Differential Gene Expression in Progression of Breast Cancer using Bioinformatics Approach

Submitted by

Rakhi Moni Saha

ID: 181-35-2416

Department of Software Engineering

Daffodil International University

Supervised by

Dr. Imran Mahmud

Associate Professor, Head

Department of Software Engineering

Daffodil International University

This Project report has been submitted in fulfillment of the requirements for the Degree of

Bachelor of Science in Software Engineering.

APPROVAL

This thesis titled on “Differential Gene Expression in Progression of Breast Cancer using Bioinformatics Approach”, submitted by Rakhi Moni Saha, ID: 181-35-2416 to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

Imran

Chairman
Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

-- Internal
Examiner 1
Afsana Begum
Assistant Professor
Department of Software Engineering
Daffodil International University

-- Internal
Examiner 2
Tapushe Rabaya Toma
Senior Lecturer
Department of Software Engineering
Daffodil International University

-- External
Examiner
Prof. Dr. Md. Saiful Islam
Professor
Institute of Information and Communication Technology (IICT)
Bangladesh University of Engineering and Technology (BUET

THESIS DECLARATION

I announce hereby that I am rendering this study document under **Dr. Imran Mahmud**, **Associate Professor, Head**, Department of Software Engineering, **Daffodil International University**. I therefore state that this work or any portion of it was not proposed here therefore for Bachelor's degree or any graduation.

Supervised by:

`

Imran

Dr. Imran Mahmud
Associate Professor, Head
Department of SWE
Daffodil International University

Submitted by:

`

Rakhi

Rakhi Moni Saha
ID: 181-35-2416
Department of SWE
Daffodil International University

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year thesis successfully.

I really grateful and wish our profound our indebtedness to **Dr. Imran Mahmud**, **Associate Professor and Head in Charge**, Department of Software Engineering, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Science*" to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism , valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express our heartiest gratitude to Dr. Imran Mahmud, Head In-Charge of the Software Engineering faculty for his kind help to finish the thesis and also to other faculty member and the staff of SWE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Breast cancer is the most well-known malignancy among women. Cancer is spread by cells in the human body. This is caused by certain DNA and RNA sequence protein data. In RNA, a gene is similar to a protein chain. We're just looking at the RNA sequences of breast cancer patients in this study, and we're attempting to figure out which genes are active at each stage of the disease. Differentially expressed genes have been identified using differential gene expression analysis. AL158154.2, F10, ELOVL6, PPEF1, IGFBP6, PAX1, AC011893.1, TMED6, AC090877.2, FAM163A, HBG2, ITLN1, HBA2, ALAS2, IL17REL genes have been downregulated and AL158154.2, F10, ELOVL6, PPEF1, IGFBP6, PAX1, AC011893.1, TMED6, AC090877.2, FAM163A, HBG2, ITLN1, HBA2, ALAS2, IL17REL genes have been upregulated. Then we tried to explain the biological and molecular roles that led to the discovery of differentially expressed genes. The main concern of this study, trying to identify novel genes in each breast cancer stage which are responsible for this stage.

TABLE OF CONTENTS

Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Background

Breast cancer is considered the most prevalent cancer in women and specifically the second most common cancer in the world. It is the leading cause of disability-adjusted life years (DALYs) in women across the world. About 1 in every 8 women at some point in their lives are affected and if it's caught early enough, there's a decent chance of recovery [1] The American Cancer Society (ACS) reports that the chance of dying from breast cancer is around 1 in 38 which is 2.6 %[2]. Breast cancer mortality did not change significantly from the 1930s to the 1970s. 2.3 million women were diagnosed with carcinoma and 6, 85,000 deaths worldwide in 2020, and 7.8 million women have been diagnosed with carcinoma in the previous five years. In nations where early detection programs were paired with multiple treatment choices to remove invasive sickness, survival rates began to improve in the 1980s. In Bangladeshi women, breast cancer is still the most frequent and deadly malignancy. According to the International Agency for Research on Cancer, over 13,000 women are diagnosed with breast cancer in Bangladesh each year, with over 7,000 dying [3].

Cancer is a condition in which some cells in the body grow out of control and spread to other regions of the body. It may begin practically anywhere in the billions of cells that make up the human body. Human cells normally expand and multiply (via a process known as cell division) to generate new cells as needed by the body. Cells die as they get old or injured, and new cells replace them. This ordered process can sometimes break down, resulting in aberrant or damaged cells growing and multiplying when they shouldn't.

Tumors, which are masses of tissue, can grow from these cells. Tumors may or may not be malignant (benign). Cancerous tumors can infect adjacent tissues and spread to other parts of the body, resulting in the formation of new tumors (a process called metastasis). Malignant tumors are another name for cancerous tumors. Many malignancies, including leukemias, create solid tumors, whereas cancers of the blood do not. Benign tumors do not penetrate or spread into neighboring tissues. Benign tumors seldom reappear after being excised, although malignant ones do. However, benign tumors can grow to be extremely enormous. Some, such as benign brain tumors, can produce significant symptoms or even be fatal[4].

Breast cancer is a disorder in which the cells of the breast get uncontrollably large. There are several types of breast cancer. The kind of breast cancer is determined by which cells in the breast become cancerous. Breast cancer can start in a variety of places in the breast. Lobules, ducts, and connective tissue are the three primary components of a breast. The glands that generate milk are known as lobules. The ducts are tubes that transport milk from the breast to the nipple. Everything is held together by connective tissue, which is made up of fibrous and fatty tissue. Breast cancer usually starts in the ducts or lobules. Breast cancer can spread to other parts of the body via blood and lymph arteries. Breast cancer is considered to have metastasized when it spreads to other regions of the body. The most prevalent types of breast cancer are described below.

Invasive ductal carcinoma is a kind of cancer that spreads throughout the body. The cancer cells start in the ducts and spread outside of them to various regions of the breast tissue.

Invasive cancer cells can also travel to other places of the body, which is known as metastasis.

Invasive lobular carcinoma is a kind of cancer that spreads throughout the body. Cancer cells start in the lobules and then travel from the lobules to nearby breast tissues. These invasive cancer cells have the potential to spread throughout the body.

Other types of breast cancer, such as Paget's disease and ovarian cancer, are less prevalent. DCIS (ductal carcinoma in situ) is a kind of breast cancer that can progress to invasive breast cancer. The cancer cells have only spread to the duct lining and have not migrated to other breast tissues [5].

The human body has trillions of cells. The growth, maturity, division, and death of these cells are all controlled by a strictly regulated cell cycle. Normal cells divide more quickly in childhood to allow a person to grow. Cells divide once they reach adulthood to replace worn-out cells and repair injuries. The cellular blueprint, or DNA and genes, which are found within the nucleus, control cell division and growth. When cells in one section of the body begin to grow out of control, cancer develops. All types of cancer, regardless of origin, are caused by abnormal cell proliferation, which results in the formation of tumors and lesions.

Breast cancer is a cancerous tumor that begins in the breast cells. There are various factors that can increase the risk of breast cancer, just as there are for other cancers. Experiments have connected estrogen exposure to DNA damage and genetic abnormalities that can lead to breast cancer. Some people are born with abnormalities in their DNA and genes such as BRCA1, BRCA2, and P53. Those who have a family history of ovarian or breast cancer are more likely to develop breast cancer. Cancer cells and cells with damaged DNA are generally sought out and destroyed by the immune system. Failure of such a strong immune defense and monitoring system could lead to breast cancer [6]. The third and last stage of tumor growth is tumor progression. The tumor cells' growth rate and invasiveness both accelerate during this period. Phenotypical alterations occur as a result of the progression, and the tumor becomes more aggressive and has a higher malignant potential.

The spread of cancer cells from the main tumor to surrounding tissues and distant organs is known as metastasis, and it is the leading cause of cancer morbidity and death. It is believed that 90 percent of cancer fatalities are caused by metastasis. In more than 50 years, this estimate hasn't altered much. Metastasis is made up of a succession of interconnected phases. Cancer cells must detach from the initial tumor, intravasate into the circulatory and lymphatic systems, avoid immune response, extravasate at distant capillary beds, then infiltrate and proliferate in distant organs to complete the metastatic cascade. Metastatic cells also create a microenvironment that promotes angiogenesis and proliferation, culminating in macroscopic secondary tumors that are malignant. Despite the fact that systemic metastasis is responsible for over 90% of cancer deaths, most cancer research does not include metastasis in the in vivo condition. The fact that around 1,500 individuals die from cancer every day demonstrates the disease's failure to be managed once it has spread throughout the body [7].

Cancer is a genetic illness, meaning it is caused by alterations in genes that regulate how our cells behave, particularly how they divide and develop. Cancer cells, on average, exhibit more genetic alterations than healthy ones. However, each person's cancer has a unique set of genetic mutations. Some of these alterations might be the effect rather than the cause of cancer. Additional alterations will occur as the malignancy progresses. Cancer cells may have diverse genetic alterations even within the same tumor.

TP53, which generates a protein that inhibits tumor development, is the most often altered gene in all malignancies. Furthermore, germline mutations in this gene can induce Li-Fraumeni syndrome, a rare genetic condition associated with an increased risk of some malignancies.

Hereditary breast and ovarian cancer syndrome is characterized by an elevated lifetime risk of breast and ovarian cancers in women due to inherited mutations in the BRCA1 and BRCA2 genes. This condition has been linked to a variety of malignancies, including pancreatic and prostate tumors, as well as male breast cancer.

PTEN is another gene that creates a protein that inhibits tumor development. Cowden syndrome is a hereditary illness in which mutations in this gene increase the risk of breast, thyroid, endometrial, and other forms of cancer [8].

Six biological characteristics developed throughout the multistep evolution of human malignancies are the hallmarks of cancer. The hallmarks serve as a framework for understanding the intricacies of neoplastic illness. Maintaining proliferative signals, evading growth suppressors, resisting cell death, enabling replicative immortality, initiating angiogenesis, and activating invasion and metastasis are only a few of them. Genome instability, which creates the genetic diversity that speeds up their acquisition, and inflammation, which supports many hallmark activities, are at the root of these hallmarks. Reprogramming of energy metabolism and escaping immune destruction are two key features of potential universality that have emerged in the previous decade. Tumors have additional layer of complexity in addition to cancer cells: they comprise a collection of recruited, presumably normal cells that help to the acquisition of characteristic qualities by forming the "tumor microenvironment." The general application of these principles will progressively influence the development of novel cancer treatments for humans [9].

The stage of your cancer relates to how far it has spread and how large the tumor is. The TNM classification system is the most extensively used cancer classification method. The TNM system is the most common way for hospitals and medical institutes to report cancer. The TNM system is as follows the T stands for the major tumor's size and extension. The major tumor is frequently referred to as the main tumor. Then the N stands for the number of cancerous lymph nodes in the area. Lastly, the M indicates whether or not the malignancy has spread. This indicates that cancer has gone beyond the initial tumor and into other regions of the body. When the TNM system is used to characterize your cancer, there will be numbers following each letter that provide more information, such as T1N0MX or T3N1M0. The meanings of the letters and digits are explained in the following:

The primary tumor (T)	Lymph nodes in the region (N)	Distant metastasis is a term that refers to the spread of cancer cells (M)
TX: There is no way to quantify the main tumor.	NX: Cancer in adjacent lymph nodes is impossible to detect.	MX: Metastasis is impossible to quantify.
T0: The primary tumor has not been located.	N0: There is no malignancy in the lymph nodes in the area.	M0: There has been no spread of cancer to other sections of the body.
T1, T2, T3, T4: The size and/or extent of the primary tumor is indicated by these letters. The larger the tumor or the more it has expanded into neighboring tissues, the higher the number	N1, N2, N3: Indicates the number and location of cancer-bearing lymph nodes. The greater the number following the N, the more cancerous lymph nodes there are.	M1: The cancer has spread throughout the body.

following the T. T's can be subdivided further to offer more information, such as T3a and T3b.		

Table 1: Description of the TNM cancer classification system.

The TNM system aids in the detailed description of cancer. But TNM combinations are classified into five less-detailed phases for various malignancies. These are Stage 0 Abnormal cells exist, but they have not spread to adjacent tissue. Also known as carcinoma in situ (CIS). CIS is not cancer, but it has the potential to become cancer. Then cancer in stages I, II, and III are present. The larger the cancer tumor and the more it has spread into neighboring tissues, the higher the number. The cancer has spread to other places of the body at this stage IV.

Another cancer staging method, which is used for all forms of cancer, divides the disease into five categories. Cancer registries utilize this staging method more frequently than doctors. However, our doctor may still refer to our cancer in one of this ways. Abnormal cells are present in situ, but they haven't disseminated to surrounding tissue. Cancer is localized- it only affects the area where it began, with no signs of spreading. Cancer has spread to adjacent lymph nodes, tissues, or organs on a regional level. Cancer has spread to sections of the body that are far away. Unknown- There is insufficient information to determine the stage [10].

Gene expression profiling is a helpful method for predicting and investigating toxicity processes. For transcriptional profiling, RNA-Seq technology has expanded as a viable alternative to standard microarray systems. When compared to microarrays, RNA-Seq detected more differentially expressed protein-coding genes and gave a larger quantitative range of expression level alterations. Both platforms found a greater number of genes that were differentially expressed (DEGs) [11].

RNA sequencing (RNA-Seq) is quickly displacing microarrays for gene expression profiling due to its increased accuracy and sensitivity. How to find a collection of transcripts that are differently expressed under distinct experimental circumstances is one of the most prevalent questions in a standard gene profiling investigation. With or without adjustments, several of the statistical approaches established for microarray data processing may be used to RNA-Seq data. Several new approaches designed exclusively for RNA-Seq data sets have recently been created.

Transcriptomics holds the key to understanding how the genome's information is translated into cellular activities, as well as how this translation process responds to changes in the environment. One of the outstanding questions in transcriptomics is to accurately quantify the abundance of each transcript within different tissues and time points, and to correlate changes in abundance to genetic and environmental perturbation in order to understand genome function and adaptation, given a transcriptome, which is a collection of all transcripts including both protein coding mRNAs and noncoding RNAs.

The method used to estimate the steady state abundance of each transcript within a transcriptome is known as transcriptome profiling or gene expression profiling. Traditional transcriptome profiling methods include quantitative RT-PCR for a few genes, or microarrays or whole genome tiling arrays for genome-wide transcriptional activity. Because of its remarkable sensitivity and accuracy (reviewed in), transcriptome profiling by RNA sequencing (RNA-Seq) has recently been the method of choice as a result of the cheap cost of next generation sequencing technology. Next-generation sequencing methods, unlike previous technologies, allow reference transcriptomes to be constructed directly from RNA-Seq data, obviating the requirement for existing reference genomes or transcriptomes. This feature is especially appealing to non-model species or microbial communities that lack high-quality references [12].

The identification of a limited number of gene signatures that define cancer phenotypes in relation to their prognosis is an important application of gene expression profile analysis. Breast cancer metastasis to different parts of the body and to different phases of metastasis can be connected to genes. IL3RA2, VCAM1, and MMP2 have been linked to aggressive breast cancer metastasis to the lung, whilst ST6GLANAC5 is a particular facilitator of breast cancer metastasis to the brain, and cytokeratin-19 has been discovered as a possible marker of breast stem cells. Nm23, KAI1, and BRMS1 are tumor suppressors connected to the prevention of tumor cells detaching from main tumors, whereas KISS1 and MKK4 are linked to decreased growth at secondary locations. Secreted phosphoprotein 1 (SPP1), S100 calcium binding proteins A4 (S100A4), and S100P and anterior gradient 2 (AGR2), a subgroup of metastasis-inducing proteins (MIPs), have also been shown to be overexpressed in patients with sporadic and metastatic breast cancer, and are linked to decreased patient survival. As a result, the up- and down-regulation of discovered gene signatures may be used to predict prognosis [13].

Gene expression that reacts to inputs or triggers; a gene regulatory mechanism. The biological mechanisms that decide which genes in a cell are actively transcribed and translated into mRNA and proteins, and under what circumstances is called differential gene expression [14].

1.2 MOTIVATION OF THE RESEARCH

Breast cancer, as stated before, is one of the most prevalent cancers worldwide, and the single most important factor of death from this type of cancer is the cancer metastasis. As the cancer progresses, there is often less chance of survival for the patients. Different types of genetic alterations are usually responsible for cancer progression. High throughput RNA-Seq technology is leveraged to measure gene expressions in cells of different conditions. Identification of novel genes that are differentially expressed in specific stages might open more opportunities to prevent cancer progression as these genes can often be used as therapeutic targets.

TCGA-BRCA is a project where thousands of breast cancer patients were recruited and their gene expression patterns of cancer tissue and normal tissue were profiled. Previous studies have mostly analyzed genes that are differentially expressed between normal tissue and cancer tissue. Some studies have analyzed differentially expressed genes between different stages of cancers. However, no study, to the best of my knowledge, has leveraged the vast dataset that was generated by the TCGA-BRCA project yet.

1.3 PROBLEM STATEMENT

The actual pathophysiology of metastasis is still poorly understood. One of the ways to elucidate the changes that tumor cells go through to reach metastasis is to measure the gene expression pattern of the cells and identify the genes that are differentially expressed in each stage. In this study, we conducted a differential gene expression analysis to unearth the genes that are differentially expressed in each stage and to discover their biological and molecular functions.

1.4 RESEARCH QUESTIONS

The research questions of this study is as follows:

1. Is there any novel gene that contributes to tumor progression?
2. If any, what are the molecular and biological functions of those genes?

1.5 RESEARCH OBJECTIVE

The main objective of this research is to find out novel differentially expressed genes from stage one to four of breast cancer. The secondary objective of this study is to elucidate the biological and molecular functions these genes play.. The main objectives are as follows:

- To collect and preprocess RNA-Seq count data of breast cancer tissues from the TCGA-BRCA project.
- To perform differential gene expression analysis to identify novel genes that are expressed specifically in each stage of cancer.
- To elucidate the biological and molecular functions of the identified differentially expressed genes.

1.6 RESEARCH SCOPE

The main scope of this research is as follows:

1.7 THESIS ORGANIZATION

A section on differential gene expression in breast cancer progression, breast cancer statistics, stage, the context behind the study, motivation for the research, issue description, research questions, and research purpose are presented in the first chapter. The following are the other elements of our research:

We will examine the literature review in our next chapter, where we will look at some previous research works on the same topic of differential gene expression, their methods, and shortcomings, as well as a comparison between my work and theirs. We shall detail our work's technique in their chapter. We will address data gathering, data pre-processing, and analysis as part of our work technique. In chapter four, the methodology's findings will be detailed. The final chapter is the chapter that brings the document to a close. We will now go to the conclusion section, where we will provide a complete summary of our efforts. We have said about what work we will undertake in the future to improve our work.

CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION

A researcher evaluates past work, research, conference papers, books, articles, and so on in a literature review. With this, one may find out what work has previously been done on the issue, summarize the entire topic, and determine what parts of the work are needed. They can work on restrictions and overcome them after evaluating to achieve better outcomes.

2.2 BREAST CANCER PROGRESSION

In situ gene expression patterns of human breast cancer in the premalignant, preinvasive, and invasive phases were studied by Xiao-Jun Ma[15] and colleagues. They gathered all gene information on breast cancer patients from the PNAS website for their investigation. The Massachusetts General Hospital provided all of the sample data. Data was filtered and normalized using the Cy3 and Cy5 channels. They employed hierarchical clustering in GENMATHS to analyze the data, which was based on the correlation coefficient between two genes, and linear discriminant analysis inside GENMATHS. For statistical analysis, the Bioconductor software is also employed. Based on early microarray study, they chose CRIP1, IFI-6–16, PNMT type 3 gene for up regulation and ELF5, NDRG2 type 2 gene for down regulation. In this work, they used a combination of laser capture microdissection, RNA amplification, and microarray technology to create epithelium specific, in situ gene expression profiles of breast cancer in the premalignant, preinvasive, and invasive phases. A group of genes was revealed with quantifiable expression levels that correlated with advanced tumor grade and the transition from DCIS to IDC. The gene RRM2 has been identified as being associated with advanced cancer stage and tumor grade. RRM2 also works with a variety of oncogenes to enhance malignancy and the potential for metastatic spread. RRM2 might thus have a dual function in sustaining fast cell proliferation while simultaneously boosting invasive growth behavior, linking increased tumor grade to the DCIS–IDC transition.

Xiao-Jun Ma[16] and colleagues found that progression from normal breast tissue to ductal carcinoma in situ (DCIS) to invasive ducal carcinoma (IDC) was accompanied by gene expression changes in all cell types within the tumor microenvironment, indicating that these cell types all play a role in tumorigenesis. They compared global gene expression changes in the stromal and epithelial compartments during the course of breast cancer from normal to preinvasive to invasive ductal carcinoma. All of the breast cancer samples were collected from the Massachusetts General Hospital between 1998 and 2001 as fresh-frozen biopsies. Patients were chosen based on the availability of patient-matched normal and tumor samples, as well as the absence of fibrocystic alteration in the normal breast lobules. They analyzed 14 patient-matched normal epithelium, normal stroma, tumor epithelium, and tumor related stroma tissues using laser capture microdissection and gene expression microarrays. Gene ontology and differential gene expression studies were carried out. During cancer development, tumor-associated stroma experiences substantial gene expression alterations, similar to those found in the malignant epithelium. Extracellular matrix components, matrix metalloproteases, and cell-cycle-related genes are among the highly elevated genes in the tumor-associated stroma. Both the tumor epithelium and the stroma showed decreased expression of cytoplasmic ribosomal proteins and increased expression of mitochondrial ribosomal proteins. Increased expression of numerous matrix metalloproteases followed the change from preinvasive to invasive growth (MMP2, MMP11 and MMP14). Furthermore, in the stroma, a gene expression signature of histological tumor grade exists, with high-grade tumors linked with increased expression of genes implicated in immune response, as seen in malignant epithelium. They concluded by claiming that the tumor microenvironment is involved in carcinogenesis even before tumor cells breach the stroma, and that it may play a key role in the shift from preinvasive to invasive growth. Malignant epithelial cells in high-grade tumors may use immune cells in the tumor stroma to promote aggressive invasive development.

Ana Cristina Vargas and Et Al[17] team published a report on gene expression analysis of tumor epithelial and stromal compartments during breast cancer development. They found gene expression profiling in 87 formalin-fixed paraffin-embedded (FFPE) samples from 17 patients, which included matched IDC, DCIS, and three types of stroma: IDC-S (three millimeters from IDC), DCIS-S (three millimeters from DCIS), and breast cancer associated-normal stroma [ten millimeters from IDC or DCIS]. Quantitative real-time PCR, immunohistochemistry, and immunofluorescence were used to confirm differential gene expression analyses. In comparison to normal stroma from reduction mammoplasties, the expression of numerous genes was down-regulated in cancer patients' stroma. They also carried out experiments on molecules involved in extracellular matrix remodeling, including as COL11A1, COL5A2, and MMP13, which were shown to be differently expressed in DCIS and IDC. In comparison to DCIS, COL11A1 was overexpressed in IDC, and it was expressed by both the epithelial and stromal compartments, but it was concentrated in invading neoplastic epithelial cells. The Human Research Ethics Committees at the Royal Brisbane and Women's Hospital (RBWH), Uniting HealthCare (The Wesley Hospital), and The University of Queensland gave their authority to access fresh frozen and Formalin fixed paraffin embedded samples locally. They utilized 17 individuals with invasive ductal carcinoma or mixed ductal lobular carcinoma. The WG-DASL test was utilized to explore gene expression pattern alterations to address the clinically significant scenario of progression from DCIS to IDC and the role performed by both the epithelial and stromal compartments in this process utilizing limma Bioconductor software packages using R. There were 58 genes that were differently expressed between IDC and DCIS, with 42 and 16 genes being up- and down-regulated in DCIS, respectively, as compared to IDC. Finally, they hypothesized that these expression alterations may aid in the shift from in situ to invasive malignancy, indicating a pivotal phase in disease progression.

The researchers used a meta-analysis of data on gene expression in metastatic and primary breast cancer tumors to validate putative prognostic and therapeutic indicators. R. That was accomplished with the help of Bell and Et Al [18]. The Genevestigator (Nebion) database was used to pull data on relative gene expression values from 12 studies on primary breast cancer and breast cancer metastasis. They employed Genevestigator software and Ingenuity technologies for the analysis. Genevestigator is a program that finds genes that are up- or down-regulated in response to a collection of perturbations. Gene sets that were differently expressed among distinct breast cancer neoplasms were identified using the conditions tool. Co-expressed gene connections computed from array data are provided by this software similarity search tool. Then, using the Ingenuity tool, the regulators upstream of the gene set were identified, and the relational functional networks were reconstructed.

Their findings reveal that transcriptional expression of the COX2 gene is dramatically downregulated in metastatic tissue relative to normal breast tissue, but not in initial tumors, and might be exploited as a differential marker in metastatic breast cancer diagnosis. When compared to primary breast cancer, RRM2 gene expression is lower in metastases, suggesting that it might be used as a marker to track breast cancer progression. MMP1, VCAM1, FZD3, VEGFC, FOXM1, and MUC1 were also shown to have significantly altered expression in breast neoplasms compared to normal breast tissue, indicating that they might be used as breast cancer onset indicators. COX2 and RRM2 are considered to be major indicators for breast cancer metastasis, they concluded. The interaction of upstream regulators of genes differentially expressed in initial breast tumors and metastasis points to p53, ER1, ERB-B2, TNF, and WNT as the most promising regulators for new complex pharmacological treatment interventions in breast cancer metastatic progression.

The transition of ductal carcinoma in situ to invasive breast cancer is linked to EMT and myoepithelia gene expression patterns, according to Erik S. Knudsen[19] and colleagues. To establish important gene expression patterns in either the epithelium or stromal compartment associated with disease development, laser capture microdissected tissue from pure DCIS and pure IBC was used. Each tissue has its own gene expression profile, and a DCIS/IBC classifier correctly separated DCIS from IBC in numerous data sets. They also discovered that the epithelial compartment, rather than the stroma, had the most significant changes in gene expression. The tumor repository at Thomas Jefferson University provided DCIS and IBC cases. The nuclear grade of DCIS was divided into three categories: low, middle, and high. IBC histological grade was determined using the Nottingham classification. The estrogen receptor, progesterone receptor, and HER2 status of DCIS and IBC were compared. Data was filtered to remove probesets with no gene annotation before analysis, and genes with multiple probesets were handled by averaging their rows and scaling by the probeset with the biggest standard deviation. Unless otherwise stated, all further analysis was carried out in Matlab. GEO data set GSE33692 has been deposited. To identify trends in the microarray data, genes with variance in the top 25% percentile were employed for both principal component analysis and hierarchical clustering. Principal component analysis was used on all samples, and the second, third, and fourth components were plotted as functions of the first component to see whether there were any natural separations that may be linked to sample tissue features. Using Pearson's correlation distance metric and average linkage, hierarchical clustering was done on both genes and samples. Then, using a two-sample t test with unequal variance, differential gene expression analysis for IBC vs DCIS samples was done inside the epithelial and stromal samples to identify genes related with progression within each of these compartments. Following this filtering phase, statistical analysis of microarrays (SAM), an alternate approach for differential gene expression analysis, was utilized to get a better estimate of the FDR, which was calculated by randomly permuting samples in the dataset. The Gene Expression Omnibus provided series matrix files and annotation for gene expression datasets GSE3893, GSE14548, and GSE26304. Psuedo-expression signatures in each GEO dataset were established by median-centering gene expression profiles, multiplying the median-centered profiles for downregulated genes by -1, and taking the average across all genes for the epithelial and stromal differential gene sets. To find commonalities between expression profiles in our microarray dataset and previously published disease progression/invasion gene sets, we employed the gene set enrichment analysis software tool. In addition, utilizing the database for annotation, visualization, and integrated discovery, functional enrichment analysis of the highest differentially expressed genes in the epithelial compartment was done. These findings suggest that variations in gene expression associated with invasive phenotype are especially important in the development of DCIS to invasive breast cancer.

Breast cancer-derived stromal fibroblasts revealed a specific gene expression pattern that distinguished them from normal breast stroma, according to Christian F. Singer and Et Al[20]. Using a 2,400 gene cDNA array, stromal fibroblasts derived from malignant tissue of ten women with invasive breast cancer and normal breast tissue of ten women with benign breast diseases were exposed to differential complementary DNA Microarray Analysis. RT-PCR was used to confirm individual gene expression patterns. In fibroblasts from malignant breast tumors, the mRNA expression of 135 genes was elevated more than 2 fold in a cDNA array that allows for the analysis of differential gene expression of more than 2,400 genes. The bulk of these genes code for cytokines that promote tumor growth, transcription factors, and cell-matrix proteins. 110 genes had their mRNA expression reduced by less than 0.5 fold. The remaining 2,155 genes were not changed in any way. The validity of the pooled gene expression profile was validated by RT-PCR on individual biopsies from breast cancer and normal breast tissues. Even in the absence of neighboring malignant epithelium, they found elevated expression of tumor promotion-associated genes, suggesting that tumor stroma has a fibroblastic subpopulation that offers a milieu that promotes tumor development and invasion.

In around 30% of human breast tumors, the proto-oncogene HER2/neu is amplified and overexpressed. This occurrence has been linked to a more aggressive phenotype in several studies. Using cDNA microarrays, Katherine S. Wilson[21] and his colleagues sought to identify genes associated with the aggressive phenotype of HER2/neu-positive breast cancer cells. Three HER2/neu-positive and three HER2/neu-negative breast cancer cell lines were used to extract RNA. For cDNA microarray analysis, ten breast carcinomas were chosen from the ICRF frozen breast cancer repository. Following surgical excision, the tumors were kept at 80°C for 8 to 10 years. All of them were diagnosed as invasive ductal carcinomas, with half of them being HER2/neu-positive. To focus on the HER2/neu pathway, they mostly employed ER positive primary tumors. For frozen section IHC, a second slice of each tumor was employed. Cell culture, tissue, RNA isolation, cDNA Microarray Hybridization, Northern Blot Analysis, and Immunohistochemistry (IHC) were employed in the analysis section. Many of the differently expressed genes have never been linked to human neoplasia before, and some of them might be new tumor suppressor or oncogenes. In cell lines and carcinomas with high HER2/neu protein levels, no genes were up-regulated, and only a few genes were down-regulated. Acidic coiled coil containing protein 1, glycogen phosphorylase BB, complement 1q, and one EST were among them. Finally, they asserted that R2/neu has emerged as a critical breast cancer biomarker since it predicts a more aggressive clinical phenotype and corresponds with a tumor's response to systemic treatment.

Shreshtha Malvia[22] and colleagues identified gene expression patterns in breast tumors from the Indian subcontinent in this work, shedding information on the pathways and genes linked to breast tumorigenesis in Indian women. They obtained 97 patient specimens who had already been histologically verified to be breast cancer patients. The data for this study came from the Indraprastha Apollo Hospital in New Delhi, India, which had been tracking breast cancer for four years. They utilized the limfit tool from the limma package for statistical analysis, and the Bioconductor package R for all forms of analysis. Microarray gene expression profiling, microarray data processing, and statistical analysis are examples. Cluster software was used to perform unsupervised hierarchical clustering on the DEGs. The adjusted probe intensities were centered on the median, Pearson correlation was used to quantify similarity, and centroid linkage clustering was done. The clustering image was also seen using JavaTree View Software. Pathway Express software was used to do gene ontology analysis. The KEGG (Kyoto Encyclopedia of Genes and Genomes) database was used to identify specific pathways for differentially expressed genes. Gene Set enrichment analysis software was also employed to get insight into the top 50 DEGs and create a heatmap. Protein-protein interaction networks were predicted using a search function for Recurring Instances of Neighboring Genes in the network analysis program 'NetworkAnalyst.' COL10A1, COL11A1, MMP1, MMP13, MMP11, GJB2, and CST1 were overexpressed genes during breast cancer, while PLIN1, FABP4, LIPE, AQP7, LEP, ADH1A, ADH1B, and CIDEC were under expressed genes. They also discovered a hub gene that interacts with the PPI network. AURKB, CENPA, TOP2A, BUB1, CCNB2, MMP1, and SPP1 were among them. The metastatic cycle is controlled by hub genes. Finally, they argued that their resources for this study were restricted, and that they could not investigate the entire Indian continent[24].

2. 3DIFFERENTIAL GENE EXPRESSION

In this section, I tried to show gene list which is differentially expressed during breast cancer.

Paper Name	**Identified Gene**
Gene expression profiles of human breast cancer progression	IDH2, FLJ10540, KNSL1, ANKT, PRO2000, 2810433K01, RAD51, **UBE2C,** ANLN, PMSCL1, STK15, CENPA, PSMD12, EST, TOPK, UBE2N, S100A10, TCEB1, EST, **RRM2,** TOP2A, CML66, FLJ10468, KIAA0165, PLK, TACC3, RTN4, EST, NEK2, MGC2721, EST, CDKN3, RACGAP1, MGC2577 , RRM2, BUB1, BIRC5, **CKS2**
Gene expression profiling of the tumor microenvironment during breast cancer progression	S100P, CYB561, SCD, RRM2, CNTNAP2, HIST1H1C, IFI27, HIST1H2BD, CDC2, RAB31, RRM2, HIST1H2BC, MELK, IFI6, IFIT1, NAT1, DHRS2, HIST1H2BC, RAB31, CYP2B6, RAB31, CEACAM6, GPC1, CAPN13, ID4, GPM6B, DMN, FOXC1, SPARCL1, ROPN1, KRT15, PHLDA1, LOC728264, CX3CL1, RGS2, GABRP, SOSTDC1, BOC, C2orf40, PHLDA1, C13orf15, KIT, HOXA9, WIF1, ID4, ELF5, SFRP1, DMD, SLC6A14
Gene expression profiling of tumour epithelial and stromal compartments during breast cancer progression	SOX10, SFRP1, KRT14, ECRG4, CA4, KLK5, ODZ2, KRT5, ZBTB16, SCARA5, INMT, KLK7, CNN1, ALDH1A2, KRT17, FMO2, KRT6B, OSR1, COL17A1, OR5P3, RNF39, MYH11, MEOX1, SNCA, ABCA6, FREM1, MGLL, PAMR1, ALOX15B, APOD, ZNF502, CDC14A, SDK2, CPXM2, TTC21B, LOXL4, ALPL, CAPN11, KLF8, ARSH, ADRB2, SNAPC1, COL5A2, PLAU, MZT1, ULK3, TTPAL, SGSM3, COL22A1, COL8A1, MMP13, GLIS3, COL12A1, GPC6, GRM8, COL10A1, GRM4, COL11A1, IDC-S versus RM-NS, ABCB1, EFCBP1, FAHD1, TUBB3, CXCR4, FLVCR2, MRAP2, CREB3L1, HIST1H2BM, USP19, DCIS-S versus RM-NS, ABCB1, EFCBP1, MRAP2, USP19, FAHD1, BC-NS versus RM-NS, ABCB1, ABCB1, USP19

Gene expression meta-analysis of potential metastatic breast cancer markers	GPX8, FST, LOX, PXDN, EHD2, HNRNPM, ATP6, DCAF6, ND2, ND3, GTSE1, HJURP, KIF2C, MKI67, TPX2, DLGAP5, DLGAP5, CCNA2, UBE2C, KIF4A, ELF3, AGR2, PIGR, TMC4, RASEF, TMC5, SLC44A4, KRT19, TSPAN1, C9orf152, ST6GALNAC1, LOC100505989, KIAA0101, TOP2A, ZWINT, DTL, CCNB2, DLGAP5, DLGAP5, MELK, BIRC5, ASPM, NUF2.
Differential gene expression profile in breast cancer-derived stromal fibroblasts	GM-CSF, PEDF, K-ras oncogene protein, EF-Ts, Ribosomal protein S12, GNAQ, 63 kDa protein kinase related to rat ERK3, Clock, Dihydrodiol dehydrogenase, Protein tyrosine phosphatase (PTPase), Osteoblast-specific factor-2 (OSF-2p1)
Differential Gene Expression Patterns in HER2/neuPositive and -Negative Breast Cancer Cell Lines and Tissues	EST, Smooth muscle myosin heavy chain isoform Smemb, EST, Vesicle trafficking protein, EST, EST, KIAA0465, KIAA0461, Spleen tyrosine kinase, EST, EST, EST, EST, EST, Max, EST, Rab13, Stanniocalcin 2, EST
Identification of key genes and pathways by bioinformatics analysis with TCGA RNA sequencing data in hepatocellular carcinoma	AURKB, AURKA, FOS, CCNB2, BIRC5, PLK1, CDC20, CCNB1, TOP2A, CDK1
Combining Serial Analysis of Gene Expression and Array Technologies to Identify Genes Differentially Expressed in Breast Cancer1	Mucin, Claudin-7, B94, HER2/neu, Neurosin, Zn-a-2-GP, Thrombospondin, NGAL/Lipocalin 2, EST, EST, Mucin, HER2/neu, Claudin-7, EST, NGAL/Lipocalin 2, Cytochrome B561, B94, EST, EST, EST, Spr1, GST-pi, EST, Ataxia telangiectasia group D, Integrin a-6, RIG like 7-1, Heparin binding protein, Cyclin D2, MEN1 region epsilon/beta, Plakophilin, Spr1, EST, GST-pi, Ataxia telangiectasia group D, RIG like 7-1, Integrin a-6, Heparin binding protein, Cyclin D2, Serine protease PRSS11, Plakophilin

Table2: Identified gene list from literature review

2.4 FUNCTIONAL ENRICHMENT ANALYSIS

Functional enrichment analysis is a technique for identifying gene or protein classes that are over-represented in a large collection of genes or proteins that may be linked to disease characteristics. This strategy use statistical approaches to identify gene groupings that are strongly enriched. DNA microarrays or RNA-Seq are still used in GSEA to compare two separate categories, but this time the focus is on a gene set rather than a single gene in a big list. Researchers look to see if the majority of the genes in the collection are found at the extremities of the list: The major variances in expression between the two categories are represented at the top and bottom of the list. If a gene set is over-expressed or under-expressed, it is thought to be associated to phenotypic variations. The following are the general steps of Creative Proteomics' enrichment analysis:

- Calculate a p-value for the number of times the proteins in the set are over-represented at the top or bottom of the list.
- The p-value is used to determine the statistical significance of a node or route.
- For multiple hypothesis testing, the P-value for each set is standardized, and a false discovery rate is determined.

2.5 CONCLUSION

There are a variety of ways that may be employed. They used a variety of techniques to identify expressed genes in tumor and non-tumor patients, including the Bioconductor package, microarray analysis, differential gene expression analysis, and so on. In this study, I also attempted to identify genes that are differently expressed in breast cancer from stage I to IV.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 DATA SOURCE

The Cancer Genome Atlas (TCGA) is a government-funded project whose goal is to produce a comprehensive "atlas" of cancer genomic profiles by cataloging and discovering key cancer-causing genetic changes. Actually this is the assembled project of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The TCGA framework is well-organized, with numerous cooperating centers in charge of sample collection and processing, followed by high-throughput sequencing and sophisticated bioinformatics data analytics. To begin, several Tissue Source Sites (TSSs) collect and send the needed biospecimens (blood, tissue) from qualified cancer patients to the Biospecimen Core Resource (BCR). The BCR then catalogs, processes, and verifies the quality and quantity of samples before sending clinical data and metadata to the Data Coordinating Center (DCC) and providing molecular analytes to Genome Characterization Centers (GCCs) and Genome Sequencing Centers (GSCs) for further genomic characterisation and high-throughput sequencing. The DCC is then used to store sequence-related data. The Genome Characterisation Centers also send trace data, sequences, and alignment maps to the Cancer Genomics Hub (CGHub), which is a secure repository run by the National Cancer Institute. Genome Data Analysis Centers and the research community have access to the genetic data created (GDACs). To encourage greater utilization of TCGA data, the GDACs give novel information-processing, analysis, and visualization tools to the entire scientific community. Furthermore, the data generated by the TCGA Research Network is centralized at the DCC and entered into public free-access databases (TCGA Portal, NCBI's Trace Archive, CGHub), allowing scientists to access cancer datasets on a continuous basis and accelerate cancer biology and related technologies advancements.

To characterize molecular profiles of human breast tumors, researchers used data from genomic DNA copy number arrays, DNA methylation, exome sequencing, mRNA arrays, microRNA sequencing, and RPPA. The existence of four main breast cancer classes was confirmed, as expected, by results from various platforms. TBX3, RUNX1, CBFB, AFF2, PIK3R1, PTPN22, PTPRD, NF1, SF3B1, and CCND3 were discovered as novel, significantly altered genes, in addition to nearly all previously implicated genes in breast cancer. The luminal A subtype had the lowest overall mutation rate, while the basal-like and HER2-positive subtypes had the highest. Genomic characterisations were also used to identify potential druggable targets. Because PIK3CA mutations are common in luminal/ER-positive tumors, inhibitors of the PI3K pathway may be useful. Somatic mutations, such as a high frequency of PIK3CA mutations, a decreased frequency of PTEN and PIK3R1 mutations, and genetic losses of PTEN and INPP4B, are possible therapeutic targets in HER2-positive tumors. Druggable mutations in the HER receptor family are another promising target. Apart from BRCA1 and BRCA2, somatic mutation study for basal-like breast tumors has not revealed a common drug-targeted mutation. However, there were significant molecular similarities between basal-like breast cancers and high-grade serous ovarian tumors, indicating a common aetiology and treatment approaches, which is corroborated by the activity of platinum analogues and taxanes in breast basal-like and serous ovarian tumors.

3.2 DATA COLLECTION & DATA PREPROCESSING

RNA-Seq count data along with the respective clinical data of the TCGA-BRCA project [39] were downloaded from the GDC Data Portal (https://portal.gdc.cancer.gov/) using the R (https://www.r-project.org/) package named ‘TCGABiolinks’ [40]. The AJCC stage of each sample was encoded in the clinical variable ‘ajcc_pathologic_stage’. The substages A, B, and C were merged into their parent stages resulting in four final stages of interest: Stage I, Stage II, Stage III, and Stage IV. Samples with missing stage information were discarded. Finally, only the samples of invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC) were retained as they are the most prevalent types of breast cancers. The information about the sample subtypes were encoded in the clinical variable ‘primary_diagnosis’. A summary of the key demographic and clinical features is shown in table 1.

Characteristics		Control	Stage I	Stage II	Stage III	Stage IV	Overall
Samples		96	160	567	221	17	1061
Age (Years)		58.01 ± 15.03	60.18 ± 12.83	59.09 ± 12.75	59.70 ± 14.24	61.71 ± 11.36	59.33 ± 13.28
Subtype	IDC	89 (92.71%)	140 (87.50%)	457 (80.60%)	152 (68.78%)	16 (94.12%)	854 (80.

							49%)
	ILC	7 (7.29%)	20 (12.50%)	110 (19.40%)	69 (31.22%)	1 (5.88%)	207 (19.51%)
Prior Treatment	Yes	1 (1.04%)	1 (0.63%)	5 (0.88%)	6 (2.71%)	0 (0.00%)	13 (1.23%)
	No	95 (98.96%)	159 (99.37%)	561 (98.94%)	215 (97.29%)	17 (100%)	1047 (98.68%)
	Not Reported	0 (0.00%)	0 (0.00%)	1 (0.18%)	0 (0.00%)	0 (0.00%)	1 (0.09%)
Vital Status	Alive	60 (62.50%)	148 (92.50%)	504 (88.89%)	184 (83.26%)	5 (29.41%)	901 (84.92%)
	Dead	36 (37.50%)	12 (7.50%)	63 (11.11%)	37 (16.74%)	12 (70.59%)	160 (15.08%)

Table 3: Summary of the important demographic and clinical features of the dataset grouped by different sample types. Percentages of the subtypes, prior treatments, and vital status are calculated within sample types.

To make the stage-wise differential analysis more concrete, cases with prior treatments and unreported prior treatments were discarded from the dataset.

TCGA Stage	TNM Classification	Cases

I		76	160
IA	T1, N0, M0	79	
IB	T0 or T1, N1, M0	5	
II		6	567
IIA	T0, N1, M0 T1, N1, M0 T2, N0, M0	323	
IIB	T2, N1, M0 T3, N0, M0	238	
III		2	221
IIIA	T0, T1, T2, or T3; N2; M0 T3, N1, M0	142	
IIIB	T4; N0, N1, or N2; M0	20	
IIIC	any T, N3, M0	57	
IV	any T, any N, M1	17	17

Table 4: Stage wise data split

3.3 DIFFERENTIAL GENE EXPRESSION ANALYSIS

Differential expression analysis is the statistical examination of normalised read count data to uncover quantifiable differences in expression levels across experimental groups. Different techniques for differential expression analysis exist, including edgeR and DESeq, which are based on negative binomial (NB) distributions, and baySeq and EBSeq, which are Bayesian approaches based on the NB model. When selecting an analytic technique, it is critical to examine the experimental design. While certain differential expression tools, such as edgeR, limma-voom, DESeq, and maSigPro, can only do pair-wise comparisons, others, such as edgeR, limma-voom, DESeq, and maSigPro, can do multiple comparison.

In below are given some information on the method of differential gene expression analysis:

Method Name	**Usage of Method**

DESeq: DESeq is a R tool for analyzing and testing differential expression in count data from high-throughput sequencing experiments like RNA-Seq.	- Use the default values if you want to be conservative. When outliers are introduced, it becomes more conservative. - TPR is often low. - FDR control is poor with two samples per condition, but acceptable with bigger sample sizes and outliers. - Requires a moderate amount of processing time, which rises significantly as the sample size grows.
edgeR: A Bioconductor module for analyzing digital gene expression data for differential expression.	- With default settings, rather liberal for small sample numbers. When outliers are introduced, it becomes more liberal. - TPR is often high. - In many situations, poor FDR control, which is exacerbated by outliers. - Requires a moderate amount of processing effort, which is mostly independent of sample size.
NBPSeq (Negative Binomial Models for RNA Sequencing Data): For two-group comparisons and regression conclusions using RNA-Sequencing Data, Negative Binomial (NB) models are used.	- For all sample sizes, be generous. When outliers are introduced, it becomes more liberal. - TPR medium - Poor FDR control, which is exacerbated by outliers. Often, the genes that are actually non-DE have the smallest p-values. - Requires a moderate amount of processing time, which rises significantly as the sample size grows.
TSPM: TSPM stands for Total Suspended Particulate Matter	- Overall performance is strongly sample-size dependant.

and is the concentration that would be obtained from a high-volume bulk sample on a filter substrate.	- For small sample sizes, liberal; outliers have little effect. - FDR control is weak for small sample sizes, but improves dramatically as the sample size grows. Outliers have little impact. - Many really non-DE genes are among the ones with the fewest p-values when all genes are overdispersed. When the counts for certain genes are Poisson distributed, the problem is solved. - Requires a moderate amount of processing effort, which is mostly independent of sample size.
Voom / vst: is a method of Bioconductor package.	- Excellent kind When outliers are introduced, I error control becomes more cautious. - For small sample sizes, low power. For bigger sample sizes, use a medium TPR. - Excellent FDR control, with the exception of simulation study B4000 0. - The inclusion of outliers has little effect. - It is computationally quick.
baySeq: This software calculates estimated posterior likelihoods of differential expression (or more complicated hypotheses) in high-throughput 'count' data, such as that produced from next-generation sequencing machines, using empirical Bayesian techniques.	- When all DE genes are controlled in the same direction, the consequences are highly varied. When DE genes are controlled in various directions, there is less variety. - TPR is low. Outliers have little impact. - In the absence of outliers, poor FDR control with two samples/condition, but good for greater sample numbers. - In the presence of outliers, FDR control is poor.

	- Processing time is sluggish, although parallelization is possible.
EBSeq: R/EBSeq is a R tool for finding genes and isoforms that are differentially expressed (DE) in an RNA-seq experiment under two or more biological circumstances.	- TPR is unaffected by sample size or the presence of outliers. - In most cases, poor FDR control; outliers have little effect. - Requires a moderate amount of processing time, which rises significantly as the sample size grows.
NOISeq : The NOISeq R package provides a complete resource for RNA-seq data analysis, including three sections: (i) count data quality control, (ii) low-count feature filtering, normalization, and batch effect correction, and (iii) DE analysis. The software provides both visualization charts and processing methods within each block to aid in the diagnosis and analysis of count datasets. The program contains a feature that allows you to quickly create a QC report pdf file that includes all of the plots specified in this section.	- It's unclear how to set the qNOISeq threshold to match a specific FDR threshold. - When the dispersion between the conditions is varied, it performs well in terms of false discovery curves (see supplementary material). - The amount of time required for computation is strongly depending on the sample size.
SAMseq: SAMseq is a simple user interface for analyzing the significance of sequencing results.	- For small sample sizes, low power. TPR is high when sample sizes are large enough. - Does well in simulation study B4000 0 as well. - The inclusion of outliers has little effect. - The amount of time required for computation is strongly depending on the sample size.

ShrinkBayes / ShrinkSeq: ShrinkBayes is a R program that may be used to analyze count-based sequencing data in complicated research designs.	- FDR control is often weak, but the user can employ a fold change threshold in the inference phase. - TPR is high. - Slow computation, however parallelization is possible.

Table 5: R Library description of bioinformatics

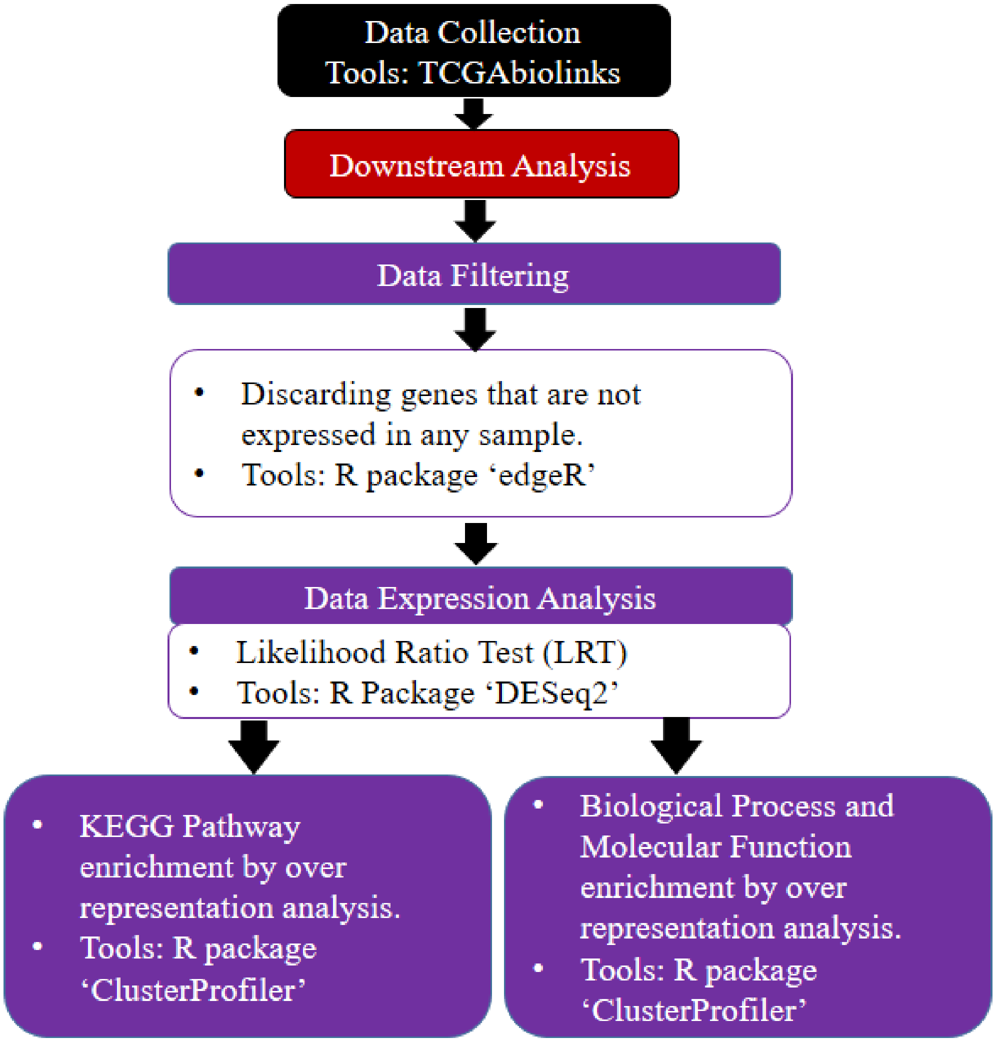
3.4 FUNCTIONAL ENRICHMENT ANALYSIS

Functional enrichment analysis also known as gene set enrichment analysis (GSEA) is a method for identifying gene or protein classes that are over-represented in a large set of genes or proteins that may be linked to disease characteristics. The strategy use statistical techniques to discover gene groupings that are highly enriched or deficient. There are several approaches for analyzing gene sets. Gene set analysis approaches are divided into three categories: over-representation analysis, functional class score, and pathway topology-based methods.

One of the most extensively used classes of gene set analysis methods is over-representation analysis (ORA), which is a natural extension of single-gene analysis. ORA is available through a variety of technologies due to its simplicity, well-established underlying statistical model, and ease of application. There are 68 gene set analysis techniques and tools mentioned, with 40 of them being ORA-based. The many components of these technologies, such as the gene set database, data visualization, and user interface, differ. ORA works with a list L of genes, each of which has been predicted to be differentially expressed using a single-gene analysis approach.

Furthermore, ORA only uses differentially expressed genes, which are typically the consequence of using a p-value threshold, and ignores all other quantitative metrics for the genes. A persistent shift in the expression of genes even those with a p-value somewhat higher than the cutoff value might help identify pathway activity. Unlike ORA, the major purpose of FCS approaches is to use all information from an expression matrix to solve the enrichment issue without depending on the physiologically flawed assumptions outlined before. To distinguish differential enrichment of gene sets, FCS approaches use an expression matrix of gene expression measurements for all genes rather than a list of differentially expressed genes. These approaches are divided into two categories: univariate and multivariate. Using each row of the expression matrix, a gene score is commonly generated for each gene in univariate FCS approaches. The gene set scores for each gene set are then calculated using these gene scores. Finally, the gene set scores are evaluated for significance, and differentially enriched gene sets are provided. Multivariate approaches derive gene set scores directly from the expression matrix, skipping the process of generating gene scores.

Not every gene in a pathway is equally vital to its function. Understanding pathway structure, such as gene product relationships, can aid in estimating a gene's contribution to pathway activity. Topology information has the potential to increase enrichment analysis accuracy. Such information regarding routes is included into topology-based pathway analysis tools. ORA-based, univariate, and multivariate methodologies may all be used to these methods. They also evaluate null hypotheses in the same way that other gene set analysis approaches do.


Data Collection
Tools: TCGAbiolinks
Downstream Analysis
Data Filtering
• Discarding genes that are not expressed in any sample.
• Tools: R package 'edgeR'
Data Expression Analysis
• Likelihood Ratio Test (LRT)
• Tools: R Package 'DESeq2'
• KEGG Pathway enrichment by over representation analysis.
• Tools: R package 'ClusterProfiler'
• Biological Process and Molecular Function enrichment by over representation analysis.
• Tools: R package 'ClusterProfiler'

Working process model in this research

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Differentially Expressed Gene List: The differential gene expression analysis yielded 33 significant genes of which 18 were downregulated and the rest of the 15 were upregulated. The list of the significant genes is given in Table.

S.N	Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	gene_name
1	ENSG00000083454	92.19242523	-0.21320086	0.458227586	23.50288563	3.17E-05	0.038053269	P2RX5
2	ENSG00000086717	175.0347521	0.265335825	0.324686082	47.06937734	3.36E-10	3.90E-06	PPEF1
3	ENSG00000120498	19.76571189	-0.083479056	0.358881622	25.82182239	1.04E-05	0.017217367	TEX11
4	ENSG00000125813	49.91015104	0.31677347	0.593996094	23.57013355	3.07E-05	0.038053269	PAX1
5	ENSG00000126218	173.4953564	0.095621968	0.267576933	22.72058533	4.62E-05	0.048931784	F10
6	ENSG00000126759	118.1290646	-0.889052566	0.331618772	24.71574341	1.77E-05	0.026779926	CFP
7	ENSG00000134201	709.9116118	-2.843726477	0.503509824	42.24908597	3.55E-09	2.38E-05	GSTM5
8	ENSG00000143340	7.309673231	0.465324955	0.405989657	29.30124471	1.94E-06	0.00481009	FAM163A
9	ENSG00000157315	29.70752435	0.371325933	0.179852629	41.94981518	4.11E-09	2.38E-05	TMED6
10	ENSG00000158578	9.523339317	1.626273845	0.46342035	26.98457823	5.93E-06	0.012896972	ALAS2
11	ENSG00000163518	10.06044842	-0.035168237	0.534252239	31.20538027	7.69E-07	0.00223084	FCRL4
12	ENSG00000163631	1564.140907	-0.235980565	0.756741486	26.54450525	7.33E-06	0.015010316	ALB
13	ENSG00000167779	1074.539377	0.299057423	0.298107476	26.31095958	8.21E-06	0.015866899	IGFBP6
14	ENSG00000170522	1381.00536	0.239876215	0.265770644	103.4785498	2.78E-22	9.66E-18	ELOVL6

15	ENSG00000170558	522.9046727	-0.934598804	0.414221592	22.70990327	4.64E-05	0.048931784	CDH2
16	ENSG00000178789	73.54030437	-0.969574389	0.340970405	24.33824489	2.12E-05	0.029544166	CD300LB
17	ENSG00000179914	4.143555964	0.832850079	0.638017568	24.40035647	2.06E-05	0.029544166	ITLN1
18	ENSG00000181195	123.2137752	-3.271160242	0.612519535	31.77868161	5.83E-07	0.001842848	PENK
19	ENSG00000188263	56.80120198	2.455128735	0.512714716	24.20612845	2.26E-05	0.030006538	IL17REL
20	ENSG00000188536	303.6220924	1.330659512	0.4684895	25.45390401	1.24E-05	0.019622673	HBA2
21	ENSG00000196565	5.53002859	0.801698616	0.426983626	37.59532538	3.44E-08	0.000149718	HBG2
22	ENSG00000202198	3797.65151	-2.816443878	0.671236274	44.66136119	1.09E-09	9.50E-06	RF00100
23	ENSG00000225972	1405.089135	-6.246085351	0.784936482	64.92477654	5.21E-14	9.05E-10	MTND1P23
24	ENSG00000226806	12.23901111	0.353680113	0.349171543	22.85440307	4.33E-05	0.048599908	AC011893.1
25	ENSG00000227170	9.943792338	-0.412041277	0.577776037	23.3787778	3.37E-05	0.039044162	AF178030.1
26	ENSG00000235304	11.35622524	-1.283226804	0.556055098	24.14591633	2.33E-05	0.030006538	LINC01281
27	ENSG00000238741	39.75484218	-4.455454669	0.532359639	31.84797583	5.63E-07	0.001842848	SCARNA7
28	ENSG00000239002	41.37635489	-5.159581377	0.77607797	29.81673696	1.51E-06	0.004036024	SCARNA10
29	ENSG00000251504	1.112549983	-0.557558608	0.858572452	35.02403417	1.20E-07	0.000465447	LINC01099
30	ENSG00000252010	53.57471963	-5.42683431	0.756274461	38.35189568	2.38E-08	0.000118329	SCARNA5
31	ENSG00000259527	141.5131665	-1.627766095	0.988054736	25.90590666	9.98E-06	0.017217367	LINC00052

32	ENSG00000259721	5.981815476	0.397852116	0.441808052	26.13825764	8.92E-06	0.016337325	AC090877.2
33	ENSG00000267559	1.649514098	0.066930435	0.296455004	27.46663584	4.70E-06	0.01090005	AL158154.2

Table 6: Differentially expressed gene from this study

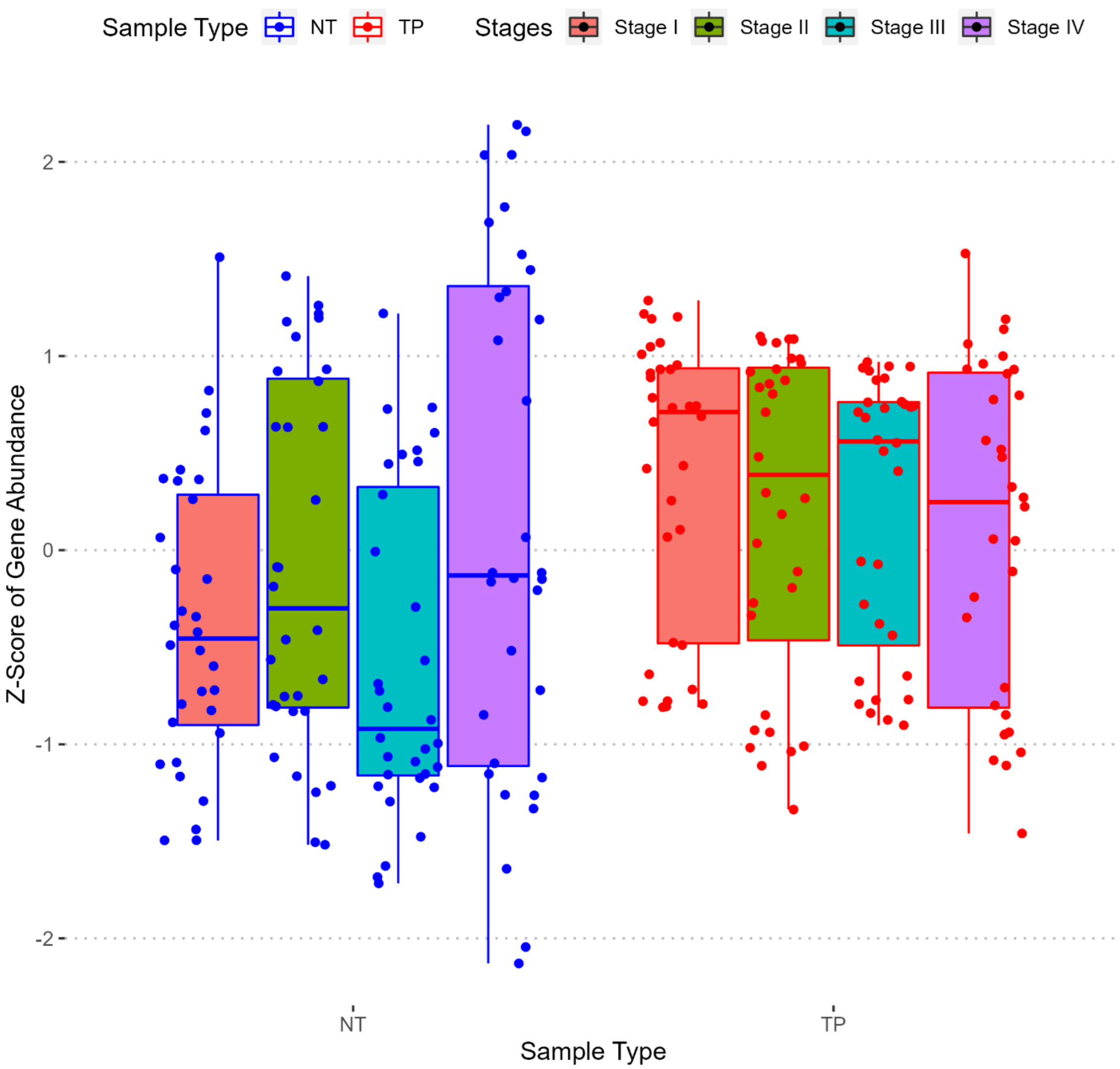


Figure 1: Upregulated and downregulated gene visualization

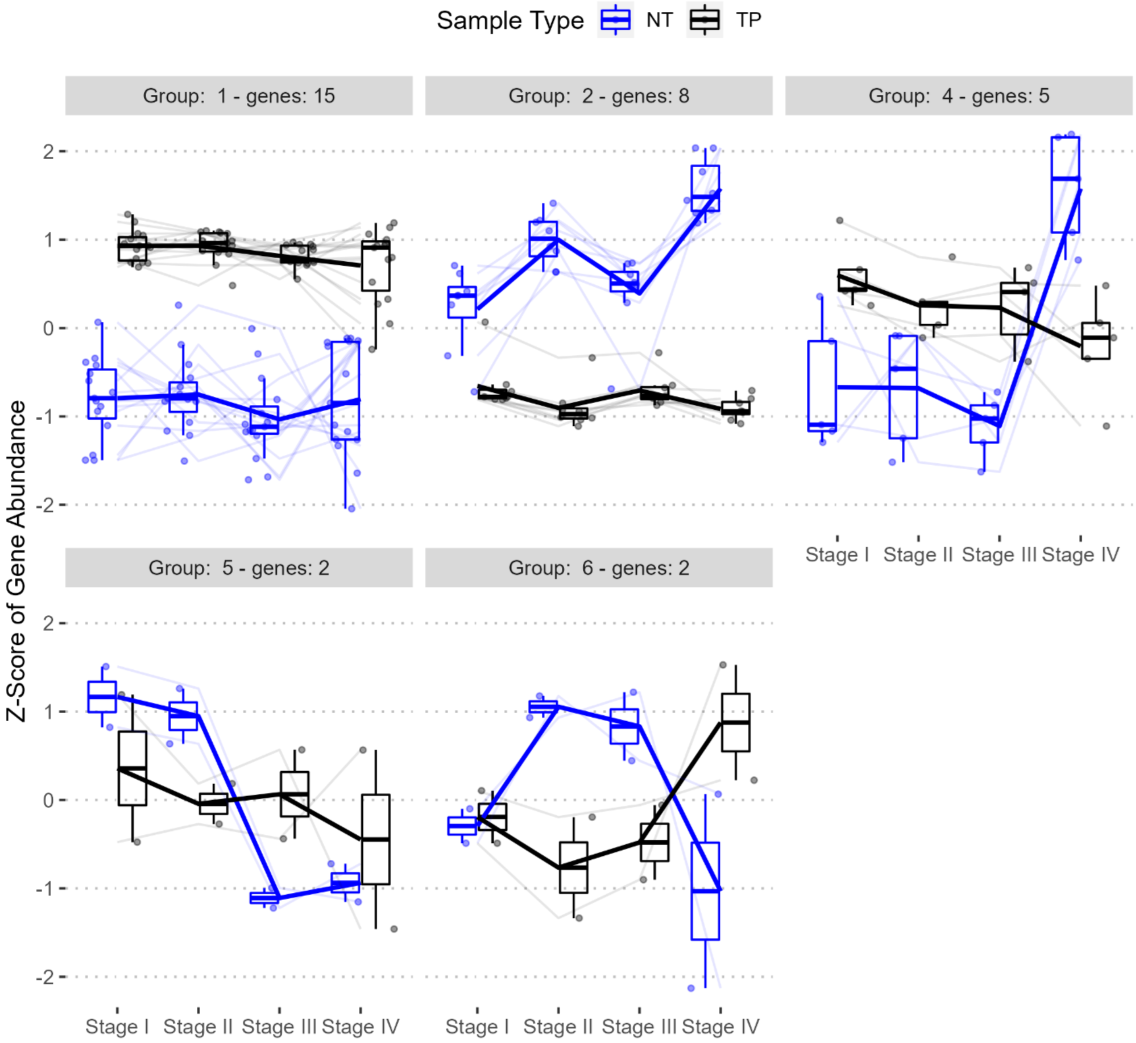


Figure 2: Upregulated and downregulated gene visualization

4.2 Functional Enrichment Analysis

Table1: GO Biological process downregulated

Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	Genes
post-translational protein	3/345	0.003426	0.067879	0	0	11.48538	65.19681	CDH2;PENK;ALB

modification (GO:0043687)								
cellular protein metabolic process (GO:0044267)	3/417	0.005817	0.067879	0	0	9.45314	48.65434	CDH2;PENK;ALB
type B pancreatic cell development (GO:0003323)	7-Jan	0.006284	0.067879	0	0	195.8431	992.8793	CDH2
detection of muscle stretch (GO:0035995)	7-Jan	0.006284	0.067879	0	0	195.8431	992.8793	CDH2
glandular epithelial cell development (GO:0002068)	7-Jan	0.006284	0.067879	0	0	195.8431	992.8793	CDH2
complement activation, alternative pathway (GO:0006957)	7-Jan	0.006284	0.067879	0	0	195.8431	992.8793	CFP
cell-cell adhesion mediated by cadherin (GO:0044331)	8-Jan	0.007179	0.067879	0	0	167.8571	828.6538	CDH2
type B pancreatic cell differentiation (GO:0003309)	9-Jan	0.008072	0.067879	0	0	146.8676	707.7996	CDH2
positive regulation of calcium-mediated signaling (GO:0050850)	13-Jan	0.01164	0.067879	0	0	97.89216	435.9406	P2RX5
synaptonemal complex assembly (GO:0007130)	15-Jan	0.01342	0.067879	0	0	83.89916	361.6911	TEX11
synaptonemal complex organization (GO:0070193)	15-Jan	0.01342	0.067879	0	0	83.89916	361.6911	TEX11

response to muscle stretch (GO:0035994)	15-Jan	0.01342	0.067879	0	0	83.89916	361.6911	CDH2
heme catabolic process (GO:0042167)	16-Jan	0.014308	0.067879	0	0	78.30196	332.5412	ALB
porphyrin-containing compound catabolic process (GO:0006787)	16-Jan	0.014308	0.067879	0	0	78.30196	332.5412	ALB
gliogenesis (GO:0042063)	17-Jan	0.015196	0.067879	0	0	73.40441	307.3228	CDH2
stem cell development (GO:0048864)	17-Jan	0.015196	0.067879	0	0	73.40441	307.3228	CDH2
high-density lipoprotein particle remodeling (GO:0034375)	18-Jan	0.016083	0.067879	0	0	69.08304	285.3111	ALB
neural crest cell differentiation (GO:0014033)	19-Jan	0.01697	0.067879	0	0	65.24183	265.9472	CDH2
detection of mechanical stimulus (GO:0050982)	19-Jan	0.01697	0.067879	0	0	65.24183	265.9472	CDH2
mitochondrion localization (GO:0051646)	19-Jan	0.01697	0.067879	0	0	65.24183	265.9472	ALB
mesenchymal cell development (GO:0014031)	19-Jan	0.01697	0.067879	0	0	65.24183	265.9472	CDH2
heme biosynthetic process (GO:0006783)	22-Jan	0.019624	0.068684	0	0	55.91317	219.7946	ALB
glutathione derivative biosynthetic process (GO:1901687)	22-Jan	0.019624	0.068684	0	0	55.91317	219.7946	GSTM5

glutathione derivative metabolic process (GO:1901685)	22-Jan	0.019624	0.068684	0	0	55.91317	219.7946	GSTM5
porphyrin-containing compound biosynthetic process (GO:0006779)	24-Jan	0.02139	0.07187	0	0	51.04604	196.2636	ALB
positive regulation of muscle cell differentiation (GO:0051149)	27-Jan	0.024033	0.074706	0	0	45.14932	168.3313	CDH2
regulation of calcium-mediated signaling (GO:0050848)	28-Jan	0.024913	0.074706	0	0	43.47495	160.526	P2RX5
homologous chromosome pairing at meiosis (GO:0007129)	29-Jan	0.025791	0.074706	0	0	41.92017	153.3319	TEX11
glial cell differentiation (GO:0010001)	29-Jan	0.025791	0.074706	0	0	41.92017	153.3319	CDH2
innate immune response (GO:0045087)	2/302	0.02965	0.081023	0	0	8.200833	28.85281	CD300LB;CFP
positive regulation of calcium ion transport into cytosol (GO:0010524)	Jan-34	0.030174	0.081023	0	0	35.55971	124.4862	P2RX5
regulation of muscle cell differentiation (GO:0051147)	Jan-35	0.031049	0.081023	0	0	34.51211	119.833	CDH2
retina homeostasis (GO:0001895)	Jan-36	0.031922	0.081023	0	0	33.52437	115.4732	ALB

positive regulation of calcium ion transport (GO:0051928)	Jan-37	0.032795	0.081023	0	0	32.5915	111.3808	P2RX5
glutathione metabolic process (GO:0006749)	Jan-43	0.038016	0.091239	0	0	27.92717	91.31448	GSTM5
neural crest cell development (GO:0014032)	Jan-45	0.039751	0.092066	0	0	26.65508	85.96583	CDH2
regulation of synaptic transmission, glutamatergic (GO:0051966)	Jan-46	0.040617	0.092066	0	0	26.06144	83.48948	CDH2
adherens junction organization (GO:0034332)	Jan-49	0.043211	0.092066	0	0	24.42892	76.74722	CDH2
regulation of complement activation (GO:0030449)	Jan-50	0.044074	0.092066	0	0	23.92917	74.70386	CFP
regulation of axonogenesis (GO:0050770)	Jan-51	0.044937	0.092066	0	0	23.44941	72.75166	CDH2
sensory perception (GO:0007600)	Jan-51	0.044937	0.092066	0	0	23.44941	72.75166	PENK
regulation of immune effector process (GO:0002697)	Jan-53	0.04666	0.09283	0	0	22.54525	69.09833	CFP
regulation of humoral immune response (GO:0002920)	Jan-54	0.04752	0.09283	0	0	22.11876	67.38709	CFP
blood vessel morphogenesis (GO:0048514)	Jan-56	0.049238	0.094001	0	0	21.3123	64.17307	CDH2
homophilic cell adhesion via	Jan-60	0.052666	0.098311	0	0	19.86341	58.47348	CDH2

plasma membrane adhesion molecules (GO:0007156)								
neuropeptide signaling pathway (GO:0007218)	Jan-63	0.05523	0.100854	0	0	18.89943	54.73757	PENK
cell-cell junction assembly (GO:0007043)	Jan-66	0.057786	0.101126	0	0	18.02443	51.38769	CDH2
cellular response to nutrient levels (GO:0031669)	Jan-66	0.057786	0.101126	0	0	18.02443	51.38769	ALB
nervous system development (GO:0007399)	2/447	0.060243	0.101366	0	0	5.487921	15.41759	P2RX5;CDH2
synapse assembly (GO:0007416)	Jan-69	0.060337	0.101366	0	0	17.22664	48.36924	CDH2
cellular protein modification process (GO:0006464)	3/1025	0.061644	0.101531	0	0	3.710372	10.33852	CDH2;PENK;ALB
cell-cell junction organization (GO:0045216)	Jan-82	0.071312	0.114354	0	0	14.45243	38.16431	CDH2
peptide metabolic process (GO:0006518)	Jan-83	0.072152	0.114354	0	0	14.27547	37.52998	GSTM5
negative regulation of cell death (GO:0060548)	Jan-94	0.081336	0.126523	0	0	12.58001	31.56529	ALB
cell junction assembly (GO:0034329)	1/102	0.087962	0.134342	0	0	11.57892	28.14659	CDH2
modulation of chemical synaptic	1/109	0.093723	0.140584	0	0	10.82462	25.62637	CDH2

transmission (GO:0050804)								
sulfur compound biosynthetic process (GO:0044272)	1/113	0.096999	0.142946	0	0	10.43592	24.34759	GSTM5
regulation of anatomical structure morphogenesis (GO:0022603)	1/123	0.105141	0.150598	0	0	9.575699	21.56884	CDH2
platelet degranulation (GO:0002576)	1/125	0.106761	0.150598	0	0	9.420304	21.07477	ALB
synapse organization (GO:0050808)	1/126	0.10757	0.150598	0	0	9.344471	20.83458	CDH2
receptor-mediated endocytosis (GO:0006898)	1/143	0.121217	0.166921	0	0	8.218724	17.34294	ALB
positive regulation of cytosolic calcium ion concentration (GO:0007204)	1/147	0.124399	0.168541	0	0	7.991942	16.65729	P2RX5
brain development (GO:0007420)	1/150	0.126779	0.169038	0	0	7.829846	16.17108	CDH2
organonitrogen compound biosynthetic process (GO:1901566)	1/158	0.133094	0.171999	0	0	7.427876	14.97977	GSTM5
cellular response to starvation (GO:0009267)	1/158	0.133094	0.171999	0	0	7.427876	14.97977	ALB
regulation of neuron projection development (GO:0010975)	1/165	0.138585	0.176381	0	0	7.108321	14.04796	CDH2

cell-cell adhesion via plasma-membrane adhesion molecules (GO:0098742)	1/170	0.142487	0.178641	0	0	6.896276	13.43742	CDH2
defense response to bacterium (GO:0042742)	1/176	0.147147	0.180289	0	0	6.657815	12.75851	CFP
regulation of immune response (GO:0050776)	1/179	0.149469	0.180289	0	0	6.544613	12.43914	CD300LB
regulated exocytosis (GO:0045055)	1/180	0.150241	0.180289	0	0	6.507723	12.33548	ALB
regulation of programmed cell death (GO:0043067)	1/194	0.160985	0.190461	0	0	6.031393	11.016	ALB
anterograde trans-synaptic signaling (GO:0098916)	1/244	0.198319	0.231372	0	0	4.778262	7.730646	PENK
positive regulation of cell differentiation (GO:0045597)	1/258	0.208488	0.239117	0	0	4.514763	7.07858	CDH2
cellular component assembly (GO:0022607)	1/261	0.210651	0.239117	0	0	4.461991	6.949783	TEX11
central nervous system development (GO:0007417)	1/268	0.215677	0.241558	0	0	4.343468	6.662769	CDH2
chemical synaptic transmission (GO:0007268)	1/306	0.242436	0.267955	0	0	3.794986	5.377566	PENK
negative regulation of	1/381	0.292739	0.319352	0	0	3.034365	3.727634	ALB

programmed cell death (GO:0043069)								
neutrophil degranulation (GO:0043312)	1/481	0.354918	0.372371	0	0	2.389951	2.475673	CFP
negative regulation of apoptotic process (GO:0043066)	1/485	0.357295	0.372371	0	0	2.369713	2.438895	ALB
neutrophil activation involved in immune response (GO:0002283)	1/485	0.357295	0.372371	0	0	2.369713	2.438895	CFP
neutrophil mediated immunity (GO:0002446)	1/488	0.359072	0.372371	0	0	2.354753	2.411816	CFP
positive regulation of intracellular signal transduction (GO:1902533)	1/546	0.392526	0.4021	0	0	2.097895	1.961851	P2RX5
positive regulation of cellular process (GO:0048522)	1/625	0.435448	0.440694	0	0	1.824849	1.517144	P2RX5
regulation of apoptotic process (GO:0042981)	1/742	0.493786	0.493786	0	0	1.527427	1.077833	ALB

Table 7: GO Biological process downregulated

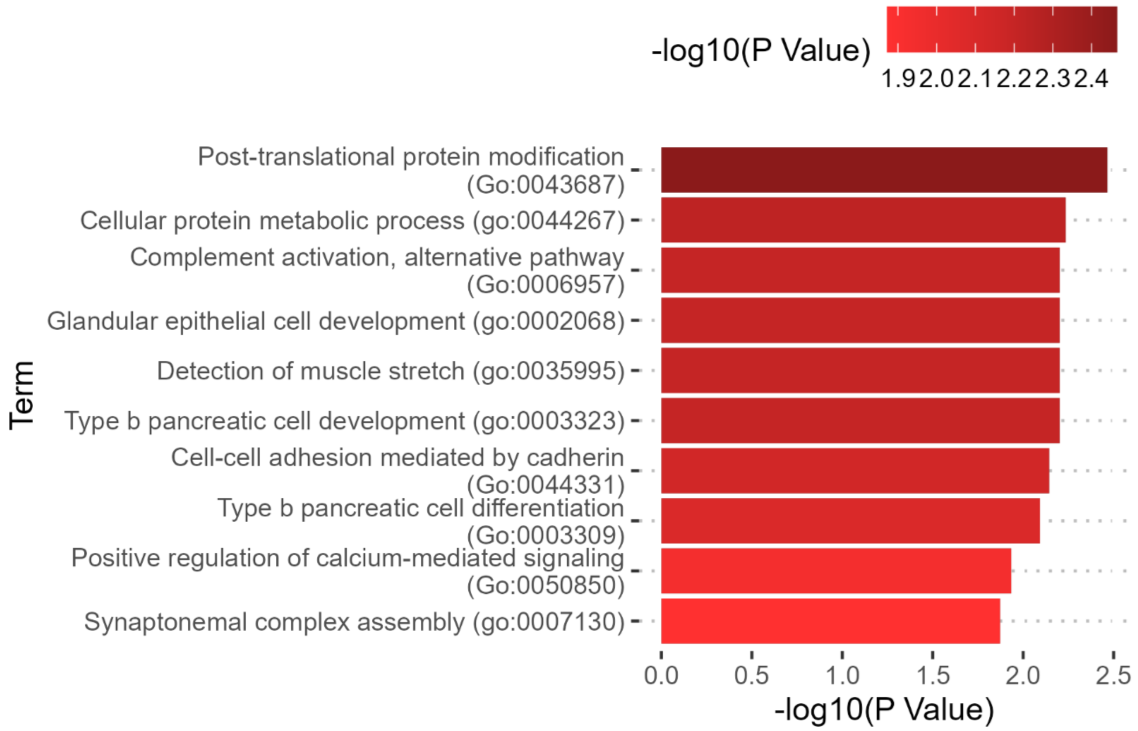


Figure 3: Significantly downregulated biological processes.

Table 2: GO Biological process upregulated

Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	Genes
hemoglobin metabolic process (GO:0020027)	5-Jan	0.003745	0.061199	0	0	356.8036	1993.609	ALAS2
oxygen homeostasis (GO:0032364)	6-Jan	0.004492	0.061199	0	0	285.4286	1542.868	ALAS2
gas homeostasis (GO:0033483)	7-Jan	0.005239	0.061199	0	0	237.8452	1249.078	ALAS2
fatty acid elongation, monounsaturated fatty acid (GO:0034625)	7-Jan	0.005239	0.061199	0	0	237.8452	1249.078	ELOVL6

fatty acid elongation, polyunsaturated fatty acid (GO:0034626)	7-Jan	0.005239	0.061199	0	0	237.8452	1249.078	ELOVL6
fatty acid elongation, saturated fatty acid (GO:0019367)	7-Jan	0.005239	0.061199	0	0	237.8452	1249.078	ELOVL6
fatty acid elongation, unsaturated fatty acid (GO:0019368)	7-Jan	0.005239	0.061199	0	0	237.8452	1249.078	ELOVL6
organonitrogen compound biosynthetic process (GO:1901566)	2/158	0.006087	0.061199	0	0	19.55523	99.76428	ALAS2;ELOVL6
negative regulation of protein activation cascade (GO:2000258)	9-Jan	0.006731	0.061199	0	0	178.3661	892.0131	F10
erythrocyte development (GO:0048821)	11-Jan	0.008221	0.061199	0	0	142.6786	685.0073	ALAS2
endoplasmic reticulum to Golgi vesicle-mediated transport (GO:0006888)	2/185	0.008255	0.061199	0	0	16.64733	79.85533	F10;TMED6
very long-chain fatty acid biosynthetic process (GO:0042761)	13-Jan	0.009709	0.061199	0	0	118.8869	551.0048	ELOVL6
fatty acid elongation (GO:0030497)	13-Jan	0.009709	0.061199	0	0	118.8869	551.0048	ELOVL6
protein trimerization (GO:0070206)	15-Jan	0.011195	0.061199	0	0	101.8929	457.7328	ITLN1
protein homotrimerization (GO:0070207)	15-Jan	0.011195	0.061199	0	0	101.8929	457.7328	ITLN1

long-chain fatty-acyl-CoA biosynthetic process (GO:0035338)	18-Jan	0.01342	0.066045	0	0	83.89916	361.6911	ELOVL6
regulation of insulin-like growth factor receptor signaling pathway (GO:0043567)	19-Jan	0.01416	0.066045	0	0	79.23413	337.3238	IGFBP6
myeloid cell development (GO:0061515)	20-Jan	0.014901	0.066045	0	0	75.06015	315.733	ALAS2
hydrogen peroxide catabolic process (GO:0042744)	21-Jan	0.01564	0.066045	0	0	71.30357	296.4745	HBG2
heme biosynthetic process (GO:0006783)	22-Jan	0.016379	0.066045	0	0	67.90476	279.2073	ALAS2
porphyrin-containing compound biosynthetic process (GO:0006779)	24-Jan	0.017856	0.066045	0	0	61.99379	249.552	ALAS2
long-chain fatty-acyl-CoA metabolic process (GO:0035336)	24-Jan	0.017856	0.066045	0	0	61.99379	249.552	ELOVL6
regulation of rhodopsin mediated signaling pathway (GO:0022400)	25-Jan	0.018593	0.066045	0	0	59.40774	236.7377	PPEF1
positive regulation of glucose import (GO:0046326)	26-Jan	0.01933	0.066045	0	0	57.02857	225.04	ITLN1
hydrogen peroxide metabolic process (GO:0042743)	29-Jan	0.021538	0.067643	0	0	50.91071	195.3923	HBG2
long-chain fatty acid biosynthetic process (GO:0042759)	30-Jan	0.022273	0.067643	0	0	49.15271	186.996	ELOVL6
fatty-acyl-CoA biosynthetic	30-Jan	0.022273	0.067643	0	0	49.15271	186.996	ELOVL6

process (GO:0046949)								
positive regulation of glucose transmembrane transport (GO:0010828)	Jan-32	0.023741	0.069204	0	0	45.97696	171.979	ITLN1
very long-chain fatty acid metabolic process (GO:0000038)	Jan-33	0.024474	0.069204	0	0	44.53795	165.2413	ELOVL6
regulation of glucose import (GO:0046324)	Jan-37	0.027403	0.074901	0	0	39.58135	142.3785	ITLN1
negative regulation of blood coagulation (GO:0030195)	Jan-40	0.029594	0.07828	0	0	36.53114	128.5967	F10
erythrocyte differentiation (GO:0030218)	Jan-46	0.033962	0.087026	0	0	31.65079	107.0596	ALAS2
myeloid cell differentiation (GO:0030099)	Jan-52	0.038311	0.095197	0	0	27.91877	91.07146	ALAS2
cellular iron ion homeostasis (GO:0006879)	Jan-58	0.042642	0.099905	0	0	24.97243	78.7857	ALAS2
membrane lipid biosynthetic process (GO:0046467)	Jan-58	0.042642	0.099905	0	0	24.97243	78.7857	ELOVL6
iron ion homeostasis (GO:0055072)	Jan-61	0.044801	0.102047	0	0	23.72024	73.66371	ALAS2
fatty acid biosynthetic process (GO:0006633)	Jan-71	0.051964	0.115164	0	0	20.32143	60.09446	ELOVL6
sphingolipid biosynthetic process (GO:0030148)	Jan-74	0.054104	0.11675	0	0	19.48337	56.83016	ELOVL6
regulation of G protein-coupled receptor signaling	Jan-82	0.059786	0.124013	0	0	17.55203	49.44378	PPEF1

pathway (GO:0008277)								
long-chain fatty acid metabolic process (GO:0001676)	Jan-83	0.060494	0.124013	0	0	17.33711	48.63423	ELOVL6
regulation of lipid metabolic process (GO:0019216)	Jan-92	0.066844	0.133689	0	0	15.61538	42.2457	ELOVL6
cellular transition metal ion homeostasis (GO:0046916)	Jan-96	0.069654	0.134165	0	0	14.95489	39.8431	ALAS2
positive regulation of cold-induced thermogenesis (GO:0120162)	Jan-97	0.070355	0.134165	0	0	14.79836	39.27787	ELOVL6
positive regulation of metabolic process (GO:0009893)	1/113	0.081506	0.151897	0	0	12.67411	31.77499	ELOVL6
sphingolipid metabolic process (GO:0006665)	1/116	0.083583	0.152306	0	0	12.34161	30.63087	ELOVL6
protein homooligomerization (GO:0051260)	1/121	0.087035	0.155149	0	0	11.8244	28.86869	ITLN1
Golgi organization (GO:0007030)	1/130	0.093217	0.155996	0	0	10.99446	26.08792	TMED6
regulation of response to external stimulus (GO:0032101)	1/130	0.093217	0.155996	0	0	10.99446	26.08792	PPEF1
regulation of primary metabolic process (GO:0080090)	1/130	0.093217	0.155996	0	0	10.99446	26.08792	ELOVL6
protein dephosphorylation (GO:0006470)	1/139	0.099361	0.162951	0	0	10.27277	23.71983	PPEF1
dephosphorylation (GO:0016311)	1/153	0.10884	0.174998	0	0	9.320019	20.67065	PPEF1
skeletal system development (GO:0001501)	1/158	0.112203	0.17671	0	0	9.020928	19.7328	PAX1

positive regulation of protein kinase B signaling (GO:0051897)	1/161	0.114215	0.17671	0	0	8.850446	19.20258	F10
negative regulation of canonical Wnt signaling pathway (GO:0090090)	1/165	0.116891	0.177501	0	0	8.63284	18.53051	IGFBP6
negative regulation of Wnt signaling pathway (GO:0030178)	1/191	0.134102	0.199935	0	0	7.441729	14.95156	IGFBP6
regulation of signal transduction (GO:0009966)	1/198	0.138682	0.200447	0	0	7.174764	14.17424	IGFBP6
endomembrane system organization (GO:0010256)	1/199	0.139335	0.200447	0	0	7.138167	14.06843	TMED6
regulation of protein kinase B signaling (GO:0051896)	1/207	0.144538	0.204347	0	0	6.858183	13.26519	F10
positive regulation of protein modification process (GO:0031401)	1/214	0.149066	0.206442	0	0	6.630449	12.62015	ITLN1
transcription, DNA-templated (GO:0006351)	1/221	0.153573	0.206442	0	0	6.417208	12.02317	PAX1
positive regulation of cell motility (GO:2000147)	1/221	0.153573	0.206442	0	0	6.417208	12.02317	F10
regulation of canonical Wnt signaling pathway (GO:0060828)	1/253	0.17389	0.226333	0	0	5.593254	9.784464	IGFBP6
positive regulation of phosphorylation (GO:0042327)	1/253	0.17389	0.226333	0	0	5.593254	9.784464	ITLN1
regulation of protein phosphorylation (GO:0001932)	1/266	0.182013	0.231968	0	0	5.315364	9.055663	ITLN1

positive regulation of cell migration (GO:0030335)	1/269	0.183877	0.231968	0	0	5.255064	8.89939	F10
cellular macromolecule biosynthetic process (GO:0034645)	1/314	0.211365	0.262605	0	0	4.489274	6.977093	ALAS2
transcription by RNA polymerase II (GO:0006366)	1/320	0.214964	0.26309	0	0	4.403493	6.769424	PAX1
cellular protein localization (GO:0034613)	1/329	0.220334	0.265697	0	0	4.280706	6.475046	TMED6
intracellular protein transport (GO:0006886)	1/336	0.224487	0.266781	0	0	4.189765	6.259256	TMED6
positive regulation of multicellular organismal process (GO:0051240)	1/345	0.229796	0.269189	0	0	4.078281	5.997377	ELOVL6
protein transport (GO:0015031)	1/369	0.243788	0.278963	0	0	3.807648	5.374334	TMED6
positive regulation of protein phosphorylation (GO:0001934)	1/371	0.244943	0.278963	0	0	3.78668	5.326837	ITLN1
negative regulation of cell population proliferation (GO:0008285)	1/379	0.249547	0.280313	0	0	3.705026	5.142973	IGFBP6
regulation of cell migration (GO:0030334)	1/408	0.26602	0.294268	0	0	3.435942	4.549824	F10
cellular protein metabolic process (GO:0044267)	1/417	0.271063	0.294268	0	0	3.360062	4.386241	IGFBP6
organelle organization (GO:0006996)	1/420	0.272737	0.294268	0	0	3.335493	4.333635	TMED6
positive regulation of intracellular signal transduction (GO:1902533)	1/546	0.339884	0.361955	0	0	2.547837	2.749499	F10

negative regulation of cellular process (GO:0048523)	1/566	0.349995	0.367943	0	0	2.45512	2.577477	IGFBP6
regulation of cell population proliferation (GO:0042127)	1/764	0.442581	0.459388	0	0	1.799476	1.466808	IGFBP6
cellular protein modification process (GO:0006464)	1/1025	0.545897	0.559545	0	0	1.322614	0.800611	PPEF1
regulation of transcription by RNA polymerase II (GO:0006357)	Jan-06	0.826869	0.832335	0	0	0.575964	0.109496	PAX1
regulation of transcription, DNA-templated (GO:0006355)	Jan-44	0.832335	0.832335	0	0	0.564996	0.103688	PAX1

Table 8: GO Biological process upregulated

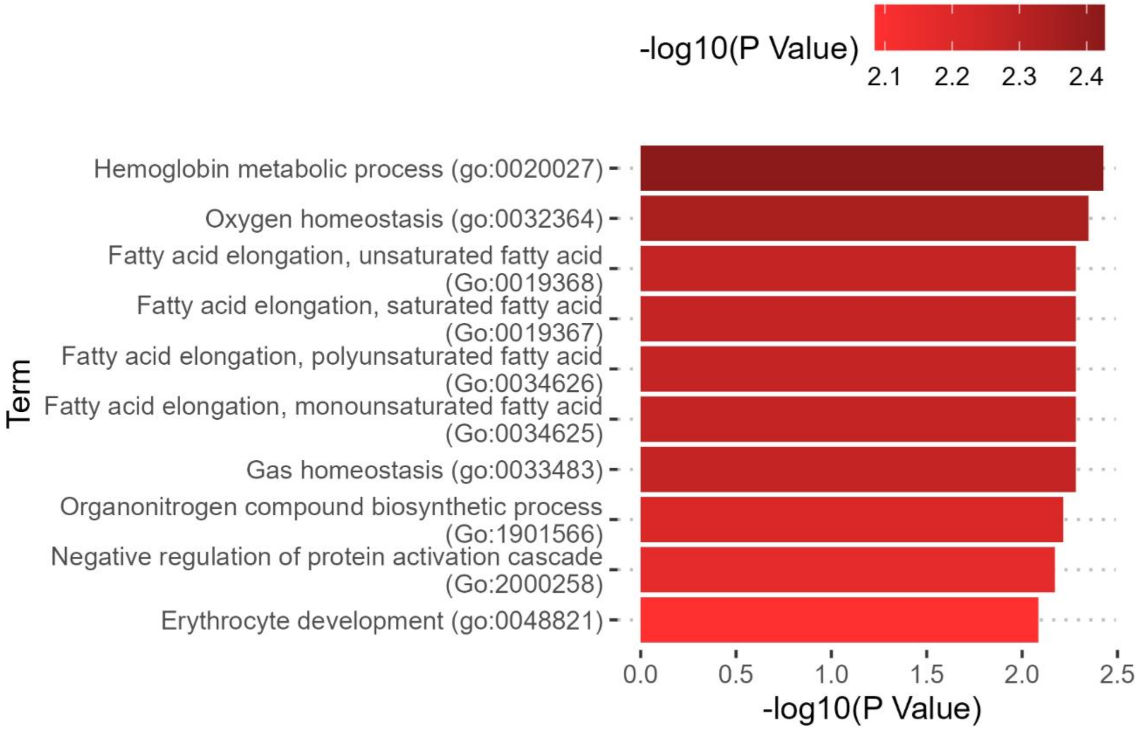


Figure 4: Significantly upregulated biological processes.

Table 3: Go molecular function downregulated

Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	Genes
opioid receptor binding (GO:0031628)	5-Jan	0.004492	0.043071	0	0	293.7941	1588.072	PENK
extracellularly ATP-gated cation channel activity (GO:0004931)	7-Jan	0.006284	0.043071	0	0	195.8431	992.8793	P2RX5
nucleotide receptor activity (GO:0016502)	7-Jan	0.006284	0.043071	0	0	195.8431	992.8793	P2RX5
ATP-gated ion channel activity (GO:0035381)	8-Jan	0.007179	0.043071	0	0	167.8571	828.6538	P2RX5
purinergic nucleotide receptor activity (GO:0001614)	15-Jan	0.01342	0.064088	0	0	83.89916	361.6911	P2RX5
pyridoxal phosphate binding (GO:0030170)	21-Jan	0.01874	0.064088	0	0	58.71176	233.5022	ALB
excitatory extracellular ligand-gated ion channel activity (GO:0005231)	21-Jan	0.01874	0.064088	0	0	58.71176	233.5022	P2RX5
neuropeptide hormone activity (GO:0005184)	26-Jan	0.023153	0.064088	0	0	46.95765	176.8256	PENK
glutathione transferase activity (GO:0004364)	27-Jan	0.024033	0.064088	0	0	45.14932	168.3313	GSTM5
copper ion binding (GO:0005507)	Jan-45	0.039751	0.088619	0	0	26.65508	85.96583	ALB
channel activity (GO:0015267)	Jan-46	0.040617	0.088619	0	0	26.06144	83.48948	P2RX5
hormone activity (GO:0005179)	Jan-78	0.067948	0.125126	0	0	15.20626	40.88978	PENK

ligand-gated cation channel activity (GO:0099094)	Jan-80	0.069632	0.125126	0	0	14.81981	39.48788	P2RX5
ion channel activity (GO:0005216)	Jan-84	0.07299	0.125126	0	0	14.10276	36.91299	P2RX5
G protein-coupled receptor binding (GO:0001664)	1/143	0.121217	0.193947	0	0	8.218724	17.34294	PENK
ATP binding (GO:0005524)	1/278	0.222804	0.325209	0	0	4.18454	6.282935	P2RX5
adenyl ribonucleotide binding (GO:0032559)	1/306	0.242436	0.325209	0	0	3.794986	5.377566	P2RX5
cadherin binding (GO:0045296)	1/322	0.253443	0.325209	0	0	3.602895	4.945398	CDH2
zinc ion binding (GO:0008270)	1/336	0.26295	0.325209	0	0	3.449868	4.608308	ALB
calcium ion binding (GO:0005509)	1/348	0.271007	0.325209	0	0	3.32853	4.345759	CDH2
transition metal ion binding (GO:0046914)	1/445	0.333154	0.373424	0	0	2.5885	2.845151	ALB
purine ribonucleoside triphosphate binding (GO:0035639)	1/460	0.342305	0.373424	0	0	2.501986	2.682259	P2RX5
metal ion binding (GO:0046872)	1/517	0.376011	0.392359	0	0	2.219106	2.170592	CDH2
DNA binding (GO:0003677)	1/811	0.525471	0.525471	0	0	1.392302	0.89589	ALB

Table 9: Go molecular function downregulated

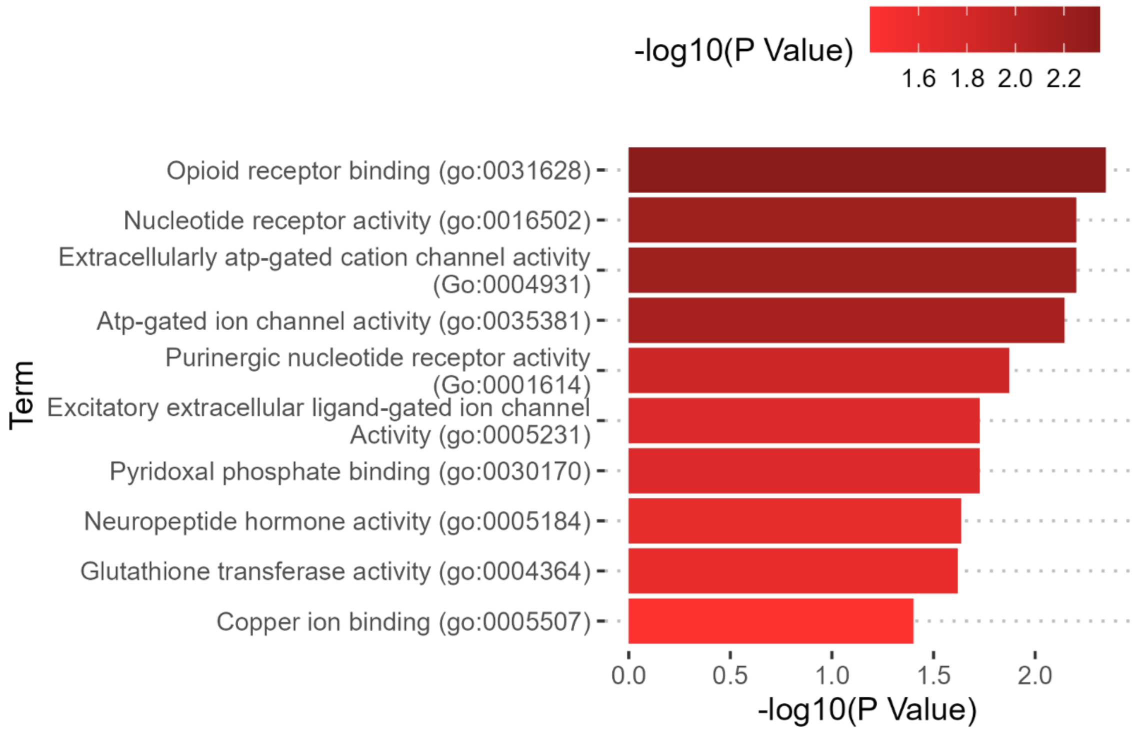


Figure 5: Significantly downregulated molecular functions.

Table 4: Go molecular function upregulated

Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	Genes
hemoglobin alpha binding (GO:0031721)	5-Jan	0.003745	0.031401	0	0	356.8036	1993.609	HBG2
insulin-like growth factor II binding (GO:0031995)	7-Jan	0.005239	0.031401	0	0	237.8452	1249.078	IGFBP6
fatty acid elongase activity (GO:0009922)	7-Jan	0.005239	0.031401	0	0	237.8452	1249.078	ELOVL6
interleukin-17 receptor activity (GO:0030368)	8-Jan	0.005985	0.031401	0	0	203.8571	1043.434	IL17REL
fatty acid synthase activity (GO:0004312)	10-Jan	0.007476	0.031401	0	0	158.5397	776.2124	ELOVL6

insulin-like growth factor I binding (GO:0031994)	13-Jan	0.009709	0.033585	0	0	118.8869	551.0048	IGFBP6
insulin-like growth factor binding (GO:0005520)	15-Jan	0.011195	0.033585	0	0	101.8929	457.7328	IGFBP6
protein serine/threonine phosphatase activity (GO:0004722)	Jan-62	0.04552	0.119489	0	0	23.33021	72.08123	PPEF1
cytokine receptor activity (GO:0004896)	Jan-88	0.064027	0.138895	0	0	16.33662	44.9004	IL17REL
heme binding (GO:0020037)	Jan-91	0.066141	0.138895	0	0	15.78968	42.88433	HBG2
serine-type endopeptidase activity (GO:0004252)	1/105	0.075946	0.144988	0	0	13.65453	35.19772	F10
serine-type peptidase activity (GO:0008236)	1/125	0.089787	0.157128	0	0	11.44067	27.57559	F10
endopeptidase activity (GO:0004175)	1/315	0.211966	0.342406	0	0	4.47475	6.941816	F10
calcium ion binding (GO:0005509)	1/348	0.231558	0.347337	0	0	4.042404	5.913739	ITLN1
metal ion binding (GO:0046872)	1/517	0.324963	0.454949	0	0	2.695044	3.029345	ITLN1
double-stranded DNA binding (GO:0003690)	1/651	0.391372	0.48946	0	0	2.124725	1.993195	PAX1
sequence-specific DNA binding (GO:0043565)	1/707	0.417275	0.48946	0	0	1.950526	1.704777	PAX1
sequence-specific double-stranded DNA binding (GO:1990837)	1/712	0.419537	0.48946	0	0	1.936307	1.681881	PAX1
RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978)	1/1149	0.588445	0.617867	0	0	1.172038	0.621499	PAX1
cis-regulatory region sequence-specific DNA binding (GO:0000987)	1/1149	0.588445	0.617867	0	0	1.172038	0.621499	PAX1

RNA polymerase II transcription regulatory region sequence-specific DNA binding (GO:0000977)	1/1359	0.652125	0.652125	0	0	0.97975	0.418861	PAX1

Table 10: Go molecular function upregulated

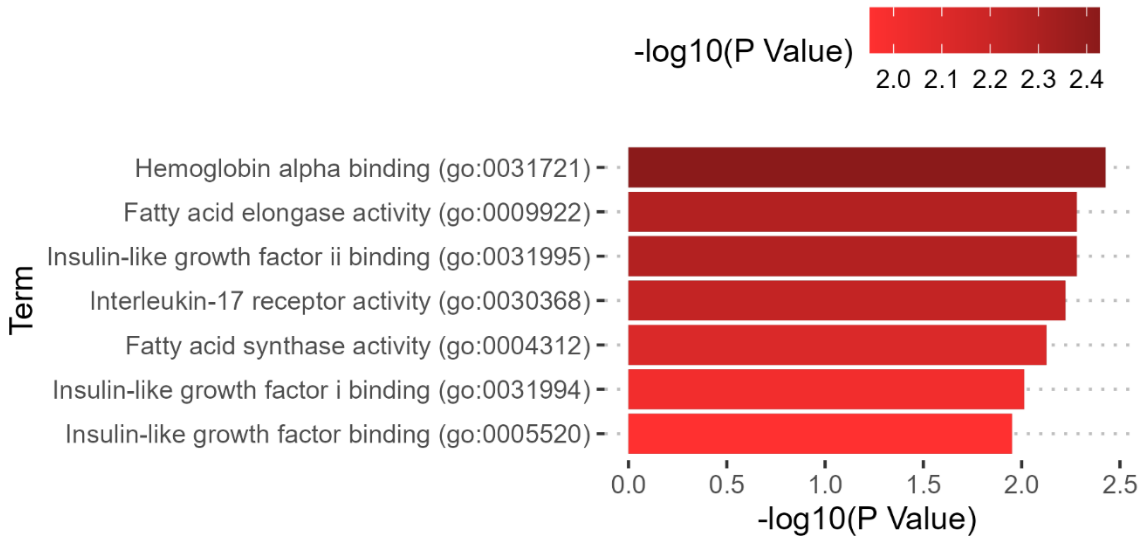


Figure 6: Significantly upregulated molecular functions.

Table 5: KEGG 2021 Downregulated

Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	Genes
Neuroactive ligand-receptor interaction	2/341	0.037046	0.174474	0	0	7.242994	23.86989	P2RX5; PENK
Glutathione metabolism	Jan-57	0.050096	0.174474	0	0	20.93067	62.66234	GSTM5
Thyroid hormone synthesis	Jan-75	0.065418	0.174474	0	0	15.82512	43.15455	ALB
Metabolism of	Jan-76	0.066262	0.174474	0	0	15.61333	42.3768	GSTM5

xenobiotics by cytochrome P450								
Arrhythmogenic right ventricular cardiomyopathy	Jan-77	0.067105	0.174474	0	0	15.40712	41.6222	CDH2
Drug metabolism	1/108	0.092902	0.201287	0	0	10.92633	25.96328	GSTM5
Fluid shear stress and atherosclerosis	1/139	0.118024	0.203439	0	0	8.458653	18.07505	GSTM5
Cell adhesion molecules	1/148	0.125193	0.203439	0	0	7.937175	16.49265	CDH2
Hepatocellular carcinoma	1/168	0.140928	0.203563	0	0	6.97957	13.67649	GSTM5
Chemical carcinogenesis	1/239	0.194657	0.230917	0	0	4.879881	7.985993	GSTM5
Calcium signaling pathway	1/240	0.195391	0.230917	0	0	4.859217	7.933898	P2RX5
Herpes simplex virus 1 infection	1/498	0.364961	0.384036	0	0	2.30619	2.324558	CFP
Pathways in cancer	1/531	0.384036	0.384036	0	0	2.158935	2.066142	GSTM5

Table 11: KEGG 2021 Downregulated

Table 6: KEGG 2021 upregulated

Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	Genes
Biosynthesis of unsaturated fatty acids	27-Jan	0.020067	0.043007	0	0	54.83242	214.3235	ELOVL6
Fatty acid elongation	27-Jan	0.020067	0.043007	0	0	54.83242	214.3235	ELOVL6

African trypanosomiasis	Jan-37	0.027403	0.043007	0	0	39.58135	142.3785	HBA2
Glycine, serine and threonine metabolism	Jan-40	0.029594	0.043007	0	0	36.53114	128.5967	ALAS2
Porphyrin and chlorophyll metabolism	Jan-43	0.03178	0.043007	0	0	33.91667	116.9759	ALAS2
Malaria	Jan-50	0.036863	0.043007	0	0	29.06122	95.91772	HBA2
Complement and coagulation cascades	Jan-85	0.061909	0.061909	0	0	16.92262	47.08033	F10

Table 12: KEGG 2021 upregulated

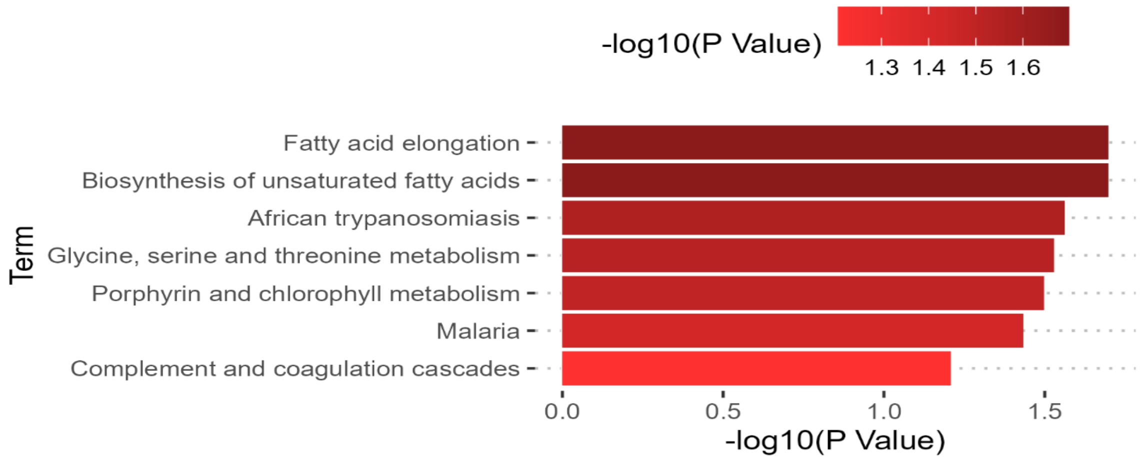


Figure 7: Significantly upregulated KEGG human pathways.

4.3 DISCUSSION

After complete the analyzation phase, I have found 15 upregulated gene and 18 downregulated gene. AL158154.2, F10, ELOVL6, PPEF1, IGFBP6, PAX1, AC011893.1, TMED6, AC090877.2, FAM163A, HBG2, ITLN1, HBA2, ALAS2, IL17REL are upregulated while MTND1P23, SCARNA5, SCARNA10, SCARNA7, PENK, GSTM5, RF00100, LINC00052, LINC01281, CD300LB, CDH2, CFP, LINC01099, AF178030.1, ALB, P2RX5, TEX11, FCRL4 are downregulated.

In breast cancer, Elovl6 is a poor prognostic predictor. In addition, Elovl6 has been linked to insulin resistance, obesity, and lipogenesis. Furthermore, the protein has been linked to nonalcoholic steatohepatitis-associated liver carcinogenesis and is increased in human hepatocellular carcinoma. Positive Elov16 expression was linked to lymph node involvement and a shorter recurrence-free survival period. Elovl6 expression, on the other hand, had no relation to the size of the initial tumor, lymph node metastases, stage, grade, estrogen receptor, progesterone receptor, HER2, or age. As a result, positive Elovl6 expression is a poor predictive indicator in patients with breast cancer who have previously had surgery, and it may one day be used as a therapeutic method, especially in the context of obesity-related illness[23]. The PPEF1 gene, which encodes a protein serine/threonine protein phosphatase, was discovered on chromosome Xp22, which has been linked to excessive cell proliferation, growth, and signaling. Furthermore, overexpression of PPEF1 boosted the tumorigenic growth of A549 cells, indicating that PPEF1 might operate as an oncogene in the formation of lung cancer by inhibiting cancer cell death. PPEF1 has also been identified as a possible target for lymphoma diagnosis and treatment, according to current studies[24]. In NCI-H1299 cells, IGFBP-6 gene is an effector of SEMA3B's tumor suppressor function. IGFBP-6 expression is suppressed by -catenin, according to research. IGFBP-6 expression was lower in normal lung tissue and linked favorably with SEMA3B expression[25]. The PAX1 gene, which has a paired domain (PD) and an octapeptide domain, is found on chromosome 20p11 (OP). It is essential for the growth and development of the bone, spine, thymus, and parathyroid gland. When comparing cervical cancer tissues to normal cervical tissues, the PAX1 gene is considerably hypermethylated (PAX1m), and the methylation level correlates favorably with tumor grade[26]. Gene

TMED6 is a potential target which found by only eQTL analysis. Its promote to colon cancer[27]. FAM163A, also known as neuroblastoma-derived secretory protein (NDSP) or C1ORF76, is a 167-amino-acid protein with a potential signal peptide that is found on chromosome 1q25.2. FAM163A was shown to be overexpressed at a greater level in neuroblastoma than in other human malignancies in previous investigations. Apart from neuroblastoma, the expression of FAM163A has been studied in a number of lung cancer cell lines, with only weak expression[28]. The ITLN1 gene is a secretory protein that has been linked to a better prognosis in individuals with advanced ovarian cancer, according to a recent study. Further research shows that ITLN1 reduces MMP1 expression and induces a metabolic change in metastatic ovarian cancer cells, suppressing lactotransferrin's influence on ovarian cancer cell invasion and proliferation. Furthermore, tumor growth rates in ovarian cancer-bearing mice treated with ITLN1 are significantly reduced[29]. In PCa metastatic samples, HBA2 is abundantly expressed. The rise in HBA2 might be linked to a higher burden of labile heme in the tumor niche, which could come from erythrocytes, dying cancer cells, or healthy cells[30]. According to the Human Protein Atlas, ALAS2 is expressed in 16 percent of lung cancer tissues. As a result, the amount of ALAS2 transcript in lung cells was also assessed. In HCC4017 cells, the amount of ALAS2 transcript was roughly five-fold higher[31].

In vitro and in vivo, the effects of lentivirus-mediated knockdown or overexpression of SCARNA10 on liver fibrosis were investigated. Furthermore, the impacts and mechanisms of SCARNA10 down-regulation or over-expression on the expression of TGF pathway genes were investigated. SCARNA10 transcript levels were higher in the serum and liver of individuals with advanced hepatic fibrosis. SCARNA10 acted as an unique positive regulator of TGF signaling in hepatic fibrogenesis by blocking PRC2 binding to the promoters of genes involved in the ECM and TGF pathways, allowing these genes to be transcribed[32]. GSTM genes are detoxifying enzymes that are involved in the deactivation of carcinogenic reactive metabolites, implying that these enzymes may play a role in carcinogenesis. A survival study of patients with gastric cancer who had varying amounts of GSTM expression is given. Patients with gastric cancer who had high GSTM5 expression had a significantly worse prognosis[33]. In several malignancies, such as breast cancer, gastric cancer, liver cancer, and lung cancer, the biological activity of long intergenic nonproteincoding RNA 52 (LINC00052) indicates that this gene can behave as either an oncogene or a tumor suppressor. When comparing colorectal cancer tissues to their surrounding tissues, LINC00052 was shown to be downregulated. In addition, both in vivo and in vitro, LINC00052 inhibited the spread of colorectal cancer cells[34]. Cadherin 2 (CDH2) is a member of the cadherin superfamily that encodes the N-cadherin protein, a traditional cadherin that maintains cell integrity and is involved in various cell signaling pathways. The expression of CDH2 has been found to be highly linked to glioma. Patients with reduced CDH2 expression had a better prognosis and were more likely to respond to temozolomide treatment. It's used to grade and treat glioma as a prognostic and predictive molecular biomarker[35]. Complement factor properdin (CFP), which encodes plasma glycoprotein, is a key regulator of the innate immune system's complement cascade. Low CFP expression was associated with poorer overall survival (OS), first progression (FP), and post progression survival (PPS), and was detrimental to the prognosis of STAD and LUAD, particularly in stages 3, T3, N2, and N3 of STAD (P0.05). Furthermore, in STAD and LUAD, CFP expression exhibited substantial positive relationships with the numbers of CD8+ T cells, CD4+ T cells, macrophages, neutrophils, and dendritic cells (DCs). In STAD and LUAD, CFP can be used as a predictive biomarker to determine prognosis and immune infiltration[36]. Tex11, an X-linked meiosis-specific gene, promotes chromosomal synapsis and meiotic recombination. Infertile males with non-obstructive azoospermia have a mutation in TEX11, and a mouse with a similar mutation has problems with meiosis[37]. FCRL4 is an immunoglobulin receptor superfamily member and one of numerous Fc receptor-like glycoproteins grouped on chromosome 1's long arm. Four extracellular C2-type immunoglobulin domains, a transmembrane domain, and a cytoplasmic domain with three immune-receptor tyrosine-based inhibitory motifs make up the encoded protein. This protein may play a part in the epithelia's memory B-cell function. Non-Hodgkin lymphoma and multiple myeloma have been linked to chromosomal abnormalities encoding this gene[38].

CHAPTER 5
CONCLUTION

The research goal was accomplished satisfactorily. With a recent work, we discovered that tumor and non-tumor patients in breast cancer had differently expressed genes. And in this investigation, we discovered unique genes that are expressed at every stage of breast cancer. We will strive to cover the protein-protein interaction (PPI) network using this dataset in future research and find hub genes that will affect the cancer pathway.

Reference

[1] "Breast cancer in women - NHS." 2019, Accessed: Jan. 08, 2022. [Online]. Available: https://www.nhs.uk/conditions/breast-cancer/.

[2] A. Felman, "Breast cancer: Symptoms, causes, stages, types, and more," What to know about breast cancer, 2021. https://www.medicalnewstoday.com/articles/37136 (accessed Jan. 08, 2022).

[3] "Central database, mass awareness to address breast cancer in Bangladesh." https://www.aa.com.tr/en/asia-pacific/central-database-mass-awareness-to-address-breast-cancer-in-bangladesh/2401288 (accessed Jan. 05, 2022).

[4] "What Is Cancer? - National Cancer Institute." https://www.cancer.gov/about-cancer/understanding/what-is-cancer (accessed Jan. 05, 2022).

[5] D. of C. P. and C. Centers for Disease Control and Prevention, "What Is Breast Cancer? | CDC," Sept. 14, pp. 1–19, 2020, Accessed: Jan. 05, 2022. [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm.

[6] B. A. Mandal, "Breast Cancer Pathophysiology," Breast Cancer Pathophysiol.,

pp. 1–3, 2019, Accessed: Jan. 05, 2022. [Online]. Available: https://www.news-medical.net/health/Breast-Cancer-Pathophysiology.aspx.

[7] T. N. Seyfried and L. C. Huysentruyt, "On the origin of cancer metastasis," Crit. Rev. Oncog., vol. 18, no. 1–2, pp. 43–73, 2013, doi: 10.1615/CritRevOncog.v18.i1-2.40.

[8] National Cancer Institution at the National Institutes of Health, "The Genetics of Cancer - National Cancer Institute," 2015. https://www.cancer.gov/about-cancer/causes-prevention/genetics (accessed Jan. 05, 2022).

[9] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," Cell, vol. 144, no. 5. Cell, pp. 646–674, Mar. 04, 2011, doi: 10.1016/j.cell.2011.02.013.

[10] National Cancer Institute, "Cancer Staging - National Cancer Institute," NCI Cancer Staging. 2015, Accessed: Jan. 05, 2022. [Online]. Available: https://www.cancer.gov/about-cancer/diagnosis-staging/staging.

[11] M. S. Rao et al., "Comparison of RNA-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies," Front. Genet., vol. 10, no. JAN, p. 636, 2019, doi: 10.3389/fgene.2018.00636.

[12] Z. Fang, J. Martin, and Z. Wang, "Statistical methods for identifying differentially expressed genes in RNA-Seq experiments," Cell and Bioscience, vol. 2, no. 1. BioMed Central, pp. 1–8, Jul. 31, 2012, doi: 10.1186/2045-3701-2-26.

[13] R. Bell, R. Barraclough, and O. Vasieva, "Gene Expression Meta-Analysis of Potential Metastatic Breast Cancer Markers," Curr. Mol. Med., vol. 17, no. 3, Aug. 2017, doi: 10.2174/1566524017666170807144946.

[14] "Differential gene expression | definition of differential gene expression by Medical dictionary."https://medical-dictionary.thefreedictionary.com/differential+gene+expression (accessed Jan. 05, 2022).

[15] M. Nacht et al., "Combining Serial Analysis of Gene Expression and Array Technologies to Identify Genes Differentially Expressed in Breast Cancer 1," 1999. [Online]. Available: http://www.lbl.gov/LBL-Programs/mrgs/review.html.

[16] X. J. Ma, S. Dahiya, E. Richardson, M. Erlander, and D. C. Sgroi, "Gene expression profiling of the tumor microenvironment during breast cancer progression," Breast Cancer Res., vol. 11, no. 1, Feb. 2009, doi: 10.1186/bcr2222.

[17] A. C. Vargas et al., "Gene expression profiling of tumour epithelial and stromal compartments during breast cancer progression," Breast Cancer Res. Treat., vol. 135, no. 1, pp. 153–165, 2012, doi: 10.1007/s10549-012-2123-4.

[18] R. Bell, R. Barraclough, and O. Vasieva, "Gene Expression Meta-Analysis of Potential Metastatic Breast Cancer Markers," Curr. Mol. Med., vol. 17, no. 3, Aug. 2017, doi: 10.2174/1566524017666170807144946.

[19] E. S. Knudsen, A. Ertel, E. Davicioni, J. Kline, G. F. Schwartz, and A. K. Witkiewicz, "Progression of ductal carcinoma in situ to invasive breast cancer is associated with gene expression programs of EMT and myoepithelia," Breast Cancer Res. Treat., vol. 133, no. 3, pp. 1009–1024, Jun. 2012, doi: 10.1007/s10549-011-1894-3.

[20] C. F. Singer et al., "Differential gene expression profile in breast cancer-derived stromal fibroblasts," Breast Cancer Res. Treat., vol. 110, no. 2, pp. 273–281, Jul. 2008,

doi: 10.1007/s10549-007-9725-2.

[21] K. S. Wilson, H. Roberts, R. Leek, A. L. Harris, and J. Geradts, "Differential gene expression patterns in HER2/neu-positive and -negative breast cancer cell lines and tissues," Am. J. Pathol., vol. 161, no. 4, pp. 1171–1185, 2002, doi: 10.1016/S0002-9440(10)64394-5.

[22] S. Malvia et al., "Study of Gene Expression Profiles of Breast Cancers in Indian Women," Sci. Rep., vol. 9, no. 1, pp. 1–15, Jul. 2019, doi: 10.1038/s41598-019-46261-1.

[23] Feng, Y. H., Chen, W. Y., Kuo, Y. H., Tung, C. L., Tsao, C. J., Shiau, A. L., & Wu, C. L. (2016). Elovl6 is a poor prognostic predictor in breast cancer. Oncology letters, 12(1), 207–212. https://doi.org/10.3892/ol.2016.4587

[24] Ye, T., Wan, X., Li, J., Feng, J., Guo, J., Li, G., & Liu, J. (2020). The Clinical Significance of PPEF1 as a Promising Biomarker and Its Potential Mechanism in Breast Cancer. OncoTargets and therapy, 13, 199–214. https://doi.org/10.2147/OTT.S229432

[25] Koyama, N., Zhang, J., Huqun, Miyazawa, H., Tanaka, T., Su, X., & Hagiwara, K. (2008). Identification of IGFBP-6 as an effector of the tumor suppressor activity of SEMA3B. Oncogene, 27(51), 6581–6589. https://doi.org/10.1038/onc.2008.263

[26] Li, X., Zhou, X., Zeng, M., Zhou, Y., Zhang, Y., Liou, Y. L., & Zhu, H. (2021). Methylation of PAX1 gene promoter in the prediction of concurrent chemo-radiotherapy efficacy in cervical cancer. Journal of Cancer, 12(17), 5136–5143. https://doi.org/10.7150/jca.57460

[27] Yao, L., Tak, Y. G., Berman, B. P., & Farnham, P. J. (2014). Functional annotation of colon cancer risk SNPs. Nature communications, 5(1), 1-13.

[28] Liu, N., Zhou, H., Zhang, X., Cai, L., Li, J., Zhao, J., ... & Miao, Y. (2019). FAM163A, a positive regulator of ERK signaling pathway, interacts with 14-3-3β and promotes cell proliferation in squamous cell lung carcinoma. OncoTargets and therapy, 12, 6393.

[29] Au-Yeung, C. L., Yeung, T. L., Achreja, A., Zhao, H., Yip, K. P., Kwan, S. Y., ... & Mok, S. C. (2020). ITLN1 modulates invasive potential and metabolic reprogramming of ovarian cancer cells in omental microenvironment. Nature communications, 11(1), 1-16.

[30] Canesin, G., Di Ruscio, A., Li, M., Ummarino, S., Hedblom, A., Choudhury, R., ... & Wegiel, B. (2020). Scavenging of labile heme by hemopexin is a key checkpoint in cancer growth and metastases. Cell reports, 32(12), 108181.

[31] Hooda, J., Cadinu, D., Alam, M. M., Shah, A., Cao, T. M., Sullivan, L. A., ... & Zhang, L. (2013). Enhanced heme function and mitochondrial respiration promote the progression of lung cancer cells. PloS one, 8(5), e63402.

[32] Zhang, K., Han, Y., Hu, Z., Zhang, Z., Shao, S., Yao, Q., ... & Hong, W. (2019). SCARNA10, a nuclear-retained long non-coding RNA, promotes liver fibrosis and serves as a potential biomarker. Theranostics, 9(12), 3622.

[33] Chen, Y., Li, B., Wang, J., Liu, J., Wang, Z., Mao, Y., ... & Chen, J. (2020). Identification and verification of the prognostic value of the glutathione S-transferase Mu genes in gastric cancer. Oncology letters, 20(4), 1-1.

[34] Yu, G., Xiong, D., Liu, Z., Li, Y., Chen, K., & Tang, H. (2019). Long noncoding RNA LINC00052 inhibits colorectal cancer metastasis by sponging microRNA-574-5p to modulate CALCOCO1 expression. Journal of cellular biochemistry, 120(10), 17258-17272.

[35] Chen, Q., Cai, J., & Jiang, C. (2018). CDH2 expression is of prognostic significance in glioma and predicts the efficacy of temozolomide therapy in patients with glioblastoma. Oncology letters, 15(5), 7415-7422.

[36] Cui, G., Le Geng, L. Z., Lin, Z., Liu, X., Miao, Z., Jiang, J., ... & Wei, F. (2021). CFP is a prognostic biomarker and correlated with immune infiltrates in Gastric Cancer and Lung Cancer. Journal of Cancer, 12(11), 3378.

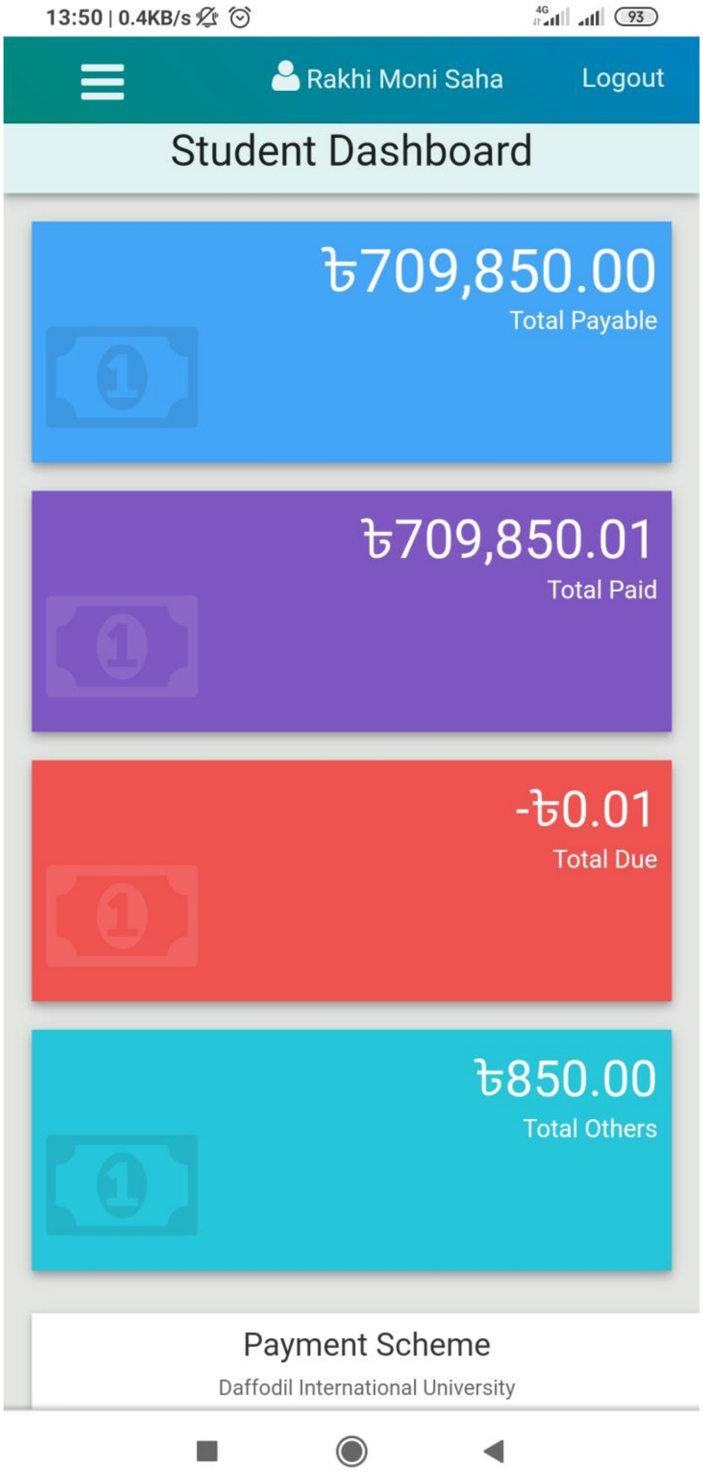
[37] Yang, F., Silber, S., Leu, N. A., Oates, R. D., Marszalek, J. D., Skaletsky, H., ... & Wang, P. J. (2015). TEX 11 is mutated in infertile men with azoospermia and regulates genome-wide recombination rates in mouse. EMBO molecular medicine, 7(9), 1198-1210.

[38] http://www.cancerindex.org/geneweb/IRTA1.htm Accessed: 2022-01-12

[39] Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nat 2012 4907418 [Internet]. 2012 Sep 23 [cited 2022 Jan 3];490(7418):61–70. Available from: https://www.nature.com/articles/nature11412

[40] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res [Internet]. 2016 May 5 [cited 2022 Jan 3];44(8):e71. Available from: https://academic.oup.com/nar/article/44/8/e71/2465925

Account Clearance:



13:50 | 0.4KB/s

Rakhi Moni Saha

Logout

Student Dashboard

৳709,850.00
Total Payable

৳709,850.01
Total Paid

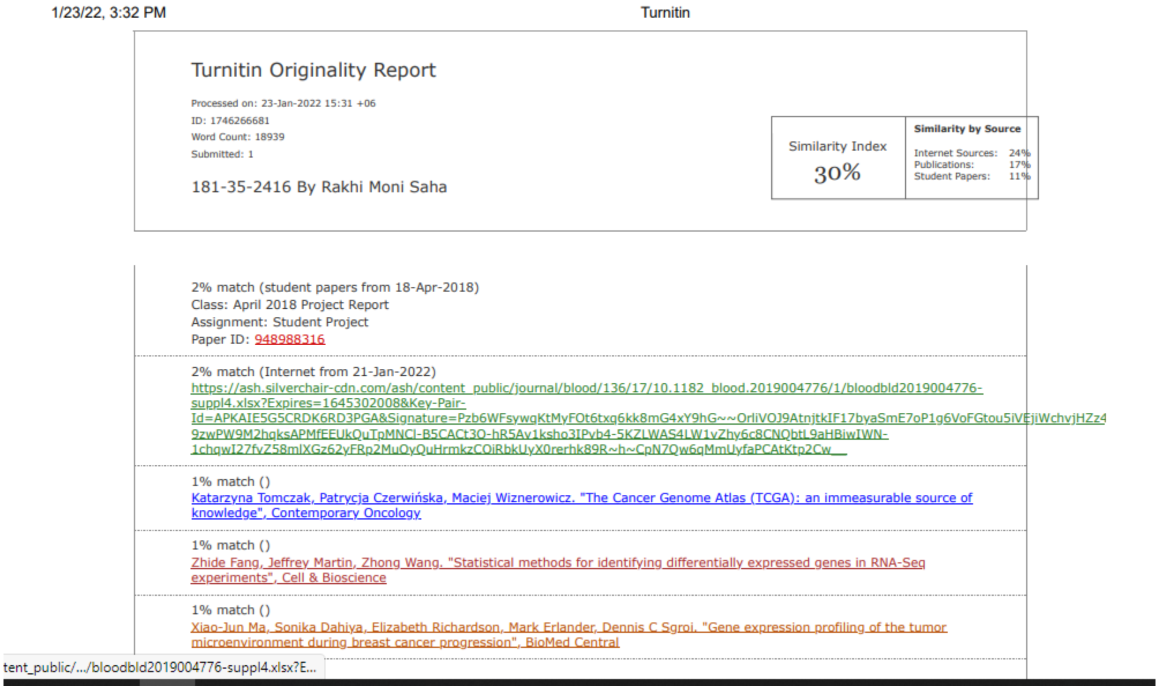
-৳0.01
Total Due

৳850.00
Total Others

Payment Scheme

Daffodil International University

Plagiarism Report



1/23/22, 3:32 PM Turnitin

Turnitin Originality Report

Processed on: 23-Jan-2022 15:31 +06
ID: 1746266681
Word Count: 18939
Submitted: 1

181-35-2416 By Rakhi Moni Saha

Similarity Index	Similarity by Source	
30%	Internet Sources:	24%
	Publications:	17%
	Student Papers:	11%

2% match (student papers from 18-Apr-2018)
Class: April 2018 Project Report
Assignment: Student Project
Paper ID: 948988316

2% match (Internet from 21-Jan-2022)
https://ash.silverchair-cdn.com/ash/content_public/journal/blood/136/17/10.1182_blood.2019004776/1/bloodbld2019004776-suppl4.xlsx?Expires=1645302008&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA&Signature=Pzb6WFsywqKtMyFOt6txq6kk8mG4xY9hG~~OrliVOJ9AtnjtkIF17byaSmE7oP1g6VoFGtou5iVEjiWchvjHZz4
9zwPW9M2hqksAPMfEEUkQuTpMNCl-B5CACt3O-hR5Av1ksho3IPvb4-5KZLWAS4LW1vZhy6c8CNQbtL9aHBiwIWN-1chqwI27fvZ58mlXGz62yFRp2MuQyQuHrmkzCOiRbkUyX0rerhk89R~h~CpN7Qw6qMmUyfaPCAtKtp2Cw__

1% match ()
Katarzyna Tomczak, Patrycja Czerwińska, Maciej Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge", Contemporary Oncology

1% match ()
Zhide Fang, Jeffrey Martin, Zhong Wang. "Statistical methods for identifying differentially expressed genes in RNA-Seq experiments", Cell & Bioscience

1% match ()
Xiao-Jun Ma, Sonika Dahiya, Elizabeth Richardson, Mark Erlander, Dennis C Sgroi. "Gene expression profiling of the tumor microenvironment during breast cancer progression", BioMed Central

tent_public/.../bloodbld2019004776-suppl4.xlsx?E...

Library Clearance