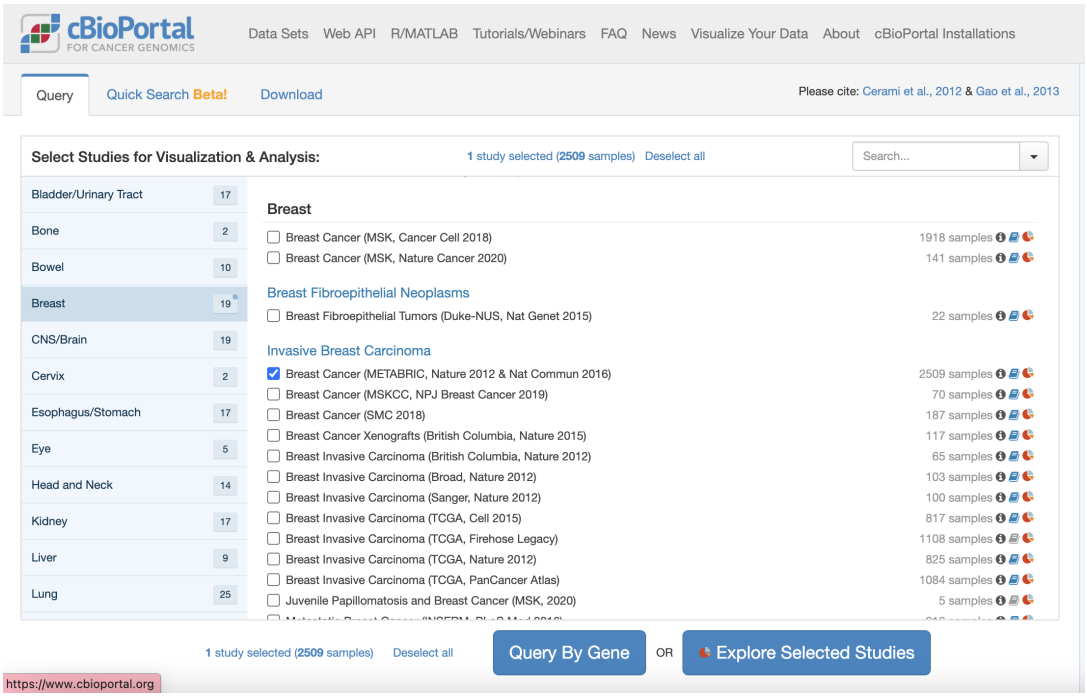


Figure 3.4: Study Selection for Breast Carcinoma

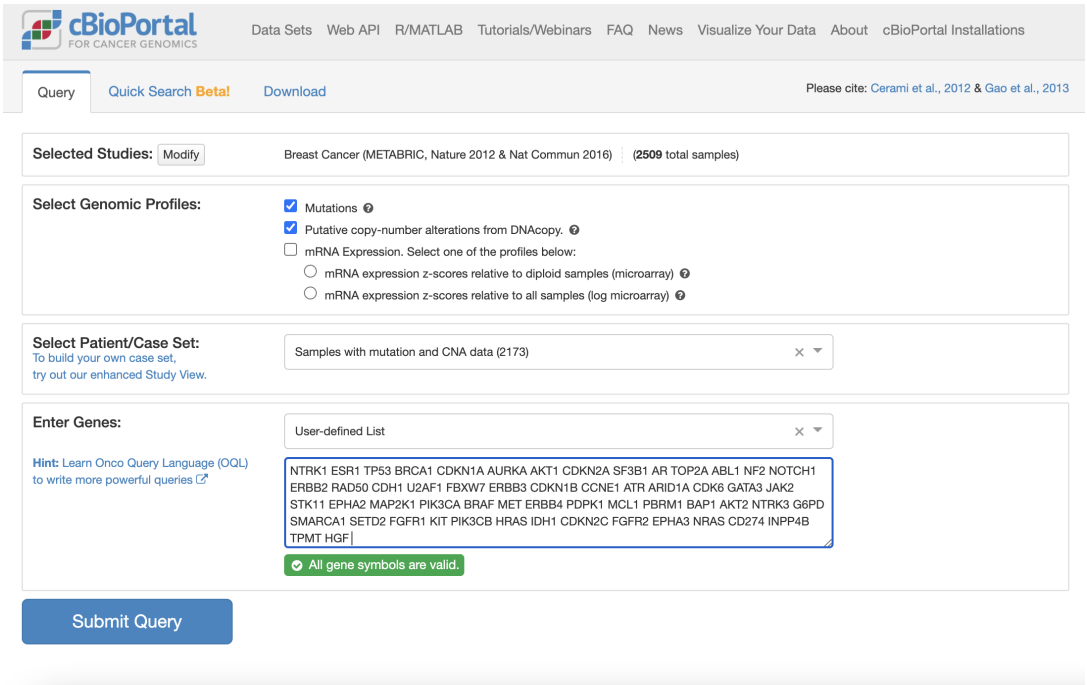


Figure 3.5: Query construction for driver nodes of breast carcinoma retrieved from CGI.

We query all driver gene information to check whether mutations related with these genes occur frequently. The studies to find frequency information are given in Table

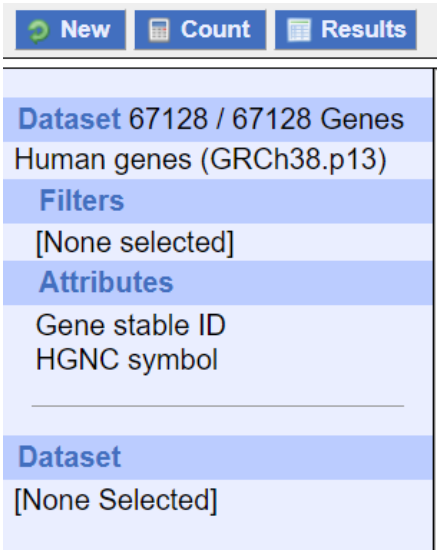
3.2

Table 3.2: Related Studies from cBioPortal.

Tissue	Study Subject	Study
Breast	Breast Carcinoma	METABRIC, Nature 2012 and Nat Commun 2016
Liver	Hepatocellular Carcinoma	TCGA, Firehose Legacy
Lymph Node	Lymphoma Cell Lines	Memorial Sloan-Kettering Cancer Center,2020
Peripheral Nerve	Malignant Peripheral Nerve Sheath Tumor	Memorial Sloan-Kettering Cancer Center, Nat Genet 2014
Ovary	Ovarian Serous Cystadenocarcinoma	TCGA, Firehose Legacy

3.3 Cross-referencing between Ensembl Gene IDs and HGNC Symbols

While nodes from TSPPI edgelist are represented in Ensembl Gene ID format, driver nodes from both CGI and cBioPortal are denoted by HGNC symbols. Then, we make a cross-reference between them via BioMart. The results shown in Figure 4.1b are exported with 67128 genes as a tsv file. After downloading data, we delete the NULL values. There are 43841 Ensembl Gene ID-HGNC Symbol pairs left after this step.



(a) Search Query Information

Gene stable ID	HGNC symbol
ENSG00000210049	MT-TF
ENSG00000211459	MT-RNR1
ENSG00000210077	MT-TV
ENSG00000210082	MT-RNR2
ENSG00000209082	MT-TL1
ENSG00000198888	MT-ND1
ENSG00000210100	MT-TI
ENSG00000210107	MT-TQ
ENSG00000210112	MT-TM
ENSG00000198763	MT-ND2

(b) Search Results

Figure 3.6: Cross-referencing between Ensembl Gene ID and HGNC symbols via BioMart

3.4 Network Similarity Measurement

In this study, we used Jaccard Similarity Index to compare similarities and differences of each TSPPI. In fact, this method was created for sets and it may be applied separately to the sets of nodes and edges of the networks. Given G(V_X, E_X) and H(V_Y, E_Y) are two networks with V_X vertices and E_X edges, V_Y vertices and E_Y edges, respectively. Node similarity between G(V_X, E_X) and H(V_Y, E_Y) is calculated as (31) while edge similarity is calculated as (32).

$$J(V_X, V_Y) = \frac{\mid V_X \cap V_Y \mid}{\mid V_X \cup V_Y \mid} \tag{31}$$

$$J(E_X, E_Y) = \frac{\mid E_X \cap E_Y \mid}{\mid E_X \cup E_Y \mid} \tag{32}$$

3.5 Network Dismantling of TSPPIs

In this part, different attack strategies by using network centrality and GND with different cost functions are introduced.

3.5.1 Network Dismantling by Using Network Centrality

Network centrality predicts the impact of a node. The following centrality calculations are done iteratively for each method. In each iteration, we remove the node with highest centrality score.

3.5.1.1 Degree Centrality

The node *v* has a degree of the number of links of this node. To obtain degree centrality, the following formula is applied.

$$\mathbf{C}_D(i) = \frac{\mathbf{d(i)}}{\mathbf{n-1}} \tag{33}$$

where $C_{\mathrm{D}}(i)$ is the degree centrality, d(i) is the degree of node *i*, n-1 is the number of nodes in the network without node *i*.

3.5.1.2 Closeness Centrality

Closeness of the node v_{x} to node v_{y} is the shortest path between them. Closeness centrality of a node is calculated by taking the inverse of the shortest distances from

this node to each other node as the formula below.

$$\mathbf{C}_C(v_i) = \left(\sum_{\phi=1}^{\gamma} \delta(v_i, v_\phi) \right)^{-1} \tag{34}$$

where $C_{\mathrm{C}}(v_{\mathrm{i}})$ denotes the closeness centrality of v_{i} while $\delta(v_i, v_\phi)$ is the shortest path

3.5.1.3 Betweenness Centrality

Betweenness centrality of a node is a metric calculated by counting the number of shortest paths including the node.

$$\mathbf{C}_B(v_i) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{35}$$

where $C_{\mathrm{B}}(v)$ is the betweenness centrality of the node v, V is the set of nodes, $\sigma(s,t)$ is the number of shortest paths, and $\sigma(s,t|v)$ is the path count passing on the node v.

3.5.1.4 Eigenvector Centrality

Eigenvector centrality gives a transitive impact of a node in a graph, rather than only taking into account its direct impact. Both the number of edges and the importance of the nodes of these links are important while calculating this measure. Let A=($a_{\mathrm{v,t}}$) be the adjacency matrix of a graph G

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t \tag{36}$$

where x_v is the relative centrality score of vertex v.

3.5.2 Network Dismantling by GND

GND attempts to locate a set of nodes that removal of these nodes results in network fragmentation into subnetworks at minimal cost. Let G(V, E) be a network with nodes V and edges E. When a set S of nodes are removed from the network, if the giant connected component (GCC) of a network contains a maximum of C nodes, then S is the C-dismantling set. The weight-cost w_i represents the removal cost of node i where w_i is a non-negative value.

Let a set of nodes $M \subseteq V$ be disconnected from the nodes from the complementary set $\overline{M}$. The node value v_i of a vector v represents whether the node i is in set M or not. v_i is 1 if $i \in M$; otherwise, it is equal to 0. Therefore, if two nodes are from the same cluster, $v_i v_j$ equals to 1. In contrast, if one of them is from M and the other is from $\overline{M}$, $v_i v_j$ equals to -1.

The goal of the classical spectral bisection is to remove minimum number of edges between the clusters M and $\overline{M}$. GND introduces a node–weighted spectral cut objective function. The function declares that the cost of breaking a link between i and j is equal to the deletion of nodes i and j. If the link (i, j) attaches nodes from different clusters, the cost is $w_i + w_j - 1$. There is a constant term -1 in this cost formula to have a refined notation. Let B be a matrix with elements $B_{i,j}$ = $A_{i,j}$ $(w_i + w_j - 1)$ where the matrix A denotes the adjacency matrix of the network. Then, matrix B can be written as $B = AW + WA - A$ where W is a diagonal matrix with the weight w_{ii} of node i. In this study, we create the W matrix by calculating the w_i values separately for the degree, betweenness, closeness, and eigenvector centralities of the node i. These centrality scores are rescaled from the range (0,1) to (1,255). In addition, the node-weighted Laplacian of the matrix is represented by $L_w = D_B - B$ where D_B is a diagonal matrix with elements $(D_B)_{ii} = \sum_{j=1}^{n} B_{ij}$. According to Courant-Fischer theorem (Fiedler, 1973), the solution of this bisection problem is the second smallest eigenvector of the node–weighted Laplacian $\lambda_2 v^{(2)} = L_w v^{(2)}$. If a node i has a corresponding element in the second smallest eigenvector is non-negative $v_i^{(2)} \geq 0$ and its neighbor j with a negative element $v_i^{(2)} < 0$, their removal will result in two subnetworks M and $\overline{M}$. $0 =\leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ are real eigenvalues corresponding to eigenvectors $v_1, v_2, ..., v_n$ of the node-weighted Laplacian L_w.

$$\|B_{abs}v\|_1 \leq \|A\|_1 \|W\|_1 \|v\|_1 + \|W\|_1 \|A\|_1 \|v\|_1 + \|A\|_1 \|v\|_1 \leq 2d_{max}(w_{max}+1) \tag{37}$$

where B_{abs} denotes the matrix consists of absolute values of B, $\|X\|_1$ denotes L_1-norm of any matrix X, d_{max} is the maximal centrality of any node of the network, and w_{max} is the largest cost in diagonal cost matrix.

$$\begin{aligned}
&|\lambda_n| \leq max_{\|v\|_1=1} \|(D_B - B)v\|_1 \\
&= max_{\|v\|_1=1} \sum_{i=1}^{n} |v_i \sum_{j=1}^{n} B_{ij} - \sum_{j=1}^{n} v_j B_{ij}| \\
&\leq max_{\|v\|_1=1} \sum_{i=1}^{n} \sum_{j=1}^{n} |v_i B_{ij}| + \sum_{i=1}^{n} \sum_{j=1}^{n} |v_j B_{ij}| \\
&= max_{\|v\|_1=1} \|B_{abs}v\|_1 + \|B_{abs}v\|_1 \leq 4d_{max}(w_{max}+1)
\end{aligned} \tag{38}$$

Since $4d_{max}(w_{max}+1) = 4d_{max}^2 + 4d_{max} \leq 4d_{max}^2 + 2d_{max}^2$, GND method uses $\lambda_n \leq 6d_{max}^2$ when $w_{max} = d_{max}$ for the sake of simplicity.

To compute second smallest eigenvector $v^{(2)}$, the matrix $\overline{L} = 6d_{max}^2 I - L_w$ which has the same eigenvectors with L_w may be considered. According to the algorithm, we start with a random vector v uniformly drawn from the unit sphere S^n. After that, we force it to be perpendicular to the largest eigenvector $v_1 = c.(1,...,1)^T$ by setting $v = v - \frac{v_1^T v}{v_1^T v_1}.v_1$. Then, we have $v = \varphi_2 v_2 + ... + \varphi_n v_n$ and $\overline{L}v = \overline{\lambda_2}\varphi_2 v_2 + ... + \overline{\lambda_n}\varphi_n v_n$. By setting $v^{(k)} = \overline{L}^k v$, we get the following equation that converges with exponential speed to some eigenvector of L with eigenvalue λ_2.

$$\begin{aligned}
\frac{v^{(k)}}{\|v^{(k)}\|} &= \frac{\overline{\lambda_2}^k \varphi_2 v_2 + ... + \overline{\lambda_n}^k \varphi_n v_n}{\left\|\overline{\lambda_2}^k \varphi_2 v_2 + ... + \overline{\lambda_n}^k \varphi_n v_n\right\|} \\
&= \frac{\varphi_2 v_2 + \left(\frac{\overline{\lambda}_3}{\overline{\lambda}_2}\right)^k \varphi_3 v_3 + ... + \left(\frac{\overline{\lambda}_n}{\overline{\lambda}_2}\right)^k \varphi_n v_n}{\left\|\varphi_2 v_2 + \left(\frac{\overline{\lambda}_3}{\overline{\lambda}_2}\right)^k \varphi_3 v_3 + ... + \left(\frac{\overline{\lambda}_n}{\overline{\lambda}_2}\right)^k \varphi_n v_n\right\|}
\end{aligned} \tag{39}$$

The last step of this method is fine tuning of the spectral approximation. To find a set of nodes from the solution set which covers all of the edges, GND method adopted a linear-time approximation algorithm for the weighted vertex cover problem (Bar-Yehuda & Even, 1981) which is a greedy algorithm that reduces costs in each iteration, and constructs a cover as a result.

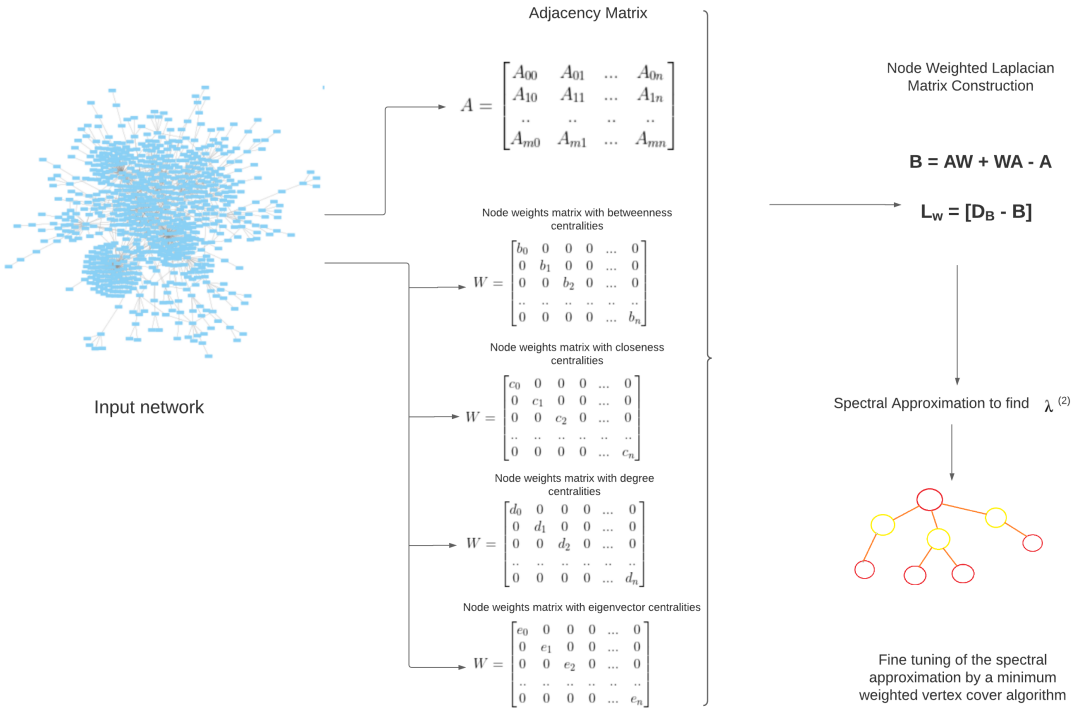


Figure 3.7: GND pipeline. First step is the construction of adjacency matrix and weight matrices. Then, the cost-weighted network B and node-weighted Laplacian matrix L_w are constructed. After that, second smallest eigenvector is found. Lastly, a weighted cover of the suggested nodes is constructed.

3.5.2.1 GND with Reinsertion (GNDR)

After application of GND method, GNDR proposes a reinsertion step which reduces costs from earlier stages to optimize network attacks. This step ensures that the nodes which do not affect the process from the initial GCC to the resulting GCC are removed from the list of the deleted nodes (Zdeborová et al., 2016). It comprises adding back the removed nodes to the subnetworks that have not reached the given number of nodes. If removed node is from a subnetwork with the maximum number of nodes, it is not added to this subnetwork.

There are two perspectives of algorithm. The first strategy follows the steps below.

1. Reverse the removed node list.
2. In each iteration,
   - I Select the next node for one TSPPI
   - II Connect the node to its neighbors left in the result of GND.
   - III If the size of the component to which it is added is smaller than the given threshold, delete the node from the removed node list.

The difference of the second strategy from the first strategy is that it starts from the beginning of the deleted nodes, not the end. Step by step explanation is below.

1. Get the removed node list.
2. In each iteration,
   - I Select the next node
   - II Connect the node to its neighbors left in the result of GND.
   - III If the size of the component to which it is added is smaller than the remaining GCC as a result of GND, delete the node from the removed node list.

After running the GND algorithm until the number of components are greater than or equal to the 20% of the initial GCC size, we implemented the first strategy to resulting subnetworks from GND to start reinserting with the nodes with higher centrality. In this study, the threshold value is the size of the resulting GCC obtained from GND.

3.6 Method Selection

To select the most suitable method for driver subnetwork identification, we need to select a method that results in a network which has features that satisfy the following criteria;

- Enough driver nodes in large connected components to be able to initiate random walks
- The GCC with enough nodes to continue random walks
- Sufficiently fragmented subnetworks

After all methods are applied to each TSPPI and the results are obtained, we create 2 metrics to understand which method is more suitable for which TSPPI. Let GCC_{res} be number of nodes in the resulting GCC, GCC_{first} be number of nodes in the initial GCC, D_{GCC} be the number of driver nodes in the GCC, D_{res} be the number of driver nodes in the resulting network, D_{total} be the number of driver nodes in the GCC of the initial network, and N be the number of all nodes in the initial TSPPI network.

$$M_1 = \frac{\frac{D_{\text{GCC}}}{D_{\text{total}}}}{\frac{GCC_{\text{res}}}{GCC_{\text{first}}}} \tag{310}$$

$$M_2 = \frac{\frac{D_{\text{GCC}}}{D_{\text{res}}}}{\frac{GCC_{\text{res}}}{N}} \tag{311}$$

3.7 Personalized PageRank

PageRank is an algorithm for measuring the relative importance of a network by assigning weights to its nodes (Page, Brin, Motwani, & Winograd, 1999). This algorithm is based on random walks on graphs. Stationary distribution consisting of random steps with a probability of p gives PageRank weights associated with each node. Personalized PageRank is a specialized version of PageRank (Jeh & Widom, 2003). The only difference is that each random step is made to the same source node. In this way, the algorithm is biased towards the given set of nodes.

3.8 Functional Enrichment Analysis

The Database for Annotation, Visualization, and Integrated Discovery (DAVID) is a way to analyze functional enrichments of a given list of nodes (Dennis et al., 2003). Since there are many components to analyze, we utilized the DAVID API services to automate this process.

Gene Ontology (GO) terms creates a controlled vocabulary for biomolecular features for functional annotation analysis. GO terms are categorized into three parts: biological process, cellular component and molecular function (Ashburner et al., 2000).

Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000) and Reactome (Joshi-Tope et al., 2005) provide pathway information of signaling, genetic and metabolic processes.

In order to obtain significant functional enrichments in each subnetwork, we used DAVID API to display the most enriched GO biological processes, and pathways in KEGG and Reactome.

CHAPTER 4

RESULTS

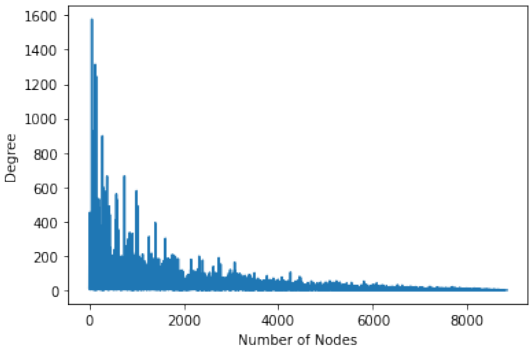
4.1 Dataset

In this study, we used breast, liver, ovary, peripheral nerve, and lymph node TSPPIs from TissueNet v.2. Initially, there were nodes with self-edges in these networks. The number of nodes and edges, before/after self edges were removed, are given in the Table 4.1.

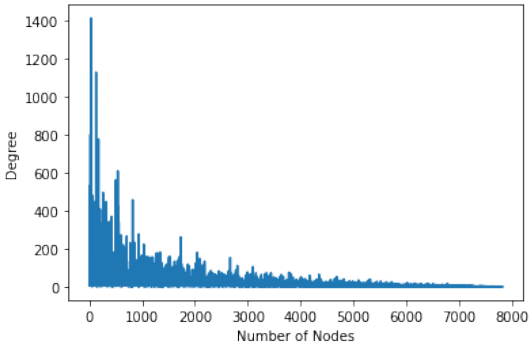
Table 4.1: Summary of the TSPPI Dataset

Tissue	Number of Nodes with Self-Edges	Number of Edges with Self-Edges	Number of Nodes without Self-Edges	Number of Edges without Self-Edges
Breast	8855	104719	8838	103231
Liver	7847	78428	7823	77117
Lymph Node	7677	85759	7645	84367
Ovary	6389	65935	6369	64775
Peripheral Nerve	5026	43002	5006	42075

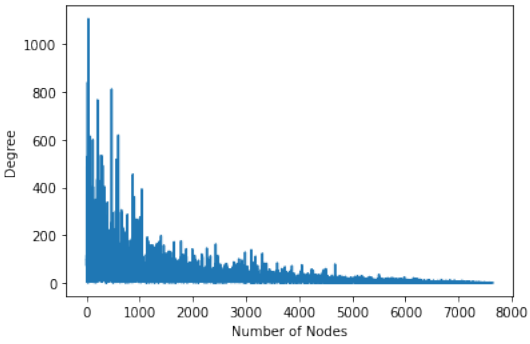
After deletion of self-edges, the degree distributions of these networks are given in Figure 4.1



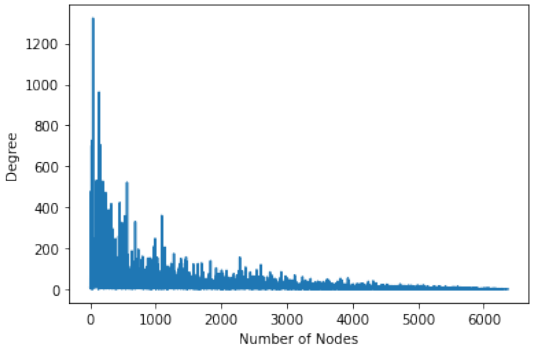
(a) Degree Distribution of Breast TSPPI



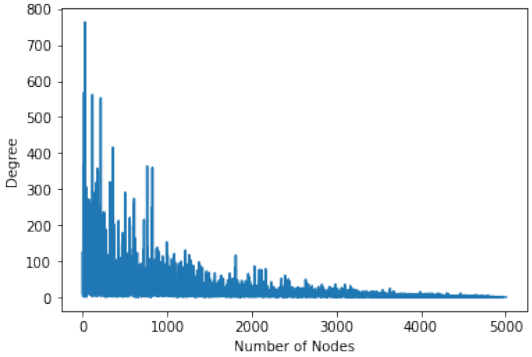
(b) Degree Distribution of Liver TSPPI



(c) Degree Distribution of Lymph Node TSPPI



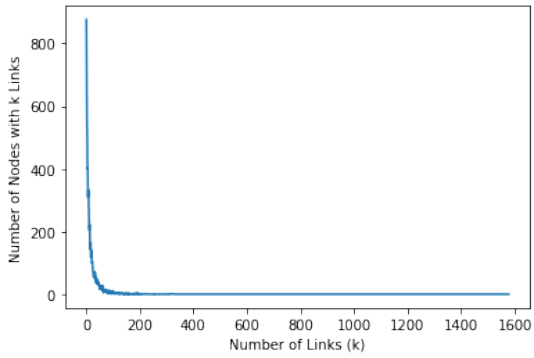
(d) Degree Distribution of Ovary TSPPI



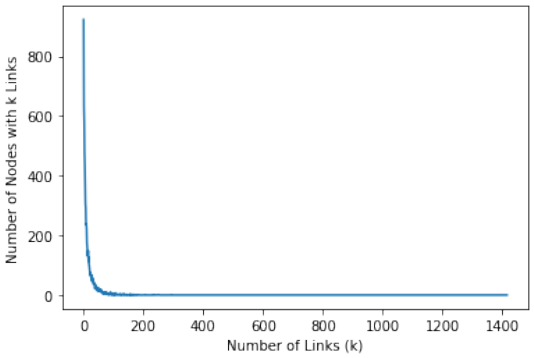
(e) Degree Distribution of Peripheral Nerve TSPPI

Figure 4.1: Degree distributions of the TSPPI networks

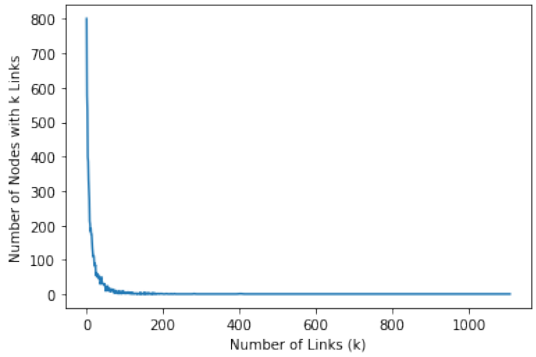
As can be seen from the Figure 4.1, while there are small numbers of highly connected nodes, the remaining nodes have very small degrees. These distributions lead us to power-law distribution ($P_{deg}(k) \propto k^{-\gamma}$) which is a common property of the scale-free networks (Figure 4.2).



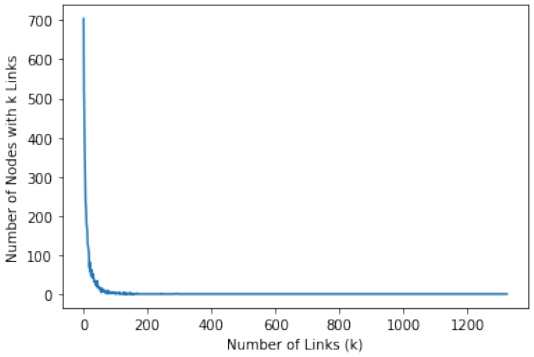
(a) Power-Law Distribution of Breast TSPPI



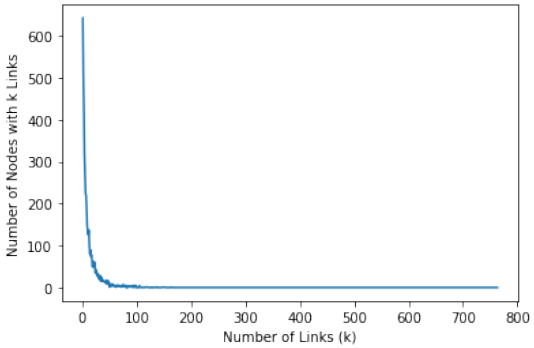
(b) Power-Law Distribution of Liver TSPPI



(c) Power-Law Distribution of Lymph Node TSPPI



(d) Power-Law Distribution of Ovary TSPPI



(e) Power-Law Distribution of Peripheral Nerve TSPPI

Figure 4.2: Power-Law Distribution of the TSPPI networks

We obtained driver node information from CGI. Firstly, we extracted "Gene"-"Primary Tumor type full name" pairs to find tissue-specific driver genes. Then, we converted given HGNC symbols to Ensembl Gene IDs. The detailed driver gene information is shown in Table 4.2

Table 4.2: Driver Gene Information from CGI Biomarkers Database

Tissue	Related Cancer	Number of Genes Associated only with the Related Cancer	Number of Genes Associated with Any Cancer Type	Number of Total Driver Genes
Breast	Breast adenocarcinoma	20	55	75
Liver	Hepatic carcinoma	2	55	57
Lymph Node	Lymphoma	6	55	61
Ovary	Ovary	7	55	62
Peripheral Nerve	Malignant peripheral nerve sheat tumor	1	55	56

After the retrieval of the information in Table 4.2, we searched them in the corresponding TSPPIs.

Table 4.3: Driver Gene Counts in TSPPIs

Tissue	Number of Driver Genes in the Largest Component	Number of Driver Genes in the TSPPI
Breast	56	56
Liver	37	37
Lymph Node	43	43
Ovary	35	35
Peripheral Nerve	19	19

We used the cBioPortal web interface to get the frequency of the driver node list retrieved from CGI in different cancer samples. Then, we filtered driver genes in Table 4.4 via their occurrences in the dataset obtained from cBioPortal. Any node with 0 occurrence removed from driver gene list. Therefore, only genes related to the certain types of cancer remained in the list. Query result of breast TSPPI from cBioPortal is shown in 4.3. Results of liver, lymph node, ovary, and peripheral nerve is in Appendix B.

Table 4.4: Final Driver Gene Counts in TSPPIs

Tissue	Number of Driver Genes in the Largest Component	Number of Driver Genes in the TSPPI
Breast	56	56
Liver	37	37
Lymph Node	34	34
Ovary	23	23
Peripheral Nerve	5	5

Figure 4.3: Breast TSPPI Driver Gene Search Result in cBioPortal

4.1.1 Centrality Rankings of Driver Nodes

We expect driver nodes to be at the top of the centrality rankings with the highest centrality scores. Although few driver nodes may have lower centrality scores, it is also necessary to consider that these nodes may be in critical positions between the hubs of the connected components. The relationships between centrality rankings of breast TSPPI are given in 4.4. The other pair-plots of previously given tissues, namely liver, lymph node, ovary, and peripheral nerve, are given in Appendix A. In addition, the relationship between node rankings and mutation frequencies is given in Figure 4.5.

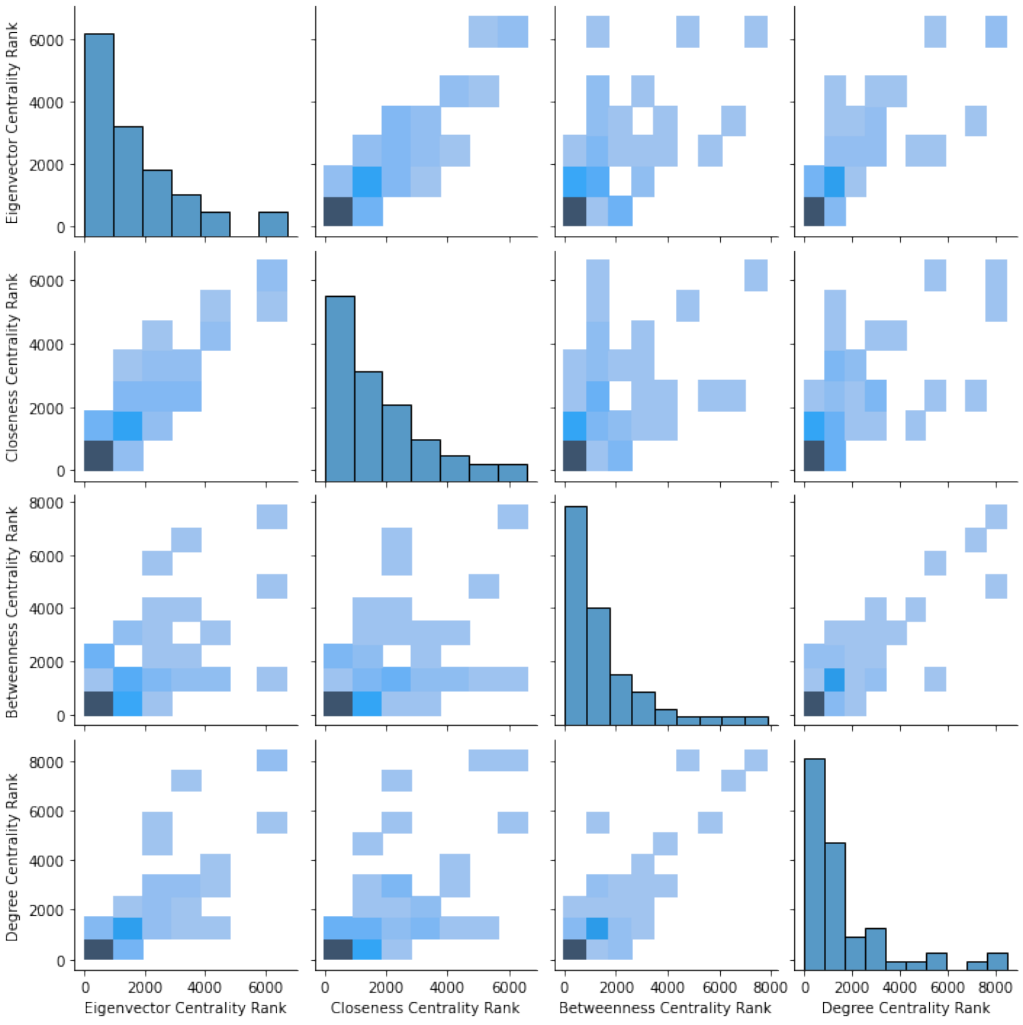


Figure 4.4: Pair-plots of Breast TSPPI.Darker colors indicate higher number of nodes within the corresponding ranks. Most driver nodes have higher centrality scores and higher rankings according to different centrality criteria. Diagonal plots represent the marginal distribution of the data for each ranking measure.

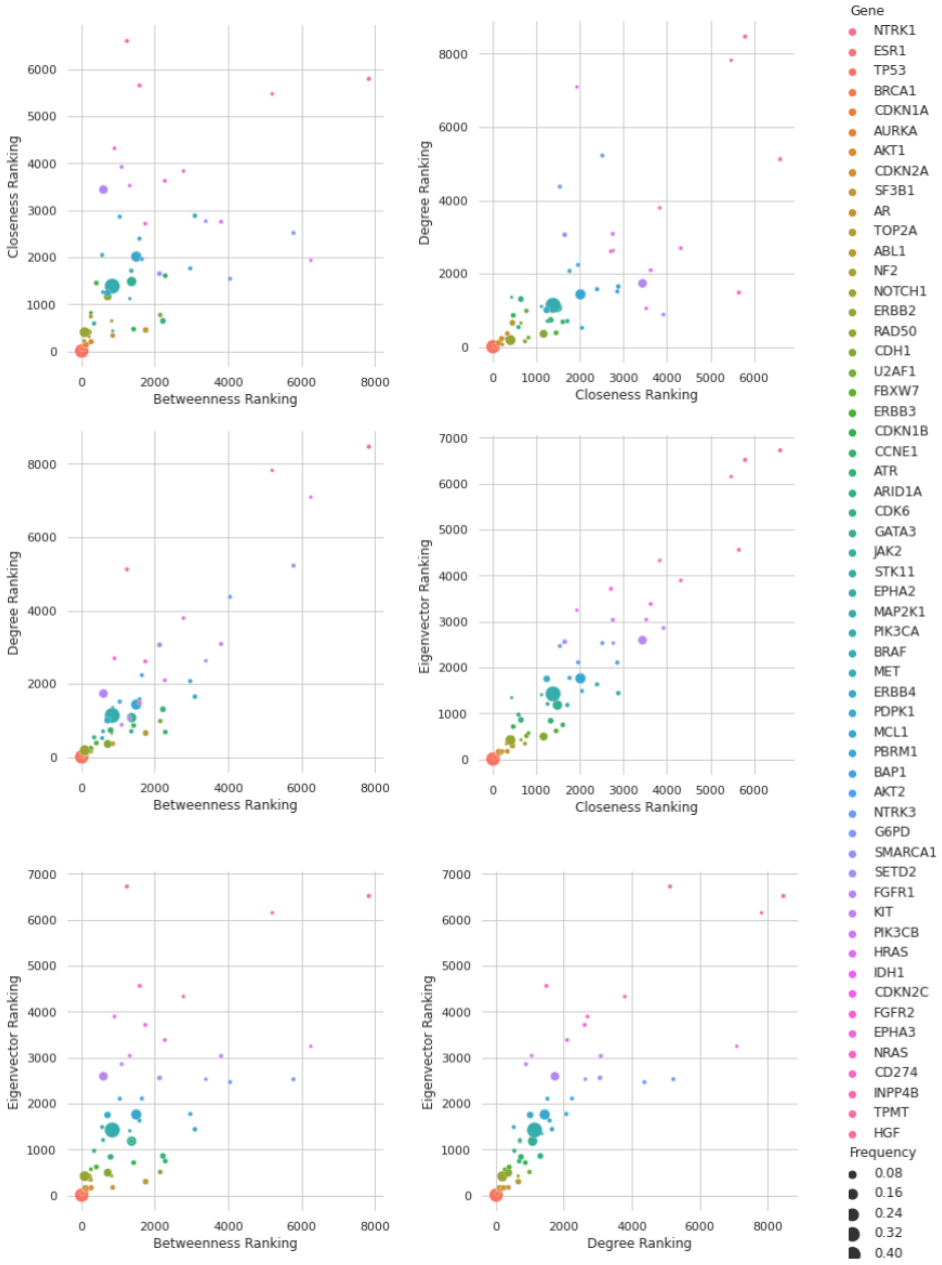
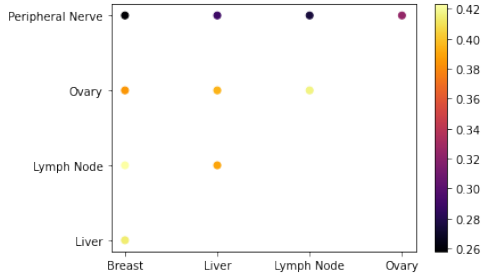


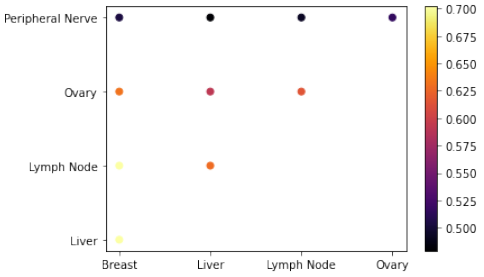
Figure 4.5: Node rankings of breast TSPPI with mutation frequency information in patients obtained from cBioPortal. Different colors indicate different genes and point sizes express the magnitude of frequency. Plots show that rankings are correlated with frequency of occurence in samples.

4.1.2 Similarity between Different TSPPIs

We used Jaccard Similarity index to show how much each TSPPI differs from tissue to tissue.



(a) Jaccard Edge Similarity Indices Between TSPPIs. The most similar network pair is Breast-Lymph Node with a Jaccard Similarity Index of 0.423.



(b) Jaccard Node Similarity Indices Between TSPPIs. The most similar network pairs are Breast-Lymph Node and Breast-Liver tissue pairs with a Jaccard Similarity Index of 0.702.

Figure 4.6: Network Similarities with respect to Jaccard Indices. Light colors indicate closer networks.

Table 4.5: Jaccard Node and Edge Similarity Indices between Each Tissue Pair

Tissue	Tissue	Jaccard Edge Similarity	Jaccard Node Similarity
Breast	Lymph Node	0.423	0.702
Lymph Node	Ovary	0.417	0.615
Breast	Liver	0.414	0.702
Liver	Ovary	0.395	0.592
Liver	Lymph Node	0.390	0.629
Breast	Ovary	0.385	0.633
Ovary	Peripheral Nerve	0.324	0.516
Liver	Peripheral Nerve	0.289	0.479
Lymph Node	Peripheral Nerve	0.275	0.493
Breast	Peripheral Nerve	0.258	0.504

As shown in the Table 4.5 and 4.6, Jaccard Similarity Index results of nodes are higher than edges. Although there are many commonalities in the gene lists of tissues, the interactions between these genes are distinctive on a tissue basis.

4.2 Network Dismantling Results

4.2.1 Network Dismantling by Different Centrality Metrics

After deletion of the node with the highest centrality value in each iteration, initial network is divided into several subcomponents. This process is shown in Figure 4.7, Figure 4.8, Figure 4.9, Figure 4.10, Figure 4.11

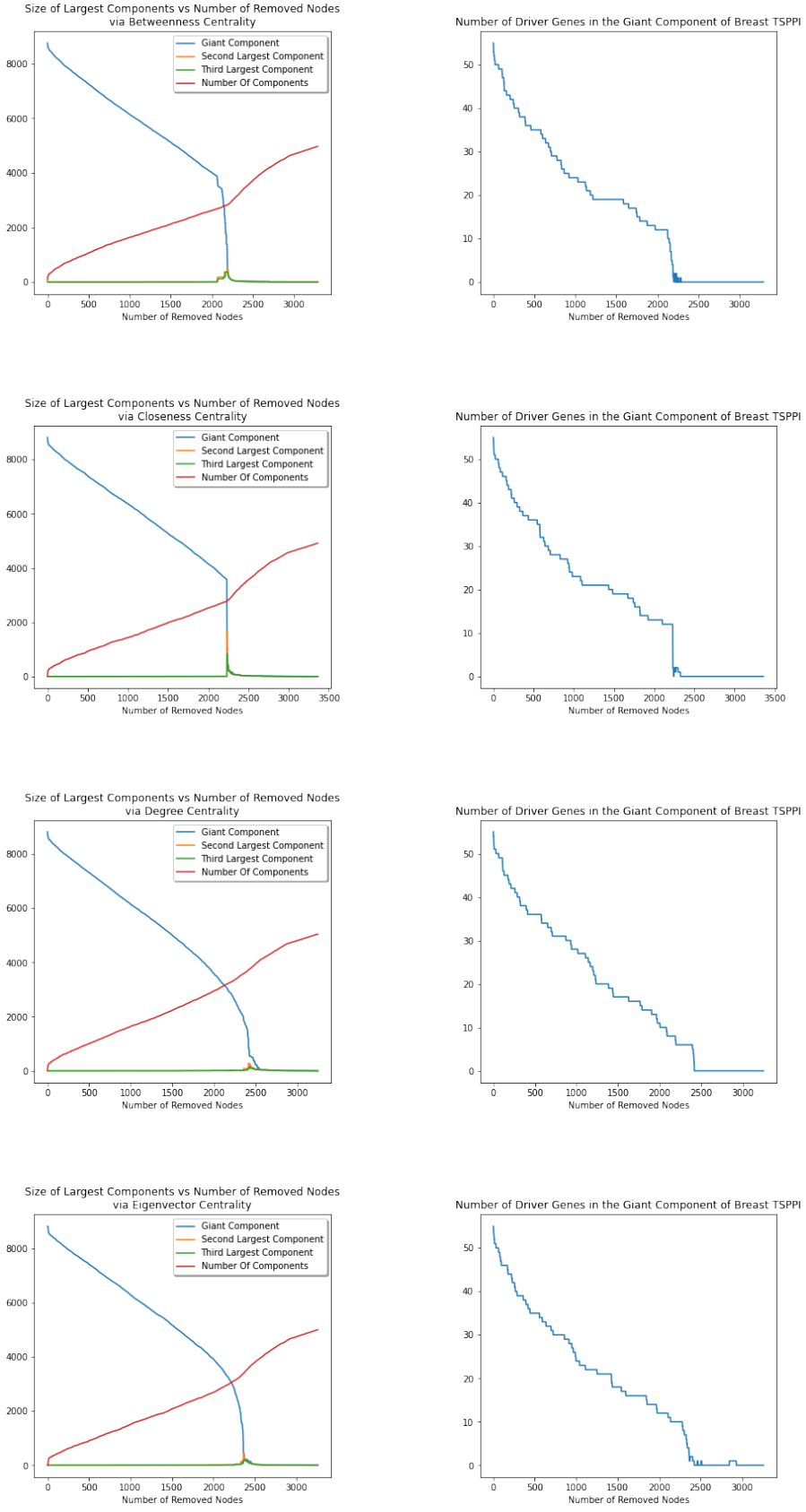


Figure 4.7: Line plots of the number of components, component sizes, and driver gene counts in breast TSPPI.

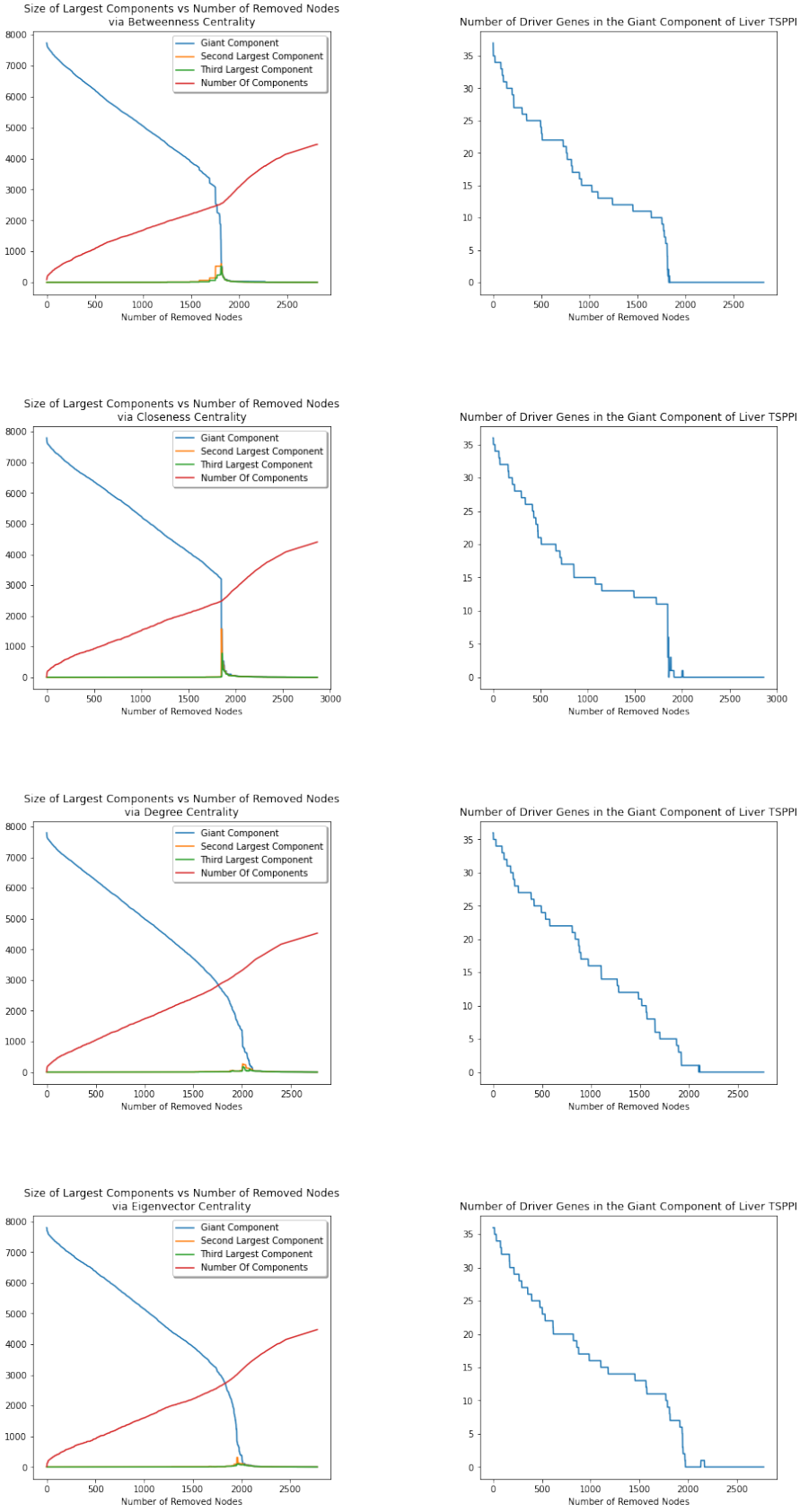


Figure 4.8: Line plots of the number of components, component sizes, and driver gene counts in liver TSPPI.

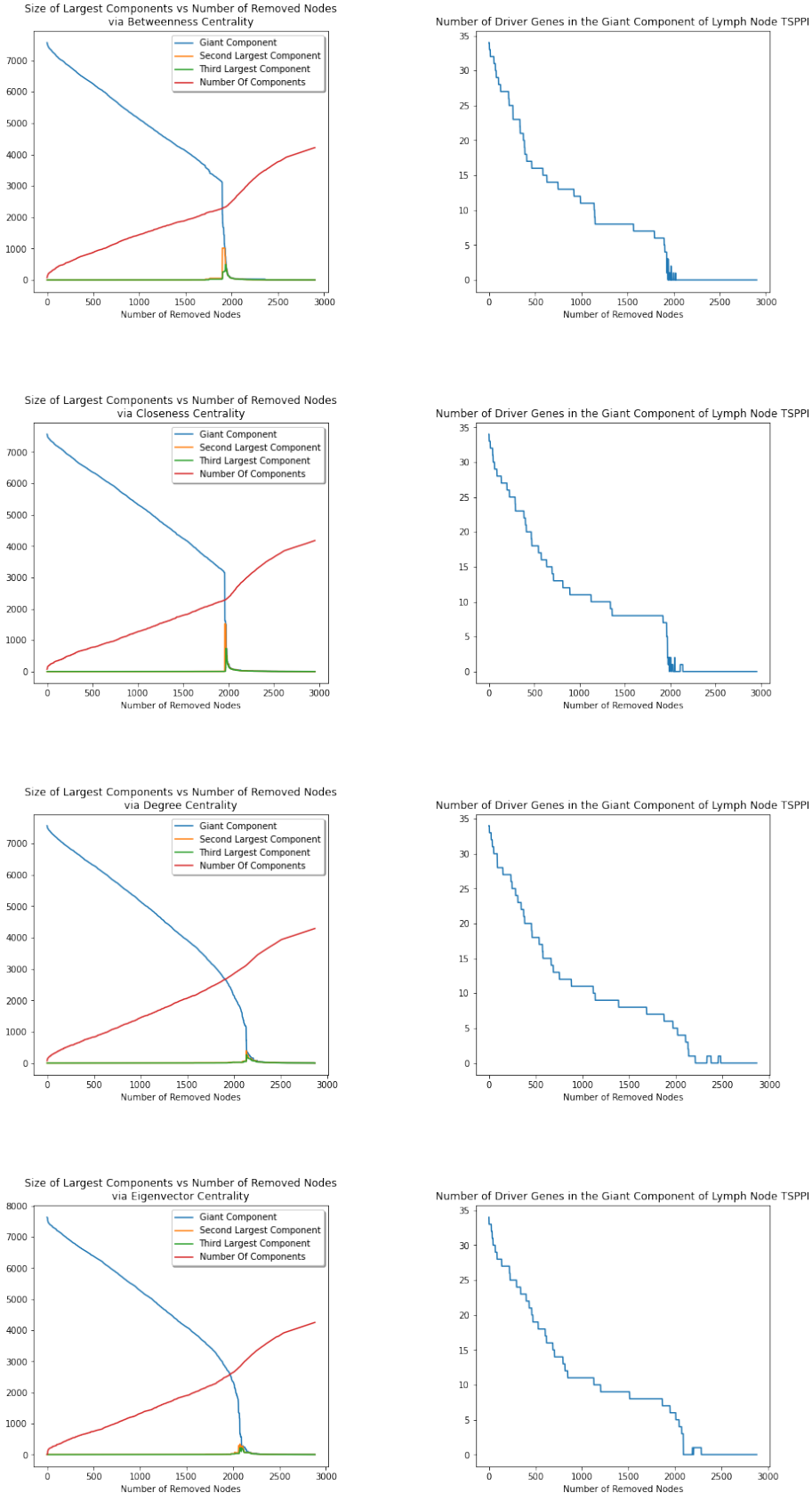


Figure 4.9: Line plots of the number of components, component sizes, and driver gene counts in lymph node TSPPI.

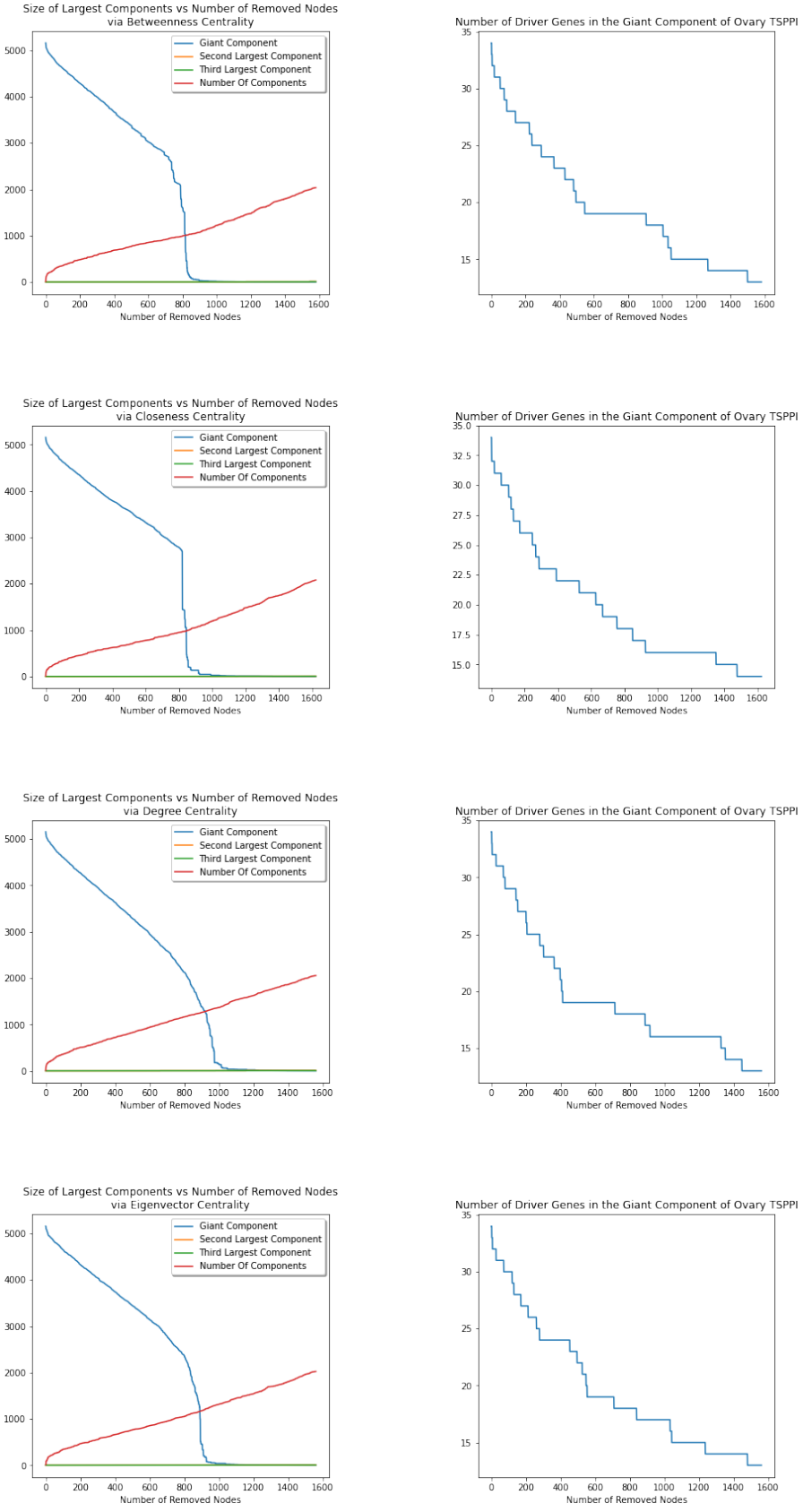


Figure 4.10: Line plots of the number of components, component sizes, and driver gene counts in ovary TSPPI.

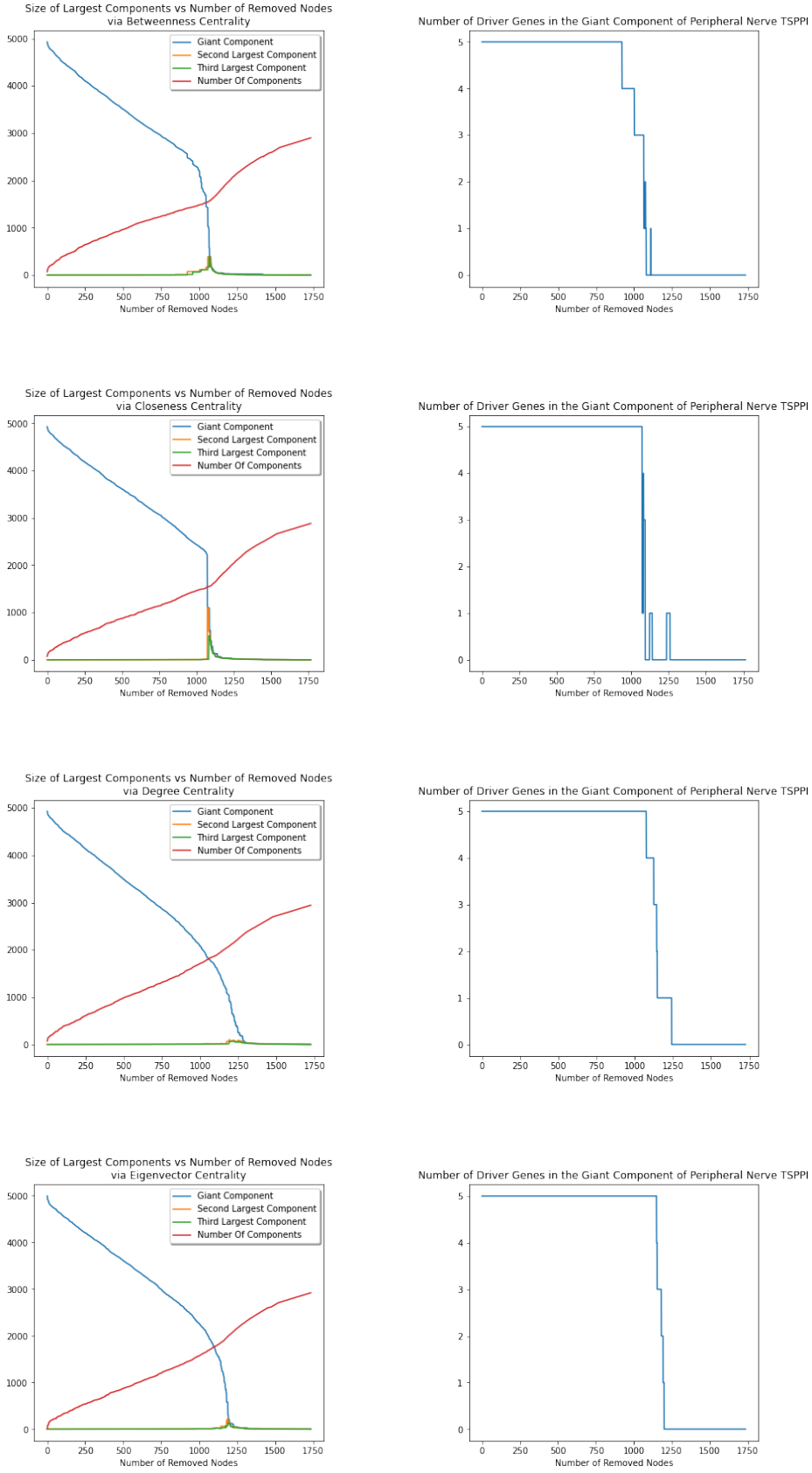


Figure 4.11: Line plots of the number of components, component sizes, and driver gene counts in peripheral nerve TSPPI.

In these scale-free networks, as the nodes are deleted at the beginning, the largest component size and number of components progress linearly, while after a certain point they decrease exponentially. It indicates that when the nodes in the hubs of the network are deleted, they begin to resemble the random network structure. In other words, when enough critical nodes in a biological network are deleted, we may see that the structure of this network is completely decomposed.

Although network centrality gives good results in dismantling the network, it takes a lot of time to calculate the centralities of all nodes in each iteration to split the network. Let k be number of iterations, n denote the number of nodes and m denote edges in a network. Computational complexities of this iterative algorithm with these metrics are as follows:

- Betweenness Centrality: $O(k(nm+n^2logn))$ time complexity (Brandes, 2001)
- Closeness Centrality: $O(knm)$ time complexity
- Degree Centrality: $O(kn^2)$ time complexity
- Eigenvector Centrality: $O(kn^3)$ time complexity

4.2.2 Network Dismantling by GND with Different Weights

Another approach to dismantle a network is GND. In this study, we modified the original GND algorithm by different weight functions by using different network centrality metrics. GND not only reduces computational complexity, but also gives results close to iterative algorithms where nodes are deleted just by looking at the centrality results in each iteration.

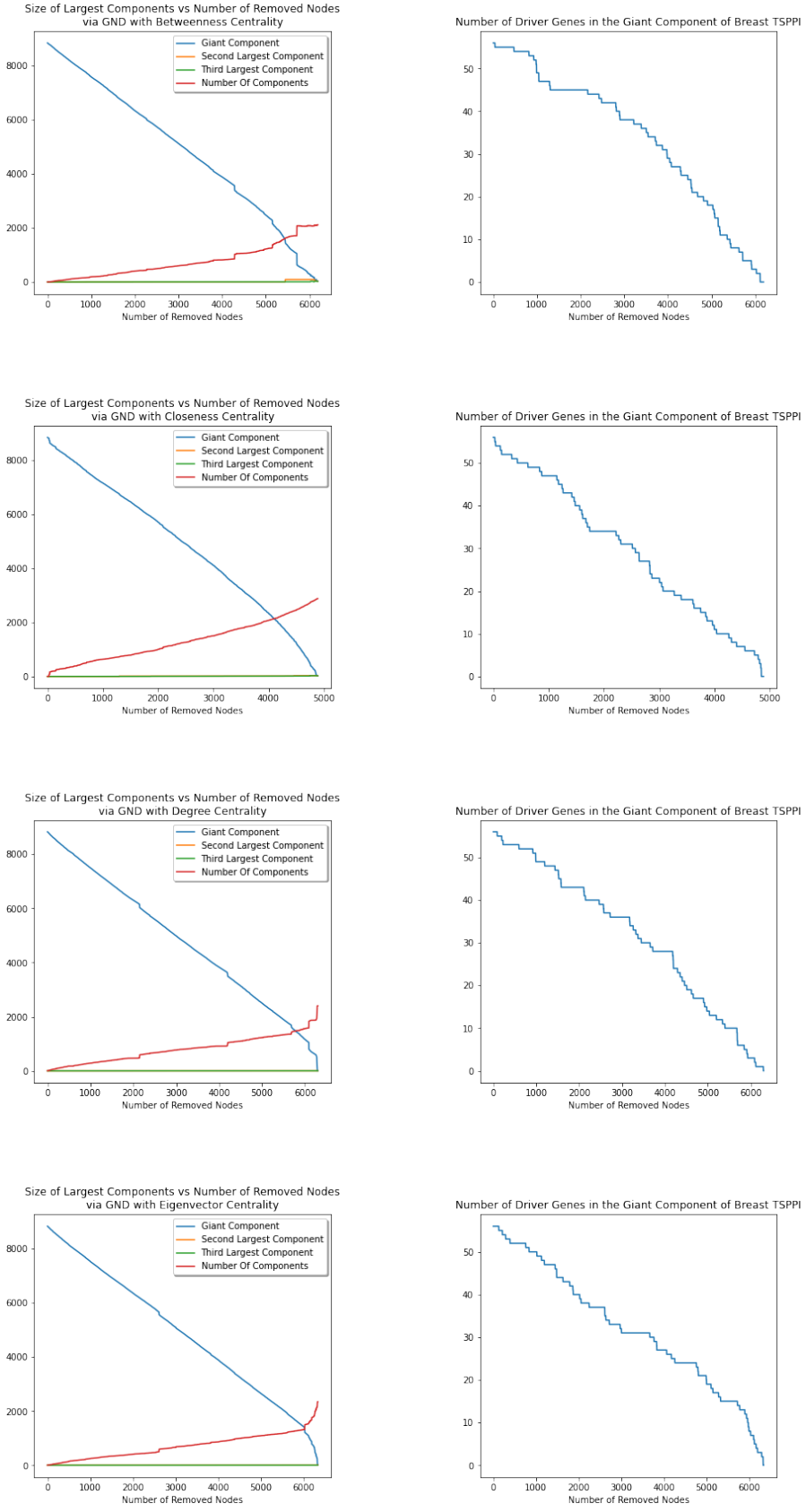


Figure 4.12: Line plots of the number of components, component sizes, and driver gene counts in breast TSPPI.

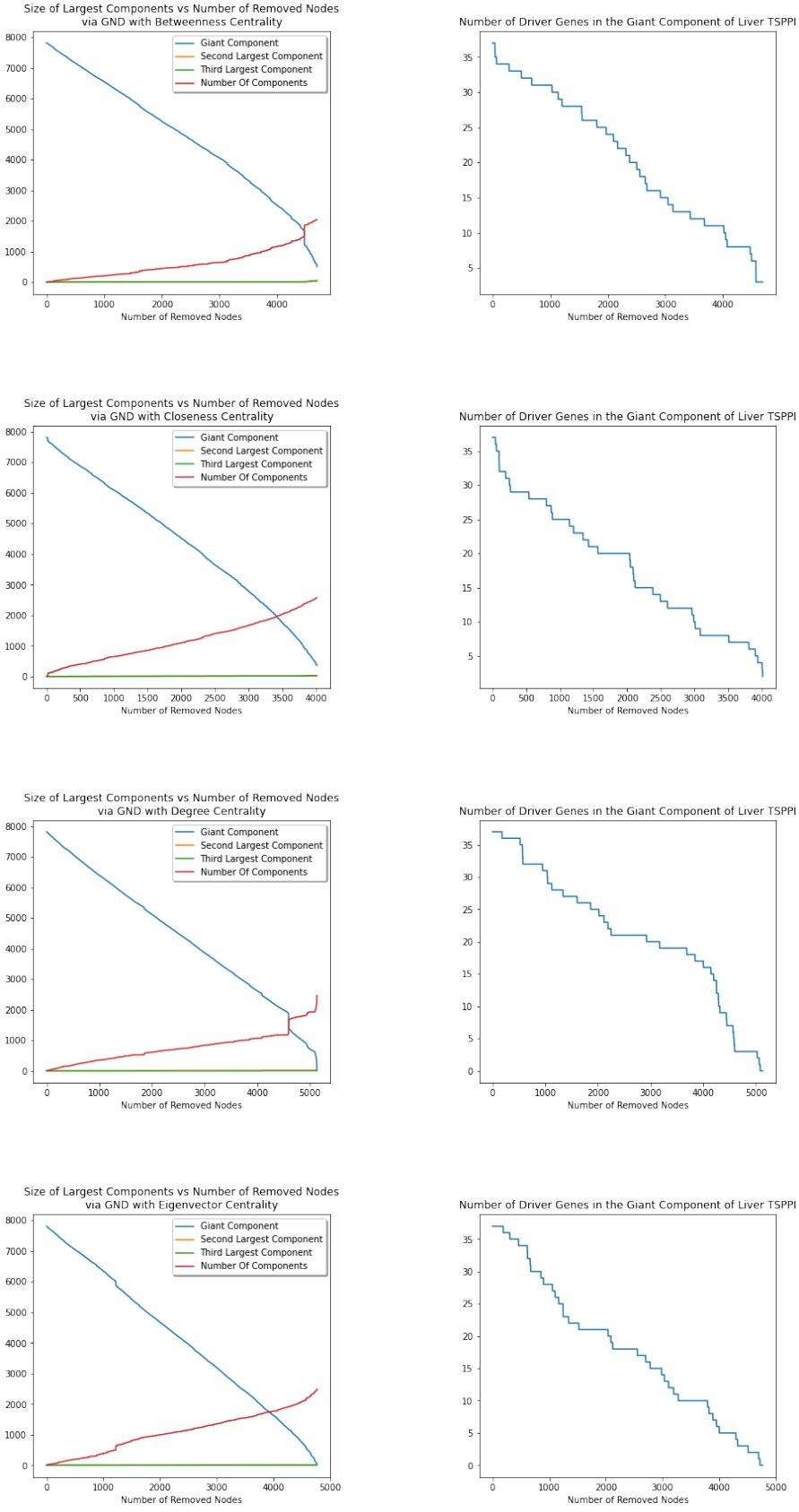


Figure 4.13: Line plots of the number of components, component sizes, and driver gene counts in liver TSPPI.

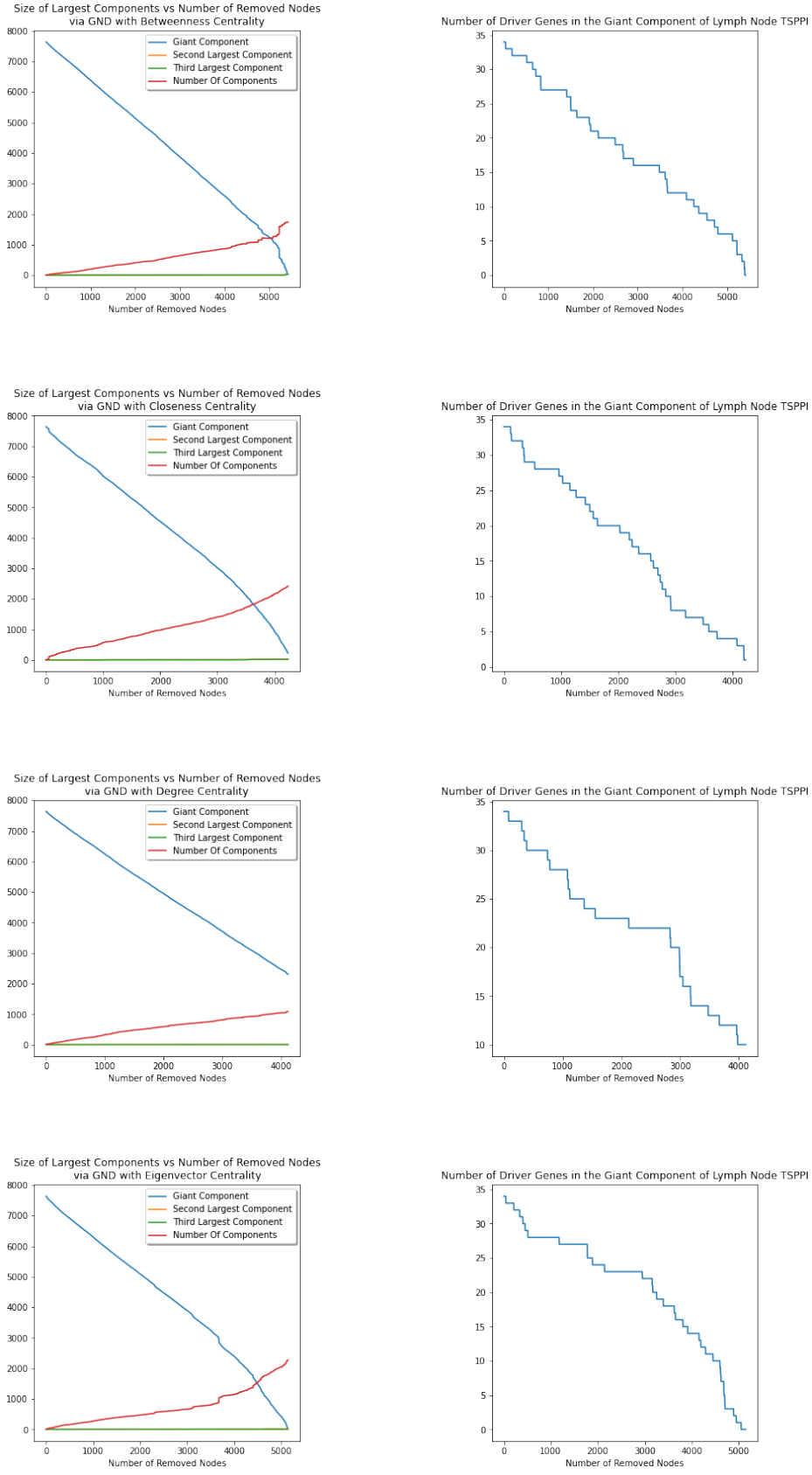


Figure 4.14: Line plots of the number of components, component sizes, and driver gene counts in lymph node TSPPI.

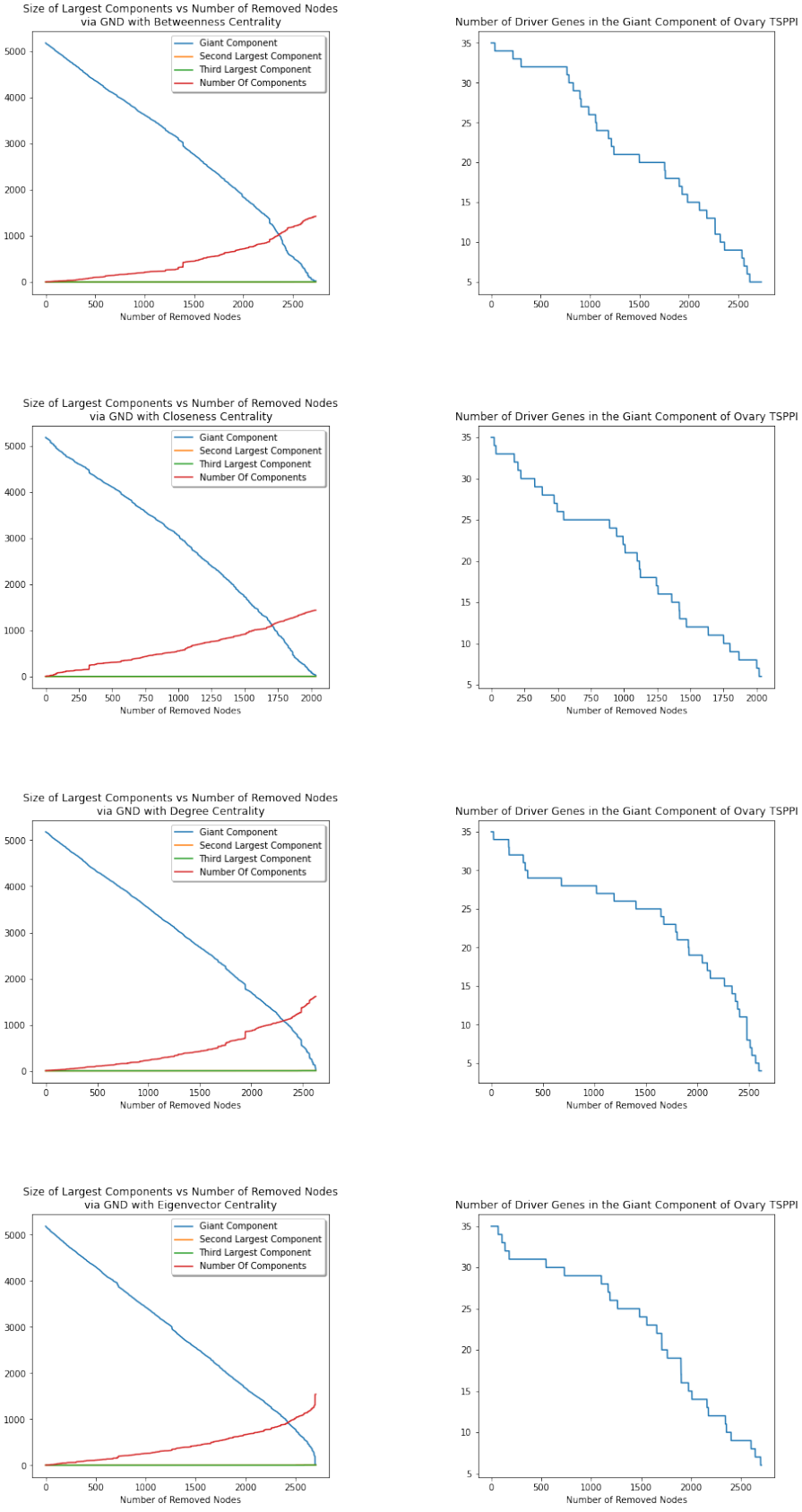


Figure 4.15: Line plots of the number of components, component sizes, and driver gene counts in ovary TSPPI.

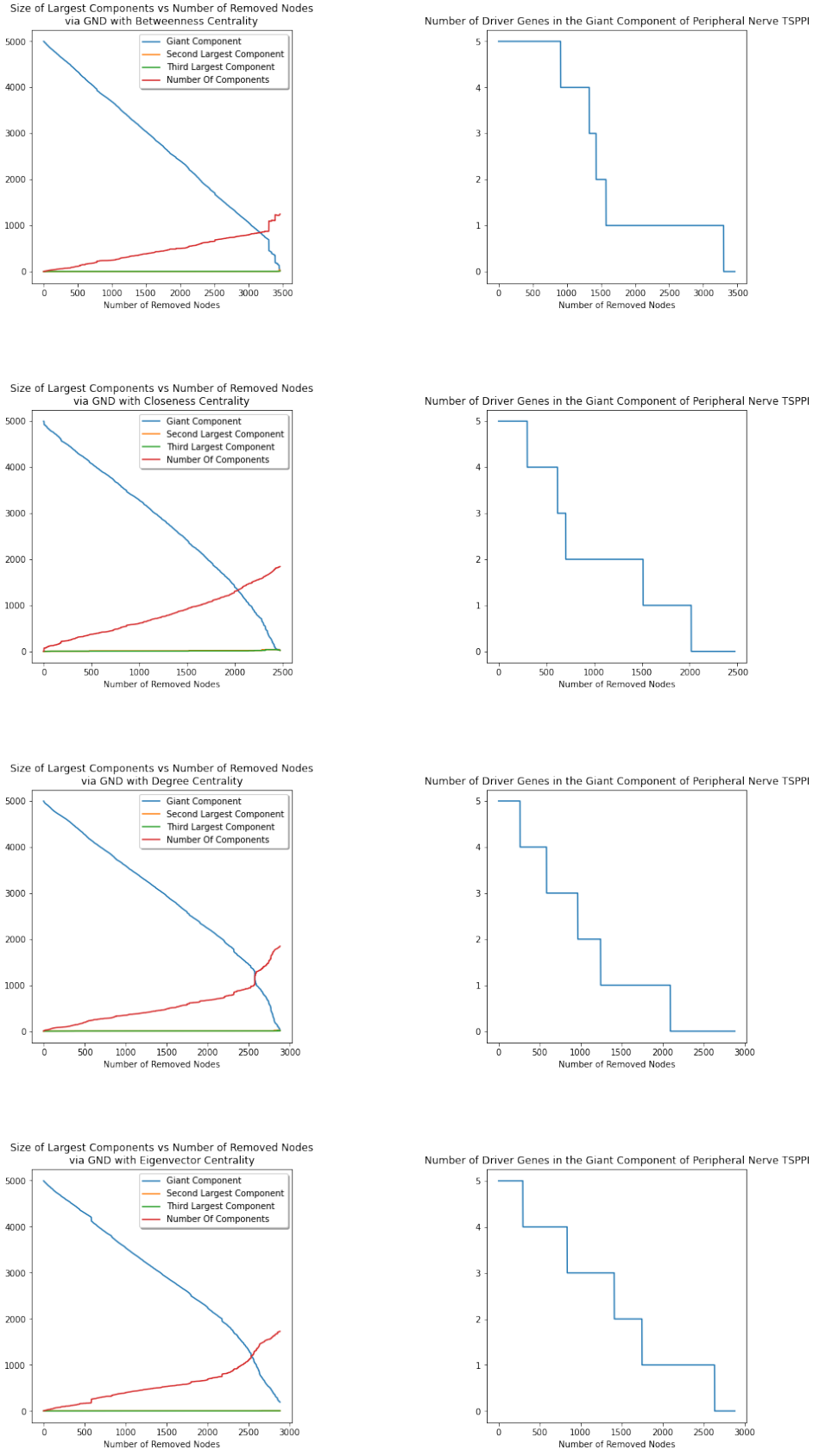


Figure 4.16: Line plots of the number of components, component sizes, and driver gene counts in peripheral nerve TSPPI.

The biggest contribution of the GND algorithm is to reduce computational complexity by using spectral approximation. Let k be number of iterations of iterative centrality algorithms, n be the number of nodes and m be the number of edges in a network. Firstly, GND constructs a Laplacian matrix with respect to centrality weights. Then, the computational complexity cost comes from the centrality measure. After this calculation, multiple nodes are deleted in each iteration by using spectral approximation (Pothen, Simon, & Liou, 1990) and minimum weighted vertex cover (Feige, Hajiaghayi, & Lee, 2008) algorithms. In this case, the k value is lower than the previously explained iterative network dismantling method. GND reduces the computation time. Spectral approximation algorithm has a computational complexity of $O(nlog^{2+\epsilon}(n))$ while minimum weighted vertex cover algorithm has a linear time complexity. We may summarize these costs as follows:

- GND with Betweenness Centrality: $O(log(k)(nm + n^2logn))$ time complexity
- GND with Closeness Centrality: $O(log(k)nm)$ time complexity
- GND with Degree Centrality: $O(log(k)n^2)$ time complexity
- GND with Eigenvector Centrality: $O(log(k)n^3)$ time complexity

4.2.3 Network Dismantling by GNDR with Different Weights

After the application of GND algorithm to all TSPPIs of interest, reinsertion is employed to reduce the removed nodes to have larger subnetworks. This process lasts until all removed nodes are tried to be reinserted.

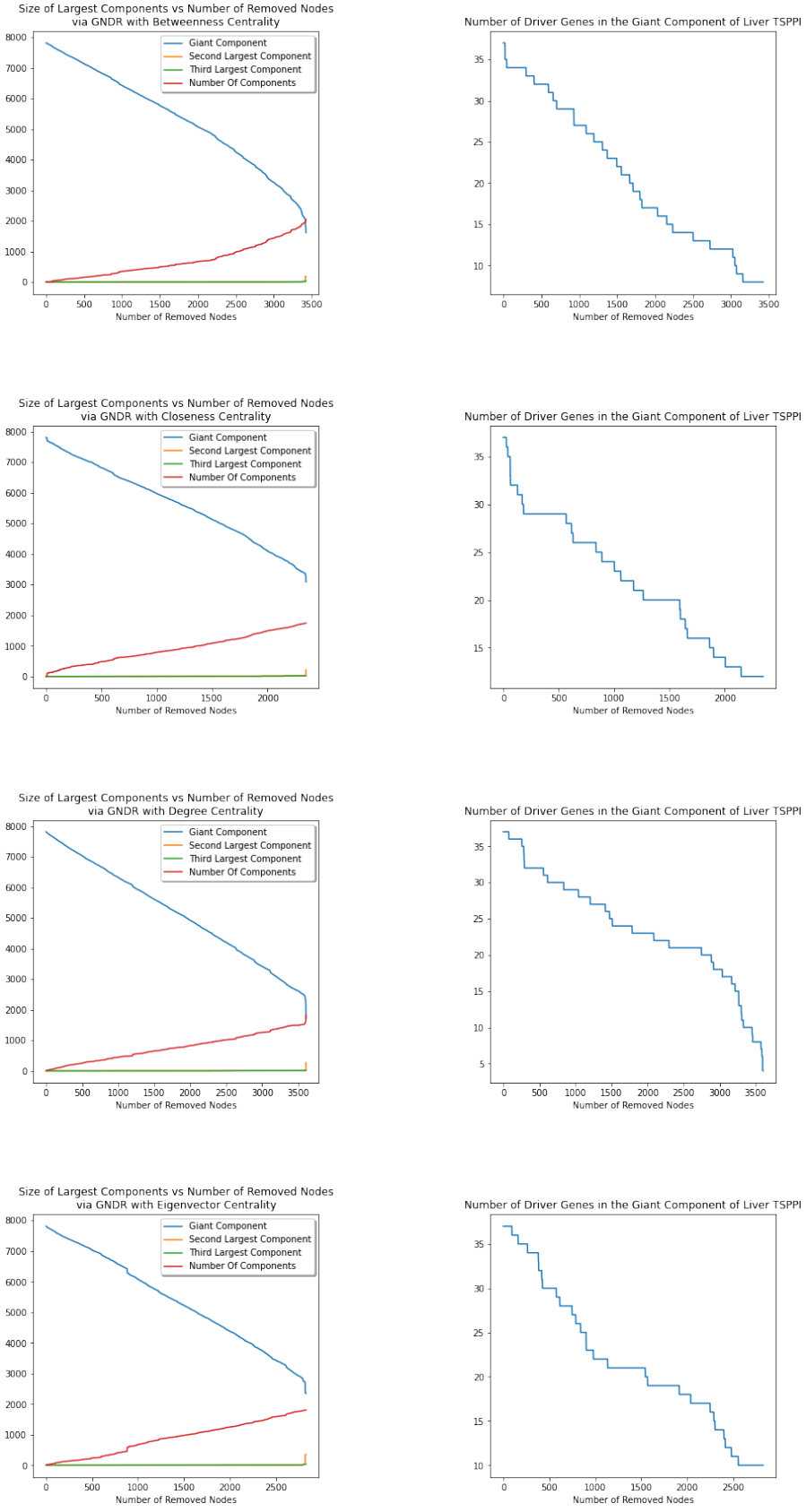


Figure 4.17: Line plots of the number of components, component sizes, and driver gene counts in breast TSPPI.

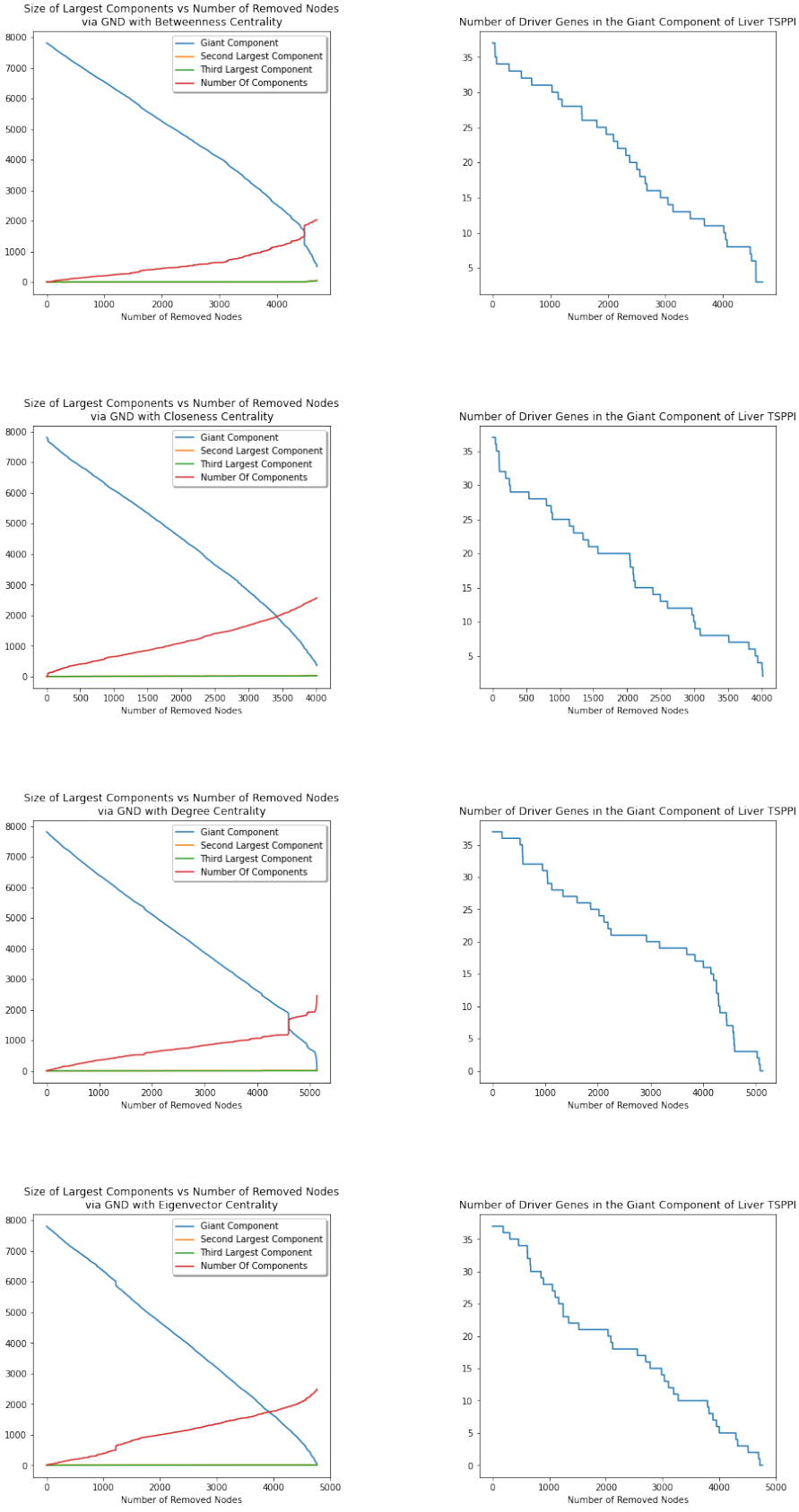


Figure 4.18: Line plots of the number of components, component sizes, and driver gene counts in liver TSPPI.

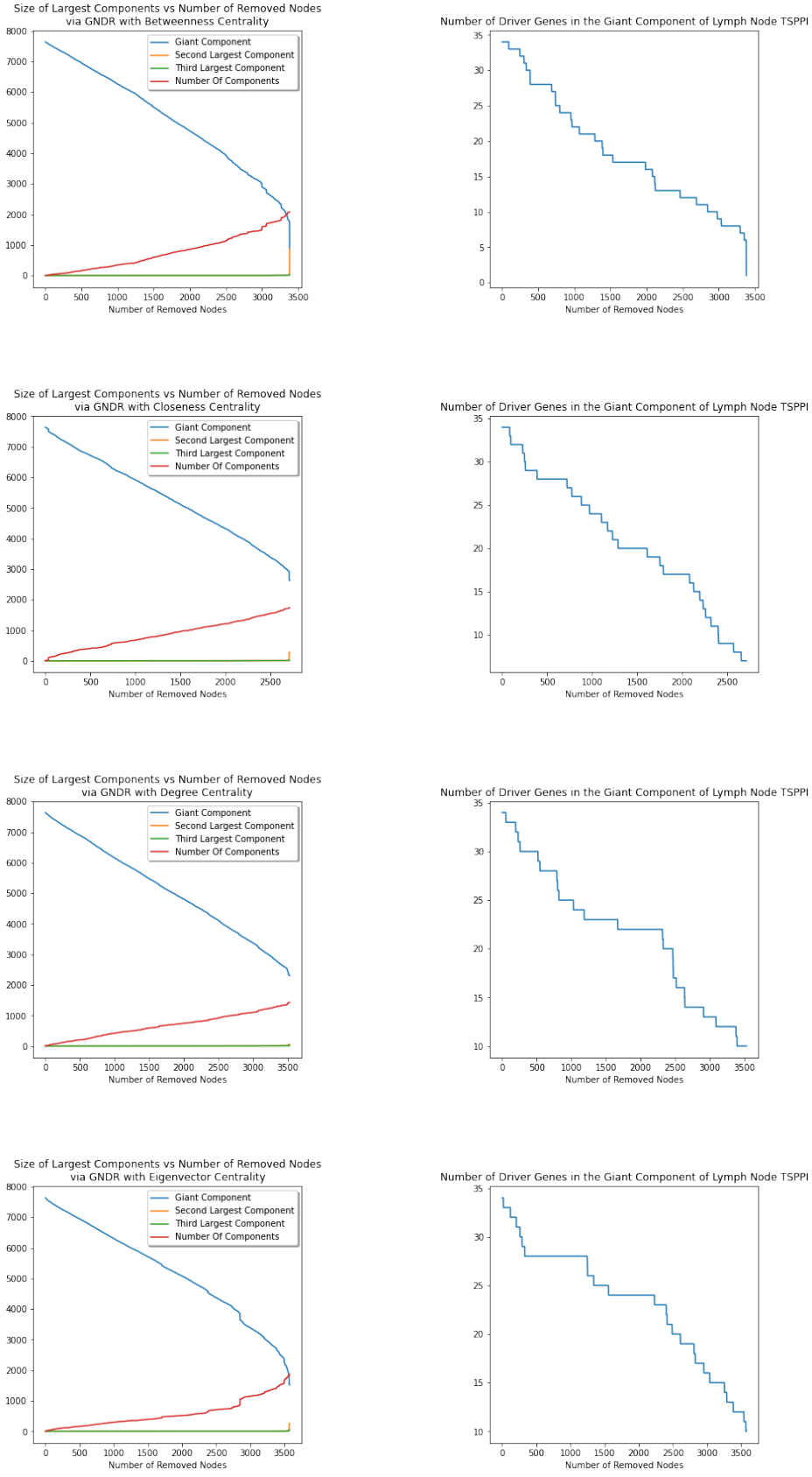


Figure 4.19: Line plots of the number of components, component sizes, and driver gene counts in lymph node TSPPI.

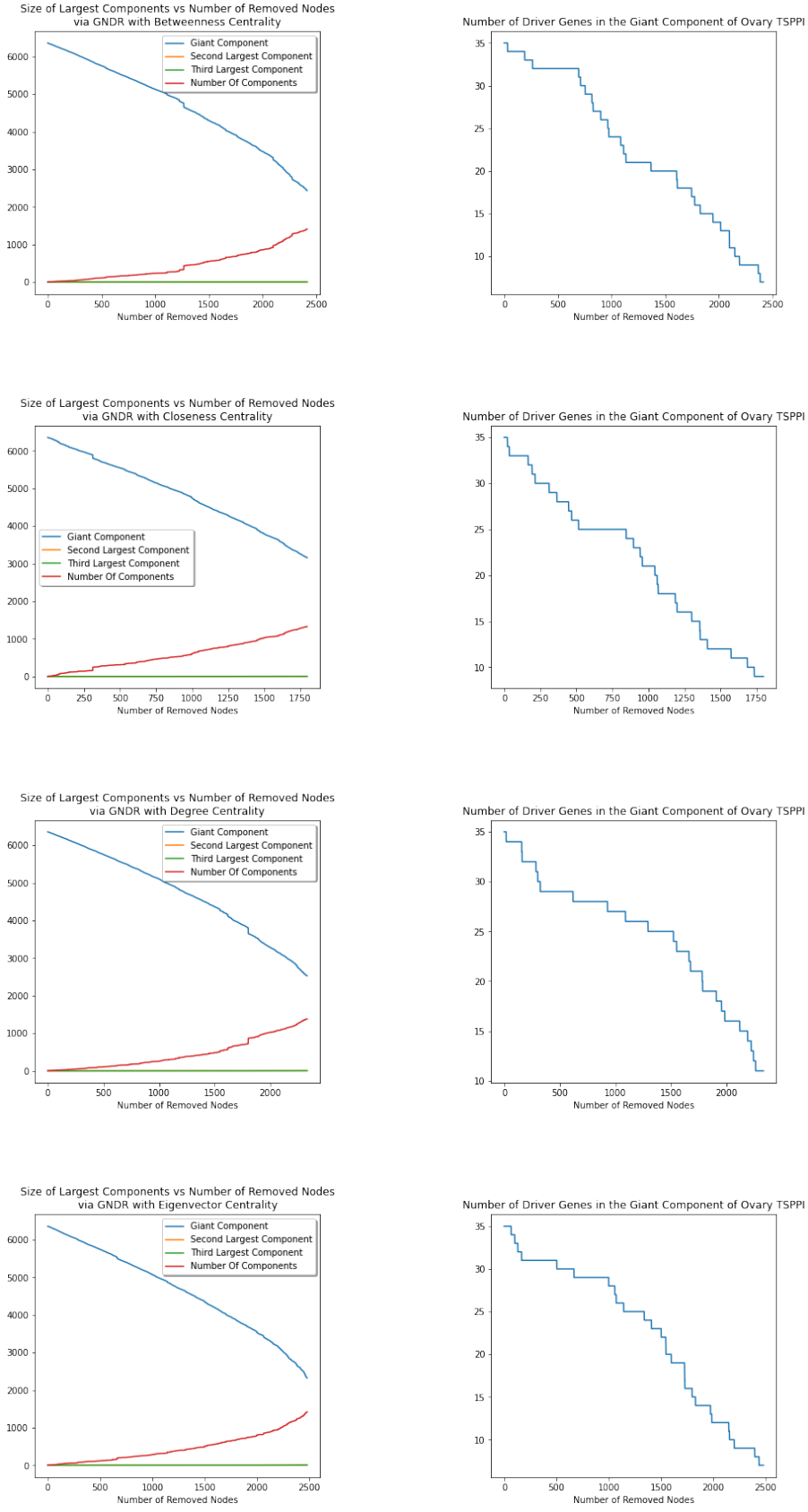


Figure 4.20: Line plots of the number of components, component sizes, and driver gene counts in ovary TSPPI.

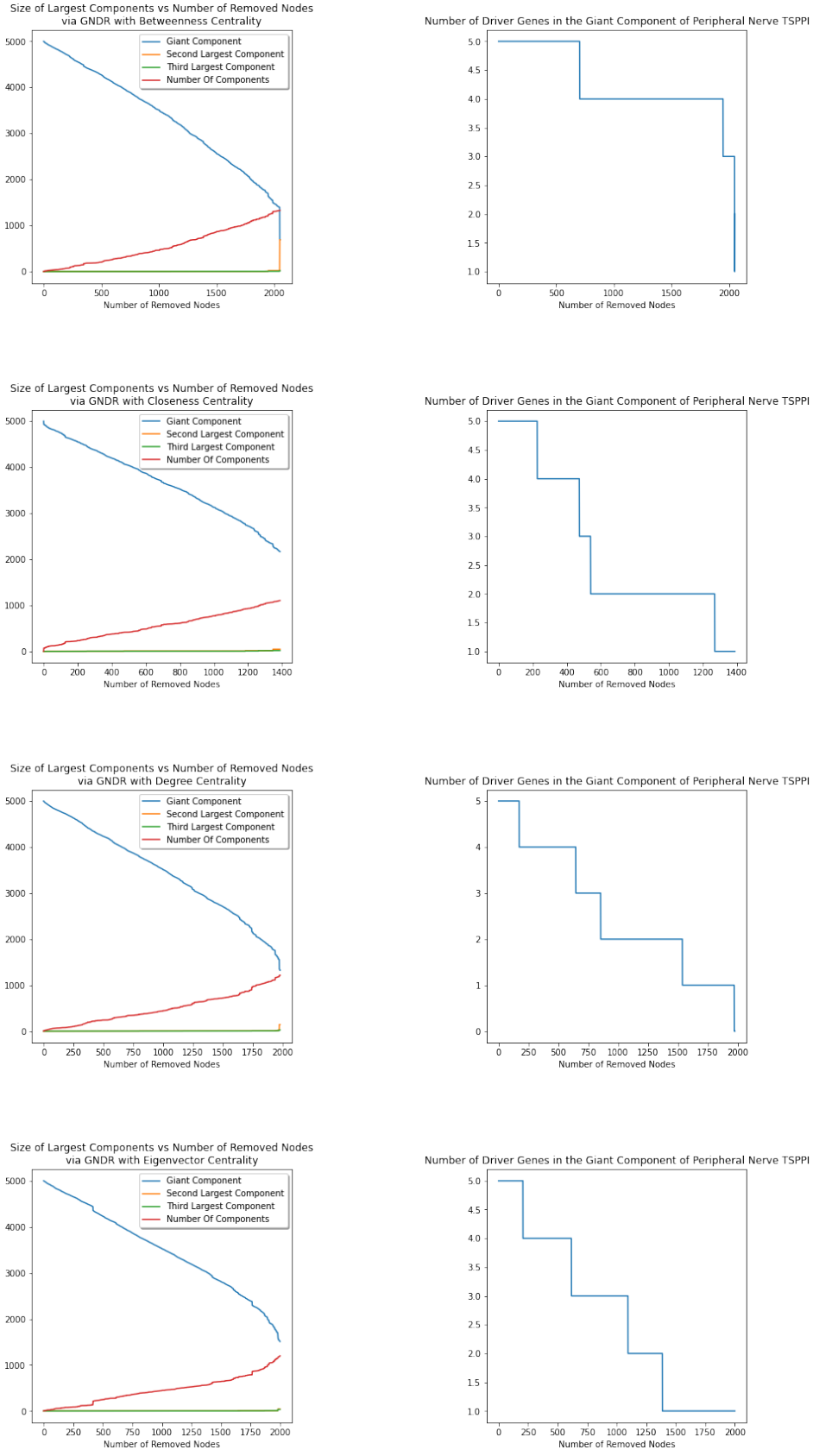


Figure 4.21: Line plots of the number of components, component sizes, and driver gene counts in peripheral nerve TSPPI.

Let k be the number of iterations, r be the number of removed nodes, n denote the number of nodes and m denote the number of edges in a network. Since we repeat the reinsertion for all nodes, the number of iterations is equal to the number of nodes, that is k = r. Since node addition cost has $O(r)$ time complexity, $O(rn)$ time complexity is added to the GND time complexities.

4.3 Comparison Between Network Dismantling Results

Usually, there are driver nodes in the hubs that hold the driver subnetworks together. For this reason, some of the drivers in the initial network should remain while the remaining driver nodes in the fragile parts of it should be deleted. In this case, the desired resulting subnetworks contain both driver nodes and non-driver nodes. The resulting networks are in Table C.1. In Table 4.6, Table 4.7, Table **??**, Table **??**, and Table **??**, we show only method comparison for each largest connected component.

Table 4.6: Method Comparison of Breast TSPPI

Method	LCC Size	Driver Node Count in LCC
Betweenness Centrality	830	2
GND with Betweenness Centrality	1052	7
GNDR with Betweenness Centrality	1052	7
Closeness Centrality	1898	6
GND with Closeness Centrality	3297	18
GNDR with Closeness Centrality	3297	18
Degree Centrality	839	0
GND with Degree Centrality	1053	3
GNDR with Degree Centrality	1053	4
Eigenvector Centrality	473	1
GND with Eigenvector Centrality	780	3
GNDR with Eigenvector Centrality	780	3

Table 4.7: Method Comparison of Liver TSPPI

Method	LCC Size	Driver Node Count in LCC
Betweenness Centrality	1405	4
GND with Betweenness Centrality	1613	8
GNDR with Betweenness Centrality	1613	8
Closeness Centrality	1590	3
GND with Closeness Centrality	3088	12
GNDR with Closeness Centrality	3088	12
Degree Centrality	1694	1
GND with Degree Centrality	1704	4
GNDR with Degree Centrality	1704	4
Eigenvector Centrality	2343	7
GND with Eigenvector Centrality	2348	3
GNDR with Eigenvector Centrality	2348	10

Table 4.8: Method Comparison of Lymph Node TSPPI

Method	LCC Size	Driver Node Count in LCC
Betweenness Centrality	778	3
GND with Betweenness Centrality	873	5
GNDR with Betweenness Centrality	873	5
Closeness Centrality	1619	5
GND with Closeness Centrality	2626	7
GNDR with Closeness Centrality	2626	7
Degree Centrality	2303	5
GND with Degree Centrality	2306	10
GNDR with Degree Centrality	2306	10
Eigenvector Centrality	1353	4
GND with Eigenvector Centrality	1527	10
GND with Eigenvector Centrality	1527	10

Table 4.9: Method Comparison of Ovary TSPPI

Method	LCC Size	Driver Node Count in LCC
Betweenness Centrality	306	19
GND with Betweenness Centrality	386	7
GNDR with Betweenness Centrality	386	7
Closeness Centrality	414	18
GND with Closeness Centrality	469	9
GNDR with Closeness Centrality	469	9
Degree Centrality	617	11
GND with Degree Centrality	676	11
GNDR with Degree Centrality	676	11
Eigenvector Centrality	204	17
GND with Eigenvector Centrality	273	7
GNDR with Eigenvector Centrality	273	7

Table 4.10: Method Comparison of Peripheral Nerve TSPPI

Method	LCC Size	Driver Node Count in LCC
Betweenness Centrality	583	1
GND with Betweenness Centrality	693	1
GNDR with Betweenness Centrality	693	2
Closeness Centrality	1110	1
GND with Closeness Centrality	2166	1
GNDR with Closeness Centrality	2166	1
Degree Centrality	1315	1
GND with Degree Centrality	1322	0
GNDR with Degree Centrality	1322	0
Eigenvector Centrality	1497	5
GND with Eigenvector Centrality	1509	1
GND with Eigenvector Centrality	1509	1

These results show that network dismantling by iterative centrality calculations performs more targeted attacks, makes dismantling of the network faster. On the other hand, it destroys the starting points for finding driver subnetworks by deleting drivers earlier. When the GND results are examined, the networks reach about the same largest component size by deleting much more nodes than methods based only on centrality. When the implementation of the reinsertion algorithm is added to GND, bringing the largest component to the specified size is achieved with less number of removals. In addition, node reinsertion from top to bottom of the removal list allows driver nodes to be reinserted to the network, while deleting them does not affect the dismantling of the largest component.

Table 4.11: Method Evaluation Results by Formula 310. Each value written in italic is the highest value of the row it is in. Each value written in bold is the highest value among the values that are calculated for a tissue.

Method	Iterative Centrality			
Centrality Metric	Betweenness	Closeness	Degree	Eigenvector
Breast	0.380	*0.499*	0	0.334
Liver	0.602	0.399	0.125	*0.632*
Lymph Node	*0.867*	0.694	0.488	0.665
Peripheral Nerve	1.717	0.902	0.761	***3.344***
Ovary	17.194	12.0397	4.937	***23.076***
Method	GND			
Centrality Metric	Betweenness	Closeness	Degree	Eigenvector
Breast	***1.050***	0.862	0.599	0.607
Liver	***1.049***	0.822	0.496	0.270
Lymph Node	1.288	0.599	0.975	***1.473***
Peripheral Nerve	*2.889*	0.462	0	0.663
Ovary	5.022	5.314	4.505	*7.100*
Method	GNDR			
Centrality Metric	Betweenness	Closeness	Degree	Eigenvector
Breast	***1.050***	0.862	0.599	0.607
Liver	***1.049***	0.822	0.496	0.270
Lymph Node	1.288	0.599	0.975	***1.473***
Peripheral Nerve	*2.889*	0.462	0	0.663
Ovary	5.022	5.314	4.505	*7.100*

Table 4.12: Method Evaluation Results by Formula 311. Each value written in italic is the highest value of the row it is in. Each value written in bold is the highest value among the values that are calculated for a tissue.

Method	Iterative Centrality			
Centrality Metric	Betweenness	Closeness	Degree	Eigenvector
Breast	1.071	*1.309*	0	1.056
Liver	*1.562*	0.871	0.349	1.486
Lymph Node	*3.692*	2.206	2.062	2.372
Peripheral Nerve	1.826	0.958	0.996	***3.509***
Ovary	13.033	9.080	6.587	***18.039***
Method	GND			
Centrality Metric	Betweenness	Closeness	Degree	Eigenvector
Breast	*2.091*	1.228	1.312	1.699
Liver	*2.092*	1.143	1.908	0.913
Lymph Node	*2.372*	2.339	1.690	1.544
Peripheral Nerve	*4.457*	2.195	N/A	2.630
Ovary	4.908	5.518	3.759	*7.497*
Method	GNDR			
Centrality Metric	Betweenness	Closeness	Degree	Eigenvector
Breast	***3.059***	1.351	1.938	1.728
Liver	***2.437***	1.187	1.422	1.527
Lymph Node	*3.514*	1.654	1.799	**2.441**
Peripheral Nerve	*3.129*	2.308	0	2.908
Ovary	5.259	5.631	3.987	*8.160*

We evaluate these results by Formula 310 for model evaluation since our primary concern is to get a driver subnetwork of the cancer related with the tissue. However, since the size of the last largest component is equal in the GND and GNDR results, the calculations of Formula 310 are equal as well. At this point, we use Formula 311, considering that driver subnetworks may be generated from other components by counting in the remaining driver nodes in other components.

- Breast TSPPI: According to Formula 310, GND with Betweenness Centrality and GNDR with Betweenness Centrality gives the best result which is around 1.05. Since these two results are equal, Formula 311 is employed in the evaluation of GND and GNDR, and GNDR Betweenness Centrality is chosen for Breast TSPPI since GNDR with Betweenness Centrality had a higher value.

- Liver TSPPI: There is a similar situation with Breast TSPPI in the liver TSPPI. For this reason, GNDR with Betweenness Centrality is chosen based on the Formula 311 result.

- Lymph Node TSPPI: For the lymph node TSPPI, there is a different situation from the above situations. According to Formula 310, GND and GNDR with Eigenvector Centrality have the best results, while according to Formula 311 Iterative Betweenness Centrality has the best result. However, as mentioned in the evaluation phase, we choose GNDR with Eigenvector Centrality because we use Formula 310 as a priority and evaluate the methods with the same values according to its results according to Formula 311.

- Peripheral Nerve TSPPI: Iterative Eigenvector Centrality performs best for Peripheral Nerve.

- Ovary TSPPI: Iterative Eigenvector Centrality has the best result.

It can be deduced that GNDR applications are effective in networks with more than 7000 nodes, while it is more appropriate to dismantle networks with less than 7000 nodes by applying iterative centrality algorithms.

4.4 Driver Subnetwork Collection by Personalized PageRank

Personalized PageRank algorithm is biased for a given set of nodes. In this algorithm, random walks start and end in this set. In this study, we marked driver gene set as the given set. While assigning 0 to other nodes, we assigned 1 to these nodes to start and end random walks in this set. We applied Personalized PageRank algorithm to obtained subnetworks which have driver nodes and more than 20 nodes in them. We selected nodes with greater than Personalized PageRank ranking score of 0.01. The results are detailed in the Table 4.13.

Table 4.13: Personalized PageRank Result

Tissue	Subnetwork Size Before Personalized PageRank	Driver Node Count	Resulting Subnetwork Size
Breast	1052	7	90
Breast	1052	1	41
Liver	1613	8	122
Lymph Node	1527	10	223
Lymph Node	260	1	36
Ovary	204	17	155
Peripheral Nerve	1497	5	84

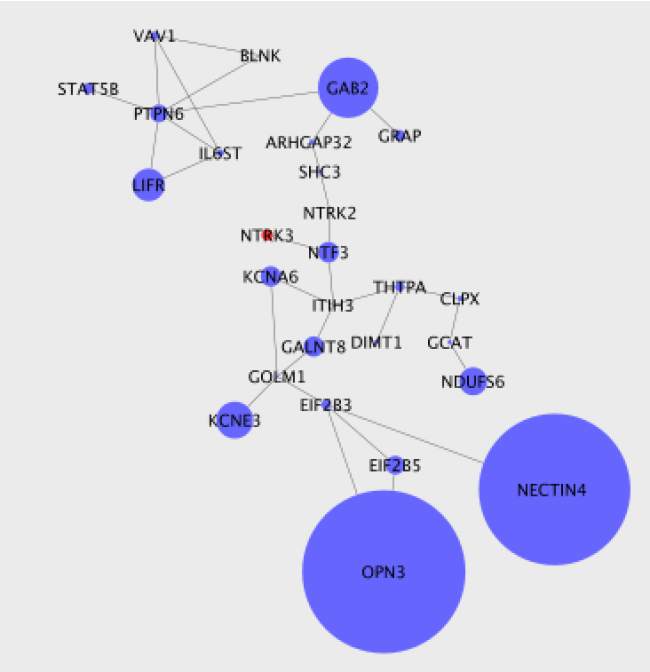


Figure 4.22: The First Driver Subnetwork of Breast TSPPI. Driver nodes are illustrated in red color, normal nodes are illustrated in blue color. Node sizes vary according to the mutation frequency information obtained from the cBioPortal. The larger the nodes, the more frequently they appear to be mutated in cancer. In this figure, only the driver nodes, the most frequent nodes, their first-order neighbors, and the nodes that form a bridge between them are visualized for simplicity.

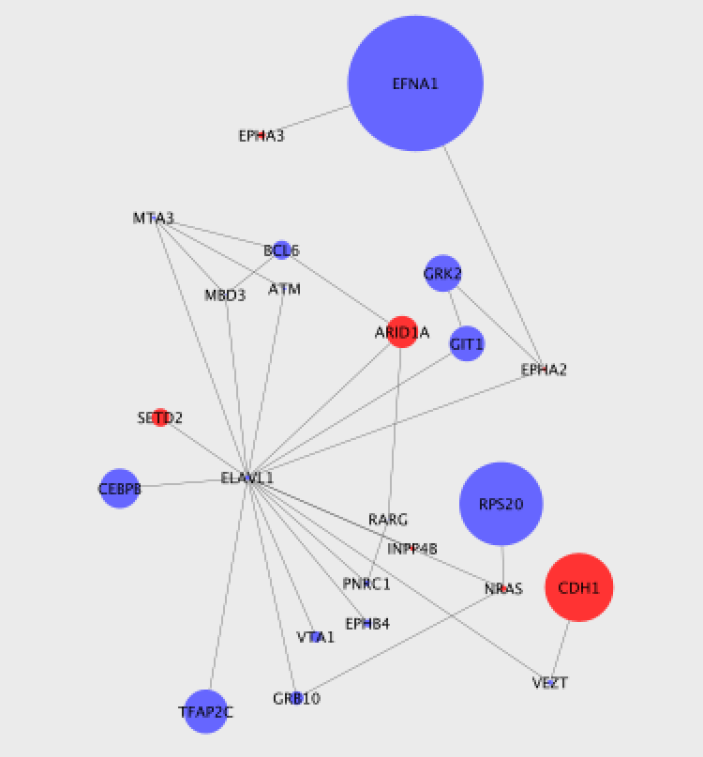


Figure 4.23: The Second Driver Subnetwork of Breast TSPPI. Driver nodes are illustrated in red color, normal nodes are illustrated in blue color. Node sizes vary according to the mutation frequency information obtained from the cBioPortal. The larger the nodes, the more frequently they appear to be mutated in cancer. In this figure, only the driver nodes, the most frequent nodes, their first-order neighbors, and the nodes that form a bridge between them are visualized for simplicity.

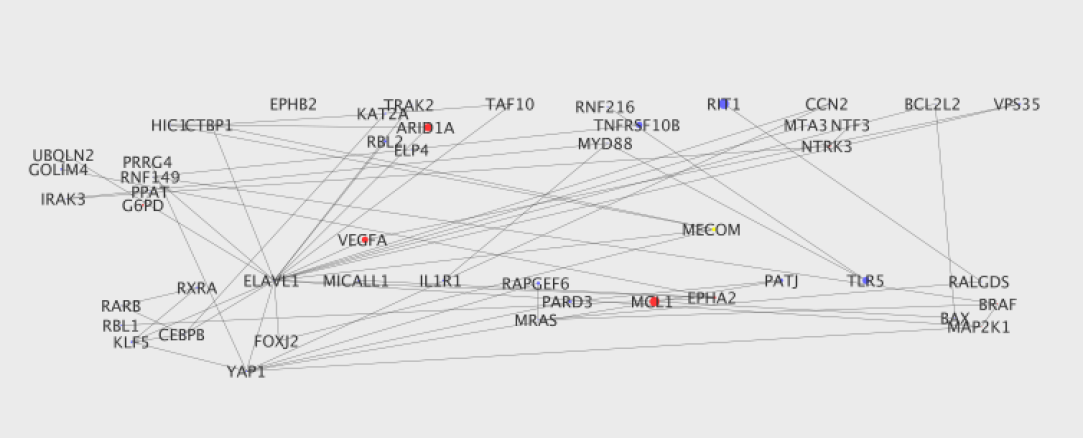


Figure 4.24: Driver Subnetwork of Liver TSPPI. Driver nodes are illustrated in red color, normal nodes are illustrated in blue color. Node sizes vary according to the mutation frequency information obtained from the cBioPortal. The larger the nodes, the more frequently they appear to be mutated in cancer.

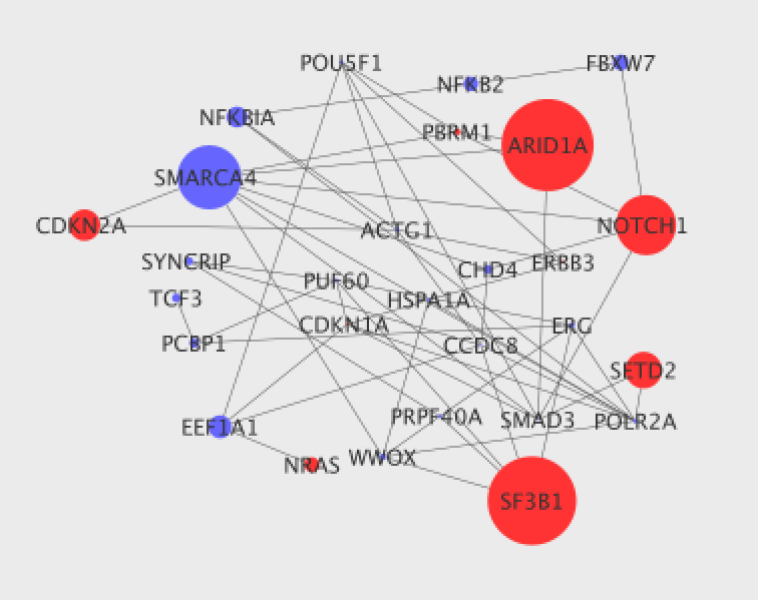


Figure 4.25: The First Driver Subnetwork of Lymph Node TSPPI. Driver nodes are illustrated in red color, normal nodes are illustrated in blue color. Node sizes vary according to the mutation frequency information obtained from the cBioPortal. The larger the nodes, the more frequently they appear to be mutated in cancer. In this figure, only the driver nodes, the most frequent nodes, their first-order neighbors, and the nodes that form a bridge between them are visualized for simplicity.

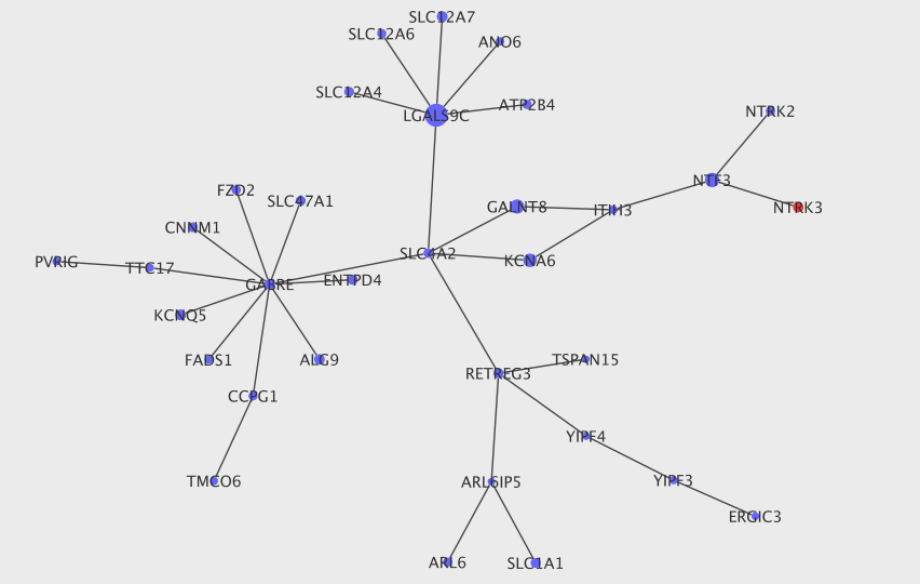


Figure 4.26: The Second Driver Subnetwork of Lymph Node TSPPI. Driver nodes are illustrated in red color, normal nodes are illustrated in blue color. Node sizes vary according to the mutation frequency information obtained from the cBioPortal. The larger the nodes, the more frequently they appear to be mutated in cancer.

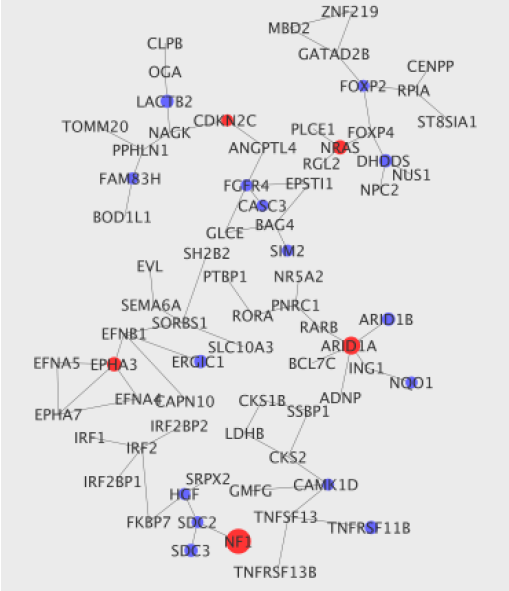


Figure 4.27: Driver Subnetwork of Peripheral Nerve TSPPI. Driver nodes are illustrated in red color, normal nodes are illustrated in blue color. Node sizes vary according to the mutation frequency information obtained from the cBioPortal. The larger the nodes, the more frequently they appear to be mutated in cancer.

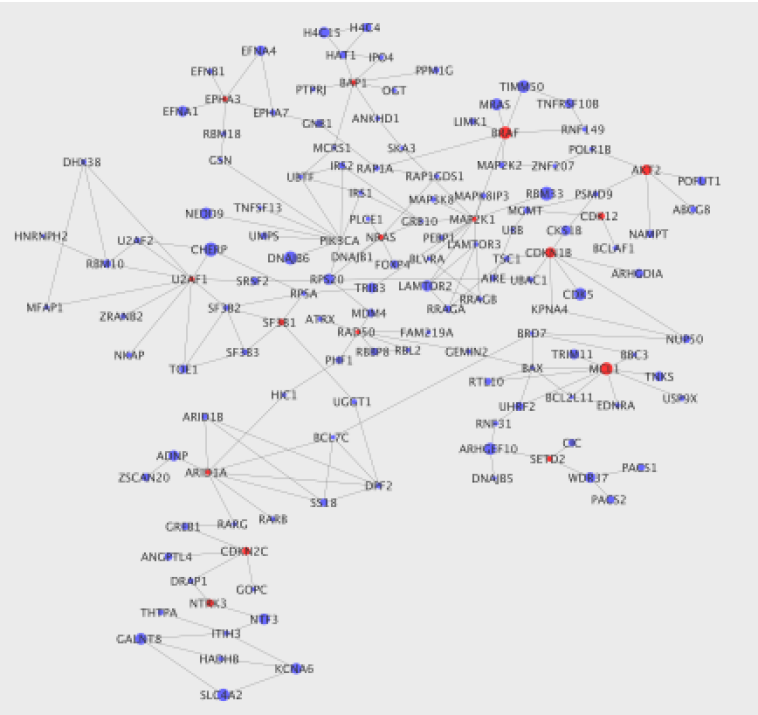


Figure 4.28: Driver Subnetwork of Ovary TSPPI. Driver nodes are illustrated in red color, normal nodes are illustrated in blue color. Node sizes vary according to the mutation frequency information obtained from the cBioPortal. The larger the nodes, the more frequently they appear to be mutated in cancer.

4.5 Functional Enrichment Analysis of the Proposed Driver Subnetworks

Functional enrichment analysis is a method that determines overrepresented genes and proteins and their pathways from given gene and protein sets. After specifying p-value cutoff as 0.02, the remaining number of pathways related to the proposed driver subnetworks are shown in the Table 4.14.

Table 4.14: Functional Enrichment Result Counts

Tissue	Number of Pathways Before Filtering	Number of Pathways After Filtering
Breast	41	17
Breast	113	58
Liver	172	69
Lymph Node	437	222
Lymph Node	17	10
Peripheral Nerve	54	23
Ovary	242	125

After obtaining these results, we repeated the functional enrichment analysis by removing the driver genes to show the effect of the driver genes on the results. Table 4.15 shows driver nodes do not actually affect count of the significant pathways in the results.

Table 4.15: Functional Enrichment Analysis Result Counts when Driver Nodes are Excluded

Tissue	Number of Pathways Before Filtering	Number of Pathways After Filtering
Breast	36	15
Breast	100	55
Liver	139	65
Lymph Node	424	220
Lymph Node	10	7
Peripheral Nerve	48	22
Ovary	166	103

As a result of the comparison, it was observed that the number of significant functional enrichment results did not decrease much. Table 4.16, Table 4.17, Table 4.18, Table 4.19, Table 4.20, Table 4.21, and Table 4.22 show functional enrichment results of all nodes in each subnetwork.

In Table 4.16, first result of functional enrichment analysis is transmembrane receptor protein tyrosine kinase signaling pathway which participates in various downstream signaling pathways such as MAPK, PI3K/Akt and JAK/STAT (Butti et al., 2018). SCF/cKIT signaling pathway leads autophosphorylation and initiation of signal trans-

duction (Lennartsson & Rönnstrand, 2012). There are pathways related with the roles of neurotrophins. Neurotrophins ensure inhibition of breast cancer cell survival, proliferation and invasion while Interleukin-3, Interleukin-5 and GM-CSF participates in the activation of the MAP kinase pathway (Kannan et al., 2000). GPVI supports tumor cell extravasation and metastasis in breast cancer cells (Mammadova-Bach et al., 2020).

Table 4.16: Functional Enrichment Analysis Results of the First Proposed Driver Subnetwork of Breast TSPPI

Pathway	P-Value
GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway	2 x 10^{-6}
Signaling by SCF-cKIT	0.000191
GO:0018108 peptidyl-tyrosine phosphorylation	0.000346
GO:0008284 positive regulation of cell proliferation	0.000495
Interleukin-3, Interleukin-5 and GM-CSF signaling	0.00243
GO:0038179 neurotrophin signaling pathway	0.004402
GO:0009967 positive regulation of signal transduction	0.007967
hsa04722:Neurotrophin signaling pathway	0.00802
GO:0032981 mitochondrial respiratory chain complex I assembly	0.008479
GO:0038165 oncostatin-M-mediated signaling pathway	0.008785
GO:0042490 mechanoreceptor differentiation	0.008785
GO:0033138 positive regulation of peptidyl-serine phosphorylation	0.010384
GO:0070120 ciliary neurotrophic factor-mediated signaling pathway	0.01097
Complex I biogenesis	0.012325
GPVI-mediated activation cascade	0.012757
hsa04630:Jak-STAT signaling pathway	0.013425
GO:0032057 negative regulation of translational initiation in response to stress	0.015325

The second proposed subnetwork of breast TSPPI comprises very important signaling pathways, pathways related to viral infection effects and cell functionality in breast cancer. Pathways such as cell proliferation, TP53 regulation, chromatin remodeling etc. are common in most cancer types. However, in these results of the subnetwork, there are pathways which are not usual for all cancer types, as well. For instance, Fanconi Anemia Pathway (Fang, Wu, Zhang, Liu, & Zhang, 2020) includes many genes from the breast cancer driver genes and breast cancer mutations may be affected by fanconi anemia. HIV (Brandão et al., 2021), insulin signaling pathway (Yang & Yee, 2012), and Circadian Clock (Cadenas et al., 2014) are the other examples.

Table 4.17: The Most Significant Functional Enrichment Analysis Results of the Second Proposed Driver Subnetwork of Breast TSPPI

Pathway	P-Value
GO:0006366 transcription from RNA polymerase II promoter	$1x10^{-5}$
GO:0006367 transcription initiation and from RNA polymerase II promoter	$1.2x10^{-5}$
GO:0010941 regulation of cell death	$2x10^{-5}$
GO:0048013 ephrin receptor signaling pathway	$7.4x10^{-5}$
GO:0006368 transcription from RNA polymerase II promoter	$7.4x10^{-5}$
GO:0036297 interstrand cross-link repair	0.000109
Fanconi Anemia Pathway	0.000141
HIV Transcription Initiation	0.000244
RNA Polymerase II HIV Promoter Escape	0.000244
RNA Polymerase II Promoter Escape	0.000244
RNA Polymerase I Transcription Initiation	0.000266
Circadian Clock	0.000312
GO:1901796 regulation of signal transduction by p53 class mediator	0.000414
RORA activates gene expression	0.000754
GO:0048096 chromatin-mediated maintenance of transcription	0.000891
GO:0006338 chromatin remodeling	0.000947
GO:0045893 positive regulation of transcription, DNA-templated	0.001165
RMTs methylate histone arginines	0.001734
RNA Polymerase II Pre-transcription Events	0.001818
hsa05168:Herpes simplex infection	0.002175
GO:0048009 insulin-like growth factor receptor signaling pathway	0.002215
GO:0042127 regulation of cell proliferation	0.002476
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	0.00339
GO:0010971 positive regulation of G2/M transition of mitotic cell cycle	0.003676
GO:0018105 peptidyl-serine phosphorylation	0.003723
GO:0006977 DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	0.003818
GO:0016032 viral process	0.004087
GO:0006974 cellular response to DNA damage stimulus	0.004089
hsa03460:Fanconi anemia pathway	0.00693
GO:0009314 response to radiation	0.008788
GO:0042149 cellular response to glucose starvation	0.009409
GO:0000077 DNA damage checkpoint	0.010048
hsa04062:Chemokine signaling pathway	0.01176
GO:0071456 cellular response to hypoxia	0.012756
GO:0006352 DNA-templated transcription, initiation	0.01427
GO:0032682 negative regulation of chemokine production	0.01511
YAP1- and WWTR1 (TAZ)-stimulated gene expression	0.016059
hsa04015:Rap1 signaling pathway	0.019013
GO:0043966 histone H3 acetylation	0.019993

Table 4.18: The Most Significant Functional Enrichment Analysis Results of the Proposed Driver Subnetwork of Liver TSPPI

Pathway	P-Value
GO:0010628 positive regulation of gene expression	$1.2\text{x}10^{-5}$
hsa04722:Neurotrophin signaling pathway	$1.7\text{x}10^{-5}$
GO:0006357 regulation of transcription from RNA polymerase II promoter	$4.6\text{x}10^{-5}$
hsa04015:Rap1 signaling pathway	0.000158
GO:0006468 protein phosphorylation	0.000273
GO:0008630 intrinsic apoptotic signaling pathway in response to DNA damage	0.000287
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	0.000326
GO:0008284 positive regulation of cell proliferation	0.000329
GO:0045892 negative regulation of transcription, DNA-templated	0.000584
GO:0006367 transcription initiation from RNA polymerase II promoter	0.000595
hsa04010:MAPK signaling pathway	0.000631
GO:0007049 cell cycle	0.000698
GO:0000187 activation of MAPK activity	0.00081
GO:0016032 viral process	0.000991
Negative feedback regulation of MAPK pathway	0.00107
hsa05200:Pathways in cancer	0.001146
GO:0010332 response to gamma radiation	0.001193
GO:0007265 Ras protein signal transduction	0.001309
hsa04151:PI3K-Akt signaling pathway	0.001512
GO:0097192 extrinsic apoptotic signaling pathway in absence of ligand	0.001564
GO:2000146 negative regulation of cell motility	0.001594
hsa04068:FoxO signaling pathway	0.001703
GO:0010629 negative regulation of gene expression	0.002423
GO:0006338 chromatin remodeling	0.002789
hsa05205:Proteoglycans in cancer	0.00288
GO:0097320 membrane tubulation	0.002884
hsa04620:Toll-like receptor signaling pathway	0.003352
hsa04668:TNF signaling pathway	0.00349
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	0.003566
GO:0001525 angiogenesis	0.004161
GO:0006915 apoptotic process	0.005231

Also, the subnetwork of liver TSPPI have common in almost all cancer types among them as well as abundant pathways. Examples of special pathways are the Toll-like receptor signaling pathway (French, Oliva, French, Li, & Bardag-Gorce, 2010), angiogenesis (Fernández et al., 2009) and the TNF signaling pathway (Liedtke & Trautwein, 2012).

As in the previous tables, there are many networks related with mRNA, DNA and other elements of cell division in Table 4.19. However, there are unique pathways for lymphoma. For example, Epstein-Barr virus infection (Brady, MacArthur, & Farrell, 2008) causes a subtype of lymphoma, called Burkitt lymphoma.

Table 4.19: The Most Significant Functional Enrichment Analysis Results of the First Proposed Driver Subnetwork of Lymph Node TSPPI

Pathway	P-Value
GO:0000398 mRNA splicing, via spliceosome	$2.19\text{x}10^{-33}$
GO:0006338 chromatin remodeling	$2.61\text{x}10^{-17}$
GO:0043044 ATP-dependent chromatin remodeling	$1.29\text{x}10^{-15}$
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	$3.68\text{x}10^{-14}$
GO:0006351 transcription, DNA-templated	$1.17\text{x}10^{-13}$
GO:0006397 mRNA processing	$1.38\text{x}10^{-13}$
GO:1901796 regulation of signal transduction by p53 class mediator	$4.46\text{x}10^{-12}$
mRNA Splicing - Minor Pathway	$8.01\text{x}10^{-12}$
GO:0045892 negative regulation of transcription, DNA-templated	$8.46\text{x}10^{-12}$
GO:0016925 protein sumoylation	$2.34\text{x}10^{-12}$
GO:0016032 viral process	$3.78\text{x}10^{-11}$
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	$1.56\text{x}10^{-10}$
GO:0016575 histone deacetylation	$2.50\text{x}10^{-10}$
GO:0010467 gene expression	$3.93\text{x}10^{-10}$
GO:0000245 spliceosomal complex assembly	$7.83\text{x}10^{-10}$
Processing of Capped Intron-Containing Pre-mRNA	$9.74\text{x}10^{-10}$
GO:0006281 DNA repair	$8.95\text{x}10^{-9}$
GO:0016569 covalent chromatin modification	$2.15\text{x}10^{-8}$
GO:0006396 RNA processing	$4.25\text{x}10^{-8}$
RMTs methylate histone arginines	$1.75\text{x}10^{-7}$
GO:0045893 positive regulation of transcription, DNA-templated	$1.88\text{x}10^{-7}$
GO:0048511 rhythmic process	$3.84\text{x}10^{-7}$
GO:1900034 regulation of cellular response to heat	$4.54\text{x}10^{-7}$
GO:0006364 rRNA processing	$6.48\text{x}10^{7}$
hsa05203:Viral carcinogenesis	$1.45\text{x}10^{-6}$
SUMOylation of DNA damage response and repair proteins	$1.60\text{x}10^{-6}$
GO:0006337 nucleosome disassembly	$1.75\text{x}10^{-6}$
GO:0000184 nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	$2.91\text{x}10^{-6}$
GO:0008285 negative regulation of cell proliferation	$3.37\text{x}10^{-6}$
GO:0006357 regulation of transcription from RNA polymerase II promoter	$3.89\text{x}10^{-6}$
GO:0006260 DNA replication	$4.85\text{x}10^{-6}$
GO:0070932 histone H3 deacetylation	$5.53\text{x}10^{-6}$
Regulation of HSF1-mediated heat shock response	$5.88\text{x}10^{-6}$
hsa04550:Signaling pathways regulating pluripotency of stem cells	$1.26\text{x}10^{-5}$
GO:0006974 cellular response to DNA damage stimulus	$1.49\text{x}10^{-5}$
hsa05169:Epstein-Barr virus infection	$1.82\text{x}10^{-5}$
GO:0032508 DNA duplex unwinding	$1.88\text{x}10^{-5}$

Table 4.20 contains chemical pathways of lymphoma such as chloride transmembrane transport, ion transport, and potassium ion transmembrane transport (Comes et al., 2015).

Table 4.20: Functional Enrichment Analysis Results of Second Proposed Driver Subnetwork of Lymph Node TSPPI

Pathway	P-Value
GO:1902476 chloride transmembrane transport	$1x10^{-6}$
GO:0071805 potassium ion transmembrane transport	$6.1x10^{-5}$
GO:0006811 ion transport	$7.3x10^{-5}$
Cation-coupled Chloride cotransporters	$7.8x10^{-5}$
GO:0033138 positive regulation of peptidyl-serine phosphorylation	0.00026
GO:0007268 chemical synaptic transmission	0.000833
GO:0038179 neurotrophin signaling pathway	0.00357
GO:0042490 mechanoreceptor differentiation	0.007128
GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway	0.012679
GO:0006884 cell volume homeostasis	0.017727

Table 4.21: The Most Significant Functional Enrichment Analysis Results of the Proposed Driver Subnetwork of Ovary TSPPI

Pathway	P-Value
MAP2K and MAPK activation	$1x10^{-8}$
hsa04722:Neurotrophin signaling pathway	$1.8x10^{-7}$
hsa04068:FoxO signaling pathway	$4.56x10^{-6}$
hsa04150:mTOR signaling pathway	$5.72x10^{-6}$
GO:0048013 ephrin receptor signaling pathway	$9x10^{-6}$
GO:0006397 mRNA processing	$2.41x10^{-5}$
hsa04151:PI3K-Akt signaling pathway	$6.32x10^{-5}$
GO:0006336 DNA replication-independent nucleosome assembly	$6.69x10^{-5}$
hsa04014:Ras signaling pathway	$8.47x10^{-5}$
GO:0008285 negative regulation of cell proliferation	0.00015799
hsa04015:Rap1 signaling pathway	0.00021779
TP53 Regulates Metabolic Genes	0.00043384
hsa04012:ErbB signaling pathway	0.00063508
hsa05230:Central carbon metabolism in cancer	0.00104368
hsa05231:Choline metabolism in cancer	0.00139167
hsa04370:VEGF signaling pathway	0.00636568
hsa04660:T cell receptor signaling pathway	0.0073232
hsa05200:Pathways in cancer	0.00789311
hsa04664:Fc epsilon RI signaling pathway	0.00931554
hsa04024:cAMP signaling pathway	0.01027012
hsa05205:Proteoglycans in cancer	0.01081811
hsa04071:Sphingolipid signaling pathway	0.01535084
hsa04650:Natural killer cell mediated cytotoxicity	0.01638564
hsa04152:AMPK signaling pathway	0.0169201

There are important pathways such as FoxO (De Brachène & Demoulin, 2016), mTOR (Mabuchi, Hisamatsu, & Kimura, 2011), ephrin receptor (Herath et al., 2006), insulin

(Lukanova et al., 2002), PI3K-Akt (Mazzoletti & Broggini, 2010), Ras (Sheppard et al., 2013), Rap1 (Y.-L. Zhang, Wang, Cheng, Ring, & Su, 2017), VEGF (Luo, Jiang, King, & Chen, 2008), and Sphingolipid (Kreitzburg, van Waardenburg, & Yoon, 2018) signaling pathways in Table 4.21. Pathways such as cell proliferation, TP53 regulation, cyclin D events, central carbon mechanism in cancer are common in most cancer types.

Table 4.22: The Most Significant Functional Enrichment Analysis Results of the Proposed Driver Subnetwork of Peripheral Nerve TSPPI

Pathway	P-Value
EPH-Ephrin signaling	$9\text{x}10^{-6}$
hsa04360:Axon guidance	$9.3\text{x}10^{-5}$
Nuclear Receptor transcription pathway	0.002278
GO:0043401 steroid hormone mediated signaling pathway	0.002356
GO:0007265 Ras protein signal transduction	0.004221
GO:0043044 ATP-dependent chromatin remodeling	0.005059
TNFs bind their physiological receptors	0.007301
GO:0006351 transcription, DNA-templated	0.007498
hsa04015:Rap1 signaling pathway	0.007746
GO:0007399 nervous system development	0.010783
GO:0050919 negative chemotaxis	0.010854
GO:0034605 cellular response to heat	0.01277
GO:0030522 intracellular receptor signaling pathway	0.01344
GO:0043525 positive regulation of neuron apoptotic process	0.017008
GO:0033209 tumor necrosis factor-mediated signaling pathway	0.017567
GO:0044772 mitotic cell cycle phase transition	0.018453
GO:0046578 regulation of Ras protein signal transduction	0.027553
GO:0051965 positive regulation of synapse assembly	0.033662
GO:0043087 regulation of GTPase activity	0.036699
GO:0048096 chromatin-mediated maintenance of transcription	0.041047
GO:2000573 positive regulation of DNA biosynthetic process	0.058752
GO:0044030 regulation of DNA methylation	0.063127
GO:0070507 regulation of microtubule cytoskeleton organization	0.071818
GO:0051893 regulation of focal adhesion assembly	0.076134
GO:0031290 retinal ganglion cell axon guidance	0.084706
GO:0001525 angiogenesis	0.085347

In Table 4.22, ephrin signaling, axon guidance, nervous system development, retinal ganglion cell axon guidance, and positive regulation of neuron apoptotic process are related with the elements of nervous system.

CHAPTER 5

DISCUSSION

Finding important submodules and candidate driver genes is a very critical task in bioinformatics. Many studies in the literature have tried to find important parts in a biological network using network centrality (Patil & Nielsen, 2005; Guo et al., 2007; Ma, Schadt, Kaplan, & Zhao, 2011; Chin et al., 2014; Zhuang, Jiang, He, Zhou, & Yue, 2015; Frainay & Jourdan, 2017; Reyna, Leiserson, & Raphael, 2018; Li, Li, Wu, Pan, & Wang, 2018; Li et al., 2018; Jiang et al., 2021). Some of these studies have tried to identify critical submodules of tissues or tissue-related diseases, including breast cancer (Zhuang et al., 2015), osteoporosis (Jiang et al., 2021) and colorectal cancer (Nibbe, Koyutürk, & Chance, 2010). In this thesis, we aim to extract various cancer-driver subnetworks, which are smaller and more meaningful parts of a network, by dismantling 5 different TSPPIs by using driver genes. While previous studies apply several methods for either the reconstruction of a TSPPI (Wong, Krishnan, & Troyanskaya, 2018; Bossi & Lehner, 2009; Liu et al., 2008) or the identification of a cancer driver subnetwork and/or cancer biomarker from a TSPPI (Ideker et al., 2002; Akavia et al., 2010; Beisser et al., 2010; Gevaert et al., 2013; Bertrand et al., 2015; Cho et al., 2016), we offer a more comprehensive pipeline using different methods with different weights using different TSPPIs and various centrality measures. For this reason, we present a methodology to interpret the results of several methods to dismantle the TSPPIs we selected, namely breast, liver, lymph node, ovary, and peripheral nerve, by using various centrality metrics as weight functions. First of all, before starting the analysis, we retrieved the networks corresponding to the selected tissues from the TissueNet v.2 database. In order to find out the driver gene counts in these networks, we separated the data we received from the Cancer Genome Interpreter according to the cancer types associated with that tissue. After adding the driver genes that are important for all cancers to the list of the driver genes for each tissue, we examined the frequency of the driver genes we obtained from cBioPortal to cause tumors in these tissues. According to the results we obtained from cBioPortal, we eliminated the genes that have never been found in clinical studies related to that tissue from among the driver genes. Since we get tissue-specific mutations, we can see that the driver genes for one tissue are imported as non-drivers for the other tissue.

To dismantle the TSPPIs, we attacked each of them with 12 different strategies and got the result we considered optimal for each. We can group these strategies under 3 headings: Iterative Centrality Metrics, GND with different weight functions, and GNDR with different weight functions. The results we obtained from iterative centrality metrics affected either the driver genes directly or their neighbors. We observed that the success order of betweenness centrality, closeness centrality, degree central-

ity and eigenvector centrality changed in each network. These results showed us the difference between TSPPIs once again. Although iterative centrality algorithms made successful targeted attacks to split the network, their computational complexities are very high and deleting most of the driver genes would cause us to exclude important pathways that driver subnetworks could contain. As an alternative strategy, we calculated the subnetworks that would be formed as a result of GND with different weight functions. In each iteration, GND takes the nodes that cover all the edges among a group of nodes it chooses and deletes them all at once. Therefore, there can be a lot of unnecessary deleted nodes with respect to the resulting subnetworks. We applied the reinsertion algorithm to detect these nodes. GND was deleting too many nodes to allow the network to reach sufficient fragmentation. On the other hand, the GNDR algorithm, significantly improved the GND results. Although GNDR outperforms iterative centrality algorithms in terms of both fragmentation rate and remaining number of driver nodes in networks with more nodes and edges, iterative centrality algorithms have been more successful in nerve TSPPI and ovary TSPPI networks that are relatively small networks with the initial node sizes less than 7000. GNDR or iterative centrality algorithms can be successful when applied to different networks for different purposes. In this case, there is no strategy that we can qualify as the "best" strategy for all networks because of the NP-hardness of the problem. For instance, *Batra et al.* compared 7 different methods to find a gold standard to this problem and concluded their study by revealing strengths and limitations of these methods (Batra et al., 2017). After the decision of the best strategy of each network, we used Personalized PageRank algorithm in order to find the most suitable hub subnetworks for each TSPPI. In these results, we found that there are nodes that are not in the list of driver nodes, but that are frequently seen to be mutated in the selected tissue related cancer. For now, most cancer drivers remain undetected. Since these results may lead to clinical justification of these nodes, finding a smaller subnetwork from the whole TSPPI is important. We also saw genes that were not labeled as biomarkers in our data, but frequently seen in mutations, in our driver subnetwork, which we eventually proposed. For example, while ARID1A and SMARCA4 mutate quite frequently in a subtype of lymphoma, in the biomarker data we obtained in our dataset, ARID1A was labeled as driver gene, while SMARCA4 was non-driver. However, in our results, we found that SMARCA4 is also a driver gene or a gene that affects driver genes (Love et al., 2012). Also, we interpreted our results with functional enrichment analysis. As a result, the pathways we revealed in the results that we consider as significant (p value < 0.02) contain both mutations that are mutated in tissue-specific cancer and common mutations in all cancer types. For example, splicing event occurs in all cancer types while Epstein-Barr virus infection causes only Burkitt lymphoma which is in the results of Lymph Node subnetwork enrichment analysis. Similarly, there is Ras protein signal transduction which is important for all tissues while axon guidance and retinal ganglion cell axon guidance are related to peripheral nerve tissue in the functional enrichment results of peripheral nerve. In addition, the two subnetworks we found for breast and lymph node were related to different pathway groups.

Our proposed pipeline may be applied the other tissues including pancreatic cancer, lung cancer, colorectal cancer etc. The only requirement is a TSPPI belonging to the tissue and known cancer driver genes of the tissue.

REFERENCES

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., ... Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. *Cell*, *143*(6), 1005–1017.

Altarelli, F., Braunstein, A., Dall'Asta, L., & Zecchina, R. (2013a). Large deviations of cascade processes on graphs. *Physical Review E*, *87*(6), 062115.

Altarelli, F., Braunstein, A., Dall'Asta, L., & Zecchina, R. (2013b). Optimizing spread dynamics on graphs by message passing. *Journal of Statistical Mechanics: Theory and Experiment*, *2013*(09), P09011.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., ... others (2004). Uniprot: the universal protein knowledgebase. *Nucleic acids research*, *32*(suppl_1), D115–D119.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... others (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, *25*(1), 25–29.

Barabási, A.-L., & Bonabeau, E. (2003). Scale-free networks. *Scientific american*, *288*(5), 60–69.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., ... Edgar, R. (2007). Ncbi geo: mining tens of millions of expression profiles—database and tools update. *Nucleic acids research*, *35*(suppl_1), D760–D765.

Barshir, R., Basha, O., Eluk, A., Smoly, I. Y., Lan, A., & Yeger-Lotem, E. (2013). The tissuenet database of human tissue protein–protein interactions. *Nucleic acids research*, *41*(D1), D841–D844.

Bar-Yehuda, R., & Even, S. (1981). A linear-time approximation algorithm for the weighted vertex cover problem. *Journal of Algorithms*, *2*(2), 198–203.

Basha, O., Barshir, R., Sharon, M., Lerman, E., Kirson, B. F., Hekselman, I., & Yeger-Lotem, E. (2017). The tissuenet v. 2 database: a quantitative view of protein-protein interactions across human tissues. *Nucleic acids research*, *45*(D1), D427–D431.

Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., ... Shah, S. P. (2012). Drivernet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome biology*, *13*(12), 1–14.

Batra, R., Alcaraz, N., Gitzhofer, K., Pauling, J., Ditzel, H. J., Hellmuth, M., ... List, M. (2017). On the performance of de novo pathway enrichment. *NPJ systems biology and applications*, *3*(1), 1–8.

Beisser, D., Klau, G. W., Dandekar, T., Müller, T., & Dittrich, M. T. (2010). Bionet: an r-package for the functional analysis of biological networks. *Bioinformatics*, *26*(8), 1129–1130.

Bell, D. C., Atkinson, J. S., & Carlson, J. W. (1999). Centrality measures for disease transmission networks. *Social networks*, *21*(1), 1–21.

Bertrand, D., Chng, K. R., Sherbaf, F. G., Kiesel, A., Chia, B. K., Sia, Y. Y., ... others (2015). Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic acids research*, *43*(7), e44–e44.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Bossi, A., & Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Molecular systems biology*, *5*(1), 260.

Brady, G., MacArthur, G., & Farrell, P. (2008). Epstein–barr virus and burkitt lymphoma. *Postgraduate medical journal*, *84*(993), 372–377.

Brandão, M., Bruzzone, M., Franzoi, M.-A., De Angelis, C., Eiger, D., Caparica, R., ... others (2021). Impact of hiv infection on baseline characteristics and survival of women with breast cancer. *AIDS*, *35*(4), 605–618.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, *25*(2), 163–177.

Braunstein, A., Dall'Asta, L., Semerjian, G., & Zdeborová, L. (2016). Network dismantling. *Proceedings of the National Academy of Sciences*, *113*(44), 12368–12373.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., ... others (2003). Arrayexpress—a public repository for microarray gene expression data at the ebi. *Nucleic acids research*, *31*(1), 68–71.

Butti, R., Das, S., Gunasekaran, V. P., Yadav, A. S., Kumar, D., & Kundu, G. C. (2018). Receptor tyrosine kinases (rtks) in breast cancer: signaling, therapeutic implications and challenges. *Molecular cancer*, *17*(1), 1–18.

Cadenas, C., van de Sandt, L., Edlund, K., Lohr, M., Hellwig, B., Marchan, R., ... Hengstler, J. G. (2014). Loss of circadian clock gene expression is associated with tumor progression in breast cancer. *Cell Cycle*, *13*(20), 3282–3291.

Cámara, P. G. (2017). Topological methods for genomics: present and future directions. *Current opinion in systems biology*, *1*, 95–101.

Cao, M., Zhang, H., Park, J., Daniels, N. M., Crovella, M. E., Cowen, L. J., & Hescott, B. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PloS one*, *8*(10), e76339.

Cheng, F., Zhao, J., & Zhao, Z. (2016). Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Briefings in bioinformatics*, *17*(4), 642–656.

Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., & Lin, C.-Y. (2014). cytohubba: identifying hub objects and sub-networks from complex interactome. *BMC systems biology*, *8*(4), 1–7.

Chisholm, R. L., Gaudet, P., Just, E. M., Pilcher, K. E., Fey, P., Merchant, S. N., & Kibbe, W. A. (2006). dictybase, the model organism database for dictyostelium discoideum. *Nucleic acids research*, *34*(suppl_1), D423–D427.

Cho, A., Shim, J. E., Kim, E., Supek, F., Lehner, B., & Lee, I. (2016). Muffinn: cancer gene discovery via network analysis of somatic mutation data. *Genome biology*, *17*(1), 1–16.

Ciriello, G., Cerami, E., Sander, C., & Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, *22*(2), 398–406.

Comes, N., Serrano-Albarras, A., Capera, J., Serrano-Novillo, C., Condom, E., y Cajal, S. R., ... Felipe, A. (2015). Involvement of potassium channels in the progression of cancer to a more malignant phenotype. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1848*(10), 2477–2492.

Consortium, G., et al. (2015). The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660.

De Brachène, A. C., & Demoulin, J.-B. (2016). Foxo transcription factors in cancer development and therapy. *Cellular and molecular life sciences*, *73*(6), 1159–1172.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). David: database for annotation, visualization, and integrated discovery. *Genome biology*, *4*(9), 1–11.

Dickson, D. (1999). *Wellcome funds cancer database.* Nature Publishing Group.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., ... others (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, *455*(7216), 1069–1075.

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., & Müller, T. (2008).

Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, *24*(13), i223–i231.

Eisenberg, E., & Levanon, E. Y. (2003). Preferential attachment in the protein network evolution. *Physical review letters*, *91*(13), 138701.

Fang, C.-B., Wu, H.-T., Zhang, M.-L., Liu, J., & Zhang, G.-J. (2020). Fanconi anemia pathway: mechanisms of breast cancer predisposition development and potential therapeutic targets. *Frontiers in cell and developmental biology*, *8*, 160.

Feige, U., Hajiaghayi, M., & Lee, J. R. (2008). Improved approximation algorithms for minimum weight vertex separators. *SIAM Journal on Computing*, *38*(2), 629–657.

Fernández, M., Semela, D., Bruix, J., Colle, I., Pinzani, M., & Bosch, J. (2009). Angiogenesis in liver disease. *Journal of hepatology*, *50*(3), 604–620.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, *23*(2), 298–305.

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., ... others (2010). Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, *39*(suppl_1), D945–D950.

Frainay, C., & Jourdan, F. (2017). Computational methods to identify metabolic subnetworks based on metabolomic profiles. *Briefings in bioinformatics*, *18*(1), 43–56.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., & Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, *296*(5568), 750–752.

Frattini, V., Trifonov, V., Chan, J. M., Castano, A., Lia, M., Abate, F., ... others (2013). The integrated landscape of driver genomic alterations in glioblastoma. *Nature genetics*, *45*(10), 1141–1149.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, *1*(3), 215–239.

French, S. W., Oliva, J., French, B. A., Li, J., & Bardag-Gorce, F. (2010). Alcohol, nutrition and liver cancer: role of toll-like receptor signaling. *World Journal of Gastroenterology: WJG*, *16*(11), 1344.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., ... others (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science signaling*, *6*(269), pl1–pl1.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... others (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), 1–16.

Gevaert, O., Villalobos, V., Sikic, B. I., & Plevritis, S. K. (2013). Identification of ovarian cancer driver genes by using module network integration of multi-omics data. *Interface focus*, *3*(4), 20130013.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., ... others (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, *15*(10), 1451–1455.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., ... others (2003). The international hapmap project.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821–7826.

Goymer, P. (2008, Sep 01). Why do we need hubs? *Nature Reviews Genetics*, *9*(9), 651-651. Retrieved from `https://doi.org/10.1038/nrg2450` doi: 10.1038/nrg2450

Guo, Z., Li, Y., Gong, X., Yao, C., Ma, W., Wang, D., ... others (2007). Edge-based scoring and searching method for identifying condition-responsive protein–protein interaction sub-network. *Bioinformatics*, *23*(16), 2121–2128.

Herath, N. I., Spanevello, M. D., Sabesan, S., Newton, T., Cummings, M., Duffy, S., ... Boyd, A. W. (2006). Over-expression of eph and ephrin genes in advanced ovarian cancer: ephrin gene expression correlates with shortened survival. *BMC cancer*, *6*(1), 1–7.

Holme, P., Kim, B. J., Yoon, C. N., & Han, S. K. (2002). Attack vulnerability of complex networks. *Physical review E*, *65*(5), 056109.

Hormozdiari, F., Salari, R., Bafna, V., & Sahinalp, S. C. (2010). Protein-protein interaction network evaluation for identifying potential drug targets. *Journal of Computational Biology*, *17*(5), 669–684.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., ... others (2002). The ensembl genome database project. *Nucleic acids research*, *30*(1), 38–41.

Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, *18*(suppl_1), S233–S240.

Jeh, G., & Widom, J. (2003). Scaling personalized web search. In *Proceedings of the 12th international conference on world wide web* (pp. 271–279).

Jeong, H., Mason, S. P., Barabási, A.-L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, *411*(6833), 41–42.

Jiang, X., Ying, P., Shen, Y., Miu, Y., Kong, W., Lu, T., & Wang, Q. (2021). Identification of critical functional modules and signaling pathways in osteoporosis. *Current Bioinformatics*, *16*(1), 90–97.

Jordan, I. K., Wolf, Y. I., & Koonin, E. V. (2003). No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC evolutionary biology*, *3*(1), 1–8.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., ... others (2005). Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, *33*(suppl_1), D428–D432.

Kanehisa, M., & Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *28*(1), 27–30.

Kannan, Y., Moriyama, M., Sugano, T., Yamate, J., Kuwamura, M., Kagaya, A., & Kiso, Y. (2000). Neurotrophic action of interleukin 3 and granulocyte-macrophage colony-stimulating factor on murine sympathetic neurons. *Neuroimmunomodulation*, *8*(3), 132–141.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature physics*, *6*(11), 888–893.

Komurov, K., Dursun, S., Erdin, S., & Ram, P. T. (2012). Netwalker: a contextual network analysis tool for functional genomics. *BMC genomics*, *13*(1), 1–9.

Kotlyar, M., Pastrello, C., Sheahan, N., & Jurisica, I. (2016). Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic acids research*, *44*(D1), D536–D541.

Kreitzburg, K. M., van Waardenburg, R. C., & Yoon, K. J. (2018). Sphingolipid metabolism and drug resistance in ovarian cancer. *Cancer drug resistance (Alhambra, Calif.)*, *1*, 181.

Langville, A. N., & Meyer, C. D. (2011). *Google's pagerank and beyond: The science of search engine rankings*. Princeton university press.

Leary, R. J., Lin, J. C., Cummins, J., Boca, S., Wood, L. D., Parsons, D. W., ... others (2008). Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proceedings of the National Academy of Sciences*, *105*(42), 16224–16229.

Leiserson, M. D., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., ... others (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, *47*(2), 106–114.

Lennartsson, J., & Rönnstrand, L. (2012). Stem cell factor receptor/c-kit: from basic science to clinical implications. *Physiological reviews*, *92*(4), 1619–1649.

Li, M., Li, W., Wu, F.-X., Pan, Y., & Wang, J. (2018). Identifying essential proteins based on sub-network partition and prioritization by integrating subcellular localization information. *Journal of theoretical biology*, *447*, 65–73.

Liedtke, C., & Trautwein, C. (2012). The role of tnf and fas dependent signaling in animal models of inflammatory liver injury and liver cancer. *European journal of cell biology*, *91*(6-7), 582–589.

Liu, X., Yu, X., Zack, D. J., Zhu, H., & Qian, J. (2008). Tiger: a database for tissue-specific gene expression and regulation. *BMC bioinformatics*, *9*(1), 1–7.

Love, C., Sun, Z., Jima, D., Li, G., Zhang, J., Miles, R., ... others (2012). The genetic landscape of mutations in burkitt lymphoma. *Nature genetics*, *44*(12), 1321–1325.

Lukanova, A., Lundin, E., Toniolo, P., Micheli, A., Akhmedkhanov, A., Rinaldi, S., ... others (2002). Circulating levels of insulin-like growth factor-i and risk of ovarian cancer. *International Journal of Cancer*, *101*(6), 549–554.

Luo, H., Jiang, B.-H., King, S. M., & Chen, Y. C. (2008). Inhibition of cell growth and vegf expression in ovarian cancer cells by flavonoids. *Nutrition and cancer*, *60*(6), 800–809.

Ma, H., Schadt, E. E., Kaplan, L. M., & Zhao, H. (2011). Cosine: Condition-specific sub-network identification using a global optimization method. *Bioinformatics*, *27*(9), 1290–1298.

Mabonga, L., & Kappo, A. P. (2019). Protein-protein interaction modulators: advances, successes and remaining challenges. *Biophysical reviews*, *11*(4), 559–581.

Mabuchi, S., Hisamatsu, T., & Kimura, T. (2011). Targeting mtor signaling pathway in ovarian cancer. *Current medicinal chemistry*, *18*(19), 2960–2968.

Mammadova-Bach, E., Gil-Pulido, J., Sarukhanyan, E., Burkard, P., Shityakov, S., Schonhart, C., ... others (2020). Platelet glycoprotein vi promotes metastasis through interaction with cancer cell–derived galectin-3. *Blood*, *135*(14), 1146–1160.

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., ... Apweiler, R. (2005). Pride: the proteomics identifications database. *Proteomics*, *5*(13), 3537–3545.

Mazzoletti, M., & Broggini, M. (2010). Pi3k/akt/mtor inhibitors in ovarian cancer. *Current medicinal chemistry*, *17*(36), 4433–4447.

Morone, F., & Makse, H. A. (2015). Influence maximization in complex networks through optimal percolation. *Nature*, *524*(7563), 65–68.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577–8582.

Nibbe, R. K., Koyutürk, M., & Chance, M. R. (2010). An integrative-omics approach to identify functional sub-networks in human colorectal cancer. *PLoS computational biology*, *6*(1), e1000639.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., ... others (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, *20*(17), 3045–3054.

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., ... others (2012). Protein interaction data curation: the international molecular exchange (imex) consortium. *Nature methods*, *9*(4), 345–350.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web.* (Tech. Rep.). Stanford InfoLab.

Patil, K. R., & Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the national academy of sciences*, *102*(8), 2685–2689.

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., & Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). *Bioinformatics*, *29*(21), 2757–2764.

Piraveenan, M., Prokopenko, M., & Hossain, L. (2013). Percolation centrality: Quantifying graph-theoretic impact of nodes during percolation in networks. *PloS one*, *8*(1), e53095.

Pothen, A., Simon, H. D., & Liou, K.-P. (1990). Partitioning sparse matrices with eigenvectors of graphs. *SIAM journal on matrix analysis and applications*, *11*(3), 430–452.

Ren, X.-L., Gleinig, N., Helbing, D., & Antulov-Fantulin, N. (2019). Generalized network dismantling. *Proceedings of the national academy of sciences*, *116*(14), 6554–6559.

Requião da Cunha, B., González-Avella, J. C., & Gonçalves, S. (2015). Fast fragmentation of networks using module-based attacks. *PloS one*, *10*(11), e0142824.

Reyna, M. A., Leiserson, M. D., & Raphael, B. J. (2018). Hierarchical hotnet:

identifying hierarchies of altered subnetworks. *Bioinformatics*, *34*(17), i972–i980.

Ruhnau, B. (2000). Eigenvector-centrality—a node-centrality? *Social networks*, *22*(4), 357–365.

Saberi, M., Khosrowabadi, R., Khatibi, A., Misic, B., & Jafari, G. (2021). Topological impact of negative links on the stability of resting-state brain network. *Scientific reports*, *11*(1), 1–14.

Saeed, R., & Deane, C. M. (2006). Protein protein interactions, evolutionary rate, abundance and age. *BMC bioinformatics*, *7*(1), 1–13.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498–2504.

Sheppard, K. E., Cullinane, C., Hannan, K. M., Wall, M., Chan, J., Barber, F., … others (2013). Synergistic inhibition of ovarian cancer cell growth by combining selective pi3k/mtor and ras/erk pathway inhibitors. *European journal of cancer*, *49*(18), 3936–3944.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). Biomart–biological queries made easy. *BMC genomics*, *10*(1), 1–12.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., & Spieth, J. (2001). Wormbase: network access to the genome and biology of caenorhabditis elegans. *Nucleic acids research*, *29*(1), 82–86.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719–724.

Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M. P., Vivancos, A., Rovira, A., … others (2018). Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome medicine*, *10*(1), 1–8.

Tian, L., Bashan, A., Shi, D.-N., & Liu, Y.-Y. (2017). Articulation points in complex networks. *Nature communications*, *8*(1), 1–9.

Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C., Chayes, J., Borgs, C., … Fraenkel, E. (2013). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *Journal of computational biology*, *20*(2), 124–136.

Uhlén, M., Björling, E., Agaton, C., Szigyarto, C. A.-K., Amini, B., Andersen, E., … others (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & cellular proteomics*, *4*(12), 1920–1932.

van den Heuvel, M. P., & Sporns, O. (2013). Network hubs in the human brain. *Trends in cognitive sciences*, *17*(12), 683–696.

Vandin, F., Upfal, E., & Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, *18*(3), 507–522.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., ... Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, *26*(12), i237–i245.

Wachi, S., Yoneda, K., & Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, *21*(23), 4205–4208.

Wandelt, S., Sun, X., Feng, D., Zanin, M., & Havlin, S. (2018). A comparative analysis of approaches to network-dismantling. *Scientific reports*, *8*(1), 1–15.

Wang, Z., Zhao, Y., Xi, J., & Du, C. (2016). Fast ranking influential nodes in complex networks using a k-shell iteration factor. *Physica A: Statistical Mechanics and its Applications*, *461*, 171–181.

Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., ... others (2002). Gramene: a resource for comparative grass genomics. *Nucleic acids research*, *30*(1), 103–105.

Weinberg, R., & Hanahan, D. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., ... Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, *45*(10), 1113–1120.

Wong, A. K., Krishnan, A., & Troyanskaya, O. G. (2018). Giant 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic acids research*, *46*(W1), W65–W70.

Xu, J., & Li, Y. (2006). Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, *22*(22), 2800–2805.

Yang, Y., & Yee, D. (2012). Targeting insulin and insulin-like growth factor signaling in breast cancer. *Journal of mammary gland biology and neoplasia*, *17*(3), 251–261.

Zdeborová, L., Zhang, P., & Zhou, H.-J. (2016). Fast and simple decycling and dismantling of networks. *Scientific reports*, *6*(1), 1–6.

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., ... others (2011). International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database*, *2011*.

Zhang, Y.-L., Wang, R.-C., Cheng, K., Ring, B. Z., & Su, L. (2017). Roles of rap1 signaling in tumor cell migration and invasion. *Cancer biology & medicine*, *14*(1), 90.

Zhao, X., Liu, F., Wang, J., Li, T., et al. (2017). Evaluating influential nodes in social networks by local centrality with a coefficient. *ISPRS International Journal of Geo-Information*, *6*(2), 35.

Zhuang, D.-Y., Jiang, L., He, Q.-Q., Zhou, P., & Yue, T. (2015). Identification of hub subnetwork based on topological features of genes in breast cancer. *International Journal of Molecular Medicine*, *35*(3), 664–674.

APPENDIX A

CENTRALITY PAIR PLOTS OF TISSUES

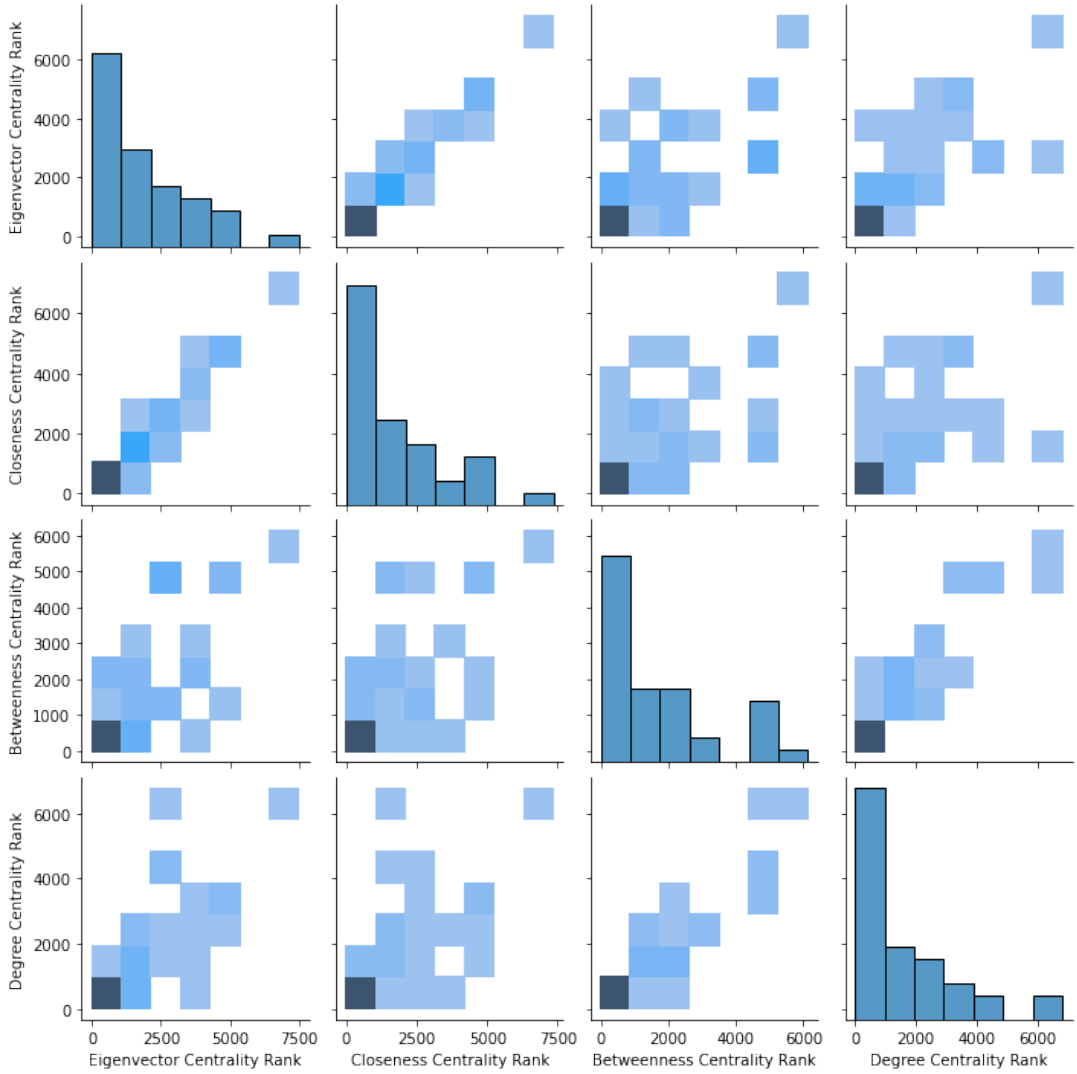


Figure A.1: Pair plots of liver TSPPI.

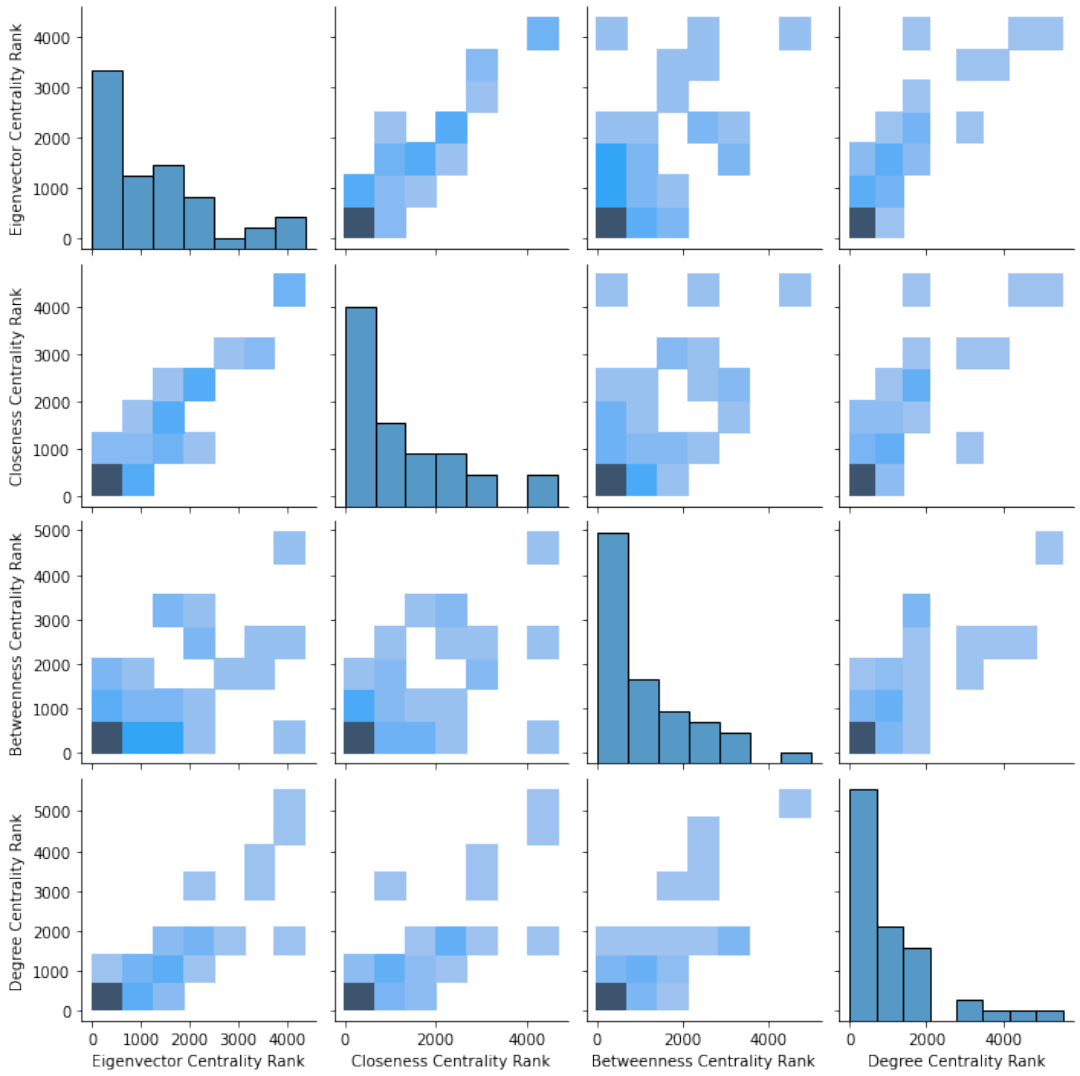


Figure A.2: Pair plots of lymph node TSPPI.

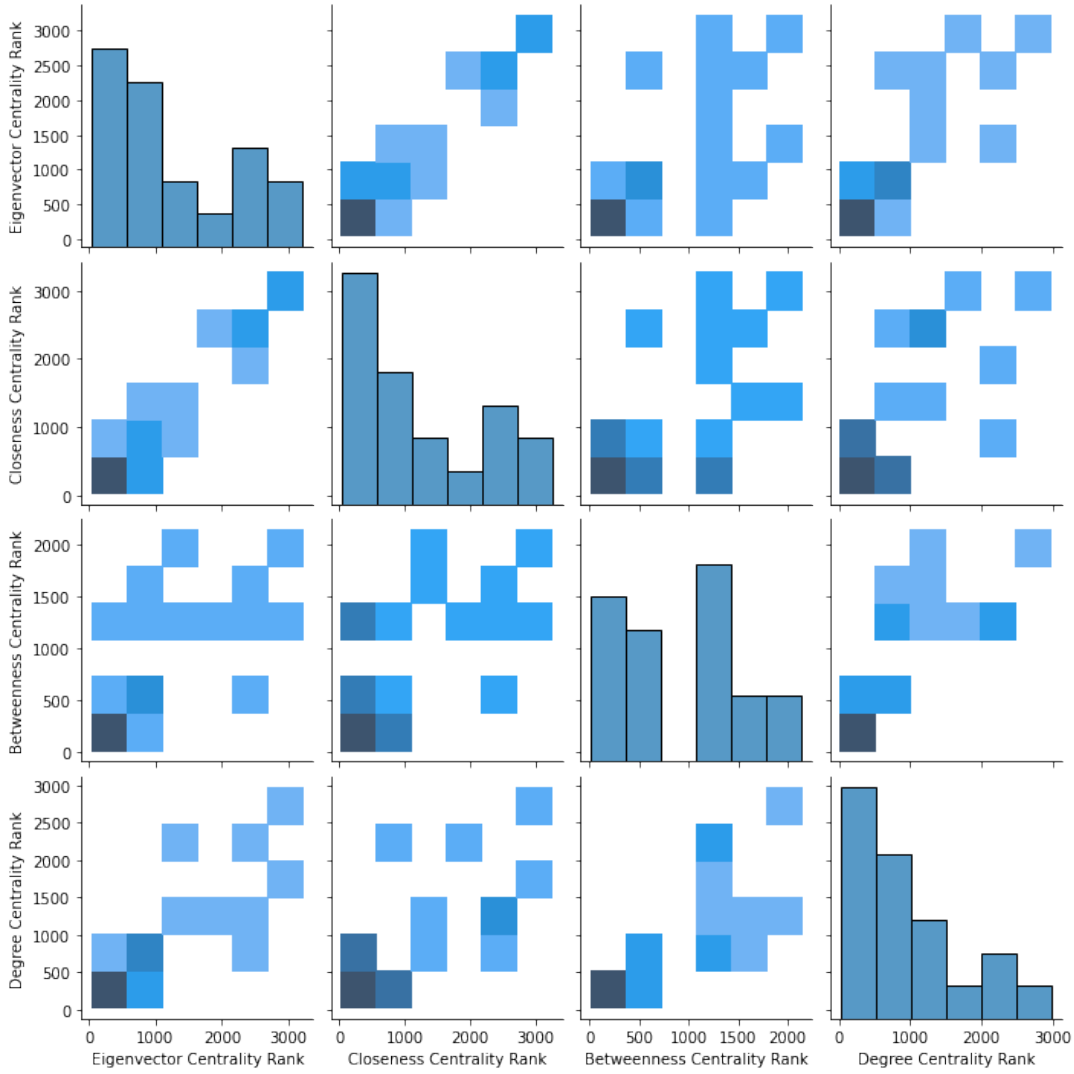


Figure A.3: Pair plots of peripheral nerve TSPPI.

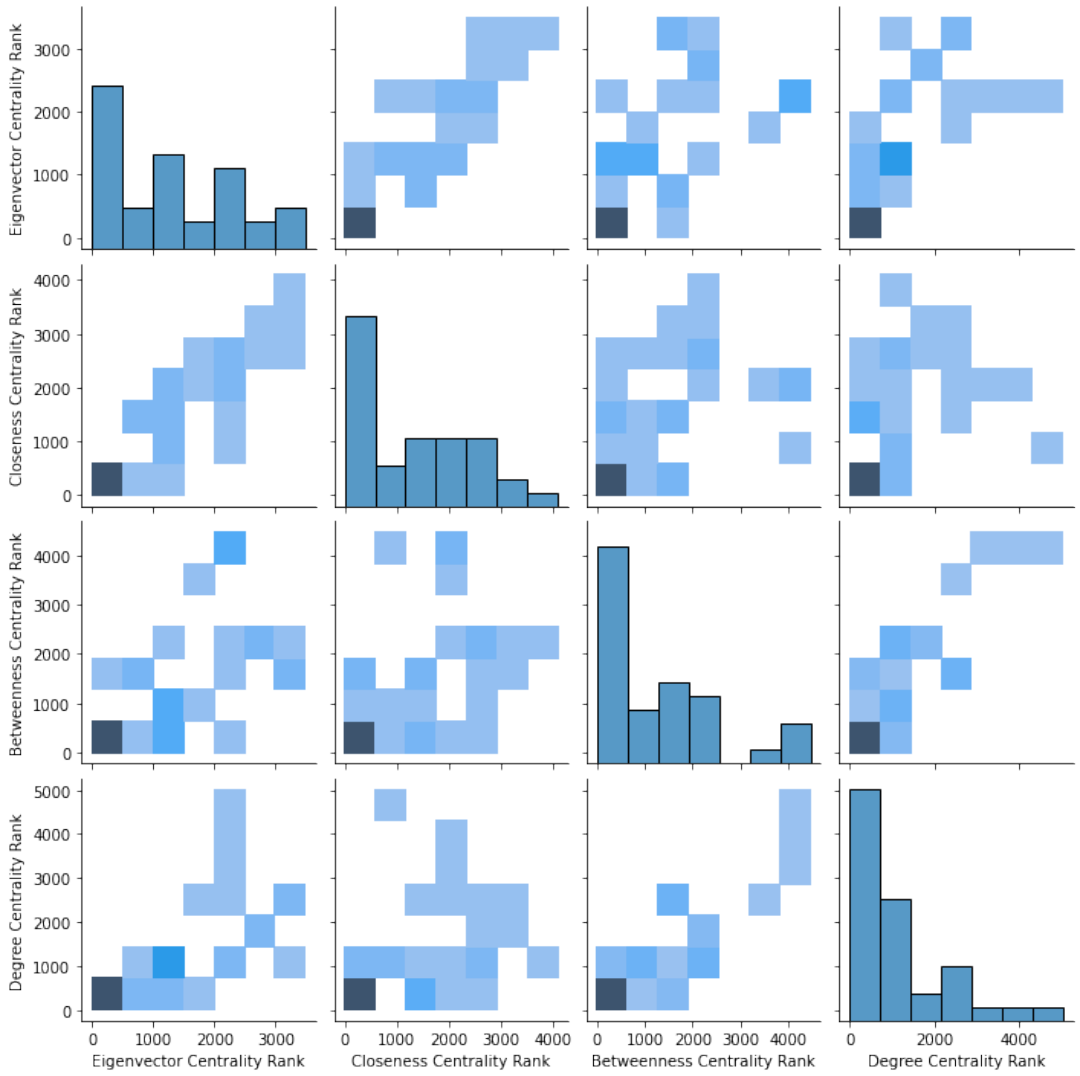


Figure A.4: Pair plots of Ovary TSPPI

APPENDIX B

RESULTS FROM CBIOPORTAL

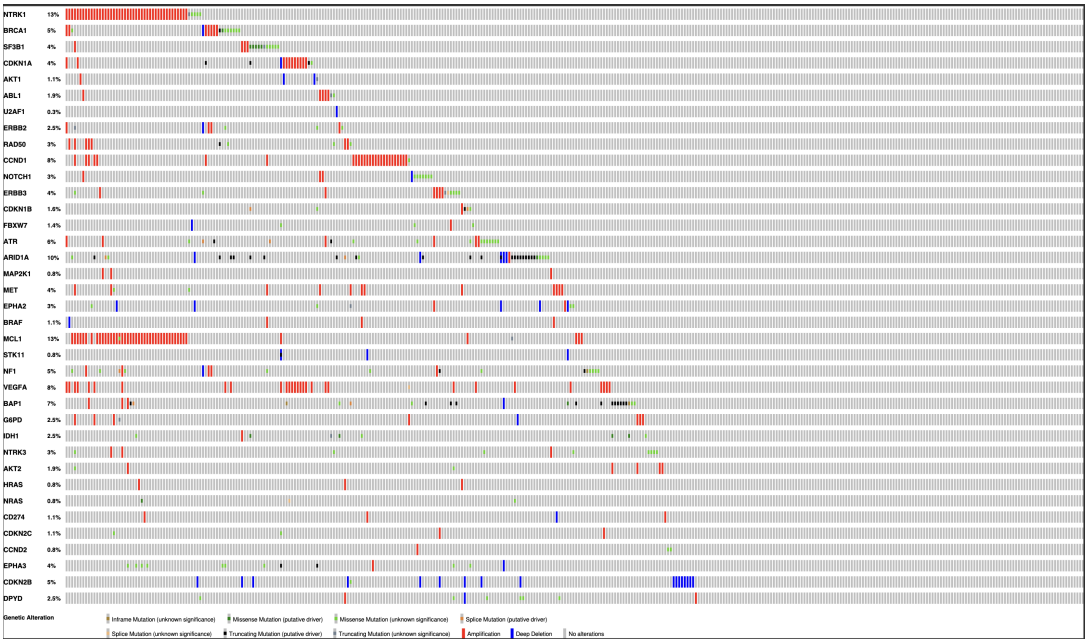


Figure B.1: cBioPortal Result of Liver TSPPI.

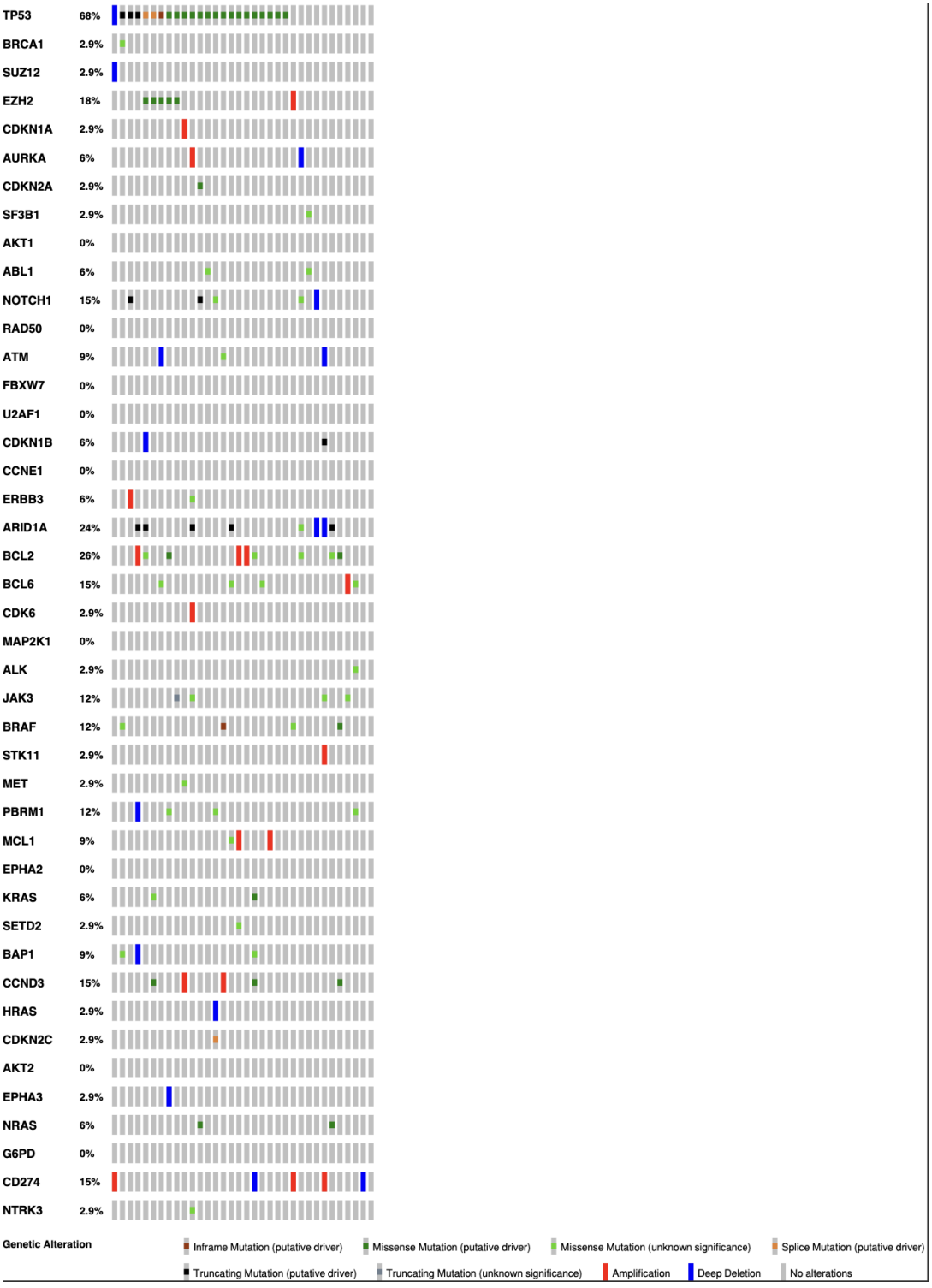


Figure B.2: cBioPortal Result of Lymph Node TSPPI.

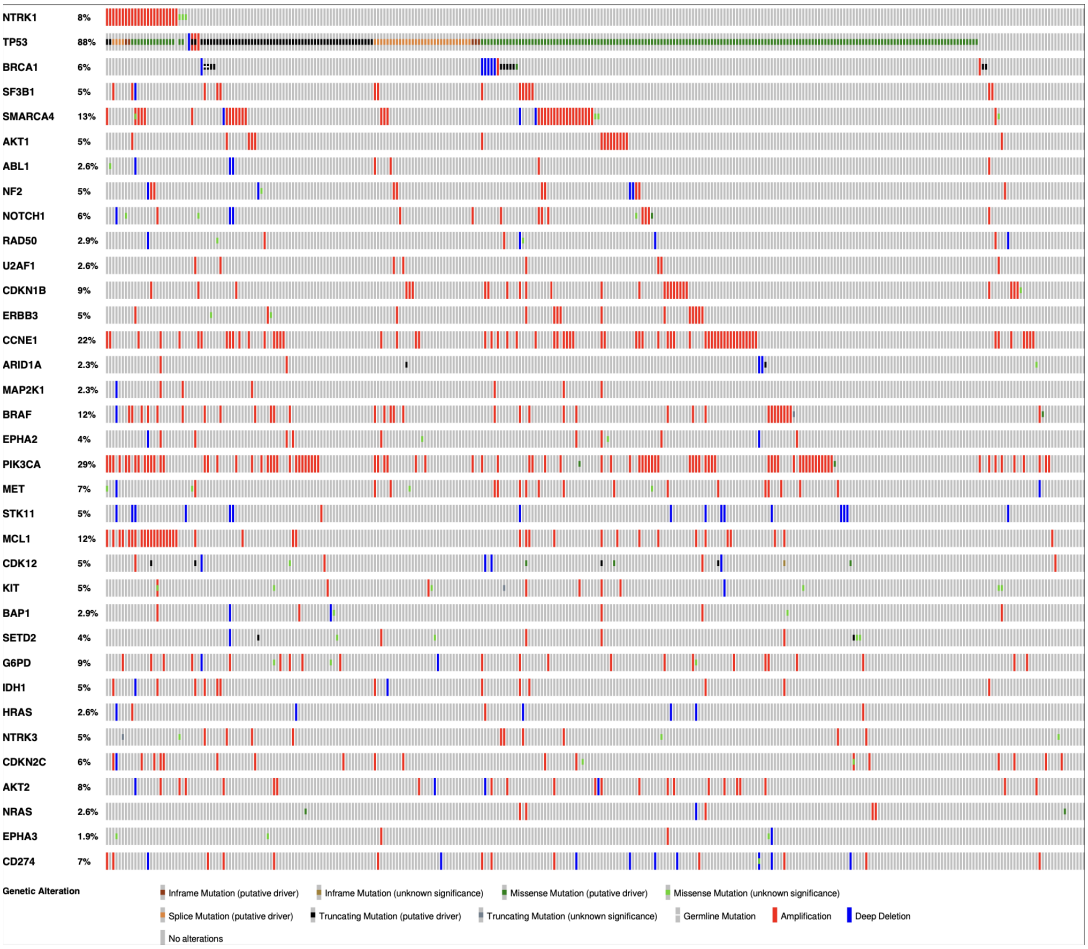


Figure B.3: cBioPortal Result of Ovary TSPPI.

Figure B.4: cBioPortal Result of Peripheral Nerve TSPPI.

APPENDIX C

NETWORK DISMANTLING RESULTS

Table C.1: Summary of Results

Tissue	Method	Largest Component Size	Number of Removed Nodes	Number of Driver Nodes in the Largest Component	Number of Remaining Driver Nodes
Breast	Betweenness Centrality	830	2190	2	15
Breast	GND with Betweenness Centrality	1052	5712	7	10
Breast	GNDR with Betweenness Centrality	1052	3797	7	11
Breast	Closeness Centrality	1898	2232	6	16
Breast	GND with Closeness Centrality	3297	3455	18	24
Breast	GNDR with Closeness Centrality	3297	2915	18	24
Breast	Degree Centrality	839	2418	0	11
Breast	GND with Degree Centrality	1053	6092	3	6
Breast	GNDR with Degree Centrality	1053	4264	4	9
Breast	Eigenvector Centrality	473	2364	1	13
Breast	GND with Eigenvector Centrality	780	6205	3	6
Breast	GNDR with Eigenvector Centrality	780	4363	3	10
Liver	Betweenness Centrality	1405	1812	4	11

Continued on next page

Table C.1 – *Continued from previous page*

Tissue	**Method**	**Largest Component Size**	**Number of Removed Nodes**	**Number of Driver Nodes in the Largest Component**	**Number of Remaining Driver Nodes**
Liver	GND with Betweenness Centrality	1613	4473	8	8
Liver	GNDR with Betweenness Centrality	1613	3425	8	9
Liver	Closeness Centrality	1590	1849	3	13
Liver	GND with Closeness Centrality	3088	2847	12	17
Liver	GNDR with Closeness Centrality	3088	2348	12	18
Liver	Degree Centrality	1694	1943	1	10
Liver	GND with Degree Centrality	1704	4596	4	4
Liver	GNDR with Degree Centrality	1704	3607	4	7
Liver	Eigenvector Centrality	2343	1877	7	12
Liver	GND with Eigenvector Centrality	2348	3560	3	6
Liver	GNDR with Eigenvector Centrality	2348	2829	10	14
Lymph Node	Betweenness Centrality	778	1932	3	6
Lymph Node	GND with Betweenness Centrality	873	5226	5	6
Lymph Node	GNDR with Betweenness Centrality	873	3382	5	7
Lymph Node	Closeness Centrality	1619	1962	5	8
Lymph Node	GND with Closeness Centrality	2626	3239	7	7
Lymph Node	GNDR with Closeness Centrality	2626	2714	7	8
Lymph Node	Degree Centrality	2303	1979	5	6

Continued on next page

Table C.1 – *Continued from previous page*

Tissue	**Method**	**Largest Component Size**	**Number of Removed Nodes**	**Number of Driver Nodes in the Largest Component**	**Number of Remaining Driver Nodes**
Lymph Node	GND with Degree Centrality	2306	4117	10	10
Lymph Node	GNDR with Degree Centrality	2306	3529	10	10
Lymph Node	Eigenvector Centrality	1353	2061	4	7
Lymph Node	GND with Eigenvector Centrality	1527	4490	10	10
Lymph Node	GND with Eigenvector Centrality	1527	3577	10	11
Peripheral Nerve	Betweenness Centrality	583	1065	1	5
Peripheral Nerve	GND with Betweenness Centrality	693	3300	1	1
Peripheral Nerve	GNDR with Betweenness Centrality	693	2051	2	4
Peripheral Nerve	Closeness Centrality	1110	1074	1	5
Peripheral Nerve	GND with Closeness Centrality	2166	1635	1	1
Peripheral Nerve	GNDR with Closeness Centrality	2166	1389	1	1
Peripheral Nerve	Degree Centrality	1315	1151	1	4
Peripheral Nerve	GND with Degree Centrality	1322	2564	0	0
Peripheral Nerve	GNDR with Degree Centrality	1322	1979	0	1
Peripheral Nerve	Eigenvector Centrality	1497	1135	5	5
Peripheral Nerve	GND with Eigenvector Centrality	1509	2420	1	1
Peripheral Nerve	GND with Eigenvector Centrality	1509	2001	1	1
Ovary	Betweenness Centrality	306	828	19	20

Continued on next page

Table C.1 – *Continued from previous page*

Tissue	**Method**	**Largest Component Size**	**Number of Removed Nodes**	**Number of Driver Nodes in the Largest Component**	**Number of Remaining Driver Nodes**
Ovary	GND with Betweenness Centrality	386	2590	7	9
Ovary	GNDR with Betweenness Centrality	386	2416	7	9
Ovary	Closeness Centrality	414	849	18	20
Ovary	GND with Closeness Centrality	469	1863	9	11
Ovary	GNDR with Closeness Centrality	469	1798	9	11
Ovary	Degree Centrality	617	962	11	11
Ovary	GND with Degree Centrality	676	2485	11	11
Ovary	GNDR with Degree Centrality	676	2331	11	11
Ovary	Eigenvector Centrality	204	913	17	19
Ovary	GND with Eigenvector Centrality	273	2687	7	8
Ovary	GNDR with Eigenvector Centrality	273	2480	7	8