

**THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY**

**INFERENCE OF DIFFERENTIAL GENE
NETWORKS**

PhD Thesis

ONUR MENDİ

ISTANBUL, 2016

**THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY**

**GRADUATE SCHOOL OF NATURAL
AND APPLIED SCIENCES
COMPUTER ENGINEERING PHD PROGRAM**

**INFERENCE OF
DIFFERENTIAL GENE NETWORKS**

PhD Thesis

ONUR MENDİ

Thesis Supervisor: PROF.DR. ADEM KARAHOCA

ISTANBUL, 2016

**THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY**

**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
PHD IN COMPUTER ENGINEERING**

Name of the thesis: Inference of Differential Gene Networks

Name/Last Name of the Student: Onur MENDİ

Date of the Defense of Thesis:

The thesis has been approved by the Graduate School of Natural and Applied Sciences

Prof.Dr. Nafiz ARICA
Graduate School Director

I certify that this thesis meets all the requirements as a thesis for the degree of Ph.D.

Asst. Prof. Tarkan AYDIN
Program Coordinator

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Ph.D.

Examining Committee Members

Signature

Thesis Supervisor
Prof.Dr. Adem KARAHOCA

Member
Prof.Dr. Salih OFLUOĞLU

Member
Asst. Prof. Tarkan AYDIN

Member
Asst. Prof. Cemal Okan ŞAKAR

Member
Asst. Prof. Oğuz KARAN

DEDICATION

This dissertation is dedicated to my lovely wife and to my family for their unconditional love and support.

Istanbul, 2016

Onur Mendi

ABSTRACT

INFERENCE OF DIFFERENTIAL GENE NETWORKS

Onur Mendi

PhD in Computer Engineering

Thesis Supervisor: Prof.Dr. Adem Karahoca

December 2016, 97 Pages

Biological systems are highly dynamic entities that behave differently under different conditions. Differential gene network analysis reveals disease-specific gene interactions from expression datasets that helps identifying the molecular interactions that underlies the progression of diseases. The purpose of this study is to present a novel differential networking approach that integrates disease-specific differential gene network with the prior biological knowledge to reveal the molecular mechanisms associated with breast cancer.

In the study, METABRIC breast cancer dataset is used to infer genome-wide breast cancer specific differential gene network. A web-based tool was developed to infer breast cancer specific gene network. In order to evaluate the results of the study, functional enrichment analyses were performed. GO and KEGG pathway enrichment analysis identified numerous pathways that may have a role in the breast cancer. Furthermore, the top genes that are identified in the breast cancer specific differential network are investigated through the literature. The findings of this study may promote the better understanding about the molecular mechanism of breast cancer and also disclose potential targets for diagnostic and effective therapies.

Keywords: Differential Network Analysis, Differential Gene Networks, Systems Biology, Disease Networks.

ÖZET

FARKSAL GEN AĞLARI ÇIKARIMI

Onur Mendi

Bilgisayar Mühendisliği Doktora Programı

Tez Danışmanı: Prof.Dr. Adem Karahoca

Aralık 2016, 97 Sayfa

Biyolojik sistemler değişik durumlarda farklı davranışlar gösteren oldukça dinamik yapılardır. Farksal gen ağları analizi, ekspresyon verisini kullanarak hastalıkların ilerlemesine neden olan moleküler etkileşimlerin belirlenmesinde önemli rol oynayan hastalığa özel gen etkileşimlerini ortaya çıkaran bir analiz türüdür. Bu çalışmanın amacı hastalığa özel farksal gene ağları ile literatürdeki biyolojik bilgileri entegre ederek göğüs kanseri ile ilişkili moleküler mekanizmaları ortaya çıkaran özgün bir farksal ağ yaklaşımı geliştirmektir.

Bu tez çalışmasında, genom seviyesinde göğüs kanserine özel farksal gen ağı çıkarımında METABRIC göğüs kanseri veri seti kullanıldı. Bu kapsamda göğüs kanserine özel gen ağı çıkaran web-tabanlı bir uygulama geliştirildi. Çalışmada elde edilen sonuçların değerlendirilmesinde, fonksiyonel zenginleştirme analizleri kullanıldı. GO ve KEGG yolak analizleri ile göğüs kanserinde önemli rol oynayan çeşitli yolaklar tespit edildi. Buna ek olarak göğüs kanserine özel farksal gen ağında önemli rol oynayan genler literatür kapsamında değerlendirildi. Bu çalışmanın bulguları, göğüs kanserinin gelişiminde rol oynayan mekanizmaların daha iyi anlaşılabilmesinin yanı sıra, hastalık tanısı ve etkili tedavi geliştirilmesinde potansiyel hedef genlerin tespitine katkı sağlayabilir.

Anahtar Kelimeler: Farksal Ağ Analizi, Farksal Gen Ağları, Sistem Biyoloji, Hastalık Ağları.

CONTENTS

TABLES	ix
FIGURES	x
ABBREVIATIONS	xii
SYMBOLS	xiii
1. INTRODUCTION	1
2. LITERATURE REVIEW	4
2.1. BIOLOGICAL BACKGROUND	4
2.1.1. Cancer	4
2.1.2. Breast Cancer	5
2.1.3. Microarray Data	6
2.1.4. Gene Regulatory Networks	8
2.1.5. Gene Characteristics	9
2.1.5.1. Transcription Factors	10
2.1.5.2. Metastatic genes	11
2.1.5.3. Prognostic genes	12
2.1.5.4. Oncogenes	12
2.2. GENE NETWORK INFERENCE	13
2.2.1. Introduction	13
2.2.2. Methods	15
2.2.2.1. Relevance Network	15
2.2.2.2. ARACNE	16
2.2.2.3. CLR	18
2.2.2.4. MRNET	18
2.2.2.5. C3NET	20
2.3. DIFFERENTIAL ANALYSIS	23
2.3.1. Single Gene Analysis	23
2.3.1.1. Differential Expression	23
2.3.1.2. Differential Variability	24
2.3.1.3. Differential Correlation	25
2.3.2. Gene-pair Analysis	25

2.3.2.1. Differential Co-expression	25
2.3.2.2. Differential Co-clustering.....	27
2.3.3. Differential Network Analysis.....	29
3. METHODOLOGY.....	35
3.1. MOTIVATION	35
3.2. THE PROPOSED IDN ALGORITHM.....	36
3.2.1. Introduction	36
3.2.2. Parameters	40
3.2.3. Gene Ontology Analysis.....	42
3.3. IMPLEMENTATION	42
3.3.1. The IDN R Package.....	42
3.3.1.1. Installation of the IDN R package	42
3.3.1.2. General guidelines for using the IDN R package.....	43
3.3.2. The IDN Web Application.....	44
3.3.2.1. System Architecture	44
3.3.2.2. Workflow	45
3.3.2.3. A Case Study of IDN Web Application.....	47
4. RESULTS	53
4.1. PRELIMINARY ANALYSES.....	53
4.1.1. Inference of Prostate Cancer Specific Differential Network.....	53
4.1.1.1. Microarray data.....	55
4.1.1.2. Results	55
4.1.2. Inference of Lung Cancer Specific Differential Network.....	64
4.1.2.1. Microarray data.....	64
4.1.2.2. Results	64
4.2. BREAST CANCER ANALYSIS	74
4.2.1. Microarray Data.....	74
4.2.2. Data Preprocessing.....	74
4.2.3. Data Integration	75

4.2.3.1. Transcription Factors.....	75
4.2.3.2. Identification of Differentially Expressed Genes	76
4.2.3.3. Prognostic Genes.....	77
4.2.3.4. Oncogenes	78
4.2.3.5. Metastatic Genes	79
4.2.4. Inference of Breast Cancer Specific Differential Network.....	80
4.2.5. Integration and ranking process.....	88
5. DISCUSSION	92
REFERENCES.....	98

TABLES

Table 2.1. Classification of transcription factor families	10
Table 4.1. GO terms of androgen stimulated prostate cancer specific differential network (top 10)	56
Table 4.2. Significant KEGG pathways in the androgen stimulated prostate cancer specific differential network.....	58
Table 4.3. Significant KEGG pathways in the largest subnetwork of the androgen stimulated prostate cancer specific differential network.....	59
Table 4.4. GO terms of non-small cell lung cancer specific differential network (top 10).....	66
Table 4.5. Significant KEGG pathways in the NSCLC specific differential network..	68
Table 4.6. Significant KEGG pathways in the largest subnetwork of the NSCLC specific differential network.....	70
Table 4.7. METABRIC Breast Cancer Datasets	74
Table 4.8. Differentially Expressed Genes of Metabrib Breast Cancer Dataset	76
Table 4.9. The prognostic gene list in Breast Cancer	77
Table 4.10. Metastatic genes associated with breast cancer	79
Table 4.11. Top five hub genes in the breast cancer specific differential network.....	82
Table 4.12. GO terms of breast cancer specific differential network (top 10).....	83
Table 4.13. Significant KEGG pathways in the breast cancer specific differential network (Top 10).....	85
Table 4.14. Significant KEGG pathways in the largest independent subnetwork of breast cancer specific differential network	88
Table 4.15. The top 20 genes identified by IDN framework that may have a crucial role in the progression of breast cancer	89
Table 4.16. The existence status of top 10 genes identified by IDN framework in the known breast cancer lists.....	91

FIGURES

Figure 2.1. The central dogma	7
Figure 2.2. A schematic view of a gene regulatory networks.....	9
Figure 2.3. The workflow of gene network inference.....	14
Figure 2.4. Working mechanism of DPI. DPI infer the most likely path of information flow regardless of significant mutual information values of all six gene pairs.....	17
Figure 2.5. Determining MI threshold, I_C	21
Figure 2.6. Visualization of the steps of C3NET	22
Figure 2.7. Instance of different gene expression correlations with the same mean expression levels.	27
Figure 2.8. Visualization of differential analysis methods of gene expression data.....	28
Figure 3.1. IDN framework overview.....	37
Figure 3.2. IDN algorithm for breast cancer.....	39
Figure 3.3. Three-tier architecture of IDN web application.....	44
Figure 3.4. Workflow of IDN-web	46
Figure 3.5. Differential gene network inference: Step 1	47
Figure 3.6. Differential gene network inference: Step 2.....	48
Figure 3.7. Differential gene network inference: Step 3.....	49
Figure 3.8. Differential gene network inference: Step 6 - Parameter selection.	50
Figure 3.9. Differential gene network inference successfully submitted.....	51
Figure 3.10. Track transaction.	51
Figure 3.11. Transaction is completed.	52
Figure 4.1. Genome-wide androgen stimulated prostate specific differential network with 891 interactions	54
Figure 4.2. Functional enrichment analysis of significantly enriched genes in the androgen stimulated prostate cancer specific differential gene network.	57
Figure 4.3. The largest connected subnetwork of the androgen stimulated prostate cancer difnet	59

Figure 4.4. Genome-wide NSCLC specific differential network with 804 interactions.	65
Figure 4.5. Functional enrichment analysis of significantly enriched genes in the NSCLC specific differential gene network	68
Figure 4.6. The largest connected subnetwork of NSCLC specific differential gene network.....	70
Figure 4.7. Breast cancer specific differential network with 1525 interactions.....	81
Figure 4.8. Top five hub genes and their targets in the breast cancer specific differential network.....	82
Figure 4.9. Functional enrichment analysis of significantly enriched genes in the breast cancer specific differential gene network.....	85
Figure 4.10. Largest subnetwork of breast cancer specific differential network with 200 interactions	87
Figure 4.11. Targets of the transcription factor YY1 (69 interactions)	90

ABBREVIATIONS

DE	:	Differential Expression
GNI	:	Gene Network Inference
GRN	:	Gene Regulatory Network
DGNI	:	Differential Gene Network Inference
MI	:	Mutual Information
MTC	:	Multiple Testing Correction
PLS	:	Partial Least Squares
TF	:	Transcription Factor

SYMBOLS

Threshold : I_c

1. INTRODUCTION

Nowadays, the identification of novel oncogenes or tumor suppressor genes has become popular in tumorigenesis studies in understanding molecular mechanisms that drive disease progression (Ren, 2015). Understanding the working mechanism of molecules in normal cell physiology and pathogenesis allows subtle drug development and helps treatment of a disease, such as cancer (Altay and Emmert-Streib, 2010a; Rual et al., 2005;). The advent of systems and network biology enable us to capture interactions occurring within a cell, which can be represented as gene networks. Computational analysis of the networks provides key insights into complex biological systems and cellular organization.

Gene and protein interaction networks can be constructed for a particular biological condition experimentally or computationally. The most commonly used method to construct a network experimentally is the yeast two-hybrid system. However, this method produces a very large number of false positives and receive significant criticism. Besides this, constructing the protein and gene association networks experimentally is an expensive and labor-intensive process. Hence, computational approaches to reverse engineer the protein and gene association networks became popular and used widely as a lucrative alternative (Gill et al., 2014a).

The inference of gene regulatory networks (GRN) is a process of estimating direct physical associations among genes from microarray gene expression data (Emmert-Streib et al., 2012). GRNs aim to capture the interactions between the molecular structures and are represented as graphs in which nodes represent genes, and edges represent molecular associations (Hecker et al., 2009). Various gene network inference algorithms are available in the literature to infer GRNs using gene expression data. Some of the most popular methods are ARACNE, CLR, C3NET, and MRNET (Margolin et al., 2006; Faith et al., 2007; Altay & Emmert-Streib, 2010a,b; Meyer et al., 2007). This is an active research area and apart from these popular ones, many other algorithms also exist.

However, both experimental and computational methods that are used for constructing gene and protein networks are static in nature. The biological systems are highly dynamic entities and the perturbations caused by environmental stresses, evolutionary changes and disease conditions result in changes in the topology of the networks. For this reason, examining the network structure under different biological settings became important to identify which parts of the network get affected by the perturbation (Gill et al., 2014a). Some examples of such different biological settings are as following:

- a. Different tissue types: e.g., normal vs. cancer (Lu et al., 2010; Ergun et al., 2007)
- b. Different stages of cancer: e.g., breast cancer stage I vs. stage IV (Iqbal et al., 2015)
- c. Different cancer subtype: e.g., squamous cell carcinoma vs adenocarcinoma non-small cell lung cancer (Bartucci et al., 2012).
- d. Different time points: e.g., two distinct time periods (Gill et al., 2010)
- e. Different subject type: e.g., male vs. female (Van Nas et al., 2009)
- f. Different race: e.g., White women vs. African American women (White-Means et al., 2015)

The biological activities at the gene level are very complex structures as genes interact with each other in a dynamic manner. A single gene can play a role in different stages of biological activities and regulate different genes at varied times (Emmert-Streib et al., 2012). Most of the existing computational methods test the changes in the expression levels of each single gene individually. However, diseases are usually consequences of interactions between multiple molecular processes, rather than an abnormality in a single gene (Menche et al., 2015).

Different associations between genes result in different cell conditions. Between two cell conditions, there are common gene interactions and also different interactions from the other. This fact can be used to identify molecular mechanisms that drive disease progression. Identifying these mechanisms let us to find new and more targeted biomarkers or drugs.

In order to find disease-specific interactions, many differential gene network inference algorithms were introduced. However, since we work in genome-wide, using these algorithms alone results many disease-specific interactions from which it is not easy to determine the main causes of the disease. In order to rank those disease-specific interactions and also the genes in the network, we developed a new approach by integrating some other available datasets for breast cancer. Our integration framework has resulted the most important genes and interactions by allowing ranking the breast cancer specific gene network. The proposed method can quantitatively identify the differences between two biological conditions (or two classes) and can provide better insights and understanding of breast cancer.

This dissertation is composed of the following five major parts. Part one contains the general introduction of differential networking methodology, the purpose of the study and the structure of the dissertation. Part two contains biological background and the review of the literature starting from gene network inference to differential network analysis. In this chapter main limitations of the existing approaches are also discussed. Part three contains the proposed differential network inference algorithm (*IDN*) including the overview and workflow of the algorithm and a case study in web based application.

In part four, firstly two preliminary analysis are presented and discussed. Following this, data collection, data pre-processing and data integration methods for breast cancer analysis are explained. Lastly, breast cancer analysis results of *IDN* approach are presented .

Finally, in part five, the discussion of the results, the limitations of the study, study conclusions and future recommendations are provided.

2. LITERATURE REVIEW

In this part of the study, firstly biological background is given. Secondly, gene network inference which is the first step in differential networking analysis was highlighted and some popular methods are presented. Finally, review of the literature starting from single gene analysis to differential network analysis are presented and discussed to understand importance of differential network analysis.

2.1. BIOLOGICAL BACKGROUND

Biological background part includes the term cancer, importance of breast cancer in women, and information about microarray data, gene regulatory networks and gene network inference.

2.1.1. Cancer

The term cancer refers to a large group of diseases that can affect any part of the body. Cancer is caused by the rapid production of abnormal cells which grow beyond normal boundaries and then invade adjacent tissues and spread to other parts of the body (World Health Organization (WHO) Fact sheet No 297).

Cancer has fast become a primary cause of mortality and morbidity in the world, especially in developed countries (Bray et al., 2012; Bař et al., 2015), and the second most frequent cause of death after heart diseases in emerging countries (Bař et al., 2015). In the WHO 2015 February fact sheet, there were about 14 million new cancer cases and 8,2 million cancer-associated deaths reported in 2012 while the number of new incidents was expected to increase by about 70% over the following two decades (WHO Fact sheet No 297). Cancer incidence rates have risen in Turkey because of environmental and individual risk factors, advancements in the registry system and improvements in the accessibility of healthcare services (Yilmaz et al., 2010).

The hallmarks of cancer acquired by different cancer types and enabling characteristics are specified as (Emmert-Streib et al., 2012):

- a. tissue invasion and sustained angiogenesis,
- b. insensitivity to anti-growth signals,
- c. self-sufficiency in growth signals,
- d. unlimited replicative potential,
- e. deregulating cellular energetics,
- f. prevention of immune destruction,
- g. avoidance of apoptosis,
- h. tumor promoting inflammation.
- i. genome mutation and instability.

2.1.2. Breast Cancer

Breast cancer is the most prevalent cancer among women worldwide, followed by colorectal, lung, cervix, and stomach cancer. The cause of 521000 of the 8.2 million cancer-associated deaths worldwide was breast cancer in 2012 (WHO Fact sheet N°297).

American Cancer Society (ACS) has estimated that nearly 246660 new invasive breast cancer cases and 61000 new cases of carcinoma in situ will be diagnosed in the USA in 2016. Moreover, ACS has reported that the number of women that will die from breast cancer will be approximately 40450. According to the other estimates 10-12% of women will have breast cancer in their life (Kleibl and Kristensen, 2016).

In Turkey, breast cancer is also the most frequent type of cancer among women according to statistics. It is followed by skin cancer, thyroid cancer, lung cancer, and stomach cancer (Yilmaz et al., 2010).

Life expectancies in high-resource nations is 80 years and over for women in Europe, North America, and Australia. Also life expectancies have increased at dramatic rates in China and India, which constitute nearly 40% of the world's population. The population of world continues to increase in size so age will be the most crucial demographic variable that affects future healthcare burdens. It will cause an increase in diseases

associated with age according to the gynecologists: problems of hormone deficiency, pelvic floor problems, and genital cancers will increase. Breast cancer will certainly increase affecting every ethnic groups in many countries (Becker, 2015).

Breast cancer isn't perceived as one disease with different histological characteristics and clinical behavior. It's comprehended as heterogeneous group of different diseases characterized by clear molecular aberrations. Research have shown that two types of breast cancers, oestrogen-receptor (ER)-positive and ER- negative, are distinct diseases at the transcriptomic level. Moreover, there might be additional molecular subtypes of these groups (Reis-Pilho and Pusztai, 2011).

There are many factors affecting development of cancer that are categorized as genetic, lifestyle and environmental. Genetic factors are very important for the breast cancer since it's known that a first-degree female relative with breast cancer doubles the risk for a proband. The more relatives with the breast cancer means the more risk. The genetic factors can be specified as: family history, personal history of breast cancer, breast condition, menoactivity, and height (Kleibl and Kristensen, 2016).

BRCA1 and BRCA2 are the top two breast cancer-associated genes, discovered in the last decade of the 20th century. There are many other genes associated with breast cancer but they didn't show a mutation frequency and clinical importance as BRCA1 and BRCA2. In general, breast cancer-associated genes are divided into three subtypes; high-penetrance, moderate- penetrance, and low-penetrance genes, according to the risk of breast cancer progression. BRCA1, BRCA2, PTEN, p53, STK11, and CDH1 are the widely known high-penetrance genes which increase breast cancer risk above four times (Kleibl and Kristensen, 2016).

2.1.3. Microarray Data

Deoxyribonucleic acid (DNA) is the basis of life for all living organisms on Earth consisting of nucleotide sequences (Gavlik and Szymczak, 2003). The DNA contains information needed by the cell to produce RNA and proteins in its genes. The information also controls when and how the genes are expressed. Then the genes

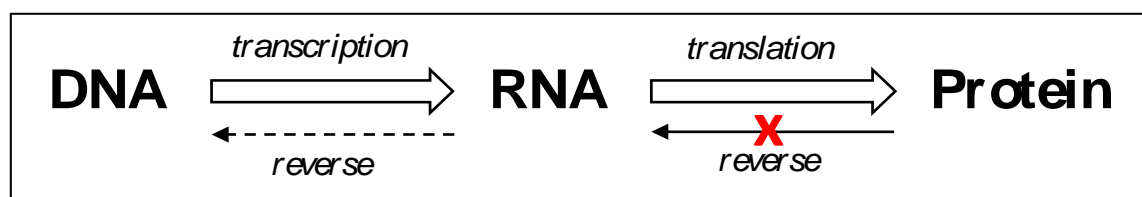
function in complex self-regulating networks within the organism giving rise to its shape and basic behavior (Burton, 2014; Reddy, 2009).

A gene can be defined as “a DNA sequence corresponding to a single norm of reaction of gene products across varying cellular conditions” (Griffiths and Neumann-Held, 1999). A copy of the DNA sequence is produced for transferring from mother cell to daughter cell, and the stored sequence information is rewritten into messenger RNA (mRNA). Afterwards the protein synthesis machinery can comprehend the genetic code (Singh et al., 2014).

The central dogma, identifies the encoding process of proteins by the genes. It describes the information flow in a biological organism including replication, transcription, and translation. Information refers to the certain detection of sequence, any of nucleic acid or of amino acid sediments’ bases in the protein (Koonin, 2012; Morange, 2008).

Francis Crick put forward the central dogma in return for the discovery of reverse transcription after integrality of information transfer from RNA to DNA in retro-transcribing genetic elements life cycle turned out to be clear (Figure 2.1). In the transcription step, DNA is copied into mRNA in the translation step, proteins are synthesized using the information carried by mRNA.

Figure 2.1. The central dogma



Source: “Koonin, E. V. 2012. Does the central dogma still stand? *Biology Direct*. 7:27”.

The central dogma situates the main exclusion principle at the translation phase. The central dogma supposes that “There is no information transfer from protein to nucleic acid” (Koonin, 2012). The possible ways of information transfer are from nucleic acid to protein, or from nucleic acid to nucleic acid. However, transfer from protein to nucleic, or transfer from protein to protein is not possible (Morange, 2008).

DNA microarray technologies were developed for the purpose of measuring the transcription levels of RNA transcripts obtained from genes within a genome in a single test. The aim is to evaluate the expression levels of many genes in the same test. Thousands of single-stranded sequences are synthesized to a glass which has a similar size microscope slide.

DNA arrays is classified in two groups; one provided by Affymetrix (<http://www.affymetrix.com/>), and the other one provided by Illumina (<http://www.illumina.com/>). The first one uses small single-stranded oligonucleotides that are synthesized in situ. The second type of arrays are used for copy-number measurements, genotyping, sequencing and detecting loss of heterozygosity (Trevino et al., 2007).

The process of microarray experiments start with labeling the extracted mRNAs and amplified cDNAs using fluorescent dyes. After then, the DNA array is incubated and washed to remove non-specific hybrids for hybridization. The attached fluorescent dyes are actuated by a laser to generate light detected by a scanner and a digital image from the microarray was produced by a scanner. The image of each spot is transformed to a numerical reading.

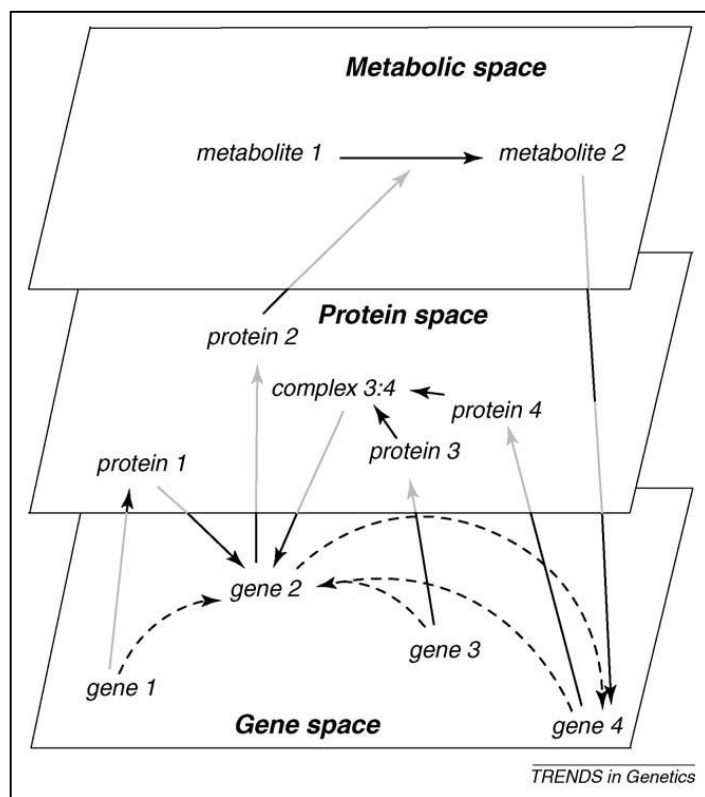
The process is completed in three steps. In the first step, the particular place and shape of each spot is found. Secondly the intensities in the identified spot are associated. Finally, the background noise is evaluated. Integrated signal is usually purged from the background noise. The final value is an integer that is proportional to the target sequence's concentration in the sample (Trevino et al., 2007).

2.1.4. Gene Regulatory Networks

A gene regulatory network (GRN) is an abstract representation of indirect gene-gene interactions that describes the ways how one gene indirectly affects all other genes which are connected to it. GRNs doesn't include the physical interactions of genes rather than a protein-protein interaction network (PPI). Gene-gene interactions are highly dependent on transcription factors (TF) (Yang, 2013)

A schematic view of a GRN is shown in Fig. 2.2. In the figure, layers represent gene activities, proteins and metabolities. In the gene activity space, there are four genes. Three of these genes (first, third and fourth gene) encode TFs. The other one (second gene) encodes a protein which starts the production process of metabolite two. The edges between nodes show biochemical processes such as metabolic conversion, transcriptional regulation and protein association. The interactions and causal effects in the gene space layer is an example of gene network concept (Yang, 2013, de la Fuente, 2010).

Figure 2.2. A schematic view of a gene regulatory networks



Source: “de la Fuente, A. 2010. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends in Genetics. 26 (7), pp. 326-33”.

2.1.5. Gene Characteristics

In this part, we presented important gene characteristics such as transcription factors, metastatic genes, prognostic genes and oncogenes which are important for the approach we presented in this dissertation.

2.1.5.1. Transcription Factors

Transcription factor (TF) is a molecule that determines if a gene's DNA is transcribed into RNA and controls the gene's activity. The gene's DNA is used by the enzyme RNA polymerase as a template. RNA is synthesized by the chemical reactions that are catalyzed by the enzyme RNA polymerase. TFs control the time, place, and the efficiency of the function of RNA polymerases. TFs play crucial role in the usual development of an organism. They also control the response to disease. TFs are a various kind of proteins which function usually in multi subunit protein complexes (Britannica.com).

Different transacting factors may interact with a common binding site. Recent studies revealed that the interaction of different factors with a common target site does not always conclude with equivalent transcriptional responses. Some factors activate transcription whereas some prevent the activation (Karin, 1990).

Transcription factors need the skill to bind to DNA and affect transcription positively or negatively, to produce their effects. Studies showed that the structure of TFs is modular and specific areas of the molecule are responsible for binding to the DNA. In Table 2.1, the categorization of TFs according to their DNA binding domains is shown (Latchman, 1997).

Table 2.1. Classification of transcription factor families

Domain	Factors containing domain
Homeobox	Numerous Drosophila homoetic genes, related genes in other organisms
POU	Oct-1, Oct-2, Pit-1, Unc-86
Paired box	Various Drosphilla segmentation genes
Cysteine-histidine zinc finger	TFIIIA, Kruppel, SP1, etc.
Cysteine-cysteine zinc finger	Steroid-thyroid hormone receptor family
Basic element	C/EBP, c-fos, c-jun, GCN4
Ets domain	Ets-1, Elk-1, SAP

Source: "Latchman, D.S. 1997. Transcription Factors: An Overview. *International Journal of Biochemistry and Cell Biology*. 29 (12), pp. 1305-1312".

2.1.5.2. Metastatic genes

Metastasis is a painful pathological and evolutionary process that causes morbidity and mortality mainly. There have been important developments in detecting and treating localized cancers in recent years. Even so, metastatic disease is the most common reason of cancer associated deaths. Early diagnosis of cancers is of capital importance for treatment when they are localized and curable (Yoshida et al., 2000). Expression of particular genetic programs is required by a tumor cell for metastasis, to enable the appropriate interactions with changing microenvironments for promoting continued survival and proliferation at secondary sites. To understand these programs and their effects on cellular interactions and signaling cascades is essential in discovering the complex metastasis process (Hurst and Welch, 2011).

Subsets of tumor cells are equipped with capabilities that are not found in their nonmetastatic counterparts. Metastatic cells turn on genes which promote metastasis. It's not easy to identify prometastatic genes because cell to accomplish numerous tasks in multiple different microenvironments is required for the ability to metastasize (Hurst and Welch, 2011). Metastasis suppressor genes are potential markers to distinguish nonmetastatic and metastatic cancer cells (Liu et al., 2001). "Metastatic genes are those in which gains in oncogene functional activity or lack of tumor suppressor genes enables cancer cells to detach, escape into the circulation, penetrate and colonize distant organs." (Alberti, 2008). Three types of metastatic genes can be identified according to their involvement level in the metastatic pathway (Alberti, 2008, Nguyen and Massague, 2007):

- a. Metastasis initiation genes: They permit the malignant cell to enter into the circulation in primary tumor, by causing tumor cell detachment, invasion and motility, and promoting the neoplastic angiogenesis. Most genes that are relevant to tumor cell motility, angiogenesis or invasion are belong to this group.
- b. Metastasis progression genes: These type of metastatic genes mediates the metastasis initiation mentioned above and play a role in metastatic colonization. Moreover, they contribute to primary tumor growth. The difference of these

genes than oncogenes are that they carry out the same cell-autonomous transforming function throughout the course of a malignant disease.

- c. Metastasis virulence genes: They are exclusively responsible for the colonization of target organs. These genes play role in metastatic colonization rather than primary tumor development.

2.1.5.3. Prognostic genes

Prognostic genes are significant in cancer research since they give information concerning clinic outcomes. Prognostic genes are important in cancer treatment and prognosis for prediction of patients' survival. Moreover, these genes give idea about molecular mechanisms of tumor progression. In a study conducted by Yang et al. (2014), the properties of prognostic genes in the gene co-expression networks of 4 different types of cancer were investigated. Their results indicate that: prognostic mRNA genes are not hub genes; prognostic mRNA genes are enriched in modules; targets of prognostic microRNAs present similar patterns; some prognostic modules are conserved across tumor types.

2.1.5.4. Oncogenes

Oncogenes are usually expressed at high levels in tumor cells and have great potential to cause cancer (Wilbur et al., 2009). Proto-oncogenes, which normally help cells to grow, become a harmful when they mutate or too many copies of them come into existence. These genes might be kept in a state of constant activity when they are not supposed to. The cell growth is uncontrolled and this may cause cancer. This bad gene is called an oncogene (Cancer.org). The first oncogene determined is SRC which was discovered in chickens in 1970. It is called the "Rous sarcoma virus". The SRC gene was then shown to be play a role in many cancer types such as colon, liver, lung, breast and pancreatic cancer in humans. There are a lot of known oncogenes in humans. They have an essential role in initiating and maintaining tumor growth and they are known as prime targets in drug development. (broadinstitute.org). Proteins which control apoptosis and cell proliferation are encoded by oncogenes. Oncogenes can be activated by structural changes caused by mutation or gene fusion (Croce, 2008).

2.2. GENE NETWORK INFERENCE

In this part, the information about gene network inference is given and some of the popular methods such as RN, ARACNE, CLR, MRNET and C3NET that are used commonly are presented.

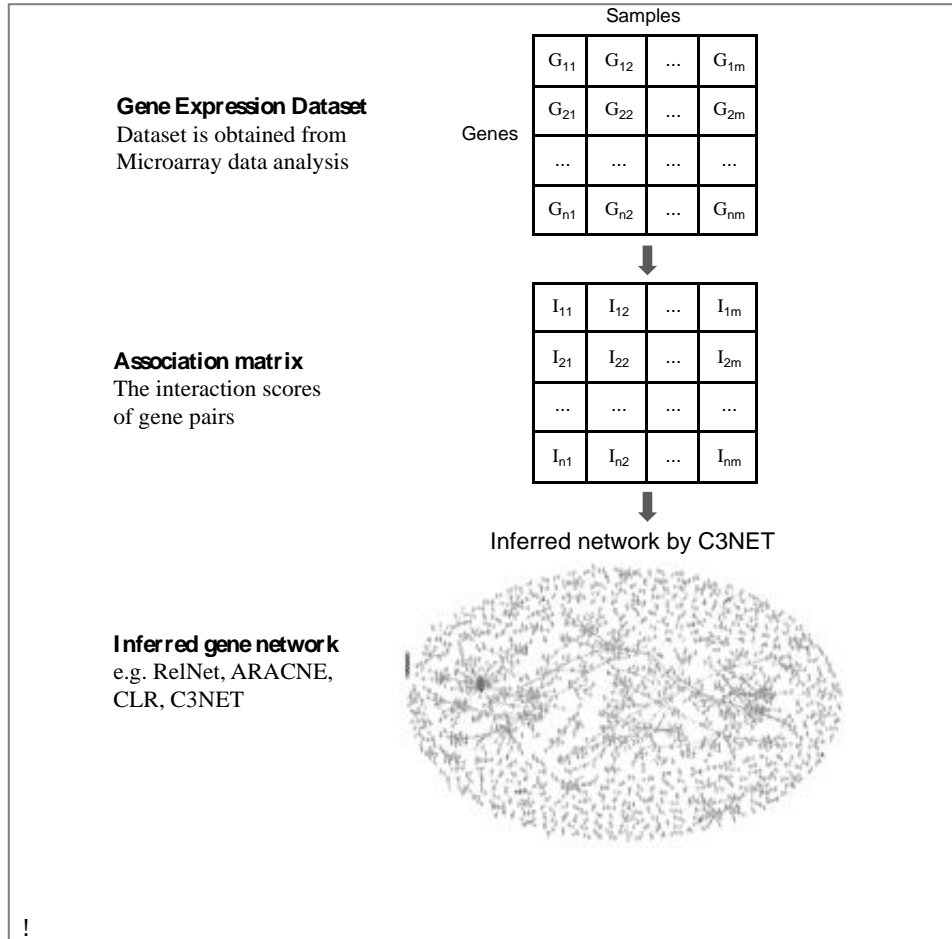
2.2.1. Introduction

The inference process of gene regulatory networks (GRN), which can be seen as a reverse engineering problem, is a process of estimating direct physical associations among genes from gene expression data (Emmert-Streib et al., 2012). This process can provide valuable information about normal cell physiology, development and pathogenesis and promote better understanding of biological and biomedical problems (Emmert-Streib & Dehmer, 2010; Rual et al., 2005; Schadt, 2009). In bioinformatics studies, gene network inference (GNI) algorithms are widely used for determining the roles of regulating and regulated genes, identifying the important genes in genetic diseases, and identifying biomarkers (Kurt et al., 2014). GRNs aim to capture the interactions between molecular structures and are represented as graphs in which nodes correspond to genes, and edges represent the dependencies between genes (Hecker et al., 2009).

In vivo or in vitro, molecular interactions can be detected accurately by classical molecular biology approaches. Unfortunately, these methods are not easy to apply and the number of associations that can be examined by these approaches is limited (Klipp et al., 2005). Gene networks such as transcriptional regulatory networks, protein networks or metabolic networks, represent patterns of dynamical operations within cells. Different kind of gene networks have different effects on the dynamical processes of cellular systems (Emmert-Streib and Dehmer, 2009a,b). Hence, gene network inference has been identified as a focal point in systems biology. However, GNI is a challenging problem because of the current very large-scale biological datasets and the noise caused by experimental and computational processes.

The steps of the gene network inference process are shown in Fig. 2.3. The dataset obtained from microarray data analysis consists of gene expression levels. Firstly, by using these pre-processed expression values, a gene expression matrix is created. In this matrix, each row represents a gene whereas each column represents a sample. The second step is estimating interaction scores of gene pairs. In this step, association score estimators such as correlation-based, entropy-based and direct mutual information (MI) estimators are used to obtain interaction scores. A dataset discretization operation is required in order to use MI estimators. At the end of the second step, a square gene association matrix is obtained. Finally, GNI algorithms are applied to this association matrix and the inference of gene regulatory network process is completed.

Figure 2.3. The workflow of gene network inference



Source: “Kurt, Z. (2013). Gen ağı çıkarım algoritmaları için en uygun ilişki kestirimcilerinin belirlenmesi. Thesis for the Ph.D. Degree. İstanbul: Yıldız Teknik Üniversitesi FBE”.

The most crucial part of GNI algorithms is to compute the dependency scores among cell structures. The interaction scores among gene pairs are determined from the gene expression datasets by the association score estimators. However, there is no commonly approved estimator that is known to provide the best performance for GNI methods. In a study conducted by (Kurt et al., 2014), Twenty seven association estimators were examined and 14 most likely estimators were evaluated. According to the study results; BS with spline order 2 (BS2), BS with spline order 3 (BS3), Kernel Density Estimator (KDE), Spearman-based Gaussian (SPG) and Pearson-based Gaussian (PBG) were found to be the best association score estimators regarding the performance and runtime. Therefore, in this thesis, Pearson-based Gaussian estimator was preferred (Kurt et al., 2014a,b).

2.2.2. Methods

The best of the methods that have been developed for inferring GRNs from microarray gene expression data are based on information theory (Gallager, 1968; Shannon and Weaver, 1949). The main principle of information-based methods is estimating mutual information (MI) values among gene pairs (Butte et al., 2000; Meyer et al., 2007). MI based algorithms are able to detect linear as well as non-linear effects between gene pairs (Li, 1990; Steuer et al., 2002). Furthermore, they enable us to work with large sample sizes such as 25000 genes (Altay et al., 2011).

2.2.2.1. Relevance Network

RN (relevance network) was one of the first algorithm that is developed for inferring GRNs from gene expression data (Butte and Kohane, 2000). This algorithm computes all mutual information scores for all pairs of genes and eliminates the edges between genes that have MI values which are not statistically significant. The approach of relevance networks (Butte et al., 2000) consists in inferring a genetic network by computing all MI scores for all gene pairs, and linking a pair of genes (xy) by an edge if MI value I_{xy} between these genes is larger than the threshold value I_t . In the resulting network, two genes connect to each other only if $I_{xy} > I_t$. If this equation is not valid, no edge is set between x and y if this. The threshold value I_t was computed by randomizing the gene expression dataset.

RN computes all pairwise interactions. Hence, the complexity of the algorithm is $O(n^2)$. Additionally, RN can set an edge between two genes which do not interact directly, but both are regulated by a third gene. For example, suppose that gene x and y are regulated by gene z . This will result in high mutual information between gene pairs (xy) , (xz) and (yz) . Therefore, RN will put an edge between x and y although these genes do not interact directly but interact through gene z (Meyer et al., 2008).

2.2.2.2. ARACNE

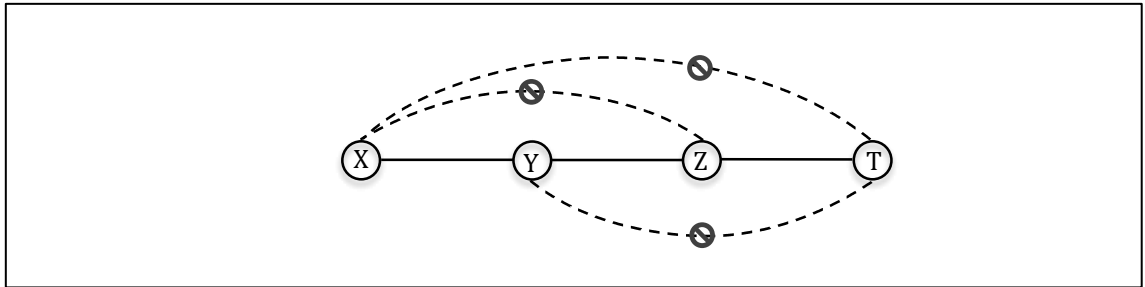
The second well-known method is the (ARACNE) “algorithm for the reconstruction of accurate cellular networks” (Margolin et al., 2006). ARACNE uses data processing inequality and, in addition to RN, it performs a second step to eliminate the least significant edge of a triplet of genes. This results in a more conservative estimation of the inferred network.

The algorithm starts with estimating the pair-wise MI values for all genes. Then it eliminates non-significant values according to the obtained threshold I_0 . This step is basically equivalent to relevance networks since it computes mutual information and declares MI values significant if $I_{xy} > I_t$. If I_{xy} is found to be significant, then an edge is included in the corresponding adjacency matrix between gene x and y , $A_{xy} = A_{yx} = 1$. In addition to the first step, ARACNE performs a second step called *data processing inequality* (DPI). The DPI is a relation between MI values that means that generally its information content can not be increased by a post-processing (24). DPI serves as a filtering step. DPI indicates that, if the interaction between gene X and gene Z occurs through another gene Y ($X \rightarrow Y \rightarrow Z$), then

$$I(X, Z) \leq \text{argmin} \{I(Y, Z), I(X, Y), \}.$$

Here, the weakest edge of the gene triplet $I(X, Z)$, corresponds to the indirect interaction and hence is eliminated by the DPI approach. The working mechanism of DPI is shown in Fig. 2.4.

Figure 2.4. Working mechanism of DPI. DPI infer the most likely path of information flow regardless of significant mutual information values of all six gene pairs. For example, $X \leftrightarrow Z$ will be eliminated because $I(Y,Z) > I(X,Z)$ and $I(X,Y) > I(X,Z)$. $Y \leftrightarrow T$ will be eliminated because $I(Y,Z) > I(Y,T)$ and $I(Z,T) > I(Y,T)$. $X \leftrightarrow T$ will be eliminated in two ways: (1) because $I(X,Y) > I(X,T)$ and $I(Y,T) > I(X,T)$, and (2) because $I(X,Z) > I(X,T)$ and $I(Z,T) > I(X,T)$.



Source: “Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*. 7 (7) Suppl 1, pp. 7.”

In this step, ARACNE tests all possible triplet gene combinations (three genes with MI values larger than I_t) and then, for each (xyz) , it eliminates the edge with the lowest MI value $I_l = I_{x'y'}$, with $(x'y') = \operatorname{argmin}\{I_{xy}, I_{yz}, I_{xz}\}$ from the adjacency matrix, if it is smaller than the second smallest MI value I_2 multiplied by a factor (Margolin et al., 2006).

$$A_{ij'} = A_{j'v} = \begin{cases} 0 & I_1 \leq I_2(1 - \epsilon) \\ 1 & \text{otherwise.} \end{cases}$$

Here, $0 \leq \epsilon \leq 1$. ϵ is the tolerance parameter. Trial studies were conducted to obtain optimal values for ϵ . For this reason, it can be said that I_0 is identified by unsupervised and ϵ is identified by supervised manner of learning. In ARACNE, each gene triplet is analyzed independently from the other triplets. Hence, it is possible that an edge can be included in the resulting network although it has been marked for removal by prior DPI applications to different triplets. Consequently, the order of examination of gene triplets does not affect the resulting network. ARACNE has a complexity in $O(n^3)$ since the method checks all triplet gene combinations (Margolin et al., 2006).

2.2.2.3. CLR

CLR (Context Likelihood of Relatedness) is another method that employs a background sensitive estimator between the gene pairs by converting MI estimates to values similar to z-scores (Faith et al., 2007).

The CLR algorithm is also an extended version of the RN approach which starts by computing the pair-wise MI values for all genes. Then it estimates the statistical likelihood of each mutual information value I_{xy} by comparing this MI value to a “background” distribution of the MI values. In particular, two z-scores are obtained for each gene pair (xy) by comparing the MI value I_{xy} with gene specific distributions, p_i and p_j . Here, p_i and p_j distributions are equivalent to the distributions of MI values related to genes x and y , respectively (Faith et al., 2007). CLR takes into account the score

$$\overline{z_{xy}} = \sqrt{z_x^2 + z_y^2}$$

by making a normality assumption about these distributions. Here, z_x and z_y are the z-scores of I_{xy} , whereas $\overline{z_{xy}}$ corresponds to the joint likelihood measure. CLR estimates *individual* thresholds by considering an *individual* background for each pair of genes differently from RN and ARACNE which uses an overall threshold I_t for each MI score between gene pairs.

CLR has $O(n^2)$ complexity since mutual information matrix is computed once for each gene pair (Faith et al., 2007).

2.2.2.4. MRNET

In addition to these methods, MRNET (maximum relevance and minimum redundancy network) (Meyer et al., 2007) uses the maximum relevance and minimum redundancy feature selection method to infer a network. The algorithm uses a scoring function to select potential association partners of a target gene Y .

The algorithm starts with ranking the set of input variables V according to a score that is the difference between the MI with the output variable Y and the average MI value with the previously ranked variables. The basic idea is ranking the direct interactions higher than indirect interactions (Meyer et al., 2007). The working mechanism is shown below.

$$X_j^S = \operatorname{argmax}_{X_j \in V \setminus S}(s_j)$$

$$s_j = I(X_j; Y) - \frac{1}{|S|} \sum_{X_k \in S} I(X_j; X_k).$$

Here, the score s_j is the difference between the MI of X_j with the target variable Y (relevance term) and the average redundancy of X_j to each already selected variable $X_k \in S$ (redundancy term). A gene is added to the set S only if the s_j is above the threshold value, s_0 and the score of gene X_j maximizes equation X_j^S . The algorithm repeats the iteration procedure until no further gene can be found that passes the threshold test. The MRNET approach consists in finding interaction partners for Y that are of maximal relevance for Y , but have a minimum redundancy for the already found interaction partners in the set S . The algorithm starts with a fully connected, undirected network among all genes and then it eliminates the edges between Y and $V \setminus S$, which have not maximized the equation X_j^S (Meyer et al., 2007). MRNET has a complexity in $O(f \times n^2)$ since it repeats the feature selection for each of the n genes (Meyer et al., 2007).

All of these methods aim to infer the entire regulatory network for a given data set. However, achieving this goal is not easy for a large sample size. Observational data may not be able to detect all dynamical associations that would allow a reliable estimation. Hence, *c3net* GNI algorithm was used in this thesis since it aims to infer only the strongest interactions among covariates (Altay and Emmert-Streib, 2010a).

2.2.2.5. C3NET

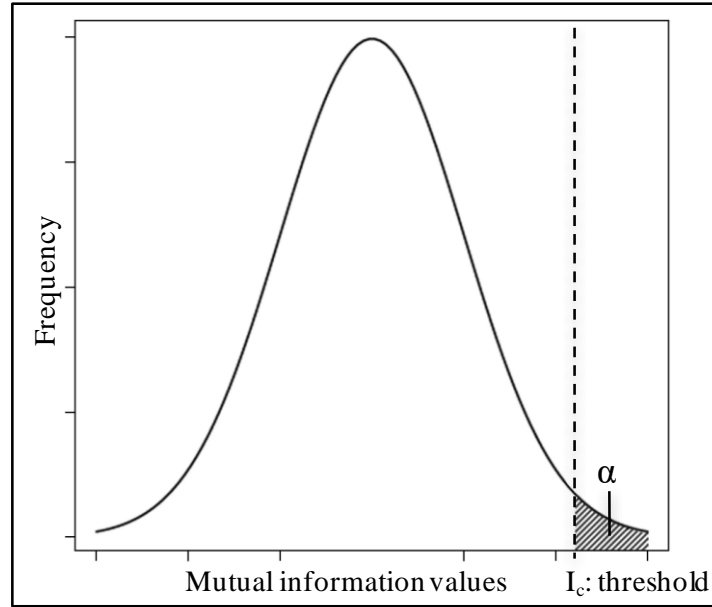
The inference algorithm C3NET have two steps. In the first step, it eliminates non-significant connections among gene pairs. Then it selects for each gene the edge with maximum mutual information (MI) value (Altay & Emmert-Streib, 2010a). The first step is similar to previous approaches, e.g., RN, ARACNE or CLR. In this step, C3NET tests the statistical significance of pair-wise mutual information values using resampling methods and eliminates non-significant edges according to a determined significance level α . Mathematical formulation of the mutual information (Cover and Thomas, 1991) of two random variables A and B is defined as

$$I(A, B) = \sum_{a \in A} \sum_{b \in B} p(a, b) \log \frac{p(a, b)}{p(a)p(b)}.$$

In order to calculate statistical threshold, C3NET uses resampling methods that estimate the distribution under the null hypothesis corresponding to a vanishing mutual information. For this purpose, it randomizes the expression data set by permuting the gene expression measurements n times and recalculating the distribution of the new pair-wise mutual information for each permutation. Then C3NET creates a vector combining these permuted mutual information matrices and determines the threshold value according to a chosen significance level α . Visualization of this vector is shown in Fig. 2.5. The vertical (Y) axis represents the frequency of mutual information values, whereas the horizontal (X) axis represents mutual information values. The threshold, denoted by I_c , is determined as the maximum mutual information value for the significant region of the null distribution, as illustrated in Fig. 2.5 by the dashed line.

Fig. 2.6 shows the principle steps of the C3NET algorithm. Primarily, C3NET creates a mutual information matrix (MIM) by estimating the mutual information values from the data. In this process, it use an estimator that allows a close approximation of the theoretical value of the population. Starting from zero matrices C , with $C_{xy} = 0$ for all $x, y \in V$ and B , with $B_{xy} = 0$ for all $x, y \in V$, C3NET thoroughly tests all pair-wise MI values I_{xy} , $x, y \in V$, and sets $C_{xy} = C_{yx} = 1$ if the null hypothesis $H_0 : I_{xy} = 0$ can be rejected, for a determined significance level α (Altay and Emmert-Streib, 2010a).

Figure 2.5. Determining MI threshold, I_C



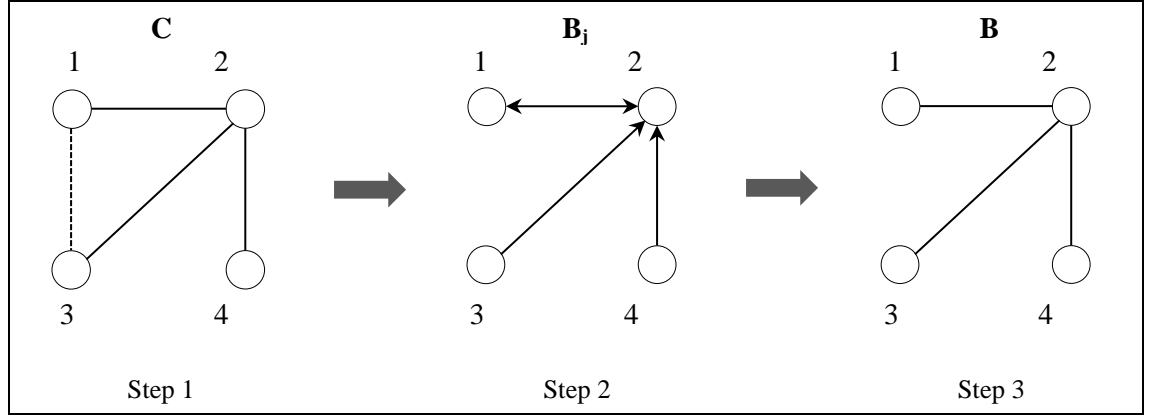
Source: This dissertation.

Secondly, the most significant connection for each gene is selected. The algorithm first determines the neighborhood N_s for all genes $x \in V$. The neighborhood of gene x is defined by $N_s(x) = \{y : C_{yx} = 1 \text{ and } y \neq x\}$. To that end, it uses the connectivity matrix C . The link corresponding to the highest mutual information value in the neighborhood for each gene is determined by using N_s and I . This link is identified by

$$j_c(x) = \operatorname{argmax}_{y \in N_s(x)} \{I_{xy}\}.$$

It is possible that all mutual information values I_{xy} for $y \in V$ are non-significant ($N_s(x) \neq \emptyset$). In this case, no index is assigned to $y_c(x)$. The algorithm constructs the adjacency matrix B of the estimated undirected network by setting $B_{xy_c(x)} = B_{y_c(x)x} = 1$ if $y_c(x)$ has been set to a valid index. The rest of the entries of B remain zero (Altay and Emmert-Streib, 2010a).

Figure 2.6. Visualization of the steps of C3NET. The edges shown in solid and dashed lines represent significant edges. In the third step, the edges in solid lines represent the edges with maximum I value.



Source: This dissertation.

Suppose that we have the MI values given by I . The MI values which are statistically significant appear as ‘1’ entries, whereas the remaining ones appear as ‘0’ entries in the corresponding connectivity matrix C .

$$I = \begin{pmatrix} 1.0 & 0.9 & 0.7 & 0.1 \\ 0.9 & 1.0 & 0.8 & 0.7 \\ 0.7 & 0.8 & 1.0 & 0.4 \\ 0.1 & 0.7 & 0.4 & 1.0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Then the algorithm determines statistically significant connections with neighboring genes with maximum MI. This is the critical step in C3NET, resulting in $y_c = (2, 1, 2, 2)$. The next step is determining auxiliary matrix B_y , directly from y_c . B_y contains exactly the edges added by each node. Due to its symmetry in its arguments, MI does not provide directional information, so the resulting adjacency matrix, B , is symmetric.

$$B_y = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

The output network represented by adjacency matrix B is a star-like network where second gene is connected to three other genes. It is obtained by the conversion of the asymmetric matrix B_y to a symmetric matrix B as shown in the example of Fig. 2.6. It is important to realize that each gene can create at most one connection. However, a same

gene $y_c(x)$ can be selected by different genes x . For this reason, the final undirected network can consist of genes having more than one connection to other genes.

2.3. DIFFERENTIAL ANALYSIS

In this part, methods for differential analysis starting from single gene analysis are presented and discussed.

2.3.1. Single Gene Analysis

Single gene based methods can be divided into three categories: (1) differential expression methods, (2) differential variability methods, and (3) differential correlation methods (Emmert-Streib et al., 2012).

2.3.1.1. Differential Expression

Differential expression (DE) analysis, which is one of the most commonly used microarray data analysis method for disease studies, selects differentially expressed genes by comparing gene expression levels between two conditions e.g., disease and healthy cells (Zheng et al., 2014). This is commonly done by testing the statistical significance of the changes in the mean expression level of each individual gene. If the mean level of expression of a given gene is significantly different (lower or higher) between treatment and control conditions, then this gene is called differentially expressed (de la Fuente, 2010).

Suppose that the mean level of expression of gene g_i in a microarray dataset for two different conditions are μ_1 and μ_2 respectively. The null and alternative hypothesis are as following:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

If the null hypothesis H_0 is rejected, then gene g_i is called *differentially expressed* (Emmert-Streib et al., 2012). This means that the mean level of expression of gene g_i is significantly different between two different conditions.

Many statistical algorithms have been developed to identify differentially expressed genes. Among these methods, most popular ones are the empirical Bayes approach (Efron et al., 2001), SAM (Chu et al., 2002) and limma (Smyth, 2005).

Diseases are usually consequences of interactions between multiple molecular processes, rather than an abnormality in a single gene (Menche et al., 2015). Genes and their products (proteins) perform their functions in coordination. However, differential expression analysis approach treats each gene individually and doesn't consider the fact that biological operations require collective work of many genes. (Yu et al., 2011; Bockmayr et al., 2013). Furthermore, without any change in its expression level, the function of the gene can be affected by the mutations and post-translational modifications and accordingly, known disease genes may not be differentially expressed in diseases (de la Fuente, 2010).

2.3.1.2. Differential Variability

The differential variability (DV) analysis aims to identify a significant change in variance of the gene expression values between disease and control samples (Prieto et al., 2006; Ho et al., 2008).

Suppose that the mean level of expression value of gene g_i in a microarray dataset is μ_c and its variance σ_c^2 for condition $c = \{1,2\}$. The null and alternative hypothesis are as following:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

If the null hypothesis H_0 is rejected, then gene g_i is called *differentially variable* (Emmert-Streib et al., 2012). In DV analysis, commonly used statistical tests are F-test (Ho et al., 2008) and ANOVA (Analysis Of Variance) (Pritchard et al., 2001).

2.3.1.3. Differential Correlation

The differential correlation analysis aims to detect changes in the dependency structure of a single gene. In this type of analysis, whole correlation vector associated with each single gene is tested to select candidate genes.

Suppose that $r_i = r_{i1}, \dots, r_{in}$ is a $n - 1$ dimensional correlation vector, which represents the correlations between the i th gene and all other remaining $n - 1$ genes in a dataset. Then, $F_{r_i}^A$ and $F_{r_i}^B$ denote $n - 1$ dimensional joint distribution functions of r_i in two different conditions. The null and alternative hypothesis are as following:

$$H_0 : F_{r_i}^A = F_{r_i}^B$$

$$H_1 : F_{r_i}^A \neq F_{r_i}^B$$

If the null hypothesis H_0 is rejected, then gene g_i is called *differentially correlated* (Hu et al., 2009).

2.3.2. Gene-pair Analysis

Differential patterns methods can be subdivided into two categories: (1) differential co-expression and (2) differential co-clustering (Odibat, 2012).

2.3.2.1. Differential Co-expression

Differential co-expression (DC) analysis, as a more advanced approach to the DE analysis, aims to identify differences in the co-expression patterns in normal and disease conditions. This is done by measuring the mean pairwise *correlation difference* between sample groups (de la Fuente, 2010). The genes whose correlated expression pattern differs between groups are defined as DC genes. A pair of gene expression datasets for normal and disease conditions are transformed into a pair of co-expression matrix. In this matrix, edges correspond to transcriptionally correlated gene pairs. Following this, the DC score is calculated for each gene. Pearson correlation coefficient is one of the most commonly used method to measure the co-expression relationships (Hu et al., 2009; Zheng et al., 2014).

$$r_{ij} = \frac{\text{cov}(x_i, x_j)}{\sqrt{\text{var}(x_i) \text{var}(x_j)}}$$

In the formula, $\text{cov}()$ represents the covariance and $\text{var}()$ represents the variance of the gene expression levels. Then, the correlation between a pair of gene expression levels is computed over the normal sample, r_{ij}^N , and over the disease sample, r_{ij}^D . The null and alternative hypothesis are as following:

$$H_{01} : r_{ij}^N = 0$$

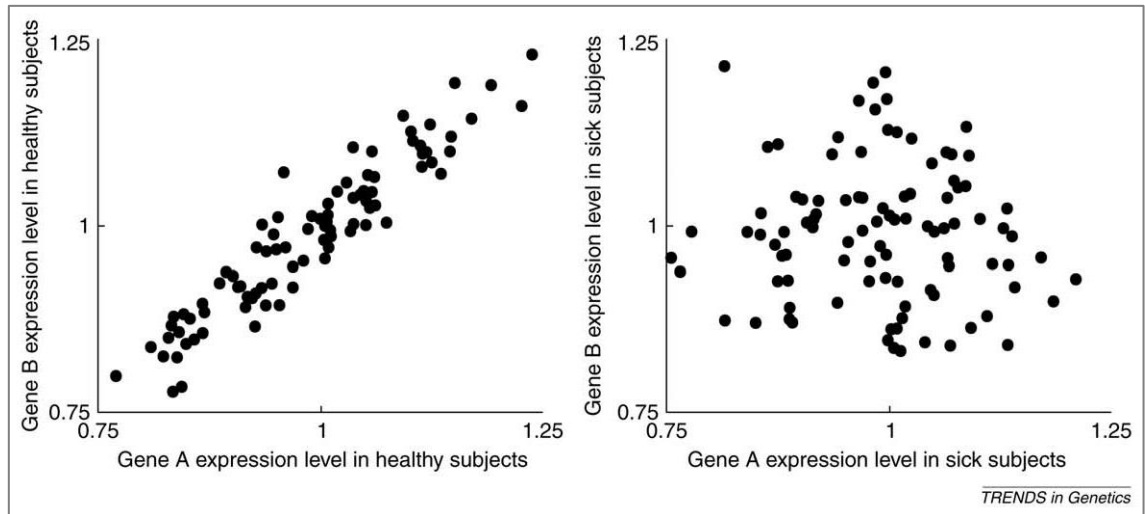
$$H_{02} : r_{ij}^D = 0$$

If none of the hypothesis is rejected, then the genes are not correlated in any sample. If both of the hypothesis are rejected but the sign of the correlations is same, then the correlation between genes are similar in both samples. If both of the hypotheses are rejected but correlations have changed signs, or only one of the hypotheses is rejected, then the pair of genes is identified to be differentially co-expressed (de la Fuente, 2010).

Correlation-based inference methods are used to construct gene co-expression networks. Co-expression networks have been widely used in the literature to uncover gene functions and investigate GRNs. Nevertheless, gene co-expression networks reveal only the gene regulatory interactions under specific conditions (Hsu et al., 2015).

Several popular methods for differential co-expression analysis are “Expected Conditional F-statistic” (ECF-statistic) (Lai et al., 2004), “Weighted Gene Coexpression Network Analysis” (WGCNA) (Zhang & Horvath, 2005), “Log Ratio of Connections” (Reverter et al., 2006), CoXpress (Watson, 2006), dCoxS (Cho et al., 2009), DCIM (Freudenberg et al., 2010), DiffCoEx (Tesson et al., 2010), “Differential Coexpression profile” (DCp) (Yu et al., 2011), “Differential Coexpression enrichment” (DCe) (Yu et al., 2011), DiffCorr (Fukushima, 2013) and “Differential Correlation in Expression for meta-module Recovery” (DICER) (Amar et al., 2013).

Figure 2.7. Instance of different gene expression correlations with the same mean expression levels.



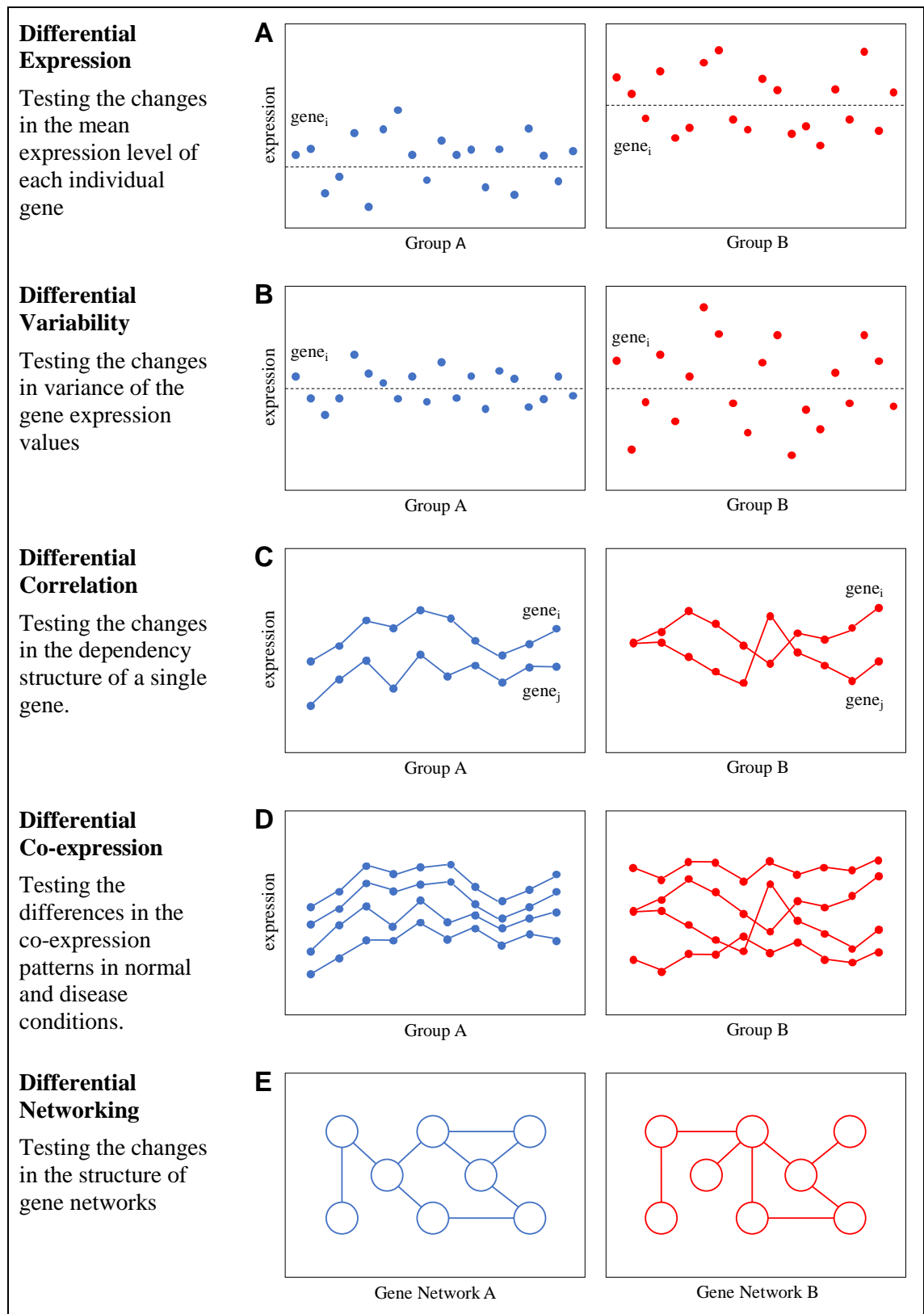
Source: “de la Fuente, A. 2010. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*. 26 (7), pp. 326-33”.

DC analysis have been successfully applied to identify condition-specific modules, a group of genes strongly correlated in one condition but not in the other. Furthermore, DC analysis may reveal the rewiring of transcriptional networks in response to disease (Hsu et al., 2015). DC methods, unlike from single gene based methods, consider the relationships between different genes.

2.3.2.2. Differential Co-clustering

A co-cluster is a set of co-expressed genes under a subset of samples. Differential Co-clustering approach is interested in a subset of rows that may be related under a subset of columns of a two dimensional data matrix which is called as co-clusters (Odibat and Reddy, 2011). The idea behind this methodology is that some genes may active only in a subset of the samples but not in the whole dataset (Wu, 2007). Differential Co-clustering methods are useful when the structure of biological samples are heterogenous and have multiple subtypes. The main difference between differential co-expression and differential co-clustering is that differential co-expression approach computes the correlation between any gene pair based on all the samples, while differential co-clustering approach use a subset of the samples (Odibat, 2012).

Figure 2.8. Visualization of differential analysis methods of gene expression data.



Source: This dissertation.

2.3.3. Differential Network Analysis

In the literature, there are two comprehensive reviews that discuss differential networking approach. The review article by de la Fuente (2010) describes the trend from DE to differential network analysis and outlines its effects on the disease related studies. In the review, DE studies and network inference approaches, which are both important in the analysis of gene expression data, were explained and the theoretical background were provided. Furthermore, some approaches for DC analysis were discussed and some additional future directions were suggested. The methodologies examined in the review mainly focus on identifying differential coexpression patterns, which is defined by the author as the first step towards identifying differential gene networks. The author mentions the importance of the comparison of the coexpression networks across multiple conditions and provide some approaches for establishing thresholds that are used to determine which edges will be included in the coexpression networks.

Some of the approaches explained in the review are the use of statistical thresholds, the global network topology and clustering coefficient. Another approach mentioned in the review is to compare the weighted coexpression networks. The author also highlights the drawback of the construction of healthy and disease coexpression networks, as they both require separate decisions and thresholds. Additionally, it is indicated that the examination of variance of gene expression distributions in normal and disease samples can be helpful. In the review, only two methods, CoExpress (Watson, 2006) and GSCA (Choi and Kendzioriski, 2009) were discussed to compare weighted coexpression networks. The author conclude that the differential networking methodology will have a crucial role in the identification of the dysfunctional regulatory networks underlying complex human diseases in the future (de la Fuente, 2010).

In another recent review, Ideker and Krogan (2012) also argue that differential networking will become a standard way of network analysis in the future. The authors review the technological progress and experimental approaches that lead to differential networking besides highlighting the biological insights derived from this type of analysis. In the review, authors emphasize that, until that day, almost all type of genetic networks have been examined under one condition despite the fact of the highly

dynamic behaviour of biological systems. They indicate the importance of understanding how these genetic networks effect, or are affected by, changes across multiple conditions, species or times. A historical timeline of differential approaches in biology was also provided in the review. Additionally, the authors discuss some quantitative methods for differential network analysis based on subtraction of interaction scores across conditions and highlight some of the statistical challenges of differential network analysis. The review mainly focus on differential network mapping and examine only one method called dE-MAP (Bandyopadhyay, 2010) that creates differential maps of genetic interactions. The authors concluded that differential network mapping will enable us to discover unexplored interactome by providing a deeper understanding of complex biological phenomena (Ideker and Krogan, 2012).

Van Landeghem et al. proposed a framework called Diffany that infers and analyzes an arbitrary set of input networks by implementing novel ontology-based algorithms. One reference network which represents the interactome of an untreated organism is always included in the input network set. The Diffany framework provides a default interaction ontology which includes genetic interactions such as regulatory, co-expression and protein-protein interactions. The algorithm was tested on a plant osmotic stress study. A regulator that is predicted by the method was experimentally confirmed (Van Landeghem et al., 2016).

Hsiao et al. presented an algorithm “modulated gene/gene set interaction analysis” (MAGIC) that identifies modulated interactions at the gene and gene set level. The MAGIC provides a statistical model to examine combinations of gene and gene set pairs. The algorithm first scores and filters gene/gene sets. In the second step it performs modulated analysis based on Fisher and inverse Fisher transformation. In the last step, it infers and visualize the modulated interaction networks. The algorithm was applied to a simulated dataset as well as to a breast cancer dataset. The authors reported that several hub genes and functional interactions were discovered (Hsiao et al., 2016).

Another algorithm that constructs differential dependency networks is “knowledge-fused differential dependency networks” (KDDN) method. KDDN is a mathematical framework that constructs differential dependency networks with significant rewiring. It integrates the measured data with the prior biological knowledge to construct both

differential and common networks. The method was applied to budding yeast *Saccharomyces cerevisiae* dataset to test the response to oxidative stress (Tian et al., 2015).

Ha et al. proposed a “pathway-based differential network analysis in genomics” (DINGO) method for predicting differential patterns between patient-specific groups. DINGO algorithm decomposes networks into global and group-specific components to estimate separate conditional dependencies between groups. The algorithm was applied to TCGA glioblastoma dataset and two networks, short-term and long-term survivors in The Cancer Genome Atlas, were extracted. Additionally, in the study, some hub genes that are related to c-Myc gene were found. The authors reported that c-Myc gene has an important role in the regulation of glioblastoma multiforme proliferation which result in shorter survival times in glioblastoma multiforme patients (Ha et al., 2015).

Differential epistasis mapping (dE-MAP) is a technique that was developed by Bandyopadhyay et al. (2010) to reveal the interaction differences between two static gene interaction networks under condition change. Seah et al. (2014) reported that manual extraction of differential summary from a dE-MAP network is time-consuming, onerous and error-prone process. Hence, they proposed a method called DiffNet that automatically generates summaries of a dE-MAP network to obtain a detailed map of functional responses due to condition change. In brief, the DiffNet algorithm leverages combination of Gene Ontology annotations and interaction data to find a group of functional sub-graphs which are highly skewed. The obtained sub-graphs represents significant functional responses emerged as a result of change in condition. However this approach can not be applied to more than two treatments (Seah et al., 2014).

Ma et al. proposed a technique called “machine learning–based differential network analysis” (mDNA) for the comparison of gene networks. Machine learning is an advanced data mining technique that creates a prediction model using prior knowledge to identify important patterns in large datasets. Before network construction, the mDNA algorithm removes non-expressed, constitutively expressed and non-informative genes by using a machine learning based filtering process. In the second step, it analyzes the retained informative genes to estimate candidate stress-related

genes based on extracted patterns. In the study, two previously unreported genes in SALK T-DNA mutagenesis lines were identified (Ma et al., 2014).

Gill et al. developed a method called dna which determines differential modular structures between two networks using connectivity scores. A connectivity score represents the strength of association between gene pairs. The algorithm use partial least squares (PLS) as a default connectivity score. The other statistical tests that can be used to compute connectivity scores are ridge regression, principal components regression and the correlation coefficient. In the study, dna is applied to two types of synthetic datasets as well as one real data set and identified some set of genes that may have important functions in obesity (Gill et al., 2010; Gill et al., 2014b).

In the study conducted by Warsow et al. (2013), a software tool ExprEssence was used to determine pre-operative breast cancer chemotherapy response. ExprEssence tool extracts sub-networks by selecting the associations of a gene/protein network that are most differentially regulated between sample groups. In the first step, the algorithm determines the link score for each association. Then a sub-network is extracted using the interactions with the largest link scores. In the study, performance of the resulting sub-network was compared against two other sub-network identifying algorithms, KeyPathwayMiner and OptDis (Warsow et al., 2013).

DINA (Differential Network Analysis) approach was developed by Gambardelle et al. (2013), to find set of genes which co-regulation of them is condition-specific. It starts from a group of condition-specific gene expression profiles. The algorithm can detect the TFs that may be responsible for the pathway condition-specific co-regulation. The authors identified several metabolic pathways as the most differentially regulated across the tissues using 30 tissue-specific gene networks in human. Transcription factors as Nuclear Receptors was identified as their main regulators and showed that a gene with unknown function (YEATS2) acts as a negative regulator of hepatocyte metabolism. The results also found that hypotheses on dysregulated pathways during disease progression can be made by using this method. DINA identified hepatocarcinoma-specific metabolic and transcriptional pathway dysregulation.

Madhamshettiwar et al. proposed the “Regulatory Module Network Inference” (RMaNI) framework for the inference, analysis and visualization of condition specific GRN and differential network analysis. The framework identifies relevant regulatory transcription factors (TF) by using the “Learning Module Networks” (LeMoNe) (Josji et al., 2009) and Regulatory Impact Factors (RIF) (Reverter-Gomez et al., 2010) algorithms. Briefly, RMaNI combines heterogeneous knowledge resources and includes a set of several bioinformatic methods such as DE analysis, module identification, regulator detection, functional enrichment analysis and visualization. In the study, RMaNI framework was applied to hepatocellular carcinoma dataset that contains normal and three disease samples (Madhamshettiwar et al., 2013).

Odibat and Reddy (2012) presented DiffRank differential network analysis algorithm in order to find the DE genes representing two biological conditions such as healthy and disease. The algorithm can be applied on directed and undirected networks since it doesn't depend on the network construction. In the study, the authors identified two structural scoring measures which are a global and a local structure measure. These are optimized by propagating the scores through the structure of network and ranking the genes based on these scores. This method identifies the changes in the edges and the change in the centrality of each gene. The utility of the algorithm was tested on synthetic and real datasets by comparing the method with the previous ones. DiffRank can be detect the local and the global changes in the topological structures between two given gene networks. In the study, the algorithm was applied to synthetic and real-world datasets.

Zhang et al. developed a method - “the differential dependency network” (DDN) - that aims to detect statistically significant topological changes in transcriptional regulatory networks between two different biological conditions. DDN algorithm use Lasso technique to learn the local dependency model that characterizes the dependencies of genes in the network and represents the local structures of a network. It estimates the statistical significance of each learned local structure by using a permutation test. The algorithm was applied to a simulated dataset as well as to a breast cancer cell line dataset and the authors concluded that the results were biologically meaningful (Zhang et al., 2009; Zhang et al., 2011).

Ma et al. (2011) presented a method that employs a scoring function jointly measuring the condition-specific changes of both individual genes and gene–gene co-expression. The algorithm uses a genetic algorithm to detect the optimal subnetwork which maximize the scoring function. This method called COSINE is useful for identifying significant subnetworks of appropriate size and meaningful biological relevance. Compared to other methods, it considers single gene’s expression variation and gene pair’s differential correlation and extracts a globally optimal sub-network that can maximize the across-group difference.

Valcarcel et al. (2011) developed a method for the differential analysis of molecular associations via network representation. Based on conventional statistical methods, there were some differences in concentration levels of lipoprotein subclasses between people with normal fasting glucose and people having prediabetes. The results showed the applicability of the approach to identify key molecular changes inaccessible to standard approaches.

Zhang et al. (2011) presented differential dependency network to determine and visualize statistically significant topological changes of transcriptional networks representing two different biological conditions. This tool provides an alternative way to define network biomarkers predictive of phenotypes. It was developed to identify the rewiring of the underlying biological network triggered by outside stimuli or different conditions using gene expression data.

A method called DEGAS was proposed by Ulitsky et al. (2010) to define connected gene subnetworks enriched for genes that are dysregulated in specimens of a disease. The method was applied to seven diseases collecting thirteen case-control gene expression datasets and the results were statistically significant. The subnetworks that are defined by DEGAS can provide a useful signature for diagnosis, possible pathways, and also offer drug intervention targets. Application of this computational technique to a large-scale protein-protein interaction network and expression data of human diseases demonstrated the method’s success. The results showed novel evidence in mRNA splicing, cell proliferation, and the 14-3-3 complex involvement in Parkinson’s progression. Compared to other methods, this method identifies more specific subnetworks that capture a significant fraction of the known disease-related pathway.

3. METHODOLOGY

In this part of the dissertation, the aim of the study, proposed framework and the implementation of the algorithm including R package and web versions are presented.

3.1. MOTIVATION

Different cell conditions are results of different associations between genes. Therefore, from each cell condition, different gene network can be illustrated. In fact, between two cell conditions there are common gene interactions and also different interactions from the other. This fact can be used to find new and more targeted biomarkers or drugs. For instance, comparing a normal cell gene network and a breast cancer gene network will result many disease-specific genes and gene interactions, in which some of them may be the main cause of the breast cancer. In order to find disease-specific interactions, a differential gene network inference algorithm, *dc3net*, was introduced in Altay et al. (2011). However, since we work in genome-wide, using this algorithm alone results many disease-specific interactions from which it is not easy to determine the main causes of the disease. In order to rank those disease-specific interactions and also the genes in the network, we developed a new approach by integrating some other available datasets for breast cancer. This integration framework has resulted the most important genes and interactions by allowing ranking the breast cancer specific gene network which was inferred from the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) breast cancer dataset (Curtis et al., 2012). We call this framework as *Integrative Differential Network (IDN)* framework in the rest of the dissertation.

Although the framework has been developed on breast cancer, IDN can be extended for other diseases as well by replacing the datasets accordingly. For a different disease, there may be more or less types of datasets than the IDN introduced in this study for breast cancer but the main integration approach can be the same. Basically, infer disease-specific gene network and integrate on the network with the related datasets found for the disease.

3.2. THE PROPOSED IDN ALGORITHM

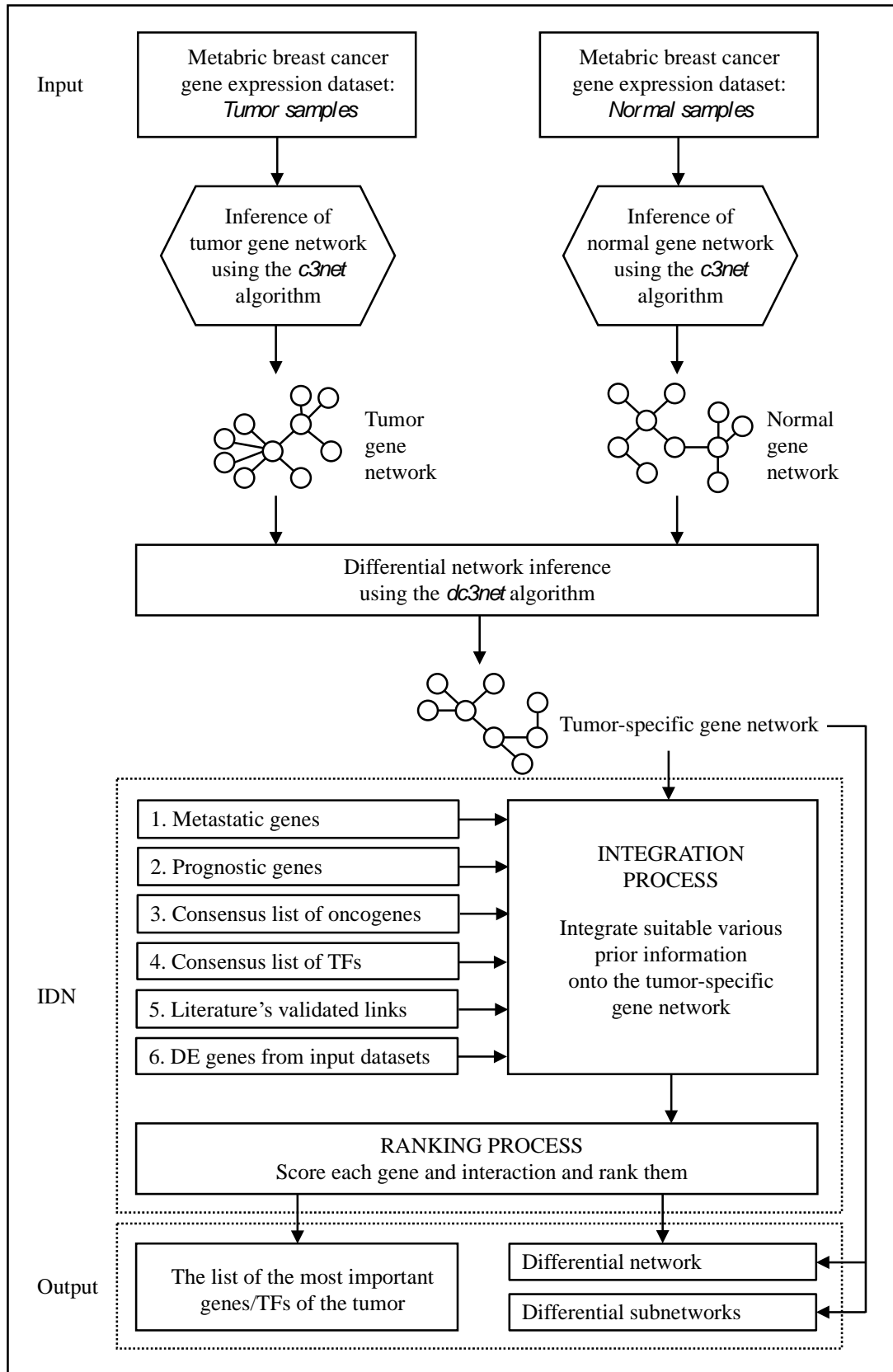
In this part, the proposed IDN approach and its parameters are presented.

3.2.1. Introduction

IDN is a novel approach for differential network analysis with its integrative framework. An outline of the IDN framework with main processes is shown in Figure 3.1. In the first step, we apply the GNI algorithm *c3net* (Altay, 2010a) to tumor and normal samples and infer tumor and normal gene networks, respectively. In the second step, we compare these gene networks with respect to mutual information (MI) values of each interaction between the two networks using the differential networking algorithm *dc3net* (Altay, 2011). In the final step, we obtain a breast cancer specific gene network which the integration of information from other datasets will be performed. Following this, IDN extracts the subnetworks from tumor-specific differential network which will let us to detect hub nodes. Hub nodes are genes that are highly connected with other genes and they were proposed to have important roles in biological development. Since hub nodes have more complex interactions than other genes, they may have crucial roles in the underlying mechanisms of disease (Guo, 2015).

In the integration step of IDN, we first integrate literature information. The interactions out of the first step of *c3net*, basically all the significant interactions, are compared with the interaction database of literature and the overlapping ones are added to the breast-specific gene network. For this process, *ganet* R software package was used (Altay and Altay, 2013) for the literature integration, in which almost all the molecular interactions in literature is combined, and in that there are computational tools to perform analysis such as overlapping, comparison and so on. The reason of integration of the validated interactions from the literature was described in (Altay and Altay, 2013), where it was hypothesized that if there is a significant association score between two genes in a cell condition and if there is a validation in the literature in any other cell condition then it is highly likely that interaction might exist in the current cell condition too. Therefore we integrate *validated interactions* from literature with significant association score in the association matrix of the GNI algorithm *c3net*. Similar literature integration approaches are popular nowadays (Olsen et al., 2014).

Figure 3.1. IDN framework overview.



Source: This dissertation.

When merging the tumor specific network and the validated literature links of the significant interactions, the resulting network has 19382 unique interactions. Then other information from the other datasets is integrated over this network. Secondly, we perform a differential network analysis over the same breast cancer datasets and obtain a list of *differentially expressed* (DE) genes in tumor condition with the p-value of 0.01. We then integrate this list of DE genes on the combined network too. Following this, we integrate the effect of *metastatic genes* using the list of under and over expressed list of genes from the dataset of (Varambally). We also integrate the effect of prognostic genes to include survival influence using the list of under and over expressed list of genes from the dataset of (Glinsky). Finally, we integrate the list of oncogenes (cancer related genes) available in the literature (<http://cancer.sanger.ac.uk>).

Since the main purpose of the algorithm is to find the most important regulators that drives the breast cancer, we focused on transcription factors (TF). We used a list of TFs assuming that it includes all of them, which we have downloaded from (<http://www.tfcats.ca/index.php>). We integrated TFs to the combined network and get a new list of TFs that exist in the combined network. We then grouped each of the TFs depends on the sub-network it belongs to. We then start scoring each TF as follows. While scoring, in order to consider the neighbor density with respect to the proximity of each TF, the link distance from each TF was limited to 2 links. This means while scoring each TF, any gene or link in two steps away affects the score of a TF and consequently the higher the neighbor density the higher the score is. Association score of each interaction, which is normalized between 0 and 1, is added to the score. Also if any of the genes from the literature integration lists exist in the two step neighborhood of a TF, then they increase the score by one. Finally the score for each TF is calculated. The scoring formula is as follows:

$$\text{Score of a TF (in 2 step neighborhood)} = N + \sum MI_s + MG + PG + DEG + OG$$

- N*: Number of neighbours
- MI*: Mutual information values
- MG*: Number of metastatic genes
- PG*: Number of prognostic genes
- DEG*: Number of differentially expressed genes
- OG*: Number of oncogenes

This way of grouping is a novel approach that was not seen in the literature with the best of our knowledge. This formula can also be extended by adding any other related datasets. Two step neighborhood might be selected as 3 or more if the scores are not high enough to rank TFs. The inputs, outputs and the main procedure of IDN framework is shown in Figure 3.2.

Figure 3.2. IDN algorithm for breast cancer

Input:	Two gene expression or correlation (dependency) matrix e.g. tumor and normal Two vector including gene names for the users who wants to work in gene level
Output:	Differential network Differential subnetworks The list of the most important genes/TFs of the tumor
Procedure:	<p>Step 1. Inference of tumor and normal gene networks <i>Infer tumor and normal gene networks using the GNI algorithm c3net. If the input datasets are correlation matrices, firstly compute mutual information matrices for each dataset</i></p> <p>Step 2. Inference of differential network <i>Infer tumor-specific gene network using the dc3net algorithm</i></p> <p>Step 3. Estimation of tumor-specific subnetworks <i>Extract subnetworks from tumor-specific differential network</i></p> <p>Step 4. Integration process <i>Integrate suitable various prior information onto the tumor-specific gene network:</i></p> <ul style="list-style-type: none"> • <i>Metastatic genes</i> • <i>Prognostic genes</i> • <i>Consensus list of oncogenes</i> • <i>Consensus list of TFs</i> • <i>Literature's validated links</i> • <i>DE genes from input datasets</i> <p>Step 5. Ranking process <i>Rank the most important genes/TFs of the tumor according to the computed scores</i></p>

3.2.2. Parameters

The required inputs of the IDN are two different gene expression data sets, e.g. tumor and normal, and optionally gene names. Users can also use pre-computed tumor and control mutual information (adjacency) matrices as input. Otherwise, the algorithm takes the two data sets and generates the matrices itself.

The MI matrices are square adjacency matrices where the MI value corresponds to the weight of interaction for each gene pair. The diagonals are set to zero to ignore self-interactions. The next step is computing row wise ranked versions of these MI matrices in descending order. Here, rank 1 corresponds to the highest mutual information value in a row of the matrix. This ranked matrices will be used in comparing and filtering the networks at the comparison step. Then the *c3net* algorithm is applied to the tumor and control MI matrices to infer gene networks of direct physical interactions of tumor and control datasets independently.

There are eight parameters users can set. The first six parameters, *method*, *cutoff*, *alpha*, *itnum*, *rankdif*, and *percentdif* are used by *c3net* and *dc3net* algorithms to control the network inference and decision filtering steps. Novice users may prefer to use the default parameters offered by the system.

The first four parameters, *method*, *cutoff*, *alpha* and *itnum*, are used to eliminate non-significant interactions in *c3net* (Altay and Emmert-Streib, 2010a). The available options for *method* parameter are “cutoff”, “justp”, “holm”, “hochberg”, “hommel”, “bonferroni”, “BH” and “BY”. *cutoff*, *alpha* and *itnum* parameters are dependent to *method* parameter. If “cutoff” is selected as the *method*, *cutoff* value must be entered which can be zero or predefined cut-off value. Zero means that mean of upper triangle will be taken as *cutoff*. If the *method* is “justp”, *alpha* and *itnum* (iteration number) parameters are need to be adjusted. The other options, “holm”, “hochberg”, “hommel”, “bonferroni”, “BH” and “BY” are multiple testing correction (MTC) methods. Users can apply MTC methods by selecting the name of MTC method. If the selected *method* is one of the MTC method, then *alpha* and *itnum* parameters are need to be set as it were in “justp” method. The default *method* was set to “cutoff” which uses mean of upper triangle of MI matrices as a significance threshold.

The next two parameters, rankDif and percentDif, are for the comparison step of dc3net. This is the core part of *dc3net* that it compares the two networks to find differential network. In the comparison step, there are four conditions that all must be validated at the same time for an edge to be included in tumor differential network.

Suppose that we check the potential interaction gene A to gene B to be included in differential network or not. As we stated above, we have been computed row wise ranked versions of the MI matrices in descending order. So we know the rank of interaction gene A to gene B in control MI matrix. The first parameter of *dc3net*, rankdif, is the predefined cutoff parameter that checks the interaction between gene A and gene B is one of the top ranked interactions in control MI matrix or not. If the rank of gene A and gene B in the ranked control MI matrix is greater than the predefined cutoff parameter, rankdif, then the first condition becomes valid for deciding it as a difnet interaction. rankdif parameter can be adjusted to any value between 1 and number of rows of control MI matrix. However, if user wants a stricter difnet, then rankdif parameter needs to be adjusted to a greater value.

The second condition is the change in MI value of interaction from gene A to gene B in the control MI matrix. Here, algorithm uses MIDif value as the cutoff parameter. MIDif is defined as percentdif times the maximum MI value of the row of geneA in the control MI matrix. Default value for the percentdif parameter is 0.6. Depends on strictness of the differential network, user can increase or decrease the second cutoff parameter. The previous two conditions compared the interaction of gene A to gene B but we also need to compare the interaction of gene B to gene A. So the algorithm validates the first and second conditions also for the interaction of gene B to gene A. In this example, if four of the conditions are validated, then *dc3net* infer this interaction as in tumor differential network and continue to perform same filtering process for all gene pairs in test network. Since the integrative part of the framework is developed special to breast cancer, the rest of the process involves score calculations.

The last two parameters, *p-value* and *step* are special to IDN algorithm. While *p-value* is used in the differential expression process, *step* parameter sets the link distance from each TF that will be used in the ranking step of IDN.

3.2.3. Gene Ontology Analysis

“Gene Ontology” (GO) enrichment analysis based on Gene Ontology database (<http://www.geneontology.org>) was performed to investigate the biological roles of the genes in the differential network (da Huang, 2009). To further assess the signalling pathway of the genes, we subsequently performed “Kyoto Encyclopedia of Genes and Genomes” (KEGG, <http://www.genome.jp/kegg>) pathway enrichment analysis. The two analysis were performed using “The Database for Annotation, Visualization and Integrated Discovery” (DAVID, <https://david.ncifcrf.gov>) which is a powerful bioinformatics tool to find out functions of interested genes (Dennis, 2003). In the enrichment analyses, the pathways with minimum gene number 5 and $p < 0.05$ are considered as significant.

3.3. IMPLEMENTATION

IDN framework was developed in the R statistical computing language which is a free software environment that runs on a wide variety of operating systems such as Linux, Windows and MacOS (<https://www.r-project.org>). Secondly, the web version of IDN was developed for the ease of use of biologists that are not computer experts or have not got computationally power computers.

3.3.1. The IDN R Package

In this part, installation steps of the *IDN* R package and general guidelines for using the the package is provided.

3.3.1.1. Installation of the IDN R package

IDN requires “R 3.2.x and later” and it depends on “*c3net*”, “*dc3net*” and “*RedeR*” packages that can be installed from the CRAN (<https://cran.r-project.org>) and Bioconductor (<https://www.bioconductor.org>) libraries. For the installation of *IDN*, users need to follow some simple installation steps.

The following commands should be executed in R console.

1. To download and install dependent packages *c3net*, *igraph* and *RedeR* from CRAN and Bioconductor (execute in R):

```
> install.packages("c3net")
> install.packages("dc3net")
> source("http://bioconductor.org/biocLite.R")
> biocLite("RedeR")
```

2. Execute the installation command for *dc3net* in R

```
> install.packages("idn1.0.tar.gz", type="source", repos=NULL)
```

3. To load the library:

```
> library("idn")
```

3.3.1.2. General guidelines for using the IDN R package

The following command would be an example of a call to the main function of *IDN*:

```
> output <- idn(tumorData, normalData, genes, difnet, pValue, step)
```

where the first two inputs are test and control datasets, e.g. tumor data and normal data, respectively. *genes* variable is a vector that includes gene names. *difnet* is the output differential network of the *dc3net* algorithm. *pValue* is the p-value that is used for differential expression step of IDN. *step* is the link distance from each TF that is used in the ranking step of IDN. This command assign the analysis results to *output* R environment variable.

Users can access to analysis results using the following commands:

```
> subnets <- output$subnets
> hubgenes <- output$subnetList
```

subnets variable contains output subnetworks and *hubgenes* variable contains hub genes that are ranked according to the computed scores.

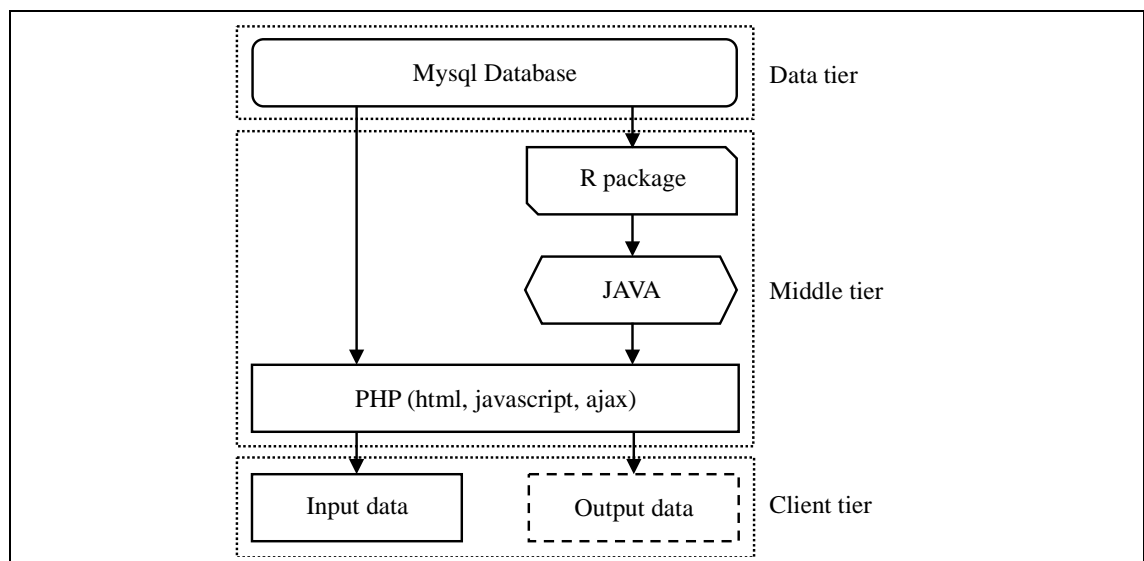
3.3.2. The IDN Web Application

IDN is a web-based application that is developed for differential network analysis of gene expression data. It is impractical and time consuming for biologists and non-technical experimentalists to install software applications on their local computers. On the other hand, these processes usually requires high computational resources. Hence, user-friendly and powerful web applications are important to overcome these drawbacks. Below, the system architecture of IDN is described.

3.3.2.1. System Architecture

Three-tier architecture model which consist of three independent layers or tiers, e.g. presentation, business logic and data tiers, is commonly used model for web-based applications. While the presentation or user services layer is the layer that user interacts with the application through a user interface, the data layer consists of data access components and manages the internal and external storage of application-related data. The business logic layer acts as an intertie which handles the communication between the other two layers and performs logical operations using computational resources (Yang, 2013). IDN web application is designed based on the widely used three-tier architecture model (Figure 3.3).

Figure 3.3. Three-tier architecture of IDN web application



Source: This dissertation.

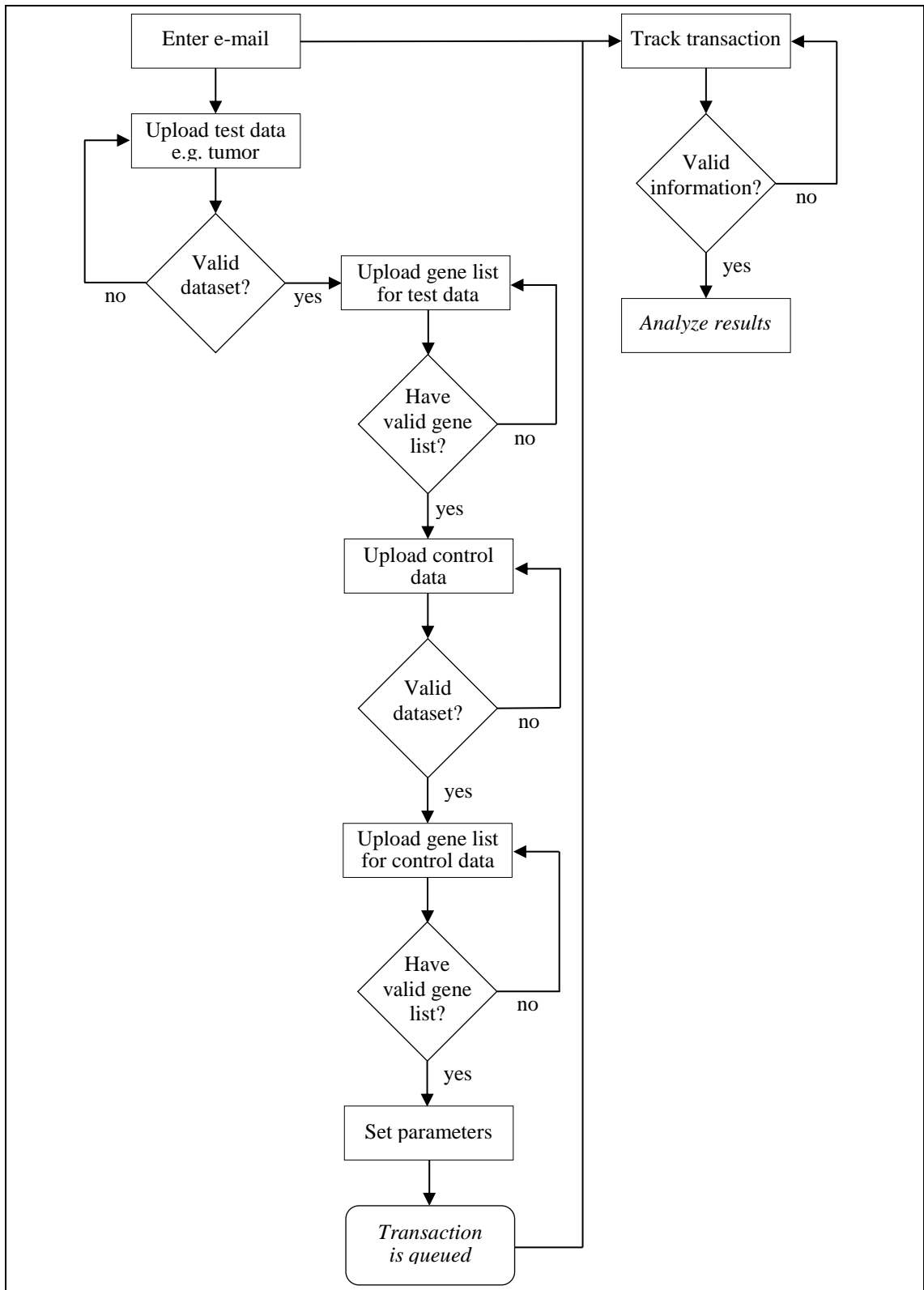
3.3.2.2. Workflow

IDN web application was originally implemented in R which is a popular statistical package among biologists. In the middle tier, PHP scripts gets the datasets and required parameters from client tier and saves the information to the Mysql database upon validation of the information. Since the differential gene networking process is a high computational task, the system queue the jobs coming from user tier and perform operations in order.

The computation time vary according to the set parameters and the size of the datasets. For this purpose, the system consistently checks the database for waiting jobs. When a new task arrives, the system calls the R package through JAVA framework and update the status of related task as processing. This let users to track the status of submitted tasks easily. When the computation is over, the system update the status of related task as completed and inform the user via e-mail. Then users can access to analyze results through the track transaction part of IDN web application using his e-mail and transaction ID information.

IDN generates three output files at the end of the analysis which are breast cancer specific differential network, differential subnetworks including the hub genes and the list of the most important genes/TFs of the tumor. Users who want to visualize the output networks can use a visualization package such as igraph (Csardi and Nepusz, 2006), Cytoscape (Shannon et al., 2003) or ReDer (Castro, 2012). In this dissertation, gene networks were plotted using RedeR software package. The detailed usage guide of IDN web application with screenshots is demonstrated at the case study part of the study.

Figure 3.4. Workflow of IDN-web



Source: This dissertation.

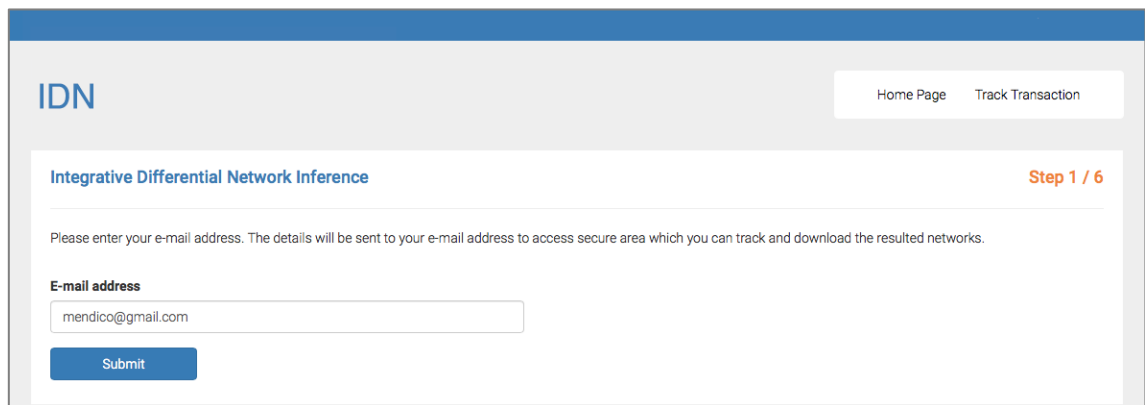
3.3.2.3. A Case Study of IDN Web Application

IDN web requires two different microarray gene expression datasets as input, one as tumor and the other as normal. Additionally, users who want to work with gene names instead of probe names have to prepare two files that include gene names for each dataset. The input files format should be CSV or TXT and should be prepared according to the instructions in the website.

Step 1

In the first step, the system asks user to enter his e-mail address. This is mandatory since users track their transactions by using the transaction id and e-mail addresses. Additionally, the system sends an e-mail to the user when the transaction completes. (Figure 3.5).

Figure 3.5. Differential gene network inference: Step 1



The screenshot shows the IDN web application interface. At the top left, the logo "IDN" is displayed. In the top right corner, there are two navigation links: "Home Page" and "Track Transaction". The main content area is titled "Integrative Differential Network Inference" and indicates "Step 1 / 6". Below the title, a message states: "Please enter your e-mail address. The details will be sent to your e-mail address to access secure area which you can track and download the resulted networks." There is a text input field labeled "E-mail address" containing the text "mendico@gmail.com". Below the input field is a blue "Submit" button.

Step 2

In the second step, the system asks user to upload test data set (e.g. tumor). Test data set must be a gene expression matrix which rows correspond to the variables (e.g. probes) and columns correspond to the samples. Users can upload data set by clicking "Select File" and then "Upload data set" button. The upload process will be shown to the user interactively (Figure 3.6).

Figure 3.6. Differential gene network inference: Step 2

Home / Step 1: User details / Step 2: Upload tumor data set

Differential Gene Network Inference: Upload test data set (e.g. tumor) Step 2 / 6

Your e-mail address is: **mendico@gmail.com**. If you mistyped your email address, you can [click here](#) to correct it.

Please upload your **test (e.g. tumor)** data set in **.csv** format where rows are variables (e.g. probes) and columns are samples.

Example gene expression data set: [Excel screenshot](#) [Notepad screenshot](#)

	A	B	C	D
1		sample1	sample2	sample3
2	probe1	0.193125	0.069375	0.106875
3	probe2	0.256875	0.055625	0.125625
4	probe3	0.205625	0.133125	0.114375
5	probe4	0.170625	0.116875	0.180625

```
,sample1,sample2,sample3
probe1,0.193125,0.069375,0.106875
probe2,0.256875,0.055625,0.125625
probe3,0.205625,0.133125,0.114375
probe4,0.170625,0.116875,0.180625
```

To download the example datasets, simply right click on [this link](#) and choose "save as" from the pop-up menu.

Data set (Comma seperated .csv or .txt file)

seçili dosya yok

The input files must be prepared in comma separated CSV or TXT file format as shown in the figures at below. Example data set can be download through the link at the same page.

Step 3

In the third step, the system gives information to the user about the uploaded test data set at the previous step (Data type, probe number and sample size). In this step, users should upload gene names correspond to the probe names in the test data set. The number of rows (the number of genes) in the gene names file must be identical to probe number. Otherwise, the system will give an error message and ask user to re-upload gene names. Users can upload gene names in CSV or TXT file format where each row represents a single gene. Example file can be download through the link at the page (Figure 3.7).

Figure 3.7. Differential gene network inference: Step 3

Home / Step 1: User details / Step 2: Upload tumor data / Step 3: Upload tumor gene names

Differential Gene Network Inference: Upload tumor gene names Step 3 / 6

Tumor data type: **Gene expression data set.**
The probe number in your data set is: **500.**
The sample size of your data set is: **52.**

You can upload gene names in **.csv** or **.txt** file format where each row represents a single gene. The number of rows (the number of genes) in the file must be identical to probe number, **500**. To download the example data set, simply right click on [this link](#) and choose "save as" from the pop-up menu.

Gene names: **Excel screenshot** **Notepad screenshot**

	A
1	gene1
2	gene2
3	gene3
4	gene4

```
gene1
gene2
gene3
gene4
```

Gene names (.csv or .txt file)

seçili dosya yok

Step 4

In the fourth step, the system asks user to upload control data set (e.g. normal). Control data set must be a gene expression matrix which rows correspond to the variables (e.g. probes) and columns correspond to the samples.

Step 5

In the fifth step, the system gives information to the user about the uploaded normal data set at the previous step (Data type, probe number and sample size). Users must upload gene names that correspond to the probe names in the control data set.

Step 6

In the sixth and last step, the system ask users to enter parameters. The detailed information about parameters was given in the parameters part of the dissertation. The default parameters recommended by the system is shown in this page. Users can click to “question mark” icons to get information about the parameters.

Some of the parameters are shown according to the selected method. Hence, all eight parameters are not shown in Figure 3.8. The differential gene network inference process starts after user clicks to “Infer Differential Network” button.

Figure 3.8. Differential gene network inference: Step 6 - Parameter selection.

Home / Step 1: User details / Step 2: Upload tumor data / Step 3: Upload tumor genes / Step 4: Upload normal data / Step 5: Upload normal genes / Step 6: Parameters

Differential Gene Network Inference Step 6 / 6

Please set required parameters. If not sure, you can use the default parameters that are dynamically change with respect to your data set.

C3NET Parameters

Method ?
Cutoff

Cutoff ?
0

DC3NET Parameters

rankDif ?
100

percentDif ?
0.6

IDN Parameters

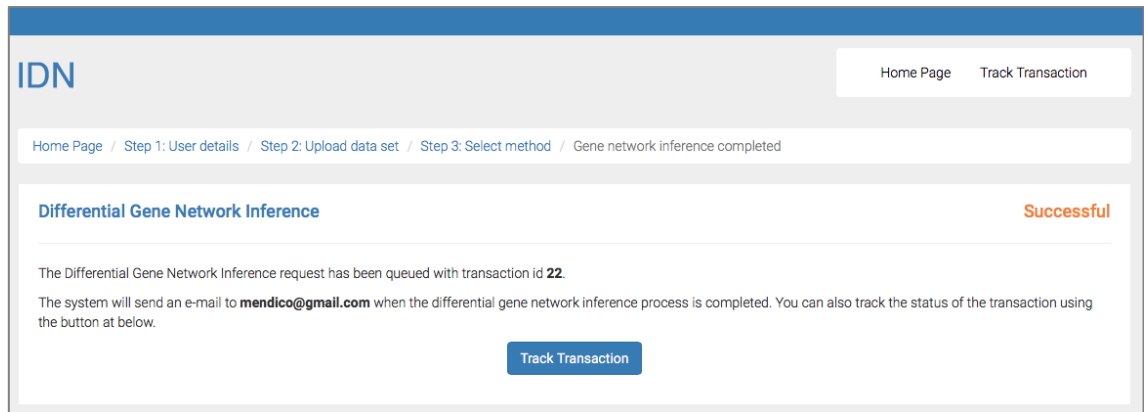
p-Value (for differential expression)
0.01

Step
2

[Infer Differential Network](#)

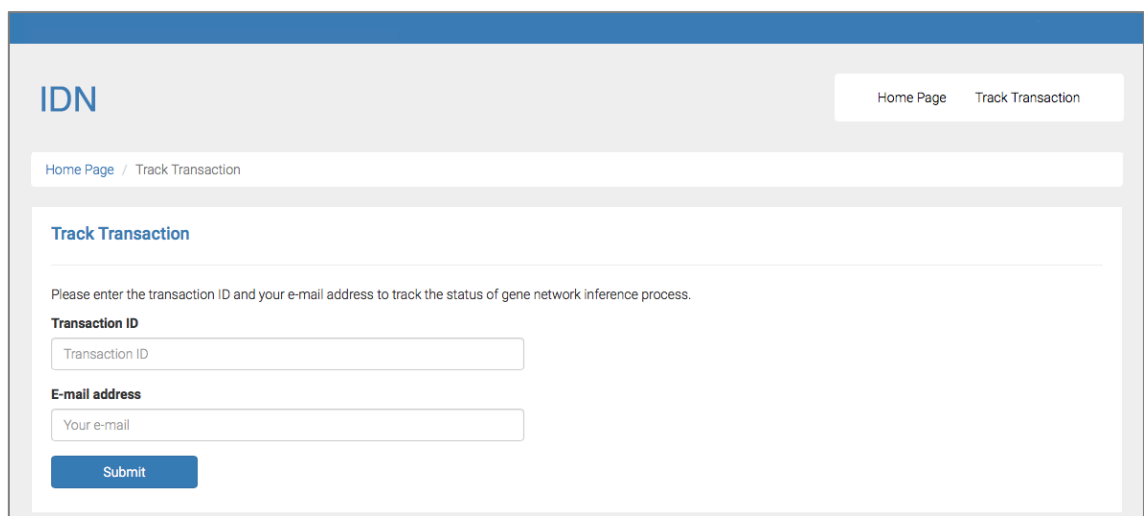
After user clicks to “Infer Differential Network” button, the system queue the transaction with the transaction ID as shown in the next page (Figure 3.9). The system sends an e-mail to the user when the transaction completes. This may take minutes or hours according to the number of waiting tasks and size of the data sets.

Figure 3.9. Differential gene network inference successfully submitted.



Users can track the transaction status through the link “Track Transaction”. The systems asks users to enter transaction Id and e-mail address to show the status of the transaction (Figure 3.10).

Figure 3.10. Track transaction.



The transaction status, “In queue”, means that the task is waiting for execution.

The transaction status, “Processing”, means that differential gene network inference task is processing. The system sends an e-mail after the execution is completed.

The transaction status, “Completed”, means that the differential gene network inference process is completed. Users can now download the inferred differential gene network through the link at the lower-left corner of the page. The details about the transaction (Date submitted, e-mail, data types, data completed and dimension (the number of interactions) of the inferred network) as well as the parameters used in the computation are also shown in this page (Figure 3.11).

Figure 3.11. Transaction is completed.

The screenshot shows the IDN web interface. At the top left is the 'IDN' logo. At the top right are links for 'Home Page' and 'Track Transaction'. Below the navigation bar is a breadcrumb trail: 'Home Page / Track Transaction / Status of Transaction'. The main content area is titled 'Status of Transaction' and has a 'Completed' status indicator in orange. A message states: 'The details of the gene network inference process is stated at below. This page update itself **automatically** so you don't need to refresh it.'

Transaction details are listed on the left:

- Transaction ID : 22
- Date submitted : 2016-11-07 12:11:08
- E-mail : mendico@gmail.com
- Date completed : 2016-11-07 12:14:29
- Differential network : (2.76 KB)
- Difnet dimension : 96
- Subnets : (5.06 KB)
- Ranked TFs : (399.00 B)

Parameters are shown in three tables:

C3NET Parameters:		Method Step 1	Cutoff value	Alpha	Iteration Number
Value:		Cutoff	0	-	-

DC3NET Parameters:		rankDif	percentDif	rankCom	percentCom
Value:		100	0.6	10	0.85

IDN Parameters:		P-Value	Step
Value:		0.01	2

A 'New transaction' button is located at the bottom right of the main content area.

4. RESULTS

In the first section of this part, the preliminary analyses that are performed on prostate cancer and lung cancer datasets are presented and discussed. Secondly, breast cancer analysis including the integration of different type of datasets and the analyses results are presented.

4.1. PRELIMINARY ANALYSES

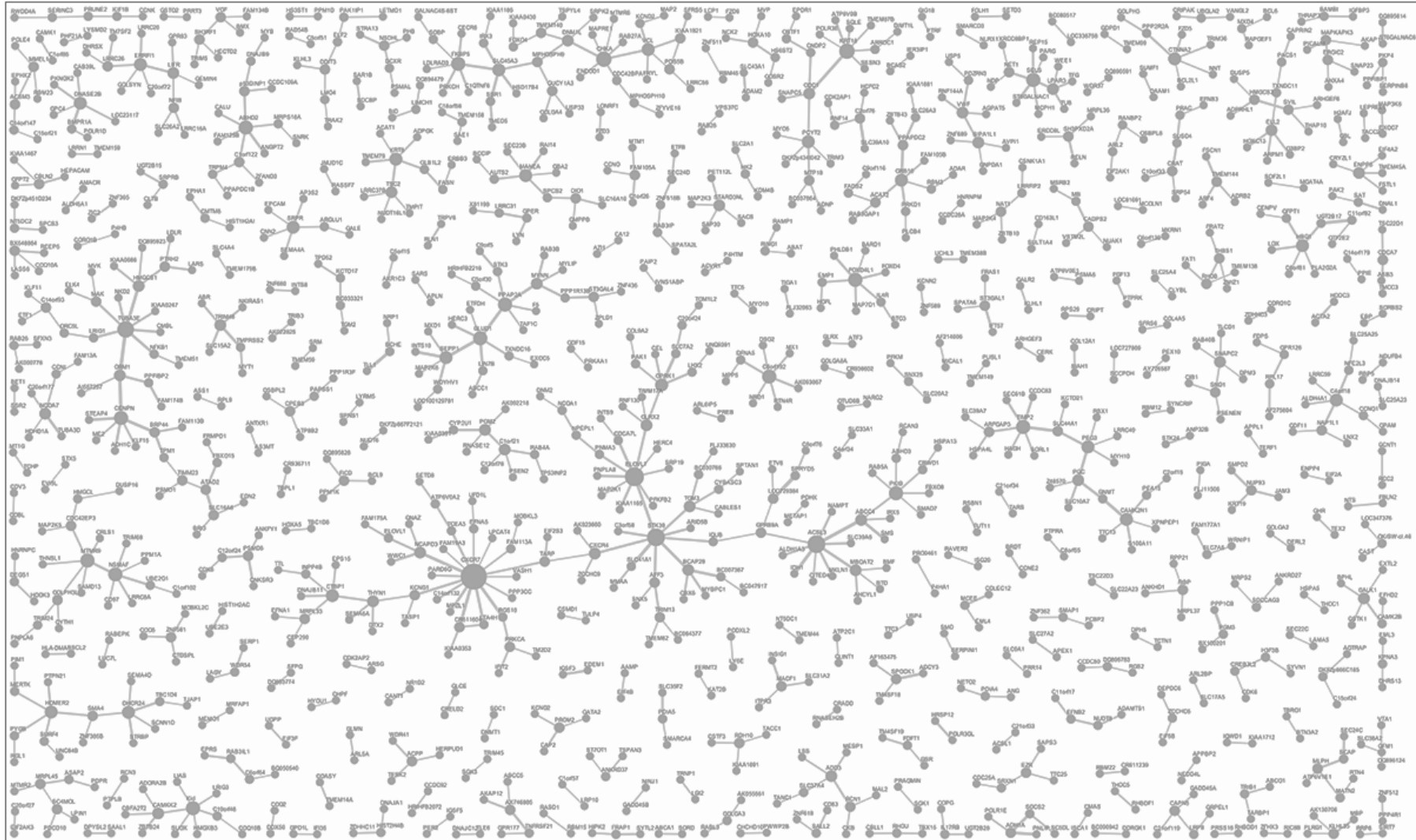
In this part of the study, two preliminary differential networking analysis was performed on prostate cancer and lung cancer datasets. These analysis were performed to test IDN framework's integration with *c3net* and *dc3net* algorithms as well as to test subnetwork extraction feature of IDN. Since the integration and ranking steps of IDN is special to breast cancer, these steps were not applied for these datasets.

4.1.1. Inference of Prostate Cancer Specific Differential Network

Prostate cancer is the second most common cancer in the male population, with an estimated 417,000 new cases diagnosed each year in Europe (Ferlay et al., 2013). The activation of androgen receptor (AR) through androgens plays a crucial role in the development and progression of prostate cancer (Kaur et al., 2016; Anantharaman et al., 2015; Choudhary et al., 2011; Massie et al., 2011).

For early detection of prostate cancer, prostate specific antigen (PSA) screening method has been used widely as a diagnostic tool (Karatat et al., 2015). However, PSA fails to discriminate indolent disease which results in over-diagnosis and this may lead to poor prognosis (Abou-Ouf et al., 2015; Ma et al., 2015; Myers et al., 2015). Furthermore, there is no evidence showing that the PSA screening reduces the incidence of death and the underlying mechanism of prostate cancer progression remains largely unknown (Cannistraci et al., 2014; Ren et al., 2015).

Figure 4.1. Genome-wide androgen stimulated prostate specific differential network with 891 interactions



4.1.1.1. Microarray data

In order to investigate the alterations in androgen stimulated prostate cancer cells compared with androgen deprived prostate cancer cells, microarray dataset GSE18684 deposited by Massie et al. (2011) was obtained from the Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo>). The expression profile included 96 samples, comprising 20 androgen deprived tissue samples and 76 tissue samples with androgen stimulated prostate cancer.

4.1.1.2. Results

The androgen stimulated prostate cancer differential gene network with 891 interactions was inferred. The largest independent subnetwork with 119 interactions were also extracted from the differential network and plotted in Figure 4.1. To investigate the functions of the genes in the androgen stimulated prostate cancer differential gene network, GO and KEGG pathway analysis were performed. A total 184 terms were retrieved from the DAVID online analytical tool.

The top ten GO terms ranked by statistical significance were listed in Table 4.1. GO analysis revealed that genes associated with sterol biosynthetic process (GO:0016126; $p=5.05 \times 10^{-8}$), protein transport (GO:0015031; $p=2.57 \times 10^{-7}$) and establishment of protein localization (GO:0045184; $p=3.80 \times 10^{-7}$) were significantly enriched top three GO terms among biological processes, while for cellular components, nucleotide binding (GO:0000166; $p=7.08 \times 10^{-5}$), purine nucleotide binding (GO:0017076; $p=5.22 \times 10^{-4}$) and purine ribonucleotide binding (GO:0032555; $p=9.11 \times 10^{-4}$) were significantly enriched, and with regards to molecular functions, genes associated with endoplasmic reticulum (GO:0005783; $p=2.66 \times 10^{-13}$), endoplasmic reticulum part (GO:0044432; $p=1.16 \times 10^{-6}$) and organelle membrane (GO:0031090; $p=1.48 \times 10^{-4}$) were significantly enriched (Table 4.1, Figure 4.2A).

Table 4.1. GO terms of androgen stimulated prostate cancer specific differential network (top 10)

GO ID	GO term	No. of genes	p
Biological processes			
GO:0016126	sterol biosynthetic process	14	5.05E-08
GO:0015031	protein transport	82	2.57E-07
GO:0045184	establishment of protein localization	82	3.80E-07
GO:0046907	intracellular transport	73	3.96E-07
GO:0006695	cholesterol biosynthetic process	11	1.25E-06
GO:0016125	sterol metabolic process	21	1.87E-06
GO:0006886	intracellular protein transport	47	2.61E-06
GO:0008104	protein localization	87	4.16E-06
GO:0034613	cellular protein localization	49	6.76E-06
GO:0008203	cholesterol metabolic process	19	7.26E-06
Cellular components			
GO:0005783	endoplasmic reticulum	117	2.66E-13
GO:0044432	endoplasmic reticulum part	46	1.16E-06
GO:0031090	organelle membrane	97	1.48E-04
GO:0005789	endoplasmic reticulum membrane	33	2.11E-04
GO:0042175	nuclear envelope-endoplasmic reticulum network	34	2.63E-04
GO:0005739	mitochondrion	92	9.73E-04
GO:0005829	cytosol	109	9.89E-04
GO:0005792	microsome	28	1.24E-03
GO:0005624	membrane fraction	71	1.61E-03
GO:0042598	vesicular fraction	28	1.90E-03
Molecular Function			
GO:0000166	nucleotide binding	174	7.08E-05
GO:0017076	purine nucleotide binding	147	5.22E-04
GO:0032555	purine ribonucleotide binding	140	9.11E-04
GO:0032553	ribonucleotide binding	140	9.11E-04
GO:0000287	magnesium ion binding	42	4.10E-03
GO:0001883	purine nucleoside binding	119	5.75E-03
GO:0003924	GTPase activity	23	6.81E-03
GO:0001882	nucleoside binding	119	7.19E-03
GO:0005524	ATP binding	110	7.74E-03
GO:0004674	protein serine/threonine kinase activity	39	8.48E-03
FDR: false discovery rate; GO: gene ontology.			

Next, the genes found in the androgen stimulated prostate cancer differential gene network were submitted to DAVID server to identify significantly enriched KEGG pathways (Kanehisa, 2000; Kanehisa, 2012). The KEGG pathways that were found significantly enriched ($p < 0.05$) are shown in Table 4.2. Pathway analysis revealed that the genes in the androgen stimulated prostate cancer difnet were significantly enriched in ten terms. The most significant three terms were those involved in steroid biosynthesis ($p = 2.80 \times 10^{-7}$), synthesis and degradation of ketone bodies ($p = 1.646523 \times 10^{-3}$), and amino sugar and nucleotide sugar metabolism ($p = 1.73 \times 10^{-3}$) processes (Figure 4.2B).

Figure 4.2. Functional enrichment analysis of significantly enriched genes in the androgen stimulated prostate cancer specific differential gene network. (A) The top 10 enriched GO categories for biological processes; (B) The top 10 enriched KEGG pathways.

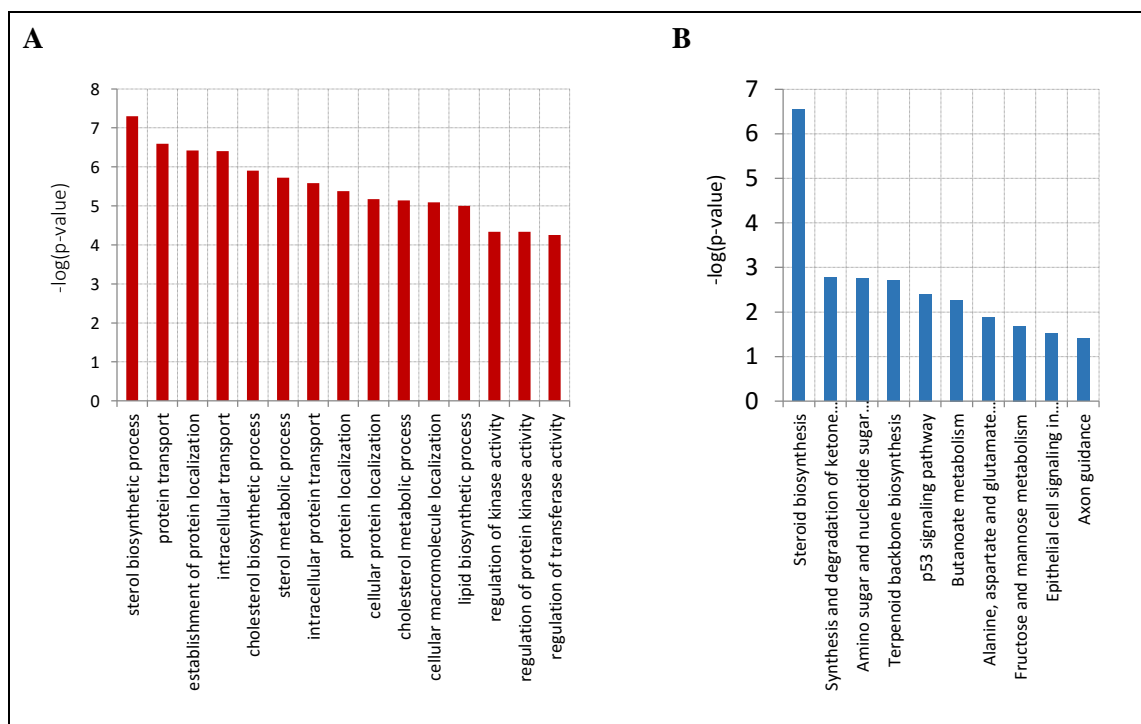


Table 4.2. Significant KEGG pathways in the androgen stimulated prostate cancer specific differential network

KEGG ID	KEGG term	No. of genes	p	Genes
hsa00100	Steroid biosynthesis	10	2.80E-07	TM7SF2, CEL, EBP, SQLE, LSS, SC5DL, DHCR24, FDFT1, SC4MOL, NSDHL
hsa00072	Synthesis and degradation of ketone bodies	5	1.69E-03	HMGCS2, HMGCS1, ACAT2, ACAT1, HMGCL
hsa00520	Amino sugar and nucleotide sugar metabolism	10	1.71E-03	PGM2, GMPPB, GALK1, PGM3, GNPDA1, CMAS, GFPT1, GFPT2, HK2, GALE
hsa00900	Terpenoid backbone biosynthesis	6	1.95E-03	HMGCS2, HMGCS1, FDPS, MVK, ACAT2, ACAT1
hsa04115	p53 signaling pathway	12	3.94E-03	CCNE2, BID, PPM1D, TSC2, SIAH1, CDK6, CCNG1, GADD45B, THBS1, IGFBP3, GADD45A, SESN3
hsa00650	Butanoate metabolism	8	5.33E-03	ACSM3, HMGCS2, ALDH5A1, HMGCS1, ABAT, ACAT2, ACAT1, HMGCL
hsa00250	Alanine, aspartate and glutamate metabolism	7	1.32E-02	ASS1, ALDH5A1, GFPT1, GLUD1, GFPT2, ABAT, ALDH4A1
hsa00051	Fructose and mannose metabolism	7	2.05E-02	MTMR2, GMPPB, SORD, PFKFB2, HK2, PFKM, MTMR6
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	10	3.03E-02	IGSF5, ATP6V0E1, LYN, ATP6V1E1, MAP2K4, NFKB1, PAK1, JAM3, ATP6V0A2, ATP6V0B
hsa04360	Axon guidance	15	3.89E-02	NRP1, EFNB3, EFNA1, EFNB2, DPYSL2, EPHA1, SEMA6A, NCK2, PAK2, CXCR4, PPP3CC, EFNA5, PAK1, SEMA4D, SEMA4A

KEGG: Kyoto Encyclopedia of genes and genomes

In order to further evaluate the biological roles of the genes in the independent subnetworks of the genome wide androgen stimulated prostate cancer difnet, KEGG analysis were performed for the largest subnetwork. As shown on Figure 4.3, this subnetwork comprises 119 interactions with CXCR7, STK39, ELOVL3 and ACSL3 at

the center of the largest hubs. KEGG analysis of the genes included in the subnetwork revealed a highly significant association with axon guidance pathway ($p=1.71 \times 10^{-3}$), which was also found significantly enriched in the whole differential network. Furthermore, pathways involved in Fc gamma R-mediated phagocytosis ($p=2.69 \times 10^{-2}$) and Endocytosis ($p=3.62 \times 10^{-2}$) were also highly enriched (Table 4.3). Interestingly, these two pathways were not found significantly enriched in the whole differential network.

Figure 4.3. The largest connected subnetwork of the androgen stimulated prostate cancer difnet. This subnetwork might have an important role in human prostate cancer as being the largest connected subnetwork with 119 edges in tumor difnet.

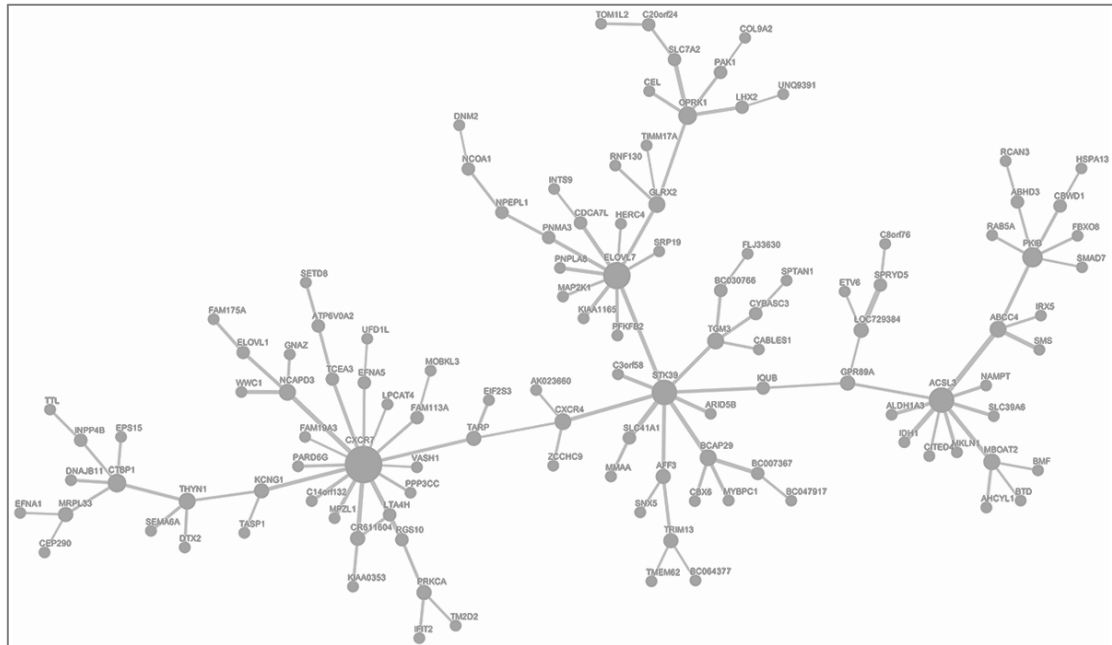


Table 4.3. Significant KEGG pathways in the largest subnetwork of the androgen stimulated prostate cancer specific differential network

KEGG ID	KEGG term	No. of genes	p	Genes
hsa04360	Axon guidance	6	1.71E-03	SEMA6A, CXCR4, EFNA1, PPP3CC, EFNA5, PAK1
hsa04666	Fc gamma R-mediated phagocytosis	4	2.69E-02	PRKCA, MAP2K1, PAK1, DNM2
hsa04144	Endocytosis	5	3.62E-02	EPS15, CXCR4, RAB5A, PARD6G, DNM2

KEGG: Kyoto Encyclopedia of genes and genomes

Top four hub nodes, identified in this trial, have been strongly associated with prostate cancer metastatic process, including CXCR7, STK39, ELOVL7 and ACSL3. Identification of hub genes involved in progression of prostate cancer may lead to the development of better diagnostic methods and providing therapeutic approaches.

According to analysis results, CXCR7 (chemokine (C-X-C motif) receptor 7) is by far the top hub gene in the androgen stimulated differential network and it is also part of the largest independent subnetwork as seen in Figure 4.3. In (Wang, 2008), it is reported that the levels of CXCR7/RDC1 expression increase on tumors being more aggressive. Also, In vitro and in vivo studies with prostate cancer cell lines propose that alterations in CXCR7/RDC1 expressions are associated with enhanced invasive and adhesive activities in addition to a survival advantage. Along other papers on CXCR7 (Zheng, 2010), it was shown that increased CXCR7 expression was found in hepatocellular carcinoma (HCC) tissues. Knockdown of CXCR7 expression by transfected with CXCR7shRNA significantly inhibited SMMC-7721 angiogenesis, adhesion and cells invasion. Moreover, down-regulation of CXCR7 expression leads to a reduction of tumor growth in a xenograft model of HCC (Zheng, 2010). Another study demonstrated that the IL-8-regulated Chemokine Receptor CXCR7 stimulates EGFR Signaling to promote prostate cancer growth (Singh, 2011). In a study conducted by Yun et al., it is reported that CXCR7 expression is increased in most of the tumor cells compared with the normal cells and is involved in cell proliferation, migration, survival, invasion and angiogenesis during the initiation and progression of many cancer types including prostate cancer (Yun, 2015). A more recent study indicated that there appeared to be disconnect of the effect of DHT on CXCL12/CXCR4/CXCR7 chemokine axis between transcriptional and translation machinery in androgen-responsive LNCaP cell line. There are many other studies that showed the strong role of CXCR7 in metastatic type cancer that strongly validates our blind foremost prediction is very likely to be true and thus needs further experimental work on its targets that we inferred in this study. However,

It was also observed that CXCR7/RDC1 levels are regulated by CXCR4 (Singh, 2011). This is a very interesting supporting information from literature for the blind estimation because in the predicted largest independent subnetwork, as shown in Figure 4.3,

CXCR7 and CXCR4 appear to be very close and interacting over only one gene. Although CXCR4 is not a hub gene, it appears to be as a bridge that connects both halves of the largest subnetwork. According to KEGG analysis, CXCR4 was found in the gene list of two different significantly enriched KEGG pathways, axon guidance and endocytosis which are strongly associated with prostate cancer (Table 4.2). Considering the prediction was made on global level, this literature confirmation seems assuring but not a coincidence. Therefore, this relation is worth experimenting in LnCap cancer too. It is also reported (Shanmugam, 2011) that inhibition of CXCR4/CXCL12 signaling axis by ursolic acid leads to suppression of metastasis in transgenic adenocarcinoma of mouse prostate model and CXCR4 induced a more aggressive phenotype in prostate cancer (Miki, 2007). In another study, it is reported that CXCR4 and CXCR7 have critical roles on mediating tumor metastasis in various types of cancers as both being a receptor for an important α -chemokine, CXCL12 (Sun, 2010). Furthermore, a more recent study concluded that CXCR4 plays a crucial role in cancer proliferation, dissemination and invasion and the inhibition of CXCR4 strongly affects prostate cancer metastatic disease (Gravina, 2015). The chief officer of Massachusetts based X4 Pharmaceuticals company recently stated that CXCR4 protein “acts as a beacon to attract cells to surround a tumor, effectively hiding the tumor from the body’s T cells that would otherwise destroy them”. He indicated that X4 company is beginning human trials using CXCR4 inhibitors which aims to develop a therapy to block the protein, CXCR4 (<http://pharmaceuticalintelligence.com/2015/12/15/are-cxc4-antagonists-making-a-comeback-in-cancer-chemotherapy>, 2015).

The second most likely prediction was STK39 (serine threonine kinase 39). Among others, in (Hendriksen, 2006) it is reported that lower mRNA expression of STK39 in primary prostate tumors was directly associated with a higher possibility of metastases following radical prostatectomy. In (Balatoni, 2009), it is stated that STK39 encoded protein SPAK, regulates cell stress responses, and microarray studies identified reduced SPAK expression in treatment-resistant breast cancers and metastatic prostate cancers, suggesting that its loss may play a role in cancer progression. They showed that epigenetic silencing of STK39 in B-cell lymphoma inhibits apoptosis from genotoxic stress in cancer. STK39 is also identified as hypertension susceptibility gene (Wang, 2008).

ELOVL7 (fatty acid elongase 7) was reported that it could play an important role in prostate cancer cell growth and survival processes through the metabolism of SVLFAs. ELOVL7 is also suggested as a promising biomarker for development of new therapies or preventive methods for prostate cancers (Tamura, 2009).

ACSL3 (acyl-CoA synthetase long-chain family member 3) was reported to be one of the androgen-regulated genes and it is shown that ACSL3 is slightly up-regulated in primary prostate tumors and strongly repressed in metastatic cancer (Marques, 2011). It also states that ACSL3, ELOVL5 and GLUD1 play a role in the production of prostatic fluid and in secretory function of the prostate. From this literature information, it worth mentioning that we blindly predicted ACSL3, ELOVL7 and GLUD1 as in top eight tumor-specific hubs, which may suggest their collaborative role in this disease from this biological process. There is also a patent that reports that the fusion genes ACSL3 and ETV1 and their expression products can be used as prognostic and diagnostic markers for prostate cancer and as clinical targets for the treatment of prostate cancer (Attard, 2008).

In order to evaluate biological functions of the genes in the differential network, GO and KEGG pathway enrichment analyses were performed. The results showed that sterol biosynthetic process was the most significantly enriched GO term for biological process. To further evaluate the biological roles of the genes in the differential network, KEGG pathway analysis was performed. According to the KEGG analysis, Steroid biosynthesis was the most significant pathway ($p=2.80 \times 10^{-7}$). It contains ten genes in our network: TM7SF2, CEL, EBP, SQLE, LSS, SC5DL, DHCR24, FDFT1, SC4MOL and NSDHL. The relation of Steroid biosynthesis and prostate cancer is reported in many studies. The ligand activation of the androgen receptor plays an important role in the progress of castration-resistant prostate cancers. The similarities and differences from glandular androgen synthesis provide direction for the development of new treatments (Migita, 2009; Sharifi, 2012; Auchus, 2012; Ferraldeschi, 2013).

The pathway with the second highest significance was the synthesis and degradation of ketone bodies pathway ($p=1.63 \times 10^{-3}$), which contains five genes: HMGCS2, HMGCS1, ACAT2, ACAT1 and HMGCL. In the study conducted by Lin et. al. (2005), synthesis and degradation of ketone bodies pathway found as up-regulated pathway in androgen-

independent CL1 cells (model for late-stage prostate cancer) when compared to androgen-dependent LNCaP (model for early-stage prostate cancer) cells.

Additionally, the other significant pathways have also examined, and found that Amino sugar and nucleotide sugar metabolism (Priolo, 2014), p53 signaling pathway (Chappell, 2012; Gupta, 2012; Stegh, 2012), Butanoate metabolism (Stoss, 2008; Romanuik, 2010), Alanine, aspartate and glutamate metabolism (Priolo, 2014), and Axon guidance (Choi, 2014) pathways were shown to be associated with the prognosis of prostate cancer. In these pathways, the p53 signaling pathway plays a critical role in cancer's response to chemotherapy and tumor growth. Inactivation of the tumor suppressor gene p53 is widely observed in more than 50% of human cancers including prostate cancer. The disruption of the p53 signaling pathway is one of the vital turning point for the survival of advanced prostate cancer cells during therapies. By enabling DNA repair, it was observed that p53 blocks cancer progression by provoking transient or permanent growth arrest (Chappell, 2012; Gupta, 2012; Stegh, 2012). However, three pathways, Terpenoid backbone biosynthesis (hsa00900), Fructose and mannose metabolism (hsa00051) and Epithelial cell signaling in *Helicobacter pylori* infection (hsa05120), have not previously been related to prostate cancer.

KEGG analysis for the largest independent subnetwork revealed much more interesting results that may show that it has the most important role in the prostate cancer. Axon guidance (hsa04360) pathway, which was also found significantly enriched in the whole differential network, is known to have tumor suppressor genes and therefore related with tumor growth. Axon guidance molecules are validated as tumor suppressor in the breast cancer and show promise as breast cancer diagnostic markers as well as potential therapeutic targets (Mehlen, 2011; Harburg, 2011). In the study conducted by Choi, axon guidance pathway was shown to be involved in prostate cancer tumorigenesis (Choi, 2014). In addition, Savli et al. reported that axon guidance signaling pathway was the most significant down-regulated canonical pathway in prostate cancer (Savli, 2008). The second significantly enriched pathway was Fc gamma R-mediated phagocytosis (hsa04666). This pathway was found as the highest significant pathway in prostate cancer and have been referred as being involved in the pathological development of prostate cancer (Jia, 2012). In the literature, the pathway endocytosis

(hsa04144), was also found related with prostate cancer. The importance of understanding the regulation between signal transduction and endocytosis pathways, and also how the breakdown of this integrated regulation contributes to cancer development was emphasized (Bonaccorsi, 2007).

4.1.2. Inference of Lung Cancer Specific Differential Network

Lung cancer is one of the most leading cause of cancer related death with an estimated 224,390 new cases and 158,080 deaths in the United States in 2016 (American Cancer Society, 2016). Based on pathological features, lung cancer is divided into two main types; small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). Approximately, 80% of all lung cancer cases are diagnosed as NSCLC (Zakaria et al., 2015). The three major subtypes of NSCLC are adenocarcinoma, squamous cell carcinoma and large cell carcinomas (Bartucci et al., 2012). In this second preliminary analyse, NSCLC specific differential network was inferred using the datasets described at below.

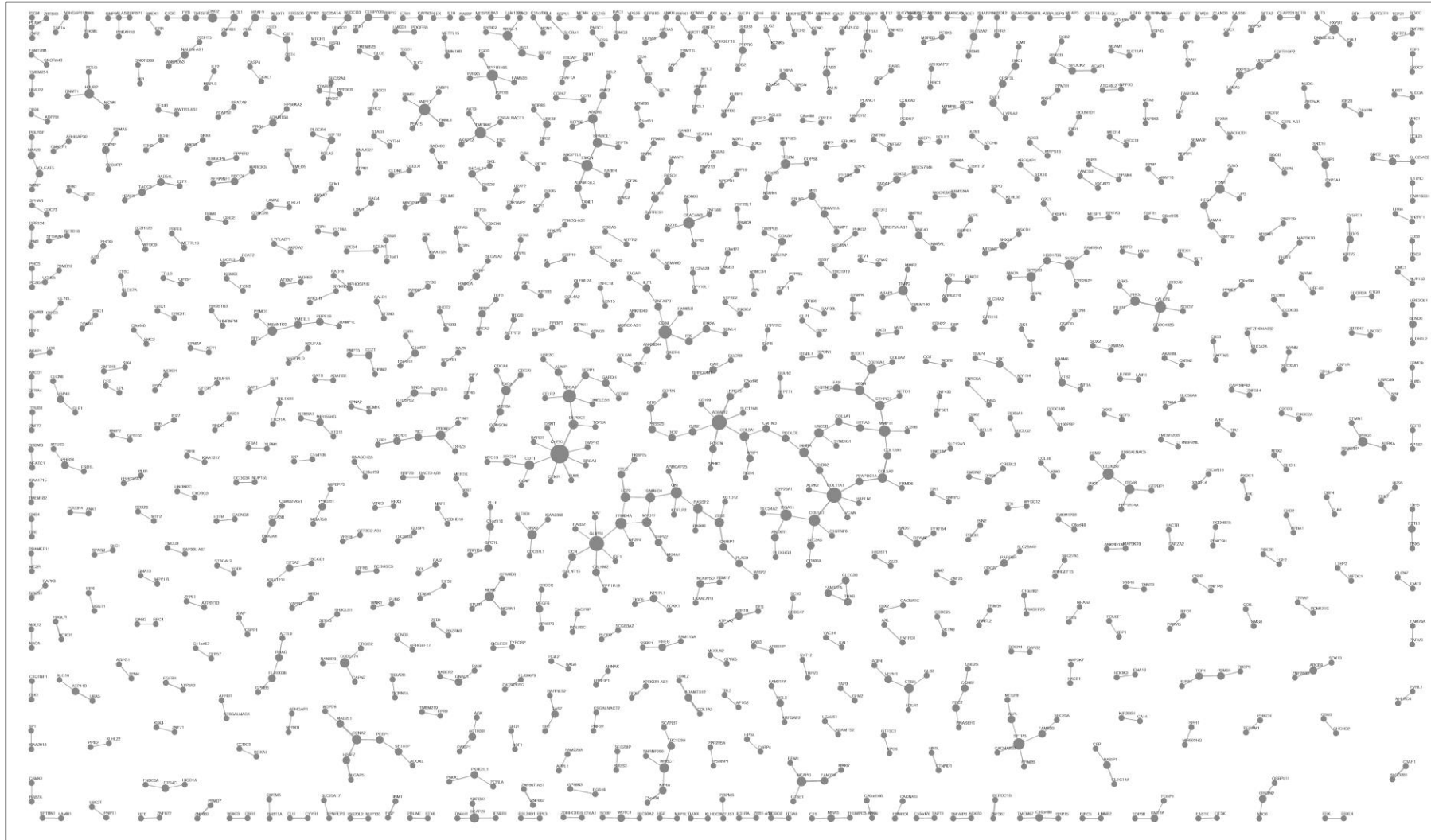
4.1.2.1. Microarray data

In order to investigate the alterations in tumor NSCLC cells compared with normal cells, microarray dataset GDS3837 deposited by Lu et al. (2010) was obtained from the Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo>). The expression profile included 120 samples, comprising 60 tissue samples with non-small cell lung cancer (NSCLC) and 60 adjacent normal lung tissue samples. The platform was GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array.

4.1.2.2. Results

The NSCLC specific differential gene network with 804 interactions was inferred. If we consider a gene that has more than five interactions as a hub gene, COL11A1, ADAM12, CHEK1 and GLIPR1 were identified as hub genes in the differential network. Additionally, the largest independent subnetwork with 50 interactions were extracted from the differential network and plotted in Figure 4.4. COL11A1 and ADAM12 genes which are the most important two hub genes in the main differential network were also detected in this subnetwork.

Figure 4.4. Genome-wide NSCLC specific differential network with 804 interactions



The top ten GO terms ranked by statistical significance were listed in Table 4.4. GO analysis revealed that genes associated with cell cycle phase (GO: 0022403; p=2.0E-08), M phase (GO: 0000279; p=2.2E-08) and cell cycle (GO: 0007049; p=4.0E-08) were significantly enriched top three GO terms among biological processes, while for molecular functions, enzyme binding (GO: 0019899; p=6.7E-06), purine ribonucleotide binding (GO: 0032555; p=8.2E-05) and ribonucleotide binding (GO: 0032553; p=8.2E-05) were significantly enriched, and with regards to cellular components, genes associated with collagen (GO: 0005581; p=2.0E-06), extracellular matrix (GO: 0031012; p=1.1E-05) and extracellular matrix part (GO: 0044420; p=1.4E-05) were significantly enriched (Table 4.4, Figure 4.5A).

Table 4.4. GO terms of non-small cell lung cancer specific differential network (top 10)

GO ID	GO term	No. of genes	p
Biological processes			
GO:0022403	cell cycle phase	62	2.0E-08
GO:0000279	M phase	53	2.2E-08
GO:0007049	cell cycle	96	4.0E-08
GO:0022402	cell cycle process	75	9.6E-08
GO:0000278	mitotic cell cycle	54	4.3E-07
GO:0006260	DNA replication	33	2.9E-06
GO:0000087	M phase of mitotic cell cycle	36	5.7E-06
GO:0030198	extracellular matrix organization	22	8.5E-06
GO:0000280	nuclear division	35	9.9E-06
GO:0007067	mitosis	35	9.9E-06
Cellular components			
GO:0005581	collagen	13	2.0E-06
GO:0031012	extracellular matrix	47	1.1E-05
GO:0044420	extracellular matrix part	23	1.4E-05
GO:0005578	proteinaceous extracellular matrix	44	1.8E-05
GO:0005583	fibrillar collagen	7	6.8E-05

GO:0005694	chromosome	55	7.9E-05
GO:0044421	extracellular region part	97	1.1E-04
GO:0031981	nuclear lumen	133	3.9E-04
GO:0000775	chromosome, centromeric region	20	8.1E-04
GO:0031974	membrane-enclosed lumen	162	8.4E-04
Molecular Function			
GO:0019899	enzyme binding	64	6.7E-06
GO:0032555	purine ribonucleotide binding	166	8.2E-05
GO:0032553	ribonucleotide binding	166	8.2E-05
GO:0001871	pattern binding	25	1.2E-04
GO:0030247	polysaccharide binding	25	1.2E-04
GO:0017076	purine nucleotide binding	170	1.8E-04
GO:0042802	identical protein binding	68	2.8E-04
GO:0032559	adenyl ribonucleotide binding	136	3.5E-04
GO:0042803	protein homodimerization activity	41	3.6E-04
GO:0005524	ATP binding	134	4.2E-04
GO: gene ontology.			

Next, the genes found in the NSCLC specific differential gene network were submitted to DAVID server to identify significantly enriched KEGG pathways (Kanehisa, 2000; Kanehisa, 2012). The top ten KEGG pathways that were found significantly enriched ($p < 0.05$) are shown in Table 4.5. Pathway analysis revealed that the genes in the NSCLC specific differential gene network were significantly enriched in sixteen terms. The most significant three terms were those involved in focal adhesion ($p = 6.8E-08$), ECM-receptor interaction ($p = 1.0E-05$), and T cell receptor signaling pathway ($p = 2.7E-03$) processes (Figure 4.5B).

Figure 4.5. Functional enrichment analysis of significantly enriched genes in the NSCLC specific differential gene network. (A) The top 10 enriched GO categories for biological processes; (B) The top 10 enriched KEGG pathways.

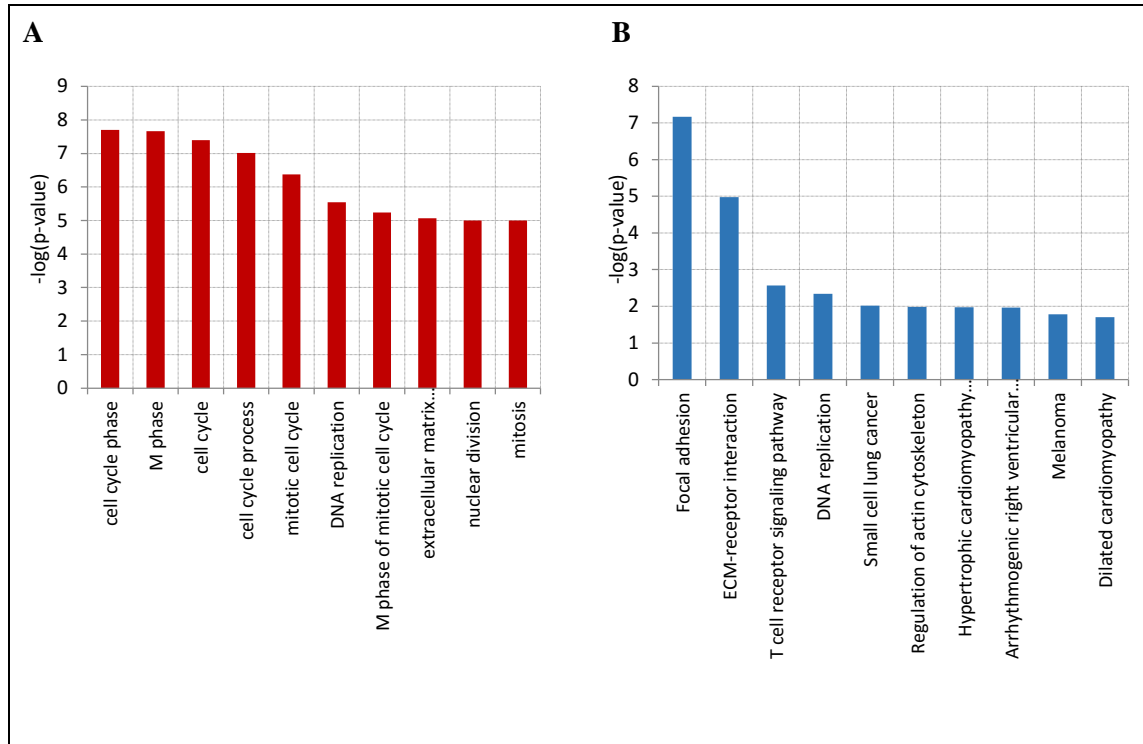


Table 4.5. Significant KEGG pathways in the NSCLC specific differential network

KEGG ID	KEGG term	No. of genes	p	Genes
hsa04510	Focal adhesion	39	6.8E-08	XIAP, COL3A1, ITGA11, ITGB5, ELK1, PAK2, BCL2, RAC1, COL6A3, PIK3CA, COL6A1, LAMB1, FIGF, COL11A1, RAPGEF1, THBS2, AKT3, PIK3R2, PARVG, COL4A2, TNXB, IGF1, RAF1, HGF, COL5A2, COL5A1, PRKCB, LAMA2, ITGA9, LAMA4, CCND3, FYN, LAMA5, ITGA8, COL1A2, PDGFRA, COL1A1, PARVB, MYLK
hsa04512	ECM-receptor interaction	20	1.0E-05	COL4A2, TNXB, COL3A1, ITGA11, ITGB5, COL5A2, COL5A1, HMMR, LAMA2, ITGA9, LAMA4, LAMA5, ITGA8, COL6A3, COL1A2, COL6A1, COL1A1, LAMB1, THBS2, COL11A1

hsa04660	T cell receptor signaling pathway	18	2.7E-03	IL4, PTPRC, ITK, CD247, RAF1, IL10, MAP3K7, PRKCQ, PAK2, FYN, NCK1, PPP3CB, PIK3CA, AKT3, CD28, NFATC1, LCP2, PIK3R2
hsa03030	DNA replication	9	4.5E-03	RFC4, SSBP1, RFC2, POLE3, RNASEH1, POLA2, RNASEH2A, MCM4, MCM6
hsa05222	Small cell lung cancer	14	9.6E-03	LAMA2, E2F2, COL4A2, LAMA4, XIAP, LAMA5, RXRB, BCL2, PIK3CA, CDK6, LAMB1, CDK2, AKT3, PIK3R2
hsa04810	Regulation of actin cytoskeleton	27	1.0E-02	GNA13, FGF6, FGFR1, FGFR4, MRAS, DIAPH3, ITGA11, IQGAP3, ITGB5, ABI2, TTLL3, PAK2, RAC1, PIK3CA, FGF2, FGD3, PIK3R2, ARHGEF6, RAF1, ARHGEF12, ITGA9, CHRM2, ITGA8, PDGFRA, MYH14, MYLK, CD14
hsa05410	Hypertrophic cardiomyopathy	14	1.1E-02	LAMA2, ITGA9, SLC8A1, DES, ATP2A2, CACNG8, ITGA8, ITGA11, SGCD, IGF1, ITGB5, CACNA1C, CACNA2D2, TPM4
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	13	1.1E-02	LAMA2, ITGA9, SLC8A1, DES, ATP2A2, CACNG8, ITGA8, ITGA11, SGCD, ITGB5, DSC2, CACNA1C, CACNA2D2
hsa05218	Melanoma	12	1.6E-02	FGF6, E2F2, FGFR1, PDGFRA, RAF1, IGF1, PIK3CA, CDK6, HGF, FGF2, AKT3, PIK3R2
hsa05414	Dilated cardiomyopathy	14	2.0E-02	LAMA2, ITGA9, SLC8A1, DES, ATP2A2, CACNG8, ITGA8, ITGA11, SGCD, IGF1, ITGB5, CACNA1C, CACNA2D2, TPM4
KEGG: Kyoto Encyclopedia of genes and genomes				

In order to investigate the biological roles of the genes in the independent subnetworks of the genome wide NSCLC specific differential gene network, we performed KEGG analysis for the largest subnetwork. As shown on Figure 4.6, this subnetwork comprises 50 interactions with COL11A1, and ADAM12 at the center of the largest hubs. KEGG analysis of the genes included in this subnetwork revealed a highly significant association with ECM-receptor interaction ($p=1.45E-08$) and focal adhesion ($p=2.68E-$

06) pathways, which were the top two significantly enriched pathways in the whole differential network.

Figure 4.6. The largest connected subnetwork of NSCLC specific differential gene network. This subnetwork might have an important role in human NSCLC as being the largest connected subnetwork with 50 edges in tumor differential network.

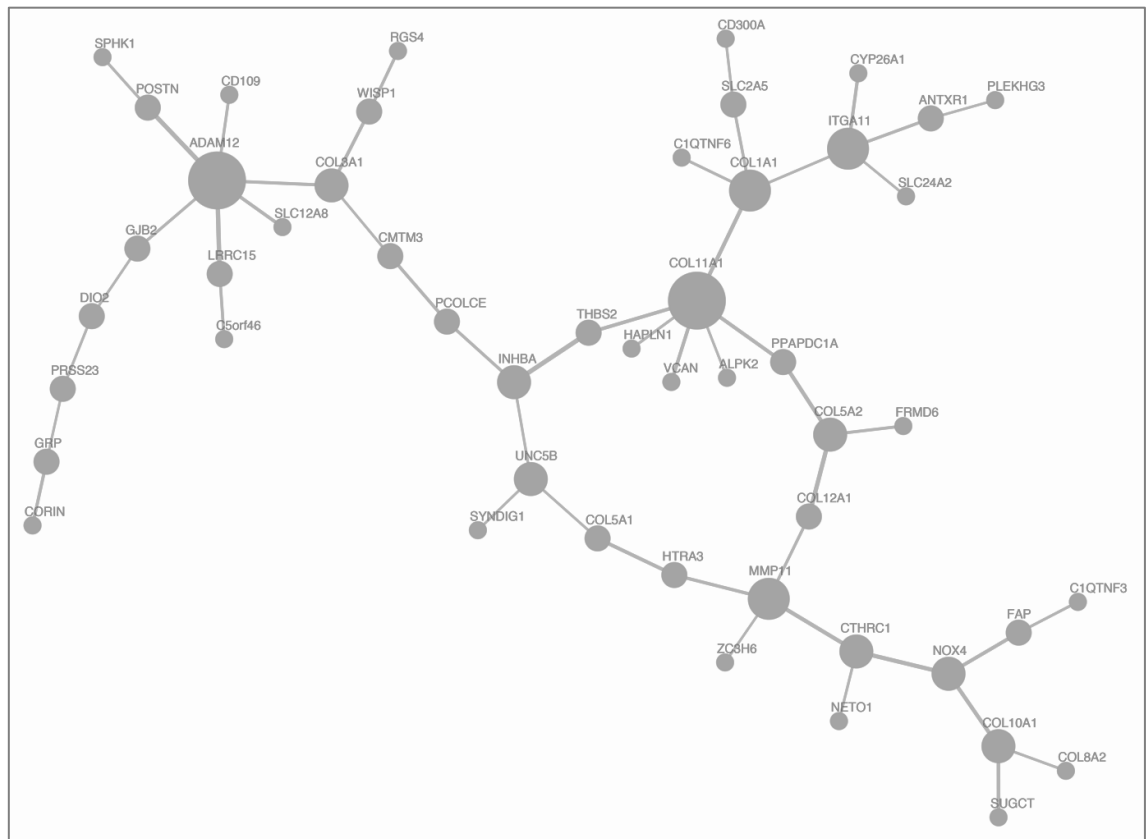


Table 4.6. Significant KEGG pathways in the largest subnetwork of the NSCLC specific differential network

KEGG ID	KEGG term	No. of genes	p	Genes
hsa04512	ECM-receptor interaction	7	1.45E-08	COL3A1, COL5A2, COL5A1, ITGA11, THBS2, COL1A1, COL11A1
hsa04510	Focal adhesion	7	2.68E-06	COL3A1, COL5A2, COL5A1, ITGA11, THBS2, COL1A1, COL11A1

KEGG: Kyoto Encyclopedia of genes and genomes

In line with previous studies, hub genes identified in the present trial have been closely associated with non-small cell lung cancer metastatic process, including COL11A1, ADAM12, CHEK1 and GLIPR1. Identification of hub genes involved in progression of NSCLC may lead to the development of better diagnostic methods and providing therapeutic approaches.

According to the analysis results, COL11A1 (Collagen, Type XI, Alpha 1) is by far the top hub gene in the NSCLC specific differential network and it is also part of the largest independent subnetwork as seen in Figure 4.6. In the literature, the COL11A1 is a widely known human gene that promotes tumor progression in many different human carcinomas (Zhang et al., 2016; Vázquez-Villa et al., 2015; Kleinert et al., 2015; Wu et al., 2014). Chong et al indicated that the overexpression of COL11A1 is highly correlated with lymph node metastasis and poor prognosis of NSCLC. Furthermore, COL11A1 and COL1A1 (encoding Collagen I alpha-1 chain protein) were found as upregulated differential expressed genes (Chong et al., 2006; Metodieva et al., 2011). This is a very interesting supporting information from literature for our blind estimation because in our predicted largest independent subnetwork, as shown in Figure 4.6, COL11A1 and COL1A1 appear to be interacting directly. Additionally, Lv and Wang strongly recommended that COL11A1 and COL1A1 may be potential targets in the treatment of smoking independent lung cancer (Lv and Wang, 2015). In another recent study, Tian et al indicated that COL11A1 might participate in the pathology of NSCLC (Tian et al., 2015). We also found that COL11A1 and COL1A1 were enriched in both ECM-receptor interaction and focal adhesion pathways which are important pathways for NSCLC.

The second most likely prediction was ADAM12 (ADAM Metallopeptidase Domain 12). Among others, Rocks et al demonstrated that production of ADAM12 is increased both at the protein and mRNA levels in human lung carcinomas. It is also suggested that overexpression of ADAM12 and a lower expression of ADAMTS-1 in NSCLC play important functions in lung cancer progression (Rocks et al, 2006). In another recent study, it is revealed that the expression levels of ADAM12 was significantly higher in small cell lung cancer (SCLC) than other ADAM genes. Furthermore, ADAM12 is

indicated as an independent prognostic factor that plays a crucial role in SCLC proliferation, invasion and metastasis (Shao et al., 2014).

CHEK1 (Checkpoint kinase 1) regulates cell cycle checkpoints and is known to be involved in the DNA damage repair (Liu et al., 2015). In tumor cells, CHK1 activation impairs the efficacy of many chemotherapeutic agents by inducing the S-phase checkpoint and by causing cell cycle arrest. Therefore, CHK1 inhibition improves survival of NSCLC patients by enhancing the therapeutic effects of chemotherapy (Bartucci et al., 2012; Fang et al., 2013). Inhibitors of CHK1 were suggested as new cancer treatment agents that can be useful in lung cancer (Syljuasen et al., 2015). Furthermore, Liu et al indicated that high expression levels of CHEK1 (Checkpoint kinase 1) in tumor tissues is significantly associated with the poor survival of NSCLC (Liu et al., 2015).

GLIPR1 (GLI pathogenesis-related 1) is known as a tumor suppressor gene which reduces cell growth and increases chemokine secretion (Ccl5) by activating immune cells (Zhang et al., 2015). In a very recent study, GLIPR1 (GLI pathogenesis-related 1) was identified as a potential therapeutic target for lung cancer by inhibiting lung cancer cell growth through suppressing ERBB3 (Erb-B2 Receptor Tyrosine Kinase 3) (Sheng et al., 2016). Taken together, these findings indicate that COL11A1, ADAM12, CHEK1 and GLIPR1 are closely associated with the progression of NSCLC.

In order to evaluate biological functions of the genes in the differential network, GO and KEGG pathway enrichment analyses were performed. The predominant enriched GO terms for biological processes, cellular components and molecular components included cell cycle phase, collagen and enzyme binding, respectively.

To further investigate the biological roles of the genes in the NSCLC specific differential network, KEGG pathway analysis was performed. KEGG analysis revealed that, ECM-receptor interaction was the most significant pathway ($p=1.45E-08$) for the largest independent subnetwork. It contains seven genes in our network: COL3A1, ITGA11, COL1A1, COL5A2, THBS2, COL11A1, COL5A1. The ECM mainly involves in tissue morphogenesis and plays a crucial role in the maintenance of cellular events such as adhesion, migration, differentiation and survival. The development and

progression of tumors is directly associated with the dysregulation of ECM (Zakaria et al., 2015, Zhang et al., 2015). The relation of ECM-receptor interaction and lung cancer has been reported in many studies. Hu and Chen indicated that ECM receptor interaction was significantly affected in lung tumor tissues and may contribute to early stages of lung adenocarcinoma irrelevant to smoking (Hu and Chen, 2015). In another study, it is demonstrated that ECM-receptor interaction pathway including COL11A1 and COL1A1 genes might be involved in lung cancer metastasis and angiogenesis (Lv and Wang, 2015).

The pathway with the second highest significance was the focal adhesion pathway ($p=2.68E-06$). The same seven genes, COL3A1, ITGA11, COL1A1, COL5A2, THBS2, COL11A1, COL5A1, were enriched in the focal adhesion pathway as in ECM-receptor interaction pathway. Focal adhesions are large protein complexes that physically connect the extracellular matrix to the cytoskeleton and regulate a number of biological processes such as cell migration, proliferation and survival. The perturbations in these processes can result in the development of malignancy and alteration of focal adhesion activity in cancer cells (McLean, 2005, Kim and Wirtz, 2013). Carelli et al indicated that the expression levels of focal adhesion kinase (FAK) are increased in tumor tissue as compared to normal lung tissue which refers to the crucial role of FAK in the progression of NSCLC (Carelli et al., 2006). Moreover, Webber et al suggested a combined treatment of Hsp90 and FAK that inhibits the growth of NSCLC cells (Webber et al., 2015). In a recent study (Howe et al., 2016), the major role of FAK in NSCLC growth and progression was emphasized and a new drug combination that targets EGFR along with FAK in NSCLC was suggested. Beside these, seven of top ten significant pathways of NSCLC specific differential network were found related with NSCLC including T cell receptor signaling (Kakimi et al., 2014), DNA replication (Zhang et al., 2014), small cell lung cancer, regulation of actin cytoskeleton (Gao et al., 2015) and melanoma (Huang et al., 2016) pathways.

4.2. BREAST CANCER ANALYSIS

4.2.1. Microarray Data

The METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) microarray dataset used in this dissertation is provided by the European Bioinformatics Institute (EBI). Upon access request and approval, the Metabric gene expression data is downloaded from the European Genome-Phenome Archive (EGA, <http://www.ebi.ac.uk/ega/>) webpage using special EGA download client.

The study accession number for the dataset is EGAS00000000083. The dataset accession IDs are shown in Table 4.7. It consists of transcriptomic information (cDNA microarrays profiling) measured using the Illumina HT-12 v3 platform, as indicated in (Curtis et al., 2012). In the normalized METABRIC gene expression dataset, *discovery* set with 997 samples and *normals* set with 144 samples were used respectively as *tumor* and *normal* datasets. Sample size over 64 is reported as sufficient to infer a gene network with maximum performance (Altay, 2012).

Table 4.7. METABRIC Breast Cancer Datasets

Dataset Accesion	Technology	Description	Samples	Type
EGAD00010000210	Illumina HT 12	Normalized expression data	997	Tumor
EGAD00010000212	Illumina HT 12	Normalized expression data	144	Normal

Source: European Genome-Phenome Archive (EGA, <https://www.ebi.ac.uk/ega/studies/EGAS00000000083>)

4.2.2. Data Preprocessing

Preprocessing is a crucial preliminary step in bioinformatics studies to prepare the dataset for the application of data analysis and to improve the performance of analysis results. In this dissertation, the following data pre-processing steps were performed:

The gene expression datasets, which was downloaded from EGA, were consisted of 48803 probes. Since the datasets were normalized before, the normalization process was

skipped. Firstly, the datasets were checked for missing values. There were no missing values in the datasets. Then probe annotation process was carried out using the *illuminaHumanv3.db* (Illumina Human HT-12 v3 annotation data) R software package (Dunning et al., 2016). This package includes all known gene symbols that correspond to probes in the Illumina HT-12 v3 platform. At the end of the annotation process, 29221 probes were annotated successfully. So the resulting datasets which were used in the rest of the analysis were consisted of 29221 probes.

Following the probe annotation process, copula transformation was performed for each of the datasets. It was reported that copula transformed datasets result in more stable estimations of the mutual information matrices. Then probe filtering operation was performed. In microarray technology, multiple probes can represent a single gene. In theory and mostly in practice those kind of probes have highest association score among them which cause an error for the inference algorithm c3net. In order to eliminate this problem, the association matrix was filtered by setting zero for the mutual information score for those probe pairs (Altay and Emmert-Streib, 2010a). Finally, mutual information (MI) matrices of each of the datasets were computed. The dimension of computed MI matrices was 29221 x 29221.

4.2.3. Data Integration

In this part, the datasets which were used in the integration part of IDN framework were described. The integrated data includes all known transcription factors, identified differential expressed genes from METABRIC breast cancer datasets, oncogenes, prognostic genes and metastatic genes associated with breast cancer.

4.2.3.1. Transcription Factors

Transcription factor is a molecule that controls the activity of a gene by determining whether the gene's DNA is transcribed into RNA. TFs are important regulators of cellular processes. A list of 1856 TFs were downloaded from TFCat (Transcription factor database) and adopted to the IDN framework (<http://www.tfcats.ca>, Access date: June, 2016).

4.2.3.2. Identification of Differentially Expressed Genes

In the genomic studies, differentially expressed (DE) genes are widely considered as potential candidates of biomarkers. The t-statistic is one of the frequently used method for identifying DE genes between two different biological conditions. (Abeel et al., 2010). In this dissertation, *limma* (Linear Models for Microarray Data) R software package was used for the analyze and identification of the differentially expressed genes between tumor and normal breast cancer cells (Ritchie et al., 2015). According to eBayes modified t-statistic analysis results, 20627 genes found to be differentially express with adjusted p value of 0.01. The top 20 differentially expressed genes are shown in the Table 4.8.

Table 4.8. Differentially Expressed Genes of Metabric Breast Cancer Dataset

Gene	logFC	<i>t</i>	P.Value	adj.P.Val	B
FXVD1	-1.78015726	-57.729493	0.00E+00	0.00E+00	770.744162
SDPR	-2.78229480	-55.706600	0.00E+00	0.00E+00	740.763584
CD300LG	-3.66903123	-55.066523	9.88e-324	9.62e-320	731.170398
ABCA9	-1.88347822	-53.967940	2.18e-316	1.59e-312	714.586282
SCARA5	-2.81934553	-53.350932	2.64e-312	0.00E-01	705.206328
LYVE1	-2.28281412	-52.200300	1.23E-304	5.98E-301	687.588957
CREB5	-1.41840590	-50.008824	7.66E-290	3.20E-286	653.591330
ABCA6	-2.19428576	-49.997155	9.20E-290	3.36E-286	653.408776
CLEC3B	-3.03235838	-49.965190	1.52E-289	4.93E-286	652.908595
CA4	-2.64780384	-48.918138	2.19E-282	6.39E-279	636.458718
RPL21P44	-0.84602167	-48.829489	8.88E-282	2.36E-278	635.060136
AQP7P1	-3.25714999	-47.843749	5.51E-275	1.24E-271	619.448092
RBPMS	-1.54516073	-47.791570	1.27E-274	2.64E-271	618.618624
ABCA8	-3.14024792	-47.233687	9.32E-271	1.81E-267	609.731198
LINC-PINT	-1.00181901	-45.880611	2.57E-261	4.41E-258	588.034583
FHL1	-3.55314298	-45.623645	1.63E-259	2.64E-256	583.892052
GPIHBP1	-2.48640897	-44.942500	1.01E-254	1.55E-251	572.878299
ARHGAP20	-1.15422218	-44.693572	5.75E-253	8.40E-250	568.841475
IGFBP6	-3.12818907	-44.562896	4.81E-252	6.70E-249	566.719834
KLHL29	-1.28006010	-44.519535	9.75E-252	1.29E-248	566.015451

4.2.3.3. Prognostic Genes

The identification of prognostic genes is a very crucial task in breast cancer since they are directly associated with the short survival times. Additionally, these genes help us to understand molecular mechanisms underlying the tumor progression. Hence, I adopted them into the IDN framework.

In a recent study conducted by Joe and Nam (2016), 26 high-expressed and 17 low-expressed genes are reported as most important prognostic factors in breast cancer. 16 genes of 26 co-expressed genes, and 8 genes of 17 low-expressed genes reported in this study were previously defined as prognostic genes in breast cancer (Table 4.9).

Table 4.9. The prognostic gene list in Breast Cancer

Type	Gene	Description
High-expressed genes	CHEK1	checkpoint kinase 1
	FOXM1	forkhead box M1
	CCNA2	cyclin A2
	CDC20	cell division cycle 20
	TTK	TTK protein kinase
	CENPA	centromere protein A
	KIF2C	kinesin family member 2C
	BUB1	BUB1, mitotic checkpoint serine/threonine kinase
	MCM6	minichromosome maintenance complex component 6
	LMNB2	lamin B2
	CDC45	cell division cycle 45
	ANLN	anillin actin binding protein
	MCM10	minichromosome maintenance 10 replication initiation factor
	CDCA8	cell division cycle associated 8
	MELK	maternal embryonic leucine zipper kinase
	CCNB2	cyclin B2
	CEP55	centrosomal protein 55 kDa
	DLGAP5	discs, large (Drosophila) homolog-associated protein 5
	HJURP	Holliday junction recognition protein
	CDCA5	cell division cycle associated 5
TRIP13	thyroid hormone receptor interactor 13	

Type	Gene	Description
High-expressed genes	GTSE1	G2 and S-phase expressed 1
	CDCA3	cell division cycle associated 3
	PRR11	proline rich 11
	FAM83D	family with sequence similarity 83 member D
	GTPBP4	GTP binding protein 4
Low-expressed genes	ESR1	estrogen receptor 1
	GATA3	GATA binding protein 3
	LRIG1	leucine-rich repeats & immunoglobulin-like domains 1
	RABEP1	rabaptin, RAB GTPase binding effector protein 1
	CIRBP	cold inducible RNA binding protein
	EVL	Enah/Vasp-like
	WDR19	WD repeat domain 19
	SCUBE2	signal peptide, CUB domain, EGF-like 2
	KIF13B	kinesin family member 13B
	TBC1D9	TBC1 domain family member 9
	ANKRA2	ankyrin repeat family A member 2
	DYNLRB2	dynein, light chain, roadblock-type 2
	NME5	NME/NM23 family member 5
	CAPN8	calpain 8
	CASC1	cancer susceptibility candidate 1
BBOF1	basal body orientation factor 1	
RUNDC1	RUN domain containing 1	

Source: “Joe, S. & Nam, H. 2016. Prognostic factor analysis for breast cancer using gene expression profiles. *BMC Med Inform Decis Mak.* **16**: 56”.

4.2.3.4. Oncogenes

Oncogenes are genes that are usually expressed at high levels in tumor cells and have great potential to cause cancer (Wilbur et al., 2009). For this reason, 518 known cancer genes were adopted into the IDN framework. The up-to-date list of oncogenes were downloaded from “Network of Cancer Genes” repository which is maintained by the Cancer Evolutionary Genomics at King's College of London (<http://ncg.kcl.ac.uk>, Access date: June, 2016).

4.2.3.5. Metastatic Genes

The metastatic genes associated with breast cancer are important promoters of metastasis in breast cancer. Fan et al. (2014) identified 61 genes as the most important metastatic genes in breast cancer. These genes which are adopted to the IDN framework are shown in (Table 4.10).

Table 4.10. Metastatic genes associated with breast cancer

Gene	Description	Gene	Description
ACTN1	actinin alpha 1	L3MBTL1	l(3)mbt-like 1
AKAP12	A-kinase anchoring protein 12	LMNA	lamin A/C
ANGPTL4	angiopoietin like 4	LPP	murein lipoprotein
BMP8B	bone morphogenetic protein 8b	MAPK7	mitogen-activated protein kinase 7
CALM1	calmodulin 1	MCAM	melanoma cell adhesion molecule
CAMK1D	calcium/calmodulin-dependent protein	MMP10	matrix metalloproteinase 10
CAV1	caveolin 1	MRC2	mannose receptor C type 2
CAV2	caveolin 2	MTMR9	myotubularin related protein 9
COL1A2	type I collagen gene	NDEL1	nudE neurodevelopment protein 1
COL5A3	collagen type V alpha 3 chain	NDUFS4	NADH:ubiquinone oxidoreductase subunit S4
C0X7A1	cytochrome c oxidase	N0X4	NADPH oxidase 4
CRISPLD2	cysteine rich secretory protein LCCL domain containing 2	PMP22	peripheral myelin protein 22
CRLP1	Cytokine Receptor Like Factor 1	PPL	periplakin
CTDSP2	CTD Small Phosphatase 2	RNASE2	ribonuclease A family member 2
CYR61	cysteine rich angiogenic inducer 61	S100A10	S100 calcium binding protein A10
DACT1	dishevelled binding antagonist of beta catenin 1	SERPINE 1	serpin family E member 1
DKK3	dickkopf WNT signaling pathway inhibitor 3	SPOCK1	SPARC/osteonectin, cwcv and kazal like domains proteoglycan 1
ECM1	extracellular matrix protein 1	SPP1	secreted phosphoprotein 1
EGR1	early growth response 1	SRPX2	sushi repeat containing protein, X-linked 2
EHD2	EH domain containing 2	STMN2	stathmin-like 2

ELF3	E74 like ETS transcription factor 3	TAC1	tachykinin precursor 1
EMP1	epithelial membrane protein 1	TAOK1	TAO kinase 1
EMX2	empty spiracles homeobox 2	TGFB1I1	transforming growth factor beta 1 induced transcript 1
FSTL3	follistatin like 3	THBS3	thrombospondin 3
GBE1	1.4-alpha-glucan branching enzyme 1	TLN2	talin 2
GLI1	GLI-Kruppel family member GLI1	TNC	tenascin C
HES1	hairy and enhancer of split 1	TPM1	tropomyosin 1
HOXD1	homeobox D1	TRIO	trio Rho guanine nucleotide exchange factor
HYAL1	hyaluronoglucosaminidase 1	WDR6	WD repeat domain 6
INSIG2	insulin induced gene 2	WFDC1	WAP four-disulfide core domain 1
ITGA5	integrin subunit alpha 5		

Source: “Fan, M., Sethuraman, A., Brown, M., Sun, W., & Pfeffer, L. M. 2014. Systematic analysis of metastasis-associated genes identifies miR-17-5p as a metastatic suppressor of basal-like breast cancer. *Breast Cancer Res Treat.* **146** (3), pp. 487-502”.

4.2.4. Inference of Breast Cancer Specific Differential Network

The breast cancer specific differential gene network with 1525 interactions was inferred and plotted in Figure 4.7. The differential network was included 2003 unique genes, 5 subnetworks with minimum 30 interactions and 5 hub genes with more than 10 targets.

In Fig. 4.7, colors are overlapped one on the other as following order. Nodes or labels are yellow if they are oncogenes. Triangles nodes represent transcription factors. Labels are red if they are significant in the first step of c3net. Labels are green if they are also differentially expressed with p-value 0.01. Edge widths vary with the correlation values.

Figure 4.7. Breast cancer specific differential network with 1525 interactions

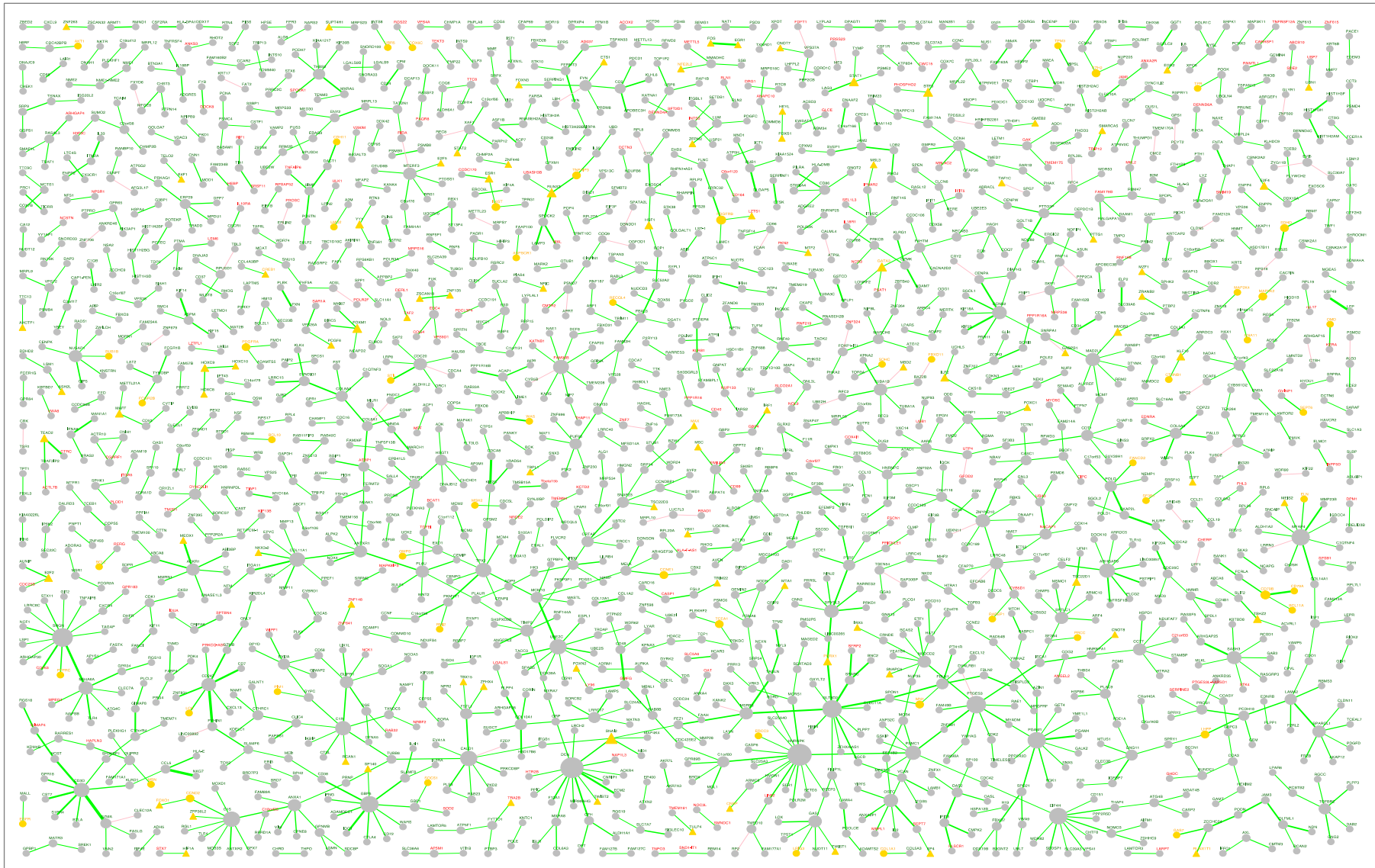
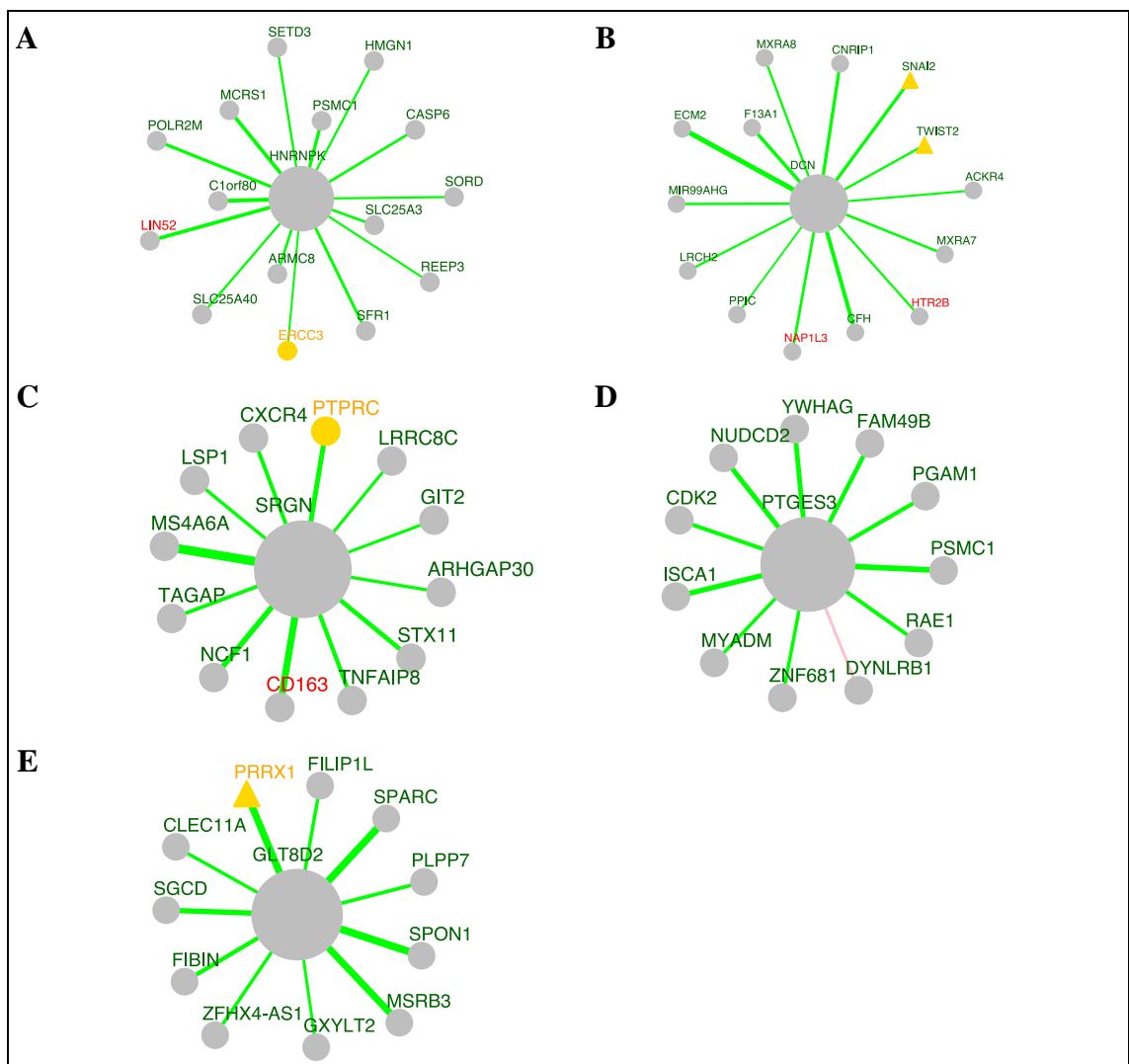


Table 4.11. Top five hub genes in the breast cancer specific differential network

Gene	Descripton	# of targets
HNRNPK	heterogeneous nuclear ribonucleoprotein K	15
DCN	decorin	14
SRGN	serglycin	12
PTGES3	prostaglandin E synthase 3	11
GLT8D2	glycosyltransferase 8 domain containing 2	11

HNRNPK, DCN, SRGN, PTGES3 and GLT8D2 genes were identified as the top five hub genes with more than 10 targets in the breast cancer specific differential network (Table 4.11). These genes and their targets are plotted in Figure 4.8.

Figure 4.8. Top five hub genes and their targets in the breast cancer specific differential network



To investigate the functions of the genes in the breast cancer specific differential gene network, GO and KEGG pathway analysis were performed. A total 699 GO terms AND 28 KEGG terms were retrieved from the DAVID. The top ten GO terms ranked by statistical significance were listed in Table 4.12. GO analysis revealed that genes associated with mitotic cell cycle (GO:0000278; $p=4.8E-08$), cell cycle (GO:0007049; $p=2.7E-15$) and cell cycle phase (GO:0022403; $p=2.9E-13$) were significantly enriched top three GO terms among biological processes, while for molecular functions, spindle (GO:0005819; $p=4.7E-09$), condensed (GO:0000793; $p=5.3E-09$) and chromosome (GO:0005694; $p=1.7E-08$) were significantly enriched, and with regards to cellular components, genes associated with ribonucleotide binding (GO:0032553; $p=3.4E-08$), purine ribonucleotide binding (GO:0032555; $p=3.4E-08$) and purine nucleotide binding (GO:0017076; $p=3.8E-08$) were significantly enriched (Table 4.12, Figure 4.9A).

Table 4.12. GO terms of breast cancer specific differential network (top 10)

GO ID	GO term	No. of genes	p
Biological processes			
GO:0000278	mitotic cell cycle	101	4.8E-16
GO:0007049	cell cycle	163	2.7E-15
GO:0022403	cell cycle phase	102	2.9E-13
GO:0022402	cell cycle process	125	5.5E-13
GO:0000279	M phase	85	4.5E-12
GO:0007067	mitosis	65	1.0E-11
GO:0000280	nuclear division	65	1.0E-11
GO:0048285	organelle fission	66	2.0E-11
GO:0000087	M phase of mitotic cell cycle	65	2.0E-11
GO:0051726	regulation of cell cycle	76	4.0E-08
Cellular components			
GO:0005819	spindle	46	4.7E-09
GO:0000793	condensed	42	5.3E-09
GO:0005694	chromosome	94	1.7E-08
GO:0043228	non-membrane-bounded	358	2.4E-08
GO:0043232	intracellular	358	2.4E-08
GO:0015630	microtubule	104	8.5E-08
GO:0044427	chromosomal	80	1.3E-07
GO:0000775	chromosome	37	2.4E-07

GO:0005654	nucleoplasm	144	8.9E-07
GO:0043233	organelle	256	2.0E-06
Molecular Function			
GO:0032553	ribonucleotide binding	253	3.4E-08
GO:0032555	purine ribonucleotide binding	253	3.4E-08
GO:0017076	purine nucleotide binding	262	3.8E-08
GO:0000166	nucleotide binding	298	5.8E-08
GO:0001883	purine nucleoside binding	213	7.0E-06
GO:0001882	nucleoside binding	214	7.8E-06
GO:0005201	extracellular matrix structural constituent	24	8.7E-06
GO:0005524	ATP binding	196	2.2E-05
GO:0032559	adenyl ribonucleotide binding	198	2.3E-05
GO:0030554	adenyl nucleotide binding	207	2.4E-05
GO: gene ontology.			

Next, the genes found in the breast cancer differential gene network were submitted to DAVID server to identify significantly enriched KEGG pathways (Kanehisa, 2000; Kanehisa, 2012). The KEGG pathways that were found significantly enriched ($p < 0.05$) are shown in Table 4.13. Pathway analysis revealed that the genes in the breast cancer specific differential network were significantly enriched in eighteen terms. The most significant three terms were those involved in cell cycle ($p = 5.4E-08$), DNA replication ($p = 9.6E-06$), and oocyte meiosis ($p = 1.2E-04$) processes (Figure 4.9B).

Figure 4.9. Functional enrichment analysis of significantly enriched genes in the breast cancer specific differential gene network. (A) The top 10 enriched GO categories for biological processes; (B) The top 10 enriched KEGG pathways.

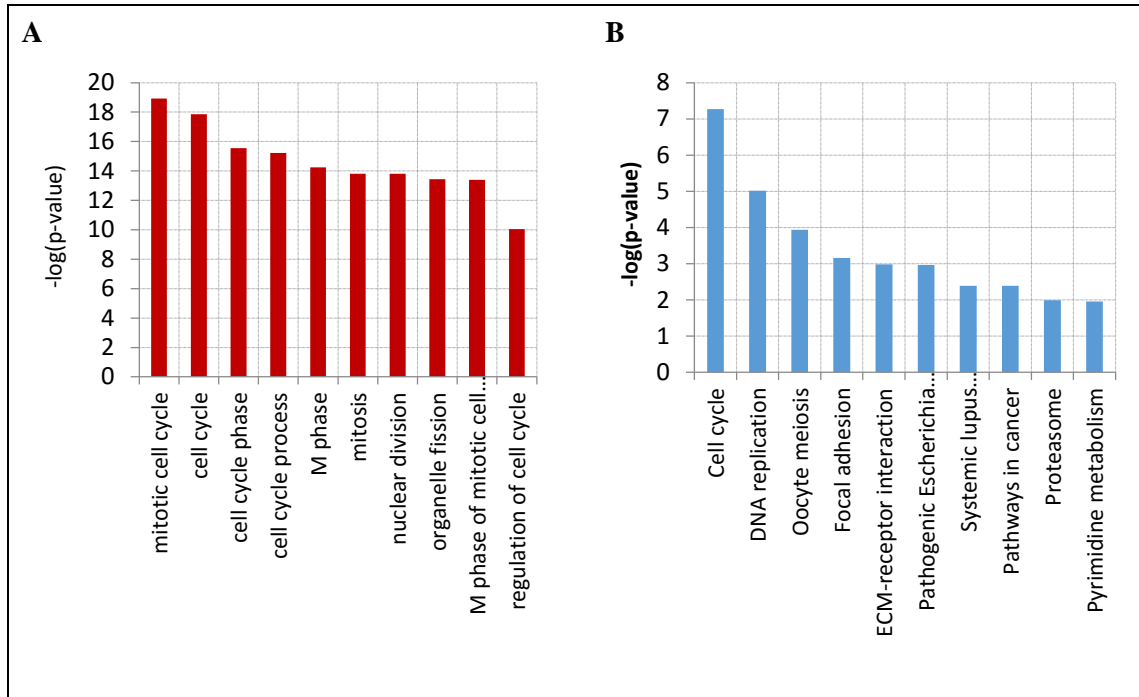


Table 4.13. Significant KEGG pathways in the breast cancer specific differential network (Top 10)

KEGG ID	KEGG term	No. of genes	p	Genes
hsa04110	Cell cycle	39	5.4E-08	E2F2, YWHAZ, E2F4, E2F5, TGFB3, PRKDC, TTK, PKMYT1, CHEK1, ANAPC10, CDC16, PTTG1, RBX1, CCNE2, CCNE1, MCM7, BUB1, CCNA2, CDK1, CCNH, SKP2, CDC20, ESPL1, MCM2, SKP1, CDC25C, MCM3, MCM4, CDK2, CDC25B, CCNB1, YWHAG, MAD2L1, CCNB2, HDAC2, CCND2, PLK1, PCNA, BUB1B
hsa03030	DNA replication	16	9.6E-06	SSBP1, POLE, MCM2, MCM3, RNASEH2A, MCM4, RNASEH2B, RFC3, RFC4, MCM7, POLE2, RFC2, POLD1, PRIM2, PCNA, FEN1
hsa04114	Oocyte meiosis	29	1.2E-04	YWHAZ, PPP2R5D, PKMYT1, AURKA, ANAPC10, CDC16, PTTG1, RBX1, CCNE2, CCNE1, IGF1R, PPP2CA, PPP2CB, BUB1, FBXO5, CDK1, SGOL1, IGF1, ESPL1, CDC20, SKP1, CDC25C, CDK2, CCNB1, REC8, YWHAG, MAD2L1, CCNB2, PLK1

hsa04510	Focal adhesion	42	6.9E-04	CAV2, ITGA11, PIP5K1C, ITGB5, ITGB1, MYL9, CTNNB1, AKT1, CDC42, IGF1R, PTK2, COMP, BCL2, COL6A3, COL6A2, RHOA, PDGFC, PDGFD, LAMB1, COL11A1, THBS2, THBS4, SPP1, EGFR, IGF1, FLNC, COL5A3, COL5A2, COL5A1, PRKCB, LAMA2, LAMA4, CCND2, FYN, ITGA5, COL1A2, PDGFRA, PDGFRB, RAPIB, LAMC1, COL1A1, CRK
hsa04512	ECM-receptor interaction	22	1.1E-03	ITGA11, ITGB5, COL5A3, COL5A2, ITGB1, COL5A1, HMMR, LAMA2, SDC1, LAMA4, ITGA5, COMP, COL6A3, COL1A2, COL6A2, LAMC1, COL1A1, LAMB1, THBS2, COL11A1, THBS4, SPP1
hsa05130	Pathogenic Escherichia coli infection	17	1.1E-03	YWHAZ, LY96, ARPC4, TLR4, ITGB1, WAS, CTNNB1, CDC42, CTTN, FYN, NCK1, RHOA, TUBA3C, TUBB6, TUBA3D, TUBA3E, TUBA1A, TUBA1B
hsa05322	Systemic lupus erythematosus	23	4.0E-03	C7, C3, C1R, C1S, HLA-DMB, HIST2H2AB, HIST2H2AC, IFNG, HIST3H2A, HIST3H2BB, HIST1H2BF, HIST1H2BG, SSB, CD40, HLA-DQA1, C1QB, CD86, FCGR2B, HLA-DPA1, HIST1H3D, FCGR2A, H3F3C, HIST1H3F, HIST1H2AM, CTSG, HIST1H3H
hsa05200	Pathways in cancer	58	4.1E-03	E2F2, TGFB3, FOXO1, FASLG, FLT3LG, CTNNB1, WNT2, AKT1, CCNE2, CDC42, FOS, MAX, CCNE1, RHOA, TPR, CSF2RA, EGFR, CTBP1, RUNX1T1, SKP2, STK4, CDK2, RAD51, PRKCB, HIF1A, PIAS4, PDGFRA, PDGFRB, LAMC1, TRAF1, CKS1B, BCL2L1, ITGB1, TPM3, RBX1, IGF1R, PTK2, BCL2, LAMB1, CSF1R, IL6, MSH2, MAP2K2, TGFB2, IGF1, BIRC5, STAT1, FZD7, LAMA2, RASSF5, LAMA4, HDAC2, PLCG1, ETS1, PLCG2, TCEB2, TCEB1, CRK
hsa03050	Proteasome	13	1.0E-02	PSMB8, PSMB9, PSMB4, PSMC6, PSMA6, PSMC4, PSME2, IFNG, PSMC1, PSMD2, PSMD4, PSMD6, PSMD7
hsa00240	Pyrimidine metabolism	21	1.1E-02	POLR2F, POLR3K, PNPT1, POLE, DTYMK, POLR1C, POLR3A, RRM2B, CMPK1, NME7, CMPK2, TYMP, NME2, NME3, POLE2, NME1-NME2, NME1, RRM2, POLD1, PRIM2, TXNRD1, UCK2, DPYD
KEGG: Kyoto Encyclopedia of genes and genomes				

In order to further evaluate the biological roles of the genes in the independent subnetworks of the genome-wide breast cancer specific differential gene network, we performed KEGG analysis for the largest subnetwork with 200 interactions. As shown on Figure 4.10, this subnetwork comprises 200 interactions with DCN and GLT8D2 at the center of the largest hubs. KEGG analysis of the genes included in this subnetwork revealed a highly significant association with ECM-receptor interaction ($p=4.4\text{-E}6$), focal adhesion ($p=4.4\text{E-}4$), complement and coagulation cascades ($p=1.1\text{-E}3$), hypertrophic cardiomyopathy (HCM) ($p=1.6\text{-E}2$) and dilated cardiomyopathy ($p=2.1\text{-E}2$) pathways.

Figure 4.10. Largest subnetwork of breast cancer specific differential network with 200 interactions

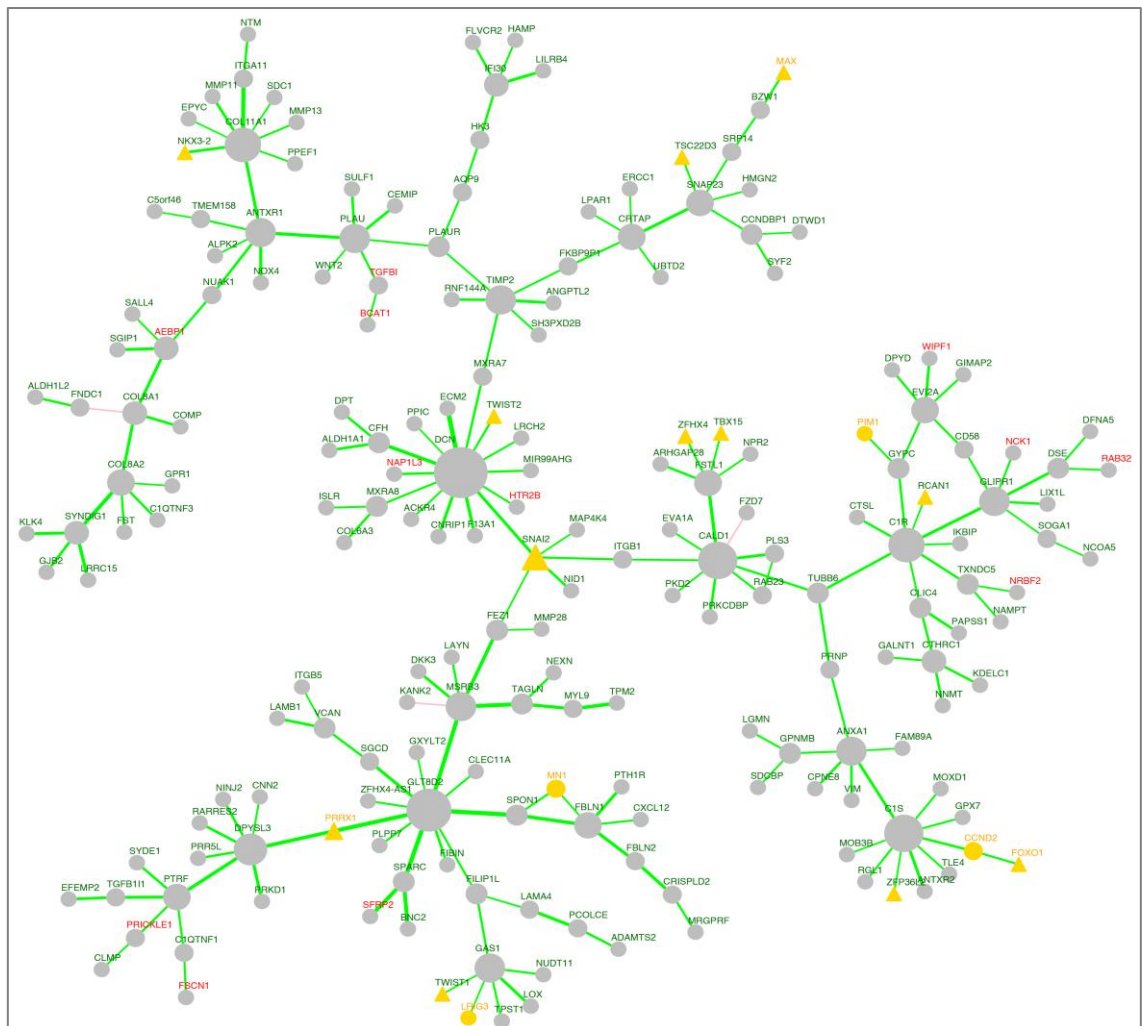


Table 4.14. Significant KEGG pathways in the largest independent subnetwork of breast cancer specific differential network

KEGG ID	KEGG term	No. of genes	p	Genes
hsa04512	ECM-receptor interaction	9	4.4E-6	LAMA4, SDC1, COMP, COL6A3, ITGA11, ITGB5, LAMB1, COL11A1, ITGB1
hsa04510	Focal adhesion	10	4.4E-4	LAMA4, CCND2, COMP, COL6A3, ITGA11, ITGB5, LAMB1, COL11A1, ITGB1, MYL9
hsa04610	Complement and coagulation cascades	6	1.1E-3	F13A1, CFH, C1R, C1S, PLAU, PLAUR
hsa05410	Hypertrophic cardiomyopathy (HCM)	5	1.6E-2	ITGA11, SGCD, ITGB5, TPM2, ITGB1
hsa05414	Dilated cardiomyopathy	5	2.1E-2	ITGA11, SGCD, ITGB5, TPM2, ITGB1
KEGG: Kyoto Encyclopedia of genes and genomes				

4.2.5. Integration and ranking process

In the integration step of IDN, the interactions out of the first step of *c3net*, are compared with the interaction database of literature including 860919 known gene-to-gene interactions and the overlapping ones are added to the breast cancer specific differential gene network. The resulting network had 5359 unique interactions. Then other information including 518 oncogenes, 43 prognostic genes and 61 metastatic genes associated with breast cancer are integrated over this network.

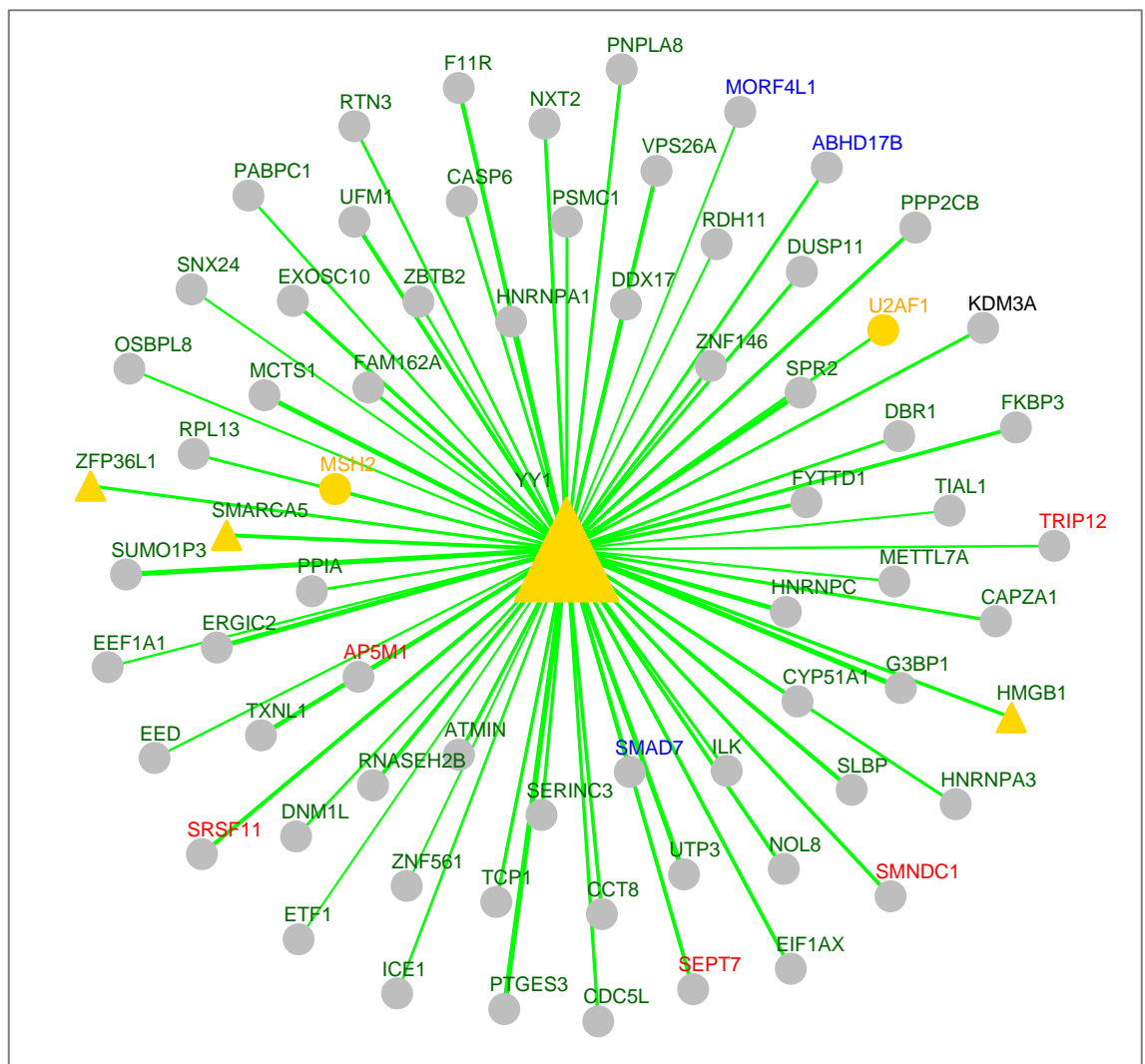
Since the main purpose of the algorithm is to find the most important regulators that drives the breast cancer, genes in the differential network were compared with the TF database of literature including 1856 TFs and scoring process were performed for the overlapping ones. In this step, 138 TFs that may have a crucial role in breast cancer were identified. Then, scoring process for each TF was performed according to the novel scoring formula of IDN. While scoring, in order to consider the neighbor density with respect to the proximity of each TF, the link distance from each TF was limited to 2 links.

Table 4.15. The top 20 genes identified by IDN framework that may have a crucial role in the progression of breast cancer

Gene	Descripton	# of links	Score
YY1	Yin yang 1	483	877.782
SMARCA5	SWI-SNF-related Matrix-associated Actin dependent Regulator of Chromatin A5	262	465.428
FOXM1	forkhead box M1	211	383.247
STAT4	signal transducer and activator of transcription 4	168	316.915
PTTG1	pituitary tumor-transforming 1	127	274.232
EOMES	eomesodermin	143	268.436
CNBP	CCHC-type zinc finger nucleic acid binding protein	134	256.344
MAX	MYC associated factor X	98	208.596
GTF2I	general transcription factor Iii	104	206.935
HMGB1	high mobility group box 1	88	189.925
HCLS1	hematopoietic cell-specific Lyn substrate 1	84	176.351
ZFP36L1	ZFP36 ring finger protein like 1	71	155.676
SNAI2	snail family transcriptional repressor 2	61	139.141
STAT1	signal transducer and activator of transcription 1	61	127.109
IRF9	interferon regulatory factor 9	70	126.154
SP140	SP140 nuclear body protein	53	125.177
IRF1	interferon regulatory factor 1	60	122.448
E2F2	E2F transcription factor 2	44	113.259
TGIF1	TGFB induced factor homeobox 1	44	102.821
HMGB2	high mobility group box 2	45	100.533

The top 20 genes that may have a crucial role in the prognosis of breast cancer is shown in Table 4.15. According to the analysis results, YY1 (YY1 transcription factor) was found to be the most important gene in breast cancer with the highest score of 2309.816. The score was computed according to 1407 links in 2 step neighborhood distance of the gene. SMARCA5, FOXM1, STAT4 and PTTG1 genes were identified as the other important genes following YY1 in the ranking table.

Figure 4.11. Targets of the transcription factor YY1 (69 interactions)



Furthermore, we checked oncogene, metastatic gene and prognostic gene lists associated with human breast cancer to verify the existence of these top 20 genes in these lists. According to this check, CNBP and MAX genes were found in oncogene list, and FOXM1 gene was found in prognostic gene list associated with breast cancer. The gene with the highest score, YY1, and the remaining six genes of top twenty were not found in these lists (Table 4.16).

Table 4.16. The existence status of top 20 genes identified by IDN framework in the known breast cancer lists

Gene	Oncogene List	Metastatic Gene List	Prognostic Gene List
YY1	No	No	No
SMARCA5	No	No	No
FOXM1	No	No	Yes
STAT4	No	No	No
PTTG1	No	No	No
EOMES	No	No	No
CNBP	Yes	No	No
MAX	Yes	No	No
GTF2I	No	No	No
HMGB1	No	No	No
HCLS1	No	No	No
ZFP36L1	No	No	No
SNAI2	No	No	No
STAT1	No	No	No
IRF9	No	No	No
SP140	No	No	No
IRF1	No	No	No
E2F2	No	No	No
TGIF1	No	No	No
HMGB2	No	No	No

5. DISCUSSION

The aim of this dissertation is to reveal mechanisms that drives disease progression in breast cancer which is the most common type of cancer in women. The IDN (integrative differential network) approach proposed in this dissertation differs significantly from the existing differential networking approaches since it combines prior knowledge in the literature with differential networking methodology in a novel way.

According to analysis results, the genes and pathways identified by the *IDN* framework were found highly associated with breast cancer. In the study, YY1, SMARCA5, FOXM1, STAT4 and PTTG1 genes were found as the most important genes in breast cancer. Accordingly, all of these blind predicted genes were found significantly associated with human breast cancer in the literature. This verifies the success of blind (unsupervised) prediction of *IDN* approach which can be easily extended to apply to the other disease datasets.

Yin yang 1 (YY1), which was found to be the most important gene in breast cancer, is a TF involved in the maintenance and initiation of DNA methylation (Qi et al. 2015). In a study conducted by Wan et al. (2012), it was observed that Yin yang 1 (YY1) is over expressed in breast cancer cells and plays a crucial role in breast cancer prognosis by regulating tumorigenesis through multiple pathways. Furthermore, it was also reported that the cooperation between YY1 and activator protein 2 results in the stimulation of expression of ERBB2 (Her2/neu) which is an oncogene highly expressed in 30% of breast cancers and mostly correlated with a malignant prognosis. In another study, the association between YY1 and Annexin A6 (AnxA6), which has multiple functions in breast cancer such as promoting the invasiveness of breast cancer cells and tumor growth, was emphasized (Qi et al. 2015).

In a study by Wang et al. (2015), it was identified that YY1 binds to the Flap endonuclease 1 (FEN1) gene which is up-regulated in breast cancer cells and suppresses the expression level of this gene. In another study that compares the GRNs of benign and malicious breast cancer samples with normal samples, YY1 was screened as a hub

gene that plays a crucial role in the whole process of breast cancer (Chen & Yang 2014). In a study by Lee et al. (2012), a strong correlation between the level of YY1 and breast cancer-associated gene 1 (BRCA1) was demonstrated. In the same study, YY1 was also reported as an important regulator of BRCA1 expression that is directly connected to the molecular etiology of breast cancer. Another recent study demonstrated that depletion of YY1 results in the inhibition of the migration, invasion, clonogenicity, and tumor formation of breast cancer tissues (Wang et al. 2016).

The second highest score is found for SWI-SNF-related Matrix-associated Actin Dependent Regulator of Chromatin A5 (SMARCA5). Hill et al. (2015) was reported that high expression level of SMARCA5, which is involved in cell proliferation and stem cell self-renewal, is directly associated with significantly short survival time, compared with those with low expression. In another recent study, it was found that SMARCA5 was highly expressed in human breast cancer cells and it was significantly associated with high proliferation, invasion, tumor size and poor survival of patients. Moreover, in a study by Chen et al. (2014), miR-100 was identified as a tumor suppressor and epithelial-mesenchymal transition (EMT) inducer by targeting SMARCA5 in breast cancer.

Forkhead box M1 (FOXM1) is a transcription factor which controls apoptosis and cell proliferation and it is increased in invasive breast cancer cells. It was reported that FOXM1 is directly associated with poor survival of breast cancer patients (Ferrer et al. 2016; Yuan & Wang 2015). In a study conducted by Hamurcu et al. (2016), it was found that FOXM1 is highly upregulated in triple negative breast cancer (TNBC) cells and the knockdown of FOXM1 by RNA interference (siRNA) results in the inhibition of eEF2K expression which is an emerging molecular target in cancer treatment that promotes to cancer proliferation, migration, tumorigenesis, disease progression and drug resistance. It was also found that high levels of FOXM1 was significantly correlated with the TNBC (Lee et al. 2016).

In another study which investigates the role of FOXM1 in breast cancer progression, it was reported that low expression levels of FOXM1 are significantly correlated with better survival in patients with estrogen receptor positive (ER+) tumors. In the study, also the critical role of FOXM1 in the development of resistance to breast cancer

therapies was emphasized (Saba et al. 2016). FOXM1 was also reported to be generally upregulated and knockdown of FOXM1 causes the inhibition of cell migration (Ye et al. 2015).

The transcription factor STAT4 (signal transducer and activator of transcription 4) was reported as playing an important role in human breast cancer physiology by regulating S100A4 mediated by HBXIP. (Liu et al., 2012; Wang et al. 2015). Kristensen et al. (2012) was identified STAT4 as the most significant difference between low and high mammographic density in healthy breast cells. In a recent study, the antitumor role of Cryptotanshinone by regulating cytotoxic CD4+ T cells through STAT4 in breast cancer was demonstrated (Li et al. 2016).

PTTG1 (pituitary tumor-transforming 1) is a gene that is over-expressed in several type of tumors and has been highly associated with tumor invasiveness and poor prognosis. Recent studies in the literature have revealed that inhibition of PTTG1 results in the suppress of tumor growth and metastasis in breast cancer (Huang et al. 2014). In a study by Han & Poon (2013), it was reported that PTTG1 increases the invasive characteristics of breast cancer cells by inducing epithelial to mesenchymal transition and progression of the cancer stem cell population. In another recent study, PTTG1 was identified as a direct target of miR-300, miR-329, miR-381 and miR-655, which may have an oncogenic role in miRNA-induced pituitary tumor progress inhibition. Additionally, PTTG1 protein levels were found as down-regulated in human breast cells and the reduction level was found highly associated with the tumor grade (Liang et al. 2015). Wang et al. (2015) reported that the over-expressed PTTG1 mRNA in tumors induce tumor recurrence and metastasis. Moreover, expression of PTTG1 in early phase of breast cancer was suggested as an invasion biomarker.

Moreover, we performed enrichment analysis for the breast cancer specific differential network. The enrichment analyses are important since they improve disease classification and reveals novel insights about a disease (Myers et al., 2015). The cell cycle pathway, which is recently suggested in the literature as an important pathway in breast cancer, was found as the most significant pathway in breast cancer specific differential network ($p < 5.4E-08$).

In a study by Wu et al. (2016), the underlying molecular pathways responsible for breast cancer was investigated. The authors reported that four pathways, cell-cycle, progesterone-mediated oocyte maturation, oocyte meiosis and p53 signalling pathways, are significantly associated with the pathogenesis of breast cancer. In another recent study, it was reported that the abnormalities in cell cycle were seen more common on the patients with metaplastic breast cancer which is a rare subtype of breast cancer. In the same study, aberrant cell-cycle pathway is suggested as an potential therapeutic target for breast cancer (Helsten et al. 2016). In another study by Zhang et al. (2016), the inhibitory role of miR-29a (microRNA-29a) was investigated by examining its role in cell cycle progression in breast cancer. The results of the study revealed that miR-29a has a growth-inhibiting function in breast cancer cells through cell cycle regulation. Zhuang et al. (2015) conducted a study to identify hub subnetwork in breast cancer using topological features of genes. In the study, the cell cycle, cluster1 and oocyte meiosis pathways were detected as the most significant subnetworks in breast cancer and hub subnetwork was constructed using the intersection of the genes involved in these three pathways.

Following the “Cell cycle” pathway, KEGG pathway analysis results showed a high enrichment of “DNA replication” ($p < 9.6E-06$), “Oocyte meiosis” ($p < 1.2E-04$), “Focal adhesion” ($p < 6.9E-04$), “ECM-receptor interaction” ($p < 1.1E-03$), “Pathogenic Escherichia coli infection” ($p < 1.1E-03$), “Systemic lupus erythematosus” ($p < 4.0E-03$), “Pathways in cancer” ($p < 4.1E-03$), “Proteasome” ($p < 1.0E-02$) and “Pyrimidine metabolism” ($p < 1.1E-02$) pathways, which have all been related to breast cancer except “Systemic lupus erythematosus” pathway. Considering the success of the comparison of the results with the literature, “Systemic lupus erythematosus” pathway may be a potential target of breast cancer.

Findings from this dissertation suggest that IDN framework has a great potential in identifying disease specific differential networks as well as gene targets in human breast cancer. Furthermore, IDN approach can be easily extended for other diseases as well by replacing the datasets accordingly.

However, this dissertation had some limitations. Previous research had reported that breast cancer has five main subtypes including luminal subtype A, luminal subtype B, ERBB2, basal-like and normal breast-like. In the present dissertation, the METABRIC tumor dataset, which includes samples of these subtypes are considered as a whole tumor dataset. Therefore, the breast cancer specific differential network, pathways and the genes identified in this study may not be differential for all subtypes.

Second limitation is about the integrated datasets. In this dissertation, the oncogene, metastatic gene and prognostic gene lists obtained from the literature for breast cancer may not cover all of the genes identified in these categories. For this reason, missing genes in these lists may affect the results.

The present study provided significant insight into the molecular mechanisms associated with breast cancer. Furthermore, GO and KEGG pathway enrichment analysis identified numerous pathways that may have a role in the breast cancer, and these findings may promote the better understanding about the molecular mechanism of this disease and also disclose potential targets for diagnostic and effective therapies.

The strongest prediction as breast cancer specific hub gene has experimental validations from the literature that reports YY1 as metastatic-level cancer indicator gene. This verifies the blind (unsupervised) prediction of IDN approach and assures the use of IDN in the other datasets. Moreover, some of our estimations in the breast cancer specific differential network may well be biomarkers or drug targets for breast cancer and awaits biologist to perform wet-lab experiments on them.

Our application not only elucidates a genome-wide tumor-specific interaction network of breast cancer but stands as a successful example for the application of other cancer types. Considering these strong confirmations, our inferred differential network might reveal the core mechanism of breast cancer.

In this dissertation, the main application of the work was on gene expression data. For future works, the algorithm may be extended for the use on the other data types such as ChIP-seq, motif data, methylation and etc.

Additionally, IDN approach can be applied to each subtype of breast cancer to identify breast cancer subtype specific differential results. Moreover, IDN approach can be extended to be applied on different kind of diseases by changing the integrated datasets.

REFERENCES

Books

- Cover, T. & Thomas, J. 1991. *Information Theory*. New York: John Wiley & Sons
- Emmert-Streib, F. & Dehmer, M. 2010. *Medical Biostatistics for Complex Diseases*.
Weinheim: Wiley-Blackwell.
- Gallager R. 1968. *Information Theory and Reliable Communication*. New York: Wiley.
- Klipp, E., Herwig, R., Kowald, H., Wierling, C., & Lehrach, H. 2005. *Systems biology
inpractice: concepts, implementation, and application*. Weinheim: Wiley-VCH
Verlag GmbH & Co. KGaA.
- Shannon, C. & Weaver, W. 1949. *The Mathematical Theory of Communication*.
University of Illinois Press.
- Smyth, G. K. 2005. Limma: linearmodels for microarray data. *Bioinformatics and
Computational Biology Solutions using R and Bioconductor*, pp. 397–420
Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., & Huber, W. New York:
Springer.
- Wilbur, B. et al. 2009. *The World of the Cell*. 7th ed. San Francisco, CA.

Periodicals

- Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. 2010. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*. **26** (3), pp. 392-398.
- Abou-Ouf, H., Zhao, L., Bismar, T. A. 2015. ERG expression in prostate cancer: biological relevance and clinical implication. *J Cancer Res Clin Oncol*.
- Alberti, C. 2008. Genetic and microenvironmental implications in prostate cancer progression and metastasis. *European Review for Medical and Pharmacological Sciences*. **12**, pp. 167-175.
- Altay, G. & Emmert-Streib, F. 2010a. Inferring the conservative causal core of gene regulatory networks. *BMC Syst. Biol.* **4**, pp. 132.
- Altay, G. & Emmert-Streib, F. 2010b. Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*. **26** (14), pp. 1738-44.
- Altay, G. & Emmert-Streib, F. 2010c. Structural influence of gene networks on their inference: analysis of C3NET. *Biology Direct*. **6**:31.
- Altay, G. 2012. Empirically determining the sample size for large-scale gene network inference algorithms. *IET Syst. Biol.* **6**, pp. 35-63.
- Altay, G., & Altay, N. 2013. Global assessment of network inference algorithms based on available literature of gene/protein interactions. *Turk J Biol.* 37:pp. 547-555.
- Altay, G., Asim, M., Markowetz, F., Neal, D. E. 2011. Differential C3NET reveals disease networks of direct physical interactions. *BMC Bioinformatics*. **12**:296.
- Amar, D., Safer, H. & Shamir, R. 2013. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput Biol.* **9**.
- Anantharaman, A. & Friedlander, T. W. 2015. Targeting the androgen receptor in metastatic castrate-resistant prostate cancer: A review. *Urol Oncol.* S1078-1439 (15) 00555-4.
- Attard, G., Clark, J., Ambrosine, L., Mills, I. G., Fisher, G. et al. 2008. Heterogeneity and clinical significance of ETV1 translocations in human prostate cancer. *Br J Cancer*. **99** (2), pp. 314-20.

- Auchus, M. L. & Auchus, R. J. 2012. Human steroid biosynthesis for the oncologist. *J Investig Med.* **60** (2), pp. 495-503.
- Balaton, C. E., Dawson, D. W., Suh, J., Sherman, M. H., Sanders, G., Hong, J. S., Frank, M. J., Malone, C. S., Said, J. W., & Teitell, M. A. 2009. Epigenetic silencing of Stk39 in B-cell lymphoma inhibits apoptosis from genotoxic stress. *Am J Pathol.* **175** (4), pp. 1653-61
- Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M. K., Chuang, R., Jaehnig, E. J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., Fiedler, D., Dutkowski, J., Guénolé, A., van Attikum, H., Shokat, K. M., Kolodner, R. D., Huh, W. K., Aebersold, R., Keogh, M. C., Krogan, N. J., & Ideker, T. 2010. Rewiring of genetic networks in response to DNA damage. *Science.* **330** (6009), pp. 1385-9.
- Bartucci, M., Svensson, S., Romania, P., Dattilo, R., Patrizii, M. et al. 2012. Therapeutic targeting of Chk1 in NSCLC stem cells during chemotherapy. *Cell Death Differ.* **19** (5), pp. 768-78.
- Bockmayr, M., Klauschen, F., Györfy, B., Denkert, C. & Budczies, J. 2013. New network topology approaches reveal differential correlation patterns in breast cancer. *BMC Systems Biology.* **7**:78.
- Bonaccorsi, L., Nosi, D., Muratori, M., Formigli, L., Forti, G., & Baldi, E. 2007. Altered endocytosis of epidermal growth factor receptor in androgen receptor positive prostate cancer cell lines. *J Mol Endocrinol.* **38** (1-2), pp. 51-66.
- Butte, A., Tamayo, P., Slonim, D., Golub, T., & Kohane, I. 2000. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci.* **97** (22), pp. 12182-6.
- Cannistraci, A., Di Pace, A. L., De Maria, R., Bonci, D. 2014. MicroRNA as new tools for prostate cancer risk assessment and therapeutic intervention: Results from clinical data set and patients' samples. *Biomed Res Int.* 146170.
- Carelli, S., Zadra, G., Vaira, V., Falleni, M., Bottiglieri, L., Nosotti, M., Di Giulio, A. M., Gorio, A., & Bosari, S. 2006. Up-regulation of focal adhesion kinase in non-small cell lung cancer. *Lung Cancer.* **53** (3), pp. 263-71.
- Castro, M. A. A. et al. 2012. RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. *Genome Biology.* **13** (4):R29.

- Chappell, W. H., Lehmann, B. D., Terrian, D. M., Abrams, S. L., Steelman, L. S., & McCubrey, J. A. 2012. p53 expression controls prostate cancer sensitivity to chemotherapy and the MDM2 inhibitor Nutlin-3. *Cell Cycle*. **11** (24), pp. 4579-4588.
- Chen, D. et al., 2014. miR-100 Induces Epithelial-Mesenchymal Transition but Suppresses Tumorigenesis, Migration and Invasion. *PLoS Genetics*, 10(2).
- Chen, D.B. & Yang, H.J., 2014. Comparison of gene regulatory networks of benign and malignant breast cancer samples with normal samples. *Genetics and Molecular Research*, 13(4), pp.9453–9462.
- Cho, S. B., Kim, J., & Kim, J. H. 2009. Identifying set-wise differential coexpression in gene expression microarray data. *BMC Bioinformatics*. **10**:109.
- Choi, Y. J., Yoo, N. J., & Lee, S. H. 2014. Down-regulation of ROBO2 Expression in Prostate Cancers. *Pathol Oncol Res*. **20** (3), pp. 517-519.
- Chong, I. W., Chang, M. Y., Chang, H. C., Yu, Y. P., Sheu, C. C., Tsai, J. R. et al. 2006. Great potential of a panel of multiple hMTH1, SPD, ITGA11 and COL11A1 markers for diagnosis of patients with non-small cell lung cancer. *Oncol Rep* **16**: 981-988.
- Choudhary, V., Kaddour-Djebbar, I., Lakshmikanthan, V., Ghazaly, T., Thangjam, G. S., Sreekumar, A., Lewis, R. W., Mills, I. G., Bollag, W. B., Kumar, M. V. 2011. Novel role of androgens in mitochondrial fission and apoptosis. *Mol Cancer Res*. **9** (8), pp. 1067-77.
- Chu, G., Narasimhan, B., Tibshirani, R., & Tusher, V. 2002. Significance analysis of microarrays (SAM) software. *Nature*. **5**, pp. 436–442.
- Croce, C. M. 2008. Oncogenes and Cancer. *The New England Journal of Medicine*. **358**, pp. 502-511.
- Csardi, G. & Nepusz, T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems*. **1695**. <http://igraph.org>
- Curtis, C., Shah, S.P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. **486** (7403), pp. 346-352.

- da Huang, W., Sherman, B. T., Lempicki, R. A. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, pp. 1-13.
- de la Fuente, A. 2010. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet.* **26** (7), pp. 326-33.
- Efron, B., Tibshirani, R., JD, S., & Tusher, V. 2001. Empirical Bayes analysis of a microarray experiment. *J AmStat Assoc.* **96** (456), pp. 1151-1160.
- Emmert-Streib, F. & Dehmer, M. 2009a. Information Processing in the Transcriptional Regulatory Network of Yeast: Functional Robustness. *BMC Systems Biology* **3**:35.
- Emmert-Streib, F. & Dehmer, M. 2009b. Predicting cell cycle regulated genes bycausal interactions. *Plos One.* **4** (8):e6633.
- Emmert-Streib, F., Glazko, G. V., Altay, G., & de Matos Simoes, R. 2012. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Frontiers in Genetics.* **3**: 8.
- Emmert-Streib, F., Tripathi, S., & de Matos Simoes, R. 2012. Harnessing the complexity of gene expression data from cancer: from single gene to structural pathway methods. *Biol Direct.* **7**: 44.
- Ergun, A., Lawrence, C. A., Kohanski, M. A., Brennen, T. A., & Collins, J. J. 2007. A network biology approach to prostate cancer. *Mol Syst Biol*, **3**:82.
- Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., & Gardner, T. S. 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PloS Biol.* **5** (1):e8.
- Fan, M., Sethuraman, A., Brown, M., Sun, W., & Pfeffer, L. M. 2014. Systematic analysis of metastasis-associated genes identifies miR-17-5p as a metastatic suppressor of basal-like breast cancer. *Breast Cancer Res Treat.* **146** (3), pp. 487-502.
- Fang, D. D., Cao, J., Jani, J. P., Tsaparikos, K., Blasina, A., Kornmann, J., Lira, M. E., Wang, J., Jirout, Z., Bingham, J., Zhu, Z., Gu, Y., Los, G., Hostomsky, Z., & Vanarsdale, T. 2013. Combined gemcitabine and CHK1 inhibitor treatment

- induces apoptosis resistance in cancer stem cell-like cells enriched with tumor spheroids from a non-small cell lung cancer cell line. *Front Med.* **7** (4), pp. 462-76.
- Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J. W., Comber, H., Forman, D., Bray, F. 2013. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer.* **49** (6), pp. 1374-403.
- Ferraldeschi, R., Sharifi, N., Auchus, R. J., & Attard, G. 2013. Molecular Pathways: Inhibiting steroid biosynthesis in prostate cancer. *Clin Cancer Res.* **19** (13), pp. 3353-3359.
- Ferrer, C.M. et al., 2016. O-GlcNAcylation regulates breast cancer metastasis via SIRT1 modulation of FOXM1 pathway. *Oncogene*, (April), pp.1–11.
- Freudenberg, J., Sivaganesan, S., Wagner, M., & Medvedovic, M. 2010. A semi-parametric Bayesian model for unsupervised differential co-expression analysis. *BMC Bioinformatics.* **11**:234.
- Fukushima, A. 2013. DiffCorr: an R package to analyze and visualize differential correlations in biological networks. *Gene.* **518** (1), pp. 209-14.
- Gambardella, G., Moretti, M.N., de Cegli, R., Cardone, L., Peron, A., & di Bernardo, D. 2013. Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics.* **29** (14), pp.1776-1785.
- Gao, Y., Li, G., Sun, L., He, Y., Li, X., Sun, Z., Wang, J., Jiang, Y., & Shi, J.. 2015. ACTN4 and the pathways associated with cell motility and adhesion contribute to the process of lung cancer metastasis to the brain. *BMC Cancer.* **15**:277.
- Gill, R., Datta, S., & Datta, S. 2014a. Differential network analysis in human cancer research. *Curr Pharm Des.* **20** (1), pp. 4-10.
- Gill, R., Datta, S., & Datta, S. 2010. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics.* **11**:95.
- Gill, R., Datta, S., & Datta, S. 2014b. dna: An R package for differential network analysis. *Bioinformatics.* **10** (4), pp. 233-4.
- Gravina, G. L., Mancini, A., Muzi, P., Ventura, L., Biordi, L., Ricevuto, E., Pompili, S., Mattei, C., Di Cesare, E., Jannini, E. A., & Festuccia, C. 2015. CXCR4 pharmacological inhibition reduces bone and soft tissue metastatic burden by

- affecting tumor growth and tumorigenic potential in prostate cancer preclinical models. *Prostate*. **75** (12), pp. 1227-46.
- Guo, X., Wang, Y., Wang, C., & Chen, J. 2015. Identification of several hub-genes associated with periodontitis using integrated microarray analysis. *Mol Med Rep*. **11** (4), pp. 2541-7.
- Gupta, K., Thakur, V. S., Bhaskaran, N., Nawab, A., Babcook, M. A., Jackson, M. W., & Gupta, S. 2012. Green tea polyphenols induce p53-dependent and p53-independent apoptosis in prostate cancer cells through two distinct mechanisms. *PLoS One*. **7** (12):e52572.
- Ha, M. J., Baladandayuthapani, V., & Do, K. 2015. DINGO: Differential Network Analysis in Genomics. *Bioinformatics*. **31** (21), pp. 3413–3420
- Hamurcu, Z. et al., 2016. FOXM1 regulates expression of eukaryotic elongation factor 2 kinase and promotes proliferation, invasion and tumorigenesis of human triple negative breast cancer cells. *Oncotarget*, 7(13), pp.1–17.
- Han, X. & Poon, R.Y.C., 2013. Critical Differences between Isoforms of Securin Reveal Mechanisms of Separase Regulation. *Molecular and Cellular Biology*, **33** (17), pp.3400–3415.
- Harburg, G. C. & Hinck, L. 2011. Navigating Breast Cancer: Axon Guidance Molecules as Breast Cancer Tumor Suppressors and Oncogenes. *J Mammary Gland Biol Neoplasia*. **16** (3), pp. 257-270.
- Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., & Guthke, R. 2009. Gene regulatory network inference: data integration in dynamic models – a review. *Biosystems*. **96** (1), pp. 86-103.
- Helsten, T. et al., 2016. Cell cycle gene alterations in 4864 tumors analyzed by next generation sequencing: Implications for targeted therapeutics. *Molecular Cancer Therapeutics*, **11**, pp.16–19.
- Hill, C. et al., 2015. HHS Public Access. Proceedings of SPIE--the International Society for Optical Engineering, 73(4), pp.389–400.
- Howe, G. A., Xiao, B., Zhao, H., Al-Zahrani, K. N., Hasim, M. S., Villeneuve, J., Sekhon, H. S., Goss, G. D., Sabourin, L. A., Dimitroulakos, J., Addison, C. L. 2016. Focal Adhesion Kinase Inhibitors in Combination with Erlotinib

- Demonstrate Enhanced Anti-Tumor Activity in Non-Small Cell Lung Cancer. *PLoS One*. **11** (3):e0150567.
- Hsiao, T. H., Chiu, Y. C., Hsu, P. Y., Lu, T. P., Lai, L. C., Tsai, M. H., Huang, T. H., Chuang, E. Y., & Chen, Y. 2016. Differential network analysis reveals the genome-wide landscape of estrogen receptor modulation in hormonal cancers. *Sci Rep*. **4** (6): 23035.
- Hsu, C., Juan, H., & Huang, H. 2015. Functional Analysis and Characterization of Differential Coexpression Networks. *Scientific Reports*. **5**.
- Hu, R., Qiu, X., Glazko, G., Klebanoz, L., & Yakovlev, A. 2009. Detecting intergene correlation changes in microarray analysis: a new approach to gene selection. *BMC Bioinformatics*. **10** (1).
- Huang, Q., Li, L., Lin, Z., Xu, W., Han, S., Zhao, C., Li, L., Cao, W., Yang, X., Wei, H., & Xiao, J. 2016. Identification of Preferentially Expressed Antigen of Melanoma as a Potential Tumor Suppressor in Lung Adenocarcinoma. *Med Sci Monit*. **22**, pp. 1837-42.
- Huang, Y. et al., 2014. Autophagy promotes radiation-induced senescence but inhibits bystander effects in human breast cancer cells. *Landes Bioscience*, **10** (7), pp.1212–1228.
- Hurst, D. R. & Welch, D. R. 2011. Metastasis Suppressor Genes: At the Interface Between the Environment and Tumor Cell Growth. *Int Rev Cell Mol Biol*. **286**, pp. 107–180.
- Iqbal, J., Ginsburg, O., Rochon, P. A., Sun, P., & Narod, S. A. 2015. Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. *JAMA*. **313** (2), pp. 165-73.
- Iwamoto, T. & Pusztai, L. 2010. Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data? *Genome Medicine*. **2** (81).
- Jia, P., Liu, Y., & Zhao, Z. 2012. Integrative pathway analysis of genome-wide association studies and gene expression data in prostate cancer. *BMC Systems Biology*. **6** (3):13.
- Joe, S. & Nam, H. 2016. Prognostic factor analysis for breast cancer using gene expression profiles. *BMC Med Inform Decis Mak*. **16**: 56.

- Joshi, A., De Smet, R., Marchal, K., Van de Peer, Y., & Michoel, T. 2009. Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics*. **25** (4), pp. 490-496.
- Kakimi, K., Matsushita, H., Murakawa, T., & Nakajima, J. 2014. $\gamma\delta$ T cell therapy for the treatment of non-small cell lung cancer. *Transl Lung Cancer Res*. **3** (1), pp. 23-33.
- Kanehisa, M. & Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. **28** (1), pp. 27-30.
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*. **40**, pp. 109-114.
- Karatas, O. F., Guzel, E., Duz, M. B., Ittmann, M., & Ozen, M. 2015. The role of ATP-binding cassette transporter genes in the progression of prostate cancer. Prostate. doi: 10.1002/pros.23137. [Epub ahead of print].
- Karin, M. 1990. Too many transcription factors: positive and negative interactions. *New Biol*. **2** (2), pp. 126-31.
- Kauffman, E. C., Robinson, V. L., Stadler, W. M., Sokoloff, M. H., & Rinker-Schaeffer, C. W. 2003. Metastasis Suppression: The Evolving Role of Metastasis Suppressor Genes for Regulating Cancer Cell Growth at the Secondary Site. *The Journal of Urology*. **169** (3), pp. 1122-1133.
- Kaur, P. & Khatik, G. L. 2016. Advancements in Non-steroidal Antiandrogens as Potential Therapeutic Agents for the Treatment of Prostate Cancer. *Mini Rev Med Chem*. [Epub ahead of print]
- Kim, D. H. & Wirtz, D. 2013. Focal adhesion size uniquely predicts cell migration. *FASEB J*. **27** (4), pp. 1351-61.
- Kleinert, R., Prenzel, K., Stoecklein, N., Alakus, H., Bollschweiler, E., Hölscher, A., & Warnecke-Eberz, U. 2015. Gene Expression of Col11A1 Is a Marker Not only for Pancreas Carcinoma But also for Adenocarcinoma of the Papilla of Vater, Discriminating Between Carcinoma and Chronic Pancreatitis. *Anticancer Res*. **35** (11), pp. 6153-8.
- Koonin, E. V. 2012. Does the central dogma still stand? *Biol Direct*. **7**: 27.

- Kristensen, V.N. et al., 2012. Integrated molecular profiles of invasive breast tumors and ductal carcinoma in situ (DCIS) reveal differential vascular and interleukin signaling. *Proc Natl Acad Sci*, **109** (8), pp.2–7.
- Kurt, Z., Aydin, N., & Altay, G. 2014a. A comprehensive comparison of association estimators for gene network inference algorithms. *Bioinformatics*. **30** (15), pp. 2142-2149.
- Kurt, Z., Aydin, N., & Altay, G. 2014b. Comprehensive review of association estimators for the inference of gene networks. *Turkish Journal of Electrical Engineering & Computer Sciences*.
- Lai, Y., Wu, B., Chen, L., & Zhao, H. 2004. A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics*. **20** (17), pp. 3146-55.
- Latchman, D.S. 1997. Transcription Factors: An Overview. *International Journal of Biochemistry and Cell Biology*. **29** (12), pp. 1305-1312.
- Lee, J.-J. et al., 2016. Expression of FOXM1 and related proteins in breast cancer molecular subtypes. *International Journal of Experimental Pathology*, 97(2), pp.170–177.
- Lee, M.-H. et al., 2012. Yin Yang 1 positively regulates BRCA1 and inhibits mammary cancer formation. *Oncogene*, 31(1), pp.116–27.
- Li, J. et al., 2016. Clinicopathological significance of STAT4 in hepatocellular carcinoma and its effect on cell growth and apoptosis. *Onco Targets and Therapy*, **9**, pp.1721–1734.
- Li, W. 1990. Mutual information functions versus correlation functions. *Journal of Statistical Physics*. **60** (5-6), pp. 823-837.
- Liang, H. et al., 2015. miR-655 inhibit pituitary tumor cell tumorigenesis and are involved in a p53 / PTTG1 regulation feedback loop. *Oncotarget*, **6** (30).
- Lin, B., White, J. T., Lu, W., Xie, T., Utleg, A. G., Yan, X., Yi, E. C., Shannon, P., Khrebtukova, I., Lange, P. H., Goodlett, D. R., Zhou, D., Vasicek, T. J., & Hood, L. 2005. Evidence for the Presence of Disease-Perturbed Networks in Prostate Cancer Cells by Genomic and Proteomic Analyses: A Systems Approach to Disease. *Cancer Res*. **65** (8), pp. 3081-91.

- Liu, B., Qu, J., Xu, F., Guo, Y., Wang, Y., Yu, H., & Qian, B. 2015. MiR-195 suppresses non-small cell lung cancer by targeting CHEK1. *Oncotarget*. **6** (11), pp. 9445-56.
- Liu, F-S., Chen, J-T., Dong, J-T., Hsieh, Y-T., Lin, A-J., Ho, E. S-C., Hung, M-J., & Lu, C-H. 2001. Metastasis Suppressor Gene Is Frequently Down-Regulated in Cervical Carcinoma. *The American Journal of Pathology*. **159** (5), pp. 1629–1634.
- Liu, S., Li, L., Zhang, Y., Zhang, Y., Zhao, Y. et. al. 2012. The Oncoprotein HBXIP Uses Two Pathways to Up-regulate S100A4 in Promotion of Growth and Migration of Breast Cancer Cells. *J Biol Chem*. **287** (36), pp. 30228–30239.
- Lu, T. P., Tsai, M. H., Lee, J. M., Hsu, C. P. et al. 2010. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev*. **19** (10), pp. 2590-7.
- Lv, M. & Wang, L. 2015. Comprehensive analysis of genes, pathways, and TFs in nonsmoking Taiwan females with lung cancer. *Exp Lung Res*. **41** (2), pp. 74-83.
- Ma, C., Xin, M., Feldmann, K. A., & Wang, X. 2014. Machine learning-based differential network analysis: a study of stress-responsive transcriptomes in Arabidopsis. *Plant Cell*. **26** (2), pp. 520-37.
- Ma, D., Zhou, Z., Yang, B., He, Q., Zhang, Q., & Zhang, X. 2015. Association of molecular biomarkers expression with biochemical recurrence in prostate cancer through tissue microarray immunostaining. *Oncol Lett*. **10** (4), pp. 2185–2191.
- Ma, H., Schadt, E.E., Kaplan, L.M., & Zhao, H. 2011. COSINE: COndition-Specific sub-NETwork identification using a global optimization method. *Bioinformatics*. **27** (9), pp. 1290-1298.
- Madhamshettiwar, P. B., Maetschke, S. R., Davis, M. J., & Ragan, M. A. 2013. RMaNI: Regulatory Module Network Inference framework. *BMC Bioinformatics*. **14** (16):S14.
- Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*. **7** (7) Suppl 1, pp. 7.

- Marques, R. B., Dits, N. F., Erkens-Schulze, S., van Ijcken, W. F., van Weerden, W. M., & Jenster, G. 2011. Modulation of androgen receptor signaling in hormonal therapy-resistant prostate cancer cell lines. *PLoS One*. **6** (8):e23144.
- Massie, C. E., Lynch, A., Ramos-Montoya, A., Boren, J. et al. 2011. The androgen receptor fuels prostate cancer by regulating central metabolism and biosynthesis. *EMBO J*. **30** (13), pp. 2719-33.
- McLean, G. W., Carragher, N. O., Avizienyte, E., Evans, J., Brunton, V. G., & Frame, M. C. 2005. The role of focal-adhesion kinase in cancer - a new therapeutic opportunity. *Nat Rev Cancer*. **5** (7), pp. 505-15.
- Mehlen, P., Delloye-Bourgeois, C., & Chédotal, A. 2011. Novel roles for Slits and netrins: axon guidance cues as anticancer targets? *Nature Reviews Cancer*. **11**:188-197.
- Menche, J., Sharma, A., Kitsak, M., Ghiassian, S., Vidal, M., et al. 2015. Uncovering disease-disease relationships through the incomplete human interactome. *Science*, **347** (6224)
- Metodieva, S. N., Nikolova, D. N., Cherneva, R. V., Dimova, II., Petrov, D. B., & Toncheva, D. I. 2011. Expression analysis of angiogenesis-related genes in Bulgarian patients with early-stage non-small cell lung cancer. *Tumori*. **97** (1), pp. 86-94.
- Meyer, P. E., Kontos, K., Latiffe, F., & Bontempi, G. 2007. Information-theoretic inference of large transcriptional regulatory networks. *EUROSIPJ Bioinform. Syst. Biol.* **2007** (1).
- Meyer, P. E., LAFitte, F., & Bontempi, G. 2008. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinformatics*. **9**:461.
- Migita, T., Ruiz, S., Fornari, A., Fiorentino, M., Priolo, C., Zadra, G., Inazuka, F. Et et al. 2009. Fatty acid synthase: a metabolic enzyme and candidate oncogene in prostate cancer. *J Natl Cancer Inst.* **101**, pp. 519-532
- Miki, J., Furusato, B., Li, H., Gu, Y., Takahashi, H., Egawa, S., Sesterhenn, I. A., McLeod, D. G., Srivastava, S., & Rhim, J. S. 2007. Identification of putative stem cell markers, CD133 and CXCR4, in hTERT-immortalized primary nonmalignant

- and malignant tumor-derived human prostate epithelial cell lines and in prostate cancer specimens. *Cancer Res.* **67** (7), pp. 3153-61.
- Myers, J. S., Lersner, A. K., Robbins, C. J., Sang, Q. A. 2015. Differentially Expressed Genes and Signature Pathways of Human Prostate Cancer. *PLoS One.* **10** (12): e0145322.
- Odibat, O. & Reddy, C. K. 2011. A Generalized Framework for Mining Arbitrarily Positioned Overlapping Co-clusters. *Proceedings of the 2011 SIAM International Conference on Data Mining.* pp. 343-354.
- Odibat, O. & Reddy, C.K. 2012. Ranking Differential Hubs In Geneco-Expression Networks. *Journal of Bioinformatics and Computational Biology.* **10** (1), pp. 1240002.
- Olsen, C., Fleming, K., Prendergast, N., Rubio, R., Emmert-Streib, F., Bontempi, G., Haibe-Kains, B., & Quackenbush, J. 2014. Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics.* **103** (5-6), pp. 329-36.
- Prieto, C., Rivas, M. J., Sanchez, J. M., Lopez-Fidalgo, J., & De Las Rivas, J. 2006. Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. *Bioinformatics.* **22** (9), pp. 1103-1110.
- Priolo, C., Pyne, S., Rose, J., Regan, E. R., Zadra, G., Photopoulos, C., Cacciatore, S., Schultz, D., Scaglia, N., McDunn, J., De Marzo, A. M., & Loda, M. 2014. AKT1 and MYC induce distinctive metabolic fingerprints in human prostate cancer. *Cancer Res.* **74** (24), pp. 7198-204.
- Pritchard, C. C., Hsu, L., Delrow, J., & Nelson, P. S. 2001. Project normal: Defining normal variance in mouse gene expression. *Proceedings of the National Academy of Sciences of the United States of America.* **98** (23), pp. 13266-13271.
- Qi, H. et al., 2015. Role of annexin A6 in cancer (Review). *Oncology Letters*, pp.1–6.
- Ren, W., Li, C, Duan, W., Du, S., Yang, F., Zhou, J., & Xing, J. 2015. MicroRNA-613 represses prostate cancer cell proliferation and invasion through targeting Frizzled7. *Biochem Biophys Res Commun.* **469** (3), pp. 633-8.
- Reverter-Gomez, A., Hudson, N. J., Nagaraj, S. H., Perez-Enciso, M., & Dalrymple, B. P. 2010. Regulatory Impact Factors: Unraveling the transcriptional regulation of complex traits from expression data. *Bioinformatics.* **26** (7), pp. 896-904.

- Reverter, A., Ingham, A., Lehnert, S. A., Tan, S. H., Wang, Y., Ratnakumar, A., & Dalrymple, B. P. 2006. Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. *Bioinformatics*. **22**, pp. 2396-404.
- 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), pp. e47.
- Rocks, N., Paulissen, G., Quesada Calvo, F., Polette, M., Gueders, M., Munaut, C., Foidart, J. M., Noel, A., Birembaut, P., & Cataldo, D. 2006. Expression of a disintegrin and metalloprotease (ADAM and ADAMTS) enzymes in human non-small-cell lung carcinomas (NSCLC). *Br J Cancer*. **94** (5), pp. 724-30.
- Romanuik, T. L. & Wang, G., 2010. Morozova O, Delaney A, Marra MA, Sadar MD. LNCaP Atlas: Gene expression associated with in vivo progression to castration-recurrent prostate cancer. *BMC Medical Genomics*. **3**:43.
- Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A. et al. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*. **437** (7062), pp. 1173-8.
- Saba, R. et al., 2016. The Role of Forkhead Box Protein M1 in Breast Cancer Progression and Resistance to Therapy. *International journal of breast cancer*, 2016, p.9768183.
- Savli, H., Szendrői, A., Romics, I., & Nagy, B. 2008. Gene network and canonical pathway analysis in prostate cancer: a microarray study. *Exp Mol Med*. **40** (2), pp. 176-185.
- Schadt E. E. 2009. Molecular networks as sensors and drivers of common human diseases. *Nature*. **461** (7261), pp. 218-23.
- Seah, B., Bhowmick, S. S., & Dewey, C. F. Jr. 2014. DiffNet: automatic differential functional summarization of dE-MAP networks. *Methods*. **69** (3), pp. 247-56.
- Shanmugam, M. K., Rajendran, P., Li, F., Nema, T., Vali, S., Abbasi, T., Kapoor, S., Sharma, A., Kumar, A. P., Ho, P. C., Hui, K. M., & Sethi, G. 2011. Ursolic acid inhibits multiple cell survival pathways leading to suppression of growth of prostate cancer xenograft in nude mice. *J Mol Med (Berl)*. **89** (7), pp. 713-27.

- 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 Res.* **13** (11), pp. 2498-504.
- Sharifi, N. & Auchus RJ. 2012. Steroid biosynthesis and prostate cancer. *Steroids.* **77** (7), pp. 719-726.
- Sheng, X., Bowen, N., & Wang, Z. 2016. GLI pathogenesis-related 1 functions as a tumor-suppressor in lung cancer. *Mol Cancer.* **15**:25.
- Singh, R. K. & Lokeshwar, B. L. 2011. The IL-8-regulated chemokine receptor CXCR7 stimulates EGFR signaling to promote prostate cancer growth. *Cancer Res.* **71** (9), pp. 3268-77.
- Stegh, A. H. 2012. Targeting the p53 signaling pathway in cancer therapy - The promises, challenges, and perils. *Expert Opin Ther Targets.* **16** (1), pp. 67-83.
- Steuer, R., Kurths, J., Daub, C. O., Weise, J., & Selbig, J. 2002. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics.* **18** (2), pp. 231-240.
- Stoss, O., Werther, M., Zielinski, D., Middel, P., Jost, N., Rüschoff, J., Henkel, T., & Albers, P. 2008. Transcriptional profiling of transurethral resection samples provides insight into molecular mechanisms of hormone refractory prostate cancer. *Prostate Cancer Prostatic Dis.* **11** (2), pp. 166-72.
- Syljuasen, R. G., Hasvold, G., Hauge, S., & Helland, A. 2015. Targeting lung cancer through inhibition of checkpoint kinases. *Front Genet.* **6**:70.
- Tamura, K., Makino, A., Hullin-Matsuda, F., Kobayashi, T., Furihata, M., Chung, S., Ashida, S., Miki, T., Fujioka, T., Shuin, T., Nakamura, Y., & Nakagawa, H. 2009. Novel lipogenic enzyme ELOVL7 is involved in prostate cancer growth through saturated long-chain fatty acid metabolism. *Cancer Res.* **69** (20), pp. 8133-40.
- Tesson, B. M., Breitling, R. & Jansen, R. C. 2010. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics.* **11**:497.
- Tian, Y., Zhang, B., Hoffman, E. P., Clarke, R., Zhang, Z., Shih, IeM., Xuan, J., Herrington, D. M., & Wang, Y. 2015. KDDN: an open-source Cytoscape app for

- constructing differential dependency networks with significant rewiring. *Bioinformatics*. **31** (2), pp. 287-9.
- Tian, Z. Q., Li, Z. H., Wen, S. W., Zhang, Y. F., Li, Y., Cheng, J. G., Wang, G. Y. 2015. Identification of Commonly Dysregulated Genes in Non-small-cell Lung Cancer by Integrated Analysis of Microarray Data and qRT-PCR Validation. *Lung*. **193** (4), pp. 583-92.
- Ulitsky, I., Krishnamurthy, A., Karp, R.M., & Shamir, R. 2010. DEGAS: De Novo Discovery of Dysregulated Pathways in Human Diseases. *PLoS ONE*. **5** (10), pp. e13367.
- Valcárcel B., Würtz P, Seich al Basatena N.K., Tukiainen T., Kangas A.J., Soininen P., Järvelin M.R., Ala-Korpela M., Ebbels T.M., & de Iorio M. 2011. A differential network approach to exploring differences between biological states: an application to prediabetes. *PLoS One*. **6** (9), e24702.
- Van Landeghem, S., Van Parys, T., Dubois, M., Inzé, D., & Van de Peer, Y. 2016. Diffany: an ontology-driven framework to infer, visualise and analyse differential molecular networks. *BMC Bioinformatics*. **17**:18.
- Van Nas, A., Guhathakurta, D., Wang, S. S., Yehya, N., Horvath, S., Zhang, B., Ingram-Drake, L., Chaudhuri, G., Schadt, E. E., Drake, T. A., Arnold, A. P., & Lusk, A. J. 2009. Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks. *Endocrinology*. **150** (3), pp. 1235-49.
- Vázquez-Villa, F., García-Ocaña, M., Galván, J. A., García-Martínez, J., García-Pravia, C., Menéndez-Rodríguez, P., González-del Rey, C., Barneo-Serra, L., de Los Toyos, J. R. 2015. COL11A1/(pro)collagen 11A1 expression is a remarkable biomarker of human invasive carcinoma-associated stromal cells and carcinoma progression. *Tumour Biol*. **36** (4), pp. 2213-22.
- Wan, M. et al., 2012. Yin Yang 1 plays an essential role in breast cancer and negatively regulates p27. *The American journal of pathology*, **180** (5), pp.2120–33.
- Wang, G. et al., 2015. Decreased STAT4 indicates poor prognosis and enhanced cell proliferation in hepatocellular carcinoma. *World Journal of Gastroenterology*, **21** (13), pp.3983–3993.
- Wang, J. et al., 2015. YY1 suppresses FEN1 over-expression and drug resistance in breast cancer. *BMC Cancer*, **15** (1), p.1043.

- Wang, J., Shiozawa, Y., Wang, J., Wang, Y., Jung, Y., Pienta, K. J., Mehra, R., Loberg, R., & Taichman, R. S. 2008. The role of CXCR7/RDC1 as a chemokine receptor for CXCL12/SDF-1 in prostate cancer. *J Biol Chem.* **283** (7), pp. 4283-94.
- Wang, M. et al., 2015. Roles of miR-186 and PTTG1 in colorectal neuroendocrine tumors. *Int J Clin Exp Med*, **8** (12), pp.22149–22157.
- Wang, Z.T. et al., 2016. Histone deacetylase inhibitors suppress mutant p53 transcription via HDAC8/YY1 signals in triple negative breast cancer cells. *Cellular Signalling*, **28** (5), pp.506–515.
- Warsow, G., Struckmann, S., Kerkhoff, C., Reimer, T., Engel, N., & Fuellen, G. 2013. Differential network analysis applied to preoperative breast cancer chemotherapy response. *PLoS One.* **8** (12):e81784.
- Watson, M. 2006. CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics.* **7**:509.
- Webber, P. J., Park, C., Qui, M., Ramalingam, S. S., Khuri, F. R., Fu, H., & Du, Y. 2015. Combination of heat shock protein 90 and focal adhesion kinase inhibitors synergistically inhibits the growth of non-small cell lung cancer cells. *Oncoscience.* **2** (9), pp. 765–776.
- White-Means, S., Rice, M., Dapremont, J., Davis, B., & Martin, J. 2015. African American Women: Surviving Breast Cancer Mortality against the Highest Odds. *Int J Environ Res Public Health.* **13** (1).
- Wu, B. 2007. Cancer outlier differential gene expression detection. *Biostatistics.* **8** (3), pp. 566–575.
- Wu, D. et al., 2016. Molecular mechanisms associated with breast cancer based on integrated gene expression profiling by bioinformatics analysis. *Journal of obstetrics and gynaecology*, **3615**, pp.1–7.
- Wu, Y. H., Chang, T. H., Huang, Y. F., Huang, H. D., & Chou, C. Y. 2014. COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene.* **33** (26), pp. 3432-40.
- Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., & Liang, H. 2014. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications.* **5**.

- Ye, X. et al., 2015. Quantitative proteomic analysis identifies new effectors of FOXM1 involved in breast cancer cell migration. *International Journal of Clinical and Experimental Pathology*, **8** (12), pp.15836–15844.
- Yoshida, B. A., Sokoloff, M. M., Welch, D. R., & Rinker-Schaeffer, C. W. 2000. Metastasis-Suppressor Genes: a Review and Perspective on an Emerging Field. *Journal of the National Cancer Institute*. **92** (21), pp. 1717-1730.
- Yu, H., Liu, B., Ye, Z., Li, C., Li, Y., & Li, Y. 2011. Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. *BMC Bioinformatics*. **12**, 315.
- Yu, H., Liu, B., Ye, Z., Li, C., Li, Y., & Li, Y. Y. 2011. Link-based quantitative methods to identify differentially coexpressed genes and gene Pairs. *BMC Bioinformatics*. **12**:315.
- Yuan, F. & Wang, W., 2015. MicroRNA-802 suppresses breast cancer proliferation through downregulation of FoxM1. *Molecular Medicine Reports*, **12** (3), pp.4647–4651.
- Yun, H. J., Ryu, H., Choi, Y. S., Song, I. C., Jo, D. Y., Kim, S., & Lee, H. J. 2015. C-X-C motif receptor 7 in gastrointestinal cancer. *Oncol Lett*. **10** (3), pp. 1227-1232.
- Zakaria, N., Yusoff, N. M., Zakaria, Z., Lim, M. N., Baharuddin, P. J., Fakiruddin, K. S., & Yahaya, B. 2015. Human non-small cell lung cancer expresses putative cancer stem cell markers and exhibits the transcriptomic profile of multipotent cells. *BMC Cancer*. **15**:84.
- Zhang, B., & Horvath, S. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. **4**.
- Zhang, B., Li, H., Riggins, R. B., Zhan, M., Xuan, J., Zhang, Z., Hoffman, E. P., Clarke, R. & Wang, Y. 2009. Differential dependency network analysis to identify condition-specific topological changes in biological networks. *Bioinformatics*. **25** (4), pp. 526-32
- Zhang, B., Tian, Y., Jin, L., Li, H., Shih, I., Madhavan, S., Clarke, R., Hoffman, E. P., Xuan, J., Hilakivi-Clarke, L., & Wang, Y. 2011. DDN: a caBIG® analytical tool for differential network analysis. *Bioinformatics*. **27** (7), pp. 1036-8.

- Zhang, D., Zhu, H., & Harpaz, N. 2016. Overexpression of $\alpha 1$ chain of type XI collagen (COL11A1) aids in the diagnosis of invasive carcinoma in endoscopically removed malignant colorectal polyps. *Pathol Res Pract.* **212** (6), pp. 545-8.
- Zhang, H., Ye, J., Weng, X., Liu, F., He, L., Zhou, D., & Liu, Y. 2015. Comparative transcriptome analysis reveals that the extracellular matrix receptor interaction contributes to the venous metastases of hepatocellular carcinoma. *Cancer Genet.* **208** (10), pp. 482-91.
- Zhang, L. J., Xiong, Y., Nilubol, N., He, M., Bommareddi, S., Zhu, X., Jia, L., Xiao, Z., Park, J. W., Xu, X., Patel, D., Willingham, M. C., Cheng, S. Y., & Kebebew, E. 2015. Testosterone regulates thyroid cancer progression by modifying tumor suppressor genes and tumor immunity. *Carcinogenesis.* **36** (4), pp. 420-8.
- Zhang, M. et al., 2016. Negative regulation of CDC42 expression and cell cycle progression by miR-29a in breast cancer. *Open Medicine*, **11** (1), pp.78–82.
- Zhang, X., Xiao, D., Wang, Z., Zou, Y., Huang, L., Lin, W., Deng, Q., Pan, H., Zhou, J., Liang, C., & He, J. 2014. MicroRNA-26a/b regulate DNA replication licensing, tumorigenesis, and prognosis by targeting CDC6 in lung cancer. *Mol Cancer Res.* **12** (11), pp. 1535-46.
- Zheng, C., Yuan, L., Sha, W. & Sun, Z. 2014. Gene differential coexpression analysis based on biweight correlation and maximum clique. *BMC Bioinformatics.* 2014; **15** (Suppl 15): S3.
- Zheng, K., Li, H. Y., Su, X. L., Wang, X. Y., Tian, T., Li, F., & Ren, G. S. 2010. Chemokine receptor CXCR7 regulates the invasion, angiogenesis and tumor growth of human hepatocellular carcinoma cells. *J Exp Clin Cancer Res.* **29**:31.
- Zhuang, D.Y. et al., 2015. Identification of hub subnetwork based on topological features of genes in breast cancer. *International Journal of Molecular Medicine*, **35** (3), pp.664–674.

Other Sources

Dunning, M., Lynch, A. & Eldridge, M. 2016. illuminaHumanv3.db: Illumina HumanHT12v3 annotation data (chip illuminaHumanv3). R package version 1.26.0.

Kurt, Z. (2013). Gen ađı ıkarım algoritmaları iin en uygun iliŐki kestirimcilerinin belirlenmesi. *Thesis for the Ph.D. Degree*. İstanbul: Yıldız Teknik Üniversitesi FBE.

Odibat, O. (2012). Differential modeling for cancer microarray data. *Thesis for the degree of Doctor of Philosophy*. Wayne State University Dissertations.

Yang, Y. (2013). Gene regulatory network analysis and web-based application development. *Thesis for the degree of Doctor of Philosophy*. The University of Southern Mississippi Dissertations.

<http://www.cancer.org/cancer/cancercauses/geneticsandcancer/genesandcancer/genes-and-cancer-oncogenes-tumor-suppressor-genes>

<https://www.broadinstitute.org/education/glossary/oncogene>

<https://global.britannica.com/science/transcription-factor>