T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

INTEGRATIVE ANALYSIS OF TRANSCRIPTOME DATA AND CELLULAR NETWORKS TO REVEAL MOLECULAR INTERACTIONS OF METASTASIS MECHANISMS IN CANCER

DİLARA UZUNER

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF SCIENCE

DEPARTMENT OF BIOENGINEERING

BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAMME

GEBZE

2020

T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

INTEGRATIVE ANALYSIS OF TRANSCRIPTOME DATA AND CELLULAR NETWORKS TO REVEAL MOLECULAR INTERACTIONS OF METASTASIS MECHANISMS IN CANCER

DİLARA UZUNER

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF SCIENCE

DEPARTMENT OF BIOENGINEERING

BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAMME

THESIS SUPERVISOR

ASSOC. PROF. TUNAHAN ÇAKIR

II. THESIS SUPERVISOR

ASSIST. PROF. DR. PINAR PİR

GEBZE

2020

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KANSER METASTASINDA ETKİLİ OLAN MOLEKÜLER ETKİLEŞİMLERİN ORTAYA ÇIKARILMASI İÇİN TRANSKRİPTOM VERİSİ VE HÜCRESEL AĞLARIN BÜTÜNLEŞİK ANALİZİ

DİLARA UZUNER
YÜKSEK LİSANS TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI
BİYOİNFORMATİK VE SİSTEM BİYOLOJİSİ PROGRAMI

DANIŞMANI
DOÇ. DR. TUNAHAN ÇAKIR
II. DANIŞMANI
DR. ÖĞR. ÜYESİ PINAR PİR

GEBZE
2020

GEBZE TEKNİK ÜNİVERSİTESİ	**YÜKSEK LİSANS JÜRİ ONAY FORMU**

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 03/01/2020 tarih ve 2020/02 sayılı kararıyla oluşturulan jüri tarafından 09/01/2020 tarihinde tez savunma sınavı yapılan Dilara UZUNER'in tez çalışması BİYOMÜHENDİSLİK Anabilim Dalında Biyoinformatik ve Sistem Biyolojisi Programında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE
(TEZ DANIŞMANI) : Doc. Dr. Tunahan Çakır

ÜYE : Dr. Öğr. Üyesi Saliha Durmuş

ÜYE : Prof. Dr. Devrim Gözüaçık

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun
....../....../......... tarih ve ..../........ sayılı kararı.

SUMMARY

Cancer is one of the most common diseases, in 21th century cause death. However, the main cause of death in cancer is not the primary tumor, it is the metastatic tumors that can appear even after decades of primary tumor. Metastatic tumors are more agressive than primary tumors, and they significantly decrease the survival time of the patients. In cancer, the functioning mechanisms of cell are disrupted, causing uncontrolled division of the cells. Therefore, conventional cancer treatments aim to kill dividing cells. Metastatic cells, after separation from the primary tumor, enter a state called "dormancy", where they may stop their division for years and then cause the relapse of disease. Consequently, dormant tumors can not be destroyed with conventional therapies. Discovery of intracellular mechanisms of the pre-metastatic dormant cells has a great importance for the complete treatment of cancer without recurrence. In recent years, mapping of transcriptome data on genome-wide interaction networks to illuminate the mechanisms within the cell is a widely used system biology method in the literature. In this study, transcriptomic data of 7 human and 3 mouse datasets selected from the literature including dormancy and cancer states were mapped to protein-protein interaction networks and gene-regulatory networks to extract subnetworks that are enriched in significantly changed genes. KeyPathwayMiner and BioNet tools were used for these analyses. The genes in the obtained subnetworks were scored based on their number of presence in the protein-protein interaction subnetworks and gene-regulatory subnetworks obtained from each dataset. The genes with high scores have been estimated to have an important role in cancer dormancy mechanism. Finally, drugs that target these important genes were identified through DrugBank-based drug repositioning analysis. The results will contribute to the elucidation of intracellular mechanisms involved in cancer dormancy and to discovery of new drugs to target dormant tumor cells.

Key Words: Bioinformatics, Systems Biology, Transcriptomics, Protein-Protein Interaction Networks (PPIN), Gene-Regulatory Networks (GRN), Metastasis.

ÖZET

Kanser, 21. yüzyılda ölüme sebep olan hastalıkların en önemlilerinden biridir. Fakat kanser hastalarının ölüm nedeni çoğunlukla birincil tümörler değil, hastalığın başlamasından yıllar sonra bile ortaya çıkabilen metastatik tümörlerdir. Bu metastatik tümörler ana tümörden daha agresiftir ve hastaların hayatta kalma sürelerini önemli oranda düşürmektedir. Kanserde hücrenin işleyişindeki mekanizmalar bozularakhücrenin kontrolsüz bölünmesine sebep olur. Bu sebeple, kanser tedavileri bu sebeple bölünen hücreleri öldürmeyi hedeflemektedir. Metastatik hücreler ise ana tümörden ayrıldıktan sonra "uyku hali" denilen bir duruma girerek bölünmelerini yıllarca durdurabilir ve sonrasında hastalığın nüks etmesine sebep olabilirler. Bu sebeple klasik tedaviler ile öldürülememektedirler. Bahsi geçen metastaz öncesi uyku halindeki hücrelerin hücre içi mekanizmalarının çözülmesi, kanser nüks etmeden tamamen tedavi edilebilmesi için büyük önem taşımaktadır. Son yıllarda, hücre içindeki mekanizmaların aydınlatılması için transkriptom verilerinin genom çaplı etkileşim ağlarına haritalanarak anlamlı değişen genlerce zengin alt-ağların bulunması literatürde çokça uygulanan bir sistem biyolojisi yöntemidir. Bu kapsamda, literatürden seçilen 7 insan ve 3 fare verisetinin kanser uyku hali ve kanser durumlarını içeren transkriptomik verileri protein etkileşim ağlarına ve gen regülasyon ağlarına haritalanarak anlamlı değişen genlerce zengin alt-ağlar bulunmuştur. Bu analizler için KeyPathwayMiner ve BioNet araçları kullanılmıştır. Elde edilen alt-ağlardaki genler, farklı verisetlerinden elde edilen protein etkileşim alt-ağları ve gen-regülasyon alt-ağlarında bulunma sayılarına göre skorlanmıştır. Bu skorlamada yüksek skorlu genler tespit edilerek kanser uyku hali mekanizmasında önemli role sahip genler tahmin edilmiştir. Son olarak elde edilen bu genleri hedef alan ilaçlar DrugBank veritabanı kullanılarak ilaç yeniden konumlandırma çalışması ile belirlenmiştir. Sonuçlar, kanser uyku halinde rol alan hücre içi mekanizmaların aydınlatılmasına ve uyku halindeki kanser hücrelerini hedefleyen yeni ilaç hedeflerinin bulunmasına katkı sağlayacaktır.

Anahtar Kelimeler: Biyoinformatik, Sistem Biyolojisi, Transkriptom, Protein-Protein Etkileşim Ağları (PPEA), Gen-Regülasyon Ağları (GRA), Metastaz.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Dr. Tunahan ÇAKIR and co-advisor Dr. Pınar PİR for the continuous support during my master study, and for their patience, motivation, and immense knowledge. Besides my advisor, I would like to thank Dr. Devrim GÖZÜAÇIK, who leads the project and inspired my thesis study.

I gratefully thank the project team for the nice collaboration environment and acknowledge TUBITAK for the financial support (Project Code: 216S489).

I would like to thank my teammates Merve KUTAY, Cemal YILDIZ and Çiğdem Hazal TİMUÇİN for their collaborative and helpful attitute. I would also like to thank Elif EMANETÇİ for her help with BioNet and for her sharing of the mouse protein-protein interaction network with me, Batuhan ÇAKIR for his help with single-cell RNAseq analysis, and İsa YÜKSEL for showing me how to solve things quickly. I am grateful to all members of Computational Systems Biology Group at Gebze Technical University for their collabration and for the fun times we spent together.

I wish to express my sincere thanks to Nurşah ERSEZEN, Buse ÇINAR and my roommates Büşra Nur ÇİÇEK and Betül ERTURAL who opened their houses for me during my master study. Especially, I would like to thank to my close friend Büşra Nur ÇİÇEK for her sincerity and supportive friendship since my undergraduate study. I would also like to thank my best friends Beyza SAĞLAM ÖZBAŞ and Büşra SAVAŞ for their endless patience, love and beautiful friendship.

Finally, I must express my very profound gratitude to my mother Rabiye ÇETİNKAYA, my father Hüseyin UZUNER, my sister Alara UZUNER and my grandmother Hatice ÇETİNKAYA for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

TABLE of CONTENTS

Page

LIST of ABBREVIATIONS and ACRONYMS

Abbreviations and Acronyms		**Explanations**
ACO	:	Ant Colony Optimization
CFSE	:	Carboxyfluorescein Succinimidyl Ester
CTC	:	Circulating Tumor Cell
DTC	:	Disseminated Tumor Cell
ECM	:	Extracellular Matrix
EMT	:	Epithelial to Mesenchymal Transition
FDR	:	False Discovery Rate
GEO	:	Gene Expression Omnibus
GO	:	Gene Ontology
GO:BP	:	Gene Ontology Biological Process
GO:MF	:	Gene Ontology Molecular Function
GR	:	Gene Regulatory
INES	:	Individual Node Exception Strategy
LRC	:	Label Retaining Cells
NCI	:	National Cancer Institute (USA)
PCA	:	Principal Component Analysis
PDX	:	Patient Derived Xenografts
PPI	:	Protein-Protein Interaction
REAC	:	Reactome
RNAseq	:	Ribonucleic Acid Sequencing
scRNA	:	Single Cell Ribonucleic Acid Sequencing
TF	:	Transcription Factor

LIST of FIGURES

LIST of TABLES

1. INTRODUCTION

Cancer is a complex disease, causing many deaths currently. Both genetic, epigenetic and environmental factors are involved in the development of this disease. In cancer, cellular mechanisms are disrupted, and this leads to uncontrolled division of cells. However, the exact mechanisms that underlie the disease are still unknown. On the other hand, most of the deaths caused by cancer are due to cancer metastasis. Metastasis is the spread of cancer from the primary tissue to other parts of the body. Cancer metastasis can occur even years after the treatment of the primary tumor. This is because metastatic cancer cells go into a special condition called “dormancy”. Dormant cells stop their division before metastasizing and protect themselves from immune system and from treatments such as chemotherapy. For this reason, it is important to target the dormant cancer cells for effective treatments.

The use of computational methods in understanding cellular mechanisms is a relatively new but rapidly developing field. Especially with the advances in sequencing technologies, it has become very easy to obtain omic data such as genome, proteome and transcriptome. In addition, many bioinformatics tools are being developed to analyse and annotate the data. Mapping of the transcriptome data on genome-wide cellular networks, such as protein-protein interaction or gene-regulatory networks, to elucidate subnetworks that are enriched in significantly changed genes is one of the methods to analyse omics data. This approach is called subnetwork discovery. There are many different tools developed for subnetwork discovery, which use different algorithms to predict subnetworks. Batra and colleagues in their study [Batra et al., 2017], by comparing subnetwork discovery tools, reported that KeyPathwayMiner (KPM) and Bionet tools led to better results than other tools.

In this study, datasets containing dormant and proliferative samples of different cancer types were downloaded from the Gene Expression Omnibus (GEO) database. Differential expression analysis was performed for each dormant-proliferative data pair. p-values of differential expression analyses were used for subnetwork discovery analyses of each dataset. Subnetwork discovery analyses were performed by KeyPathwayMiner and tools. Then, each gene was scored based on the number of subnetworks they appeared. Genes were filtered based on this score, and high-scoring genes were suggested as drug target candidates. Finally, drugs targeting these candidate genes were identified by drug repositioning.

The main aim of the study is to elucidate cellular mechanisms that cause adaptation of dormant cells to survive and to identify candidate drugs to destroy dormant cancer cells for efficient cancer therapy. Chapter 2 includes literature information on cancer, dormant cancer cells and computational techniques. Chapter 3 explains the datasets and methods that were used in the analyses. Chapter 4 presents results on protein-protein interaction networks and Chapter 5 includes results of gene-regulaory networks. Major results are discussed in Chapter 6, and Chapter 7 includes the conclusions of the study and recommendations for future work.

2.BACKGROUND ASPECTS

2.1. Cancer

Cancer is one of the common causes of deaths in the 21st century. Globally, about 1 in 6 deaths is owing to cancer, according to the World Health Organization's (WHO) 2018 data. The incidence was 18.1 million new cancer cases, and mortality was 9.6 million cancer deaths in 2018 [Bray et al., 2018]. Cancer is a collection of diseases that may have different molecular types depending on the tissue of origin. In brief, cancer is malfunctioning of the cells, which proliferate unrestrainedly. Molecular mechanisms that are responsible for cancer progression are quite complex. Impairments in the immune system, impaired metabolism or some viruses have been shown to cause cancer [Mutch et al., 2018]. At the molecular level, gene mutations and epigenetic alterations play an important role in cancer etiology. Figure 2.1 [Garnham et al., 2019] shows the known hallmarks of cancer and involved major genes.

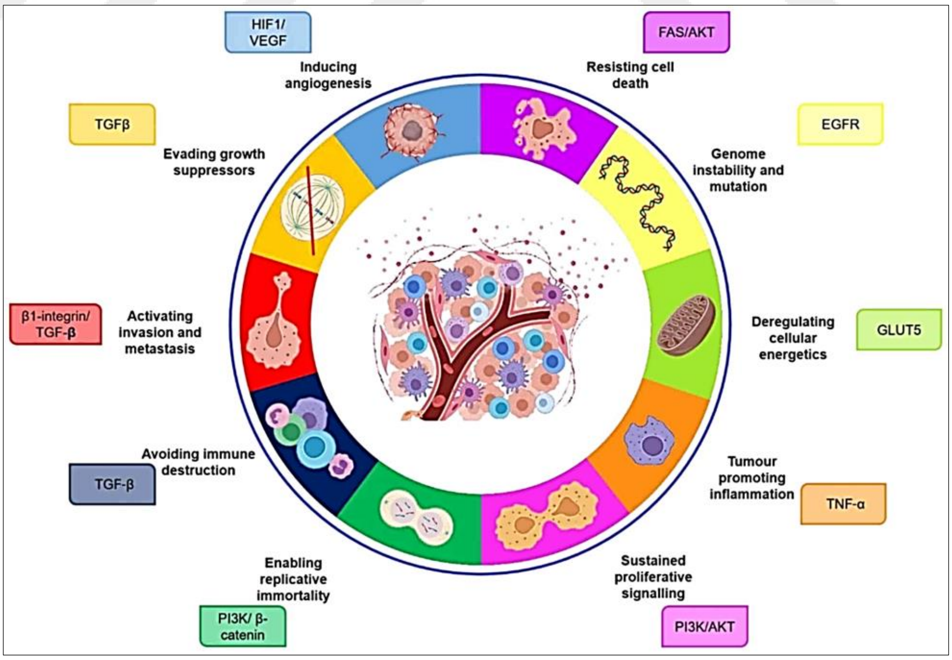


Figure 2.1: Hallmarks of cancer.

As the disease progresses, cancer cells need an appropriate environment for growth and survival. Therefore, they modulate their extracellular matrix (ECM).

Another challenge in cancer progression is hypoxia. When tumor mass increases, oxygen access decreases for inner cells, and they get hypoxic [Ayob and Ramasamy, 2018]. In this case, angiogenesis is activated to form new vessels to feed cancer cells. In cancer cells, autophagy has controversial roles. It has both suppressing and supporting roles in tumor cells [Bishop and Bradshaw, 2018]. Evidence suggests that autophagy suppresses tumor cells in the initiation of disease, but if the disease is progressed autophagy helps cancer cells to survive in stress conditions such as hypoxia by promoting quiescent state [Singh et al., 2018].

Early-diagnosed cancer is largely curable. However, in later stages when cancer has already progressed and metastasized to other tissues, it often shows more aggressive and untreatable profile. Even when cancer is diagnosed at an early stage and treated, residual tumor cells can remain in the body and cause recurrence and metastasis of the disease [Ayob and Ramasamy, 2018].

2.2. Metastasis

The cause of most of the cancer-related deaths is metastasis, which is the spreading of cancer cells to different tissues. Although significant advances have been made in clinical oncology, metastasis remains a major challenge in cancer therapy.

Metastasis mechanisms can be summarized in 7 steps. First step is local invasion of the tumor cells, second one is intravasation from local tissue to blood vessels, third is dissemination in circulation, fourth is arrest at microvasculature of a distant tissue, fifth is extravasation from blood vessel to distant tissue, sixth is micrometastasis and seventh is macrometastasis [Gomis and Gawrzak, 2016]. In the first step, tumor cells proliferate in the primary tissue. In the second step, epithelial to mesenchymal transition occurs [Jie et al., 2017]. One of the most important events in metastasis progression is epithelial to mesenchymal transition (EMT). In EMT, cancer cells change their epithelial character to mesenchymal. It means that they are separated from primary tissue and become motile and, thus, they can diffuse to the blood circulation. EMT allows cancer cells to gain migratory and invasive character. [Liu et al., 2017]. After EMT stage, cancer cells are referred as circulating tumor cells (CTCs). Some of these CTCs are capable of migrating to distant tissues, and they are called disseminated tumor cells (DTCs). The metastatic tissue imposes many stress factors on DTCs such as immune response. Therefore, some of DTCs enter dormancy to be able to survive in the metastatic tissue and adapt the conditions [Dasgupta et al., 2017]. Dormant DTCs in a metastatic tissue constitute the micrometastasis step [Nakanishi et al., 2004]. When micrometastatic cells find a suitable environment, they start proliferation again, and macrometastasis occurs. Figure 2.2 [Aguirre-Ghiso et al., 2004] shows the progression of metastasis.

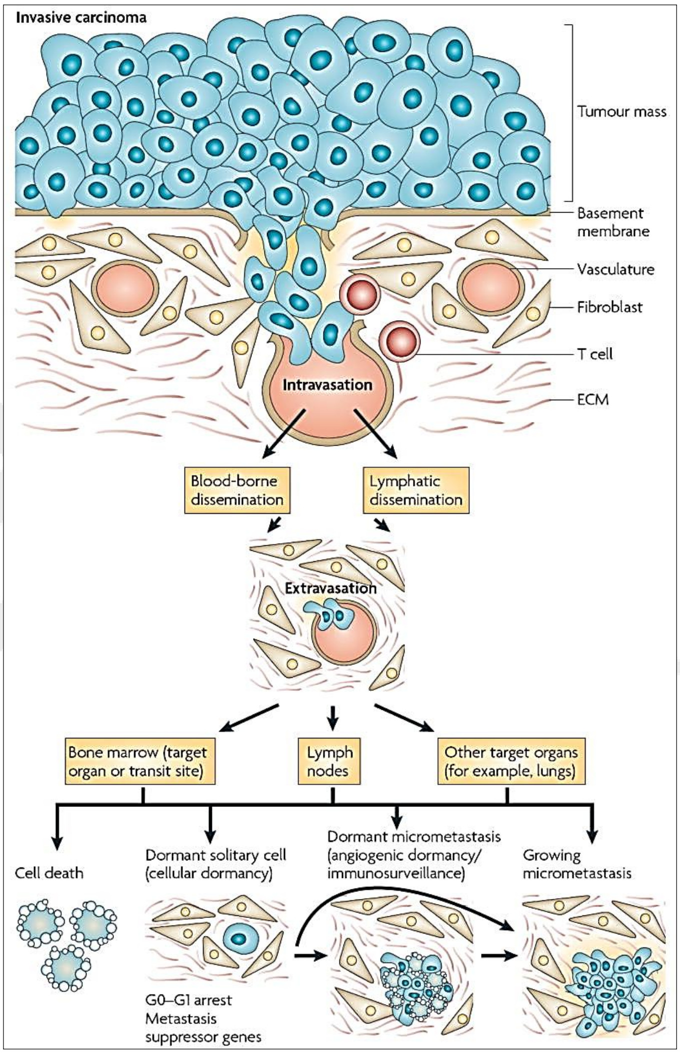


Figure 2.2: Steps of metastasis of invasive carcinoma.

2.3. Cancer Cell Dormancy

Tumor dormancy is a stage of cancer in which cancer cells are present but not clinically detectable [Paez et al., 2012]. Dormancy provides an advantage to cancer cells to escape from treatment and stay asymptomatic for years or even decades after treatment. Cancer cells can stay dormant in primary tumor or they get dormant when they disseminate [Senft and Ronai, 2016]. Cancer cell dormancy is a complex event and there are several underlying mechanisms. Recent studies suggested that there are two mechanisms of cancer cell dormancy. First mechanism is "tumor mass dormancy" and the second one is "cellular dormancy" [Endo and Inoue, 2019]. Tumor mass dormancy also is also divided into two branches based on the form of formation; "angiogenic dormancy" and "immunologic dormancy" (Figure 2.3).

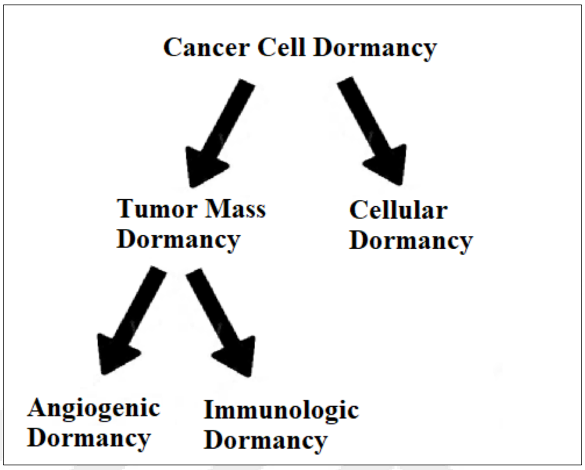


Figure 2.3: Types of cancer cell dormancy.

Besides, cellular dormancy is not only observed in tumor cells in human. Some stem-cell types in human, such as hematopoietic stem cells, also switch from dormancy to proliferative stage naturally to keep their self-renewal potentials [Gomis and Gawrzak, 2016].

2.3.1. Tumor Mass Dormancy

Tumor mass dormancy is defined as the balance between death and proliferation of cancer cells [Endo and Inoue, 2019]. As a result of this balance, tumor mass remains unchanged for a period of time There are two suggested mechanisms; angiogenic dormancy and immunomediated dormancy [Endo and Inoue, 2019]. Since tumor cells need oxygen and nutrients to survive, angiogenesis is important for them. If tumor cells cannot reach blood vessels, they are unable to survive. Angiogenic dormancy is the equilibrium between proliferation and death rate of tumor cells due to lack of blood vessel connection [Endo and Inoue, 2019]. However, recent studies also showed that CTCs mostly seed in oxygen-rich environment. Therefore, until tumor reaches a certain mass there should be other active dormancy mechanisms [Yeh and Ramaswamy, 2015]. The other proposed tumor mass dormancy mechanism is immunomediated dormancy. In this hypothesis, immune system endeavors to destroy tumor cells but it does not succeed completely, and there is an equilibrium between proliferation and death of tumor cells [Paez et al., 2012]. Interestingly, recent studies showed that G0 arrest is seen in not only cellular dormancy but also tumor mass dormancy [Gomis and Gawrzak, 2016]. Figure 2.4 [Yadav et al., 2018] shows the mechanisms of tumor mass dormancy.

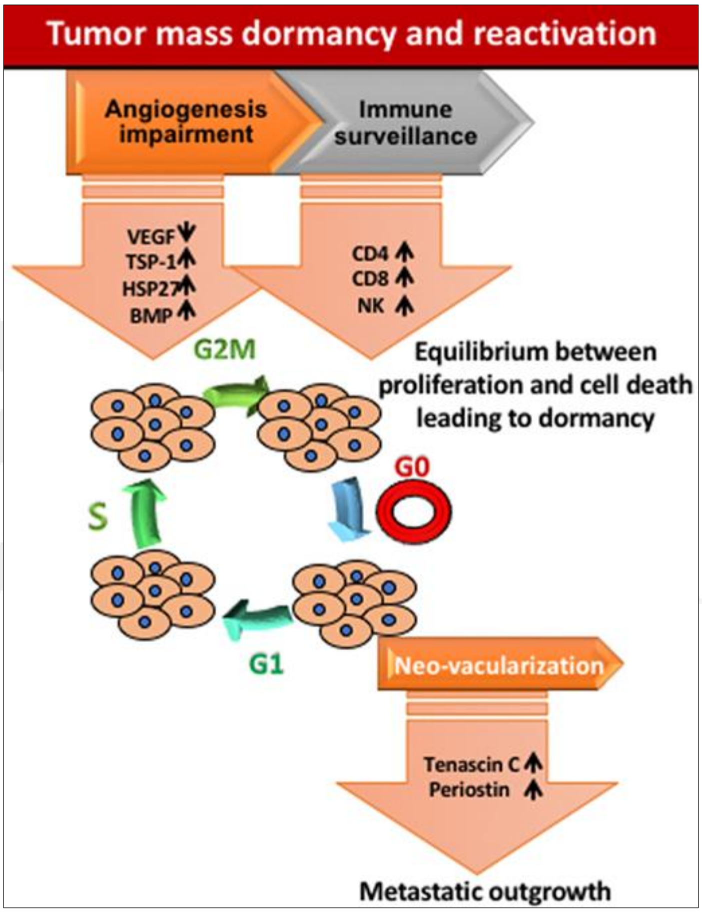


Figure 2.4: Cellular mechanisms of tumor mass dormancy.

2.3.2. Cellular Dormancy

In cellular dormancy, which is also known as solitary cell dormancy, growth and death rates of cancer cells are minimum [Endo and Inoue, 2019]. Cells are arrested in G0/G1 transition of cell cycle. Ki67, which is a cell proliferation marker, is also used as cellular dormancy marker. Cells in cellular dormancy are Ki67 negative because they stop proliferation [Aguirre-Ghiso et al., 2004]. Cells are in a stationary phase but this stationary phase is reversible [Gomis and Gawrzak, 2016]. Cells can exit from G0/G1 arrest and continue to proliferate.

There are several biological pathways that play a role in cellular dormancy. The cell decides whether or not to enter the cell cycle in late G1 phase. Cell cycle progression continues via the dissociation of the transcription factor E2F from Rb in G1 [Yeh and Ramaswamy, 2015]. Studies show that the two signaling pathways that directly activate the cell cycle, Ras-MEK-ERK/MAP and PI(3)K-AKT, are downregulated during cancer cell dormancy [Aguirre-Ghiso et al., 2004, Dey-Guha et al., 2011, Lu et al., 2008]. In addition, signaling through the extracellular matrix is another important pathway for cellular dormancy. Integrins, which are transmembrane receptors, communicate with ECM and activate several intracellular pathways. Inactivation of integrin signaling in metastatic cancer cells stops the proliferation of cancer cells [Shibue and Weinberg, 2009]. Also, stress response such as autophagy stimulates cell dormancy [Yeh and Ramaswamy, 2015]. Figure 2.5 [Yadav et al., 2018] shows the mechanisms of cellular dormancy.

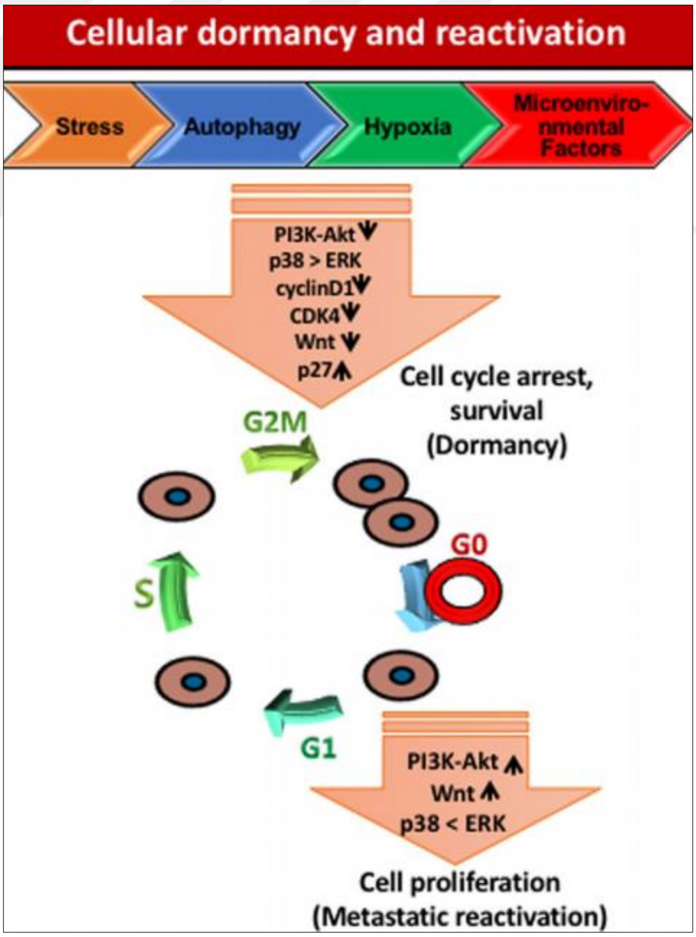


Figure 2.5: Cellular mechanisms of cellular dormancy.

2.4. Transcriptome

Human genome includes gene coding sequences (exomes) and non-coding sequences. Copying of gene encoding regions to create RNA is called transcription. The whole set of transcribed RNA in the cell is called transcriptome. Human genome consists of 3 billion base pairs. Only 30-45 million base pairs within the whole genome correspond to exome, the coding regions. Therefore, transcriptome is a very powerful tool to study the genetic disorders.

There are two major techniques to collect transcriptome data: microarrays and RNA-sequencing. Microarray technology was firstly introduced by Patrick Brown and their colleagues in 1995 [Schena et al., 1995]. Microarray contains tiny spots that include oligonucleotides with known sequences. Due to the hybridization of sample fragments to array, expression levels of sample fragments can be measured. RNA microarrays are used to measure mRNA expression levels to identify disease causing genes, pathways or gene networks [Sealfon and Chu, 2011]. Next-generation sequencing (NGS) allows the sequencing of millions of small fragments in parallel, and gathers these fragments by mapping the individual fragments to the reference genome by using bioinformatics tools [Behjati and Tarpey, 2013]. The use of NGS approach in RNA sequencing (RNA-seq) is another alternative technique to obtain transcriptome data. By RNA-seq, all transcripts in the cell can be profiled and quantified after alignment to the reference genome [Wang et al., 2009]. Recent developments in sequencing technologies and experimental methods enable the quantification of transcriptome of individual cells. This technology is called single-cell RNA-seq (scRNA-seq). Thanks to scRNAseq, transcriptome of rare cell types in tissue population can be quantified.

2.5. Biological Networks

Network-based methods are among frequently used systems biology techniques. Networks consist of nodes and edges. Nodes are the main components of a network, and edges represent the interactions between nodes. Networks are used for the modelling of biological systems [Lotfi Shahreza et al., 2018]. In biological networks, nodes can represent biological molecules such as genes and/or proteins. Edges represent the interactions between these components, such as the regulatory relationship between a gene and a protein, or protein-protein interactions. Number of edges owned by a node is called the degree of the node. In a biological network, some nodes have higher degrees than others. The high-degree nodes are called hub nodes. Since hub nodes interact with many nodes, they are considered as important nodes in the network. Biological networks show scale-free behavior, which means that the number of hub proteins are low and the number of non-hub proteins are high [Fionda, 2019]. Figure 2.6 shows an example of scale free network and network elements.

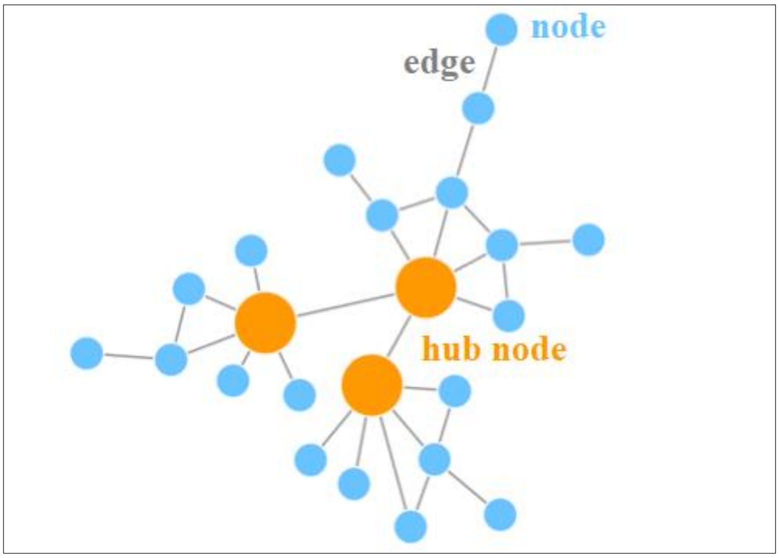


Figure 2.6: Example of a scale-free network.

There are several databases for access to known biomolecular interactions, such as BioGRID [Chatr-aryamontri et al., 2017], KEGG [Kanehisa et al., 2015], IntAct [Hermjakob, 2004], MINT [Chatr-aryamontri et al., 2007], TRRUST [Han et al., 2018], and RegNetwork [Liu et al., 2015]. In these databases, information on biological interactions such as protein-protein interactions or DNA-gene interactions are stored. Information on interactions can be obtained by different techniques such as experimental methods, text mining algorithms or artificial intelligence. Experimental methods can be used to measure physical (direct) or functional (indirect) interactions. In addition, text mining can be used to scan the articles in the literature with machine learning algorithms to detect the interactions reported in articles.

2.5.1. Protein-Protein Interaction (PPI) Networks

A protein-protein interaction network includes the whole set of protein-protein interactions of a given organism [Fionda, 2019]. In PPI networks, nodes represent proteins and edges represent interactions between proteins.

Main techniques to infer PPIs of a given organism are affinity purification followed by mass spectrometry and yeast two hybrid system. Because experimental methods are expensive and they have the possibility of detecting false positive interactions, computational methods are also used to predict PPIs [Huang et al., 2016]. Identified protein-protein interactions are collected in PPI databases for easy access by researchers.

PPI networks play an important role in systemic understanding of cellular and metabolic pathways of organisms to discover new therapies for the treatment of diseases [Folador et al., 2019]. Especially, computational methods that integrate omics data with PPI networks are widely used to discover proteins that are linked to disease etiology.

2.5.2. Gene Regulatory (GR) Networks

Gene regulatory networks are constructed by the connecting the transcription factors (TFs) to their downstream genes. These networks describe the molecular regulation of a specific cell function such as cell cycle control [Halfon, 2017]. In GR networks, nodes represent transcription factors and genes; edges represent regulation between the TF and the gene. In many aspects, GR networks are similar to PPI networks. Gene-regulatory interactions are inferred experimentally or predicted computationally. Widely used experimental methods are CHIPseq and RNAseq [Angelini and Costa, 2014]. Unlike PPI networks, edges in GR networks are directed. Edges originate from TFs and end in the genes regulated by the TFs in question.

In complex diseases such as cancer, regulation mechanisms have crucial roles in disease progression. Therefore, the identification of gene-regulatory networks underlying the disease mechanisms is important for treatment and drug development.

2.6. Integration of Transcriptome Data with Biological Networks

As mentioned in a previous section, with the developing technology, genome and exome data have become very easy to access. For instance, data from more than 120,000 datasets are freely available at the microarray and next-generation sequencing data repository, Gene Expression Omnibus (GEO) [Barrett et al., 2013]. However, currently, the main challenge is processing and interpretation of this very large data. For this purpose, many bioinformatics tools have been developed and continue to be developed based on different systems biology techniques.

Network models that are formed by mapping high-throughput data onto interaction networks are useful tools to analyze omics data in a systems context. The mapping means numerical values that come from statistical analysis of high-throughput data, i.e. p-value of Gene A, are matched with the corresponding nodes in interaction networks, i.e. node of Gene A. By mapping the omics data such as gene expression data on a molecular interaction network, the relations between genes become easy to see [Kedaigle and Fraenkel, 2018]. Therefore, researchers can focus on specific parts of a biological system, which are more affected than others. Also, network models have potential to reveal novel genes that are not differentially expressed, and, therefore, cannot be detected by the analysis of only differentially expressed genes (DEG). For example, to identify markers of metastasis within gene expression data, Chuang et. al. applied protein-interaction-network-based approach in breast cancer studies and reported that known breast cancer related genes, which are typically not detected through differential expression analysis, play a crucial role in the protein network by interconnecting many differentially expressed genes [Chuang et al., 2007].

In literature, different techniques were proposed for integrating omics data with biological networks. One of these techniques is active-module detection through network projection of omics data [Mitra et al., 2013],known also as subnetwork discovery. Bioinformatics tools that are based on this technique maps omics data on large-scale interaction networks to infer functional subnetworks. Some tools that use this technique are BioNet [Beisser et al., 2010], jActiveModules [Cline et al., 2007] and KeyPathwayMiner [Alcaraz et al., 2016]. Yang et. al. applied BioNet to discover subnetworks that are affected when prostate cancer associated transcript 1 (PCAT1) is silenced. They found that PCAT1 plays an important role in the cell cycle and tumorigenesis of prostate cancer cells by mediating the expression of genes that were recently associated with cancer [Yang et al., 2019]. In another study, to understand the role of BORIS gene in ovarian cancer, the authors formed BORIS knockdown and knockout models and analyzed subnetworks using KeyPathwayMiner [Salgado-Albarrán et al., 2019]. BORIS is proposed by NCI as one of the most promising cancer antigens because its deregulation is related to oncogenic properties. BORIS is reported to be regulator of many genes that act as a transcriptional suppressor [Salgado-Albarrán et al., 2019].

2.7. Drug Repositioning

The term "drug repositioning" was firstly introduced by Ashburn and Thor in 2004 [Ashburn and Thor, 2004]. Drug repositioning, also called drug repurposing, is discovering new indication of an existing drug which is different from original medical indication. Drug repositioning is faster, cheaper and safer alternative of traditional de novo drug discovery pipelines [Dudley et al., 2011]. Due to these advantages, drug repositioning studies are gradually increasing in the literature. There are different approaches for drug repositioning in the literature such as computational approches, experimental approaches and mixed approaches [Xue et al., 2018]. Experimental drug repositioning methods are more challenging because mechanisms for most of the drugs are unknown [Moridi et al., 2019]. Computational approaches do not require information on the mechanisms of drugs, and they are faster and cheaper than experimental methods.

Computational drug repositioning approaches aim to find new uses of known drugs using systems biology and bioinformatics tools. Thanks to advances in sequencing technologies, both genomic databases such as GEO [Barrett et al., 2013], GenBank [Clark et al., 2016], PDB [Burley et al., 2018] and disease and drug databases such as KEGG [Kanehisa et al., 2015], DrugBank [Wishart et al., 2018], OMIM [Amberger et al., 2015] have been developed. Consequently, the computational tools for processing the information in these databases are developing rapidly. Thus, computational drug repositioning studies rapidly increased over the last decade. DrugBank is a free-access database that comprise drugs and drug targets. It includes small molecule drugs, approved biologics (proteins, peptides, vaccines, and allergenics) and experimental (discovery-phase) drugs [Wishart et al., 2018]. One approach of drug repositioning is comparing puroposed drug target gene list with DrugBank drug target list. By this method drugs that target proposed genes can be elucidated.

3. METHODS

3.1. Downloading Transcriptome Datasets from GEO

Transcriptome data of proliferative, dormant or post-dormant cancer cells of human or mouse origin were obtained from GEO database [Barrett et al., 2013]. Microarray datasets were downloaded from GEO using R programming language, through "getGEO" function of GEOquery library in BioConductor package in R [Davis and Meltzer, 2007]. This function only needs GSE ID of a transcriptomic dataset to download it. getGEO function has an optional input, "AnnotGPL=TRUE", which can be used to download and implement GPL file to GSE data. GSE code is a unique ID for each dataset entry in GEO. A GSE-coded dataset includes the expression data of all samples in a study and meta-information about that study. In GSE-coded datasets, genes are represented with their probe IDs. To convert theprobe IDs to gene symbols or gene IDs, GPL file is required. GPL ID is a unique ID for each platform entry. Platform records include probe IDs for that particular microarray design and information on corresponding genessuch as gene names, Entrez IDs. RNAseq and single cell RNAseq data were downloaded from GEO database manually by using GSM IDs. GSM ID is a unique ID for each sample in a study. If a dataset has 20 samples, for example, the dataset will have a single GSE ID, and each of the samples will have separate GSM ID. In GEO database, there is an individual webpage for each GSM ID. The GSM file of each sample of an RNAseq dataset was downloaded separately from the supplementary information section of the GSM webpage in GEO database, and they were later combined in an Excel file. The Excel file was read by R, and Ensembl IDs of the genes were converted to Entrez IDs by using "mapIDs" function of "AnnotationDbi" library in R [Pagès et al., 2019]. This function receives 4 inputs. The first input is, the annotation data of the organism for which the genes to be transformed are available. The second input, list of input genes. The third input specifies the gene ID format of the input genes, for example ENSEMBL. The fourth input specifies the desired gene ID format for the output,, for example ENTREZ ID. Once these inputs are given, the function returns the converted list of gene IDs in the desired format.

3.2. Transcriptome Datasets Used in this Study

In order to study the changes of cancer cells in the dormant state, the GEO database was searched for the datasets containing the word "dormancy". The datasets found were filtered to include only RNAseq or microarray data and to be human or mouse data. While the data were chosen to be suitable for statistical analysis, those which did not include both cancer and dormant conditions and those which did not have at least 2 replicates for each conditions were eliminated. A total of 10 datasets were found to meet all these criteria. Table 3.1 gives basic features of the 10 datasets that were chosen from literature for this thesis study. A short information is given below for each dataset.

Table 3.1: Selected transcriptome datasets and their features.

GEO Code	**Organism**	**Cell Type**	**Experimental Method**	**Reference**
GSE35947	*Homo sapiens*	Bladder Cancer	Microarray	[Hurst et al., 2013]
GSE83142	*Homo sapiens*	Acute Lymphoblastic Leukemia	RNAseq, scRNAseq	[Ebinger et al., 2016]
GSE114012	*Homo sapiens*	Colorectal Cancer Spheroids	RNAseq	[Buczacki et al., 2018]
GSE102483	*Homo sapiens*	Acute Myeloid Leukemia	Microarray	[Al-Asadi et al., 2017]
GSE20611	*Homo sapiens*	Breast Cancer	Microarray	[Lu et al., 2011]
GSE64262	*Homo sapiens*	Prostate Cancer	Microarray	[Ruppender et al., 2015]
GSE77379	*Homo sapiens*	Prostate Cancer	Microarray	[Shiozawa et al., 2016]
GSE57695	*Mus musculus*	Myeloma Cells	Microarray	[Lawson et al., 2015]
GSE112094	*Mus musculus*	Breast Cancer	RNAseq	[Vera-Ramirez et al., 2018]
GSE131890	*Mus musculus*	Breast Cancer	Microarray	[La Belle Flynn et al., 2019]

GSE35947 bladder cancer dataset includes two cell lines; JB-V and J82 [Hurst et al., 2013]. The experimental design comprises growth of cell lines on different extracellular matrix material. One of these ECM material is referred as Plastic, which includes collagen. In plastic ECM material, tumor cells are proliferative. Another

ECM material is called SISgel. In SISgel, the malignant phenotype of cancer cells is suppressed, and the cells are in dormant state. Transcriptomic data of the growth of JB-V and J82 cell lines on SISgel and Plastic (dormant and proliferative) material were used in this study. The dataset includes 3 dormant and 3 proliferative samples for each cell line.

GSE83142 dataset includes both RNAseq and single-cell RNAseq data of Acute lymphoblastic leukemia cell line in dormant and proliferative conditions [Ebinger et al., 2016]. In the experimental design of the study, a proliferation-sensitive dye, CFSE, was used in patient-derived xenografts. Human acute lmphoblastic leukemia cells were isolated and stained with proliferation-sensitive dyes. These cells were then implemented to mice. Proliferation-sensitive dyes remain stable in mice for a long time. Loss of the dye was used to distinguish slowly and rapidly growing cells. If dye remain in cells, these cells were called label-retaining cells (LRC) and they were considered as dormant. The LRC undergoes up to three divisions within 21 days. If the dye does not retain cells,these cells called non-label-retaining cells (non-LRC) and considered as proliferative. RNAseq data consists of 5 dormant and 5 proliferative samples, and scRNAseq data consists of 15 dormant and 35proliferative samples.

GSE114012 dataset includes six different colorectal cancer cell lines, which are DLD1, HCT15, HT55, SW948, RKO, and SW48 cell lines [Buczacki et al., 2018]. The cell lines were labeled with dye CFSE and cultured 6 days. After 6 days, dormant cells had remained colored (LRC), but proliferative cells were color-free. Then, RNAseq data analyses were performed for each cell line in both dormant and proliferative conditions. The dataset included 4 proliferative and 4 dormant samples for each cell line.

GSE102483 dataset includes microarray data of acute myeloid leukemia cells in dormant and proliferative conditions [Al-Asadi et al., 2017]. In that study, cancer cells were brought to dormancy by TGFB1 and rapamycin treatment. Non-treated cancer cells were proliferative, and they were used as control. There are 3 dormant and 3 proliferative samples in the dataset.

GSE20611 dataset includes dormant and post-dormant microarray data of breast cancer cells [Lu et al., 2011]. In the experimental design, cells that lack bone metastasis gene signature including genes such as CXCR4, IL11, CTGF, MMP1, and OPN were derived from a single breast cancer cell line, MDA-MB-231, by *in vivo* selection strategy. These single-cell derived populations were unable to produce any bone metastases within 100 days after injection into the mouse. Microarray data of these new cell lines represented dormant state gene expression. After more than 4 months, cells had made metastasis. Microarray data of these post-dormant cell lines represented post-dormancy gene expression. This data includes 3 dormant and 5 post-dormant samples.

GSE64262 dataset includes microarray data of 3 dormant and 3 proliferative samples of patient-derived xenograft (PDX) lines [Ruppender et al., 2015]. In this study, dormant PDX cells were obtained from human bone marrow microenvironment *in vitro*. Then, the dormant cells were re-activated by cell-cell contact on bone marrow stroma.

GSE77379 dataset includes microarray data of two different prostate cancer cell lines [Shiozawa et al., 2016]. These cell lines are PC3 and C42B cell lines. These cell lines were grown *in vitro* both in culture medium and grown *in vivo* by injecting into mice. *In vivo* CSC-free cells were given to the mouse and only the cells leading to the bone marrow were found to be dormant. The dataset includes *in vivo* and *in vitro* proliferative and dormant conditions of both cell lines.In the corresponding article, the authors showed that when disseminated tumor cells arrive at the hematopoietic stem cell (HSCs) ‘niche’ they gain cancer stem cell functions and can switch to dormant state. They selected dormant samples (cancer stem cells) as CD133+/CD44+ cells, and proliferative samples as CD133-/CD44-cells. CD133 and CD44 are surface proteins used as cancer stem cell markers.

GSE57695 dataset includes microarray analysis of mouse-origin multiple myeloma cell line, 5TGM1 [Lawson et al., 2015]. This dataset consists of 4 proliferative (control) and 4 dormant samples. In the experimental design, cells were labeled with DiD dye, which labels cell membrane. After 28 days of labeling, in proliferative cells label tailed off while passing into daughter cells, Therefore DiD-High cells, which retained dye indicated dormant cells, and DiD-Neg cells, which lost dye, indicated proliferative cells.

GSE112094 dataset includes RNAseq analysis of D2.OR murine mammary cell line. D2.OR cells show dormant behavior when growing on basal membrane extracts (BME), and they show proliferative behavior when growing on BME + Collagen. This dataset consists of 3 proliferative and 3 dormant samples.

GSE131890 dataset is a breast cancer microarray dataset. This dataset includes two different breast cancer cell lines, and each cell line was grown separately in 2D and 3D cultures. D2.OR cell line shows proliferative behavior in 2D culture and shows dormant behavior in 3D culture. D2.A1 cell line shows proliferative behavior in 2D culture and metastatic behavior in 3D culture. In the dataset, some samples are from spindle morphology cells and some are from epithelial morphology. Because 3D dormant samples were only collected from spindle form, only samples from spindle morphology were used in this thesis study.

Some datasets include more than two conditions to be compared or more than one cell line. All different pairwise comparisons were analyzed separately. Table 3.2 shows all analyzed pairwise comparisons.. In total, there are 21 comparisons coming from the data from 10 datasets.

Table 3.2: Compared conditions of transcriptome datasets.

GEO Code	Compared Conditions
GSE35947	JB-V cell line, dormant vs proliferative
GSE35947	J82 cell line, dormant vs proliferative
GSE83142	RNAseq analysis, dormant vs proliferative
GSE83142	scRNAseq analysis, dormant vs proliferative
GSE114012	DLD1 cell line, dormant vs proliferative
GSE114012	HCT15 cell line, dormant vs proliferative
GSE114012	HT55 cell line, dormant vs proliferative
GSE114012	RKO cell line, dormant vs proliferative
GSE114012	SW48 cell line, dormant vs proliferative
GSE114012	SW948 cell line, dormant vs proliferative
GSE102483	dormant vs proliferative
GSE20611	dormant vs post-dormant
GSE64262	dormant vs proliferative
GE77379	PC3 cell line *in vivo*, dormant vs proliferative
GE77379	PC3 cell line *in vitro*, dormant vs proliferative
GE77379	C42B cell line *in vivo*, dormant vs proliferative
GE77379	C42B cell line *in vitro*, dormant vs proliferative
GSE57695	dormant vs proliferative
GSE112094	dormant vs proliferative
GSE131890	D2.OR (3D) vs D2.A1 (3D) cell line, dormant vs proliferative
GSE131890	D2.OR (3D) vs D2.OR (2D) cell line, dormant vs proliferative

3.3. Statistical Analysis of the Transcriptome Datasets

After datasets were downloaded from GEO, Principle Component Analysis

(PCA) was performed in R to find if there are any outlier samples. “prcomp” function in R was used to perform PCA analysis. “fviz_pca_var” function of “factoextra” [Kassambara and Mundt, 2019] library was used for the visualization of PCA results.Then, differential expression analysis was applied to all 21 compared conditions to obtain p-values, adjusted p-values and fold changes of genes. p-value shows the possibility of having made a mistake when a gene is chosen as significantly changed. Therefore a cut-off for p-value is needed to classify genes as significantly changed or unchanged. The standard cut-offs used in the literature are 0.01 or 0.05. However, in large data such as RNAseq and microarray, if a standard p-value cut-off is used, the probability of finding false positive results increases due to large number of hypothesis tests applied to the large set of genes. For example, if 3000 significantly changed genes were found with a cut-off value of 0.01, probably 30 of them would have been found to be significantly changed even though they did not change significantly. Therefore, p-values should be adjusted again considering the size of the data set, thus reducing the false positive detection rate in the large data. There are various methods to adjust the p-value. One of the most commonly used p-value adjustment methods is Benjamini-Hochberg, which is also called as FDR [Benjamini and Hochberg, 1995]. In the analyses, the adjusted p-values of both microarray and RNAseq data were calculated using FDR. The R package Limma [Ritchie et al., 2015] was used to identify differentially expressed genes for microarray data. Limma analyses were performed using the GEO2R tool of GEO database, which also generates corresponding R scripts. The R package DESeq2 [Love et al., 2014] was used for the differential expression analysis of RNAseq data. For the differential expression analysis of single-cell RNAseq data, DESeq2 was combined with R package zinbwave [Risso et al., 2018]. Zero-Inflated Negative Binomial-based Wanted Variation Extraction (ZINB-WaVE) considers low-signal level and high-zero read nature of single cell data in the identification of differentially expressed genes. In all datasets, genes that have no gene name information were removed from the datasets. In a microarray design, some genes are represented with multiple probes. Differential expression analysis will result in multiple p-values for those genes. These genes were unified by choosing the smallest p-value for further analysis.

3.4. Cellular Networks

Large-scale human protein-protein interaction (PPI) network, mouse PPI network, human gene-regulatory (GR) network and mouse GR network were used in the thesis for mapping transcriptome data. Human PPI network was downloaded from BioGRID 3.5.166 (November, 2018) [Chatr-aryamontri et al., 2017]. Replicated edges and self-loops of interactome were removed by using Cytoscape. Entrez ID of each gene in the interactome was obtained by using bioDBnet [Mudunuri et al., 2009], and integrated into interactome. Final human PPI network consists of 17,241 edges and 292,471 interactions. Mouse PPI network was obtained from a thesis study in our research group [Emanetçi, 2019], which combined and unified PPI data from BioGRID [Chatr-aryamontri et al., 2017], MINT [Chatr-aryamontri et al., 2007], IntAct [Hermjakob, 2004], UniProt [Consortium, 2018] and iRefIndex [Razick et al., 2008] databases. Replicated edges and self-loops of mouse PPI network were removed, and gene names were converted to Entrez ID as described above. The final mouse PPI network consists of 7.713 nodes and 24,830 interactions. In all PPI networks in this thesis study, proteins were represented by their coding genes.

Genome-scale interactions obtained from databases mentioned above do not include all biological interactions since they are still incomplete. These databases are updated frequently to include newly identified interactions. Therefore, some of the differentially expressed genes in the studied datasets do not have any corresponding interaction information in the PPI networks mentioned above. Within the scope of this thesis, to reduce the missing interaction information, the genes that are available in transcriptome datasets but not available in the organism-specific PPI network were detected. These genes were scored based on the number of times they were identified as significantly changed in dormancy versus control state in 21 pairwase comparisons. This analysis was performed for human and mouse PPI networks separately.

Gene-regulatory networks of both human and mouse were downloaded from TRRUST version2 [Han et al., 2018] and RegNetwork [Liu et al., 2015] databases. These two databases were combined, and duplicate edges and nodes and self-loops were removed. miRNA interactions, if any, were removed. Human GR network in TRRUST database includes 2,862 nodes (genes), 9,396 edges and 795 transcription factors. Human GR network in RegNetwork database includes 5,124 nodes, 12,500 edges and 305 transcription factors. The final human GR network consists of 6,261 nodes, 19,146 edges and 945 transcription factors. Mouse GR network includes 2,439 nodes, 6,464 edges and 826 transcription factors in TRRUST database and mouse GR network includes 2,780 nodes, 7,863 edges and 681 transcription factors in RegNetwork database. The final mouse GR network consists of 4,010 nodes, 13,485 edges and 1,070 transcription factors (Table 3.3).

Table 3.3: Statistics of gene regulatory networks used in this thesis study.

	Human			**Mouse**		
	# of Nodes	# of Edges	# of TFs	# of Nodes	# of Edges	# of TFs
TRRUST	2862	9396	795	2439	6464	826
RegNetwork	5124	12500	305	2780	7863	681
Merged	6261	19146	945	4010	13485	1070
TFless	6048	2530468	-	3767	4622077	-

Transcription factors can modulate target genes in response to a condition change even if their own expression levels remain unchanged. Therefore, TFs have an important role in the network even if they do not change significantly. Subnetwork discovery algorithms scan large networks and find subnetworks that are enriched with significantly changed genes.Since the expression level of transcription factors does not change significantly, they may not be present in the subnetworks. However, lack of TFs, which are the most important elements of gene regulatory networks, in sub-networks may cause missing information. Therefore, TFless GR networks were alternatively reconstructed to avoid missing information in further analyses. In the TFless GR networks reconstructed in this study, all genes that were affected by the same TF were represented to be in interaction with each other, and all TF interactions were removed from the network. In this way, the direct effect of the transcription factor is eliminated, but gene regulation information is protected. Figure 3.1 shows the reconstruction diagram of TFless GR networks. The TFless GR network of human consists of 6048 nodes and 2,530,468 edges. The TFless GR network of mouse consists of 3767 nodes and 4,622,077 edges (Table 3.3).

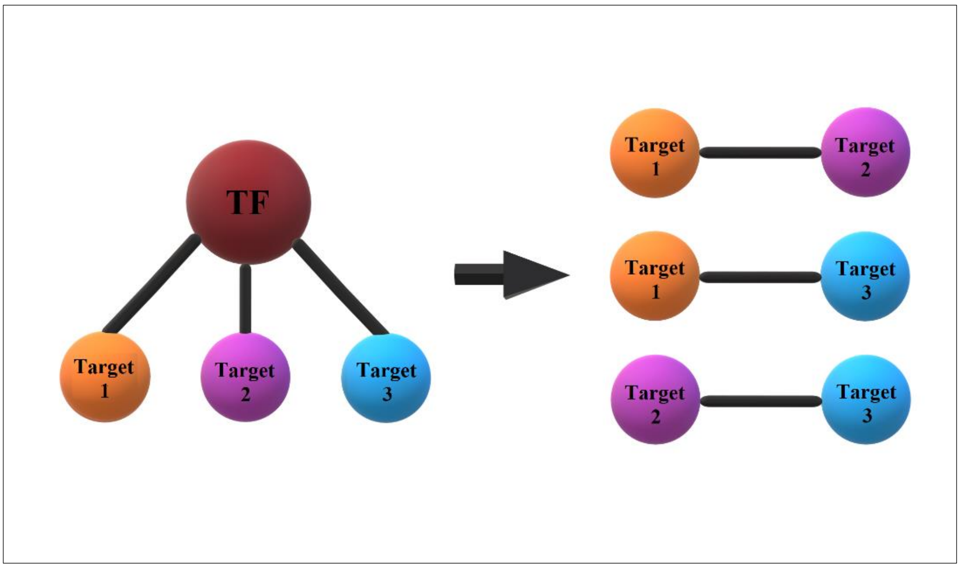


Figure 3.1: Reconstruction of TFless gene-regulatory networks.

3.5. Subnetwork Discovery Methods

Transcriptome datasets provide enormous amount of data. To interpret these data by identifying functional modules, subnetwork discovery methods were used in this study. Two subnetwork discovery tools were selected from literature, KeyPathwayMiner (KPM) and BioNet. As mentioned in Section 2.6, there are several computational tools for subnetwork discovery in literature. Batra et al. performed comparative analyses of different tools to show the power and weakness of these tools. They demonstrated that KPM and BioNet were among the most powerful subnetwork discovery tools [Batra et al., 2017].

KPM and BioNet were used to discover case-specific functional subnetworks by mapping transcriptome data, which are described in Section 3.1, on interactome data described in Section 3.2. Some datasets include more than one comparative conditions as mentioned before. Subnetwork discovery analysis was performed for all compared conditions of datasets given in Table 3.2, and the data were mapped on PPI network, GR network and TFless GR network separately by using KPM and BioNet tools.

3.5.1. KeyPathwayMiner (KPM)

KPM is a subnetwork discovery tool, which maps transcriptome data on interactome to identify subnetworks enriched in significantly changed genes. KPM accepts list of genes and their p-values in binary format as input. The p-values must be binarized based on the significance cut-off. Significant genes were assigned the value of 1, and non-significant genes were assigned the value of 0 for binarization. The adjusted p-values of genes for each compared condition were binarized by using an appropriate cut-off for each comparison. The appropriate cut-off values were chosen such that a similar-size subnetwork will be obtained for all comparisons. The main objective in selecting cut-off values specific to each compared condition is to keep the subnetwork dimensions as close as possible to each other. The cut-off value of each comparison is shown in Table 3.4. The cut-off values of each compared condition remained the same across analyses based on protein-protein interactome and gene-regulatory interactome. In GSE773790, GSE64262, GSE57695 and GSE131890, p-values were used instead of adjusted p-values because KPM could not identify any subnetwork when adjusted p-values with a cut-off of 0.05 was used.

Table 3.4: Adjusted p-value cut-off values for data binarization in KPM based analysis.

GEO Code	**Compared Conditions**	**Cut-off value**
GSE35947	J82 cell line, dormant vs proliferative	0.01
GSE35947	JB-V cell line, dormant vs proliferative	0.01
GSE83142	RNAseq analysis, dormant vs proliferative	0.001
GSE83142	scRNAseq analysis, dormant vs proliferative	0.001
GSE114012	DLD1 cell line, dormant vs proliferative	0.0001
GSE114012	HCT15 cell line, dormant vs proliferative	0.005
GSE114012	HT55 cell line, dormant vs proliferative	0.005
GSE114012	RKO cell line, dormant vs proliferative	0.00001
GSE114012	SW48 cell line, dormant vs proliferative	0.01
GSE114012	SW948 cell line, dormant vs proliferative	0.05
GSE102483	dormant vs proliferative	0.005
GSE20611	dormant vs post-dormant	0.005
GSE64262	dormant vs proliferative	0.01*
GE77379	PC3 cell line *in vivo*, dormant vs proliferative	0.05*
GE77379	PC3 cell line *in vitro*, dormant vs proliferative	0.05*
GE77379	C42B cell line *in vivo*, dormant vs proliferative	0.05*
GE77379	C42B cell line *in vitro*, dormant vs proliferative	0.05*
GSE57695	dormant vs proliferative	0.001*
GSE112094	dormant vs proliferative	0.001

Table 3.4: Continued.

GSE131890	D2.OR (3D) vs D2.A1 (3D) cell line, dormant vs proliferative	0.01*
GSE131890	D2.OR (3D) vs D2.OR (2D) cell line, dormant vs proliferative	0.01*
*p-values were used instead of adjusted p-values		

To discover an optimal subnetwork, KPM uses the following parameters:

1. Algorithm: KPM can use two different algorithms to identify a subnetwork. Greedy algorithm chooses the best choice (shorter path) on every step of iteration, and continues to next step to reach final solution. The greedy algoritm does not consider the previous choices to improve the final solution. [Curtis, 2003]. The other algorithm is Ant Colony Optimization Algorithm (ACO). ACO is based on the principle that a food-seeking ant will find the shortest path based on the level of pheromone on the path [Alcaraz et al., 2012]. In this thesis study, both algorithms were applied on GSE35947 and GSE83142 datasets, and same subnetworks were obtained. Therefore, only the ACO algorithm was used in the analysis of all other comparisons.

2. K value: K is a parameter that specifies the maximum number of exception-genes (not significantly-changed genes) allowed to be present in identified subnetworks. K value is only used as an input when node exception strategy is chosen as INES (individual node exception strategy). Node exception strategy determines how to calculate the presence of exception genes in the subnetwork. INES calculates an individual score for each gene. A higher K value corresponds to a bigger-size subnetwork. To optimize K values, GSE35947 dataset was run with different K values and the appropriate K value was chosen and applied to all analyses. K value was chosen in all PPI and TFless GR network analyses as “2” to keep network sizes optimal and not to ignore non-significant genes that bind many significantly-changed nodes. In all GR network analyses, K value was chosen as “8”. A value higher than the value used in PPI and TFless GR network analyses was chosen, because transcription factors are important nodes in GR network subnetworks and they usually do not show significant expression at mRNA level.

KPM has different versions; Cytoscape plug-in, Web server or stand-alone application. In this study, all alternative versions of KPM were used for GSE35947 dataset, and the stand-alone application was chosen for further use since it is the fastest one. KPM 4.0 version was used in this study.

3.5.2. BioNet

BioNet is another subnetwork discovery tool in the literature. BioNet is an R package and can be used via R. BioNet Bioconductor 3.7 version was used in this thesis study. The working principle is quite similar to KPM. However, the algorithm, parameters, and format that BioNet accepts are different from KPM. BioNet uses Heinz algorithm to find optimum subnetworks after it maps the transcriptome data on interactome. It accepts gene list with their p-values, without any binarization. It uses False Discovery Rate (FDR) cut-off to reconstruct subnetworks rich in significantly-changed genes. FDR is a p-value correction method to eliminate false-positive results. FDR values for each pairwise comparison selected such that similar-size subnetworks will be produced. The compared conditions and FDR values are shown in Table 3.5. The FDR values of each compared condition remained the same across the analyses based onprotein-protein and gene-regulatory interactomes.

Table 3.5: FDR cut-off values used in BioNet analysis.

GEO Code	**Compared Conditions**	**FDR Value**
GSE35947	JB82 cell line, dormant vs proliferative	0.005
GSE35947	JB-V cell line, dormant vs proliferative	0.005
GSE83142	RNAseq analysis, dormant vs proliferative	0.001
GSE83142	scRNAseq analysis, dormant vs proliferative	0.001
GSE114012	DLD1 cell line, dormant vs proliferative	0.0001
GSE114012	HCT15 cell line, dormant vs proliferative	0.005
GSE114012	HT55 cell line, dormant vs proliferative	0.001
GSE114012	RKO cell line, dormant vs proliferative	0.00001
GSE114012	SW48 cell line, dormant vs proliferative	0.005
GSE114012	SW948 cell line, dormant vs proliferative	0.05
GSE102483	dormant vs proliferative	0.001
GSE20611	dormant vs post-dormant	0.001
GSE64262	dormant vs proliferative	0.05
GE77379	PC3 cell line *in vivo*, dormant vs proliferative	0.05
GE77379	PC3 cell line *in vitro*, dormant vs proliferative	0.05
GE77379	C42B cell line *in vivo*, dormant vs proliferative	0.05
GE77379	C42B cell line *in vitro*, dormant vs proliferative	0.05
GSE57695	dormant vs proliferative	0.05
GSE112094	dormant vs proliferative	0.001

Table 3.5: Continued.

GSE131890	D2.OR (3D) vs D2.A1 (3D) cell line, dormant vs proliferative	0.05
GSE131890	D2.OR (3D) vs D2.OR (2D) cell line, dormant vs proliferative	0.05

3.6. Interpretation of Subnetworks

Once the subnetworks are identified through BioNet or KPM algorithms, the main issue is to interpret them and to extract useful information. In this study, three different approaches were used to interpret subnetworks, as explained in detail in the following subsections. The first approach is Gene Set Enrichment Analysis. This approach finds the biological terms enriched in given gene lists. In another approach, the lists of genes that are associated with autophagy, cell cycle, extracellular matrix, and dormancy were extracted. Then, each discovered subnetwork was checked in terms of their overlap with these lists. This approach demonstrates, for example, whether subnetworks include dormancy genes or genes involved in mechanisms known to be effective in dormancy. Finally, genes were scored based on number of times they appeared in the discovered subnetworks as significantly-changed. Here, the aim was to find possible target genes which changed significantly in the case of dormancy and appeared in subnetworks in multiple studies.

3.6.1. Gene Set Enrichment Analysis

After subnetworks were discovered, enrichment analysis was performed by g:Profiler online tool [Raudvere et al., 2019]. The aim of this application is determining whether genes in subnetworks are really related with dormancy mechanism or not. g:Profiler is an online tool. g:GOST tool was used to perform functional enrichment analysis on input gene list. g:GOST retrieves data from Gene Ontology for gene ontology enrichment, from KEGG, Reactome and WikiPathways for pathway enrichment; from miRTarBase for miRNA targets, from TRANSFAC for regulatory motif matches, from Human Protein Atlas for tissue specificity, from CORUM for protein complexes and from Human Phenotype Ontology for human disease phenotypes. Thus, on one run, g:GOST can perform multiple enrichment analyses.

All subnetworks were analyzed separately. Organism option was selected as *Homo sapiens* if subnetwork is from human dataset and *Mus musculus* if subnetwork is from mouse dataset. Other options were left at their default values. The list of genes from each subnetwork was pasted to the "query" tab and organism was selected as 'human' for JB-V and J82 cell lines. Organism option was selected as 'mouse' if the dataset is a mouse dataset.

3.6.2. Dormancy, Autophagy, Extracellular Matrix and Cell Cycle Genes in Subnetworks

Lists of genes that are associated with autophagy, cell cycle, and extracellular matrix terms were downloaded via AmiGO from GeneOntology Database [Carbon et al., 2009]. Each term was quick searched separately. "Genes and gene products" section was chosen on the menu. The organism was filtered as *Homo sapiens* and gene lists were downloaded. Because AmiGO data does not include Entrez IDs, Entrez IDs were obtained by using db2db tool of BioDBnet [Mudunuri et al., 2009] and added to gene lists. Then, all steps were repeated for mouse lists. Only organism filter of AmiGO was changed to *Mus musculus*. The dormancy list was created with the manual curation by Cemal Yıldız (un-published) as follows: Pubmed articles were searched with the following keywords; "cancer", "dormancy" and "metastasis". Abstracts of about 200 articles were scanned manually to find cancer dormancy marker genes in human. Mouse homologs of dormancy genes were identified via Ensembl/BioMart [Kinsella et al., 2011] and used as mouse dormancy marker list. Table 3.6 shows the numbers of genes found in the lists of each term.

Table 3.6: Number of genes in autophagy, cell cycle, extracellular matrix, and dormancy lists.

Term	**Number of genes (Human)**	**Number of genes (Mouse)**
Autophagy	475	378
Cell cycle	1858	1698
Extracellular matrix	5140	712
Dormancy	18	18

The gene lists of dormancy, autophagy, ECM and cell cycle were compared with the node lists of each subnetwork using R. Finally, nodes that are related with these specific terms for each subnetwork were written to CSV (comma separated values) file.

3.6.3. Integration of Subnetworks

To find the most important genes that appeared in most of the subnetworks, a score table was generated. Firstly, all genes that appeared in at least one subnetwork were collected in a list for human and mouse separately. Then, human homologs of mouse genes were identified by using BioMart by Ensembl [Kinsella et al., 2011]. To find human homologs of mouse genes, firstly “EnsemblGenes 97” database and “Mouse genes (GRCm38.p6)” dataset were chosen. Gene lists in the form of as Entrez IDs was imported using the filters on “Input external references ID list” section. Attributes were selected as “Gene name” and “Human gene name” and results were exported as CSV file. Then, human gene names in the file were converted to Entrez IDs again by using BioMart. To this aim, “Human genes (GRCh38.p12)”, was chosen as the dataset. Human gene names were imported using “Gene Names” filters. Attributes were selected as “Gene name” and “NCBI gene accession” and results were exported as CSV. Human Entrez IDs were concatenated to mouse gene lists based on their homologous genes. Then, human and mouse lists were merged by using Human Entrez IDs.

All genes in the merged list were scored based on the number of their appearance in subnetworks. The genes get a score of 1 for each subnetwork in which they are found. And, the final score of a gene is calculated by the sum of its scores. Three different scoring approaches were used. In the first scoring, called “overall score” the gene gets a score of 1 if it is identified with at least one of the BioNet or KPM tools in a comparison. In this scoring, it is not taken into account whether the gene changes significantly in those datasets. In the second scoring, called “significance score” the gene gets a score of 1 if it is identified with at least one of the BioNet or KPM tools in a comparison and should be changed significantly in the related dataset. Finally, genes get score for the number of articles in which they were found. In this scoring, called “article score” a gene gets only one point for all comparisons of a single dataset. For example, a gene gets one point if it is found in both JB-V and J82 subnetworks. In addition , the genes were compared with dormancy, autophagy, ECM and cell cycle lists and get TRUE if it is found in list, get FALSE if it is not. The score table includes genes, significance score, significant upregulation score, significant downregulation score, non-significance score, article score, names of upregulated datasets, names of downregulated articles and whether the gene is a dormancy, autophagy, ECM or cell cycle gene. Integrated score tables were generated for PPI network and GR network analyses separately. Filtering was done after preparing the score table. Genes were filtered if found in at least 2 articles and, significance score was at least 3. If significance score is 3, it must be upregulated in all comparisons, If significance score is 4 it must be upregulated at least in 3 comparisons. If significance score is 6, it must be upregulated in at least in 4 comparisons. If significance score is 7, it must be upregulated at least 5 comparisons. All calculations and rearrangements were performed in R. Then genes in the filtered score table were mapped to related genome scale network on Cytoscape [Shannon et al., 2003]. Thus, interactions between filtered genes were displayed.

When transcriptome data was mapped to interactome, some genes could not be mapped because current interactome does not have these genes as a node. To not to lose the information of these genes, the genes that were identified to be significantly changed at least in one transcriptome dataset were listed separately for human and mouse. Then, as described above, human homolog of mouse genes were found via BioMart and gene lists were combined. After that, each gene was scored according to the number of datasets in which it is significantly-changed. The number of datasets in which a gene is significantly up-regulated and down-regulated and whether the gene was found in dormancy, autophagy, ECM and/or cell cycle gene lists were also written to the score table. This score table was only constructed for PPI network. All calculations and rearrangements were again performed in R.

3.7. Drug Repositioning

Drug repositioning was applied to find the drugs that target the genes that have high scores on integrated subnetwork score lists. Drug repositioning was applied on PPI network, GR network and TFless GR network genes separately.

Firstly, drug-gene interactions were downloaded from “Downloads” section of DrugBank database website. The list includes “Drug Name”, ”Drugbank ID”, “Target Human Gene Symbol”, “ Target Human Gene ID” , “Targeted Organisms” and “ Drug Group” (approved,investigational etc.). Genes in filtered score lists of PPI network, GR network and TFless GR network were used as target gene lists. All lists were compared with drug-gene list of DrugBank on R through “Target Human Gene ID” column by using the “merge” function of R. The drugs that target genes in the filtered score listswere extracted.

4. Protein-Protein Interaction Subnetworks

4.1. Data Preparation and Subnetwork Discovery

In this study, 10 datasets were analyzed to find subnetworks that are enriched with significantly changed genes. These 10 datasets include 21 compared conditions (Table 3.2). Firstly, Principal Component Analysis (PCA) was performed on each compared set to show if data from same conditions clustered together. If there is a sample that clusters with samples from the other condition (outlier), that sample was removed from the set. Figure 4.1 shows the PCA results of J82 and JB-V cell lines of GSE35947 dataset. PCA plots of other datasets were shown in supplementary data.

As seen in Figure 4.1, proliferative and dormant samples of the two cell lines are separated from each other and samples from each group clustered together. Based on PCA results, one control sample from GSE57695 dataset, one LRC (dormant) and one BULK (control) sample from HT55 cell line of GSE114012 dataset were identified as outliers and removed from the datasets.

After outliers were removed, p-values and fold changes were calculated by appropriate tools as mentioned in Section 3.1. For instance, GSE83142 dataset has 2 different experiments, which lead to 2 different compared conditions. One of the experiments is bulk RNAseq of acute lymphoblastic leukemia cell line in dormant and proliferative conditions. Deseq2 package was used to calculate p-values, adjusted p-values and fold changes of this condition. The other experiment is single-cell RNAseq of acute lymphoblastic leukemia cell line in dormant and proliferative conditions. Zinbwave package with Deseq2 package were used to calculate p-values, adjusted p-values and fold changes of this condition. Calculated p-values were used to perform BioNet subnetwork discovery analysis. Also, calculated adjusted p-values were used to perform KeyPathwayMiner (KPM) subnetwork analysis.

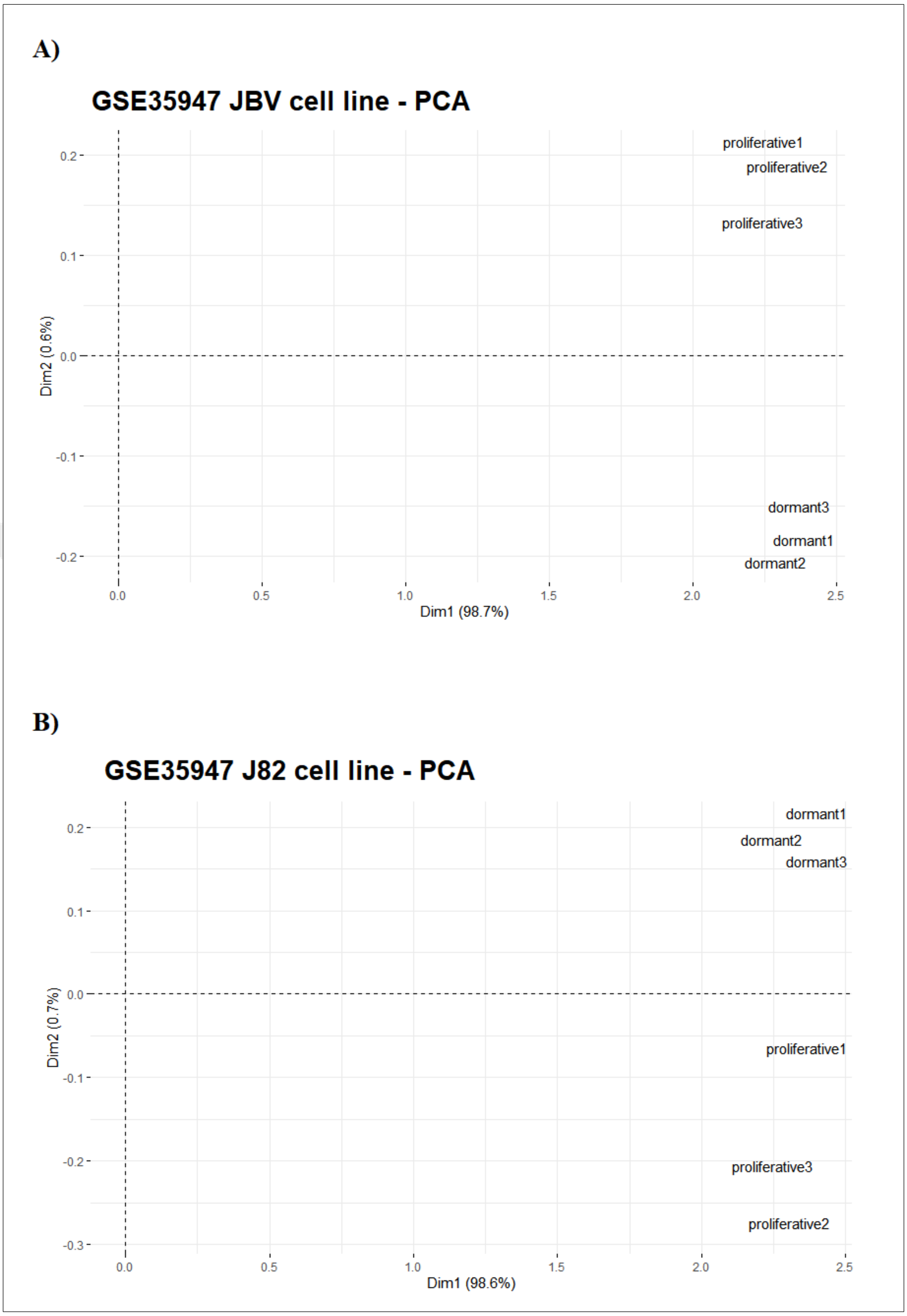


Figure 4.1: PCA graphics of GSE35947 dataset. A) JB-V cell line, B)J82 cell line.

KPM can run on three different platforms; on Cytoscape plug-in, on Web server or as a stand-alone application. The following strategy was followed to decide on the platform of KPM to be used. To confirm that all these 3 tools give the same results, the tutorial data and network of KPM were downloaded from https://keypathwayminer.compbio.sdu.dk/ and used to compare the three versions. All of them gave the same subnetworks with the tutorial data. Then KPM analysis of J82 and JB-V cell lines subnetwork analysis performed on KPM Cytoscape app, KPM web tool and KPM stand-alone version with same parameters. Among these 3 applications, Cytoscape app can not finish the analysis on our computers. Web tool also works slowly. The fastest tool is the stand-alone version. Thus, stand-alone version of KPM was used in further analysis. Then, parameters of KPM were optimized. For this purpose, for J82 and JB-V cell lines the stand-alone version of KPM was applied with different parameters. As a result of the optimization, ACO and GREEDY algorithms gave the same results. In addition, the INES strategy allows to control the subnetwork size since it allows to use the K parameter. The K parameter was determined according to the size of subnetworks. The parameters and the subnetwork sizes are shown in Table 4.1. Based on these results, K=2, INES and ACO parameters were used in the analysis of all protein-protein interaction subnetwork analyses in this thesis study.

Table 4.1: Parameters and subnetwork sizes for parameter optimization of KPM.

J82					**JB-V**				
Strategy	**Algorithm**	**K**	**Node**	**Edge**	**Strategy**	**Algorithm**	**K**	**Node**	**Edge**
INES	ACO	0	78	78	INES	ACO	0	156	346
INES	ACO	1	117	129	INES	ACO	1	170	390
INES	**ACO**	**2**	**139**	**171**	**INES**	**ACO**	**2**	**180**	**414**
INES	ACO	3	157	213	INES	ACO	3	189	449
INES	ACO	4	174	245	INES	ACO	4	196	480
INES	ACO	5	188	286	INES	ACO	5	202	506
INES	GREEDY	0	78	78	INES	GREEDY	0	156	346
INES	GREEDY	1	117	129	INES	GREEDY	1	170	390
INES	GREEDY	2	139	171	INES	GREEDY	2	180	414
INES	GREEDY	3	157	213	INES	GREEDY	3	190	456
INES	GREEDY	4	174	245	INES	GREEDY	4	198	501
INES	GREEDY	5	188	286	INES	GREEDY	5	204	519
GLONE	ACO	-	78	78	GLONE	ACO	-	156	346
GLONE	GREEDY	-	78	78	GLONE	GREEDY	-	156	346

KPM subnetwork discovery analyses of J82 and JB-V cell lines of GSE35947 dataset were performed with the chosen parameters. Binarization cut-off was selected as 0.01 for adjusted p-values for both cell lines. At this cut-off, 339 genes were found to be differentially expressed in J82 cell line and 283 genes differentially expressed in JB-V cell line. 96.74 percent of transcriptome data was mapped to PPI network. Figure 4.2 shows the largest subnetworks discovered by PPI KPM analysis for J82 and JB-V cell lines.

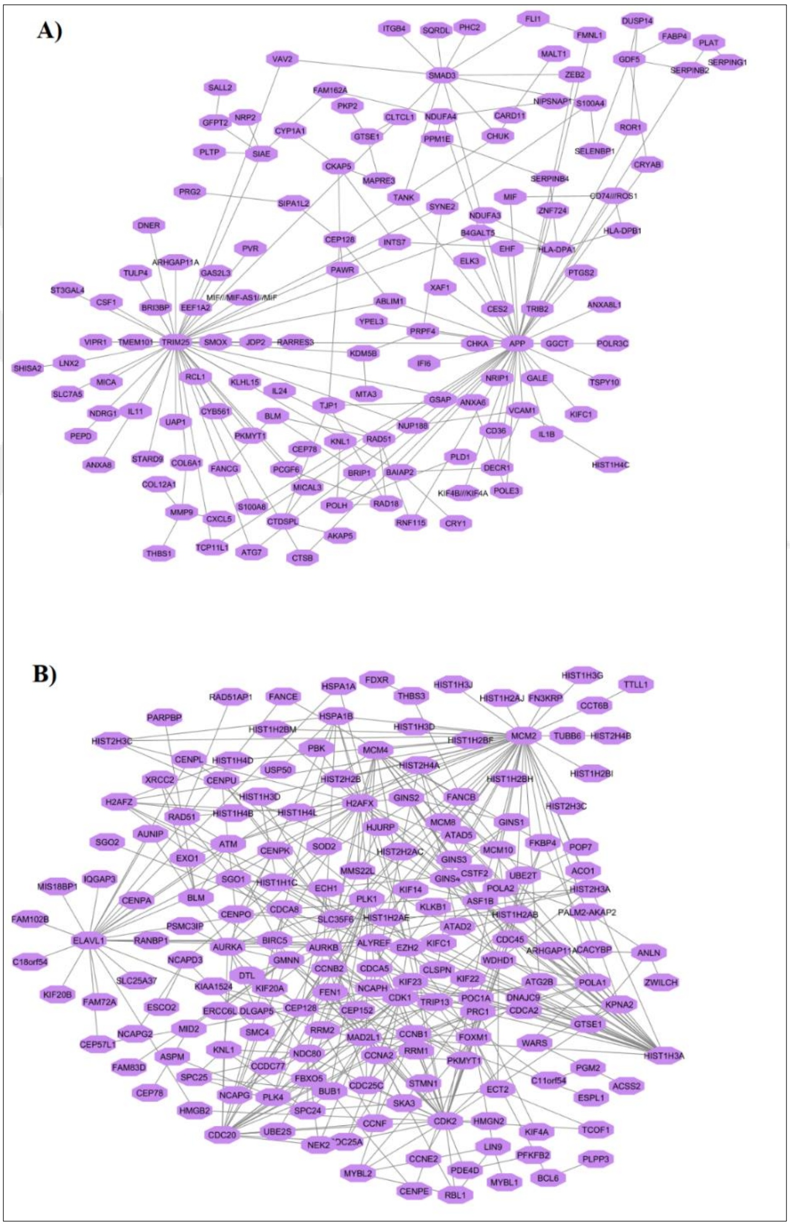


Figure 4.2: KPM PPI subnetworks for dataset GSE35947. A)J82 (139 nodes and 171 edges), B)JB-V (180 nodes and 414 edges)

BioNet subnetwork discovery analysis wasalso performed with the same dataset. FDR value was chosen as 0.005 for both J82 and JB-V cell lines. The network after transcriptome data mapping to PPI network consists of 17221 genes and 605562 interactions. Figure 4.3 shows the subnetworks discovered by BioNet analysis for J82 and JB-V cell lines.

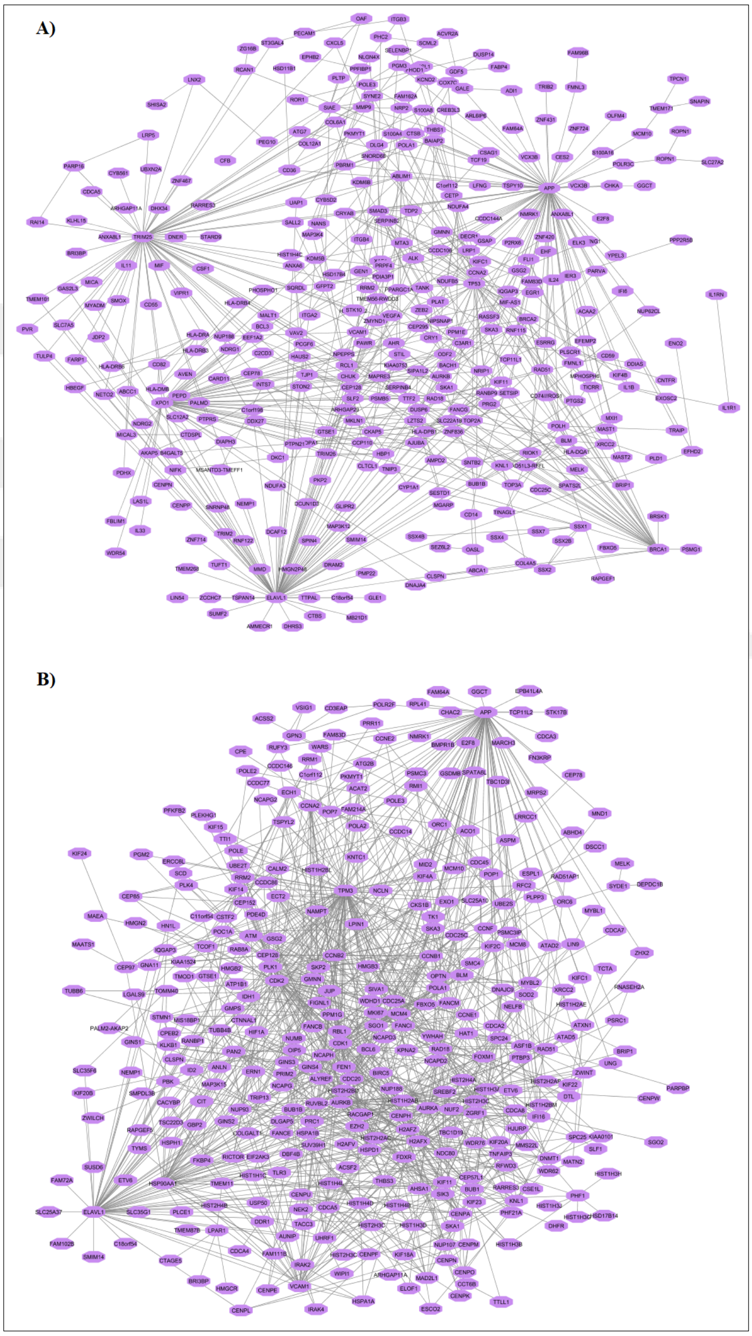


Figure 4.3: BioNet PPI subnetworks for dataset GSE35947. A)J82 (374 nodes and 650 edges), B)JB-V (366 nodes and 1109 edges)

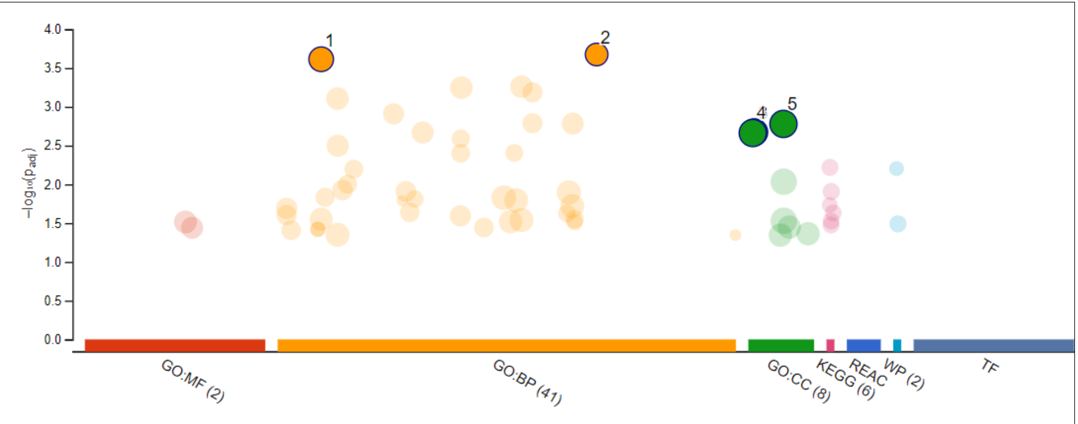
Same workflow was applied to all compared conditions given in Table 3.2 with parameters that were mentioned in Section 3. The corresponding cut-off values used for each comparison are given in Table 3.4 for KPM analysis, and in Table 3.5 for BioNet analysis. For eight of the comparisons, BioNet could not identify any subnetwork. As a result, 34 protein-protein interaction subnetworks were discovered from 21 compared conditions. The sizes of these subnetworks are shown in Table 4.2.

Table 4.2: Subnetwork sizes of PPI subnetworks.

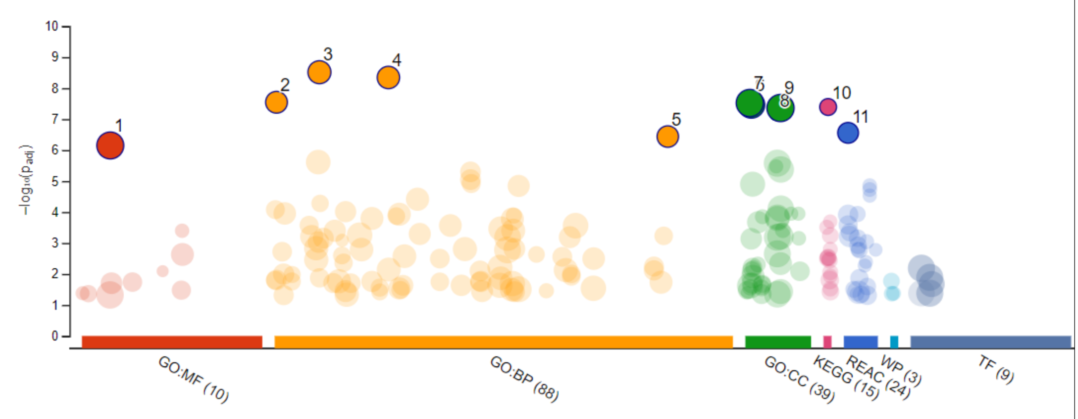
			KPM		BioNet	
#	**GEO Code**	**Compared Conditions**	**Nodes**	**Edges**	**Nodes**	**Edges**
1	GSE35947	J82 cell line, dormant vs proliferative	139	171	374	650
2	GSE35947	JB-V cell line, dormant vs proliferative	180	414	366	1109
3	GSE83142	RNAseq analysis, dormant vs proliferative	113	210	164	378
4	GSE83142	scRNAseq analysis, dormant vs proliferative	119	266	164	495
5	GSE114012	DLD1 cell line, dormant vs proliferative	338	1412	426	3922
6	GSE114012	HCT15 cell line, dormant vs proliferative	78	94	112	362
7	GSE114012	HT55 cell line, dormant vs proliferative	203	311	166	542
8	GSE114012	RKO cell line, dormant vs proliferative	1074	6594	1262	16784
9	GSE114012	SW48 cell line, dormant vs proliferative	70	82	82	216
10	GSE114012	SW948 cell line, dormant vs proliferative	131	162	253	814
11	GSE102483	dormant vs proliferative	90	115	80	204
12	GSE20611	dormant vs post-dormant	340	928	612	5090
13	GSE64262	dormant vs proliferative	144	305	-	-
14	GE77379	PC3 cell line *in vivo*, dormant vs proliferative	383	951	-	-
15	GE77379	PC3 cell line *in vitro*, dormant vs proliferative	765	1611	-	-
16	GE77379	C42B cell line *in vivo*, dormant vs proliferative	235	326	-	-
17	GE77379	C42B cell line *in vitro*, dormant vs proliferative	19	18	-	-
18	GSE57695	dormant vs proliferative	145	184	-	-
19	GSE112094	dormant vs proliferative	178	202	241	336
20	GSE131890	D2.OR (3D) vs D2.A1 (3D) cell line, dormant vs proliferative	83	96	-	-
21	GSE131890	D2.OR (3D) vs D2.OR (2D) cell line, dormant vs proliferative	31	33	-	-

4.2. g: Profiler Analysis of Subnetworks

Enrichment analysis was performed for all subnetworks to determine the enriched pathways and molecular functions of the genes in the subnetworks. g:Profiler online tool is a useful tool to perform enrichment analysis on different lists such as molecular function, pathway, disease relation. To perform enrichment analysis g:GOST tool of g:Profiler was used. At the end of analysis, the tool gives graphical results. In this graph, enriched terms are illusturated as filled circles and plotted according to p-value of enrichment. Also, in detailed results section, the tool gives the list of terms and p-values under the headers of related database. Figure 4.4 and Figure 4.5 shows the graphic results of g:Profiler of JB-V and J82 cell lines for both KPM and BioNet analyses. Table 4.3 shows a summary of g:Profiler results of J82 cell line KPM subnetwork. Full results are shown in supplementary file.

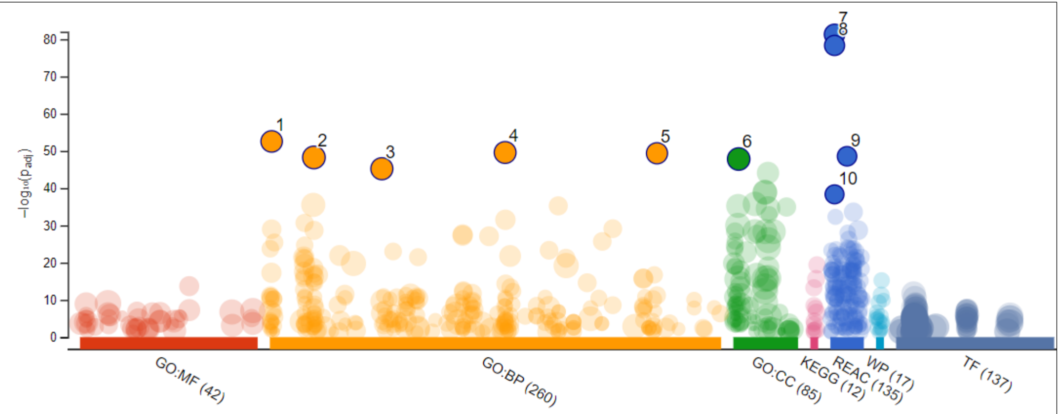


ID	Source	Term ID	Term Name	p_{adj} (J82_KPM)
1	GO:BP	GO:0006950	response to stress	2.434×10^{-4}
2	GO:BP	GO:0080134	regulation of response to stress	2.113×10^{-4}
3	GO:CC	GO:0005737	cytoplasm	2.125×10^{-3}
4	GO:CC	GO:0005622	intracellular	2.178×10^{-3}
5	GO:CC	GO:0044424	intracellular part	1.661×10^{-3}

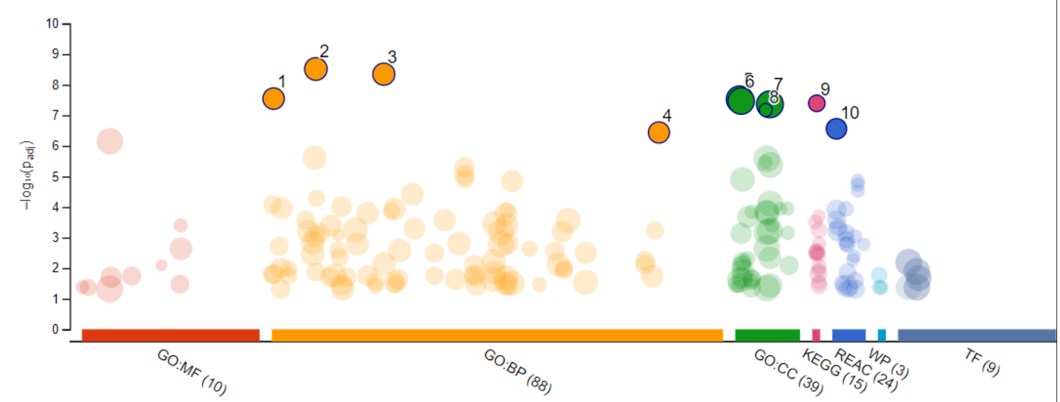


ID	Source	Term ID	Term Name	p_{adj} (J82_BioNet)
1	GO:MF	GO:0005515	protein binding	7.085×10^{-7}
2	GO:BP	GO:0000278	mitotic cell cycle	2.853×10^{-8}
3	GO:BP	GO:0007049	cell cycle	3.056×10^{-9}
4	GO:BP	GO:0022402	cell cycle process	4.523×10^{-9}
5	GO:BP	GO:1903047	mitotic cell cycle process	3.649×10^{-7}
6	GO:CC	GO:0005737	cytoplasm	3.437×10^{-8}
7	GO:CC	GO:0005622	intracellular	3.012×10^{-8}
8	GO:CC	GO:0042613	MHC class II protein complex	6.717×10^{-8}
9	GO:CC	GO:0044424	intracellular part	4.375×10^{-8}
10	KEGG	KEGG:04640	Hematopoietic cell lineage	4.059×10^{-8}
11	REAC	REAC:R-HSA-16...	Cell Cycle	2.773×10^{-7}

Figure 4.4: g:Profiler results of PPI subnetworks of J82 cell line for dataset GSE35947. A) KPM , B) BioNet



ID	Source	Term ID	Term Name	p_{adj} (JB-V_KPM)
1	GO:BP	GO:0000278	mitotic cell cycle	2.763×10^{-53}
2	GO:BP	GO:0007049	cell cycle	5.484×10^{-49}
3	GO:BP	GO:0022402	cell cycle process	5.924×10^{-46}
4	GO:BP	GO:0051276	chromosome organization	2.623×10^{-50}
5	GO:BP	GO:1903047	mitotic cell cycle process	4.193×10^{-50}
6	GO:CC	GO:0005694	chromosome	1.478×10^{-48}
7	REAC	REAC:R-HSA-16...	Cell Cycle	4.940×10^{-82}
8	REAC	REAC:R-HSA-69...	Cell Cycle, Mitotic	4.588×10^{-79}
9	REAC	REAC:R-HSA-68...	M Phase	2.678×10^{-49}
10	REAC	REAC:R-HSA-69...	Cell Cycle Checkpoints	4.470×10^{-39}



ID	Source	Term ID	Term Name	p_{adj} (JB-V_BioNet)
1	GO:BP	GO:0000278	mitotic cell cycle	2.853×10^{-8}
2	GO:BP	GO:0007049	cell cycle	3.056×10^{-9}
3	GO:BP	GO:0022402	cell cycle process	4.523×10^{-9}
4	GO:BP	GO:1903047	mitotic cell cycle process	3.649×10^{-7}
5	GO:CC	GO:0005622	intracellular	3.012×10^{-8}
6	GO:CC	GO:0005737	cytoplasm	3.437×10^{-8}
7	GO:CC	GO:0044424	intracellular part	4.375×10^{-8}
8	GO:CC	GO:0042613	MHC class II protein complex	6.717×10^{-8}
9	KEGG	KEGG:04640	Hematopoietic cell lineage	4.059×10^{-8}
10	REAC	REAC:R-HSA-16...	Cell Cycle	2.773×10^{-7}

Figure 4.5: g:Profiler results of PPI subnetworks of J-BV cell line for dataset GSE35947. A) KPM , B) BioNet

Figure 4.4 and Figure 4.5 show that most of the terms are related with cell cycle, stress response and immune system. These results are expected results because it is known that the cell cycle changes in dormancy. The immune system and stress response are the mechanisms that cells change to survive in dormant condition and adapt metastatic niche.

Table 4.3: g:Profiler results of PPI subnetwork of J82 cell line for dataset GSE35947.

Source	**Term Name**	**Adjusted p-value**
GO:MF	identical protein binding	3.06E-02
GO:MF	protein-containing complex binding	3.64E-02
GO:BP	regulation of response to stress	2.11E-04
GO:BP	response to stress	2.43E-04
GO:BP	response to other organism	5.49E-04
GO:BP	response to external biotic stimulus	5.65E-04
GO:BP	positive regulation of NF-kappaB transcription factor activity	3.93E-03
KEGG	NF-kappa B signaling pathway	6.05E-03
KEGG	TNF signaling pathway	1.24E-02
KEGG	Fanconi anemia pathway	1.84E-02
KEGG	Rheumatoid arthritis	2.34E-02
KEGG	IL-17 signaling pathway	2.98E-02
KEGG	Hematopoietic cell lineage	3.34E-02
WP	Photodynamic therapy-induced NF-kB survival signaling	6.27E-03
WP	Senescence and Autophagy in Cancer	3.24E-02

4.3. Integration of Subnetworks

Subsequently, all genes in all subnetworks were combined to form a new gene list. When forming this list, human homolog of the mouse genes were found via the BioMart online tool and the combining was performed through these homologous genes. Genes without a human homologue remained only as a mouse gene. After this list was made, each gene was scored according to the number of subnetworks it appeared in total. By combining homologous mouse and human genes a priori, scoring could include the total number of occurrences of each gene in both human and mouse subnetworks. The resulting score list contains 4527 genes. Since this number was too high for the interpretation of the genes in the list, the genes were filtered as indicated in section 3.6.3. The filtered score table includes 76 genes based on significance score filter. Table 4.4 shows the filtered score table and Table 4.5 shows a summary of g:Profiler results of filtered gene list.

Table 4.4: Filtered score table of all PPI subnetworks obtained in this study.

Human ID	Human Gene Symbol	Significance Score	Overall Score (significant or not)	Article Score	Dormancy genes?	Autophagy genes?	ECM genes?	Cell Cycle genes?
1191	CLU	9	9	4	FALSE	TRUE	TRUE	FALSE
3006	HIST1H1C	6	6	4	FALSE	FALSE	FALSE	FALSE
351	APP	6	16	8	FALSE	FALSE	TRUE	TRUE
8349	HIST2H2BE	6	6	4	FALSE	FALSE	TRUE	FALSE
10133	OPTN	6	6	4	FALSE	TRUE	FALSE	TRUE
114907	FBXO32	6	6	2	FALSE	FALSE	FALSE	FALSE
5329	PLAUR	5	5	3	FALSE	FALSE	TRUE	FALSE
6187	RPS2	5	5	3	FALSE	FALSE	TRUE	FALSE
7057	THBS1	5	5	3	TRUE	FALSE	TRUE	TRUE
718	C3	5	5	4	FALSE	FALSE	TRUE	FALSE
10769	PLK2	5	5	2	FALSE	TRUE	FALSE	TRUE
23710	GABARAPL1	5	5	2	FALSE	TRUE	TRUE	FALSE
1030	CDKN2B	5	5	3	FALSE	FALSE	TRUE	TRUE
4035	LRP1	5	5	4	FALSE	FALSE	TRUE	FALSE
1508	CTSB	5	5	3	FALSE	FALSE	TRUE	FALSE
2634	GBP2	5	5	3	FALSE	FALSE	FALSE	FALSE
1604	CD55	4	4	2	FALSE	FALSE	TRUE	FALSE
3140	MR1	4	4	2	FALSE	FALSE	TRUE	FALSE
9572	NR1D1	4	4	2	FALSE	FALSE	FALSE	FALSE
9709	HERPUD1	4	4	2	FALSE	FALSE	FALSE	FALSE
57568	SIPA1L2	4	4	2	FALSE	FALSE	FALSE	FALSE
59	ACTA2	4	4	3	FALSE	FALSE	TRUE	FALSE
604	BCL6	4	4	3	FALSE	FALSE	FALSE	TRUE
2879	GPX4	4	4	3	FALSE	FALSE	TRUE	FALSE
3162	HMOX1	4	4	3	FALSE	TRUE	TRUE	FALSE
3383	ICAM1	4	4	3	FALSE	FALSE	TRUE	FALSE
6318	SERPINB4	4	4	3	FALSE	FALSE	TRUE	FALSE
64061	TSPYL2	4	4	3	FALSE	FALSE	FALSE	TRUE
972	ROS1	4	4	4	FALSE	FALSE	TRUE	FALSE
7422	VEGFA	4	4	4	FALSE	FALSE	TRUE	FALSE
10014	HDAC5	4	4	4	FALSE	FALSE	FALSE	FALSE
1893	ECM1	4	4	2	FALSE	FALSE	TRUE	FALSE
5252	PHF1	4	4	2	FALSE	FALSE	FALSE	FALSE
6520	SLC3A2	4	4	2	FALSE	FALSE	TRUE	FALSE
7077	TIMP2	4	4	2	FALSE	FALSE	TRUE	TRUE
9619	ABCG1	4	4	2	FALSE	FALSE	TRUE	FALSE
1917	EEF1A2	4	4	3	FALSE	TRUE	FALSE	FALSE
4088	SMAD3	4	4	3	FALSE	FALSE	TRUE	TRUE
8878	SQSTM1	4	4	3	FALSE	TRUE	TRUE	FALSE
10765	KDM5B	4	4	3	FALSE	FALSE	FALSE	FALSE

Table 4.4: Continued.

94241	TP53INP1	4	4	3	FALSE	TRUE	FALSE	TRUE
960	CD44	4	4	4	FALSE	FALSE	TRUE	FALSE
5360	PLTP	4	4	4	FALSE	FALSE	TRUE	FALSE
301	ANXA1	3	3	2	FALSE	FALSE	TRUE	TRUE
360	AQP3	3	3	2	FALSE	FALSE	TRUE	FALSE
378	ARF4	3	3	2	FALSE	FALSE	TRUE	FALSE
780	DDR1	3	3	2	FALSE	FALSE	TRUE	FALSE
2026	ENO2	3	3	2	FALSE	FALSE	TRUE	FALSE
2934	GSN	3	3	2	FALSE	FALSE	TRUE	FALSE
3248	HPGD	3	3	2	FALSE	FALSE	TRUE	TRUE
4535	MT-ND1	3	3	2	FALSE	FALSE	FALSE	FALSE
4854	NOTCH3	3	3	2	FALSE	FALSE	TRUE	FALSE
9462	RASAL2	3	3	2	FALSE	FALSE	FALSE	FALSE
10397	NDRG1	3	3	2	FALSE	FALSE	TRUE	FALSE
10580	SORBS1	3	3	2	FALSE	FALSE	FALSE	FALSE
23135	KDM6B	3	3	2	FALSE	FALSE	FALSE	FALSE
26470	SEZ6L2	3	3	2	FALSE	FALSE	FALSE	FALSE
57062	DDX24	3	3	2	FALSE	FALSE	FALSE	FALSE
64393	ZMAT3	3	3	2	FALSE	FALSE	FALSE	FALSE
201895	SMIM14	3	3	2	FALSE	FALSE	FALSE	FALSE
1043	CD52	3	3	3	FALSE	FALSE	TRUE	FALSE
1628	DBP	3	3	3	FALSE	FALSE	FALSE	FALSE
3119	HLA-DRB6	3	3	3	FALSE	FALSE	FALSE	FALSE
3939	LDHA	3	3	3	FALSE	FALSE	TRUE	FALSE
3956	LGALS1	3	3	3	FALSE	FALSE	TRUE	FALSE
5802	PTPRS	3	3	3	FALSE	FALSE	TRUE	FALSE
6175	RPLP0	3	3	3	FALSE	FALSE	TRUE	FALSE
6205	RPS11	3	3	3	FALSE	FALSE	TRUE	FALSE
7832	BTG2	3	3	3	FALSE	FALSE	TRUE	TRUE
23521	RPL13AP5	3	3	3	FALSE	FALSE	FALSE	FALSE
57326	PBXIP1	3	3	3	FALSE	FALSE	FALSE	FALSE
84162	KIAA1109	3	3	3	FALSE	FALSE	FALSE	FALSE
91010	FMNL3	3	3	3	FALSE	FALSE	FALSE	FALSE
253635	GPATCH11	3	3	3	FALSE	FALSE	FALSE	FALSE
7316	UBC	3	8	6	FALSE	FALSE	TRUE	FALSE

Genes in the filtered gene list are mainly associated with extracellular matrix and cell cycle. These two terms are known to play a major role in dormancy. Besides, genes associated with the autophagy mechanism, which is thought to be important in the dormancy mechanism, have been found. In addition, only one gene was directly associated with dormancy. This may be because the manually curated dormancy gene list contains only 18 genes. In g: profiler analysis, terms such as cell communication and autophagy in cancer were found in correlation with these results.

Table 4.5: g:Profiler result of filtered gene list of all PPI subnetworks.

Source	**Term Name**	**Adjusted p-value**
GO:MF	enzyme binding	5.03E-04
GO:MF	proteoglycan binding	1.02E-03
GO:MF	collagen binding	1.41E-02
GO:MF	protein binding	4.10E-02
GO:BP	regulation of apoptotic process	3.25E-10
GO:BP	regulation of programmed cell death	4.75E-10
GO:BP	response to stress	1.82E-09
GO:BP	response to chemical	1.97E-06
GO:BP	response to external stimulus	1.98E-06
GO:BP	immune system process	3.51E-06
GO:BP	positive regulation of cellular metabolic process	9.23E-06
GO:BP	positive regulation of cell death	9.38E-06
GO:BP	defense response	1.05E-05
GO:BP	regulation of cell communication	2.02E-02
GO:BP	organonitrogen compound metabolic process	2.10E-02
GO:BP	protein modification process	2.21E-02
GO:BP	cellular protein modification process	2.21E-02
GO:BP	negative regulation of intracellular signal transduction	2.24E-02
GO:BP	regulation of cell adhesion	2.32E-02
KEGG	FoxO signaling pathway	2.31E-02
REAC	Eukaryotic Translation Elongation	1.26E-02
WP	Senescence and Autophagy in Cancer	2.36E-03

Then, genes in the filtered score table were mapped to the Human genome-scale PPI network. 47 out of 76 genes were found to interact. Figure 4.6 shows the networks of filtered genes.

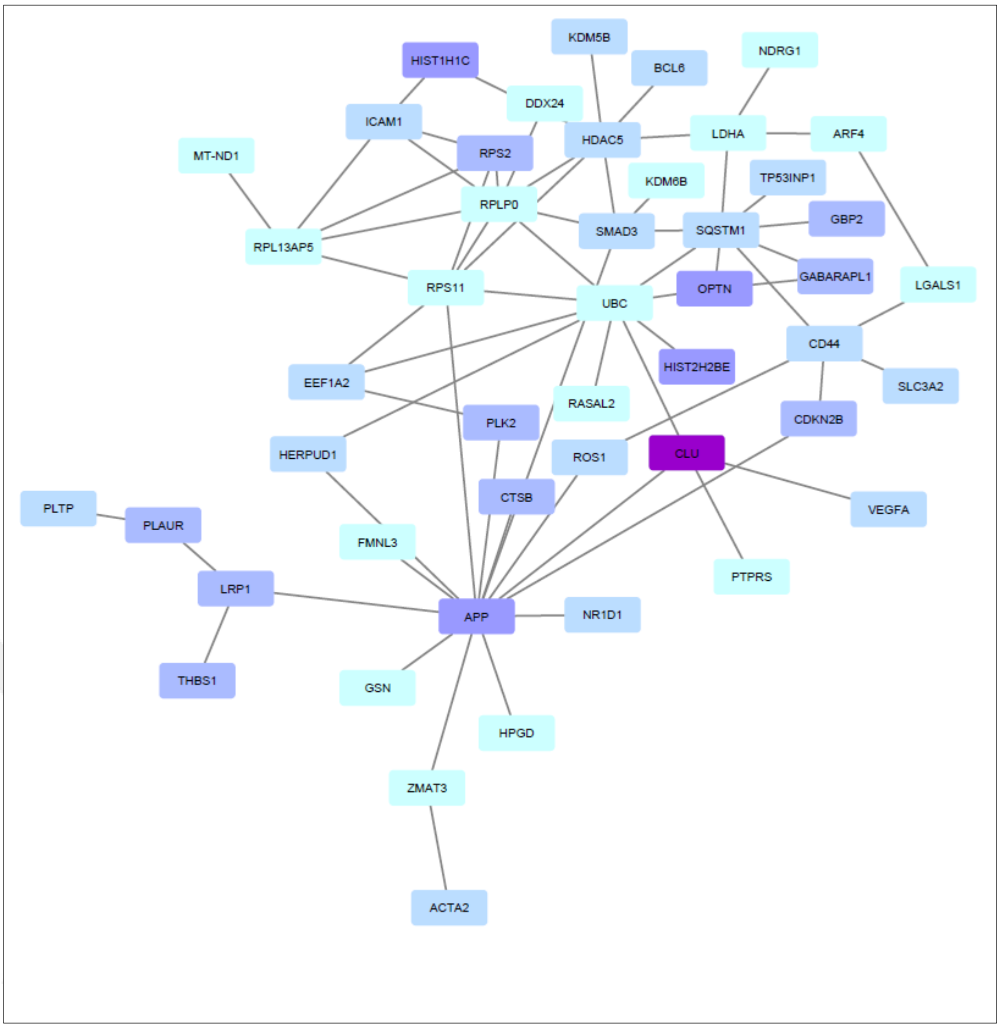


Figure 4.6: Network of genes from the filtered PPI score list.

4.4. Drug Repositioning Analysis of PPI Subnetworks

The genes that are frequently found in the subnetworks can be important in dormancy mechanism in different cancers. Therefore, drugs that target these genes can help destroy the dormant cancer cells. For this purpose, drug repositioning analysis, which has been used widely in recent years in the literature and allows the use of currently approved drugs for different indications, were applied. In this analysis, the genes included in the filtered score list were selected as the target genes. Drugs targeting these genes were retrieved from DrugBank database data. Table 4.6 shows the drugs that target at least 2 filtered PPI subnetwork genes from Table 4.4, number of genes the drug target from the filtered list, the target genes, and the clinical testing stage of these drugs.

Table 4.6: The drugs that target at least 2 genes from the filtered score table (Table 4.4) of PPI subnetworks.

Drug	**Score**	**Genes**	**Group**
COPPER	6	APP,C3,CLU,GSN, LDHA,RPS2	approved, investigational
ZINC CHLORIDE	4	APP,C3,CLU,GSN	approved,
ZINC ACETATE	4	APP,C3,CLU,GSN	approved,
ZINC	4	APP,C3,CLU,GSN	approved,
NADH	4	HMOX1,HPGD,LD HA,MT-ND1	approved, investigational
FOSTAMATINIB	3	DDR1,PLK2,ROS1	approved
TENECTEPLASE	2	LRP1,PLAUR	investigational
PHENETHYL ISOTHIOCYANATE	2	ACTA2,MT-ND1	investigational
N-FORMYLMETHIONINE	2	MT-ND1,UBC	experimental
LANOTEPLASE	2	LRP1,PLAUR	investigational
HYALURONIC ACID	2	CD44,ICAM1	approved
ARTENIMOL	2	LDHA,LGALS1	experimental,

The drugs in Table 4.6 have been shown to target related genes, but many of them are unknown pharmacological action or mechanism of action on genes.

4.5. Genes Not Found in PPI Network

Interactome databases store experimentally validated interactions or interactions predicted by methods such as text-mining. Only experimentally validated interactions are used in this thesis. However, the information contained in these databases does not include all protein-protein interactions due to the technical difficulties of obtaining data and the lack of knowledge of the interaction of all proteins yet. For some genes in the transcriptome data, there is no interaction information in the interactome data. In order not to ignore these proteins which may be important, as also mentioned in Section 3.6.3, genes selected from any transcriptome data but not found in the interactome were identified among the datasets selected from the literature and scored according to the number of transcriptome datasets in which they were significantly changed. Table 4.7 shows human and mouse genes with highest significance scores.

Table 4.7: Human and mouse genes not found in PPI network with highest score. (F:FALSE, T:TRUE)

Human							**Mouse**						
Entrez ID	**Gene Symbol**	**Significance Score (total: 17)**	**Dormancy gene?**	**Autophagy gene?**	**ECM gene?**	**Cell Cycle gene?**	**Entrez ID**	**Gene Symbol**	**Significance Score (total: 4)**	**Dormancy gene?**	**Autophagy gene?**	**ECM gene?**	**Cell Cycle gene?**
79815	NIPAL2	7	F	F	F	F	53313	Atp2a3	4	F	F	F	F
60526	LDAH	7	F	F	F	F	30955	Pik3cg	4	F	F	F	F
100507246	SNHG16	7	F	F	F	F	11305	Abca2	4	F	F	F	F
85315	PAQR8	7	F	F	F	F	13197	Gadd45a	4	F	F	F	T
84798	C19orf48	7	F	F	F	F	77938	Fam53b	4	F	F	F	F
54842	MFSD6	7	F	F	F	F	14456	Gas6	4	F	F	F	F
92579	G6PC3	7	F	F	F	F	68818	Zfand2b	3	F	F	F	F
3128	HLA-DRB6	6	F	F	F	F	227333	Dgkd	3	F	F	F	F
57567	ZNF319	6	F	F	F	F	226691	Ifi207	3	F	F	F	F
100533105	C8orf44-SGK3	6	F	F	F	F	57319	Smpdl3a	3	F	F	F	F

The highest scoring genes that not found in PPI were generally associated with none of the terms autophagy, ECM, cell cycle, and dormancy. This shows that the absence of these genes does not cause a great loss of information in the analyses.

5. Gene-Regulatory Subnetworks

5.1. Data Preparation and Subnetwork Discovery

Transcriptome data that was processed for PPI subnetwork discovery (Section 4.1) were used in these analyses. In gene-regulatory subnetwork discovery, only interactome data was different from the PPI subnetwork discovery method. Gene-regulatory networks include transcription factor-gene interactions and they are directed networks. It means edge is directed from transcription factor to target gene of the transcription factor. Besides, some parameters of the subnetwork discovery methods were changed. To decide on the optimal value of the K parameter, different K values were tried on JB-V and J82 cell lines. Figure 5.1 shows the K value versus TF number of subnetworks and, Table 5.1 shows the subnetwork sizes for each K value. K value was chosen as 8 based on this analysis since higher K value allows existance of more transcription factors in subnetwork and size of subnetwork is not so big or small when K value was selectad as 8. In BioNet, network type parameter was chosen as ‘directed’ instead of ‘undirected’ in GR network analyses.

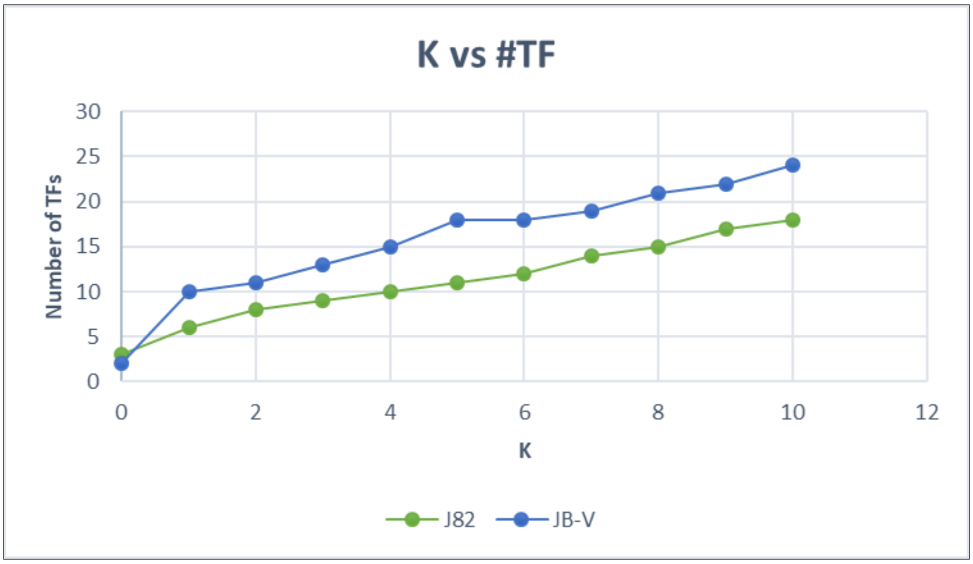


Figure 5.1: K versus number of TFs for J82 and JB-V GR subnetworks.

Table 5.1: Subnetwork sizes of J82 and JB-V GR cell lines for K parameter optimization.

	J82			**JB-V**		
K	**#TF**	**Nodes**	**Edges**	**#TF**	**Nodes**	**Edges**
0	3	7	6	2	5	4
1	6	32	33	10	56	61
2	8	55	62	11	69	98
3	9	67	72	13	80	123
4	10	77	95	15	88	137
5	11	85	118	18	95	151
6	12	93	126	18	101	172
7	14	98	136	19	106	189
8	**15**	**102**	**148**	**21**	**111**	**217**
9	17	106	147	22	115	229
10	18	110	161	24	118	237

KPM subnetwork discovery analysis of J82 and JB-V cell lines of GSE35947 dataset was performed with K=8, L=0, INES and ACO parameters. Binarization cut-off was selected as 0.01 for adjusted p-values for both cell lines. Based on this cut-off, 339 genes were differentially expressed in J82 cell line and 283 genes were differentially expressed in JB-V cell line. 97 percent of transcriptome data was mapped to GR network. Figure 5.2 shows the biggest subnetworks discovered by KPM analysis for J82 and JB-V cell lines.

BioNet subnetwork discovery analysis was also performed with the same dataset. FDR value was chosen as 0.005 for both J82 and JB-V cell lines. The network after transcriptome data mapping to PPI network consists of 6073 genes and 18263 interactions. Figure 5.3 shows the subnetworks discovered by BioNet analysis for J82 and JB-V cell lines respectively.

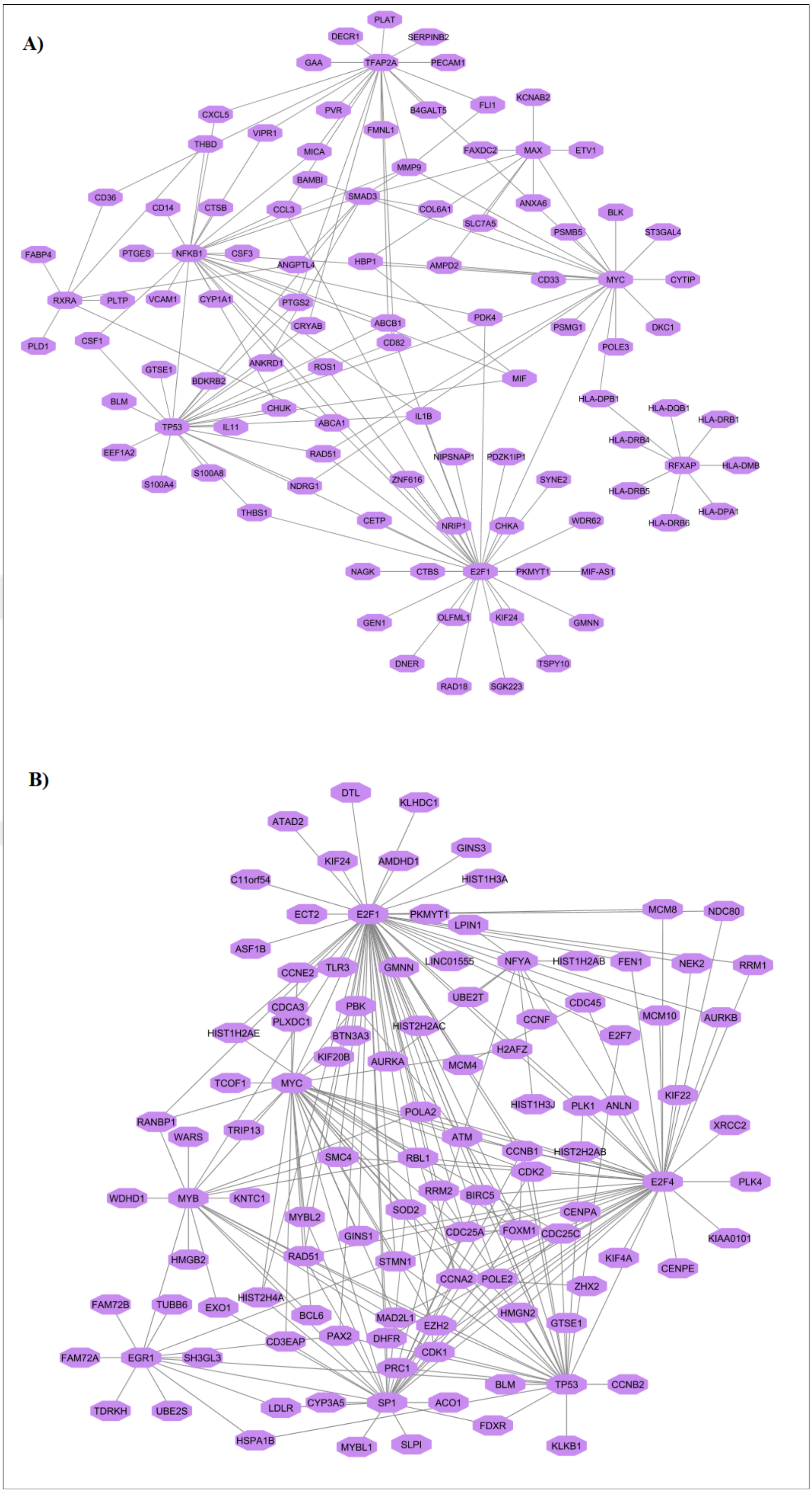


Figure 5.2: KPM GR subnetworks for dataset GSE35947. A)J82 (102 nodes and 148 edges), B)JB-V (111 nodes and 217 edges)

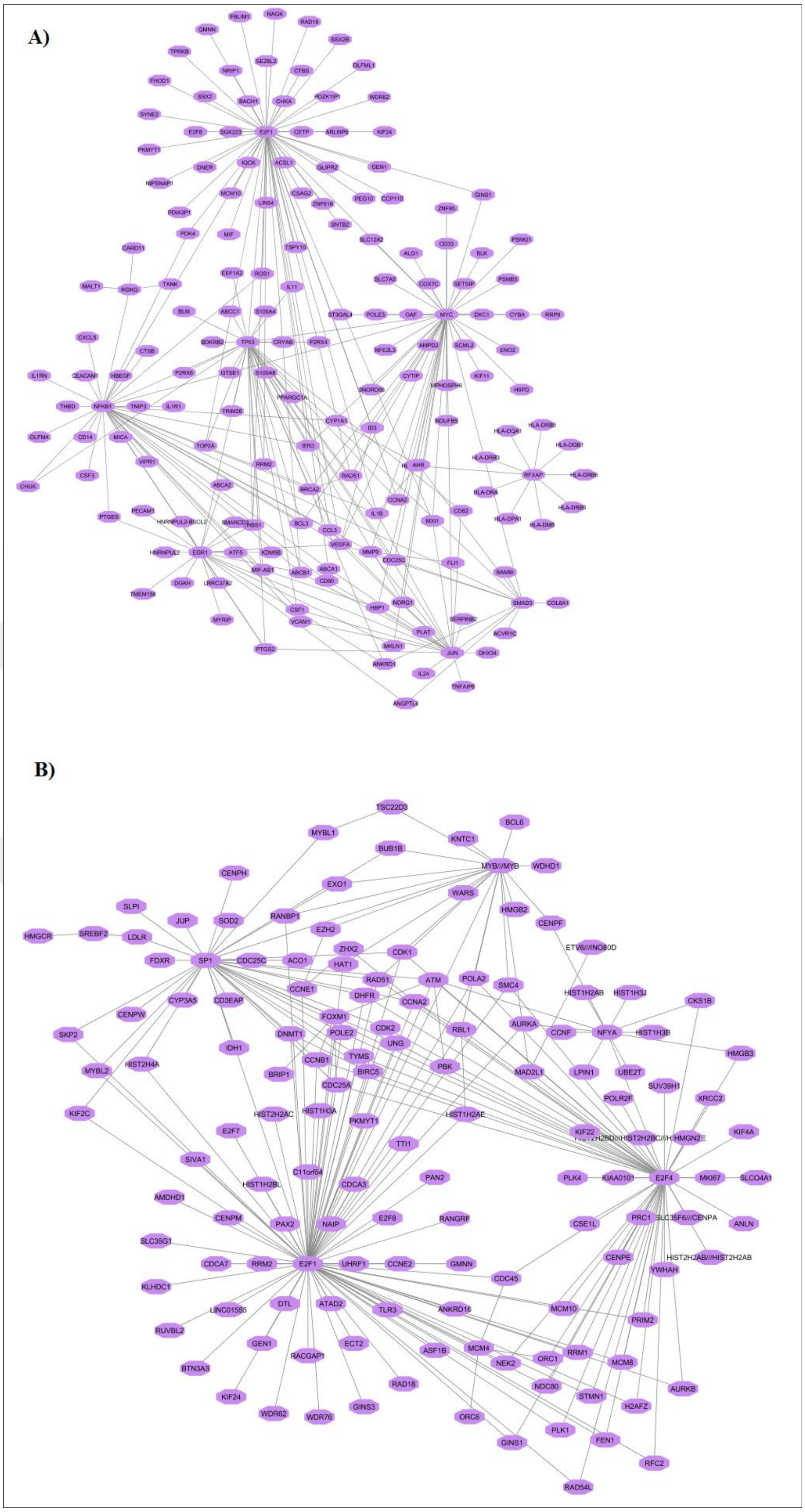


Figure 5.3: BioNet GR subnetworks for dataset GSE35947. A)J82 (168 nodes and 241 edges), B)JB-V (144 nodes and 218 edges.

The same workflow was applied to all the compared conditions, with parameters that were mentioned in Section 3. The corresponding cut-off values used for each comparison are given in Table 3.4 for KPM analysis, and in Table 3.5 for BioNet analysis. For seven of the comparisons BioNet could not identify any subnetworks. As a result, 35 gene-regulatory subnetworks were discovered from 21 compared conditions. The sizes of these subnetworks are shown in Table 5.2.

Table 5.2: Subnetwork sizes of GR subnetworks.

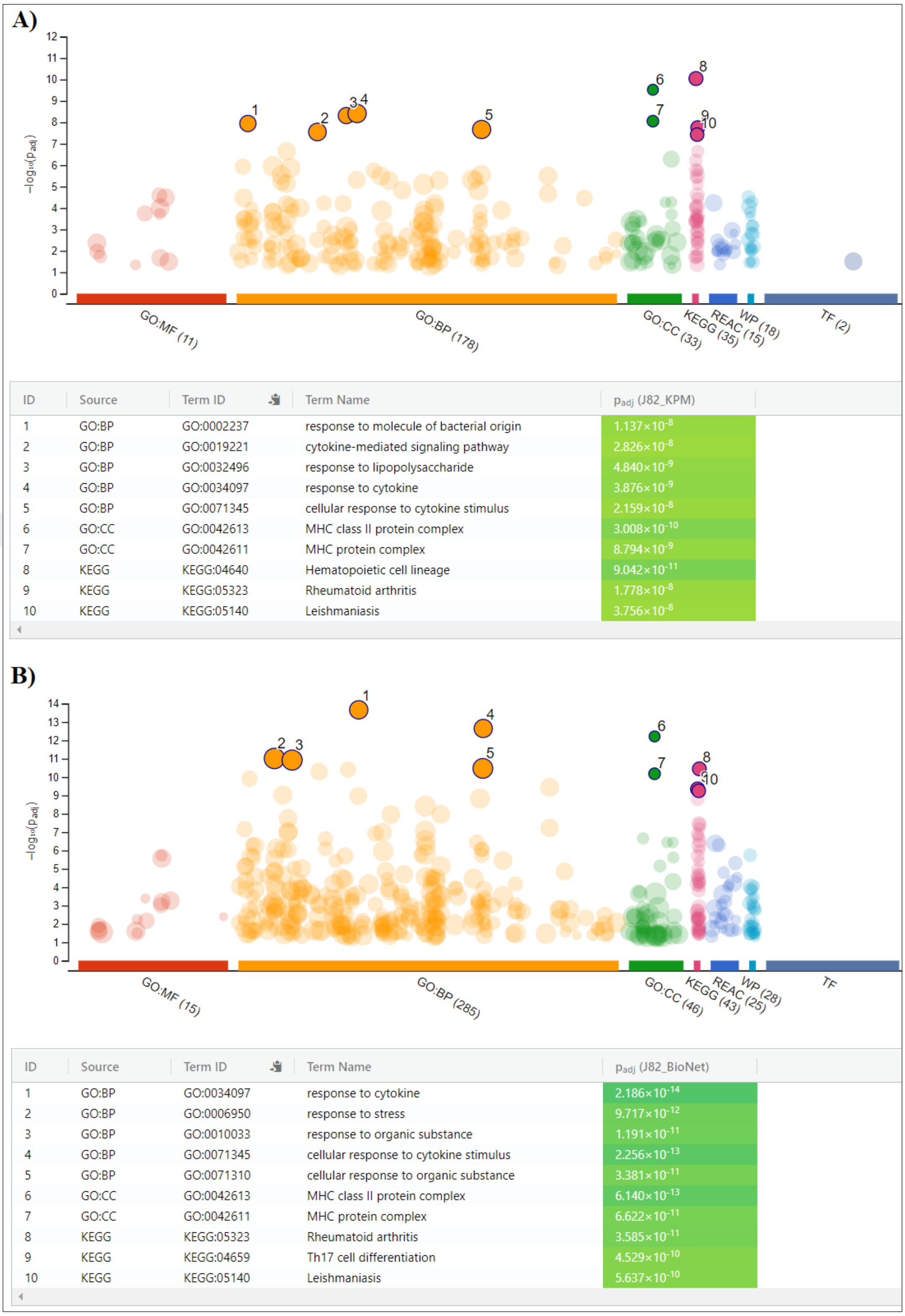
			KPM			BioNet		
#	**GEO Code**	**Compared Conditions**	**Nodes**	**Edges**	**#TF**	**Nodes**	**Edges**	**# TF**
1	GSE35947	J82 cell line, dormant vs proliferative	102	148	15	168	241	24
2	GSE35947	JB-V cell line, dormant vs proliferative	111	217	21	144	218	26
3	GSE83142	RNAseq analysis, dormant vs proliferative	67	122	14	74	132	14
4	GSE83142	scRNAseq analysis, dormant vs proliferative	59	79	14	50	56	8
5	GSE114012	DLD1 cell line, dormant vs proliferative	195	375	39	338	686	65
6	GSE114012	HCT15 cell line, dormant vs proliferative	55	78	13	39	48	9
7	GSE114012	HT55 cell line, dormant vs proliferative	156	284	36	294	447	61
8	GSE114012	RKO cell line, dormant vs proliferative	497	1116	89	534	1170	98
9	GSE114012	SW48 cell line, dormant vs proliferative	77	134	18	65	95	15
10	GSE114012	SW948 cell line, dormant vs proliferative	100	173	24	116	172	29
11	GSE102483	dormant vs proliferative	87	148	22	47	56	14

Table 5.2: Continued.

12	GSE20611	dormant vs post-dormant	170	265	28	37	42	5
13	GSE64262	dormant vs proliferative	377	652	18	-	-	-
14	GE77379	PC3 cell line *in vivo*, dormant vs proliferative	186	287	28	-	-	-
15	GE77379	PC3 cell line *in vitro*, dormant vs proliferative	338	595	48	-	-	-
16	GE77379	C42B cell line *in vivo*, dormant vs proliferative	121	181	27	-	-	-
17	GE77379	C42B cell line *in vitro*, dormant vs proliferative	19	22	10	-	-	-
18	GSE57695	dormant vs proliferative	107	156	32	71	93	19
19	GSE112094	dormant vs proliferative	209	322	59	183	296	50
20	GSE131890	D2.OR (3D) vs D2.A1 (3D) cell line, dormant vs proliferative	57	68	19	-	-	-
21	GSE131890	D2.OR (3D) vs D2.OR (2D) cell line, dormant vs proliferative	24	25	6	-	-	-

5.2. g: Profiler Analysis of Subnetworks

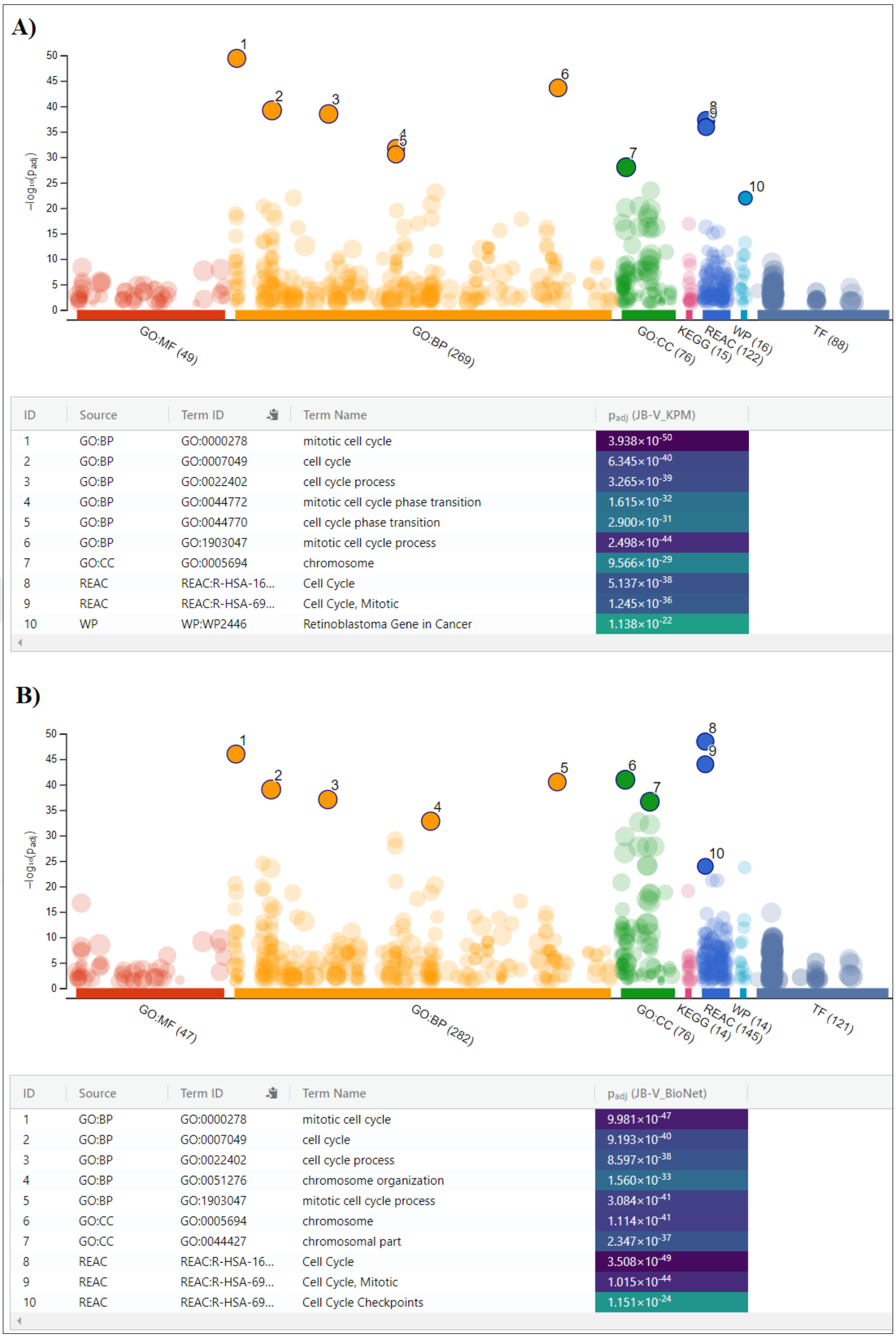
After subnetworks were discovered, enrichment analyses were performed by g:GOST tool of g:Profiler as mentioned previously. Figure 5.4 and Figure 5.5 shows the graphic results of g:Profiler of JB-V and J82 cell lines. Table 5.3 shows the a summary of g:Profiler results of J82 cell line GR subnetwork.



ID	Source	Term ID	Term Name	p_{adj} (J82_KPM)
1	GO:BP	GO:0002237	response to molecule of bacterial origin	1.137×10^{-8}
2	GO:BP	GO:0019221	cytokine-mediated signaling pathway	2.826×10^{-8}
3	GO:BP	GO:0032496	response to lipopolysaccharide	4.840×10^{-9}
4	GO:BP	GO:0034097	response to cytokine	3.876×10^{-9}
5	GO:BP	GO:0071345	cellular response to cytokine stimulus	2.159×10^{-8}
6	GO:CC	GO:0042613	MHC class II protein complex	3.008×10^{-10}
7	GO:CC	GO:0042611	MHC protein complex	8.794×10^{-9}
8	KEGG	KEGG:04640	Hematopoietic cell lineage	9.042×10^{-11}
9	KEGG	KEGG:05323	Rheumatoid arthritis	1.778×10^{-8}
10	KEGG	KEGG:05140	Leishmaniasis	3.756×10^{-8}

ID	Source	Term ID	Term Name	p_{adj} (J82_BioNet)
1	GO:BP	GO:0034097	response to cytokine	2.186×10^{-14}
2	GO:BP	GO:0006950	response to stress	9.717×10^{-12}
3	GO:BP	GO:0010033	response to organic substance	1.191×10^{-11}
4	GO:BP	GO:0071345	cellular response to cytokine stimulus	2.256×10^{-13}
5	GO:BP	GO:0071310	cellular response to organic substance	3.381×10^{-11}
6	GO:CC	GO:0042613	MHC class II protein complex	6.140×10^{-13}
7	GO:CC	GO:0042611	MHC protein complex	6.622×10^{-11}
8	KEGG	KEGG:05323	Rheumatoid arthritis	3.585×10^{-11}
9	KEGG	KEGG:04659	Th17 cell differentiation	4.529×10^{-10}
10	KEGG	KEGG:05140	Leishmaniasis	5.637×10^{-10}

Figure 5.4: g:Profiler results of GR subnetworks of J82 cell line for dataset GSE35947. A) KPM , B) BioNet



ID	Source	Term ID	Term Name	p_{adj} (JB-V_KPM)
1	GO:BP	GO:0000278	mitotic cell cycle	3.938×10^{-50}
2	GO:BP	GO:0007049	cell cycle	6.345×10^{-40}
3	GO:BP	GO:0022402	cell cycle process	3.265×10^{-39}
4	GO:BP	GO:0044772	mitotic cell cycle phase transition	1.615×10^{-32}
5	GO:BP	GO:0044770	cell cycle phase transition	2.900×10^{-31}
6	GO:BP	GO:1903047	mitotic cell cycle process	2.498×10^{-44}
7	GO:CC	GO:0005694	chromosome	9.566×10^{-29}
8	REAC	REAC:R-HSA-16...	Cell Cycle	5.137×10^{-38}
9	REAC	REAC:R-HSA-69...	Cell Cycle, Mitotic	1.245×10^{-36}
10	WP	WP:WP2446	Retinoblastoma Gene in Cancer	1.138×10^{-22}

ID	Source	Term ID	Term Name	p_{adj} (JB-V_BioNet)
1	GO:BP	GO:0000278	mitotic cell cycle	9.981×10^{-47}
2	GO:BP	GO:0007049	cell cycle	9.193×10^{-40}
3	GO:BP	GO:0022402	cell cycle process	8.597×10^{-38}
4	GO:BP	GO:0051276	chromosome organization	1.560×10^{-33}
5	GO:BP	GO:1903047	mitotic cell cycle process	3.084×10^{-41}
6	GO:CC	GO:0005694	chromosome	1.114×10^{-41}
7	GO:CC	GO:0044427	chromosomal part	2.347×10^{-37}
8	REAC	REAC:R-HSA-16...	Cell Cycle	3.508×10^{-49}
9	REAC	REAC:R-HSA-69...	Cell Cycle, Mitotic	1.015×10^{-44}
10	REAC	REAC:R-HSA-69...	Cell Cycle Checkpoints	1.151×10^{-24}

Figure 5.5: g:Profiler results of GR subnetworks of J-BV cell line for dataset GSE35947. A) KPM , B) BioNet

According to Figure 5.4 and Figure 5.5, in J82 cell line, mostly immune and immune system related disease terms and terms related to cytokines which play an important role in intercellular communication were found. In the JB-V cell line, there are mainly cell cycle terms. The reason for this difference may be that different cell lines become dormant through different cellular mechanisms.

Table 5.3: g:Profiler results of GR subnetwork of J82 cell line for dataset GSE35947.

Source	Term Name	Adjusted p-value
GO:MF	peptide binding	2.55E-05
GO:MF	protein-containing complex binding	3.25E-05
GO:MF	identical protein binding	0.000100864
GO:BP	response to cytokine	3.88E-09
GO:BP	response to lipopolysaccharide	4.84E-09
GO:BP	response to molecule of bacterial origin	1.14E-08
GO:BP	cellular response to cytokine stimulus	2.16E-08
GO:BP	cytokine-mediated signaling pathway	2.83E-08
GO:BP	response to abiotic stimulus	2.27E-07
GO:BP	response to stress	1.12E-06
GO:BP	response to hypoxia	1.17E-06
GO:BP	response to organic substance	1.35E-06
GO:BP	response to decreased oxygen levels	1.77E-06
GO:BP	cellular response to organic substance	2.78E-06
GO:BP	cell population proliferation	2.91E-06
GO:BP	regulation of cell population proliferation	3.19E-06
GO:BP	cell differentiation	0.001951015
GO:BP	interleukin-1 beta production	0.001968143
GO:BP	regulation of immune system process	0.002010233
GO:BP	negative regulation of cell cycle	0.002046997
GO:BP	regulation of response to external stimulus	0.002186582
GO:BP	regulation of cytokine production	0.002457829
GO:BP	cellular response to abiotic stimulus	0.002584872
GO:BP	cellular response to environmental stimulus	0.002584872
GO:BP	negative regulation of cell death	0.002633523
GO:BP	DNA damage response, signal transduction by p53 class mediator	0.002720417
GO:BP	myeloid leukocyte activation	0.002732002
GO:BP	response to antibiotic	0.002819711
GO:BP	regulation of lipid metabolic process	0.003018191
GO:CC	external side of apical plasma membrane	0.049954897
KEGG	Hematopoietic cell lineage	9.04E-11
KEGG	Rheumatoid arthritis	1.78E-08
KEGG	Leishmaniasis	3.76E-08
KEGG	Inflammatory bowel disease (IBD)	2.25E-07
KEGG	Antigen processing and presentation	5.99E-07
KEGG	Th17 cell differentiation	1.68E-06
KEGG	Graft-versus-host disease	1.74E-06
KEGG	Human T-cell leukemia virus 1 infection	3.23E-06
KEGG	Type I diabetes mellitus	3.79E-06

Table 5.3: Continued.

REAC	Cytokine Signaling in Immune system	5.78E-05
REAC	Signaling by Interleukins	0.001135102
REAC	TFAP2 (AP-2) family regulates transcription of cell cycle factors	0.001480797
REAC	Downstream TCR signaling	0.003258966
REAC	Translocation of ZAP-70 to Immunological synapse	0.004879366
REAC	Interleukin-4 and Interleukin-13 signaling	0.006963825
REAC	Immune System	0.006987582
WP	Senescence and Autophagy in Cancer	0.00157859
WP	Vitamin D Receptor Pathway	0.001780933
WP	Spinal Cord Injury	0.003036575
WP	Small cell lung cancer	0.006587222
WP	Bladder Cancer	0.006808241

5.3. Integration of Subnetworks

Score tables of genes that appeared in GR and TFless GR subnetworks were formed as mentioned in Section 4.3 and filtered separately. The unfiltered table includes 2197 genes for GR subnetworks while the filtered table includes 71 genes. Table 5.4 shows the filtered score table of GR subnetworks. Table 5.5 shows the a part of g:Profiler analysis result of genes that found in Table 5.4.

Table 5.4: Filtered score table of all GR subnetworks obtained in this study.

Human ID	**Human Gene Symbol**	**Significance Score**	**Overall Score (significant or not)**	**Article Score**	**Dormancy genes?**	**Autophagy genes?**	**ECM genes?**	**Cell Cycle genes?**
1191	CLU	7	7	3	FALSE	TRUE	TRUE	FALSE
1030	CDKN2B	6	6	3	FALSE	FALSE	TRUE	TRUE
8349	HIST2H2	6	6	4	FALSE	FALSE	TRUE	FALSE
114907	FBXO32	6	6	2	FALSE	FALSE	FALSE	FALSE
1508	CTSB	6	6	3	FALSE	FALSE	TRUE	FALSE
7057	THBS1	5	5	3	TRUE	FALSE	TRUE	TRUE
7422	VEGFA	5	5	5	FALSE	FALSE	TRUE	FALSE
10769	PLK2	5	5	2	FALSE	TRUE	FALSE	TRUE
90427	BMF	5	5	3	FALSE	TRUE	FALSE	FALSE

Table 5.4: Continued.

2034	EPAS1	5	5	3	FALSE	FALSE	FALSE	FALSE
5329	PLAUR	5	5	3	FALSE	FALSE	TRUE	FALSE
3106	HLA-F	5	5	2	FALSE	FALSE	FALSE	FALSE
9619	ABCG1	5	5	3	FALSE	FALSE	TRUE	FALSE
26959	HBP1	5	5	3	FALSE	FALSE	FALSE	TRUE
4088	SMAD3	5	5	3	FALSE	FALSE	TRUE	TRUE
7832	BTG2	5	5	4	FALSE	FALSE	TRUE	TRUE
1893	ECM1	5	5	4	FALSE	FALSE	TRUE	FALSE
94241	TP53INP	5	5	4	FALSE	TRUE	FALSE	TRUE
3913	LAMB2	4	4	2	FALSE	FALSE	TRUE	FALSE
4323	MMP14	4	4	2	FALSE	FALSE	TRUE	FALSE
9572	NR1D1	4	4	2	FALSE	FALSE	FALSE	FALSE
489	ATP2A3	4	4	3	FALSE	FALSE	FALSE	FALSE
604	BCL6	4	4	3	FALSE	FALSE	FALSE	TRUE
2879	GPX4	4	4	3	FALSE	FALSE	TRUE	FALSE
3162	HMOX1	4	4	3	FALSE	TRUE	TRUE	FALSE
639	PRDM1	4	4	3	FALSE	FALSE	FALSE	FALSE
5054	SERPINE	4	4	3	FALSE	FALSE	TRUE	FALSE
8878	SQSTM1	4	4	3	FALSE	TRUE	TRUE	FALSE
972	ROS1	4	4	4	FALSE	FALSE	TRUE	FALSE
3553	IL1B	4	4	4	FALSE	FALSE	TRUE	TRUE
3956	LGALS1	4	4	4	FALSE	FALSE	TRUE	FALSE
84617	TUBB6	4	4	4	FALSE	FALSE	TRUE	TRUE
467	ATF3	4	4	2	FALSE	FALSE	TRUE	FALSE
10370	CITED2	4	4	2	FALSE	FALSE	FALSE	TRUE
80315	CPEB4	4	4	2	FALSE	FALSE	TRUE	FALSE
64065	PERP	4	4	2	FALSE	FALSE	FALSE	FALSE
26470	SEZ6L2	4	4	2	FALSE	FALSE	FALSE	FALSE
6520	SLC3A2	4	4	2	FALSE	FALSE	TRUE	FALSE
19	ABCA1	4	4	3	FALSE	FALSE	TRUE	FALSE
9828	ARHGEF	4	4	3	FALSE	FALSE	FALSE	FALSE
1917	EEF1A2	4	4	3	FALSE	TRUE	FALSE	FALSE
2026	ENO2	4	4	3	FALSE	FALSE	TRUE	FALSE
355	FAS	4	4	3	FALSE	FALSE	TRUE	FALSE
2348	FOLR1	4	4	3	FALSE	FALSE	TRUE	FALSE
10397	NDRG1	4	4	3	FALSE	FALSE	TRUE	FALSE
960	CD44	4	4	4	FALSE	FALSE	TRUE	FALSE
10018	BCL2L11	3	3	2	FALSE	TRUE	FALSE	TRUE
901	CCNG2	3	3	2	FALSE	FALSE	FALSE	TRUE
1490	CTGF	3	3	2	FALSE	FALSE	TRUE	TRUE
1543	CYP1A1	3	3	2	FALSE	FALSE	TRUE	TRUE
1643	DDB2	3	3	2	FALSE	FALSE	FALSE	FALSE
3383	ICAM1	3	3	2	FALSE	FALSE	TRUE	FALSE
23135	KDM6B	3	3	2	FALSE	FALSE	FALSE	FALSE

Table 5.4: Continued.

9314	KLF4	3	4	2	FALSE	FALSE	FALSE	TRUE
81631	MAP1LC	3	3	2	FALSE	TRUE	TRUE	FALSE
9462	RASAL2	3	3	2	FALSE	FALSE	FALSE	FALSE
6590	SLPI	3	3	2	FALSE	FALSE	TRUE	FALSE
7008	TEF	3	3	2	FALSE	FALSE	FALSE	FALSE
7077	TIMP2	3	3	2	FALSE	FALSE	TRUE	TRUE
219699	UNC5B	3	3	2	FALSE	FALSE	FALSE	FALSE
64393	ZMAT3	3	3	2	FALSE	FALSE	FALSE	FALSE
51129	ANGPTL	3	3	3	FALSE	FALSE	TRUE	FALSE
360	AQP3	3	3	3	FALSE	FALSE	TRUE	FALSE
8553	BHLHE4(	3	3	3	FALSE	FALSE	FALSE	FALSE
6347	CCL2	3	3	3	FALSE	FALSE	TRUE	FALSE
10129	FRY	3	3	3	FALSE	FALSE	FALSE	FALSE
10765	KDM5B	3	3	3	FALSE	FALSE	FALSE	FALSE
7277	TUBA4A	3	3	3	FALSE	FALSE	TRUE	TRUE

Genes found in the filtered list, similar to PPI analysis, are mainly associated with autophagy, ECM, and the cell cycle. However, no genes associated with dormancy has been found. This may be because there are few genes in the dormancy list. But terms related tometastasis and dormancy mechanism like “angiogenesis”, “hypoxia” and “cell population proliferation” are found in g:profiler analysis of filtered gene list.

Table 5.5: g:Profiler result of filtered gene list of all GR subnetworks.

Source	**Term Name**	**Adjusted p-value**
GO:BP	response to stress	1.7E-06
GO:BP	vasculature development	1.9E-06
GO:BP	circulatory system development	1.9E-06
GO:BP	anatomical structure formation involved in morphogenesis	2.0E-06
GO:BP	cardiovascular system development	2.1E-06
GO:BP	Angiogenesis	5.7E-06
GO:BP	animal organ development	1.8E-05
GO:BP	cell population proliferation	1.8E-05
GO:BP	regulation of cell adhesion	2.0E-05
GO:BP	response to hypoxia	0.00027
GO:BP	immune system development	0.00032
GO:BP	response to decreased oxygen levels	0.00033
GO:BP	positive regulation of gene expression	0.00034

Table 5.5: Continued.

REAC	FOXO-mediated transcription	0.02346
REAC	Regulation of gene expression by Hypoxia-inducible Factor	0.0259
REAC	Signaling by TGF-beta Receptor Complex	0.02919
WP	Vitamin D Receptor Pathway	0.00194
WP	Senescence and Autophagy in Cancer	0.00472
WP	Platelet-mediated interactions with vascular and circulating cells	0.0445

Then, the filtered lists based on significance score were mapped on the genome scale GR network and connected genes were extracted. 15 out of 71 genes interact in the GR networks. Figure 5.6 shows the network of filtered genes of GR subnetworks.

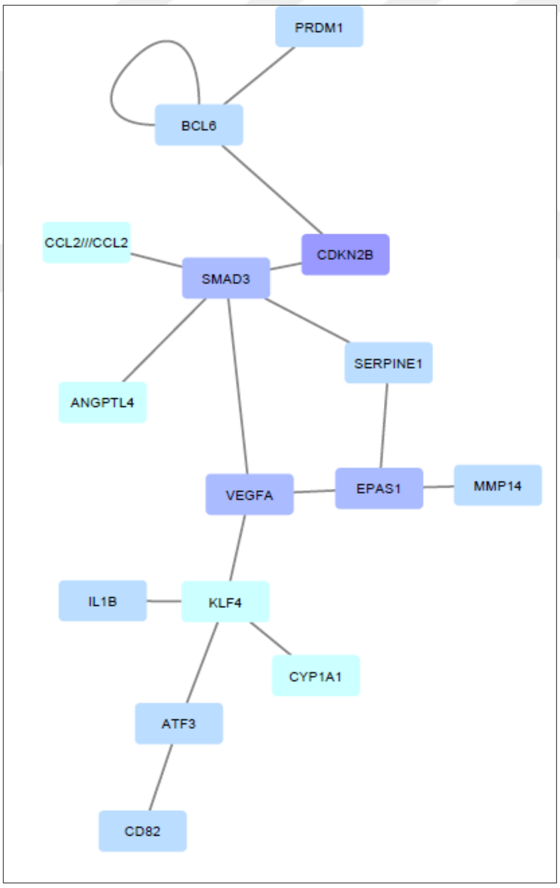


Figure 5.6: Network of genes from the filtered GR score list.

5.4. Drug Repositioning Analysis of GR Subnetworks

Drug repositioning analysis was performed similarly to Section 4.4, which was for protein interaction subnetworks was also performed for gene regulation subnetworks. For this purpose, the genes included in the filtered score list were used as target genes, and drugs that target the target genes were identified by comparing them with gene-drug interactions from the DrugBank database. Drugs targeting at least two of the genes in the target gene list are shown in Table 5.6.

Table 5.6: The drugs that target at least 2 genes from the filtered score table (Table 5.4) of GR subnetworks.

Drug	**Score**	**Genes**	**Group**
UROKINASE	2	PLAUR,SERPINE1	approved, investigational
TENECTEPLASE	2	PLAUR,SERPINE1	approved
RETEPLASE	2	PLAUR,SERPINE1	approved, investigational
MINOCYCLINE	2	IL1B,VEGFA	approved, investigational
HYALURONIC ACID	2	CD44,ICAM1	approved
FOSTAMATINIB	2	ROS1,PLK2	approved, investigational
FORESKIN KERATINOCYTE (NEONATAL)	2	IL1B,VEGFA	approved
CYT997	2	TUBB6,TUBA4A	investigational
COPPER	2	CLU,SERPINE1	approved, investigational
CHONDROITIN SULFATE	2	VEGFA,CCL2	approved, investigational, nutraceutical
ATP	2	ABCG1, ABCA1	investigational, nutraceutical
ARTENIMOL	2	LGALS1,TUBB6	experimental, investigational
ANISTREPLASE	2	PLAUR,SERPINE1	approved
ALTEPLASE	2	PLAUR,SERPINE1	approved

Similar to PPI results, the drugs in Table 5.6 have been shown to target related genes, but many of them are unknown pharmacological action or mechanism of action on genes.

5.5. TFless Gene Regulatory Networks

Transcription factors are key players in the cells for gene expression regulation. Although the target genes alter their gene expression levels in response to condition changes, gene expression levels of TFs are usually constant in the cell. Based on this phenomenon, the GR network downloaded from TRRUST and RegNetwork were used to generate a TFless GR network which was also used for subnetwork discovery. The aim of the construction of TFless gene regulatory network is to remove TFs, which do not show significant changes in the transcriptome data, from the network while keeping their regulatory effect still in the network (Figure 3.1). K value was selected as 2 in the KPM analysis of TFless GR networks.BioNet network type was chosen as 'undirected' because TFless network is the TF-eliminated version of GR network and the edges do not have direction as mentioned in Section 3.4. 96.9 percent of transcriptome was mapped to TFless GR network in the KPM analysis of J82 and JB-V cell lines.. Table 5.7 shows the sizes of subnetwork obtained by mapping all comparisons to TFless GR networks.

Table 5.7: Subnetwork sizes of TFless GR subnetworks.

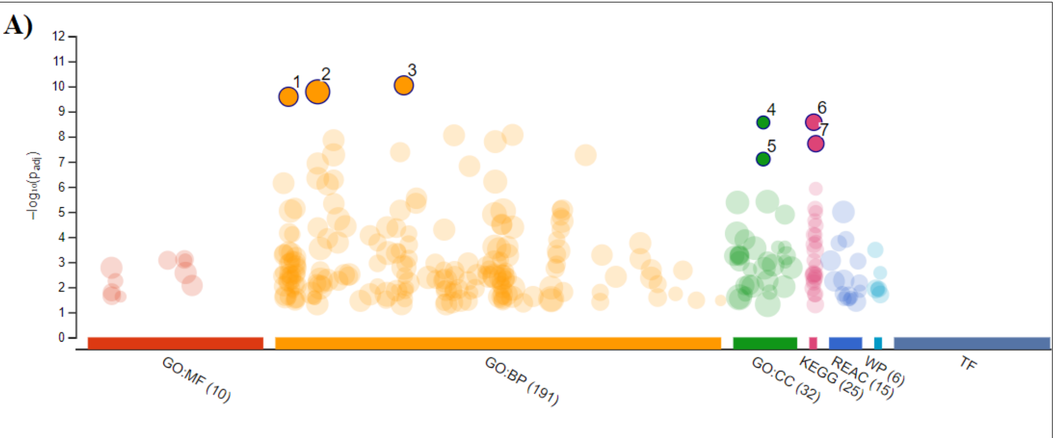
			KPM		**BioNet**	
#	**GEO Code**	**Compared Conditions**	**Nodes**	**Edges**	**Nodes**	**Edges**
1	GSE35947	J82 cell line, dormant vs proliferative	138	1817	258	6403
2	GSE35947	JB-V cell line, dormant vs proliferative	130	2464	213	5762
3	GSE83142	RNAseq analysis, dormant vs proliferative	78	918	99	1431
4	GSE83142	scRNAseq analysis, dormant vs proliferative	66	559	90	1119
5	GSE114012	DLD1 cell line, dormant vs proliferative	233	7583	406	22751
6	GSE114012	HCT15 cell line, dormant vs proliferative	64	315	82	520
7	GSE114012	HT55 cell line, dormant vs proliferative	206	4015	479	17650

Table 5.7: Continued.

8	GSE114012	RKO cell line, dormant vs proliferative	545	33781	628	43254
9	GSE114012	SW48 cell line, dormant vs proliferative	94	956	87	912
10	GSE114012	SW948 cell line, dormant vs proliferative	124	1527	194	4259
11	GSE102483	dormant vs proliferative	105	1145	85	761
12	GSE20611	dormant vs post-dormant	239	4530	57	371
13	GSE64262	dormant vs proliferative	86	725	-	-
14	GE77379	PC3 cell line *in vivo*, dormant vs proliferative	256	5662	-	-
15	GE77379	PC3 cell line *in vitro*, dormant vs proliferative	389	13353	-	-
16	GE77379	C42B cell line *in vivo*, dormant vs proliferative	160	2106	-	-
17	GE77379	C42B cell line *in vitro*, dormant vs proliferative	19	40	-	-
18	GSE57695	dormant vs proliferative	153	1341	326	5323
19	GSE112094	dormant vs proliferative	303	4649	318	5116
20	GSE131890	D2.OR (3D) vs D2.A1 (3D) cell line, dormant vs proliferative	133	1180	-	-
21	GSE131890	D2.OR (3D) vs D2.OR (2D) cell line, dormant vs proliferative	65	300	-	-

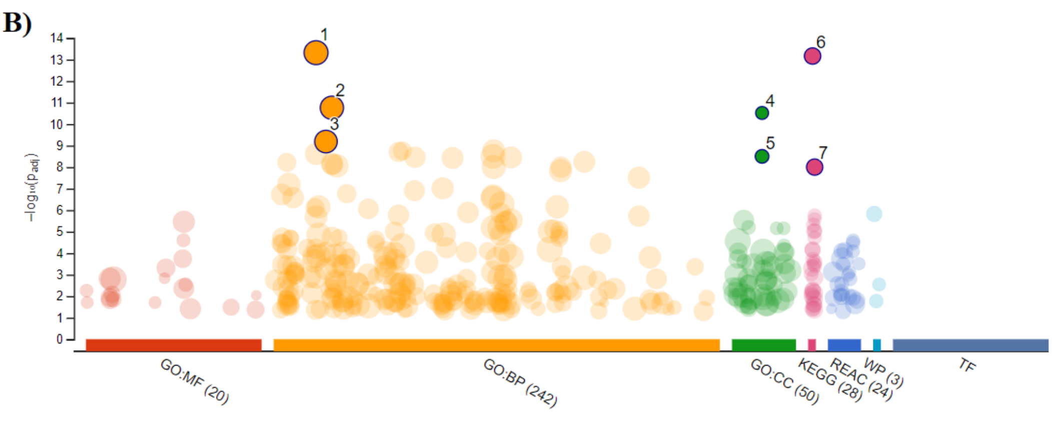
After subnetworks were discovered, g:Profiler analysis was performed as described previously. Figure 5.7 and Figure 5.8 shows the g:Profiler results of J82 and JB-V cell lines. Table 5.8 shows the a part of g:Profiler results of J82 cell line TFless GR subnetwork.

A)



ID	Source	Term ID	Term Name	p_{adj} (J82_KPM)
1	GO:BP	GO:0002237	response to molecule of bacterial origin	2.592×10^{-10}
2	GO:BP	GO:0006950	response to stress	1.640×10^{-10}
3	GO:BP	GO:0032496	response to lipopolysaccharide	9.088×10^{-11}
4	GO:CC	GO:0042613	MHC class II protein complex	2.729×10^{-9}
5	GO:CC	GO:0042611	MHC protein complex	7.878×10^{-8}
6	KEGG	KEGG:04640	Hematopoietic cell lineage	2.679×10^{-9}
7	KEGG	KEGG:05323	Rheumatoid arthritis	1.918×10^{-8}

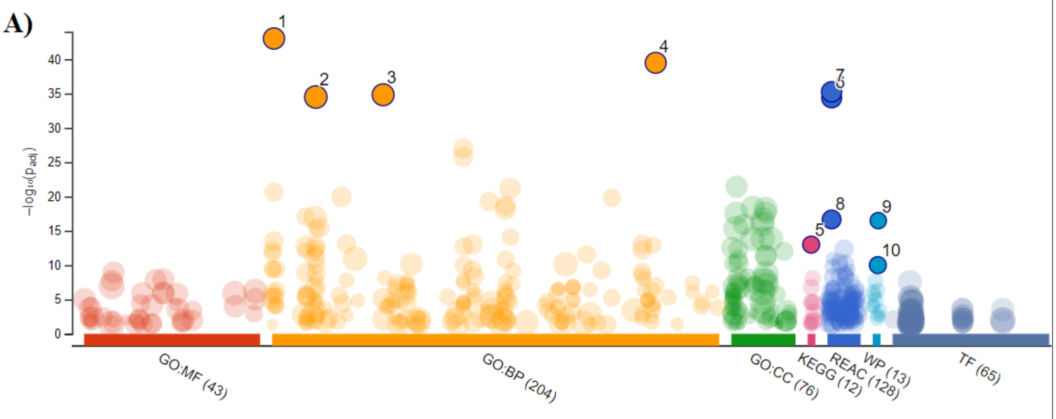
B)



ID	Source	Term ID	Term Name	p_{adj} (J82_BioNet)
1	GO:BP	GO:0006950	response to stress	4.791×10^{-14}
2	GO:BP	GO:0009605	response to external stimulus	1.739×10^{-11}
3	GO:BP	GO:0008283	cell population proliferation	6.483×10^{-10}
4	GO:CC	GO:0042613	MHC class II protein complex	3.034×10^{-11}
5	GO:CC	GO:0042611	MHC protein complex	3.165×10^{-9}
6	KEGG	KEGG:04640	Hematopoietic cell lineage	6.844×10^{-14}
7	KEGG	KEGG:05323	Rheumatoid arthritis	9.982×10^{-9}

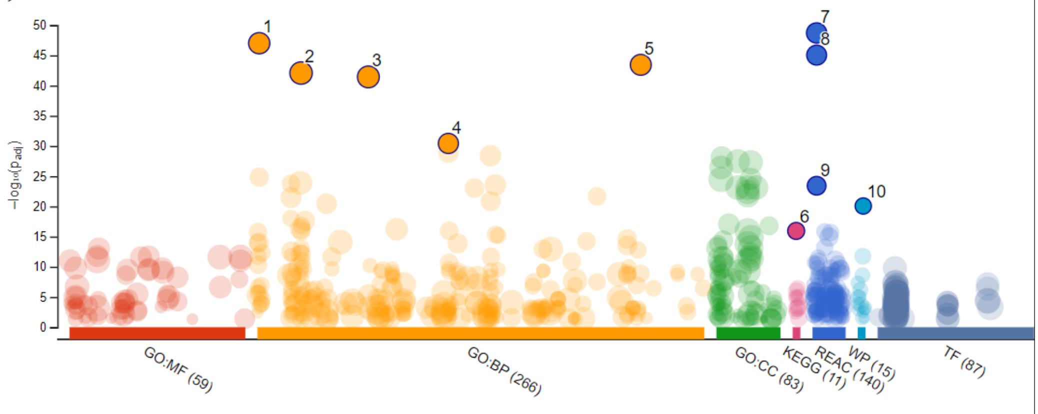
Figure 5.7: g:Profiler results of TFless GR subnetworks of J82 cell line for dataset GSE35947. A) KPM , B) BioNet

A)



ID	Source	Term ID	Term Name	p_{adj} (JB-V_KPM)
1	GO:BP	GO:0000278	mitotic cell cycle	9.432×10^{-44}
2	GO:BP	GO:0007049	cell cycle	3.151×10^{-35}
3	GO:BP	GO:0022402	cell cycle process	1.482×10^{-35}
4	GO:BP	GO:1903047	mitotic cell cycle process	3.292×10^{-40}
5	KEGG	KEGG:04110	Cell cycle	1.014×10^{-13}
6	REAC	REAC:R-HSA-69...	Cell Cycle, Mitotic	4.220×10^{-35}
7	REAC	REAC:R-HSA-16...	Cell Cycle	5.486×10^{-36}
8	REAC	REAC:R-HSA-69...	Cell Cycle Checkpoints	2.170×10^{-17}
9	WP	WP:WP2446	Retinoblastoma Gene in Cancer	3.203×10^{-17}
10	WP	WP:WP179	Cell Cycle	9.758×10^{-11}

B)



ID	Source	Term ID	Term Name	p_{adj} (JBV_BioNet)
1	GO:BP	GO:0000278	mitotic cell cycle	9.693×10^{-48}
2	GO:BP	GO:0007049	cell cycle	8.367×10^{-43}
3	GO:BP	GO:0022402	cell cycle process	3.665×10^{-42}
4	GO:BP	GO:0044772	mitotic cell cycle phase transition	3.772×10^{-31}
5	GO:BP	GO:1903047	mitotic cell cycle process	3.653×10^{-44}
6	KEGG	KEGG:04110	Cell cycle	1.152×10^{-16}
7	REAC	REAC:R-HSA-16...	Cell Cycle	1.937×10^{-49}
8	REAC	REAC:R-HSA-69...	Cell Cycle, Mitotic	9.049×10^{-46}
9	REAC	REAC:R-HSA-69...	Cell Cycle Checkpoints	3.824×10^{-24}
10	WP	WP:WP2446	Retinoblastoma Gene in Cancer	8.619×10^{-21}

Figure 5.8: g:Profiler results of TFless GR subnetworks of JB-V cell line for dataset GSE35947. A) KPM , B) BioNet

TFless GRN g: profiler results are very similar to results of GR network analysis for both cell lines. Majority of terms are related to stress and immune system in J82 cell line, and related to cell cycle in JB-V cell line.

Table 5.8: g:Profiler results of TFless GR subnetwork of J82 cell line for dataset GSE35947.

Source	**Term Name**	**Adjusted p-value**
GO:MF	peptide binding	0.000783645
GO:MF	amide binding	0.000847658
GO:MF	peptide antigen binding	0.00085265
GO:BP	response to stress	1.64E-10
GO:BP	response to molecule of bacterial origin	2.59E-10
GO:BP	response to other organism	8.55E-09
GO:BP	immune response	4.60E-07
GO:BP	response to bacterium	5.23E-07
GO:BP	regulation of response to stimulus	6.23E-07
GO:BP	cell activation	7.18E-07
GO:BP	cell population proliferation	7.89E-07
GO:BP	activation of immune response	0.000375134
GO:BP	leukocyte mediated immunity	0.000429494
GO:BP	cell differentiation	0.000437353
GO:BP	regulation of defense response	0.000446737
KEGG	Hematopoietic cell lineage	2.68E-09
KEGG	Rheumatoid arthritis	1.92E-08
KEGG	Asthma	1.19574E-06
REAC	Immune System	1.00654E-05
REAC	MHC class II antigen presentation	0.000125356
REAC	Downstream TCR signaling	0.000177672
REAC	Adaptive Immune System	0.00087662
WP	Allograft Rejection	0.000327908
WP	Cytokines and Inflammatory Response	0.002690271
WP	Nuclear Receptors Meta-Pathway	0.011714112
WP	Senescence and Autophagy in Cancer	0.011852769

Finally, a score table was generated and filtered for TFless subnetworks as described previously. The unfiltered table includes 3006 genes for TFless subnetworks while the filtered table includes 111 genes. Table 5.9 shows the filtered score table of TFless subnetworks.

Table 5.9: Filtered score table of all TFless GR subnetworks obtained in this study.

Human ID	Human Gene Symbol	Significant Score	Score (significant or not)	Article Score	Dormancy gene?	Autophagy genes?	ECM genes?	Cell Cycle genes?
1191	CLU	8	8	4	FALSE	TRUE	TRUE	FALSE
4758	NEU1	7	7	3	FALSE	FALSE	TRUE	FALSE
1728	NQO1	6	6	5	FALSE	FALSE	TRUE	FALSE
3304	HSPA1B	6	6	4	FALSE	FALSE	TRUE	TRUE
9572	NR1D1	6	6	4	FALSE	FALSE	FALSE	FALSE
1030	CDKN2B	6	6	3	FALSE	FALSE	TRUE	TRUE
8349	HIST2H2BE	6	6	4	FALSE	FALSE	TRUE	FALSE
114907	FBXO32	6	6	2	FALSE	FALSE	FALSE	FALSE
1508	CTSB	6	6	3	FALSE	FALSE	TRUE	FALSE
9619	ABCG1	6	6	4	FALSE	FALSE	TRUE	FALSE
10133	OPTN	6	6	5	FALSE	TRUE	FALSE	TRUE
7057	THBS1	5	5	3	TRUE	FALSE	TRUE	TRUE
718	C3	5	5	4	FALSE	FALSE	TRUE	FALSE
7422	VEGFA	5	5	5	FALSE	FALSE	TRUE	FALSE
10769	PLK2	5	5	2	FALSE	TRUE	FALSE	TRUE
90427	BMF	5	5	3	FALSE	TRUE	FALSE	FALSE
10370	CITED2	5	5	3	FALSE	FALSE	FALSE	TRUE
2034	EPAS1	5	5	3	FALSE	FALSE	FALSE	FALSE
2548	GAA	5	5	3	FALSE	FALSE	TRUE	FALSE
5329	PLAUR	5	5	3	FALSE	FALSE	TRUE	FALSE
4035	LRP1	5	5	4	FALSE	FALSE	TRUE	FALSE
23135	KDM6B	5	5	4	FALSE	FALSE	FALSE	FALSE
3106	HLA-F	5	5	2	FALSE	FALSE	FALSE	FALSE
19	ABCA1	5	5	3	FALSE	FALSE	TRUE	FALSE
4088	SMAD3	5	5	3	FALSE	FALSE	TRUE	TRUE
7077	TIMP2	5	5	3	FALSE	FALSE	TRUE	TRUE
7832	BTG2	5	5	4	FALSE	FALSE	TRUE	TRUE
1893	ECM1	5	5	4	FALSE	FALSE	TRUE	FALSE
94241	TP53INP1	5	5	4	FALSE	TRUE	FALSE	TRUE
9590	AKAP12	4	4	2	FALSE	FALSE	FALSE	FALSE
25805	BAMBI	4	4	2	FALSE	FALSE	FALSE	FALSE
3659	IRF1	4	4	2	FALSE	FALSE	FALSE	TRUE
3913	LAMB2	4	4	2	FALSE	FALSE	TRUE	FALSE
55818	KDM3A	4	4	2	FALSE	FALSE	FALSE	FALSE
4323	MMP14	4	4	2	FALSE	FALSE	TRUE	FALSE
653361	NCF1C	4	4	2	FALSE	FALSE	FALSE	FALSE
59	ACTA2	4	4	3	FALSE	FALSE	TRUE	FALSE
489	ATP2A3	4	4	3	FALSE	FALSE	FALSE	FALSE
604	BCL6	4	4	3	FALSE	FALSE	FALSE	TRUE
966	CD59	4	4	3	FALSE	FALSE	TRUE	FALSE

Table 5.9: Continued.

1026	CDKN1A	4	6	3	TRUE	FALSE	TRUE	TRUE
4189	DNAJB9	4	4	3	FALSE	FALSE	TRUE	FALSE
2879	GPX4	4	4	3	FALSE	FALSE	TRUE	FALSE
3162	HMOX1	4	4	3	FALSE	TRUE	TRUE	FALSE
639	PRDM1	4	4	3	FALSE	FALSE	FALSE	FALSE
8878	SQSTM1	4	4	3	FALSE	TRUE	TRUE	FALSE
5054	SERPINE1	4	4	3	FALSE	FALSE	TRUE	FALSE
972	ROS1	4	4	4	FALSE	FALSE	TRUE	FALSE
1628	DBP	4	4	4	FALSE	FALSE	FALSE	FALSE
2217	FCGRT	4	4	4	FALSE	FALSE	FALSE	FALSE
3956	LGALS1	4	4	4	FALSE	FALSE	TRUE	FALSE
3553	IL1B	4	4	4	FALSE	FALSE	TRUE	TRUE
3119	HLA-DRB6	4	4	4	FALSE	FALSE	FALSE	FALSE
84617	TUBB6	4	4	4	FALSE	FALSE	TRUE	TRUE
467	ATF3	4	4	2	FALSE	FALSE	TRUE	FALSE
80315	CPEB4	4	4	2	FALSE	FALSE	TRUE	FALSE
26959	HBP1	4	4	2	FALSE	FALSE	FALSE	TRUE
64065	PERP	4	4	2	FALSE	FALSE	FALSE	FALSE
5252	PHF1	4	4	2	FALSE	FALSE	FALSE	FALSE
26470	SEZ6L2	4	4	2	FALSE	FALSE	FALSE	FALSE
6520	SLC3A2	4	4	2	FALSE	FALSE	TRUE	FALSE
219699	UNC5B	4	4	2	FALSE	FALSE	FALSE	FALSE
2026	ENO2	4	4	3	FALSE	FALSE	TRUE	FALSE
1917	EEF1A2	4	4	3	FALSE	TRUE	FALSE	FALSE
56204	FAM214A	4	4	3	FALSE	FALSE	FALSE	FALSE
355	FAS	4	4	3	FALSE	FALSE	TRUE	FALSE
2348	FOLR1	4	4	3	FALSE	FALSE	TRUE	FALSE
10765	KDM5B	4	4	3	FALSE	FALSE	FALSE	FALSE
10397	NDRG1	4	4	3	FALSE	FALSE	TRUE	FALSE
4854	NOTCH3	4	4	3	FALSE	FALSE	TRUE	FALSE
9828	ARHGEF17	4	4	3	FALSE	FALSE	FALSE	FALSE
960	CD44	4	4	4	FALSE	FALSE	TRUE	FALSE
5360	PLTP	4	4	4	FALSE	FALSE	TRUE	FALSE
8992	ATP6V0E1	3	3	2	FALSE	TRUE	FALSE	FALSE
10018	BCL2L11	3	3	2	FALSE	TRUE	FALSE	TRUE
1490	CTGF	3	3	2	FALSE	FALSE	TRUE	TRUE
901	CCNG2	3	3	2	FALSE	FALSE	FALSE	TRUE
1543	CYP1A1	3	3	2	FALSE	FALSE	TRUE	TRUE
1643	DDB2	3	3	2	FALSE	FALSE	FALSE	FALSE
3958	LGALS3	3	3	2	FALSE	FALSE	TRUE	FALSE
8347	HIST1H2BC	3	3	2	FALSE	FALSE	TRUE	FALSE
85236	HIST1H2BK	3	3	2	FALSE	FALSE	TRUE	FALSE
3383	ICAM1	3	3	2	FALSE	FALSE	TRUE	FALSE
23338	JADE2	3	3	2	FALSE	FALSE	TRUE	FALSE
8425	LTBP4	3	3	2	FALSE	FALSE	TRUE	FALSE
9961	MVP	3	3	2	FALSE	FALSE	TRUE	FALSE
81631	MAP1LC3B2	3	3	2	FALSE	TRUE	TRUE	FALSE
1397	-	3	3	2	FALSE	FALSE	FALSE	FALSE

Table 5.9: Continued.

9462	RASAL2	3	3	2	FALSE	FALSE	FALSE	FALSE
6590	SLPI	3	3	2	FALSE	FALSE	TRUE	FALSE
10580	SORBS1	3	3	2	FALSE	FALSE	FALSE	FALSE
7008	TEF	3	3	2	FALSE	FALSE	FALSE	FALSE
64393	ZMAT3	3	3	2	FALSE	FALSE	FALSE	FALSE
51129	ANGPTL4	3	3	3	FALSE	FALSE	TRUE	FALSE
360	AQP3	3	3	3	FALSE	FALSE	TRUE	FALSE
8553	BHLHE40	3	3	3	FALSE	FALSE	FALSE	FALSE
84699	CREB3L3	3	3	3	FALSE	FALSE	FALSE	FALSE
6347	CCL2	3	3	3	FALSE	FALSE	TRUE	FALSE
9451	EIF2AK3	3	3	3	FALSE	FALSE	TRUE	FALSE
10129	FRY	3	3	3	FALSE	FALSE	FALSE	FALSE
3488	IGFBP5	3	3	3	FALSE	FALSE	TRUE	FALSE
3108	HLA-DMA	3	3	3	FALSE	FALSE	FALSE	FALSE
4320	MMP11	3	3	3	FALSE	FALSE	TRUE	FALSE
57326	PBXIP1	3	3	3	FALSE	FALSE	FALSE	FALSE
5802	PTPRS	3	3	3	FALSE	FALSE	TRUE	FALSE
6662	SOX9	3	3	3	FALSE	FALSE	TRUE	TRUE
1831	TSC22D3	3	3	3	FALSE	FALSE	FALSE	FALSE
7277	TUBA4A	3	3	3	FALSE	FALSE	TRUE	TRUE

When Table 5.9 compared with Table 5.4, it is seen that the genes obtained in TFless GR network analysis cover all but one ("Kruppel like factor 4" (KLF4)) of the genes obtained in GR nework analysis. In the g: Profiler analysis of the GSE35947 dataset, very similar results were obtained for both TFless GR network and GR network. Accordingly, it can be concluded that although the number of genes in TFless GR network is high, the subnetworks are functionally similar to the GR network.

Then filtered lists were mapped on human genome scale TFless GR network and interacting parts were extracted. 110 out of 110 genes interact in TFless networks. There were 111 genes in filtered list but only 110 genes were mapped to human network because one gene (Crip2) was only found in the mouse network. Figure 5.9 shows the network of filtered genes of TFless subnetworks.

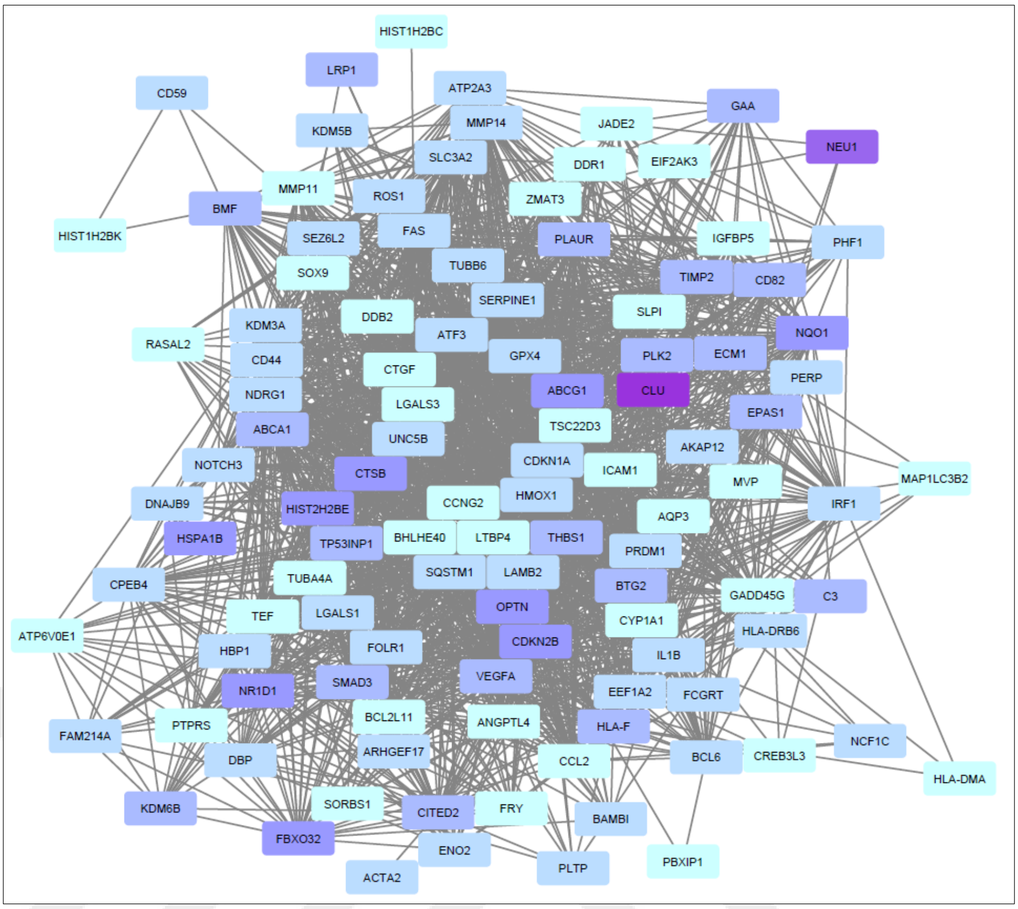


Figure 5.9: Network of genes that found in filtered TFless GR network score list.

Finally, the drug repositioning study, which was also applied for the results of PPI and GR networks, was also applied to TFless GR subnetworks. For this purpose, the genes in the filtered score table were used as target genes. Drugs targeting these genes were identified using drug-gene interaction data from the DrugBank database. Table 5.10 shows the list of drugs targeting at least 2 of the target genes.

Table 5.10: The drugs that target at least 2 genes from the filtered score table (Table 5.9) of TFless GR subnetworks.

Drug	Score	Genes	Group
COPPER	4	C3, HIST1H2BC, CLU, SERPINE1	approved, investigational
FOSTAMATINIB	3	DDR1, ROS1, PLK2	approved, investigational
TENECTEPLASE	3	SERPINE1, PLAUR, LRP1	approved
ALTEPLASE	2	SERPINE1, PLAUR	approved
ANISTREPLASE	2	SERPINE1, PLAUR	approved
ARTENIMOL	2	TUBB6,LGALS1	experimental, investigational
ATP	2	ABCA1, ABCG1	investigational, nutraceutical
CHONDROITIN SULFATE	2	CCL2,VEGFA	approved, investigational, nutraceutical
CYT997	2	TUBA4A, TUBB6	investigational
FORESKIN KERATINOCYTE (NEONATAL)	2	IL1B,VEGFA	approved
HYALURONIC ACID	2	CD44,ICAM1	approved
LANOTEPLASE	2	LRP1,PLAUR	investigational
MARIMASTAT	2	MMP14,MMP11	investigational
MINOCYCLINE	2	IL1B,VEGFA	approved, investigational
PHENETHYL ISOTHIOCYANATE	2	ACTA2, TUBA4A	investigational
RETEPLASE	2	PLAUR, SERPINE1	approved, investigational
UROKINASE	2	SERPINE1,PLAUR	approved, investigational, withdrawn
ZINC	2	CLU,C3	approved, investigational
ZINC ACETATE	2	CLU,C3	approved, investigational
ZINC CHLORIDE	2	CLU,C3	approved, investigational

The drug candidates found as a result of TFless GR network analyses include the whole drug candidates found in the GR network analysis. Considering that TFless GR network target gene list cover GR network target gene lists, this is an expected result.

6. DISCUSSION

Cancer is one of the most important lethal diseases. The most important reason for cancer to be fatal is metastasis of cancer, which is due to the dissemination of cancer cells to other body parts. In addition, these metastatic cells can resist treatment and remain silent without showing signs of disease for years and then relapse. For these reasons, understandingthe mechanisms underlying metastatisis crucial for developing efficient therapies. To achieve this purpose in this study, computational techniques were used to reveal cellular mechanisms that have role in metastasis.

Firstly, cancer datasets were chosen from GEO database [Barrett et al., 2012]. All dataset includes dormant and proliferative cancer samples. Dormant versus proliferative conditions were chosen for comparison to understand the changes between cancer and pre-metastatic cells. In this study, a comparative computational analysis of cancer and dormant conditions in humans and mice was performed to understand the changes in the cellular mechanisms of cancer cells before relapse, called macro metastasis.

In the computational analysis of datasets, two subnetwork discovery methods, KPM and BioNet based on the study by Batra and collegues, which showed their superiority [Batra et al., 2017].

Different compared conditions in all datasets were analyzed by these two subnetwork discovery tools. As shown in Table 4.2, Table 5,2 and Table 5.4, the success of finding subnetwork of KPM, especially in mouse datasets, was higher than BioNet. This may be due to the differences in algorithms used by the two tools and the way they use the data. For example, KPM receives the expression data in a binarized manner, consisting of 1 and 0, and data is usually binarized by using adjusted p-values. In BioNet, unadjusted p-values are used, and the tool adjusts the p-values . In addition, because of these characteristics, KPM considers all of the genes that change significantly as if they have changed significantly at the same rate, and ignores the difference between p-value 0.01 and p-value 0.00001. BioNet uses this information because it takes p-values directly. Also, KPM considers unmapped nodes of interactome as exception nodes, which are the nodes that didnot change significantly. However, BioNet ignores these nodes in its analyses.

The binarization cut-off used in KPM when creating subnetworks and FDR values used in BioNet were adjusted to keep subnetwork sizes similar in all analyses. However, since the p-value ranges of each compared conditions were different from each other, the subnetwork dimensions were not found to be similar in all analyses. Although the subnetwork dimensions are different, the results of g: Profiler have similar terms between different compared conditions. Moreoversimilar terms were obtained in the PPI, GR network and TFless GR network results of the same compared condition. This shows that the subnetworks obtained in the analyses are consistent although they come from different studies and belong to different types of cancer and even different organisms.

In g: Profiler analyses, the most common terms in both PPI, GR network and TFless GR network in different datasets are terms related to cell cycle and response to stress of cells. These are expected results because, as mentioned earlier, the cells stop dividing in the case of dormancy. Furthermore, tumor cells separated from the main tissue are exposed to many stresses, such as hypoxia and immune system, in the new tissue they reach. They change their cellular mechanisms in order to survive and adapt to the environment [Endo and Inoue, 2019, Gomis and Gawrzak, 2016].

Subsequently, to analyse subnetworks deeply, the genes common in many subnetworks were identified. These genes were filtered so that appearing at least in 2 datasets and having a subnetwork score of at least 3 were retained. Fold changes were also checked and only the genes that showed increased mRNA abundance in the majority of subnetworks were kept for further analysis. These filtered gene lists were identified separately for PPI network, GR network and TFless GR network. As a result, it was observed that all genes in the GR network filtered list except the "Kruppel like factor 4" (KLF4) gene were included in the TFless GR network list. Therefore, it is reasonable to take only the genes in the GR network list as a result of the GR network analysis. When the genes in the PPI and GR lists are compared, 34 genes are common in these two lists. These common genes are shown in Table 6.1.

Table 6.1 : Common filtered genes in PPI and GR subnetwork score lists.

Human Genesymbol	**Human Gene ID**	**PPI Score**	**GR Score**
CLU	1191	7	9
HIST2H2BE	8349	6	6
FBXO32	114907	6	6
CDKN2B	1030	6	5
CTSB	1508	6	5
PLAUR	5329	5	5
THBS1	7057	5	5
PLK2	10769	5	5
ECM1	1893	5	4
SMAD3	4088	5	4
VEGFA	7422	5	4
ABCG1	9619	5	4
TP53INP1	94241	5	4
BCL6	604	4	4
CD44	960	4	4
ROS1	972	4	4
EEF1A2	1917	4	4
GPX4	2879	4	4
HMOX1	3162	4	4
SLC3A2	6520	4	4
SQSTM1	8878	4	4
NR1D1	9572	4	4
ICAM1	3383	3	4
TIMP2	7077	3	4
KDM5B	10765	3	4
BTG2	7832	5	3
ENO2	2026	4	3
LGALS1	3956	4	3
NDRG1	10397	4	3
SEZ6L2	26470	4	3
AQP3	360	3	3
RASAL2	9462	3	3
KDM6B	23135	3	3
ZMAT3	64393	3	3

The top 10 of these common genes with the highest scores are those previously associated with cancer metastasis and/or treatment resistance. For example, inhibition of CLU (clusterin) reduces the sensitivity of prostate cancer cells to chemotherapy and

CLU is highly expressed in drug-resistant cancer cells [Patterson et al., 2006, Wilson and Zoubeidi, 2017]. In addition, the CLU gene has been shown to induce EMT in lung cancer [Jin et al., 2017]. Nayak and collegues showed that HIST1H2BE is overexpressed in treatment resistant breast cancer cells [Nayak et al., 2015]. FBXO32 is ubiquitinated to stabilize CtBP1, which induces transcription of genes for creating suitable microenvironment for EMT progression [Sahu et al., 2017]. Chen and collegues reported that CDKN2B expression is upregulated in gastric cancer and it is correlated with terms like invasion and distant metastasis [Chen et al., 2019]. Also, Roshan and collegues demonstrated that downregulation of CDKN2A and CDKN2B is associated with the recurrence of oral squamous cell carcinoma [Deepak Roshan et al., 2019]. CTSB has roles in cell invasion and metastasis of colorectal cancer [Bian et al., 2016].

Obviosly, genes that were identified in these analyses are involved in mechanisms such as metastasis and resistance to treatment, which are characteristics of dormant cells. Thus, drugs targeting the proteins expressed by these genes can be used to destroy the dormant cancer cells. Developing new drugs is both costly and time-consuming process. However, drug repositioning studies allow the use of existing drugs for other indications. Thus, treatments can be improved quickly and cheaper. In this thesis study, drugs and genes targeted by drugs were downloaded from DrugBank database for drug repositioning. The proposed target genes in filtered score lists were compared with DrugBank drug target list and drugs that target proposed genes were identified. Table 6.2 shows the common drugs for PPI and GR subnetwork analyses.

Table 6.2: Common drugs in PPI and GR subnetwork score lists

Drug	**Target Genes Comes From PPI list**	**Target Genes Comes From GR list**
ARTENIMOL	LDHA, LGALS1	LGALS1, TUBB6
COPPER	APP, C3, CLU, GSN, LDHA, RPS2	CLU, SERPINE1
FOSTAMATINIB	DDR1, PLK2, ROS1	ROS1, PLK2
HYALURONIC ACID	CD44, ICAM1	CD44, ICAM1
TENECTEPLASE	LRP1, PLAUR	PLAUR, SERPINE1

The 5 drugs shown in Table 6.2 target the genes that have changed significantly in case of dormancy in different cancer samples analysed within the scope of this thesis. Hyaluronic acid and Tenectaplase have an activating effect on the targeted genes. However, in this thesis study, selected drug target genes are genes that are upregulated in case of dormancy. For this reason, it may not be appropriate to use these two drugs for the destruction of dormant cells. The pharmacological action of Artenimol and Copper on target genes is unknown. Therefore, further research is needed to propose that these drugs can be used to destroy dormant cells.Fostamatinib has inhibitor effect on all target genes. Therefore, Fostamatinib is a strong candidate for destroying cancer cells.

7. CONCLUSIONS

The analysis of genes that change significantly in the transcriptome data of disease and healthy conditions is a widely used method to understand mechanism of diseases [Jiang et al., 2015]. However, more complex methods are needed to elucidate the molecular mechanisms of complex organisms affected in diseases. Nowadays, network-based analysis of transcriptome data is a commonly used method to investigate the etiology of diseases [Batra et al., 2017]. In this approach, the information of the gene expression is combined with the molecular interaction information of the organism of interest. Molecular interactions can be protein interactions, transcription factor-gene interactions, or metabolomic interactions.

Within the scope of this thesis, transcriptome data of different types of cancer in human and mouse were used to perform differential expression analysis of dormancy and cancer conditions, and obtained p-values were mapped on protein-protein interaction and gene-regulatory networks of the relevant organism by two different subnetwork discovery tools, KPM and BioNet, to extract sub-networks that are enrichd in significantly changed genes. The results show that the two different subnetwork discovery methods KPM and BioNet give similar results.KPM performs better in finding subnetworks in some cases where BioNet cannot find a subnetwork.

In this study, 10 different datasets and 21 different dormancy-cancer comparisons were analysed and subnetworks of PPI and GR networks were identified. Then, the genes in the subnetwork were combined to determine the most common genes in all subnetworks. The analyzed datasets belong to seven different cancer types. Therefore, the genes which are found to be significantly changed in many subnetworks have changed in dormancy in different cancer types. It is thought that the identified genes with high score in different subnetworks can be used as a drug target or biomarker for the destruction of dormant cells in cancer. The killing of dormant cells can reduce the risk of metastasis after treatment and increase survival time of patients. For this reason, drug re-positioning analysis was performed in order to find drugs that target important genes in dormancy. The identified drugs as a result of this analysis can be used to destroy dormant cancer cells. The next step is to perform experimental validation of the effect of the identified drugs.

Furthermore, the gene expression data of dormant cancer cells is very limited in the literature. In this thesis, all datasets found in the literature containing dormant and proliferative samples that are suitable for computational analysis were used. However, due to limited data, cancer-specific analyses or metastatic-level specific analyses could not be performed. In future studies, with the increase in the data set in the literature and the development of computational methods, more specific analyses will contribute to a better understanding of the dormancy mechanism.

REFERENCES

Aguirre-Ghiso J. A., Ossowski L., Rosenbaum S. K., (2004), “Green Fluorescent Protein Tagging of Extracellular Signal-Regulated Kinase and p38 Pathways Reveals Novel Dynamics of Pathway Activation during Primary and Metastatic Growth”, Cancer Research, 64 (20), 7336.

Al-Asadi M. G., Brindle G., Castellanos M., May S. T., Mills K. I., Russell N. H., Seedhouse C. H., Pallis M., (2017), “A molecular signature of dormancy in CD34(+)CD38(-) acute myeloid leukaemia cells”, Oncotarget, 8 (67), 111405-111418.

Alcaraz N., Friedrich T., Kotzing T., Krohmer A., Muller J., Pauling J., Baumbach J., (2012), “Efficient key pathway mining: combining networks and OMICS data”, Integr Biol (Camb), 4 (7), 756-764.

Alcaraz N., List M., Dissing-Hansen M., Rehmsmeier M., Tan Q., Mollenhauer J., Ditzel H. J., Baumbach J., (2016), “Robust de novo pathway enrichment with KeyPathwayMiner 5”, F1000Res, 5, 1531.

Amberger J. S., Bocchini C. A., Schiettecatte F., Scott A. F., Hamosh A., (2015), “OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders”, Nucleic Acids Research, 43 (Database issue), D789-D798.

Angelini C., Costa V., (2014), “Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems”, Frontiers in Cell and Developmental Biology, 2 (51).

Ashburn T. T., Thor K. B., (2004), “Drug repositioning: identifying and developing new uses for existing drugs”, Nat Rev Drug Discov, 3 (8), 673-683.

Ayob A. Z., Ramasamy T. S., (2018), “Cancer stem cells as key drivers of tumour progression”, J Biomed Sci, 25 (1), 20.

Barrett T., Wilhite S. E., Ledoux P., Evangelista C., Kim I. F., Tomashevsky M., Marshall K. A., Phillippy K. H., Sherman P. M., Holko M., Yefanov A., Lee H., Zhang N., Robertson C. L., Serova N., Davis S., Soboleva A., (2012), “NCBI GEO: archive for functional genomics data sets—update”, Nucleic Acids Research, 41 (D1), D991-D995.

Barrett T., Wilhite S. E., Ledoux P., Evangelista C., Kim I. F., Tomashevsky M., Marshall K. A., Phillippy K. H., Sherman P. M., Holko M., Yefanov A., Lee H., Zhang N., Robertson C. L., Serova N., Davis S., Soboleva A., (2013), “NCBI GEO: archive for functional genomics data sets--update”, Nucleic acids research, 41 (Database issue), D991-D995.

Batra R., Alcaraz N., Gitzhofer K., Pauling J., Ditzel H. J., Hellmuth M., Baumbach J., List M., (2017), “On the performance of de novo pathway enrichment”, NPJ Syst Biol Appl, 3, 6.

Behjati S., Tarpey P. S., (2013), “What is next generation sequencing?”, Arch Dis Child Educ Pract Ed, 98 (6), 236-238.

Beisser D., Klau G. W., Dandekar T., Muller T., Dittrich M. T., (2010), “BioNet: an R-Package for the functional analysis of biological networks”, Bioinformatics, 26 (8), 1129-1130.

Benjamini Y., Hochberg Y., (1995), “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”, Journal of the Royal Statistical Society. Series B (Methodological), 57 (1), 289-300.

Bian B., Mongrain S., Cagnol S., Langlois M. J., Boulanger J., Bernatchez G., Carrier J. C., Boudreau F., Rivard N., (2016), “Cathepsin B promotes colorectal tumorigenesis, cell invasion, and metastasis”, Mol Carcinog, 55 (5), 671-687.

[Bishop E., Bradshaw T. D., (2018), “Autophagy modulation: a prudent approach in cancer treatment?”, Cancer Chemother Pharmacol.

Bray F., Ferlay J., Soerjomataram I., Siegel R. L., Torre L. A., Jemal A., (2018), “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”, CA Cancer J Clin, 68 (6), 394-424.

Buczacki S. J. A., Popova S., Biggs E., Koukorava C., Buzzelli J., Vermeulen L., Hazelwood L., Francies H., Garnett M. J., Winton D. J., (2018), “Itraconazole targets cell cycle heterogeneity in colorectal cancer”, The Journal of experimental medicine, 215 (7), 1891-1912.

Burley S. K., Berman H. M., Bhikadiya C., Bi C., Chen L., Di Costanzo L., Christie C., Dalenberg K., Duarte J. M., Dutta S., Feng Z., Ghosh S., Goodsell D. S., Green R. K., Guranović V., Guzenko D., Hudson B. P., Kalro T., Liang Y., Lowe R., Namkoong H., Peisach E., Periskova I., Prlić A., Randle C., Rose A., Rose P., Sala R., Sekharan M., Shao C., Tan L., Tao Y.-P., Valasatava Y., Voigt M., Westbrook J., Woo J., Yang H., Young J., Zhuravleva M., Zardecki C., (2018), “RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy”, Nucleic Acids Research, 47 (D1), D464-D474.

Carbon S., Ireland A., Mungall C. J., Shu S., Marshall B., Lewis S., Ami G. O. H., Web Presence Working G., (2009), “AmiGO: online access to ontology and annotation data”, Bioinformatics, 25 (2), 288-289.

Chatr-aryamontri A., Ceol A., Palazzi L. M., Nardelli G., Schneider M. V., Castagnoli L., Cesareni G., (2007), “MINT: the Molecular INTeraction database”, Nucleic Acids Research, 35 (Database), D572-D574.

Chatr-aryamontri A., Oughtred R., Boucher L., Rust J., Chang C., Kolas N. K., O'Donnell L., Oster S., Theesfeld C., Sellam A., Stark C., Breitkreutz B.-J., Dolinski K., Tyers M., (2017), “The BioGRID interaction database: 2017 update”, Nucleic Acids Research, 45 (D1), D369-D379.

Chen X., Yu X., Shen E., (2019), “Overexpression of CDKN2B is involved in poor gastric cancer prognosis”, J Cell Biochem, 120 (12), 19825-19831.

Chuang H. Y., Lee E., Liu Y. T., Lee D., Ideker T., (2007), “Network-based classification of breast cancer metastasis”, Molecular Systems Biology, 3 (1).

Clark K., Karsch-Mizrachi I., Lipman D. J., Ostell J., Sayers E. W., (2016), “GenBank”, Nucleic Acids Research, 44 (D1), D67-D72.

Cline M. S., Smoot M., Cerami E., Kuchinsky A., Landys N., Workman C., Christmas R., Avila-Campilo I., Creech M., Gross B., Hanspers K., Isserlin R., Kelley R., Killcoyne S., Lotia S., Maere S., Morris J., Ono K., Pavlovic V., Pico A. R., Vailaya A., Wang P.-L., Adler A., Conklin B. R., Hood L., Kuiper M., Sander C., Schmulevich I., Schwikowski B., Warner G. J., Ideker T., Bader G. D., (2007), “Integration of biological networks and gene expression data using Cytoscape”, Nature protocols, 2 (10), 2366-2382.

Consortium T. U., (2018), “UniProt: a worldwide hub of protein knowledge”, Nucleic Acids Research, 47 (D1), D506-D515.

Curtis S. A., (2003), “The classification of greedy algorithms”, Science of Computer Programming, 49 (1-3), 125-157.

Dasgupta A., Lim A. R., Ghajar C. M., (2017), “Circulating and disseminated tumor cells: harbingers or initiators of metastasis?”, Mol Oncol, 11 (1), 40-61.

Davis S., Meltzer P. S., (2007), “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor”, Bioinformatics, 23 (14), 1846-1847.

Deepak Roshan V. G., Sinto M. S., Vargees B. T., Kannan S., (2019), “Loss of CDKN2A and CDKN2B expression is associated with disease recurrence in oral cancer”, J Oral Maxillofac Pathol, 23 (1), 82-89.

Dey-Guha I., Wolfer A., Yeh A. C., G Albeck J., Darp R., Leon E., Wulfkuhle J., Petricoin E. F., 3rd, Wittner B. S., Ramaswamy S., (2011), “Asymmetric cancer cell division regulated by AKT”, Proceedings of the National Academy of Sciences of the United States of America, 108 (31), 12845-12850.

Dudley J. T., Deshpande T., Butte A. J., (2011), “Exploiting drug-disease relationships for computational drug repositioning”, Brief Bioinform, 12 (4), 303-311.

Ebinger S., Özdemir E. Z., Ziegenhain C., Tiedt S., Castro Alves C., Grunert M., Dworzak M., Lutz C., Turati V. A., Enver T., Horny H.-P., Sotlar K., Parekh S., Spiekermann K., Hiddemann W., Schepers A., Polzer B., Kirsch S., Hoffmann M., Knapp B., Hasenauer J., Pfeifer H., Panzer-Grümayer R., Enard W., Gires O., Jeremias I., (2016), “Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia”, Cancer cell, 30 (6), 849-862.

Emanetçi E., (2019), “Network-based analysis of cognitive impairment and memory deficits from transcriptome data / Transkriptom verileri kullanılarak hafıza problemlerinin ve bilişsel bozuklukların hücresel ağlara dayalı analizi”, Gebze Technical University.

Endo H., Inoue M., (2019), “Dormancy in cancer”, Cancer Sci, 110 (2), 474-480.

Fionda V. (2019), “Networks in Biology”. “Encyclopedia of Bioinformatics and Computational Biology”.

Folador E. L., Tiwari S., Da Paz Barbosa C. E., Jamal S. B., Da Costa Schulze M., Barh D., Azevedo V. (2019), “Protein-Protein Interactions: An Overview”. “Encyclopedia of Bioinformatics and Computational Biology”.

Garnham R., Scott E., Livermore K., Munkley J., (2019), “ST6GAL1: A key player in cancer (Review)”, Oncology Letters.

Gomis R. R., Gawrzak S., (2016), “Tumor cell dormancy”, Mol Oncol.

Halfon M. S., (2017), “Perspectives on Gene Regulatory Network Evolution”, Trends Genet, 33 (7), 436-447.

Han H., Cho J.-W., Lee S., Yun A., Kim H., Bae D., Yang S., Kim C. Y., Lee M., Kim E., Lee S., Kang B., Jeong D., Kim Y., Jeon H.-N., Jung H., Nam S., Chung M., Kim J.-H., Lee I., (2018), “TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions”, Nucleic Acids Research, 46 (D1), D380-D386.

Hermjakob H., (2004), “IntAct: an open source molecular interaction database”, Nucleic Acids Research, 32 (90001), 452D-455.

Huang L., Liao L., Wu C. H., (2016), “Inference of protein-protein interaction networks from multiple heterogeneous data”, EURASIP J Bioinform Syst Biol, 2016 (1), 8.

Hurst R. E., Hauser P. J., Kyker K. D., Heinlen J. E., Hodde J. P., Hiles M. C., Kosanke S. D., Dozmorov M., Ihnat M. A., (2013), “Suppression and activation of the malignant phenotype by extracellular matrix in xenograft models of bladder cancer: a model for tumor cell “dormancy”“, PLoS One, 8 (5), e64181.

Jiang Z., Zhou X., Li R., Michal J. J., Zhang S., Dodson M. V., Zhang Z., Harland R. M., (2015), “Whole transcriptome analysis with sequencing: methods, challenges and potential solutions”, Cellular and molecular life sciences : CMLS, 72 (18), 3425-3439.

Jie X.-X., Zhang X.-Y., Xu C.-J., (2017), “Epithelial-to-mesenchymal transition, circulating tumor cells and cancer metastasis: Mechanisms and clinical applications”, Oncotarget, 8 (46), 81558-81571.

Jin R., Chen X., Han D., Luo X., Li H., (2017), “Clusterin modulates transdifferentiation of non-small-cell lung cancer”, BMC Cancer, 17 (1), 661.

Kanehisa M., Sato Y., Kawashima M., Furumichi M., Tanabe M., (2015), “KEGG as a reference resource for gene and protein annotation”, Nucleic Acids Research, 44 (D1), D457-D462.

Kassambara A., Mundt F., factoextra: Extract and Visualize the Results of Multivariate Data Analyses, in: (Ed.)^(Eds.), R package, 2019, pp.

Kedaigle A. J., Fraenkel E. (2018), “Discovering Altered Regulation and Signaling

Through Network-based Integration of Transcriptomic, Epigenomic, and Proteomic Tumor Data”. “Cancer Systems Biology”.

Kinsella R. J., Kahari A., Haider S., Zamora J., Proctor G., Spudich G., Almeida-King J., Staines D., Derwent P., Kerhornou A., Kersey P., Flicek P., (2011), “Ensembl BioMarts: a hub for data retrieval across taxonomic space”, Database (Oxford), 2011, bar030.

La Belle Flynn A., Calhoun B. C., Sharma A., Chang J. C., Almasan A., Schiemann W. P., (2019), “Autophagy inhibition elicits emergence from metastatic dormancy by inducing and stabilizing Pfkfb3 expression”, Nature Communications, 10 (1), 3668-3668.

Lawson M. A., McDonald M. M., Kovacic N., Hua Khoo W., Terry R. L., Down J., Kaplan W., Paton-Hough J., Fellows C., Pettitt J. A., Neil Dear T., Van Valckenborgh E., Baldock P. A., Rogers M. J., Eaton C. L., Vanderkerken K., Pettit A. R., Quinn J. M. W., Zannettino A. C. W., Phan T. G., Croucher P. I., (2015), “Osteoclasts control reactivation of dormant myeloma cells by remodelling the endosteal niche”, Nature Communications, 6, 8983-8983.

Liu Q., Zhang H., Jiang X., Qian C., Liu Z., Luo D., (2017), “Factors involved in cancer metastasis: a better understanding to “seed and soil” hypothesis”, Mol Cancer, 16 (1), 176.

Liu Z.-P., Wu C., Miao H., Wu H., (2015), “RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse”, Database, 2015.

Lotfi Shahreza M., Ghadiri N., Mousavi S. R., Varshosaz J., Green J. R., (2018), “A review of network-based approaches to drug repositioning”, Brief Bioinform, 19 (5), 878-892.

Love M. I., Huber W., Anders S., (2014), “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”, Genome Biology, 15 (12), 550.

Lu X., Mu E., Wei Y., Riethdorf S., Yang Q., Yuan M., Yan J., Hua Y., Tiede B. J., Lu X., Haffty B. G., Pantel K., Massagué J., Kang Y., (2011), “VCAM-1 promotes osteolytic expansion of indolent bone micrometastasis of breast cancer by engaging α4β1-positive osteoclast progenitors”, Cancer cell, 20 (6), 701-714.

Lu Z., Luo R. Z., Lu Y., Zhang X., Yu Q., Khare S., Kondo S., Kondo Y., Yu Y., Mills G. B., Liao W. S. L., Bast R. C., Jr., (2008), “The tumor suppressor gene ARHI regulates autophagy and tumor dormancy in human ovarian cancer cells”, The Journal of clinical investigation, 118 (12), 3917-3929.

Mitra K., Carvunis A. R., Ramesh S. K., Ideker T., (2013), “Integrative approaches for finding modular structure in biological networks”, Nat Rev Genet, 14 (10), 719-732.

Moridi M., Ghadirinia M., Sharifi-Zarchi A., Zare-Mirakabad F., (2019), “The assessment of efficient representation of drug features using deep learning for drug repositioning”, BMC bioinformatics, 20 (1), 577.

Mudunuri U., Che A., Yi M., Stephens R. M., (2009), “bioDBnet: the biological database network”, Bioinformatics, 25 (4), 555-556.

Mutch D. G., Babb S. A., Di Saia P. J. (2018), “19 - Genes and Cancer: Genetic Counseling and Clinical Management”. In: P. J. DiSaia, W. T. Creasman, R. S. Mannel, D. S. McMeekin and D. G. Mutch, Editors, “Clinical Gynecologic Oncology (Ninth Edition)”, Elsevier.

Nakanishi H., Kodera Y., Tatematsu M., (2004), “Molecular Method to Quantitatively Detect Micrometastases and its Clinical Significance in Gastrointestinal Malignancies”, Advances in clinical chemistry, 38, 87-110.

Nayak S. R., Harrington E., Boone D., Hartmaier R., Chen J., Pathiraja T. N., Cooper K. L., Fine J. L., Sanfilippo J., Davidson N. E., Lee A. V., Dabbs D., Oesterreich S., (2015), “A Role for Histone H2B Variants in Endocrine-Resistant Breast Cancer”, Horm Cancer, 6 (5-6), 214-224.

Paez D., Labonte M. J., Bohanes P., Zhang W., Benhanim L., Ning Y., Wakatsuki T., Loupakis F., Lenz H. J., (2012), “Cancer dormancy: a model of early dissemination and late cancer recurrence”, Clin Cancer Res, 18 (3), 645-653.

Pagès H., Carlson M., Falcon S., Li N., AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor., in: (Ed.)^(Eds.), 2019, pp.

Patterson S. G., Wei S., Chen X., Sallman D. A., Gilvary D. L., Zhong B., Pow-Sang J., Yeatman T., Djeu J. Y., (2006), “Novel role of Stat1 in the development of docetaxel resistance in prostate tumor cells”, Oncogene, 25 (45), 6113-6122.

Raudvere U., Kolberg L., Kuzmin I., Arak T., Adler P., Peterson H., Vilo J., (2019), “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)”, Nucleic Acids Res, 47 (W1), W191-W198.

Razick S., Magklaras G., Donaldson I. M., (2008), “iRefIndex: a consolidated protein interaction database with provenance”, BMC bioinformatics, 9, 405-405.

Risso D., Perraudeau F., Gribkova S., Dudoit S., Vert J.-P., (2018), “A general and flexible method for signal extraction from single-cell RNA-seq data”, Nature Communications, 9 (1), 284.

Ritchie M. E., Phipson B., Wu D., Hu Y., Law C. W., Shi W., Smyth G. K., (2015), “limma powers differential expression analyses for RNA-sequencing and microarray studies”, Nucleic Acids Research, 43 (7), e47-e47.

Ruppender N., Larson S., Lakely B., Kollath L., Brown L., Coleman I., Coleman R., Nguyen H., Nelson P. S., Corey E., Snyder L. A., Vessella R. L., Morrissey C., Lam H.-M., (2015), “Cellular Adhesion Promotes Prostate Cancer Cells Escape from Dormancy”, PloS one, 10 (6), e0130565-e0130565.

Sahu S. K., Tiwari N., Pataskar A., Zhuang Y., Borisova M., Diken M., Strand S., Beli P., Tiwari V. K., (2017), “FBXO32 promotes microenvironment underlying epithelial-mesenchymal transition via CtBP1 during tumour metastasis and brain development”, Nat Commun, 8 (1), 1523.

Salgado-Albarrán M., González-Barrios R., Guerra-Calderas L., Alcaraz N., Estefanía Sánchez-Correa T., Castro-Hernández C., Sánchez-Pérez Y., Aréchaga-Ocampo E., García-Carrancá A., Cantú de León D., Herrera L. A., Baumbach J., Soto-Reyes E., (2019), "The epigenetic factor BORIS (CTCFL) controls the androgen receptor regulatory network in ovarian cancer", Oncogenesis, 8 (8).

Schena M., Shalon D., Davis R. W., Brown P. O., (1995), "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray", Science, 270 (5235), 467.

Sealfon S. C., Chu T. T. (2011), "RNA and DNA Microarrays". "Biological Microarrays".

Senft D., Ronai Z. e. A., (2016), "Adaptive Stress Responses During Tumor Metastasis and Dormancy", Trends in Cancer, 2 (8), 429-442.

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

Shibue T., Weinberg R. A., (2009), "Integrin β1-focal adhesion kinase signaling directs the proliferation of metastatic cancer cells disseminated in the lungs", Proceedings of the National Academy of Sciences, 106 (25), 10290.

Shiozawa Y., Berry J. E., Eber M. R., Jung Y., Yumoto K., Cackowski F. C., Yoon H. J., Parsana P., Mehra R., Wang J., McGee S., Lee E., Nagrath S., Pienta K. J., Taichman R. S., (2016), "The marrow niche controls the cancer stem cell phenotype of disseminated prostate cancer", Oncotarget, 7 (27), 41217-41232.

Singh S. S., Vats S., Chia A. Y., Tan T. Z., Deng S., Ong M. S., Arfuso F., Yap C. T., Goh B. C., Sethi G., Huang R. Y., Shen H. M., Manjithaya R., Kumar A. P., (2018), "Dual role of autophagy in hallmarks of cancer", Oncogene, 37 (9), 1142-1158.

Vera-Ramirez L., Vodnala S. K., Nini R., Hunter K. W., Green J. E., (2018), "Autophagy promotes the survival of dormant breast cancer cells and metastatic tumour recurrence", Nat Commun, 9 (1), 1944.

Wang Z., Gerstein M., Snyder M., (2009), "RNA-Seq: a revolutionary tool for transcriptomics", Nature Reviews Genetics, 10 (1), 57-63.

Wilson M. R., Zoubeidi A., (2017), "Clusterin as a therapeutic target", Expert Opin Ther Targets, 21 (2), 201-213.

Wishart D. S., Feunang Y. D., Guo A. C., Lo E. J., Marcu A., Grant J. R., Sajed T., Johnson D., Li C., Sayeeda Z., Assempour N., Iynkkaran I., Liu Y., Maciejewski A., Gale N., Wilson A., Chin L., Cummings R., Le D., Pon A., Knox C., Wilson M., (2018), "DrugBank 5.0: a major update to the DrugBank database for 2018", Nucleic Acids Research, 46 (D1), D1074-D1082.

Xue H., Li J., Xie H., Wang Y., (2018), "Review of Drug Repositioning Approaches and Resources", Int J Biol Sci, 14 (10), 1232-1244.

Yadav A. S., Pandey P. R., Butti R., Radharani N. N. V., Roy S., Bhalara S. R., Gorain M., Kundu G. C., Kumar D., (2018), “The Biology and Therapeutic Implications of Tumor Dormancy and Reactivation”, Frontiers in Oncology, 8 (72).

Yang Y., Yang W., Jin L., (2019), “The Role of Long Non-coding RNA Prostate Cancer-Associated Transcript 1 in Prostate Cancer”, Journal of Computational Biology.

Yeh A. C., Ramaswamy S., (2015), “Mechanisms of Cancer Cell Dormancy--Another Hallmark of Cancer?”, Cancer Res, 75 (23), 5014-5022.

BIOGRAPHY

Dilara Uzuner was born in Üsküdar in October 19, 1995. She graduated from Gebze Technical University Department of Molecular Biology and Genetics in 2017. She is an M.Sc. student in Bioinformatics and Systems Biology Program under Bioengineering Department at Gebze Technical University Graduate School of Natural and Applied Sciences.

APPENDIX

Appendix A: Other Appendix (CD)