PATHWAY ANALYSIS OF DISEASE-GENE ASSOCIATED NETWORK IN THE HUMAN BREAST CANCER

By

Lamia

Student ID: 14176002

A thesis submitted to the Department of Mathematics and Natural Sciences in partial fulfillment of the requirements for the degree of Master of Science in Biotechnology

Department of Mathematics and Natural Sciences

BRAC University

January 2020

Declaration

It is hereby declared that

1. The thesis submitted is my/our original work while completing a degree at BRAC University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material that has been accepted or submitted for any other degree or diploma at a university or other institution.

4. I have acknowledged all of the main sources of help.

Lamia

_____

Lamia

Student ID: 14176002

Approval

The thesis/project titled "pathway analysis of disease-gene associated network in the breast cancer" submitted by Lamia (Student ID 14176002) has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Science on January 20, 2020.

Examining Committee:

Supervisor: _____

(Member**)**

Dr. Mahboob Hossain
Professor, Microbiology Discipline
MNS Department BRAC University, Dhaka

Supervisor: _____
(Member)

RasifAjwad
Lecturer,
Department of Computer Science & Engineering
BRAC University, Dhaka

Program Coordinator: _____
(Member**)**

Iftekhar Bin Naser, PHD
Assistant Professor, MNS department
BRAC University, Dhaka

External Expert Examiner: _____
(Member)

Departmental Head: _____
(Chair**)**

A F M Yusuf Haider, PhD
Professor and Chairperson
MNS Department
BRAC University, Dhaka

Ethics Statement

This is to certify that the thesis titled "Pathway analysis of cancer-driving gene network in the human breast cancer" has been carried out through computer technology bioinformatics. So no need to be approved by the Ethical Committee.

Abstract

The research of human genes and diseases is very interrelated and can lead to an improvement in healthcare, disease diagnostics, and drug discovery. In this work, a system was established to construct similarity measures of gene pair mutation for human breast cancer and then performed network analysis to identify disease-related genes. The overlapping position of the interacting genes was used to calculate their similarity co-efficient. Using this similarity coefficient the co-occur genes were analyzed and built up a network of the gene cluster. Finally, a significant pathway was detected which was followed by the genes in a cluster. In this study, the process of constructing the gene regulatory networkin breast cancer was refined. A network topography for measuring gene-pair mutation similarity had been taken using their position where they overlap to induce a significantly mutated network. We aim to evaluate whether the identified network can be used as a biomarkerfor predicting breast cancer patient endurance. Common genes were estimated for different cancer types (i.e. lung cancer, prostate cancer, and breast cancer) from Gene bank. On a breast cancer case study, the system predicted an average 80% breast-related genes. These common genes were matched with reference breast cancer genes from clinical data in cBioPortal. Using the position of the gene pair in the genome similarity coefficient was measured. After that gene clusters were detected using similarity score. Finally, we identified the JAK-STAT signaling pathway in which clustered genes were enriched. It was found that 3 out of 4 datasets contained the MTOR NEDD9 EPOR gene cluster. This gene cluster followed the JAK-STAT signaling KEGG pathway. The JAK-STAT pathway played a vital role in cytokine-mediated immune responses, mainly cytokine receptors and they were able to polarize T-helper cells.

Other gene clusters were SMAD4 PDGFRA KIT EGFR KDR ERBB3 and ERCC2 COL18A1 ERCC1. They followed the MAPK signaling pathway and the nucleotide excision repair pathway. Dysregulations in both pathways played a vital role in various cancer development.

Our research showed that this study has the potential to identify disease-gene associated networks as a biological marker that may be useful to breast cancer patients for selecting the finest treatment. These common genes can be found in different cancers so that we can compare our work in case of other (lung cancer and prostate cancer) cancer types.

Keywords: Similarity coefficient; overlapping; cancer-driving mutated network; network topography; JAK-STAT pathway.

Acknowledgment

At first, I would like to thank Almighty. I went through so many difficulties in my life. Without his help and blessings, nothing was possible. My prayers were granted and he made everything easy in the period of my study.

I would like to convey my indebtedness to Professor Dr. A F M Yusuf Haider, Chairperson, MNS department for allowing me to pursue my postgraduate studies in the department of MNS and for his constant guidance and help throughout my entire period of study in the department.

I am overwhelmed to express my respect, utmost gratitude and deepest thanks to my supervisor Rasif ajwad.Lecturer, Department of Computer Science & Engineering, BRAC University, Dhaka for his continuous guidance and inspiration. He is an amazing supervisor. It was never easy to accomplish my research work without his endless support, constructive criticism and encouragement.

I would like to convey the deepest thank to Dr. Mahboob Hossain. Professor, Microbiology Discipline, Department of Mathematics and Natural Sciences, BRAC University, Dhaka who provided me valuable time with his constant knowledge, sound advice and for his inspiration. He gave me the most privilege whenever I faced any problem or had an inquiry about my work. He consistently gave me important knowledge that is related to my work and always gave me the right guidance whenever I need. My words are very little to commend his contribution throughout my student life.

I need to express my deepest gratitude to my parents and my husband for keeping faith in me and also giving boundless support and spontaneous inspiration during the period of my thesis work. It would not have been possible to accomplish my work without their support.

The Author
Department of Mathematics and
Natural Sciences
BRAC University, January 2020.

Contents

List of Figure

List of Table

List of Abbreviations

CNV	Copy Number Variation
HR	Hormone Receptor
DNA	Deoxyribonucleic Acid
ER	Estrogen Receptor
NSCLC	Non-Small Cell Lung Cancer
SCLC	Small Cell Lung Cancer
mRNA	Messenger Ribonucleic Acid
NGS	Next Generation Sequencing
PR	Progesterone Receptor
GO	Gene Ontology
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variants
BP	Biological Process
MF	Molecular Function
BIC	Breast Invasive Carcinoma
KEGG	Kyotos Encyclopedia Genes and Genome
JAK-STAT	Janus Kinase-Signaling Transducer and Activator of Transcription
MAPK	Mitogen-Activated Protein Kinase
NER	Nucleotide Excision Repair

Chapter 1

Background and Introduction

Background and Introduction

1.1. Gene

Genes are comprised of a sequence of nucleotides in DNA or RNA. Genes encode molecules to make protein. Many of the genes are not responsible for protein-coding. The size of genes varies depending on the DNA bases. The human Genome project defines that humans have between 20,000 and 25000 genes. Every person carries two sets of each gene. One is from a mother and the other is from the father. Most of the genes are the same but few numbers of genes vary among people. There is the same gene with small differences in their nucleotide sequences which form allele. These differences make different functional features among people (Jonathan et al., 2011).

Genes have two regions including promoter regions and alternating regions. The alternation region includes introns (noncoding sequence) and exons (coding sequence). The genes are transcribed from DNA to RNA to produce a functional protein. There is a specific method where introns are removed and exons are spliced together. Then pieces of RNA sequences are translated into an amino acid chain(Jonathan et al., 2011).

In human genes are located inside the cell nucleus. Even a few genes are found in the mitochondria which differ from the genes found in the nucleus.

So, a gene is a small part of DNA that give instructions for protein molecule and also act as information storage. They provide information for individual characteristics, for example, eye and hair color which differ individually.

1.2. Chemical structure of the gene

Genes are made up of Deoxyribonucleic acid (DNA), some viruses are the exception. They have genes consists of ribonucleic acid (RNA). DNA molecule is a double helix that is comprised of two chains of nucleotides in a twisted form. Each chain is composed of sugars and phosphates. These two chains are in twisted form with nitrogenous bases. These are adenine (A), guanine (G), cytosine (C) and thymine (T). Here adenine (A) makes a bond with thymine (T). Similarly, cytosine (C) makes a bond with guanine (G). Through this DNA molecule, the information is passed from one generation to the next one.

In this method, bonds are broken inside twisted DNA and the chain becomes unwind and also creates a free nucleotide line. This process makes two identical DNA molecules from one original (Jonathan et al., 2011).

1.3. Gene mutation

When the number of nucleotide bases of the gene is interrupted mutations occur. Nucleotides in DNA sequence can be duplicated, deleted, rearranged or replaced. There is a particular effect for each alternate. When mutations occur small or no effect can be found. Because of mutation, great change happens which causes diseases like cancer. Mutation can happen from a small single DNA building block to a large section of a chromosome that contains genes. Each cell has to depend on proteins to accelerate its functions. Gene mutation prevents protein function. If this protein function has the importance for the human body it will disrupt the normal function of the cell and causes disease (Jonathan et al, 2011).

1.4. Genetic Variations

Genetic variation refers to a diversification of DNA nucleotide sequences inside genes. It is defined as the difference in nucleotide sequences between individuals or between populations.Genetic Variation can be found in germ cells i.e. sperm and egg, and also in somatic (all other) cells. The particular source of somatic genetic variation is mutation. Single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) are responsible for the variations where copy numbers in a specific DNA region are responsible for cancer development in somatic genetic mutations. Genetic variation is occurred by a mutation which is a continual diversity in the chemical structure of chromosomes in the form of gain or loss in copies of DNA segments. As a result, abnormal proteins are created or prevent protein formation and causes uncontrollable growth of cancer cells.

For example, four different nucleotides build up all the DNAs inside the human genome. They are Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). Variation occurs when the number of DNA copies resembles the deletion or duplication of one or more segments in the DNA. If we consider a DNA segment that normally has GCAATCG, this might have a variation such as GCAATAATCG (a duplication of "AAT") or GCCG (a deletion of “AAT”) (Jonathan et al., 2011).

1.5. Single Nucleotide Polymorphism

One of the most common types of genetic variations is Single nucleotide polymorphism, (SNPs) in the DNA sequence of the human genome. Every SNP comprises nucleotide which is a different building block. For example, an SNP may replace the nucleotide adenine (A) with the nucleotide guanine (G) in that specific position in a certain sequence of DNA.

SNPs can be found in every 200-300 nucleotides on averagethat means there are about 10 million SNPs in the human genome. They are commonly placed in the DNA sequence of genes. Genes that are responsible for a particular type of disease scientists are now able to identify those genes using the SNPs as a biological marker. When a gene can not proceed with its function and somewhat develop a disease, SNPs are present inside a gene or in the regulatory arena near a gene.

As the majority of variations are not engaged to alter cellular function and thus have no effect to create disease, some SNPs have been revealed to contribute to cancer development. SNPs can influence people with great diversity in common traits, such as eye color, hair, etc. In the specific region of DNA, SNP can be found which acts as chromosomal tags. These regions are the specific area where a human disease or disorder can be immersed. These disease-related SNPs are useful for diagnostic purposes. Even drug development is possible by identifying variations that are involved in the disease. This approach will be most appreciated by doing genetic screening for specific SNP in the human genome which can be used to select particular drugs. (Ajwad, 2017).

1.6. Copy Number Variations

Genetic variations are multiple types that consist of human variability. Because of these genetic variations, DNA nucleotide sequence and chromosome structure change SNPs are common that are frequently observed in the DNA sequence of genes. The human genome is comprised of two sets of 23 chromosomes, one set comes with mother and another set comes with father. But recent analysis has explained that a large sequence of DNA bases can differ in copy number. This copy number variations can either be deleted or duplicated in the arrangement of genes on chromosomes. The genes that are always can be found in two copies per genome, now have been observed in one, three or more than three copies. Often genes are altogether disappeared in a specific region of DNA. These changes are indications of different diseases such as Autism, Alzheimer's, Thalassemia, Parkinson's disease, etc.

All mutations are not engaged in ruining the human cells. Only a few driver mutations can cause disease development based on the location of the mutation in the genes.

Identification of driver mutation is the main challenge in cancer genomics research. Different methods (next-generation sequencing and microarray techniques) are used to identify driver mutations. Still, computational methods to analyze the huge amount of CNV data need to develop to accelerate the data (Ajwad, 2017).

1.7. Chromosomal Rearrangement

A type of mutation that occurs in the chromosomes and produces chromosomal abnormality by changing the structure of the native chromosomes. The changes vary. There are different changes such as deletion, duplication, inversion, and translocation.

Maximum chromosomal rearrangements happen during fertilization. When the chromosome pair breaks during the recombination process and does not repair the breaks then pieces of chromosomes may be rearranged. If these chromosomes participate in fertilization, the embryo result in extra, missing or rearranged chromosomal pieces. As the baby grows the cell division occurs and then rearrangement is transferred to all or some of the baby's cells. This type of chromosomal rearrangement is the source of genetic disease and cancer (Louise et al., 2014).

Chapter 2

Literature review

Literature review

2.1. Cancer

The human body is made of cells that grow and divide. According to these cells body function varies. Every cell has a specific function and also its life cycle. After some time, old cells are replaced by new cells. This process is damaged by cancer and causes abnormal growth of cells. (Maira et al.,2006).

DNA mutation or changes are responsible for this abnormality. Every cell contains a specific gene that holds DNA. So cell functions are regulated by the instructions of the gene. When DNA mutation disrupts then cells divide and grow abnormally and can lead to cancer. Human death is gradually increasing due to cancer.

Cancer can be found when genetic changes can not proceed systematically. Then cells start to grow abnormally. As a result, a tumor may be formed by these cells. A tumor can be two types that are cancerous and benign. A malignant tumor is cancerous, it means, this type of tumor can be developed and expanded to any part of the body. But if the tumor only grows and does not spread which is called a benign tumor. (Maira et al., 2006).

2.2. Type of Cancer

Four types of cancers are classified based on where it starts.

Carcinomas: Mainly carcinoma is identified in the surface of internal organs and glands. Carcinomas usually contain solid tumors. Carcinomas are prostate cancer, breast cancer, lung cancer, and colorectal cancer.

Sarcomas: Sarcoma can be found in fat, muscles, nerves, tendons, joints, blood vessels, lymph vessels, cartilage, or bone.

Leukemias: Blood cancer is called Leukemia. There are four types of leukemia. These areacute lymphocytic leukemia, chronic lymphocytic leukemia, acute myeloid leukemia, and chronic myeloid leukemia.

Lymphomas: This type of cancer can be developed in the lymphatic system. The lymphatic system is comprised of vessels and glands that protect the body from infection. There are two types of lymphomas: Hodgkin lymphoma and non-Hodgkin lymphoma (The Healthline Editorial Team, 2016).

2.3. Carcinomas

Cancers are various types and the most common type is carcinoma. Carcinoma starts from the tissue or cell that covers the organ. For example, prostate, lung, liver, kidney, breast, etc.

Carcinomas are also different type and occur in many parts of the body. These are basal cell carcinoma, squamous cell carcinoma, renal cell carcinoma, ductal carcinoma in situ, invasive ductal carcinoma and adenocarcinoma (Laura et al., 2018).

2.4. Lung cancer

Lung cancer is also known as lung carcinoma. Based on the size and injured cells there are two types of lung cancer are non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). The lung cancer has many stages depends on spreading from the lungs to lymph nodes or other organs. The lung is a large organ. Identification of the early stage of lung cancer is difficult. It can grow a tumor for a long time before they detect it. When symptoms occur like coughing and fatigue most people ignore thinking because of another cause. That's why difficult to find out lung cancer in an early stage(Laura et al., 2018).

Three types of lung cancer have been identified. 85% of lung cancer is adenocarcinoma occur in both men and women. 25% of lung cancer is squamous carcinoma and 10% is large cell carcinoma.Even NSCLC has four stages based on the spread area of cancer. Remaining 15% of lung cancer is small cell lung cancer (SCLC). They grow fast than NSCLC tumors. SCLC has two stages limited stage and extensive stage(William et al., 2000).

If there are any genetic mutation which causes cancer this genetic information helps to predict cancer. Based on this genetic information treatment can be specified and medicine also can be personalized.

The most frequently mutated genes in NSCLC are TP53, KRAS, EGFR, ALK, CDKN2A, and STK11 (Greulich, 2010). When a cancer cell is identified in the lungs is called primary

lung cancer. Sometimes cancer from other parts of the body spread or metastasize to the lung is called secondary lung cancer. For example, many of the genes of breast cancer have been identified in the lungs which do not indicate lung cancer and need to prescribe treatment for breast cancer, not for lung cancer.

2.5. Prostate cancer

Prostate cancer is identified when cells in the prostate gland begin to grow abnormally. Prostate cancer is developed when the growth of the cells in the prostate gland are uncontrolled. Annually 1,600,000 cases and 366,000 deaths for prostate cancer have been reported as it is the most common noncutaneous cancer in men worldwide(Torre et al. 2015). In spite of the recent progress, men are significantly affected by this cancer, with mistreatment of inherently benign disease and incompetent therapies for metastatic prostate cancer (Guocan et al., 2018). The rates of clinical prostate cancer in the Asian population are lowest. In the Hispanic Caucasian population, this cancer has intermediate rates and the highest rates are identified in African American populations. This proposed that environmental factors play a vital role in prostate cancer (Rebbeck, 2017).In different levels of genetic and epigenetic changes arise. Genetic modifications have suggested its use as biomarkers, especially for breast and ovarian cancer. To conclude the high risk of the development of these disease mutation in BRCA1 and BRCA2 have been identified. RNase L (HPC1, 1q22), MSR1 (8p), ELAC2/ HPC2 (17p11) genes are used as biomarkers for prostate cancer (Joke et al., 2010).

2.6. Breast Cancer

2.6.1.Genetic Changes in Breast Cancer

Genes are found in chromosomes which are small segments of DNA. They carry the instructions for functional proteins. The structure and function of all the cells are maintained by these proteins in the human body. When these genes are mutated, they damage the cell functionalities. That causes uncontrollable cell growth and division which leads to cancer formation. So cancer is the uncontrolled growth of abnormal cells which develops the complex human disease.

The cells conduct a series of genetic and epigenetic alteration which causes modification of gene activities of several genes, Al these changes occur in the single somatic cell and are also known as somatic mutations. Due to the involvement of genetic factors, genetic mutations can be found in all cells which can lead to the risk of breast cancer. These genetic changes are defined as germline mutations. The difference in somatic mutation and germline mutation is that somatic mutation occurs in the single body cell and can't be inherited whereas germline mutation occurs in gametes and can be inherited from parents (Rizzolo et al., 2011).

There are various risk factors for breast cancer among them lifestyle and hormonal factors are commonly identified. About 15–20% of breast cancer is familial and the affected women with this disease are relatives. The heritable elements in these families are consequential, mostly the women who are with low age are affected by breast cancer (Ajwad, 2017).

2.6.2.Identifying breast cancer susceptibility genes

Breast cancer represents 25% of all cancer. Studies show that the percentage of breast cancer is inherited by about 15-20%. The susceptibility genes for breast cancer are classified into three groups based on their mutation frequency and position in the human genome.

1. High penetrance group
2. Moderate penetrance group and
3. Low penetrance group

BRCA1 and BRCA2 are both high-risk susceptible genes for breast cancer. A few numbers of genes represent moderate to high risk for breast cancer. TP53 is a germline mutation and tumor suppressor gene has been discovered in cancer families which is responsible for Li-Fraumeni syndrome (LFS). But it has been found as the most common dormant form. The moderate-risk genes are CHEK2, ATM & PALB2. SQL is the new breast cancer susceptible to a moderate risk gene. High-risk genes are responsible for DNA damage and other pathways. The function of low-risk genes is uncertain but they are essential for gene expression.Protein formation from these genes is directly related to cell formation and DNA damage repair which means a mutation in these genes can lead to uncontrollable cell growth (Camilla et al.,2019).

So, in total high-risk genes are BRCA1, BRCA2, TP53, STK11, CD1, and PTEN identified for approximately 20% of the inherited breast cancer risk. Moderate risk genes are identified about up to 5% of the inherited familial breast cancer risk. Low-risk genes are more than 180 which are identified as 18% of the familial risk. But still, most of the genetic framework of familial breast cancer is uncertain (Camilla et al.,2019).

2.6.3.Type of Breast cancer

In many cases, female hormones are responsible for some breast cancers. The hormones are estrogen and progesterone. There are different receptors in breast cancer cells. Depending on the cancer cells respond to different receptors can catch specific hormones. From this view, four different types of breast cancer are identified.

1. Estrogen receptor (ER) positive: In this breast cancer, the cells are positive to estrogen hormone. These cells have the receptor which allows them to grow estrogen hormone. Endocrine hormone therapy can suppress cancer cell growth.
2. Progesterone receptor (PR) positive: The cancer cells are sensitive to progesterone. Receptors in the cancer cells grant them to grow in response to the progesterone hormone. The growth of cancer cells can be inhibited by doing treatment with endocrine therapy.
3. Hormone receptor (HR) negative: In this cancer type cells do not have hormone receptors. So it can't be treated with endocrine therapy but the aim is blocking hormones in the body (Ajwad, 2017).

2.7. Mutation Analysis Approaches

There are so many advantages of next-generation sequencing that has been widely used in genomic research. The traditional sequencing methods (such as Sanger sequencing method)are replaced by next-generation sequencing, and many complex disease research is possible using this method. So this technology has the advantages of high speed, high throughput and high efficiency in breast cancer. It has been massively used in different cancers (such as prostate

cancer, lung cancer, pancreatic cancer, liver cancer, etc.), especially in breast cancer. In this technology, constructing "libraries" of short DNA fragments which are then sequenced. Using this technology genetic alteration can be identified(Hahn et al., 2002). One of the common approaches to analyze the genomic alterations is to look for recurrently mutated genes, which are mutated in a large group of the patients. The main idea behind this approach is that if the genes that are mutated in a large fraction of patients can be identified, they will correspond to non-random mutations. However, this approach is challenging. Although some cancer genes have higher mutation frequency rates (e.g., TP53) than non-cancerous genes, most of them have much lower mutation frequencies (Vandin et al., 2010). Instead of a single gene, driver mutations can target groups of genes in networks or pathways (Vogelstein et al., 2004).

Current network-based approaches have one key limitation is that they can't assign the same mutation genes into different networks although overlapped networks are possible. Because of this reason we are applying two network-based clustering algorithms to analyze breast cancer genomic data for identifying cancer-driving mutated networks. The approach is called ClusterOne (Nepusz et al., 2012). In this study a new mutation analysis method was designed by taking network topography into account for measuring gene pair's mutation similarity to infer cancer-driving mutated subnetworks.

2.8. Motivation and Research Objectives

2.8.1. Motivation

Breast cancer diagnosis based on gene expression profile is an important program to illustrate patient care. Genes are connected in a network that encodes protein functions or pathway information. In this article, we develop a bioinformatics analysis method where common genes are collected from different cancers (lung cancer, breast cancer, and prostate cancer) to identify the disease-gene associatednetwork which follows a specific pathway.

2.8.2. Hypothesis

We hypothesize a disease-gene associated network with highly mutated genes which are collected from different cancer types (lung cancer, breast cancer, and prostate cancer). These genes are predicted as 80% of breast-related genes. Common genes are also matched with reference genes from clinical data of breast cancer patients. Using their overlapping position has to identify the similar coefficient of gene pair which will help to develop a network of highly mutated genes. Finally, we will establish a significant pathway of that developed gene network. This approach will lead to the network-based biomarkers for breast cancer survivors that may be useful to choose the finest treatment for cancer survival patients. Thisdisease-gene associatednetwork can be compared in the case of other (lung cancer and prostate cancer) cancer data analysis.

2.8.3. Research Objectives

At first, we will refine the top-ranked genes for different types (lung, prostate and breast cancer) of cancers. On a breast cancer case study, the system predicts 80% breast-related genes. These genes will be matched with reference clinical data of breast cancer patients. Secondly, we will investigate the disease - gene relationship based on the similarity coefficient of gene pair. We will present a computational method to create a network with highly mutated genes and their significant pathway. This network will be used as a biomarker for breast cancer survival patients.

Chapter 3

Materials and Methods

Materials and Methods

3.1. Preprocess of data extraction

3.1.1. Gene bank

Gene bank was a database that provides all the information on the DNA sequence. This information was the most up-to-date. From gene bank, we retrieved the total number of established genes in lung cancer 2880, breast cancer 4153 and prostate cancer is 2775. Then we collected the common genes among these cancer types. The number of common genes was 1097. That means these genes were present in the lung, breast, and prostate simultaneously. To build up a network we collected information from these genes were gene ID, symbol, start point, endpoint, and frequency.

3.1.2. cBioPortal

cBioPortalwas a cancer genomics software from where we collected genes as reference causing breast cancer from different datasets. These were (Breast Invasive Carcinoma (Koboldt et al., Nature 2012), Breast Cancer (Razaviet al., 2018), Breast Invasive Carcinoma (Hoadleyet al.,2018. Ellrott et al.,2018. Taylor et al.,2018. Gao et al., 2018. Liu et al.,2018. Sanchez-Vega et al., 2018.) andBreast Invasive Carcinoma (Ciriello et al., 2015)). These datasets have been described below.

Breast Invasive Carcinoma (Koboldt et al., Nature 2012)

The project was Breast Invasive Carcinoma and the cases were 825. To collect the mutated genes from these sample several methods had been done. The methods werewhole-exome sequencing (510 samples with matched normal), genomic DNA copy number arrays, DNA methylation, messenger RNA arrays, microRNA sequencing, and reverse-phase protein arrays analysis. The number of mutated genes was 507.

Breast Cancer (Razavi et al., 2018)

Breast cancer mutated genes were 1918 and the patient number was 1756. These were the highly mutated genes engaged with the mitogen-activated protein kinase (MAPK) pathway and in the estrogen receptor transcriptional machinery. In endocrine-resistant tumors activating ERBB2 mutations and NF1 loss-of-function mutations were more than twice. Mutations in other MAPK pathway genes (EGFR, KRAS, among others) and estrogen receptor transcriptional regulators (MYC, CTCF, FOXA1, and TBX3) were also enriched. Altogether, these mutations were present in 22% of tumors, ESR1 mutations are exclusive, and correlate with a shorter duration of response to consecutive hormonal therapies.

Breast Invasive Carcinoma (Hoadley et al.,2018. Ellrott et al.,2018. Taylor et al.,2018. Gao et al., 2018. Liu et al.,2018. Sanchez-Vega et al., 2018.)

The data were collected from different papers where the number of patients was 1084 and the samples of mutated genes were 1066.

Breast Invasive Carcinoma (Ciriello et al., 2015)

The project wasThe Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma. The number of patients was 816 and the number of mutated gene samples from breast tumor of Invasive lobular carcinoma was 817. It was the second-highest prevalent subtype of Breast Invasive Carcinoma.

3.2. Data extraction

The official list of genes were downloaded to develop a cancer-driving mutated network in breast cancer. The common genes of different cancers (lung, breast and prostate cancer) were required from gene bank. The number of total genes was in lung cancer (2880), prostate cancer (2775) and breast cancer (4153). Then we collected the common genes among these cancer types. The number of common genes was (1097). That means these genes were present in the lung, breast, and prostate simultaneously. In this work, we also collected genes as reference causing breast cancer from different datasets. These were Breast Invasive Carcinoma (Koboldtet al., Nature 2012), Breast Cancer (Razavi et al., 2018), Breast Invasive Carcinoma (Hoadley et al.,2018. Elliott et al.,2018. Taylor et al., 2018. Gao et al., 2018. Liu et al.2018. Sanchez-Vega et al., 2018.) andBreast Invasive Carcinoma (Ciriello et al., 2015) using cBioPortal. Then matched common genes with reference genes from different datasets. Finally, we found four different datasets. The total number of matched genes in different datasets was 546, 197, 841 and 752.

3.3. Method

To build up a disease-gene relationshipnetwork we established a method that has been given below in **Figure 1.** Concisely speaking, at first, matched genes and their information were collected those were gene ID, symbol, start point, endpoint, and frequency. Then we identified the overlapping genes calculating the start and endpoint of the genes. As we have already established the frequency of these matched genes, we can easily calculate gene pair specific mutation similarity scores. To use these scores we created a disease-gene network from ClusterONE. Many gene clusters were identified. Finally, we analyzed the pathway of the gene clusters according to the lowest p-value and found remarkable results.

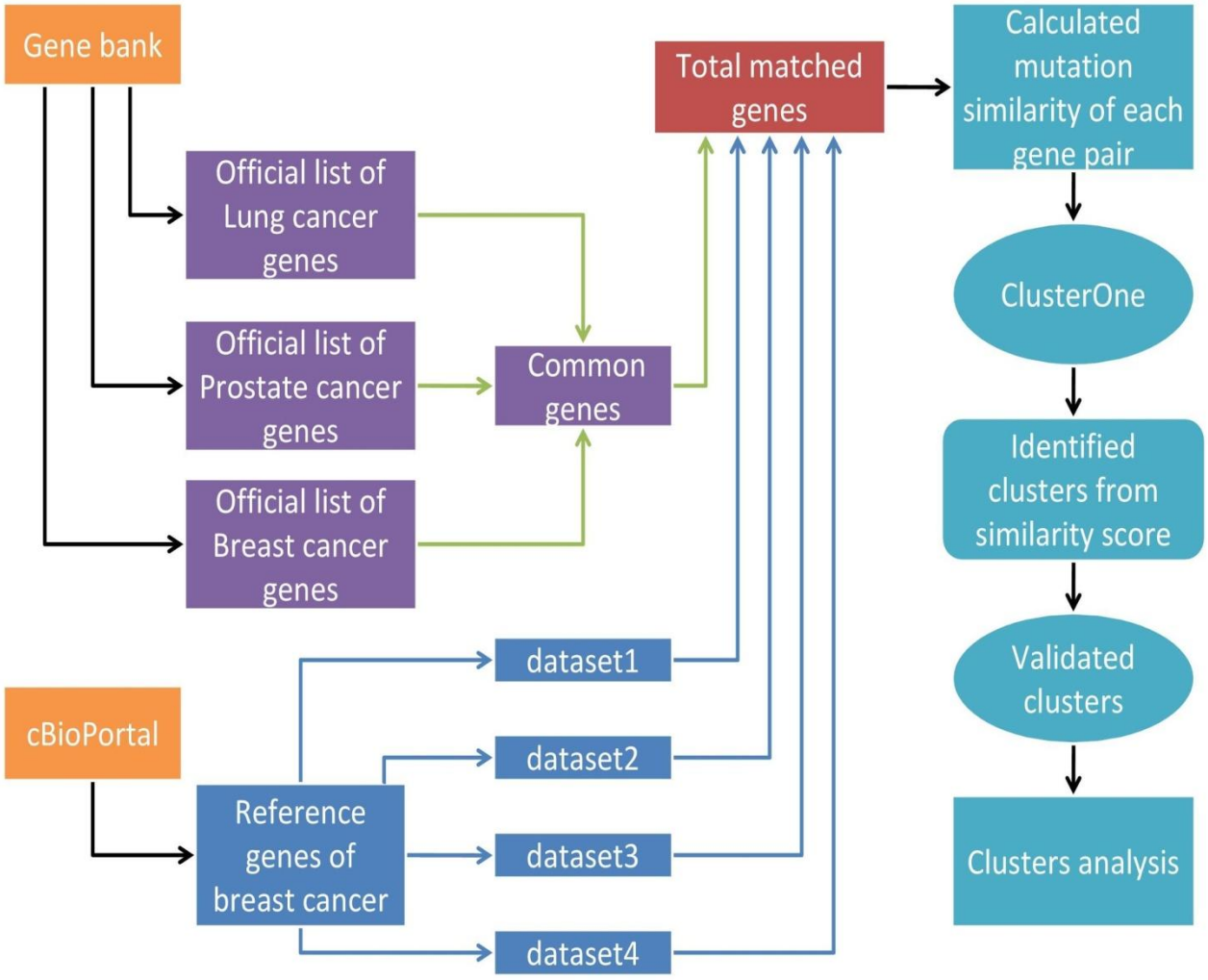


Figure 1: A way of analysis to identifythe disease-gene associatednetwork.

3.4. Overlapping genes

Overlapping genes are defined as chromosomal locations of two adjacent gene loci overlapping partially or entirely with each other by sharing a common genomic region(Chao-Hsin, 2019).

In the case of two overlapping genes, different proteins are involved in the overlap events. DNA codes for the amino acids in their overlapping portion. These overlapping events can be noticed in viruses, prokaryotes, and also in eukaryotes (Maxime, 2014). Using genome-wide annotations of genes and mRNA features human overlapping genes are identified and analyzed in a specific way. More than 10% of human genes are engaged in gene overlap incidents (Meng-Ru et al., 2012). Even many overlapping genes are responsible for the infection.

Both strands of the human genome are used for transcription. Two types of overlapping events happen; two genes overlapping on the same strand, and two genes overlapping on opposing strands. Furthermore, using the relative positions of the two genes, overlapping genes patterns can be classified (Tomohiro et al., 2007).

3.5. Identification of overlapping genes

In this study, genes of breast cancer were retrieved from Gene bank. These genes have been already established with the start point and endpoint in Gene bank. After getting matched genes we identified the start point and endpoint for each matched genes. From these points, we can easily identify overlapping genes. For example, if the start point of g2 remains in between the start point and endpoint of g1, then these two genes overlap with each other. In this way, two or more genes can share the same region when overlap. Using database overlapping genes were

identified by matching their start point whichremains in between the start and endpoint of that next before gene.

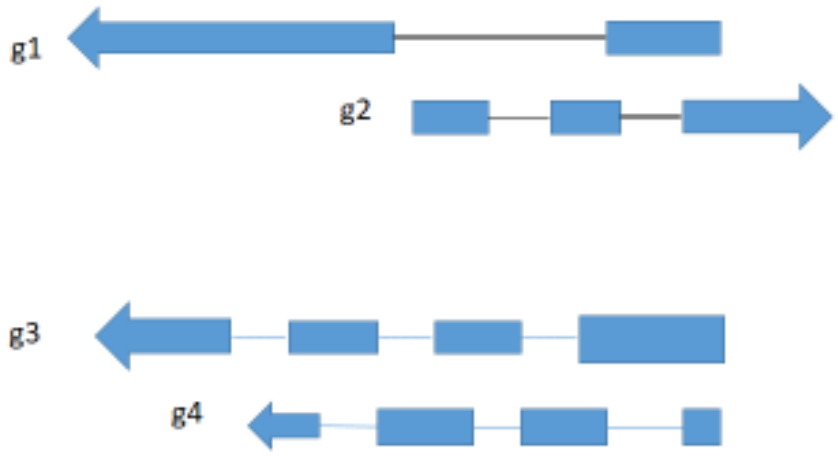


Figure 2: Overlapping genes. Genes sharing the same locus but coding for different proteins. Start point of g2 lies in between start and end point of g1 and overlap with each other.

3.6.Gene-specific mutation frequency

The next step was to find out the frequency of gene-specific mutation. As the frequency of the genes was given in gene bank data. The desired frequency of the genes was directly retrieved from gene bank data. Using this frequency gene pair specific mutation similarity score could be calculated.

3.7. Calculation of gene-pair specific mutation similarity Score

Using gene mutation frequencies retrieved from the gene bank similarity score of the gene pair was calculated to build up a disease-related gene network. We applied this similarity score to create the network. Each pair similarity score is used as an input to ClusterONE, which has been described in Section 3.1.9. The authors said that it is the finest approach to get improved results as the ClusterONE algorithm shows a constant weight (=1) for each interaction in the network (Nepusz et al., 2012).

For the calculation, we used the similarity coefficient formula (IBM Knowledge Center, 2012) to measure gene pair mutation similarity score:

$$\text{Similarity Coefficient} = \frac{(C_{IJ})2}{(C_I \times C_J)} \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where,

C_I is the frequency of g1 or the records in which the mutation I occur.

C_J is the frequency of g2 or the records in which the mutation J occurs.

C_{IJ} is the frequency of g1 and g2 or the records in which both g1 and g2 mutations I and J co-occur.

For this work, we applied the gene-pair the similarity formula as a network edge weight. We have treated the gene similarity score to the edges of the network.

3.8. Identification of disease-gene associated networks

We have used the ClusterONE algorithm to identify disease-gene associated network from the similarity score of a gene pair for all the interactions. ClusterONE is an algorithm that detects clusters or groups of overlapping genes in the gene interaction network (Nepusz et al., 2012). We briefly describe this algorithm as follows.

ClusterONE: It is an algorithm that is used to build up a gene interaction network with a group of genes carrying high cohesiveness. In the author's statement, each group defines cohesiveness making a protein complex that has two structural properties of a subgraph: the interaction between its subunits and the separation from the rest of the network. For a group of proteins$P$, the cohesiveness (P)is defined by:

$$(P) = \frac{total\ \ weig\,ht\ for\ \ internal\ \ \ edges}{total\ \ weig\,ht\ for\ \ internal\ \ \ edges\ + total\ \ weig\,ht\ for\ \ boundary\ \ \ edges + P} \quad \ldots\ldots\ldots(2)$$

There are two terms have been which defined the edges of the network. 'Internal edges' are used to define the given group and 'boundary edges' have a connection with the rest of the network. The constant □ is a term that clarifies the uncertainty in the data.

This algorithm mainly consists of three steps. First, as the first seed protein has to be selected with the highest degree by following a greedy approach and begins to make a cohesive group from the primary seed. While choosing the next seed, the algorithm considers all the proteins that are not available in any other network (protein complexes). After that highest degree node is selected again. This process goes on until no proteins left to consider. If the cohesive group creates any vertex during the growth process it can be removed if necessary. So vertex will not be seen in any of the clusters or groups.

The next step is, according to the overlapping score pair of groups are marked. The authors merge when overlap score (ω) more than 0.3. For two protein sets X and Y, the author defined ω as:

$$\omega(X,Y)=|X\cap Y|2|\ X\ ||\ Y|\ \ldots\ldots\ (3)$$

According to the author's statement, a pair of groups must be merged one after another or spontaneously. In this first approach, the problem is overlapping score needs to be recalculated after each merging happens in each interaction. To ignore this problem, ClusterONE uses the concurrent approach. ClusterONE constructs an overlap graph from the collection of cohesive groups. In the graph, each node represents a cohesive group and two nodes are connected if the overlap score is larger than the given value ($\omega \geq 0.3$). Nodes can be directly (one-to-one) or indirectly (via paths of edges) promoted to protein complex candidate. If no edges exist in the node then merging is not possible.

In the final step, the algorithm gets rid of complexes that have a density below a specific value. In our study, we consider the overlapping gene pairs and the weight of each gene pair is based on the gene-pair mutation similarity coefficient.

3.9. Parameters used to run ClusterONE

To run ClusterONE, we only need a similarity score of gene pairs. No need to use any additional parameters.

3.10. Pathway analysis

After identification of the gene network of breast cancer pathway analysis using the gene list of that network was implemented. We used the Enrichr (Kuleshov et al., 2016) software to detect gene ontologies (biological processes and molecular functions) and biological pathways (KEGG). More than 100 updated genes are available in the Enrichr software for analysis. This software is popular for the pathway enrichment analysis to identify the specific pathway in which the overlapped genes act. These overlapped genes are from ClusterONE which makes networks.

Pathway enrichment analysis encloses the knowledge of gene sets and these genes are involved in biological pathways or biological functions. The priority is to analyze the gene from the gene list that represents the gene sets. There is an enrichment score that quantifies the degree of representation (Subramanian et al., 2005). The analysis consists of three different parts. First, using the gene list (e.g. the whole human genome) for both input and background the enrichment score is calculated based on the number of gene representations from the gene set. Second, the statistical implication of the calculated score is predicted. Finally, the peak for each gene is detected after establishing the scores for each gene set.

Chapter 4

Results

Results

Thousands of genes are responsible for causing cancer types and it is impossible to find out any single gene in the diagnosis for cancer. Scientists discovered that every type of cancer was caused by different types of genes. There are huge variations in the genes. These variations play a vital role in cancer research. So it is not possible to explain any single gene for cancer diagnosis.

In this study at first 1097 common genes (Table 1) were retrieved from lung cancer, prostate cancer and breast cancer from Gene bank. All information on these genes is available in the gene bank. In this work, 80% of genes are related to breast cancer. Again we have found clinical datasets of breast cancer patients as reference data. We collected this data from cBioPortal. (Table 2).

After that, the common genes were matched with the genes of four clinical datasets of breast cancer patients and developed four matched genes datasets. These four datasets of matched genes are given below in the table form. (Table 3, Table 4, Table 5, Table 6).

These matched genes are already established with gene ID, start point, endpoint, and frequency as they were matched with the common genes.

According to the start point and endpoint, we found out overlapping genes in a pair for these four datasets individually. Using the equation of similarity coefficient we calculated similarity score of these gene pairs

Table 1.Common genes have been collected from lung cancer, prostate cancer, and breast cancer.

Common Gene						
ABCA1	B2M	CASP10	EEF2	HPSE	MARCKS	TNK2
ABCB1	BAD	CASP3	EFEMP1	HRAS	MB	TOP1
ABCC1	BAG1	CASP8	EGF	HSD17B1	MCAM	UBE3A
ABCC4	BAG3	CASP9	EGFR	HYAL1	MDM4	UCA1
ABCG2	BAK1	CAT	EGLN1	IGFBP7	NPRL2	ULK1
ABL1	BARD1	CAV1	EGR1	IKBKB	NQO1	VEGFD
ACE	BAX	CAVIN1	EIF2S1	IL10	NR1I2	VHL
ACHE	BBC3	CBL	FOXP3	IRS2	NUDT1	VIM
ACKR3	BCAR1	CBX5	FOXR2	ISG15	OLFM4	VTCN1
ACTA2	BCL11A	CBX7	FSCN1	ITGA2	P2RX7	WNT5A
ADAM10	BCL2	DKK2	FSTL1	JAK2	P2RY2	WT1
ADAM12	BCL2A1	DKK3	FYN	JUN	PBK	WWOX
ADAM15	BCL2L1	DLX4	FZD8	KCNMA1	RRM2	XRCC4
ADAM17	BCL3	DNMT1	GADD45A	KDR	RSF1	XRCC6
ADAM9	BHLHE41	DNMT3A	GAPDH	KEAP1	RUNX2	YAP1
ADAMTS1	BIRC2	DNMT3B	GAS5	KMT2D	S100P	YBX1
ADAR	BIRC3	DPP4	GATA2	LATS2	SAA1	YY1
ADIPOQ	BIRC5	DPYD	GLS	LCN2	SDC1	ZBTB33
ADRB2	BIRC7	DSC3	GNA12	LDHA	TBXT	ZBTB7A
AGER	BLM	DUSP1	GSTM3	LEF1	TCF7L2	ZEB1

Table 2. Clinical datasets of breast cancer patients as a reference from cBioPortal.

Link of Reference data
1. Breast Invasive Carcinoma (Koboldt et al., Nature 2012).
2. Breast Cancer (Razavi et al., 2018).
3. Breast Invasive Carcinoma (Hoadley et al.,2018. Ellrott et al.,2018. Taylor et al.,2018. Gao et al., 2018. Liu et al.,2018. Sanchez-Vega et al., 2018.) and
4. Breast Invasive Carcinoma (Ciriello et al., 2015).

Table 3. **Matched genes from common genes and reference genes (Dataset 1)**

GeneID	Matched Genes	start_point	end_point	Freq	newFreq
10273	STUB1	680410	682801	0.002	0.2
977	CD151	832952	838835	0.002	0.2
1855	DVL1	1335278	1349418	0.004	0.4
5176	SERPINF1	1761925	1777574	0.002	0.2
7468	NSD2	1871393	1982207	0.012	1.2
5590	PRKCZ	2050411	2185399	0.002	0.2
4521	NUDT1	2242222	2251145	0.002	0.2
2305	FOXM1	2857680	2877174	0.006	0.6
7161	TP73	3652516	3736201	0.002	0.2
1316	KLF6	3775996	3785281	0.002	0.2

In table 3 GeneID is fixed for a specific gene in a particular locus. Matched genes are the gene's symbol. Start point means from where the transcription of the gene begins and endpoint means where the transcription of the gene stopped. The frequency of a gene means allele frequency at a particular locus in a defined population and new frequency is the diverted form of frequency without percentage.

Table 4. Matched genes from common genes and reference genes (Dataset2)

GeneID	Matched Genes	start_point	end_point	Freq	new freq
3265	HRAS	532242	535576	0.004	0.4
6794	STK11	1205778	1228431	0.014	1.4
7015	TERT	1253148	1295068	0.015	1.5
2261	FGFR3	1793293	1808872	0.004	0.4
7468	NSD2	1871393	1982207	0.001	0.1
894	CCND2	4273762	4305353	0.002	0.2
3717	JAK2	4985086	5128183	0.009	0.9
29126	CD274	5450503	5470567	0.002	0.2
5879	RAC1	6374527	6403967	0.003	0.3
23081	KDM4C	6720863	7175648	<0.1%	#VALUE!

In table 4GeneID is fixed for a specific gene in a particular locus. Matched genes are the gene's symbol. Start point means from where the transcription of the gene begins and endpoint means where the transcription of the gene stopped. The frequency of a gene means allele frequency at a particular locus in a defined population and new frequency is the divertedform of frequencywithout percentage.

Table 5. Matched genes from common genes and reference genes (Dataset3)

GeneID	Matched Genes	start_point	end_point	Freq	new freq
23410	SIRT3	215030	236931	0.002	0.2
3265	HRAS	532242	535576	0.006	0.6
5176	SERPINF1	1761925	1777574	0.004	0.4
2261	FGFR3	1793293	1808872	0.002	0.2
7468	NSD2	1871393	1982207	0.007	0.7
5590	PRKCZ	2050411	2185399	0.004	0.4
4521	NUDT1	2242222	2251145	0.003	0.3
5434	POLR2E	1086574	1095392	0.002	0.2
6794	STK11	1205778	1228431	0.007	0.7
7015	TERT	1253148	1295068	0.006	0.6

In table 5 GeneID is fixed for a specific gene in a particular locus. Matched genes are the gene's symbol. Start point means from where the transcription of the gene begins and endpoint means where the transcription of the gene stopped. The frequency of a gene means allele frequency at a particular locus in a defined population and new frequency is the diverted form of frequency without percentage.

Table 6. Matched genes from common genes and reference genes (Dataset4)

GeneID	Matched Genes	start_point	end_point	Freq	new freq
23410	SIRT3	215030	236931	0.002	0.2
682	BSG	571283	583493	0.002	0.2
3665	IRF7	612555	615999	0.001	0.1
7298	TYMS	657653	673578	0.001	0.1
10273	STUB1	680410	682801	0.006	0.6
977	CD151	832952	838835	0.002	0.2
6794	STK11	1205778	1228431	0.002	0.2
7015	TERT	1253148	1295068	0.002	0.2
1855	DVL1	1335278	1349418	0.004	0.4
5176	SERPINF1	1761925	1777574	0.002	0.2

In table 6 GeneID is fixed for a specific gene in a particular locus. Matched genes are the gene's symbol. Start point means from where the transcription of the gene begins and endpoint means where the transcription of the gene stopped. The frequency of a gene means allele frequency at a particular locus in a defined population and new frequency is the diverted form of frequency without percentage.

.

4.1. Similarity score of gene pairs:

Table 7. The similarity score of gene pairs for dataset 1

g1	g2	Similarity Score
COL1A1	AKAP4	0.6
PRKDC	SMAD2	0.4
MTOR	NEDD9	0.7
NEDD9	EPOR	0.3
PROM1	NCOR1	0.4
HDAC1	BRCA2	0.5
BRCA1	RET	0.5
CDH1	MDM2	0.4
ITGB4	MYO6	0.4
AKT1	ABCA1	0.7

Following the similarity coefficient equation, we have to put these genes pair within-group name g1 and g2. g2 is the group contains the genes whose start point must be found in between the start and endpoint of the genes of group g1. After calculation, we can find a similarity score using the similarity coefficient formula.

Table 8. The similarity score of gene pairs for dataset 2

g1	g2	Similarity Score
RAC1	KDM4C	0.8
VHL	TYK2	0.7
KDM4C	INSR	0.9
ID3	KRAS	0.9
DDR1	BCL2L1	0.5
CEBPA	CCL2	0.8
CCL2	NFKBIA	0.9
MST1R	CEBPB	0.9
TIMP2	MSH3	0.8
PGR	TRPC6	0.8

Following the similarity coefficient equation, we have to put these genes pair within-group name g1 and g2. g2 is the group contains the genes whose start point must be found in between the start and endpoint of the genes of group g1. After calculation, we can find a similarity score using the similarity coefficient formula.

Table 9. The similarity score of gene pairs for dataset 3

g1	g2	Similarity Score
GNA12	SGTA	0.8
MMP26	AKR1C3	0.8
CD274	FSCN1	0.8
INSR	CLDN7	0.9
GDF15	NAT2	0.5
NAT2	LDHA	0.8
E2F3	RHOB	0.9
RHOB	APEX1	0.8
PEBP4	IL6	0.7
MAP3K8	HMGB1	0.8

Following the similarity coefficient equation, we have to put these genes pair within-group name g1 and g2. g2 is the group contains the genes whose start point must be found in between the start and endpoint of the genes of group g1. After calculation, we can find a similarity score using the similarity coefficient formula.

Table 10. The similarity score of gene pairs for dataset 4

g1	g2	Similarity Score
CTCF	GSTP1	0.6
GOLGA2	ETS1	0.6
JAK1	GNA13	0.5
LEPR	AXIN2	0.4
AKT1	ABCA1	0.4
DICER1	PON1	0.5
MTOR	NEDD9	0.7
NEDD9	EPOR	0.6
NCOR1	GRPR	0.5
CDH1	MDM2	0.5

Following the similarity coefficient equation, we have to put these genes pair within-group name g1 and g2. g2 is the group contains the genes whose start point must be found in between the start and endpoint of the genes of group g1. After the calculation, we can find a similarity score using the similarity coefficient formula.

4.2. Result from ClusterONE

Using similarity coefficient we run ClusterONE and found many complex gene clusters with high cohesiveness. Each cluster is a protein complex. According to datasets, we detected different clusters of genes for different datasets.

For Dataset 1, we detected only one complex gene cluster. MTOR, NEDD9, and EPOR make a complex gene network as they have high cohesiveness and their similarity score is not less than 0.3.

Table 11. Detected 1 complex gene cluster for dataset1

1. MTOR NEDD9 EPOR

Table 12. Detected 37 complexes gene clusters for dataset 2.

Gene Cluster	p-value from enrichr
1. HRAS STK11 TERT FGFR3	0.00002171
2. CCND2 JAK2 CD274 RAC1	0.00002171
3. INSR TP53 ERRFI1	0.00007701
4. AURKB DNMT1 VHL	0.007034
5. KEAP1 SMARCA4 MTOR	5.82E-07
6. RAF1 CDKN1B CALR	0.00004264
7. NOTCH3 BRD4 NCOR1 E2F3 LATS2 MAPK1 CDKN2A	2.78E-07
8. KRAS DNMT3A CDK8	0.0006617
9. TEK FLT1 CHEK2 ALK NF2	2.03E-04
10. MAPK3 TGFBR2 MDC1	0.00003825
11. BRCA2 WT1 DNMT3B TAP1 CEBPA	8.45E-04
12. NFKBIA IL7R CDKN1A MLH1 PIM1	1.28E-08
13. PIM1 SRC FOXA1 FGFR1	7.63E-05
14. SRC FOXA1 FGFR1 RICTOR ERBB2 AKT2 RARA FOXO1	3.59E-08
15. RAD51 TOP1 EP300 CTNNB1 AXL STAT3 BRCA1 RET TP53BP1	1.48E-06
16. B2M MUTYH ERCC2 EPAS1 BBC3 EPCAM MSH2 NCOA3 SMAD2	1.48E-06
17. MSH2 NCOA3 SMAD2 RB1 KMT2D RHOA	7.53E-08
18. SMAD4 PDGFRA KIT EGFR KDR ERBB3	**4.00E-09**
19. AURKA MAP3K1 GLI1 CDK4	4.26E-04
20. RAD51C JUN GNAS PPM1D	3.52E-05
21. MEN1 JAK1 EPHA5 AXIN2	2.85E-04
22. SMAD3 AR CTCF PIK3R1 CDH1 MDM2	7.53E-08
23. CCND1 MED12 SOX9 CD276 PAK1	3.30E-04
24. MSH3 HGF FASN	1.95E-03
25. PTEN EPHA3 NBN BLM SYK	4.08E-05
26. EPHA7 DICER1 IGF1R	7.33E-03
27. EIF4E TGFBR1 PGR	1.08E-02
28. YAP1 BIRC3 IGF1 ERCC5 AKT1 PIK3CG KLF4 ATM IRS2 FYN	3.69E-07
29. PTPN11 APC TCF7L2 MET	1.61E-06
30. ROS1 CBL GSK3B XIAP	7.26E-05
31. MYC GATA2 ABL1 TSC1 RXRA	3.32E-05
32. CXCR4 NOTCH1 TRAF2 PIK3CB BRAF INPP4B	7.53E-08
33. INPP4B EZH2 PDGFRB MCL1	1.31E-05
34. RHEB ESR1 DDR2	9.97E-05
35. FGFR4 NFE2L2 PIK3CA FLT4 SOX2 TP63 CASP8	2.58E-07
36. SOX2 TP63 CASP8 CTLA4 MDM4	1.27E-04
37. IDH1 ERBB4 BARD1 PARP1 IRS1 PDCD1 AKT3	1.55E-04

***The p-value means a binomial distribution and independence for the probability of any gene belonging to any set.p-value must be less than (≤0.05). The smaller the p-value and the pathway will be more significant.**

Table 13. Detected 59 complexes gene clusters for dataset3.

Gene Cluster	p-value from enrichr
1. STIM1 EEF2 ZBTB7A	0.01789
2. JAK2 MMP26 AKR1C3	0.003895
3. EPB41L3 CD274 FSCN1	0.02159
4. ENO2 INSR CLDN7	0.00007401
5. MTOR NEDD9 EPOR	0.0001946
6. ETV6 MTHFR CTSB	0.002997
7. NFIB PRKACA XPC	0.005241
8. CASP9 BMX TUSC3	0.007428
9. ABCC1 PROM1 NCOR1 GRPR EPHA2	0.0008445
10. SHMT1 SATB1 GDF15 NAT2 LDHA	0.002997
11. E2F3 RHOB APEX1	0.001249
12. ROCK1 LATS2 NDRG2 IFNB1	0.004942
13. IL6 MMP14 PRMT5	0.0007448
14. RARB KRAS DNMT3A UBE3A	0.00008960
15. ALK NF2 CCNE1 MVP	0.00006408
16. CPD MAP3K8 HMGB1 TGFBR2	0.004059
17. ZEB1 MICA MICB	0.001276
18. HDAC1 BRCA2 WT1	0.0001272
19. STIM1 EEF2 ZBTB7A	0.0002565
20. DNMT3B TIMP3 TAP1	0.0006617
21. NRP1 RXRB BAG1	0.005540
22. PARD3 HMGA1 CCL2 CAT	0.00005319
23. PPARD HMOX1 NFKBIA	0.00001851
24. ITGA9 POSTN FOXA1 LGALS1	0.01432
25. RICTOR APOBEC3B ADAM9	0.02263
26. CTNNB1 AXL UCHL1 SFRP1	0.0003682
27. IKBKB STAT3 POLB	0.00003211
28. GHR CEACAM1 YBX1	0.02410
29. **ERCC2 COL18A1 ERCC1**	**0.00001619**
30. PRKCE UBE2C MMP9 CD40	0.0005099
31. PRKDC SMAD2 VDR	0.0001139
32. COL1A1 CEBPB AKAP4	0.01225
33. SPAG9 KLK10 KLK11 SMAD4 KLK13	0.01787
34. PDGFRA CBX5 POLI HNRNPA1	0.01076
35. IL6ST CDK2 ERBB3	0.0002995
36. MAP3K1 TRIM25 REST	0.00003614

Gene Cluster	p-value from enrichr
37. PDE4D FHIT RPS6KB1 ANXA2 BCL11A PPM1D SIX1 CDK1	0.0002456
38. ESR2 BAD ESRRA	0.0003529
39. JAK1 GNA13 WIF1	0.002066
40. EPHA5 LEPR AXIN2 RELA	0.0003752
41. AR CTCF GSTP1	0.00007007
42. GLCE CCNB1 CDK7	0.00006962
43. NCOA2 FADD CTTN	0.0001139
44. KCNMA1 PAK1 PTPN12	0.008228
45. TK1 BIRC5 ANXA3 ROBO1	0.01017
46. TIMP2 PAQR3 NFATC1	0.01135
47. CDH13 XRCC4 HPSE	0.009719
48. ABCC4 DICER1 PON1	0.001949
49. DPYD MTDH LAPTM4B	0.006735
50. IGF1R RGMB EIF4E	0.002847
51. ALDH1A3 ACHE YWHAZ	0.00007401
52. NFKB1 TMSB15A MMP7	0.002548
53. APC IL1A AMOT	0.0003555
54. CADM1 NGF TRPS1	0.0001898
55. PVT1 ENG PTPRK RUVBL1	0.01493
56. RUVBL1 MKI67 GOLGA2 ETS1 GATA2	0.03123
57. MGAT5 PPP2CA RXRA	0.01713
58. ESR1 SPARC RORC MAGEA4	0.0001781
59. MAP3K1 TRIM25 REST	0.006186

***The p-value means a binomial distribution and independence for the probability of any gene belonging to any set.p-value must be less than (≤0.05). The smaller the p-value and the pathway will be more significant.**

Table 14. Detected 02 complexes gene clusters for dataset 4.

Gene Cluster	p-value from enrichr
1. MTOR NEDD9 EPOR	**0.0001946**
2. MSH2 NCOA3 TIMP1	0.002066

***The p-value means a binomial distribution and independence for the probability of any gene belonging to any set.p-value must be less than (≤0.05). The smaller the p-value and the pathway will be more significant.**

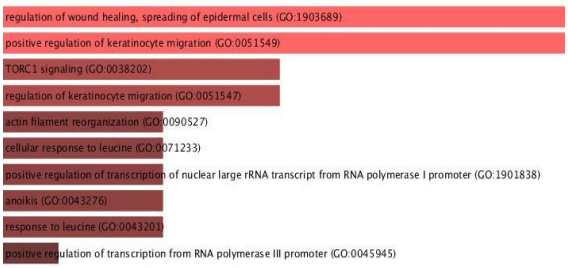
3 out of 4 datasets contain a common gene cluster that is MTOR NEDD9 EPOR. The clusters of these genes are associated with breast cancer. These clusters are based on the similarity score of the gene pair and the score value must not be less than 0.3.

Clusters consist of multiple genes with coordinated biological functions and/or correlated expressions.Genes in a cluster are related. Each cluster is a group of two or more genes that encode a specific protein or polypeptide. Even a few parts of the DNA sequence of each gene within a cluster are found to be the same. The size of the cluster varies from a few genes to several hundred numbers.

4.3. Enrichment analysis

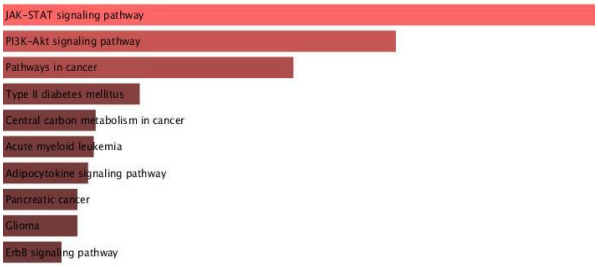
We performed enrichment analysis of 4 sets of complex gene clusters and we took the smallest p-value for individual datasets via the enrichr software.

For dataset 1, we found only 1 complex gene cluster MTOR NEDD9 EPOR . These genes are significantly responsible for the regulation of wound healing, spreading of epidermal cells biological process, cytokine receptor activity molecular function, JAK-STAT signaling KEGG pathway. The result has been given below.



A. GO Biological Process (GO BP)

B. GO Molecular Function (GO MF)

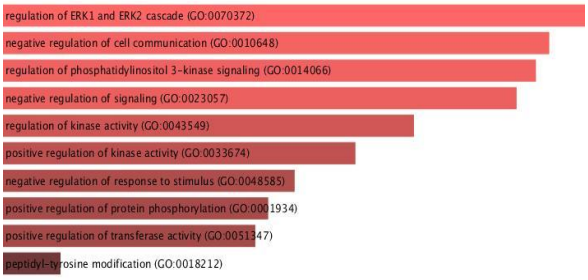


C.Kyoto Encyclopedia Genes and Genomes (KEGG)

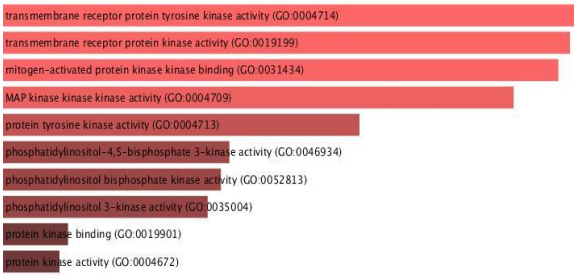
Figure 3.Geneset Enrichment analysis of the disease-gene associated network (Dataset 1).

Graph bars are arranged by p-value ranking. The bar length represents the significance of that particular gene set.

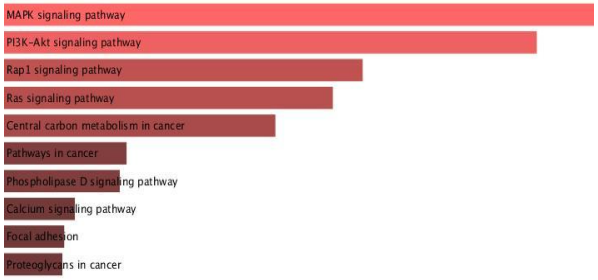
For dataset 2, we have found 37 complex gene clusters. According to the lowest p-value 4.00E-09from enrichment analysis, we identified SMAD4 PDGFRA KIT EGFR KDR ERBB3 genesin a cluster that belong to a particular pathway. They are responsible for the regulation of ERK1 and ERK2 cascade biological process, transmembrane receptor protein tyrosine kinase activity molecular function and MAPK signaling pathway. The result has been given below.



A. GO Biological Process (GO BP)



B. GO Molecular Function (GO MF)

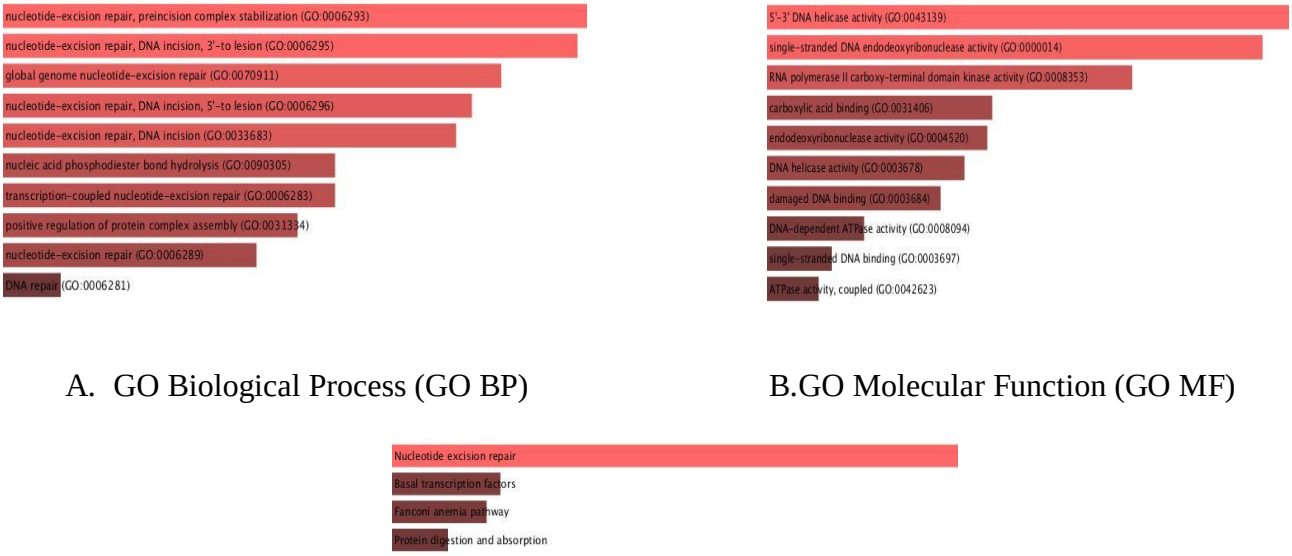


C. Kyoto Encyclopedia Genes andGenomes (KEGG)

Figure 4. Geneset Enrichment analysis of the disease-gene associated network (Dataset 2).

Graph bars are arranged by p-value ranking. The bar length represents the significance of that particular gene set.

For dataset 3, we have found 59 complex gene clusters. According to the lowest p-value 1.62E-05 from enrichment analysis, we identified ERCC2 COL18A1ERCC1 genes in a cluster that belong to a particular pathway. They are responsible for the nucleotide-excision repair, pre-incision complex stabilization biological process, 5′-3′ DNA helicase activity molecular function and nucleotide excision repair pathway. The result has been given below.



A. GO Biological Process (GO BP)

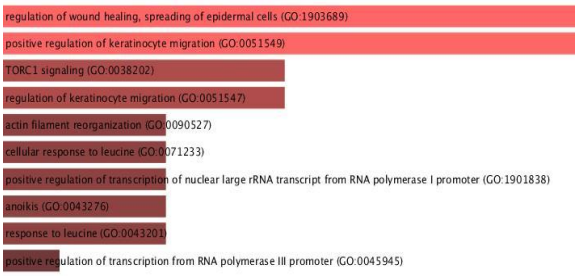
B.GO Molecular Function (GO MF)

C. Kyotos Encyclopedia Genes and Genomes

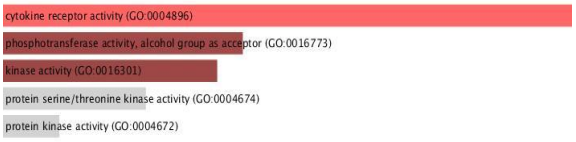
Figure 5.GenesetEnrichment analysis of the disease-gene associated network (Dataset 3).

Graph bars are arranged by p-value ranking. The bar length represents the significance of that particular gene set.

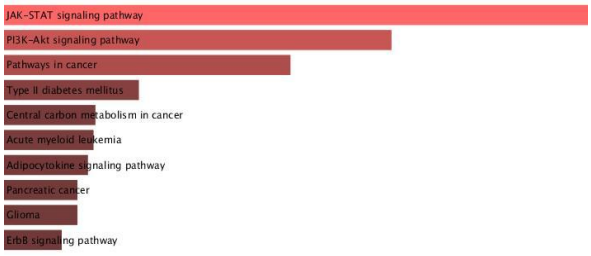
For dataset 4, we have found 2 complex gene clusters. According to the lowest p-value 0.000195, we have identified MTOR NEDD9 EPOR genes in a cluster that belong to a particular pathway the same as dataset 1. These genes are significantly responsible for the regulation of wound healing, spreading of epidermal cells biological process, cytokine receptor activity molecular function, JAK-STAT signaling KEGG pathway. The result has been given below.



A.GO Biological Process (GO BP)



B. GO Molecular Function (GO MF)



C. Kyotos Encyclopedia Genes and Genomes

Figure 6. Geneset Enrichment analysis of the disease-gene associated network (Dataset 4).

Graph bars are arranged by p-value ranking. The bar length represents the significance of that particular gene set.

Chapter 5

Discussion

Discussion

In the present study, genes were collected from lung cancer, prostate cancer, and breast cancer, finally matched with reference genes and retrieved matched genes. As 80% of genes were from breast cancer but the common gene cluster MTOR NEDD9 EPOR was found mostly related to lung cancer and colon carcinomas (Pranabananda et al., 2013). This common gene cluster belongs to the JAK-STAT pathway which is common for various cancers. Dysregulation of the JAK-STAT pathway in T helper cells may result in different types of immune disorders (Farhadet al., 2017). 3 out of 4 datasets contain the MTOR NEDD9 EPOR gene cluster. The JAK-STAT pathway plays a vital role in cytokine-mediated immune responses. Many studies in the JAK-STAT pathway have enlightened its roles in different types of cellular processes such as proliferation, apoptosis, and migration, and have identified dysregulation of the JAK-STAT pathway in a variety of cancer (Pranabananda et al., 2013).

The Signal Transducers and Activators of Transcription (STAT) interfere in tumor suppression through an epigenetic mechanism. STAT5A is engaged in stabilizing heterochromatin, hence giving protection from cancer. This rising site of research may cause drug development in the argument against cancer (Pranabananda et al., 2013).

In the present study, other gene clusters are SMAD4 PDGFRA KIT EGFR KDR ERBB3 and ERCC2 COL18A1 ERCC1 from dataset 2 and dataset 3 with the lowest p-value. They follow the MAPK signaling pathway and the nucleotide excision repair pathway. Both of these pathways play a vital role in various cancer development.

In the case of dataset2, the SMAD4 PDGFRA KIT EGFR KDR ERBB3gene cluster follows the MAPK signaling pathway. Many studies found that the mitogen-activated protein

kinase (MAPK) pathway plays an important role in the regulation of gene expression, cellular growth, and survival. Abnormal MAPK signaling may cause increased or uncontrolled cell proliferation and resistance to apoptosis (Weibin et al., 2019). It is a complex signaling pathway involved with oncogenesis, tumor progression and drug resistance. Various cancers are diagnosed due to dysregulation of the MAPK signaling pathway through multiple mechanisms, including abnormal expression of pathway receptors and/or genetic mutations that occur activation of receptors and downstream signaling molecules where the specific stimuli are absent. As this pathway is resistant to drugs, particularly because of the high degree of interactions and possible compensatory responses.The multi-targeted approach is applied to eradicate the event of resistance over activation of compensatory pathways which is related to MAPK (Cornelia et al., 2019).

In the case of dataset 3, the ERCC2 COL18A1 ERCC1 gene cluster follows the nucleotide excision repair pathway. Nucleotide excision repair (NER) is the most important pathway used by mammals to abolish bulky DNA lesions, for example, those produced by UV light, environmental mutagens, and some cancer chemotherapeutic adducts from DNA. For cancer development, the main factor is DNA damage which has been identified long ago. When inaccurate DNA repair causes mutations or chromosomal peculiarity that affects oncogenes and tumor suppressor genes as a result cancerous growth occurs when cells undergo malignant transformation. Various types of cancer are possible because of the mutations in specific DNA repair system (Alessandro et al., 2015).

Another study found that seven genes RDH5, RDH8, RDH10, RDH11, RDH12, RDH13, RDH14 from beast cancer associated with retinoid metabolic biological process, retinol dehydrogenase activity molecular function and retinol metabolism KEGG pathway (*P*-value

<1E-06) (Ajwad,2017). Retinoid receptors initiate various effects of retinoids, such as metabolize estrogen in human breast carcinomas (Ginestier et al., 2009; Suzuki et al., 2001). So that, retinoids (such as vitamin A and its natural and synthetic analogs) have been used as potential chemotherapeutic or chemopreventive agents because of their difference, anti-proliferative, pro-apoptotic properties.Retinoids may have a potential role in chemotherapy for breast cancer patients (Ajwad,2017).In this study we found three genes MTOR NEDD9 EPOR in 3 out of 4 datasets that are related to the regulation of wound healing and spreading of epidermal cells biological process, cytokine receptor activity molecular function and JAK-STAT signaling KEGG pathway.The Janus kinase (JAK)-signal transducer and activator of transcription (STAT) pathway is important for immune system development in the human body, mainly cytokine receptors and they can polarize T-helper cells. T-helper cells are responsible to help the other cells to recognize any foreign substance like cytokine. Cytokine activates T-cells and B-cells. If any dysregulation occurs in this pathway, then the patient must be suspected of lung cancer, colon cancer or breast cancer. STAT has five types of protein. One of them STAT3 and STAT5 regulates the growth of cancer cells. To develop therapies main focus on suppressing the activity of STATs. Witacnistinis a natural product that constrains the binding of STAT3 and STAT5 to the cytoplasmic region of receptors, thus inhibiting the recruitment of STATs for phosphorylation (Zhang et al, 2014). It has been processed to reduce the activity of STAT3 and STAT5 in cancer cell culture. The natural product small-molecule Wit inhibits the recruitment of STAT3 and STAT5 to growth factor and cytokine receptors, tyrosine phosphorylation, nuclear translocation, and DNA binding, that results in inactivation and inhibition of malignant transformation of STAT3 and STAT5 (S J Thomas et al., 2015).Witacnistin may be a potential anticancer drug not only for breast cancer but also for lung cancer and colon cancer

Conclusion

As genes interact with each other very closely, so any type of cancer is a result of coordinated dysfunction of closely connected gene network enriched with all information of cancer mutation.

We established an analysis to detect disease-gene associated networks using lung cancer, breast cancer, and prostate cancer. For this, we identified the reference genes from clinical data of breast cancer patients and found matched genes from common genes and reference genes. After detecting matched genes we identified overlapping genes and made gene pairs to calculate the similarity coefficient. Using these similarity scores and mutation frequency of those matched genes we created gene clusters or networks. Each network has a pathway. So according to p-value ranking, we took the lowest p-value containing gene cluster for analysis. Finally, we found a gene cluster that is common in 3 out of 4 datasets and enriched with the JAK-STAT signaling pathway.The JAK-STAT pathway plays a vital role in cytokine-mediated immune responses. Many studies in the JAK-STAT pathway have enlightened its roles in different types of cellular processes such as proliferation, apoptosis and migration, and have identified dysregulation of the JAK-STAT pathway in a variety of cancer. (Pranabananda et al., 2013). Witacnistin is a natural product that has the ability to protect this pathway from cancer. This natural product may be a potential anticancer drug not only for breast cancer but also for lung cancer and colon cancer.

In this work, an interesting feature has been noticed. Though 80% of genes are related to breast cancer the common gene cluster pathway is identified for lung cancer and colon carcinomas.

So this common gene cluster can be the biomarker for breast cancer, lung cancer, and colon carcinomas.

6.2. Future Works

Further studies must be needed on these pathways to develop new drugs for better treatment. Survival plot should be performed as well to analyze the mutation pattern of the genes in the disease-gene associated network with the cancer patients.

References

1. AjwadRasif. (2017). Identification of significantly mutated subnetworks in the breast cancer genome. 1-70, http://hdl.handle.net/1993/32634.
2. Alessandro Torgovnick and Björn Schumacher. (2015). DNA repair mechanisms in cancer. Front. Genet. 6:157, 1-15, https://doi.org/10.3389/fgene.2015.00157.
3. Camilla Wendt & Sara Margolin. (2019). Identifying breast cancer susceptibility genes – a review of the genetic background in familial breast cancer. ActaOncologica, 58:2, 135-146, DOI: 10.1080/0284186X.2018.1529428.
4. Chao-Hsin Chen, Chao-Yu Pan & Wen-chang Lin. (2019). Overlapping protein-coding genes in the human genome and their coincidental expression in tissues. 9:13377, 1-15, https://doi.org/10.1038/s41598-019-49802-w.
5. Ciriello G1, Gatza ML2, Beck AH3, Wilkerson MD4, Rhie SK5, Pastore A6, Zhang H7, McLellan M8, Yau C9, Kandoth C10, Bowlby R11, Shen H12, Hayat S6, Fieldhouse R6, Lester SC3, Tse GM13, Factor RE14, Collins LC3, Allison KH15, Chen YY16, Jensen K17, Johnson NB3, Österreich S18, et.al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. 163(2):506-19, doi: 10.1016/j.cell.2015.09.033.
6. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M, Sofia HJ, Hutter C, Getz G, Wheeler D, Ding L; MC3 Working Group; Cancer Genome Atlas Research Network. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. 28;6(3):271-281.e7, doi: 10.1016/j.cels.2018.03.002.

7. FarhadSeif, Majid Khoshmirsafa, HosseinAazami, MonirehMohsenzadegan, GholamrezaSedighi and MohammadaliBahar. (2017).The role of JAK-STAT signaling pathway and its regulators in the fate of T helper cells. 15:23, 1-13, DOI 10.1186/s12964-017-0177-y.
8. Gao Q, Liang WW, Foltz SM, Mutharasu G, Jayasinghe RG, Cao S, Liao WW, Reynolds SM, Wyczalkowski MA, Yao L, Yu L, Sun SQ; Fusion Analysis Working Group; Cancer Genome Atlas Research Network, Chen K, Lazar AJ, Fields RC, Wendl MC, Van Tine BA, Vij R, Chen F, Nykter M, Shmulevich I, Ding L. (2018). Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. 23(1):227-238.e3, doi: 10.1016/j.celrep.2018.03.050.
9. Ginestier, C., Wicinski, J., Cervera, N., Monville, F., Finetti, P., Bertucci, F., …Charafe-Jauffret, E. (2009). Retinoid signaling regulates breast cancer stem cell differentiation. Cell Cycle (Georgetown, Tex.), 8(20), 3297–302. http://doi.org/10.4161/cc.8.20.9761.
10. Greulich H. (2010). The genomics of lung adenocarcinoma: opportunities for targeted therapies. 1(12):1200-10,doi: 10.1177/1947601911407324.
11. Guocan Wang,Di Zhao, Denise J. Spring, and Ronald A. DePinho. (2018). Genetics and biology of prostate cancer. 32(17-18): 1105–1140, doi: 10.1101/gad.315739.118.
12. Hahn, W. C., & Weinberg, R. A. (2002). Modelling the molecular circuitry of cancer. Nature Reviews. Cancer, 2(5), 331–41, http://doi.org/10.1038/nrc795.
13. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, Akbani R, Bowlby R, Wong CK, Wiznerowicz M, Sanchez-Vega F, Robertson AG, Schneider BG, Lawrence MS, Noushmehr H, Malta

TM; Cancer Genome Atlas Network, Stuart JM, Benz CC, Laird PW. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. 173(2):291-304.e6, doi: 10.1016/j.cell.2018.03.022.

https://books.google.ca/books/about/The_genetic_basis_of_human_cancer.html?id=pYG09OPbXp0C&redir_esc=y.

14. IBM Knowledge Center. (2012). Calculating Similarity Link Values. https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.ta.help/tmwb_cluster_similarity.htm.
15. Joke Beuten,1,2 Jonathan A.L. Gelfond,3 Jennifer L. Franke,2 Stacey Shook,2 Teresa L. Johnson-Pais,1 Ian M. Thompson,4 and Robin J. Leach1,2,4.(2010). Single and Multivariate Associations of MSR1, ELAC2, and RNASE L with Prostate Cancer in an Ethnic Diverse Cohort of Men. 19(2): 588–599, doi: 10.1158/1055-9965.EPI-09-0864.
16. Jonathan P. Beauchamp, David Cesarini, Magnus Johannesson, Matthijs J. H. M. van der Loos, Philipp D. Koellinger, Patrick J. F. Groenen, James H. Fowler, J. NielsRosenquist, A. Roy Thurik, and Nicholas A. Christakis. (2011). Molecular Genetics and Economics. Volume 25, 57–82, http://www.aeaweb.org/articles.php?doi=10.1257/jep.25.4.57.
17. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, Fulton LL, Dooling DJ, Ding L, Mardis ER, Wilson RK, Ally A, Balasundaram M, Butterfield YS, Carlsen R, Carter C, Chu A, Chuah E, Chun HJ, Coope RJ, Dhalla N, Guin R, Hirst C, Hirst M, Holt RA, Lee D, et.al. (2012). Comprehensive molecular portraits of human breast tumors. 490(7418):61-70, doi: 10.1038/nature11412.

18. Laura J. Martin, MD. (2018). What Is Carcinoma? https://www.webmd.com/cancer/what-is-carcinoma#1.
19. Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V, Thomas, J. L., Raphael, B. J. (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, *47*(2), 106–114, http://doi.org/10.1038/ng.3168.
20. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, Omberg L, Wolf DM, Shriver CD, Thorsson V; Cancer Genome Atlas Research Network, Hu H. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. 173(2):400-416.e11, doi: 10.1016/j.cell.2018.02.052.
21. Louise Harewood and Peter Fraser. (2014). The impact of chromosomal rearrangements on regulation of gene expression. Human Molecular Genetics, Vol. 23, Review Issue 1 R76–R82, doi:10.1093/hmg/ddu278.
22. M -S Maira, M A Pearson, D Fabbro, and C Garcı ´a-Echeverrı ´a. (2006). Cancer Biology. ISBN (Volume 7) 0-08-044520-9; pp. 1–32, DOI: 10.1016/B0-08-045044-X/00202-9.
23. MaximeHuvet and Michael PH Stumpf. (2014). Overlapping genes: a window on gene evolvability. BMC Genomics 2014, 15:721, http://www.biomedcentral.com/1471-2164/15/721.
24. Meng-Ru Ho, Kuo-Wang Tsai, Wen- chang Lin. (2012). A unified framework of overlapping genes: Towards the origination and endogenic regulation. Genomics 100 (2012)231–239, doi:10.1016/j.ygeno.2012.06.011.

25. Nepusz, T., Yu, H., &Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods, 9(5), 471–472, http://doi.org/10.1038/nmeth.1938.
26. Pranabananda Dutta and Willis X Li. (2013). Role of the JAK-STAT Signaling Pathway in Cancer. 1-10, DOI: 10.1002/9780470015902.a0025214.
27. Razavi P1, Chang MT2, Xu G3, Bandlamudi C4, Ross DS5, Vasan N1, Cai Y5, Bielski CM4, Donoghue MTA4, Jonsson P2, Penson A2, Shen R6, Pareja F5, Kundra R4, Middha S5, Cheng ML7, Zehir A5, Kandoth C4, Patel R4, et.al. (2018). The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers. 34(3):427-438.e6, doi: 10.1016/j.ccell.2018.08.008.
28. RebbeckTR . (2017). Prostate Cancer Genetics: Variation by Race, Ethnicity, and Geography. 27(1):3-10, doi: 10.1016/j.semradonc.2016.08.002.Epub 2016 Aug 26.
29. Rizzolo, P., Silvestri, V., Falchetti, M., &Ottini, L. (2011). Inherited and acquired alterations in development of breast cancer. The Application of Clinical Genetics, 4, 145–58. http://doi.org/10.2147/TACG.S13226.
30. S J Thomas, J A Snowden,M P Zeidler, and S J Danson. (2015). The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. Br J Cancer. 2015 Jul 28; 113(3): 365–371.
31. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, Chakravarty D, Daian F, Gao Q, Bailey MH, Liang WW, Foltz SM, Shmulevich I, Ding L, et. al. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. 173(2):321-337.e10, doi: 10.1016/j.cell.2018.03.035.

32. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545–50, http://doi.org/10.1073/pnas.0506580102
33. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, Schumacher SE, Wang C, Hu H, Liu J, Lazar AJ; Cancer Genome Atlas Research Network, Cherniack AD, Beroukhim R, Meyerson M. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. 33(4):676-689.e3, doi: 10.1016/j.ccell.2018.03.007. Epub 2018 Apr 2.
34. The Healthline Editorial Team. (2016). What Do You Want to Know About Cancer? Medically reviewed by University of Illinois-Chicago, College of Medicine on February 19, 2016. https://www.healthline.com/health/cancer.
35. Tomohiro Nakayama, Satoshi Asai, Yasuo Takahashi, OtoMaekawa, Yasuji Kasama. (2007). Overlapping of Genes in the Human Genome.3(1): 14–19,https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3614620/pdf/IJBS-3-14.pdf
36. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. (2015). Global cancer statistics, 2012. CA Cancer J Clin 65: 87–108,doi: 10.3322/caac.21262, Epub 2015 Feb 4.
37. Vandin, F., Upfal, E., & Raphael, B. J. (2010). Algorithms for detecting significantly mutated pathways in cancer. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6044 LNBI(3), 506–521, http://doi.org/10.1007/978-3-642-12683-3_33.

38. Vogelstein, B., &Kinzler, K. W. (2002). The genetic basis of human cancer. McGraw-Hill, Medical Pub. Division. Retrieved from
39. Vogelstein, B., &Kinzler, K. W. (2004). Cancer genes and the pathways they control. Nature Medicine, 10(8), 789–99, http://doi.org/10.1038/nm1087.
40. Weibin Du, Huahui Hu, Jiangsong Zhang, GuanaiBao, Rongliang Chen, and RenfuQuan. (2019). The Mechanism of MAPK Signal Transduction Pathway Involved with Electroacupuncture Treatment for Different Diseases. Volume 2019, Article ID 8138017, 10 pages. https://doi.org/10.1155/2019/8138017.
41. WILLIAM N. ROM, JOHN G. HAY, THEODORE C. LEE, YIXING JIANG, and KAM-MENG TCHOU-WONG. (1999). Molecular and Genetic Aspects of Lung Cancer. Vol 161, pp 1355–1367, 2000 Internet address: www.atsjournals.org.
42. Zhang X, Blaskovich MA, Forinash KD, Sebti SM. (2014). Withacnistin inhibits recruitment of STAT3 and STAT5 to growth factor and cytokine receptors and induces regression of breast tumours. Br J Cancer. 2014;111 (5:894–902).