T.C.

ANTALYA BİLİM UNIVERSITY

INSTITUTE OF POSTGRADUATE EDUCATION

ELECTRICAL AND COMPUTER ENGINEERING

THESIS PROGRAM

RANKING CANCER DRIVERS VIA BETWEENNESS-BASED OUTLIER DETECTION AND RANDOM WALKS

DISSERTATION

Prepared By

Aissa HOUDJEDJ

ANTALYA–2021

T.C.

ANTALYA BİLİM UNIVERSITY

INSTITUTE OF POSTGRADUATE EDUCATION

ELECTRICAL AND COMPUTER ENGINEERING

THESIS PROGRAM

RANKING CANCER DRIVERS VIA BETWEENNESS-BASED OUTLIER DETECTION AND RANDOM WALKS

DISSERTATION

Prepared By

Aissa HOUDJEDJ

Dissertation Advisors

Prof. Dr. Cesim Erten

Doç. Dr. Hilal Kazan

ANTALYA–2021

APPROVAL/NOTIFICATION FORM

ANTALYA BILIM UNIVERSITY

INSTITUTE OF POST-GRADUATE EDUCATION

Aissa HOUDJEDJ, a M.Sc. student of Antalya Bılım University, Institute of Post Graduate Education, Electrical and Computer Engineering owning student ID 171222003, successfully defended the thesis/dissertation entitled ”Ranking Cancer Drivers via Betweenness-based Outlier Detection and Random Walks”, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Academic Tittle, Name-Surname, Signature

Jury Member (Chairman): Prof. Dr. Cesim Erten (Advisor) ,.....

Jury Member: Doç. Dr. Hilal Kazan (Co-advisor) ,.....

Jury Member: Prof. Dr. Melih Güney ,.....

Jury Member: Asst. Prof. Shahram Taheri ,.....

Jury Member: Asst. Prof. Aslı Bay ,.....

Thesis Submission Date : 25 / 12 / 2020

Thesis Defence Exam Date : 22 / 01 / 2021

Director of The Institute: ,.....

ÖZET

KANSER SÜRÜCÜ GENLERININ ARASINDALIK BAZLI AYKIRILIK TANIMI VE RASTGELE YÜRÜYÜŞLE TESPITI

Son yıllardaki kanser genom çalışmaları yüksek sayıda kanser genomu için detaylı moleküler veri üretmiştir. Ortaya çıkan önemli bir problem bu verileri kullanarak kanser sürücü genleri tespit etmektir.

Bu tezde, genomik veriyi protein-protein etkileşim ağlarıyla birleştirerek kanser sürücü genleri tespit eden BetweenNet isimli işlemsel bir yöntem önerilmektedir. BetweenNet, hastaya özgü oluşturulmuş ağları arasındalık merkeziliği tabanlı bir metrik ile inceleyerek etkinliği değişmiş aykırı genleri bulmaktadır. Mutasyona uğramış genlerle etkinliği değişmiş genlerin arasındaki ilişkileri iki parçalı bir çizgede tanımlayıp, bu çizgede rastgele yürüyüş algoritması uygulayarak mutasyona uğramış genleri sürücülük potansiyeline göre sıralamaktadır. BetweenNet yöntemi varolan benzer yöntemlerle akciğer, meme ve pan-kanser verileri kullanılarak karşılaştırılmıştır. Değerlendirmelerimiz BetweenNet'in bilinen kanser genlerini bulmada daha başarılı olduğunu göstermektedir. Ayrıca, bilinen kanser genleriyle BetweenNet tarafından sıralanan genlerin Gene Ontology terimleri ve referans ağlar bakımından birbiriyle önemli derecede örtüştüğü tespit edilmiştir.

Anahtar sözcükler: Sürücü genleri, iki parçalı çizge, arasındalık merkeziliği, ağ difüzyonu, Protein Protein Etkilesim.

ABSTRACT

RANKING CANCER DRIVERS VIA BETWEENNESS-BASED OUTLIER DETECTION AND RANDOM WALKS

Recent cancer genomic studies have generated detailed molecular data on a large number of cancer patients. A key remaining problem in cancer genomics is the identification of driver genes.

We propose BetweenNet, a computational approach that integrates genomic data with a protein-protein interaction network to identify cancer driver genes. BetweenNet utilizes a measure based on betweenness centrality on patient specific networks to identify the so-called *outlier genes* that correspond to dysregulated genes for each patient. Setting up the relationship between the mutated genes and the outliers through a bipartite graph, it employs a random-walk process on the graph, which provides the final prioritization of the mutated genes. We compare BetweenNet against state-of-the art cancer gene prioritization methods on lung, breast, and pan-cancer datasets.

Our evaluations show that BetweenNet is better at recovering known cancer genes based on multiple reference databases. Additionally, we show that the Gene Ontology terms and the reference pathways enriched in BetweenNet ranked genes and those that are enriched in known cancer genes overlap significantly when compared to the overlaps achieved by the rankings of the alternative methods.

Keywords: Driver genes, bipartite graph, betweenness centrality, network diffusion, Protein-Protein Interaction.

DEDICATION AND ACKNOWLEDGMENT

I would like to express my sincere gratitude to my advisors, Prof. Dr. Cesim Erten, and Assoc. Prof. Dr. Hilal Kazan for their support, and endless guidance, without which I could not have completed this dissertation. I am extremely grateful for the knowledge and experience. They were very patient with me, and they spent a huge time discussing the various subjects of this thesis.

I am also grateful for all the team in the lab, Rafsan Ahmed, Ilyes Baali, Yacine Marouf, and Ahmed Amine Taleb Bahmed, for their help, support, and all the good time we spent together.

My deep and sincere gratitude to the best persons in my life, my mother, my father, my brother, and my sister. I am grateful to my father for always being there for me as a friend. I am forever indebted to you my family for giving me the opportunities and experiences that have made me who I am. I am dedicating this work to you and I hope I can always make you very proud and happy.

Finally, Words cannot describe how I should thank you Ahlam, my wonderful wife, she has encouraged and helped me. Thank you for standing by me despite all the circumstances. I hope this dissertation makes you proud!.

INDEX

PREFACE

I hereby declare that this master's thesis titled "Ranking Cancer Drivers via Betweenness-based Outlier Detection and Random Walks" has been written by myself under the academic rules and ethical conduct of the Antalya Bilim University. I also declare that the work attached to this declaration complies with the university requirements and is my work. I also declare that all materials used in this thesis consist of the mentioned resources in the reference list. I verify all these with my honor.

25/12/2020

Aissa HOUDJEDJ

List of Figures

List of Table

ABBREVIATIONS

CGC Cancer Gene Census

GO Gene Ontology

KEGG Kyoto Encyclopedia of Genes and Genomes

NCG Network of Cancer Gene

PPI Protein-Protein Interaction

SNV Single Nucleotide Variant

TCGA The Cancer Genome Atlas

CHAPTER 1

1. Introduction

Cancer is a complex disease arising in many cases from the effects of multiple genetic changes that give rise to pathway dysregulation through alterations in copy number, DNA methylation, gene expression, and molecular function [3, 4]. Recent cancer genomics projects such as The Cancer Genome Atlas The Cancer Genome Atlas (TCGA) have created a comprehensive catalog of somatic mutations across all major cancer types. A key current challenge in cancer genomics is to distinguish driver mutations that are causal for cancer progression from passenger mutations that do not confer any selective advantage. Consequently, several computational methods have been proposed for the identification of cancer driver genes or driver modules of genes by integrating mutations data with various other types of genetic data [5, 6, 7, 8, 9, 10, 11, 12]; see [13, 14, 15, 16] for recent comprehensive evaluations and surveys on the topic.

Rather than outputting a set of candidate driver genes or modules, a subclass of cancer driver identification methods output a prioritized list of genes ranked by their cancer-driving potential. Early approaches in this group have utilized the mutation frequency of each gene by comparing it with background mutation rates [17, 18, 19]. However, with a careful review of the existing cancer catalogues it is easy to observe that most tumors share only a small portion of the set of all mutated genes, giving rise to the so called *tumor heterogeneity problem*; methods solely based on mutation rates suffer from low sensitivity due to the existence of long-tail of infrequently mutated genes [6, 20].

One strategy that aims to tackle the long-tail phenomenon is to move from a mutation-centric point of view to a *guilts by association* viewpoint where a correlation between

differentially expressed genes and mutated genes are sought. This strategy assumes that even though different sets of genes are mutated in different patients, each of the candidate driver mutations tend to affect a large number of differentially expressed genes. Masica and Karchin present one of the early models based on such a strategy by employing statistical methods for setting up the correlation between mutated genes and the differentially expressed genes to identify candidate drivers [3]. Many different models follow a similar trail by further incorporating biological pathway/network information for setting up such a correlation [8, 2, 21, 22, 23, 24].

1.1 Contributions

We provide an overview of the research contributions in this thesis towards the problem of driver genes prioritization in cancer. First, we propose BetweenNet algorithm for cancer driver gene prioritization. However different from other methods, it determines outlier genes based on the betweenness centrality values of the genes in personalized networks. The second contribution of BetweenNet is the employment of a random-walk process on the resulting influence bipartite graph. Random-walks have been utilized in this context previously [22, 24]. However, our application of random walk with restart on the whole influence graph is quite different from the two-step or three-step employment of the diffusion process on a per patient basis used in the alternative methods. Lastly, we provide extensive evaluations to confirm that BetweenNet outperforms the alternative methods in recovering known reference genes and in providing functionally coherent rankings when compared to the enriched GO terms or the enriched known functional pathways.

1.2 Thesis Organization

The thesis is organized as follows:

- Chapter 2, is a biological and computational background review. We review cancer biology, and briefly discuss types of somatic mutations. Also, we review methods related to driver genes identification in cancer.
- In Chapter 3, we provide a computational problem definition to model this biological phenomenon. We discuss the computational complexity of the algorithm, then present the details of the proposed efficient algorithm.
- In Chapter 4, we present the results of our algorithm in comparison to state-of-the-art algorithms and discuss the key biological insights of the results.
- Finally, in Chapter 5, we summarize the thesis and discuss future research directions.

This thesis is based on a published preprint paper submitted to *bioRxiv* [25].

CHAPTER 2

2. Background

2.1 Biological Background

2.1.1 Cancer

Cancer is a complex disease in which cells start to grow out of control and cripple normal cells. The mechanism of cell division in unicellular organisms is reproduction; and maintenance in multicellular organisms, it is done by receiving signals instructing them to divide, differentiate or die. However, cancerous cells are not able to stop dividing and die, due to a lack of components that instruct them. The spread of cancerous cells to other vital organs in a process known as *metastasis*. Some cancers cause a faster cell growth, while other cancers lead to a slower cell divide and growth. Since malicious growth can occur in virtually all locations of the body, there are over 100 different types of cancers. [26] states that the Conventional Molecular Networks that are shared among mammalian cells are controlling the reproduction, differentiation, and cell death. The transformation of healthy tissues to tumor cells is caused by the mutations that target genes in the Conventional Molecular Networks.

2.1.2 Types of Mutations in Cancer

Genes control the development of an organ. Mutations can affect the structure of an encoded protein or can lead to a change in its expression. Where every transformation in the DNA sequence will affect all copies of that encoded protein. However, a mutation can

be particularly damaging to a cell or organism. Cancer arises as a result of somatically acquired changes in the DNA of cancer cells. However, not all the somatic mutations are involved in cancer arise, some mutations have no contribution at all.

Casually, a driver mutation is involved in the development of a cancer cell, and it is highly selected in the microenvironment of the tissue in which cancer arises. Where a driver mutation is leading to the damage of the cell and it is highly not required for maintenance of final cancer but it must have been selected at some point along the lineage of cancer development.

The mutation that has not been selected and not contributed to cancer development is called a passenger mutation. These mutations can be found within the cancer genome because some mutations occur without functional consequences that often occur during cell division.

Mutations in the genome occur in different forms, from single nucleotide substitution to complete chromosome changes. These variations can be classified into three categories, single nucleotide variants Single Nucleotide Variant (SNV), copy number aberrations (CNA), and structural variants [27].

The main objective is the identification of driver mutations among the passenger mutations that appear in the same cancer genome.

2.1.3 Single Nucleotide Variants

SNV mutations may occur, due to a substation of a single base in the DNA sequence (Figure 2.1-A), it can be common in a population and rare for other populations. If the variant is present in at least 1% of the population, the mutations are named single nucleotide polymorphisms (SNPs). Indels are the insertion or deletion of a small segment of bases from the genome (Figure 2.1-B).

At the coding region, SNVs that occur could result in either $synonymous$ or $nonsynonymous$. Synonymous mutation is a substitution that does not make any changes in amino acid that do not affect the protein structure is named $Synonymous$

mutation, however, mutations that affect the protein function and structure are named $nonsynonymous$ mutations. $Nonsynonymous$ mutations are categorized into two types $missense$ and $nonsense$. $Missens$e is the substitution that changes the encoded amino acid and resulting in different protein structures. On the other hand, $nonsense$ mutation changes the encoded amino acid to a stop codon, thereby terminating protein synthesis prematurely. But not all mutations can affect the proteins directly, such as the mutation that occurs at the genome's noncoding regions affect gene regulation mechanisms. In our project we used only SNVs, more about data processing is detailed in $Chapter3$.

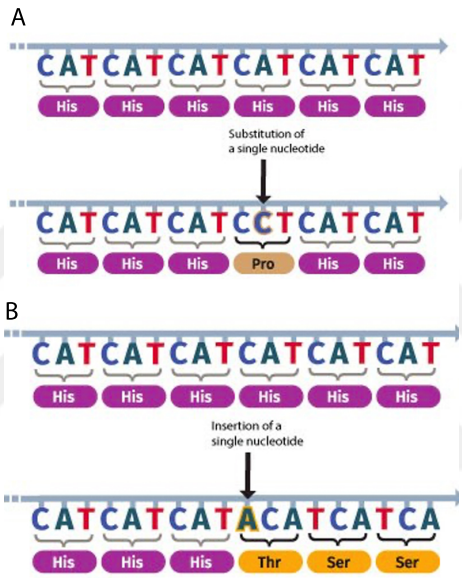

Figure 2.1: Illustration of Single Nucleotide Variants (SNVs), where A) a single nucleotide substitution B) Illustration of Indels that could produce incorrect amino acid sequence. [1])

2.1.4 Protein-protein interactions

Protein-protein interactions Protein-Protein Interaction (PPI) are integral to understanding the useful interactions and connections between proteins. The development of drugs and therapies are based on the identification of protein interactions [28], because PPIs play a critical role in cellular functions and biological processes in all organisms. These PPIs are constructed using many physiochemical and intensive computational experimental techniques. Nodes in PPIs correspond to proteins, and interactions between them are described by edges. As a result of PPIs identified a small number of proteins, there is a

huge need for new techniques to predict the non-discovered PPIs and validate the existing and experimental results.

2.1.5 Biological pathways

Thousands of molecules are part of cells, most of these molecules are proteins and small molecule compounds that work and interact together to perform some cellular tasks such as responding to the outer environment. Cells receive chemical cues from either inside or outside the body prompted by stress for example. As a reaction to these cues, cells are sending and receiving signals as well through the biological pathways. However, molecules that make up the biological signals (such as proteins) are interacting with each other as well as with signals.

A biological pathway is a set of molecular events that ends in the creation of a new molecular component or change in a cellular state, and it is a sequence of interaction between molecules in a cell.

Biological pathways can take actions over short or long distances, where some cells are sending signals to a nearby cell to produce, or repair any damage. And there are some other cells that produce substances that travel through the blood to distant cells. However, biological pathways sometimes are not performing properly, which causes many diseases such as cancer or diabetes. Biological pathways are classified into many types, the most known types are involved in metabolism, gene expression, gene regulation, and signal transduction.

2.1.5.1 Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG [29] projects aim to join both genomic information and higher-order functional information together, it consists of multiple manually curated databases (15 different databases) and a computationally generated database by linking genes in the genome with a network of interacting molecules in the cell.

KEGG is defined by three databases, PATHWAY, GENES, and LIGAND [30], which

forms the reference knowledge base to understand higher-level systemic functions of the cell and the organism, including metabolism, other cellular processes, organismal functions, and human diseases, where PATHWAY represents the higher-order functions in terms of the network of interacting molecules, GENES is the collection of gene catalogs for sequenced genomes, and LIGAND represents the chemical compounds in the cell, and enzyme molecules and its reactions.

The architecture of the KEGG database is similar to the version defined in their previous version [29] , where it gives the chance for users to enter the system in both ways, top-down that starts from functional (pathway) or bottom-up that start from genomic information.

2.1.5.2 Reactome

Reactome [31] is a project that also consists of manually curated databases, aims to provide bioinformatics and computational tools to visualize, interpret, and analyze pathway knowledge. As biological information became so complex in recent years, a huge need to develop databases to validate and interpret biological studies. Reactome is a free open-source database, consists of relational signaling and metabolic molecules and their relations that are organized into biological pathways and processes. The reaction is the main unit in Reactome data. Where, many entities participating in reactions form a network of biological interactions such as (nucleic acids, proteins, complexes, vaccines…etc) and are grouped into pathways. There are many biological pathways in Reactome, include classical intermediary metabolism, signaling, transcriptional regulation, apoptosis, and disease.

The process of collecting a Reactome pathway is quite the same as editing a scientific review. In which experts from different domains provide their expertise, then a curator formalizes it into the database structure, afterward, another expert from another domain will review the representation. Then, an evidence system keeps tracking and ensures that the assertions are backed up by the primary literature.

Finally, the Reactome database is designed to validate and interpret the results of experimental studies and research, and it is used by bioinformaticians to develop and validate their algorithms to mine knowledge from genomic information.

2.1.6 Gene Ontology (GO)

The Gene Ontology (GO) is a project that aims to build and use ontologies to map genes and their products to a biological annotation in bioinformatic centers [32, 33, 34] and organism databases using computational algorithms that extract the relationship from scientific literature

A GO annotation term is a link that describes how a gene product type is related to a molecular function, and biological process, in which it describes the capabilities and contributions of a gene product. The GO Ontologies described attributes of genes and gene products by categorizing them into three key domains, molecular function, biological process, and cellular component using ontology [35, 36, 37, 38, 39].

GO structure is described in terms of a graph, consists of GO terms as nodes and relationships are defined by edges. Also, GO child terms are more specific and specialized than parent terms, and a child can have more than one parent term.

The GO annotations are used to validate and analyze functionalities from very large datasets. While it is used to analyze results from high-throughput studies, where they provide a set of genes, using GO annotations, researchers are able to validate their results, determine which function is significantly over- or under-represented in their results.

2.1.7 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) project [40] aims to discover and catalog mutations that lead to cancer arise, by using genome sequencing and bioinformatics tools, to discover therapies and treatments and a better understanding of the disease [41]. The structure of the TCGA database is well defined and organized in which many centers are involved in collecting and processing sample data using sophisticated bioinformatics data analyses. Many Tissue Source Sites (TSSs) are responsible for collecting tissues from patients and bring them to the Biospecimen Core Resource (BCR) to verify and process them. Afterward, the processed data is submitted to the Data Coordinating Center (DCC) to provide genomic characterization and sequencing.

TCGA data provides studies and data for 33 different cancer types from 11,328 samples, where the studied cancer is chosen based on the poor prognosis and availability of samples. The identifications of alteration among the selected samples are done by molecularly characterizing matched paired data (tumor-normal tissues). Also, the TCGA database provides molecular data from different types of analyses such as DNA sequencing, RNA sequencing, gene expressions, exon expression, miRNA expression.

The generated data are publicly and freely available to the research community, to allow scientists and researchers to access them and speed up advancements in cancer discovery.

2.2 Computational Background

2.2.1 Graph theoretical measurements

A graph is a set of nodes connected by edges, where edges between nodes can be either weighted or unweighted. Also, edges can be directed or undirected. Mathematically, many measures are presented to describe the properties of graphs, classified into two categories, a global measure that refers to global properties of a graph represented by a single number, and a nodal measure that refers to properties of the nodes in a graph, such as: degree, strength, path length, clustering coefficient, closeness centrality, betweenness centrality, global efficiency, and many other measurements.

2.2.2 Bipartite Graphs

A bipartite graph is a graph with two disjoint and independent sets U and V, in which an edge can exist only between a vertex from U to V but no edges exist between nodes of the same set U or V. A bipartite graph is a special case of a k-partite graph with k = 2. The $Figure\ 2.2$ shows a bipartite graph example, with vertices colored based on the disjoint set it belongs to.

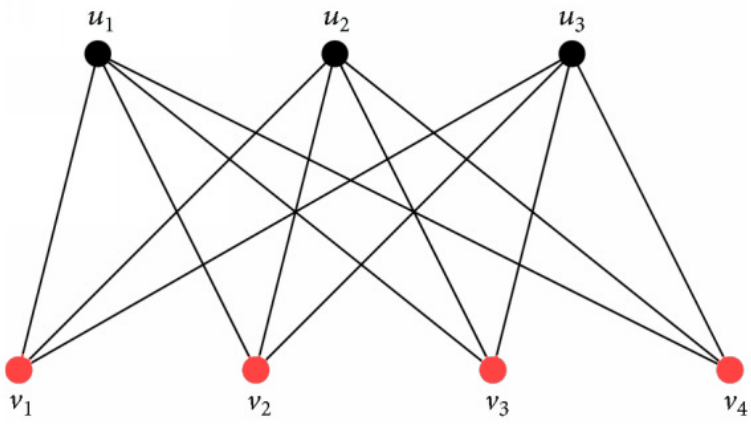

Figure 2.2: A representation of a bipartite graph.

2.2.3 Review of methods for Identifying Driver Genes

2.2.3.1 Betweenness

Betweenness [42] is one of the alternative models that is used to discover novel driver mutations. in which it measures the importance to which a node lies on paths between other nodes.

Previously, graph centralities have been employed in the context of identifying cancer genes such as Jaccard index, degree centrality, and graph-theoretical distances.Dopazo and Erten; used mutations data, PPI network, and employed various graph centralities measurements to discover cancer genes [8]. among all employed measurements, betweenness differentiate driver genes than the rest of methods.

Dopazo and Erten; used paired TCGA samples with both normal and tumor samples, mutation data, and PPI network H. for each sample i, a pair of graphs are created, N_i and T_i, where the normal graph N_i is a subgraph of H that consists of only expressed genes (genes with normalized count value less than 1) in the normal sample i, whereas the tumor graph T_i consist of expressed genes in the tumor sample i and non-mutated genes as well.

For each sample i, betweenness is calculated for both N_i and T_i, as follows:

$$bw_G(v) = \sum_{\forall s,t \in V, s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.1}$$

where, σ_{st} is the total number of shortest paths between nodes s,t, and σ_v is the number of such paths that go through the node v.

To rank genes, for each gene, a weight function is defined as follows:

$$W_{bw}(v) = \sum_{\forall N_i, T_i \in P} |bw_{N_i}(v) - bw_{T_i}(v)| \tag{2.2}$$

2.2.3.2 DriverNet

DriverNet [2] is one of the first methods that implement bipartite graph to prioritize driver genes, by evaluating their impact on their gene expression. This algorithm uses real-valued gene expression data to generate outliers data, by transforming a gene expression data into a binary matrix O'(i, j) that indicates whether a gene i is an outlier gene from population-level distribution for that specific gene in patient j. Also, a binary mutation data matrix $M(i, j)$ and PPI network are used as input to their algorithm. The mutation data M can be in the form of somatic point mutations, indels, copy number changes, or possibly epigenomic events.

DriverNet is formulated in a bipartite graph as shown in $Figure\ 2.3$, where the left nodes of the graph are the mutated genes from M, and nodes in the right partition represent the set of outlier genes across all samples which are multiple sets of differentially expressed genes from O', where each set represents a patient(represented as red nodes). Edges are drawn if for each patient P_k, a mutation i in the left partition is mutated in P_k, and is known to have interaction with an outlier j in the right partition.

The aim of the algorithm is to identify genes from the left partition that cover the maximum outlier genes in the right partition. So, genes are ranked based on their degree in the constructed bipartite graph. A mutated gene is selected then removed with all outliers connected to it, repeatedly until it covers all outliers in the right partition.

Many outlying genes can not be defined and explained due to the PPI network. Finally,

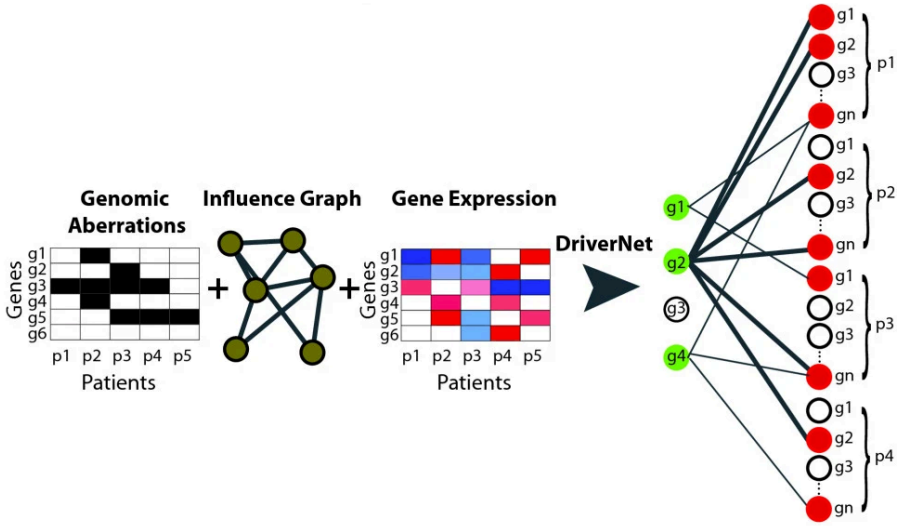

Figure 2.3: A summary of DriverNet approach. Nodes on the left partition correspond to the mutations nodes connected to the outliers on the right partition for each patient [2].

the algorithm uses a greedy algorithm to solve the optimization problem since it is almost identified as the minimum set cover problem, which is NP-hard.

2.2.3.3 Subdyquency

Subdyquency [24] method is similar to DriverNet method, based on random walk to prioritize driver genes. The algorithm framework is formulated in a bipartite graph same as DriverNet, where mutated genes on the right side correspond to all genes that being at least mutated in one patient, and the left partition corresponds to genes whose expression is significantly different with respect to the cohort. Edges are drawn similarly to DriverNet method's criteria.

Since a pair of proteins can interact only if they are in the same subcellular compartment, moreover, proteins are performing their functions if they are located in the same subcellular compartments. Following these assumptions, all edges are assigned weights based on Tang's [43] method, by measuring the number of proteins in each compartment C_x [44], this measurement denotes the importance of that compartment C_X, then normalize the value of C_X with the largest size of compartment C_m. The significance score S_C

is calculated as follows:

$$SC(i) = \frac{C_X(i)}{C_M} \tag{2.3}$$

After the normalization process, the final SC's values range between 0 and 1, where compartments with the larger size of proteins are more important than other compartments with small sizes, so, edge weight is assigned for each pair of mutated gene i and an outlier j is defined as follows:

$$Wi,j = \begin{cases} max(SC(I)), & \text{if } SLoc(i,j) \neq, \\ SC(C_N), & \text{Otherwise} \end{cases} \tag{2.4}$$

If both mutation i and outlier j interact in the same compartment, then the edge weight is equal to the maximum significance score of their shared compartments. Otherwise, the edge weight is equal to the minimum compartment significance scores among all compartments denoted by C_N. $M(i)$ is a mutation frequency of gene i and $O(j)$ is also an outlying frequency of gene j in the cohort of patients. Finally, the authors propose to simulate a random walk on the bipartite graph using the three following steps:

$$R_m(i). = a * M(i) + (1-a) * \sum_{j=1}^{m} W_{ij} * O(j) \tag{2.5}$$

$$R_o(i). = a * O(i) + (1-a) * \sum_{i=1}^{m} W_{ij} * R_m(j) \tag{2.6}$$

$$R_m(i). = a * M(i) + (1-a) * \sum_{j=1}^{m} W_{ij} * R_o(j) \tag{2.7}$$

Ranking genes is based on calculating scores that are derived from the summation of R_pm vectors among patients and rank them based on the final scores.

2.2.3.4 DawnRank

DawnRank [21] is also a random walk based method, that implements Google's PageRank approach [45]. If a mutated M has a large direct or indirect connectivity to differentially expressed genes, then that gene has a higher impact score. The idea behind driver genes tends to have a high degree in gene networks [46, 47] recommends that PageRank would be significant to prioritize driver genes. DawnRank uses a directed adjacency matrix and the absolute differential expression vector. To further differentiate gene-gene interactions, the binary adjacency matrix is expanded to construct edge weights. The rank of each gene is defined iteratively as follows:

$$r_j^{t+1} = (1 - d_j) f_j + d_j \sum_{i+1}^{n} \frac{A_{ji} r_i^t}{deg_i}, 1 \leq j \leq N \tag{2.8}$$

where r^t is the rank in the t^{th} iteration, this ranking is affected by the degree of a gene j and a damping factor of the gene $0 \leq d_j \leq 1$. However, d is the extent to which the gene j ranking depends on the structure of the graph (the higher the d_j the higher dependency on the graph), and the parameter f is the absolute differential expression value of j. also, deg_i corresponds to the in-degree of gene i that is different from PageRank algorithm that uses the out-degree value instead. To handle the zero-one gap problem (when ranking a gene with 0 incoming edges), a dynamic damping factor was used [48], where each gene has its own damping factor. As the number of incoming edges of a gene j increases, the damping factor slowly increases as well to assimilate more connectivity information into the ranking of the gene.

Finally, the algorithm stops when it converges. The convergence is achieved when the magnitude of the difference of the ranks between the current iteration t and the previous iteration $t - 1$ falls below a $threshold = 0.001$.

2.2.3.5 IntDriver

IntDriver is one of the cohort level algorithms that prioritize driver genes from somatic mutations by utilizing the logic of matrix factorization to estimate a mutation score. Two

types of functional information are used as input, GO similarity of genes and binary adjacency matrix. A Frobenius norm-based regularization is used in the mutation scoring method to prevent overfitting [49]. The model is based on the optimization of the following function:

$$min_{U,V}||X - UV^T||_{F^2} + \lambda_n Tr\{V^T L_N V\} + \lambda_S Tr\{V^T L_S V\} \\ +\lambda_V ||V||_{F^2} \;\; s.t \;\; U \in \{0,1\}^{p \times k}$$

The matrix U is the sample assignment matrix, $U = [u_1 ... u_, k] = s \times k$, and the matrix V is the mutation score matrix, $V = [v_1 ... v_, k] = g \times k$. The reconstruction of the mutation matrix X is denoted by the matrixUV^T in which k is the rank of the matrix UV^T. Based on these matrices, we can calculate and measure scores for each gene. The idea of the algorithm is able to detect rare genes (genes with low mutation frequencies). The Laplacian matrix of any gene interaction definition is: $L_N = D_N - A_N$, where A_N stands for the adjacency matrix of the interaction network, and D_N consists of gene degrees. Also, the matrix L_S is the Laplacian GO similarity matrix, defined as $L_S = D_S - S$, where the values of GO similarity are scores for each gene pairs, and D_S is the summation of rows or columns of the similarity matrix for the related gene. The default values for the algorithm parameters are $\lambda_N = 0.3$,$\lambda_S = 0.7$ and $\lambda_V = 0.01$. $||V||F^2$ stands for the Frobenius norm regularization that prevents overfitting on the mutation scores.

Finally, genes are ranked as follows:

$$Score(j) = max\{V_{jk^j} k^j = 1, ...k\}, \;\; j = 1, ...; g \tag{2.9}$$

where, for each gene j, the maximum score is selected as a driver gene's final score.

CHAPTER 3

3. Materials and Methods

3.1 Introduction

In this chapter, we describe the input data used in our project. And also, we describe the details of the main steps of the BetweenNet algorithm. Figure 3.1 provides an overview of the algorithm.

3.2 Input Data

In order to construct the pan-cancer cohort, we first identify the cancer types that have more than 10 paired measurements from normal and tumor samples in the TCGA cohort [40] as shown in Table 3.1. We then take the union of all the samples from these cancer types to form the cohort. In addition to the pan-cancer data, we perform separate evaluations on two cancer types. These are breast cancer (BRCA) with 110 samples, and lung cancer (LUSC + LUAD) with 61 samples.

We download the gene expression (RSEM normalized values [50]) and somatic mutation data for these patients from the Firebrowse database (*http://firebrowse.org*; version 2016_01_28). We exclude the silent mutations in the calculation of mutation frequencies. In addition to the gene expression and mutations data, we also employ protein-protein interactions data which we gather from the *H. Sapiens* PPI network of the IntAct database [51] on 18th June, 2020.

We preprocess the IntAct network so that both interactors are proteins and both are

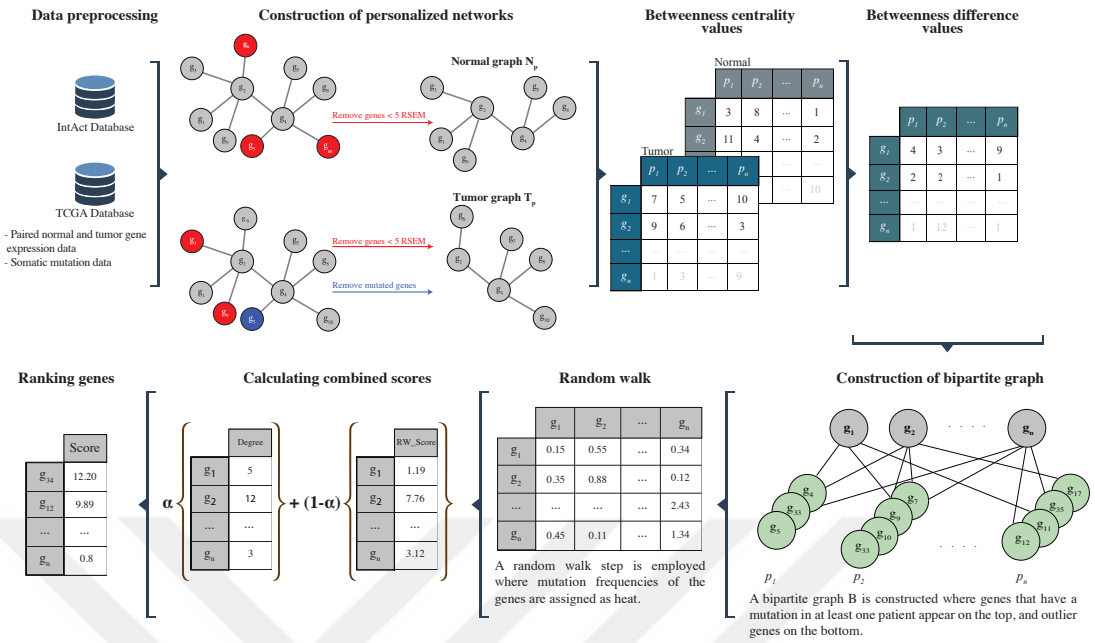

Figure 3.1: Main steps of the BetweenNet algorithm.

from the human genome to avoid human-virus interactions. Also, we only include the interactions where the type is ”physical association” or one of its descendants. Next, we convert UniProt ids to gene symbols where we merge multiple UniProt ids for the same protein to a single id. The resulting network contains 15,345 nodes and 113,524 edges.

3.3 Graph Construction

Let $G = (V, E)$ represent the reference *H. Sapiens* PPI network where each vertex $u_i \in V$ denotes a gene i whose expression gives rise to the corresponding protein in the network. Each undirected edge $(u_i, u_j) \in E$ denotes the interaction among the proteins corresponding to the genes i, j. Let P represent the set of patient samples. For each patient $p \in P$, we define two graphs N_p and T_p that represent the PPI networks of the normal and tumor samples, respectively. To construct N_p, we start with the reference PPI network G and remove the nodes that correspond to the genes that are not expressed in the normal sample of the patient p.

We deem genes with normalized count values less than 5 as not expressed. To construct

Table 3.1: Number of samples in each cancer type within the pan-cancer dataset

Cancer type	Samples
STAD	27
LUSC	16
KIRC	65
KICH	25
KIRP	28
LIHC	49
LUAD	45
BRCA	110
ESCA	11
PRAD	43
HNSC	37
THCA	46
COADREAD	25
STES	38
BLCA	15

T_p, we remove two sets of genes: (i) genes with normalized count value less than 5 in the tumor sample; (ii) genes that contain non-silent mutations in the tumor sample.

3.4 Betweenness Calculation

The standard definition of the betweenness centrality ignores the length of a shortest path. Since considering very long paths as functional relations may not be biologically meaningful, we use a variant of the betweenness centrality called *k-betweenness*, where only shortest paths of length $\leq k$ are included in the calculations [52]. Given an unweighted graph $G = (V, E)$, k-betweenness value of a node that corresponds to gene i is defined as follows:

$$\sum_{\forall s,t \in V, s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{3.1}$$

where σ_{st} is the number of shortest paths of length $\leq k$ between genes s and t, and $\sigma_{st}(i)$ is the number of such paths that pass through gene i. We utilize the algorithm presented in Brandes *et al.* to efficiently calculate the k-betweenness values [53]. Let $B^N_{p,i}$ and $B^T_{p,i}$ denote the k-betweenness centrality values of the gene i in the N_p and T_p graphs of the patient p, respectively. We define $B^{diff}_{p,i}$ as $|B^N_{p,i} - B^T_{p,i}|$.

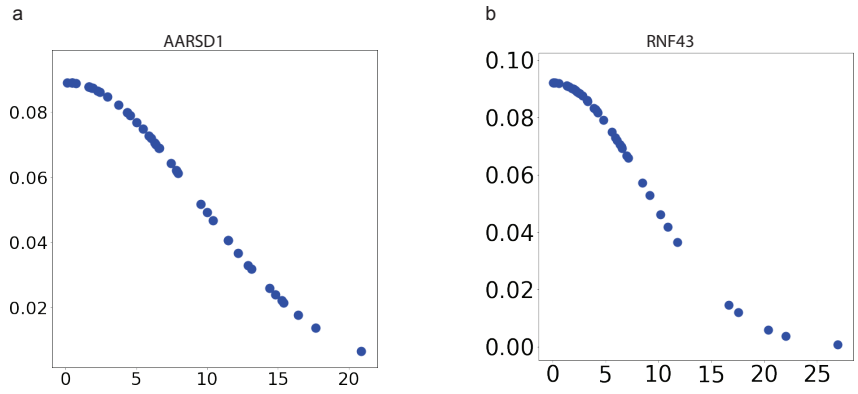

Figure 3.2: The distribution of betweenness difference values for two selected genes AARSD1 and RNF43.

3.5 Selection of outliers

For each gene i which exist in both normal and tumor networks, we plot the $B_{p,i}^{diff}$ values across all the patients. We observe that the distribution can be approximated with a truncated normal distribution (Figure 3.2).

We use the *truncnorm* function in Python to estimate the mean and standard deviation of the distribution. A gene i is defined as an *outlier* in patient p, if $B_{p,i}^{diff}$ is greater than t standard deviations from the mean. We repeat this process for each gene and construct a set of outlier genes for each patient.

3.6 BetweenNet algorithm

3.6.1 Bipartite graph

Similar to DriverNet, we construct a bipartite graph B that models the relationship between the set of mutated genes and the outliers. The *mutations partition* of the bipartite graph consists of the genes that have a mutation in at least one patient and the *outliers partition* consists of the outlier genes of all the patients in the cohort. Note that a gene j can be an outlier for multiple patients. In such a case, each occurrence of a gene is represented with a distinct node in the outliers partition of B. Assuming j is an outlier gene for

patient p, let u_j^p be the node corresponding to it in the outliers partition. For a node u_i in the mutations partition, edge (u_i, u_j^p) is inserted in B, if gene i is mutated in p and (u_i, u_j) is an edge in G.

3.6.2 Random Walk

We apply a random walk on the bipartite graph B. The mutation frequencies of the genes are assigned as initial heat values to be diffused throughout the network during random walk. Let $MF(i)$ denote the mutation frequency of gene i, that is, the number of patients where i has a non-silent mutation divided by the total number of patients. Note that heat values are assigned to genes on both sides of the bipartite graph. The random walk starts at a node u_i in B and at each time step moves to one of u_i's neighbors with probability $1 - \beta$ $(0 \leq \beta \leq 1)$. The walk can also restart from u_i with probability β, called the *restart probability*. This process can be defined by a transition matrix T which is constructed by setting $T_{ij} = \frac{1}{deg(u_j)}$ if $(u_i, u_j) \in E$, and $T_{ij} = 0$ otherwise. Here, $deg(u_j)$ corresponds to the degree of the node u_j. Thus T_{ij} can be interpreted as the probability that a simple random walk will transition from u_j to u_i. The random walk process can also be considered as a network propagation process by the equation, $F_{t+1} = (1 - \beta)TF_t + \beta F_0$, where F_t is the distribution of walkers after t steps and F_0 is the diagonal matrix with initial heat values, that is $F_0[i, i] = MF(i)$. We compute the final distribution of the walk by calculating the F matrix iteratively until convergence.

3.6.3 Ranking Genes

Genes in the *mutations partition* of the bipartite graph B are prioritized by a score that combines both degree information and the edge weights that are inferred with random walk. Assuming that $w_{in}(u_i)$ indicates the sum of incoming edge weights for gene i after random walk, the combined score for gene i can be defined as follows:

$$S_i = \alpha \frac{deg(u_i)}{\max_{\forall u_j \in mutations} deg(u_j)} + (1 - \alpha) \frac{w_{in}(u_i)}{max_{\forall u_j \in mutations} w_{in}(u_j)} \tag{3.2}$$

Note that the $w_{in}(u_i)$ for gene i corresponds to summing the corresponding row of F for gene i after convergence. Once a gene is selected, we remove the corresponding node and

its neighbors in B. After each such update of the B graph, the maximum degree value and the degrees of all the genes are computed again, whereas the w_in values are pre-computed and remain fixed throughout the ranking procedure.

CHAPTER 4

4. Results and Discussion

We implemented the betweenness centrality measurement algorithm in C++ using the *LEDA* (Library of Efficient Data types and Algorithms) library. The remaining steps are implemented in Python using *NetworkX* library. All the code and necessary datasets are available at https://github.com/abu-compbio/BetweenNET. We compare BetweenNet results against those of five other existing cancer driver prioritization methods: DriverNet, Subdyquency, DawnRank, IntDriver, and Dopazo and Erten's prioritization method based on betweenness centrality values, hereafter named only *Betweenness*. Note that for the Betweenness method, although the original method ranks all genes, here we only rank mutated genes using the same method for a fair comparison, since all the other methods under consideration are designed to rank mutated genes only. DriverNet is chosen due to its close connection to our work. DawnRank and Subdyquency are included as they extend and improve over DriverNet. Betweenness is included as a baseline since our method utilizes a variation of betweenness differences in identifying outlier genes. Finally, IntDriver is included to represent the performance of a distinct strategy that is based on matrix factorization. We evaluate the methods with three datasets: lung cancer, breast cancer, and pan-cancer samples.

4.1 Validation

4.1.1 Compiling reference gene sets

We compile known cancer genes from the databases Cancer Gene Census (CGC) [54], Network of Cancer Gene (NCG) [55] and CancerMine [56]. From CGC, we obtain a list of 723 genes that are found to be associated with cancer. We further identify the genes with mutation frequencies $\leq$ 2%, namely the *rare drivers*. Since the number of paired samples for breast and lung cancer is small, we use all available samples in breast and lung cohorts to compute the mutation frequencies of genes for defining *rare drivers*. For pan-cancer dataset, the number of paired samples is much larger. Therefore, we calculate mutation frequencies with paired samples only. Furthermore, we filter the genes according to the *Tumour Types* column to define cancer type specific gene sets for breast and lung cancer. We also compile cancer type specific genes from NCG by filtering the *primary site* column. Because these cancer type specific reference gene sets are small we take the union of CGC and NCG cancer type specific reference genes. The third repository, CancerMine, uses text-mining to catalogue cancer associated genes where it also extracts information about the type of the cancer. We compile two lists of genes that have at least 3 and 5 citations, respectively. Hereafter, these two reference gene sets are named *CancerMine3* and *CancerMine5*. The number of genes in each reference set for each cancer type (i.e., lung, breast) and for pan-cancer cohort are available in the Appendix A Supplementary Tables A.4-A.6. For lung cancer, we are unable to use *CancerMine5* as a reference due to its small size.

4.1.2 Enrichment analysis with Gene Ontology and pathway databases

For Gene Ontology (GO) [57] term analysis, we use *goatools*. We download go-basic.obo file from http://geneontology.org/docs/download-ontology/ on June 26th of 2019. We restrict the gene annotations to level 5 by ignoring the higher-level annotations and replacing the deeper-level category annotations with their ancestors at the restricted level.

For the pathway analysis, we use the *AllEnricher* tool with Reactome and Kyoto Encyclopedia of Genes and Genomes (KEGG) [58] pathways. Both *goatools* and *AllEnricher* use Fisher's exact test to calculate p-values and False Discovery Rate (FDR) for multiple testing correction. We use 0.05 as the p-value cutoff to determine significant enrichments.

4.2 Results

4.2.1 Sensitivity of BetweenNet to its parameter settings

We assess the sensitivity of BetweenNet to its parameterization by varying the parameters t, β, α and k for lung, breast and pan-cancer samples (Appendix A Supplementary Figures A.4 to A.14). Among them, the largest change is observed when the outlier detection threshold t is increased from 0.5 to larger values. Varying the other parameters results in minimal changes where the changes can only be discerned at the 5th decimal point and beyond. We choose the following setting as it leads to the best performance: $t = 0.5$, $\beta = 0.4$, $\alpha = 0.5$, $k = 3$. Overall, these tests show that BetweenNet is robust to a variety of parameter settings.

4.2.2 Evaluations with respect to reference cancer gene sets

We first compare the methods based on their ability to recover the sets of known cancer genes. For this, we compute true positive and false positive rates for the top 1000 genes and calculate the area under the ROC (AUROC). Figure 4.1 shows the ROCs obtained from lung cancer data. In Figure 4.1-a all CGC genes are used as reference, whereas in Figure 4.1-b genes with mutation frequencies $\leq 2\%$, namely the *rare drivers*, are included. Figure 4.1-c is obtained with CancerMine3 as the reference set. BetweenNet achieves a higher AUROC value than all the alternatives for CGC and CancerMine3 reference sets and it has the same AUROC value with DawnRank for CGC-rare reference sets. For CGC and CancerMine3, the ranking of the other methods is the same. Namely, the second ranked method is DawnRank which is followed by Subdyquency and DriverNet with similar performance with respect to each other. Finally, Betweenness and IntDriver are the

worst ranking methods. On the other hand, for the CGC-rare reference set, BetweenNet and DawnRank both perform the best whereas the second ranking method is Betweenness. This is followed by DriverNet and IntDriver which have the same AUROC value. Subdyquency performs significantly worse than all the other methods for the CGC-rare reference set. The fact that Betweenness performs much better than DriverNet, Subdyquency and IntDriver is interesting and suggests that most existing methods perform much better in retrieving drivers with larger mutation frequencies. Comparisons using the union of CGC-Lung and NCG-Lung reference sets show that BetweenNet has a significantly better performance than all the other models in retrieving lung cancer specific reference gene sets (Appendix A Supplementary Figure A.1). Here, Subdyquency ranks second, which is followed by DawnRank, DriverNet, Betweenness, and IntDriver. Overall, these results illustrate the superiority of BetweenNet as it can find both rare and common drivers in lung cancer accurately.

Figure A.3 depicts analogous results for the breast cancer data. BetweenNet achieves the top performance with CGC and CGC-rare reference sets. For both reference sets, the ranking of the other methods from best to worst is the same and as follows: DawnRank, Subdyquency, Betweenness, DriverNet, and IntDriver. For CancerMine3, DawnRanks shows the best performance. BetweenNet's AUROC value is slightly worse than DawnRank. This is followed by BetweenNess and Subdyquency. As in the other results of breast cancer, DriverNet and IntDriver are the worst performing methods. Subdyquency ranks the best in retrieving breast cancer specific reference gene sets (union of CGC-Breast and NCG-Breast)(Appendix A Supplementary Figure A.2-a). BetweenNet's performance is slightly worse than Subdyquency. The other methods rank as follows: DawnRank, DriverNet, Betweenness, IntDriver. Results with respect to the CancerMine5 reference set are similar to those obtained with CancerMine3 reference set and are available in Appendix A Supplementary Figure A.2-b.

Lastly, Figure 4.3 shows the results with respect to the pan-cancer dataset. For the CGC reference gene set, Subdyquency performs the best. BetweenNet ranks the second, which is followed by DawnRank, DriverNet, and IntDriver, respectively. Interestingly, Betweenness performs the worst in this evaluation. The employed methods rank differently when the reference set is changed to CGC-rare. BetweenNet and DawnRank have

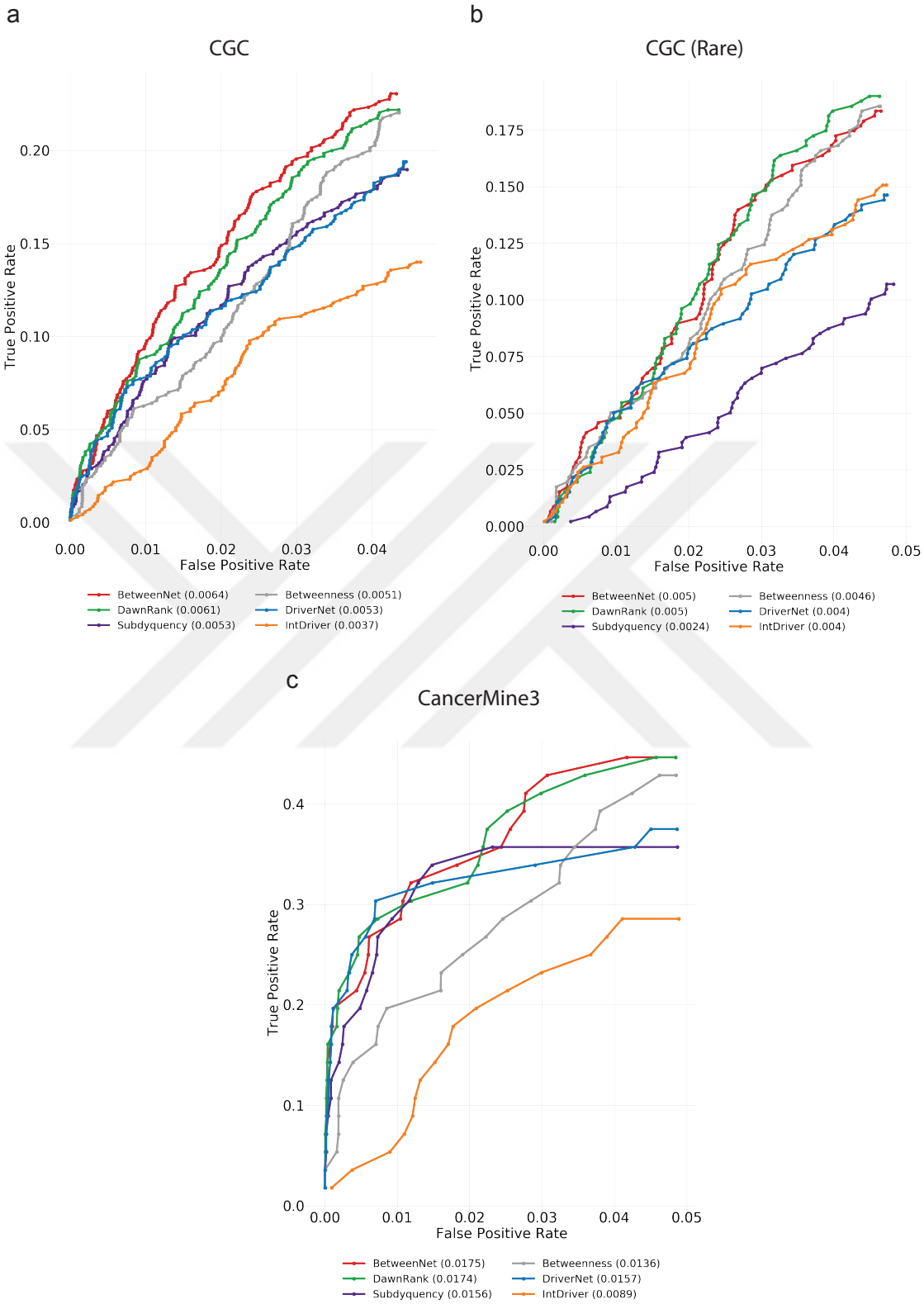

Figure 4.1: The fraction of recovered reference genes is shown with a ROC curve for lung cancer data a) *CGC* genes are used as reference. b) *CGC* rare genes are used as reference. c) *CancerMine3* genes are used as reference.

a similar performance and perform the best. DriverNet and Betweenness rank the second and third, respectively. Subdyquency ranks the fourth which is surprising given its top performance with the full CGC reference set. For the CancerMine3 reference set, Subdyquency's AUROC is the highest. BetweenNet's performance is slightly worse than Subdyquency. DawnRank and DriverNet give the same performance and rank third. Betweenness and IntDriver are the worst performing methods. Results with CancerMine5 are similar to those of CGC and CancerMine3, respectively. These are available in Appendix A Supplementary Figure A.3.

4.2.3 Evaluations based on functional and pathway analysis

Reference cancer driver gene sets might be incomplete and biased. As such, rather than only finding exact matches between the output gene sets and the reference gene sets, we also define other metrics that measure how well the associated functions of the genes of the two sets match. One such metric is based on GO consistency (GOC) and the other is based on pathway information. For the former, we find the GO terms enriched in the output gene sets and in the reference gene sets, and check whether the corresponding GO terms overlap. The underlying assumption is that the reference cancer genes and the predicted cancer genes should have similar biological functions. We find the enriched GO terms in the ranked gene sets of varying total sizes from 100 to 500 in the increments of 100 for each method under consideration. We repeat the same GO term enrichment analysis with the reference gene set. We then compute the GOC value between the enriched GO terms of the ranked gene set and those of the reference set, which is defined as the ratio between the size of the intersection of the two sets and the size of the union [59]. Figure 4.4 shows the GOC values calculated for each cancer type and pan-cancer cohort. We observe that BetweenNet ranked genes for lung cancer perform the best for almost all total size values. Here, Subdyquency's low performance is notable since it performs similar to DriverNet and Betweenness in retrieving CGC genes for lung cancer. For breast cancer, DawnRank ranks the best in four out of five total size values, whereas BetweenNet is the second best. This is followed by Betweenness, Subdyquency and DriverNet, respectively. Finally, IntDriver performs significantly worse than the other methods. For pan-cancer

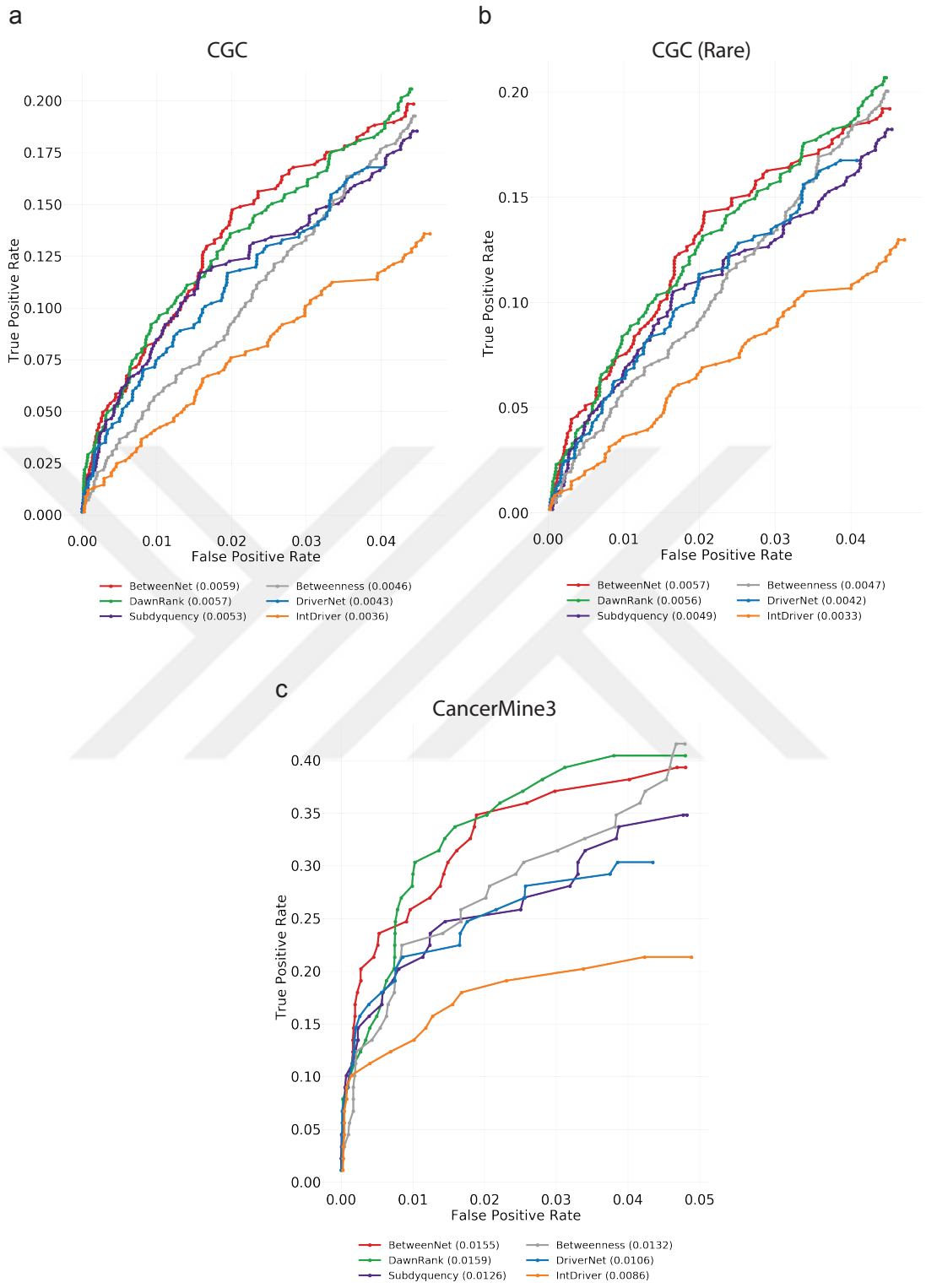

Figure 4.2: The fraction of recovered reference genes is shown with a ROC curve for breast cancer data a) *CGC* genes are used as reference. b) *CGC* rare genes are used as reference. c) *CancerMine3* genes are used as reference.

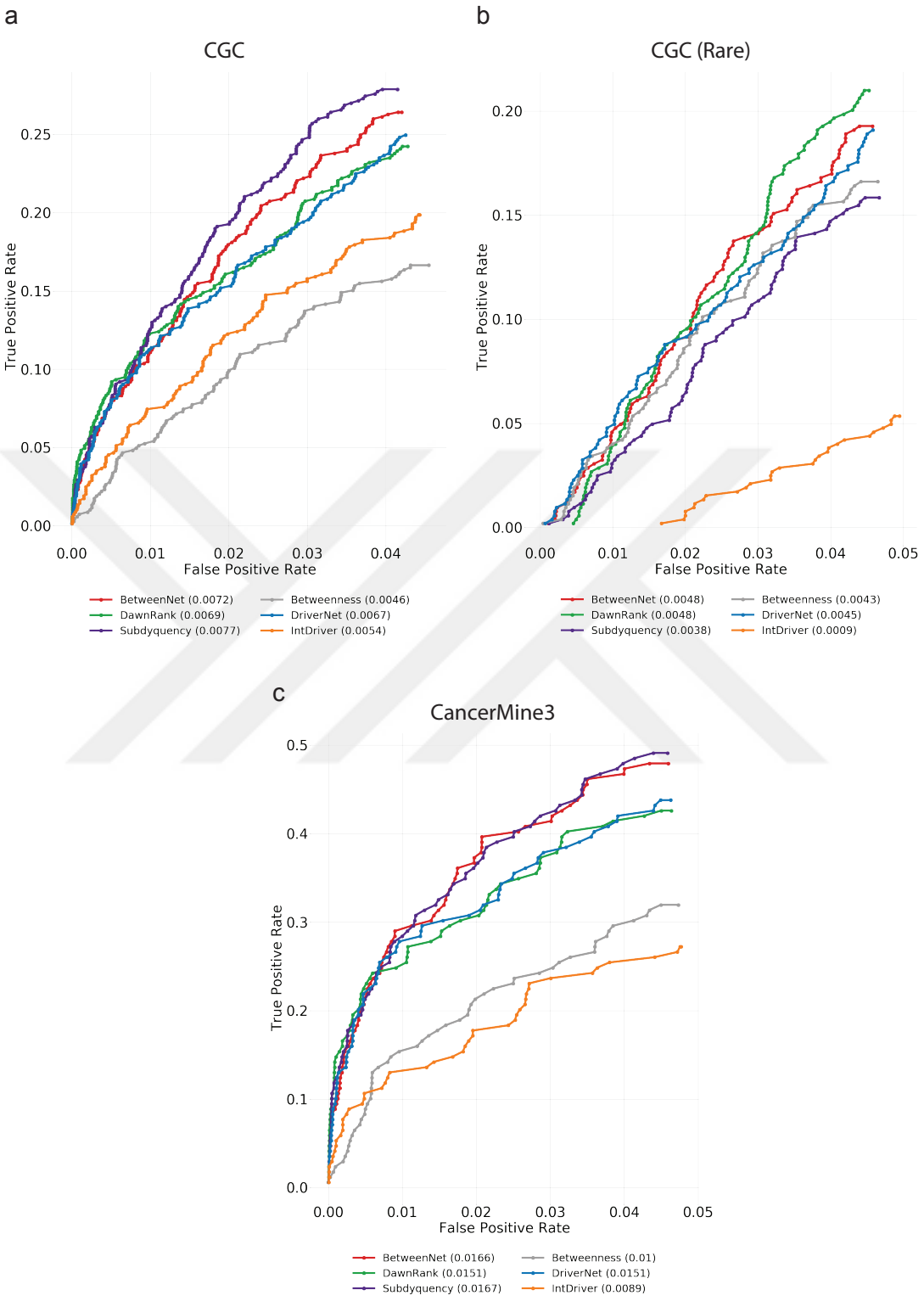

Figure 4.3: The fraction of recovered reference genes is shown with a ROC plot for pan-cancer data a) *CGC* genes are used as reference. b) *CGC* rare genes are used as reference. c) *CancerMine3* genes are used as reference.

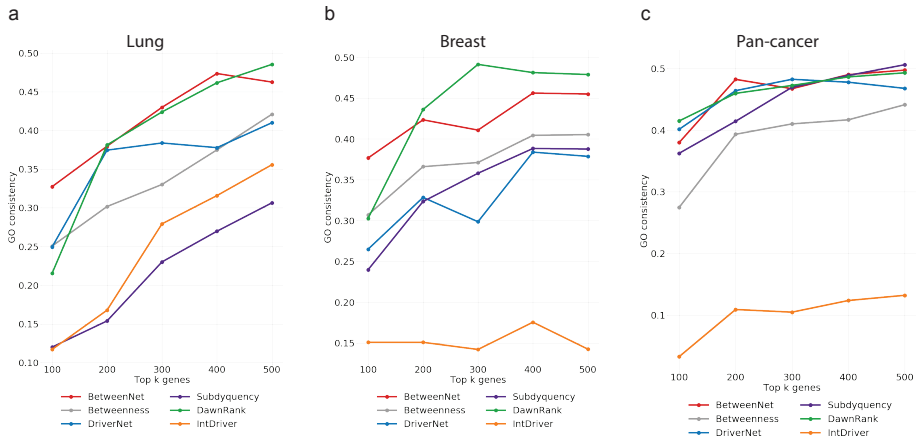

Figure 4.4: GO consistency values for a) lung cancer b) breast cancer c) pan-cancer cohort.

data, there is no clear winner. BetweenNet, DriverNet and DawnRank perform close to each other whereas Subdyquency's performance is the best for total sizes of 400 and 500. Similar to the results obtained from breast cancer data, IntDriver's performance is notably worse than all the other methods.

We repeat the same type of analysis with pathways as well, this time replacing GO term enrichment with pathway enrichment. Namely, we identify the pathways enriched in the reference set of genes and the set of genes output by a ranking method. We then compute the number of pathways common in both of these sets. Figure 4.5 shows the results with Reactome reference pathways for all cancer types. For lung cancer, the best method varies for each total size value, where BetweenNet outperforms the other models with a large margin for total sizes 100 and 300. On the other hand, DawnRank results in the top consistency values for total sizes 200 and 400. Finally, for a total size 400, these two methods share the same performance. For breast cancer, we observe similar results where BetweenNet and DawnRank perform the top. Here, DriverNet's performance is notably worse than the other methods. For pan-cancer, BetweenNet gives the top consistency value in four out of five cases. Interestingly, Subdyquency ranks lower than BetweenNet, DawnRank and DriverNet in contrary to its top performance in evaluations with respect to CGC on pan-cancer data. It is less difficult to identify the top performing method when KEGG pathways are used as reference Figure 4.5. For lung cancer, BetweenNet gives the best performance for four out five cases, whereas for breast cancer and pan-cancer data

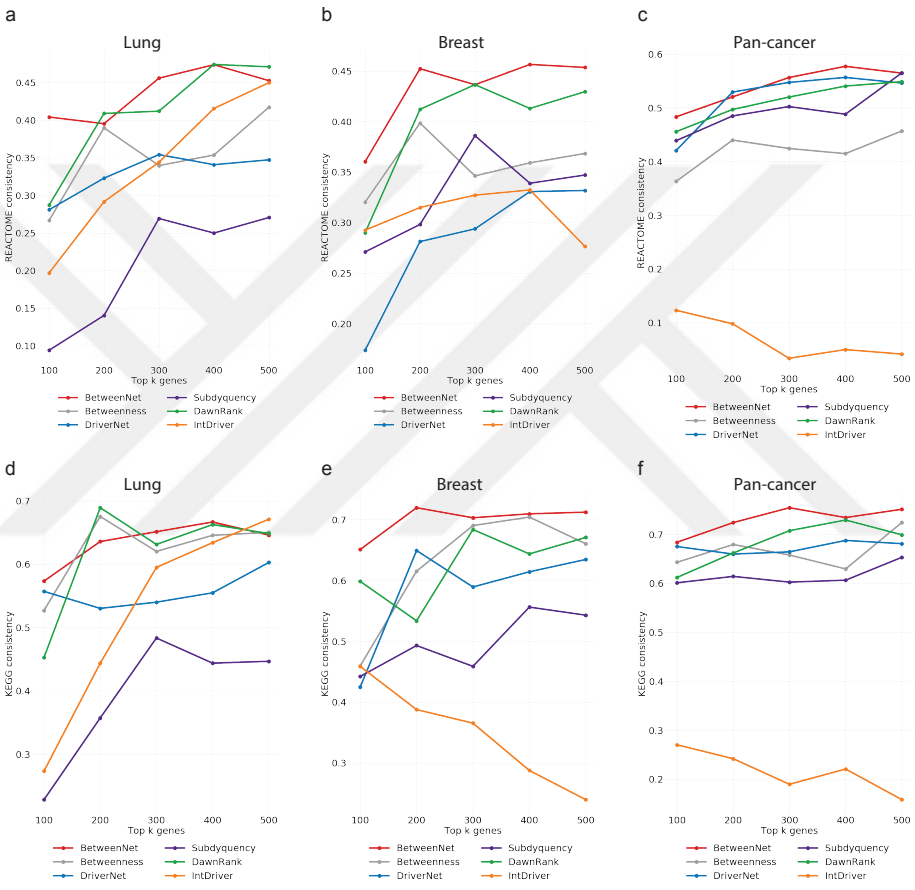

Figure 4.5: Reactome pathway consistency values for a) lung cancer b) breast cancer c) pan-cancer cohort.

BetweenNet ranks top for all total size values. Subdyquency's low performance is again notable in these evaluations.

4.3 Analysis of BetweenNet ranked genes

We further explore BetweenNet's top 30 ranking genes for each dataset (Appendix A Supplementary tables A.1 to A.3). Among the CGC genes that appear in our top 30 genes for breast cancer, EWSR1 can be found by BetweenNet only whereas ERBB2 and HSP90AB1 can be found by BetweenNet and Betweenness only. We observe that these three genes have lower mutation frequencies than the other CGC genes that appear in our top 30 genes. Namely, HSP90AB1 is mutated in a single patient and the other two genes are mutated in two patients. Similarly, for lung cancer CGC genes SMAD2 and REL can only be detected by BetweenNet and Betweenness within the top 30 ranking genes. Again, these genes have the lowest mutation frequencies among the CGC genes in our top 30 ranking genes. In order for BetweenNet to rank these genes higher than other genes with larger mutation frequencies, many connections must exist between these genes and the outlier sets of the patients that they are mutated in. The fact that these genes cannot be recovered by DriverNet or Subdyquency suggests that defining outlier genes based on betweenness centrality provides an advantage over defining them based on gene expression.

We also check the top 30 ranking genes that do not appear in CGC. Among these, LRRK2 consistently ranks within our top 30 genes for lung cancer, breast cancer, and pan-cancer datasets. LRRK2 is also ranked within the top 30 genes by DawnRank, DriverNet and Betweenness for breast cancer dataset; and by all the other methods except IntDriver for lung cancer and pan-cancer datasets. Indeed, multiple studies have reported that individuals with LRRK2 mutations have an increased risk of developing cancers [60, 61, 62]. Another gene which is ranked among our top 30 genes for all three datasets is RIF1. RIF1 is also identified by DriverNet and DawnRank in all three datasets. Supporting this finding, RIF1 is recently shown to promote tumor growth and cancer stem cell-like traits in non-small-cell lung carcinoma by activating the Wnt/β-catenin signaling pathway. On the other extreme, there are also genes which are only identified by BetweenNet. MAGED1

is one such example for breast cancer. Tian et al have shown that BRCA2 suppresses cell proliferation via stabilizing its downstream target MAGED1 [63]. As such, MAGED1 is strongly associated with cancer development by mediating the growth-suppressing function of BRCA2 [64].

4.4 Discussion

Having shown that BetweenNet performs better than existing methods in most of the evaluations, we also investigate the added value of defining outliers based on betweenness centrality. To this end, we replace the outliers of BetweenNet with the outliers found by DriverNet for the same data sets. We observe that BetweenNet performs significantly better than its modified version and DriverNet for all cancer types and for all reference gene sets (Appendix A Supplementary Figures 14-17). These results show that the outlier detection strategy of BetweenNet is critical to its performance.

Lastly, we analyze the running time requirements of the main steps of the BetweenNet algorithm. Computing the betweenness values of all the nodes in an unweighted graph of n nodes and m edges require $O(nm)$ time, since starting from each node a breadth-first search (BFS) is executed until completion to find the shortest path distances necessary for the betweenness values. However since we only consider the shortest paths within a diameter of k, the number of edges traversed at each BFS is bounded by δ^{2k+2}, where δ denotes the maximum degree of all the nodes in the graph G representing the input PPI network. Thus the running time of the betweenness step of BetweenNet is $O(|P||V|\delta^{2k+2})$. Let a denote the average number of outliers per patient and μ denote the number of genes mutated at least once in the set of samples P. The running time of the random-walk step is bounded by $O((|P|a+\mu)^3 r)$, where r denotes the number of times the F matrix is calculated iteratively until convergence. We observe the a and μ values of 870 and 4,335 for lung cancer; 390 and 4,096 for breast cancer; 627 and 11,105 for pan-cancer datasets respectively. We observe that the random walk converges after 3 iterations for all three datasets. For the actual ranking step, the main operations are those of iteratively selecting and removing the maximum rank mutated gene and updating the current ranks of the remaining mutated genes that are also connected to the outliers of the removed

gene. Although a more efficient structure such as a priority queue could be employed, since this is not the dominantly time-consuming step of the algorithm we opt for simple node deletions from B followed by a linear search for maximum ranking mutated gene. A removed mutated gene is on average incident to $O(\delta a)$ edges in B. Thus a single removal of a mutated gene and its neighbors in B and the following degree updates costs $O(\delta^2 a+\mu)$ time and the overall running time of the actual ranking step is $O(\mu(\delta^2 a + \mu))$.

CHAPTER 5

5. Conclusion

5.1 Conclusions

We propose BetweenNet, a novel cancer driver gene prioritization approach that integrates genomic data with the connectivity within PPI networks. One contribution of BetweenNet is the identification of patient specific dysregulated genes with a measure based on betweenness centrality on personalized networks. BetweenNet ranks mutated genes by their effects on dysregulated genes. To characterize these effects, a bipartite influence graph is formed to represent the relations between the mutated genes and dysregulated genes in each patient. Another contribution of BetweenNet is the employment of a random-walk process on the resulting influence bipartite graph. Through careful comparisons, we show that both the use of betweenness centrality metric and the employment of random walk have added values in the identification of cancer driver genes. We also demonstrate that BetweenNet outperforms the alternative methods in recovering known reference genes and in providing functionally coherent rankings with three large-scale TCGA datasets: lung cancer, breast cancer, and pan-cancer samples. Additionally, we find that many of our top ranking genes that do not appear in reference cancer gene sets have roles in cancer development based on existing literature. Taken together, our results indicate that BetweenNet effectively integrates genomic data and connectivity information to prioritize cancer driver genes.

5.2 Future Work

BetweenNet can be improved in several ways. Our approach is to use paired data (samples with both normal and tumor tissues), which makes it critical and quite limited, due to small cohort sizes, which pushes forward to think more about providing or constructing computational models that predict and generate normal samples for the missing ones.

Another problem is related to the use of bulk expression data, in which non-cancerous cells can be present in the data, many research and publications are suggesting to switch to single-cell data to be used instead of bulk-expression data.

In terms of running time, our method is implemented in both C++ and python, where python parts are quite consuming time and memory inefficiently, which makes the use of our method quite limited to small data only. To prevent such a problem in the future, we may try to switch to C++.

Finally, our method shows less performance in terms of GO evaluation and AUROC values for pan-cancer comparing to the other methods. Introducing more information in our scoring function can improve our results quite better.

Bibliography

[1] "Types of variants," *Garvan Institute of Medical Research*, 11 2018. (Online) www.garvan.org.au/research/kinghorn-centre-for-clinical-genomics/learn-about-genomics/for-gp/genetics-refresher-1/types-of-variants (Access Date:22.12.2020).

[2] A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D. G. Huntsman, C. Caldas, S. A. Aparicio, and S. P. Shah, "DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer," *Genome Biology*, vol. 13, no. 12, p. R124, 2012.

[3] D. Masica and R. Karchin, "Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival," *Cancer research*, vol. 71, pp. 4550–61, 05 2011.

[4] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes.," *Science (New York, N.Y.)*, vol. 339, pp. 1546–1558, Mar. 2013.

[5] S. Erten, G. Bebek, and M. Koyuturk, "Vavien: An algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks," *J Comput Biol*, vol. 18, pp. 1561–74, 11 2011.

[6] M. Lawrence, P. Stojanov, P. Polak, G. V Kryukov, K. Cibulskis, A. Sivachenko, S. L Carter, C. Stewart, C. H Mermel, S. Roberts, A. Kiezun, P. S Hammerman, A. McKenna, Y. Drier, L. Zou, A. H Ramos, T. J Pugh, N. Stransky, E. Helman,

and G. Getz, “Mutational heterogeneity in cancer and the search for new cancer-associated genes,” *Nature*, vol. 499, 06 2013.

[7] H. Yang, Q. Wei, X. Zhong, H. Yang, and B. Li, “Cancer driver gene discovery through an integrative genomics approach in a non-parametric bayesian framework,” *Bioinformatics*, vol. 33, no. 4, pp. 483–490, 2017.

[8] J. Dopazo and C. Erten, “Graph-theoretical comparison of normal and tumor networks in identifying BRCA genes,” *BMC Systems Biology*, vol. 11, nov 2017.

[9] C. A. Miller, S. H. Settle, E. P. Sulman, K. D. Aldape, and A. Milosavljevic, “Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors,” *BMC Medical Genomics*, vol. 4, p. 34, Apr 2011.

[10] F. Vandin, E. Upfal, and B. J. Raphael, “*De Novo* discovery of mutated driver pathways in cancer,” in *Research in Computational Molecular Biology - 15th Annual International Conference, RECOMB 2011, Vancouver, BC, Canada, March 28-31, 2011. Proceedings*, pp. 499–500, 2011.

[11] M. D. M. Leiserson, F. Vandin, H.-T. Wu, J. R. Dobson, J. V. Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M. S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G. A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, and B. J. Raphael, “Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes,” *Nature Genetics*, vol. 47, pp. 106–114, Dec. 2014.

[12] B. Liu, C. Wu, X. Shen, and W. Pan, “A novel and efficient algorithm for de novo discovery of mutated driver pathways in cancer,” *Ann. Appl. Stat.*, vol. 11, pp. 1481–1512, 09 2017.

[13] C. M. Dimitrakopoulos and N. Beerenwinkel, “Computational approaches for the identification of cancer genes and pathways,” 2017. Published online 11 November 2016.

[14] J. Zhang and S. Zhang, “The discovery of mutated driver pathways in cancer: Models and algorithms,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, pp. 988–998, May 2018.

[15] M. Bailey, C. Tokheim, E. Porta, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. Wendl, J. Kim, B. Reardon, K. s. Ng, K. Jeong, S. Cao, Z. Wang, J. Gao, Q. Gao, F. Wang, E. M. Liu, L. Mularoni, and A. Mariamidze, "Comprehensive characterization of cancer driver genes and mutations," *Cell*, vol. 173, pp. 371–385.e18, 04 2018.

[16] C. Tokheim, N. Papadopoulos, K. Kinzler, B. Vogelstein, and R. Karchin, "Evaluating the evaluation of cancer driver genes," *Proceedings of the National Academy of Sciences*, vol. 113, p. 201616440, 11 2016.

[17] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding, "MuSiC: Identifying mutational significance in cancer genomes," *Genome Research*, vol. 22, pp. 1589–1598, aug 2012.

[18] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. Dicara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, 2013.

[19] E. Hodis, I. R. Watson, G. V. Kryukov, S. T. Arold, M. Imielinski, J. P. Theurillat, E. Nickerson, D. Auclair, L. Li, C. Place, D. Dicara, A. H. Ramos, M. S. Lawrence, K. Cibulskis, A. Sivachenko, D. Voet, G. Saksena, N. Stransky, R. C. Onofrio, W. Winckler, K. Ardlie, N. Wagle, J. Wargo, K. Chong, D. L. Morton, K. Stemke-Hale, G. Chen, M. Noble, M. Meyerson, J. E. Ladbury, M. A. Davies, J. E. Gershenwald, S. N. Wagner, D. S. Hoon, D. Schadendorf, E. S. Lander, S. B.

Gabriel, G. Getz, L. A. Garraway, and L. Chin, “A landscape of driver mutations in melanoma,” *Cell*, vol. 150, pp. 251–263, jul 2012.

[20] M. D. M. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael, “Simultaneous identification of multiple driver pathways in cancer,” *PLOS Comput Biol*, vol. 9, pp. 1–15, 05 2013.

[21] J. P. Hou and J. Ma, “Dawnrank: discovering personalized driver genes in cancer,” *Genome Medicine*, vol. 6, no. 56, pp. 1–16, 2014.

[22] K. Shi, L. Gao, and B. Wang, “Discovering potential cancer driver genes by an integrated network-based approach,” *Mol. BioSyst.*, vol. 12, 07 2016.

[23] P.-J. Wei, D. Zhang, J. Xia, and C.-H. Zheng, “Lndriver: Identifying driver genes by integrating mutation and expression data based on gene-gene interaction network,” *BMC Bioinformatics*, vol. 17, pp. 221–230, 12 2016.

[24] J. Song, W. Peng, and F. Wang, “A random walk-based method to identify driver genes by integrating the subcellular localization and variation frequency into bipartite graph,” *BMC Bioinformatics*, vol. 20, p. 238, dec 2019.

[25] C. Erten, A. Houdjedj, and H. Kazan, “Ranking cancer drivers via betweenness-based outlier detection and random walks,” *bioRxiv*, 2020.

[26] J.-D. J. Han, “Understanding biological functions through molecular networks,” *Cell Research*, vol. 18, pp. 224–237, Feb. 2008.

[27] L. E. MacConaill and L. A. Garraway, “Clinical Implications of the Cancer Genome,” *Journal of Clinical Oncology*, vol. 28, pp. 5219–5228, Dec. 2010.

[28] M. Shatnawi, “Chapter 6 - review of recent protein-protein interaction techniques,” in *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology* (Q. N. Tran and H. Arabnia, eds.), Emerging Trends in Computer Science and Applied Computing, pp. 99 – 121, Boston: Morgan Kaufmann, 2015.

[29] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Research*, vol. 28, pp. 27–30, 01 2000.

[30] S. Goto, T. Nishioka, and M. Kanehisa, “LIGAND: chemical database of enzyme reactions,” *Nucleic Acids Research*, vol. 28, pp. 380–382, 01 2000.

[31] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D’Eustachio, “The reactome pathway knowledgebase,” *Nucleic Acids Research*, vol. 48, pp. D498–D503, 11 2019.

[32] J. A. Blake, J. T. Eppig, C. J. Bult, J. A. Kadin, and M. G. D. G. Richardson, Joel E., “The Mouse Genome Database (MGD): updates and enhancements,” *Nucleic Acids Research*, vol. 34, pp. D562–D567, 01 2006.

[33] G. Grumbling and T. F. C. Strelets, Victor, “FlyBase: anatomical data, images and queries,” *Nucleic Acids Research*, vol. 34, pp. D484–D488, 01 2006.

[34] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, “The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology,” *Nucleic Acids Research*, vol. 32, pp. D262–D266, 01 2004.

[35] T. R. Gruber, “A translation approach to portable ontology specifications,” *Knowledge Acquisition*, vol. 5, no. 2, pp. 199 – 220, 1993.

[36] D. M. Jones and R. C. Paton, “Toward principles for the representation of hierarchical knowledge in formal ontologies,” *Data Knowledge Engineering*, vol. 31, no. 2, pp. 99 – 113, 1999.

[37] R. Stevens, C. A. Goble, and S. Bechhofer, “Ontology-based knowledge representation for bioinformatics,” *Briefings in Bioinformatics*, vol. 1, pp. 398–414, 11 2000.

[38] S. E. Lewis, “Gene ontology: looking backwards and forwards,” *Genome Biology*, vol. 6, no. 1, p. 103, 2004.

[39] G. O. Consortium, “The Gene Ontology (GO) database and informatics resource,” *Nucleic Acids Research*, vol. 32, pp. D258–D261, 01 2004.

[40] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemporary oncology (Poznan, Poland)*, vol. 19, no. 1A, pp. A68–A77, 2015.

[41] L. Chin, J. Andersen, and P. Futreal, "Cancer genomics: From discovery science to personalized medicine," *Nature medicine*, vol. 17, pp. 297–303, 03 2011.

[42] L. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 03 1977.

[43] H. Tang, Y. Chen, X. Liu, S. Wang, Y. Lv, D. Wu, Q. Wang, M. Luo, and H. Deng, "Downregulation of HSP60 disrupts mitochondrial proteostasis to promote tumorigenesis and progression in clear cell renal cell carcinoma," *Oncotarget*, vol. 7, no. 25, pp. 38822–38834, 2016.

[44] X. Peng, J. Wang, J. Wang, F.-X. Wu, and Y. Pan, "Rechecking the centrality-lethality rule in the scope of protein subcellular localization interaction networks," *PLOS ONE*, vol. 10, pp. 1–22, 06 2015.

[45] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[46] E. Khurana, Y. Fu, V. Colonna, X. Mu, H. Kang, T. Lappalainen, A. Sboner, L. Lochovsky, J. Chen, A. Harmanci, J. Das, A. Abyzov, S. Balasubramanian, K. Beal, D. Chakravarty, D. Challis, Y. Chen, D. Clarke, L. Clarke, F. Cunningham, U. Evani, P. Flicek, R. Fragoza, E. Garrison, R. Gibbs, Z. Gümüş, J. Herrero, N. Kitabayashi, Y. Kong, K. Lage, V. Liluashvili, S. Lipkin, D. MacArthur, G. Marth, D. Muzny, T. Pers, G. Ritchie, J. Rosenfeld, C. Sisu, X. Wei, M. Wilson, Y. Xue, F. Yu, E. Dermitzakis, H. Yu, M. Rubin, C. Tyler-Smith, and M. Gerstein, "Integrative annotation of variants from 1092 humans: Application to cancer genomics," *Science*, vol. 342, no. 6154, 2013.

[47] M. D'Antonio and F. Ciccarelli, "Integrated analysis of recurrent properties of cancer genes to identify novel drivers," *Genome Biology*, vol. 14, May 2013.

[48] X. Wang, T. Tao, J.-T. Sun, A. Shakery, and C. Zhai, "Dirichletrank: Solving the zero-one gap problem of pagerank," *ACM Trans. Inf. Syst.*, vol. 26, Apr. 2008.

[49] T. Fukuma and F. Toriumi, "Meta learning approach to collaborative filtering for tackling learning from sparse data and handling uncertainty," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 35, no. 5, pp. F–JC3$_1$ $-$ -9, 2020.

[50] B. Li and C. N. Dewey, "RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, p. 323, aug 2011.

[51] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, and H. Hermjakob, "The mintact project–intact as a common curation platform for 11 molecular interaction databases," *Nucleic acids research*, vol. 42, p. D358—63, January 2014.

[52] S. P. Borgatti and M. G. Everett, "A graph-theoretic perspective on centrality," *Social Networks*, vol. 28, no. 4, pp. 466 – 484, 2006.

[53] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," *Social Networks*, vol. 30, pp. 136–145, may 2008.

[54] S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. G. Cole, S. Ward, E. Dawson, L. Ponting, R. Stefancsik, B. Harsha, C. Y. Kok, M. Jia, H. Jubb, Z. Sondka, S. Thompson, T. De, and P. J. Campbell, "COSMIC: somatic cancer genetics at high-resolution," *Nucleic Acids Research*, vol. 45, pp. D777–D783, jan 2017.

[55] D. Repana, J. Nulsen, L. Dressler, M. Bortolomeazzi, S. K. Venkata, A. Tourna, A. Yakovleva, T. Palmieri, and F. D. Ciccarelli, "The Network of Cancer Genes (NCG): A comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens 06 Biological Sciences 0604 Genetics 11 Medical and Health Sciences 1112 Oncology and Carcinogenesis 06 Biological Sciences 0601 Biochemistry and Cell Biology," *Genome Biology*, vol. 20, p. 1, jan 2019.

[56] J. Lever, E. Y. Zhao, J. Grewal, M. R. Jones, and S. J. M. Jones, "CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer.," *Nature methods*, vol. 16, no. 6, pp. 505–507, 2019.

[57] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nature genetics*, vol. 25, pp. 25–9, may 2000.

[58] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research*, vol. 45, pp. D353–D361, jan 2017.

[59] A. E. Aladağ and C. Erten, "SPINAL: scalable protein interaction network alignment," *Bioinformatics*, vol. 29, pp. 917–924, apr 2013.

[60] R. Saunders-Pullman, M. J. Barrett, K. M. Stanley, M. S. Luciano, V. Shanker, L. Severt, A. Hunt, D. Raymond, L. J. Ozelius, and S. B. Bressman, "LRRK2 G2019S mutations are associated with an increased cancer risk in Parkinson disease," *Movement Disorders*, vol. 25, pp. 2536–2541, nov 2010.

[61] I. Agalliu, M. San Luciano, A. MirelmanMD, N. Giladi, B. Waro, J. Aasly, R. Inzelberg, S. Hassin-Baer, E. Friedman, J. Ruiz-Martinez, J. F. Marti-Masso, A. Orr-Urtreger, S. Bressman, and R. Saunders-Pullman, "Higher frequency of certain cancers in LRRK2 G2019S mutation carriers with Parkinson disease a pooled analysis," *JAMA Neurology*, vol. 72, pp. 58–65, jan 2015.

[62] R. Inzelberg, O. S. Cohen, J. Aharon-Peretz, I. Schlesinger, R. Gershoni-Baruch, R. Djaldetti, Z. Nitsan, L. Ephraty, O. Tunkel, E. Kozlova, L. Inzelberg, N. Kaplan, T. Fixler Mehr, A. Mory, E. Dagan, E. Schechtman, E. Friedman, and S. Hassin-Baer, "The LRRK2 G2019S mutation is associated with Parkinson disease and concomitant non-skin cancers," mar 2012.

[63] X. X. Tian, D. Rai, J. Li, C. Zou, Y. Bai, D. Wazer, V. Band, and Q. Gao, “BRCA2 suppresses cell proliferation via stabilizing MAGE-D1,” *Cancer Research*, vol. 65, pp. 4747–4753, jun 2005.

[64] Q. Du, Y. Zhang, X. X. Tian, Y. Li, and W. G. Fang, “Mage-D1 inhibits proliferation, migration and invasion of human breast cancer cells,” *Oncology Reports*, vol. 22, no. 3, pp. 659–665, 2009.

Appendix A

Supplementary

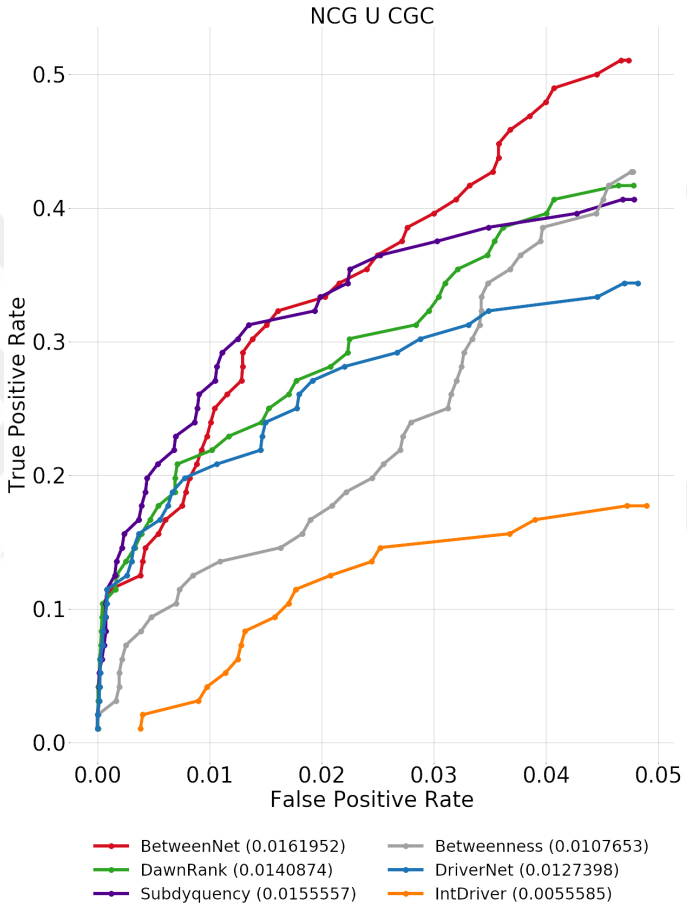

Figure A.1: The fraction of recovered reference genes is shown with a ROC curve for lung cancer data where the union of *CGC (Lung)* and *NCG (Lung)* genes are used as reference.

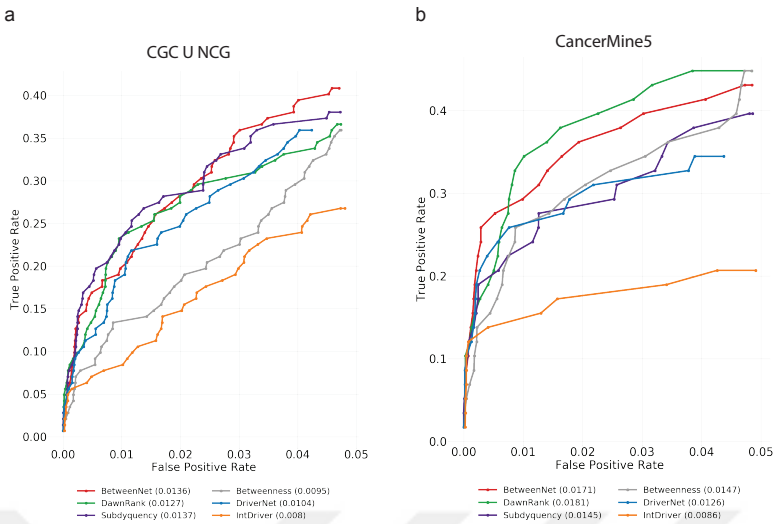

Figure A.2: The fraction of recovered reference genes is shown with a ROC curve for breast cancer data where a) the union of *CGC (Breast)* and *NCG (Breast)*, b) *CancerMine5* genes are used as reference.

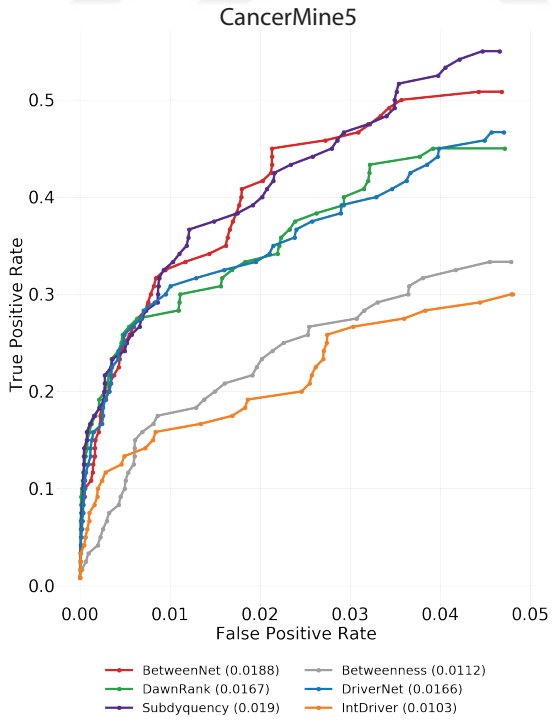

Figure A.3: The fraction of recovered reference genes is shown with a ROC curve for pan-cancer data where *CancerMine5* genes are used as reference.

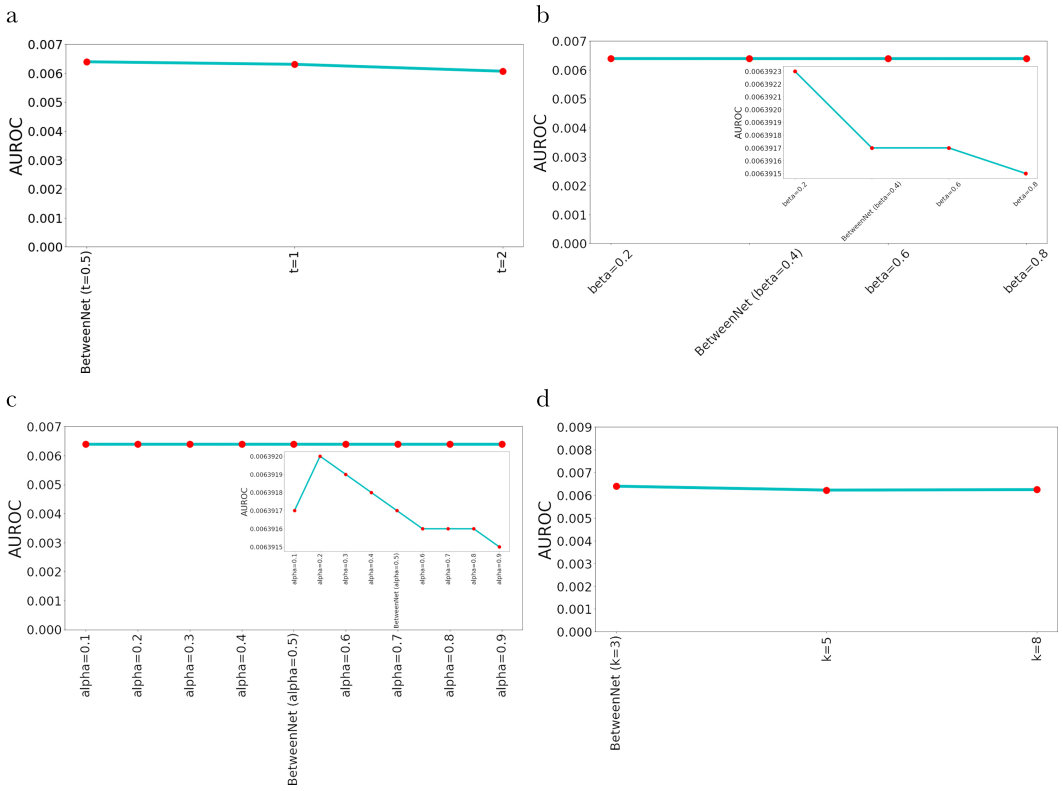

Figure A.4: Sensitivity test of parameters of BetweenNet on lung cancer data when *CGC* is used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

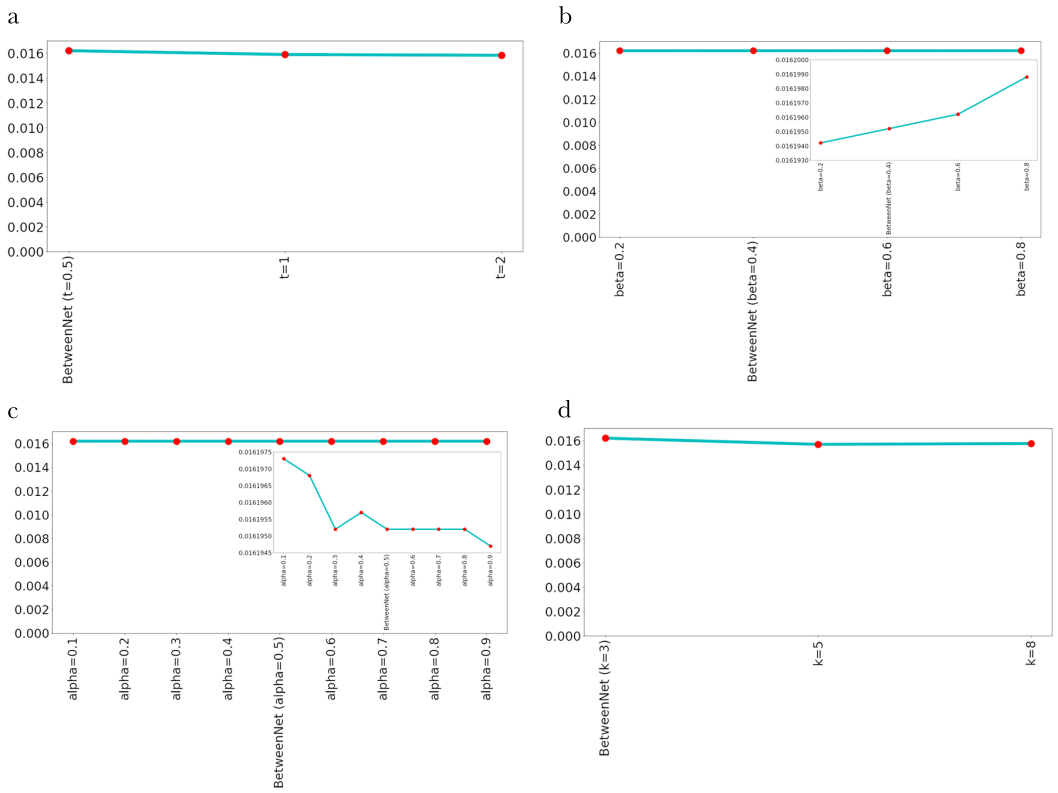

Figure A.5: Sensitivity test of parameters of BetweenNet on lung cancer data when the union of *CGC (Lung)* and *NCG (Lung)* is used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

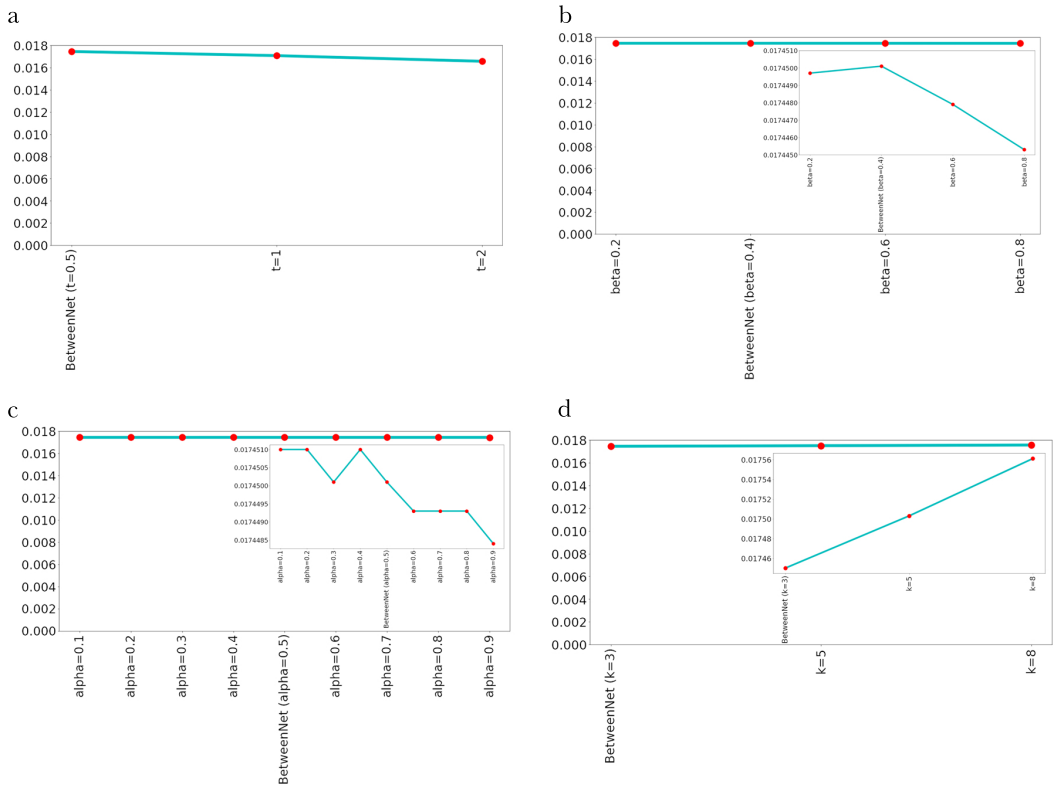

Figure A.6: Sensitivity test of parameters of BetweenNet on lung cancer data when *CancerMine3* is used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

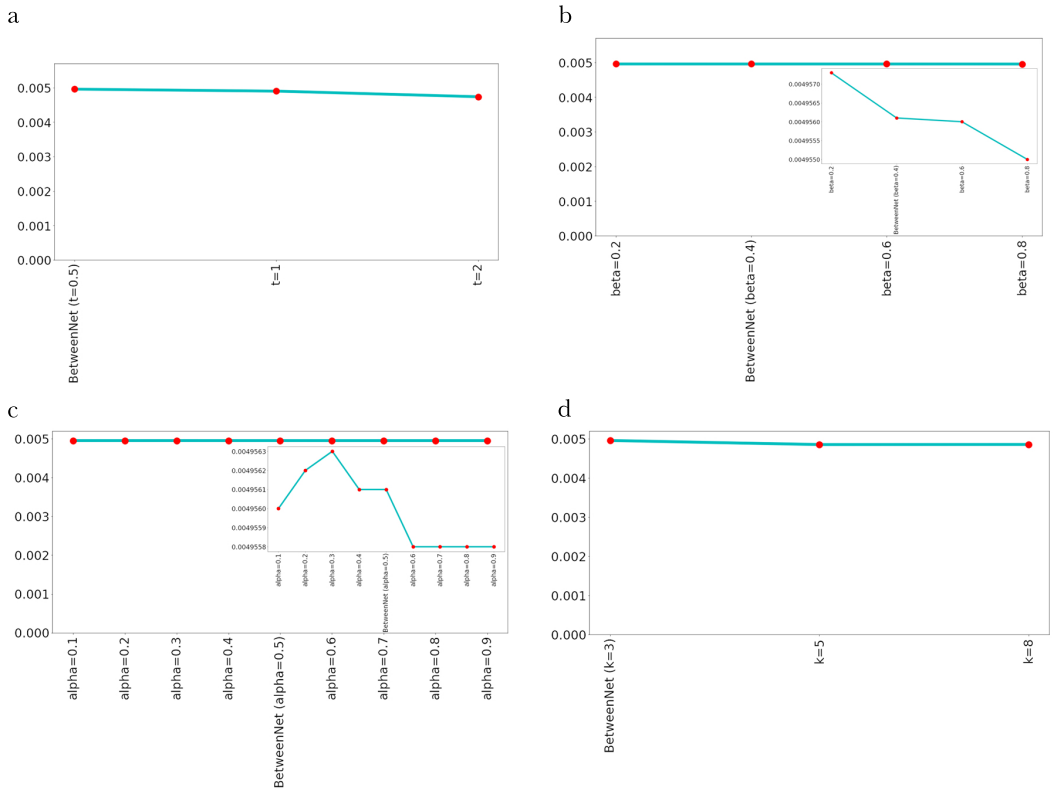

Figure A.7: Sensitivity test of parameters of BetweenNet on lung cancer data when *CGC* rare genes are used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

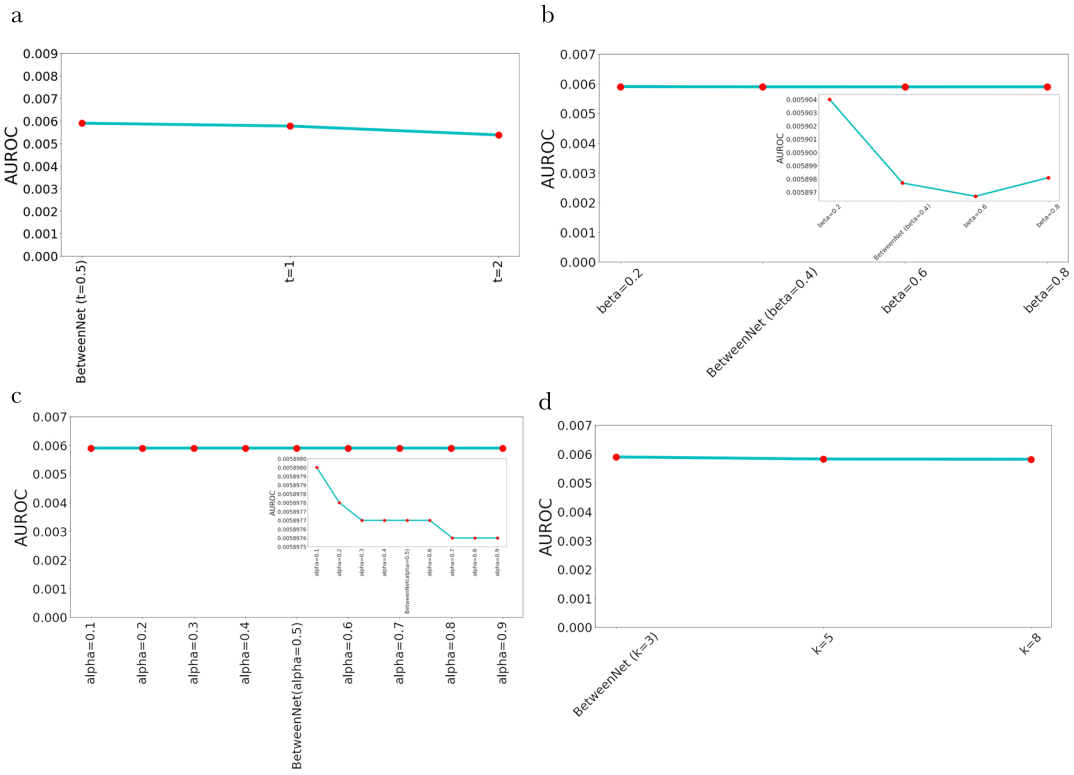

Figure A.8: Sensitivity test of parameters of BetweenNet on breast cancer data when *CGC* is used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

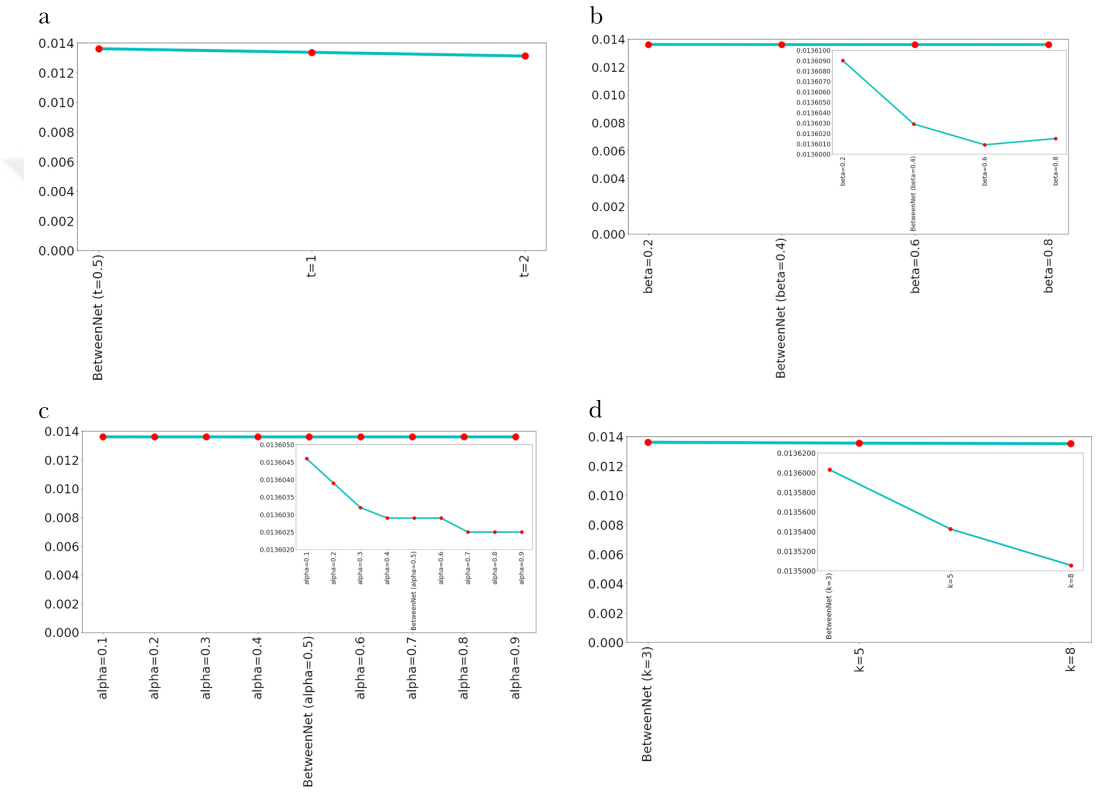

Figure A.9: Sensitivity test of parameters of BetweenNet on breast cancer data when the union of *CGC (Breast)* and *NCG (Breast)* is used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

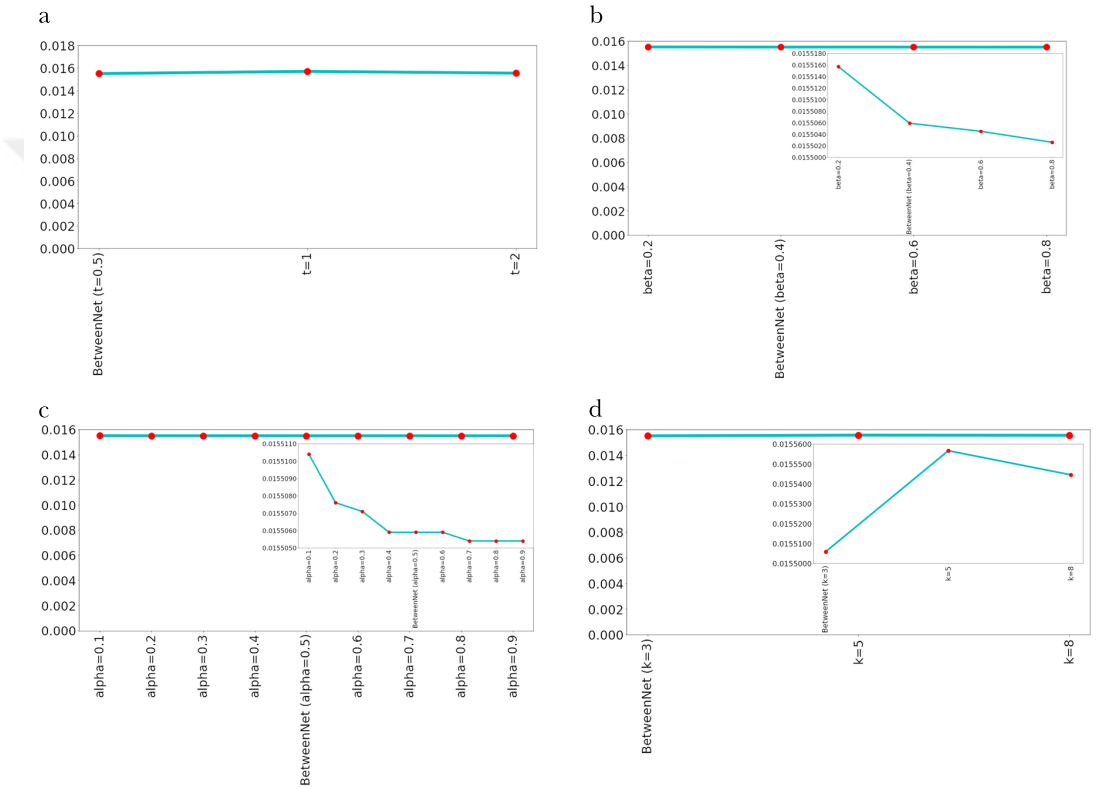

Figure A.10: Sensitivity test of parameters of BetweenNet on breast cancer data when *CancerMine3* is used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

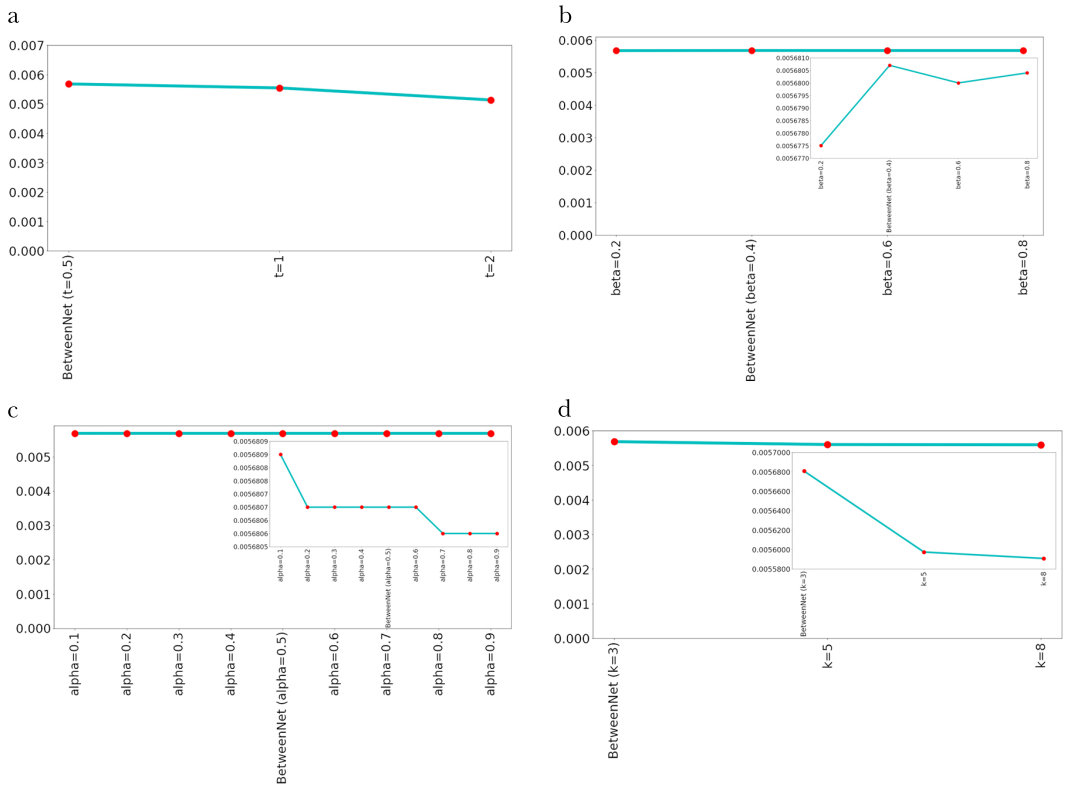

Figure A.11: Sensitivity test of parameters of BetweenNet on breast cancer data when *CGC* rare genes are used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

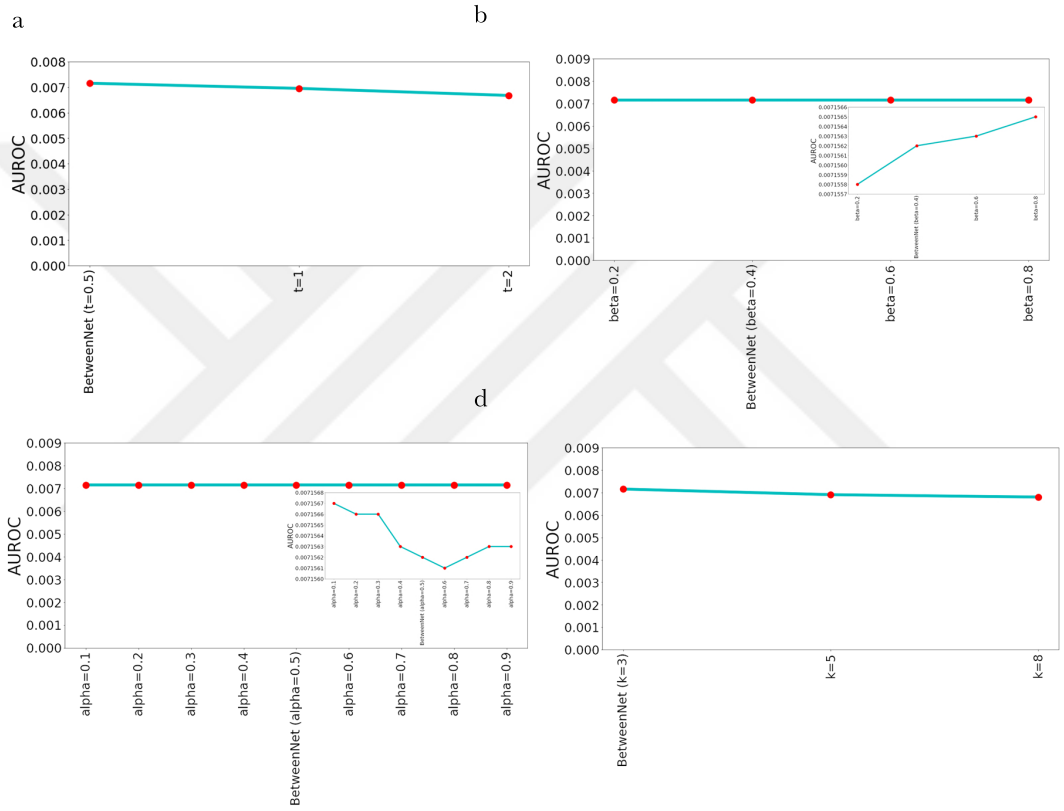

Figure A.12: Sensitivity test of parameters of BetweenNet on pan cancer data when *CGC* is used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

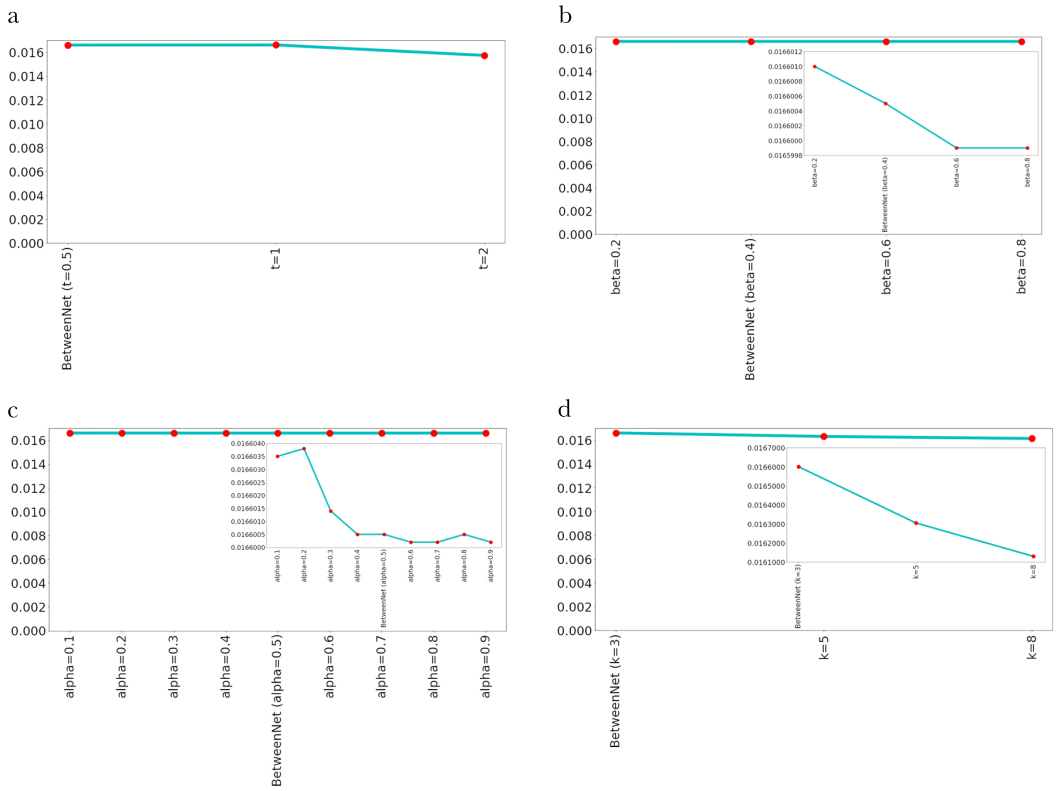

Figure A.13: Sensitivity test of parameters of BetweenNet on pan cancer data when *CancerMine3* is used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

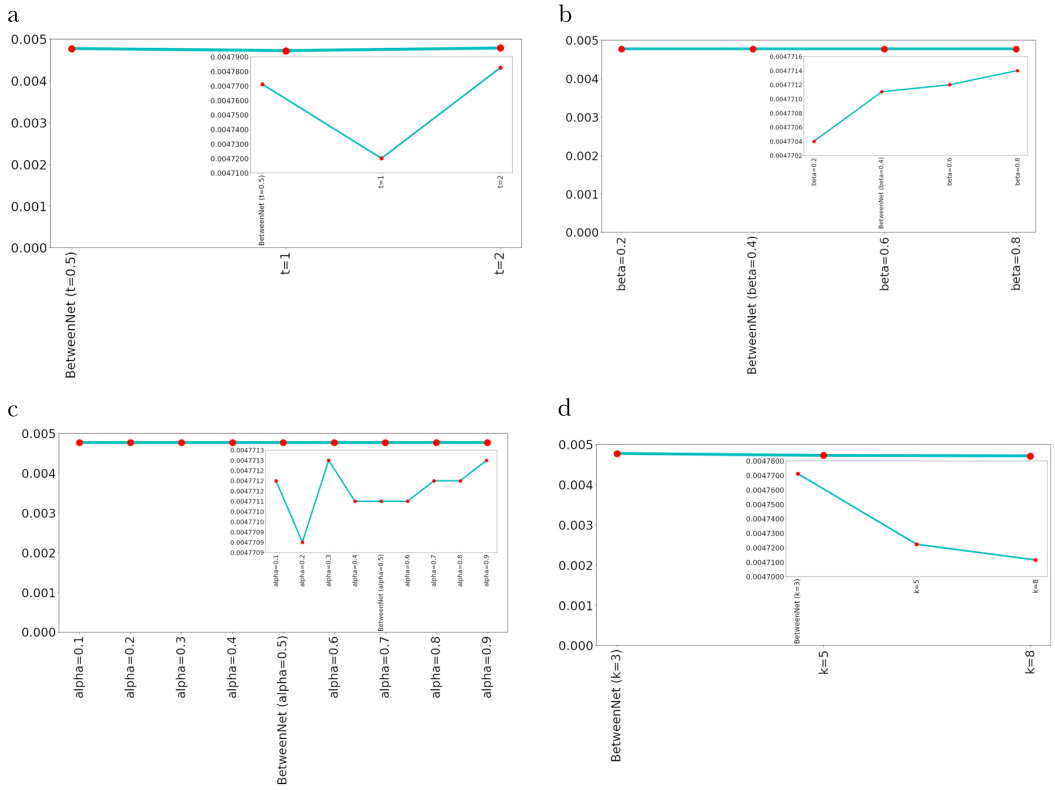

Figure A.14: Sensitivity test of parameters of BetweenNet on pan cancer data when *CGC* rare genes are used as reference. a) AUROC values with different thresholds for defining outlier genes b) AUROC values with different restart probability values (β) c) AUROC values with different alpha values d) AUROC values with different k values.

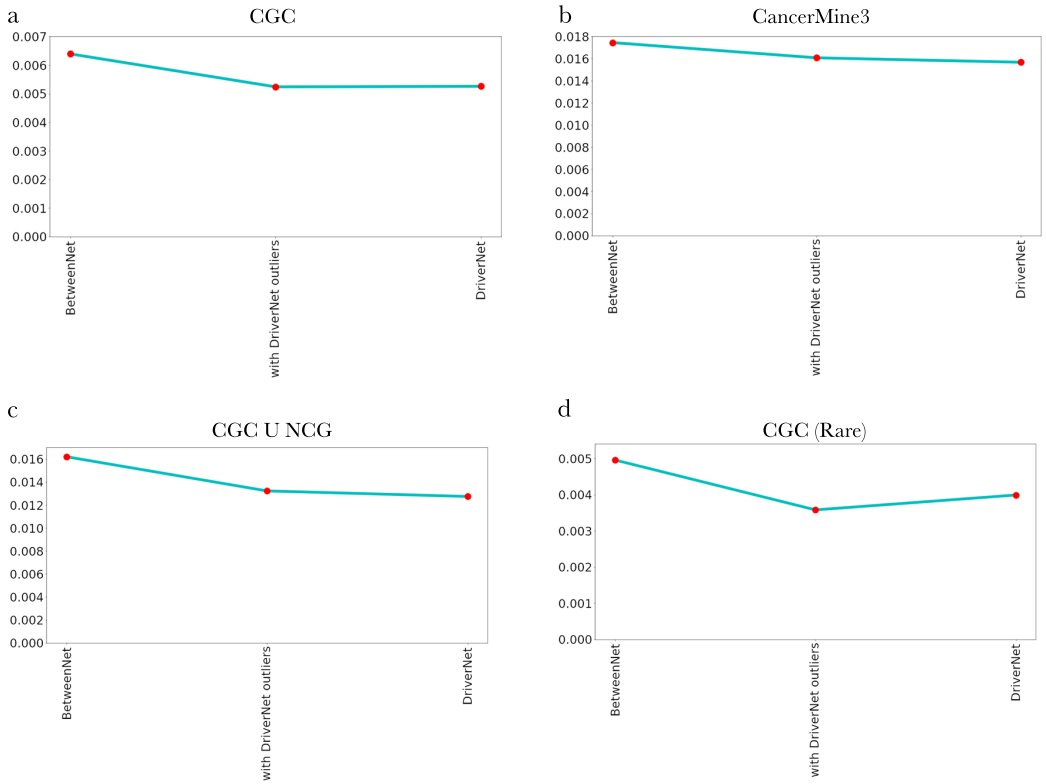

Figure A.15: AUROC values of BetweenNet, DriverNet and a modified version of BetweenNet where outliers are defined according to DriverNet's outlier definition method. where a) *CGC* b) *CancerMine3* c) the union of *CGC (Lung)* and *NCG (Lung)*, d) *CGC rare* genes are used as reference.

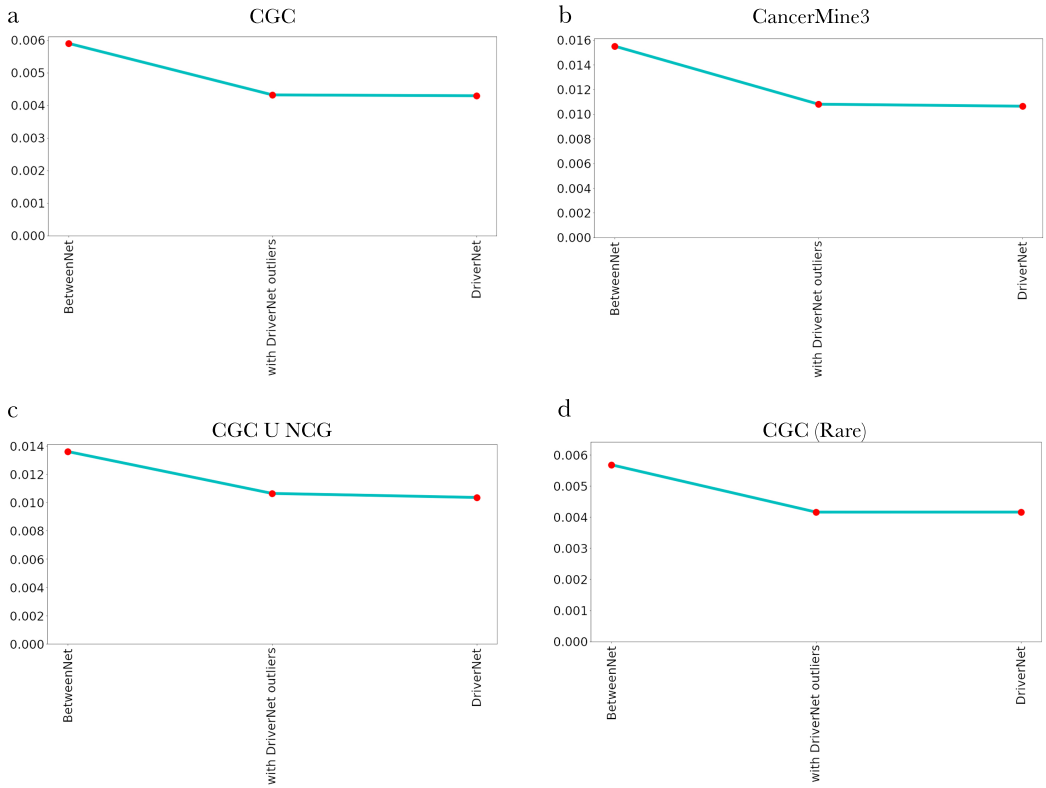

Figure A.16: AUROC values of BetweenNet, DriverNet and a modified version of BetweenNet where outliers are defined according to DriverNet's outlier definition method. where a) *CGC* b) *CancerMine3* c) the union of *CGC (Breast)* and *NCG (Breast)*, d) *CGC rare* genes are used as reference.

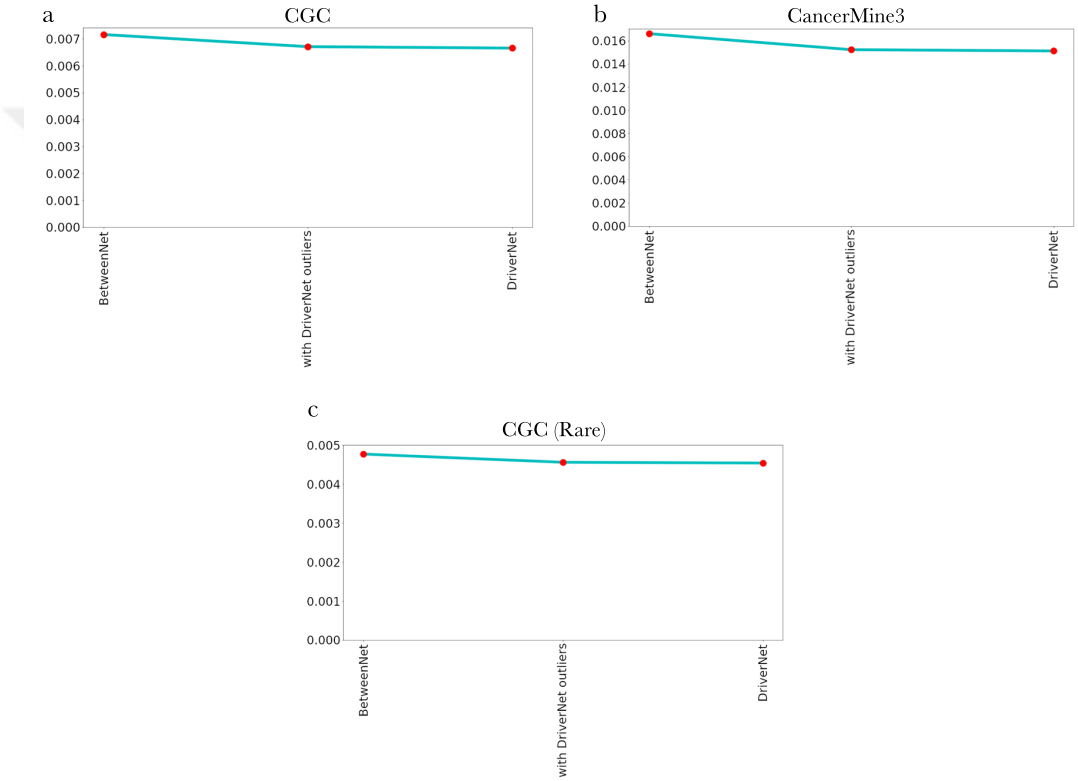

Figure A.17: AUROC values of BetweenNet, DriverNet and a modified version of BetweenNet where outliers are defined according to DriverNet's outlier definition method. where a) *CGC* b) *CancerMine3* and c) *CGC rare* genes are used as reference.

Table A.1: The statistics of top 30 lung cancer driver genes identified by our method

Gene	Rank	CGC	the number of patients with mutations	# of other methods that also rank this gene within the top 30
TP53	1	1	28	4
EGFR	2	1	9	4
TTN	3	0	34	4
LRRK2	4	0	7	4
KEAP1	5	1	11	3
DISC1	6	0	1	2
STK11	7	1	6	2
ATXN1	8	0	2	2
KRAS	9	1	14	3
MET	10	1	6	3
SMAD2	11	1	2	1
PLEC	12	0	4	3
TRAF6	13	0	1	2
NF1	14	1	8	3
RB1	15	1	5	3
IKBKE	16	0	1	1
POT1	17	1	2	2
NLRP12	18	0	6	2
SMAD4	19	1	4	3
RIF1	20	0	4	2
REL	21	1	2	1
ZDHHC17	22	0	2	2
HOXA1	23	0	2	1
FYN	24	0	1	1
APC	25	1	2	1
ANK2	26	0	13	1
MYH9	27	1	5	1
AHNAK	28	0	6	1
IKZF3	29	0	2	1
DYSF	30	0	7	2

Table A.2: The statistics of top 30 breast cancer driver genes identified by our method

Gene	Rank	CGC	the number of patients with mutations	# of other methods that also rank this gene within the top 30
TP53	1	1	37	5
TTN	2	0	20	5
MAP3K1	3	1	12	4
PIK3CA	4	1	34	4
GOLGA2	5	0	2	3
FMR1	6	0	2	2
PICK1	7	0	2	3
ERBB2	8	1	2	1
LZTS2	9	0	3	3
MEOX2	10	0	1	2
RIF1	11	0	5	3
ABL1	12	1	3	4
LRRK2	13	0	2	3
ERBB3	14	1	4	4
DISC1	15	0	1	2
PIK3R1	16	1	3	2
HSP90AB1	17	1	1	1
SETDB1	18	1	3	3
IKBKE	19	0	1	2
ATXN1	20	0	1	2
RB1	21	1	3	2
PLCG1	22	1	3	1
YWHAG	23	0	1	0
EWSR1	24	1	2	0
CDH1	25	1	5	2
APP	26	0	2	2
SRGAP2	27	0	4	2
TSG101	28	0	2	1
DLG1	29	0	3	2
MAGED1	30	0	2	0

Table A.3: The statistics of top 30 pan cancer driver genes identified by our method

Gene	Rank	CGC	the number of patients with mutations	# of other methods that also rank this gene within the top 30
TP53	1	1	201	5
TTN	2	0	187	5
EGFR	3	1	31	4
VHL	4	1	45	5
LRRK2	5	0	34	4
APC	6	1	45	4
GOLGA2	7	0	13	3
CTNNB1	8	1	32	3
ERBB2	9	1	17	3
HLA-B	10	0	10	1
HTT	11	0	15	2
EP300	12	1	21	3
ATXN1	13	0	11	1
SMAD4	14	1	31	3
ERBB3	15	1	21	3
IKBKE	16	0	7	1
PLEC	17	0	21	2
RB1	18	1	33	3
MAP3K1	19	1	21	2
PIK3R1	20	1	15	1
RIF1	21	0	23	2
PIK3CA	22	1	91	4
HSP90AB1	23	1	4	2
NF1	24	1	38	4
DISC1	25	0	3	1
MCC	26	0	7	1
FMR1	27	0	6	0
TNIK	28	0	10	1
TRAF6	29	0	4	0
CSNK2A1	30	0	9	0

Table A.4: Size of lung reference sets

Datasets	Genes	# of enriched GO terms	# of enriched Reactome pathways	# of enriched KEGG pathways
CGC	723	1353	564	144
CGC (Lung) UNCG (LUNG)	96	-	-	-
CancerMine3	58	-	-	-
CancerMine5	33	-	-	-

Table A.5: Size of breast reference sets

Datasets	Genes	# of enriched GO terms	# of enriched Reactome pathways	# of enriched KEGG pathways
CGC	723	1353	564	144
CGC (Breast) UNCG (Breast)	142	-	-	-
CancerMine3	89	-	-	-
CancerMine5	58	-	-	-

Table A.6: Size of pan-cancer reference sets.

Datasets	Genes	# of enriched GO terms	# of enriched Reactome pathways	# of enriched KEGG pathways
CGC	723	1353	564	144
CancerMine3	169	-	-	-
CancerMine5	120	-	-	-