Transcriptomics Analysis of Differential Gene Expression between ER+ and HER2+ Subtypes of Breast Cancer

By

Muhammad Shakir

CIIT/SP21-BSI-037/ISB

BS Thesis

In

Bioinformatics

COMSATS University Islamabad, Pakistan

Fall 2024

COMSATS University Islamabad

Transcriptomics Analysis of Differential Gene Expression Between ER+ And HER2+ Subtypes of Breast Cancer

A Thesis Presented to

COMSATS University Islamabad

In Partial Fulfilment

Of the requirement for the Degree of

BS (Bioinformatics)

By

Muhammad Shakir

CIIT/SP21-BSI-037/ISB

Fall 2024

Transcriptomics Analysis of Differential Gene Expression Between ER+ And HER2+ Subtypes of Breast Cancer

A Graduate Thesis submitted to the Department of Bioscience as partial fulfilment of the requirement for the award of Degree of BS (Bioinformatics).

Name	Registration Number
Muhammad Shakir	CIIT/SP21-BSI-037/ISB

Supervisor

Dr. Adul Rauf Siddiqi

Associate Professor

Department of Biosciences

COMSATS University Islamabad

Final Approval

This thesis

Transcriptomics Analysis of Differential Gene Expression Between ER+ And HER2+ Subtypes of Breast Cancer

Muhammad Shakir

CIIT/SP21-BSI-037/ISB

Has been approved

For the COMSATS University Islamabad

External Examiner: _____

Dr. Faheem Tahir
Project Director NIH – CDC Project
National Institute of Health Pakistan

Supervisor: _____

Dr. Abdul Rauf Siddiqi
Department of Biosciences
COMSATS University Islamabad

Head of the Department: _____

Prof. Dr. Ijaz Ali
Department of Biosciences
COMSAT University Islamabad

DECLARATION

I, Muhammad Shakir (Registration #: CIIT/SP21-BSI-037/ISB) hereby declare that I have done the work presented in this thesis, during the scheduled period of study. I also declare that I have not taken any material from any source except referred to wherever due to that amount of plagiarism is within acceptable range. If a violation of HEC rules on research has occurred in this thesis. I shall be liable to punishable action under the plagiarism rule of HEC.

Date:_____

Muhammad Shakir
CIIT/SP21-BSI-037/ISB

Certificate

It is certified that Muhammad Shakir (CIIT/SP21-BSI-037/ISB) has carried out all the work related to this thesis under my supervision at the Department of Biosciences, COMSATS University, Islamabad.

Date:_____

Supervisor:

_____

Dr. Muhammad Rauf Siddiqi
Associate Professor
Department of Biosciences
COMSATS University
Islamabad

Head of the Department:

_____

Prof. Dr. Ijaz Ali
Department of Biosciences
COMSATS University Islamabad.

DEDICATION

Dedicated to ALLAH Almighty, Holy Prophet (P.B.U.H) and My Beloved Parents whose prayers, affection and encouragement made me able to get success.

Acknowledgement

I have no words to express my deepest sense of thanks fullness To **All Mighty Allah** who blessed upon me in the potential to complete this task.

Considerable admiration and deepest gratitude for the help and support of my all-faculty members for being so kind and patient in teaching me also, I am deeply grateful to **Dr. Abdul Rauf Siddiqi** whose help encouraging suggestions and inspiration helped me during the time of research, and in the writing of this Research Thesis.

I also express my deepest gratitude for the continuous guidance and encouragement of my humble senior **Alishbah Saddiqa.** And Individuals for all the cooperation and assistance in collecting statistics and giving their utmost support.

Finally, simple expressions of gratitude fall short of expressing our **parents'** genuine love and support. Their unwavering support in the form of their prayers, well wishes, and encouragement has been crucial in helping us conclude our work.

Muhammad Shakir

CIIT/SP21-BSI-037/ISB

ABSTRACT

Breast cancer is a fatal disease that arises from uncontrolled division of cancerous cells in breast tissue, that leading to tumors formation in the breast. Just like other cancers, breast cancer often originates in the breast area but can spread to other body parts, forming other tumors. There are many factors contribute to breast cancer, including hormonal and genetic aspects, though many genetic links remain unknown. Breast cancer is the second most common and deadly cancer in women afterwards skin cancer, and the mortality rate is increasing worldwide due to delayed or incorrect diagnoses and inappropriate treatment.

This research focuses specifically on HER2+ and ER + breast cancer Subtypes. HER2+ breast cancer is driven by an overexpression of the HER2 protein, while ER+ breast cancer is influenced by estrogen receptors that encourage cancer cell growth. Understanding these types is crucial for developing effective treatments and improving diagnosis.

In this research, I obtained dataset from an Array Express and processed them using python programing. That Dataset is pre-processed by Python libraries such as Pandas and NumPy Biopython facilitate data manipulation, while visualization tools like Matplotlib and Seaborn can help to represent the Data graphically. By preprocessing of that dataset, I identified the common genes that was directly involved in triple negative breast cancer. further network analysis carried out by Cytoscape to map significant pathways and interactions and identified 4309 DEGs (Differentially Expressed Genes). And pathway analysis was done by KEGG Software. Shedding light on the genetic factors specific to HER+2 and ER+ breast cancer types. This research aims to aid in developing reliable and swift diagnostic methods, ultimately improving survival rates for breast cancer patients.

TABLE OF CONTENTS

LIST OF FIGURES:

LIST OF TABLES:

LIST OF ABBREVIATIONS:

HER2 +	Human Epidermal Recepter2
ER+	Estrogen receptor positive
RNA Seq	RNA Sequencing

CHAPTER 1

INTRODUCTION

1.1 Cancer:

Cancer is the conditions of the uncontrol division of the cells that can affect any part of the body and is characterized by the rapid growth and spread of cancerous cells (National Cancer Institute, 2021). A tumor is a mass of tissue consisting of these abnormal cells that might invade and harm tissue in humans and are resistant to the body's natural system of regulation (World Health Organization, 2021). Starting in a single cell, cancer develops by a series of genetic abnormalities and epigenetic changes that promote cell division, prevent cell death, and allow invasive growth. Genetic mutations, immunological conditions, or internal factors including hormones, radiation, and certain chemicals might all cause these changes(Hausman, 2019).

There are two major types of tumors:

- **Benign:** Benign tumors have minimal growth as compared to the Malignant tumor and it don't spread out in body or invade the other tissues. They develop slowly, that makes surgical removal simpler and easier. They are non-invasive and are not harmful for the surrounding tissues just like nerve cells and blood vessels. (Siegel, Miller, & Jemal, 2020)**.**
- **Malignant tumor:** While Malignant tumor have ability to spread throughout the body by the help of lymphatic system and blood vessels and invade the other tissues of the body.it also have an ability to move from one part to another part of the body. Due to that characteristic, it's become challenge for the treatment. Surgical removal of this type of cancer. (Hanahan & Weinberg, 2011).

Cancer is essentially a conflict between scheduled cell death that is called apoptosis, which happens during the body's regular biological processes, and cell

growth. Mutations in the tumor suppressor and promoter genes, which can be inherited. This disorder can be characterized by a variety of genetic and epigenetic modifications, including mutations in tumor suppressor genes, which can be inherited over the course of a person's lifespan. Because cancer is caused by modified genes, this shows that it is a genetic disease (Stratton, Campbell, &Futreal,2009).

The mutations in genes that cause cancer can be carried by following mutation.

- **Spontaneous mutations:** When cancerous cells divide randomly during replication, then they cause errors in DNA replication. That error caused spontaneous mutations.
- **Environmental exposure:** Damaging or altering DNA due to exposure to specific compounds or environmental factors classified, such as carcinogens, including radiation, certain chemicals, tobacco smoke, and chronic inflammation.
- **Viruses:** A few types of viruses could cause mutation in DNA sequence, which can lead to cancer.
- **Inherited Mutations:** there are some people who receive defective genes from their descendants, which increase the risk of receiving certain types of cancer.
- **Age:** The accumulation of genetic mutation with passage of time It is also the fundamental reason why the risk of cancer grows with age.

Cancer is a challenging disease to understand and treat. The reason is that it is complex and heterogeneity, both within and between distinct tumors. A better diagnosis and quality of life for cancer patients is now possible thanks to new opportunities for targeted therapies and personalized medicine that have been made possible by improvements in our understanding of the molecular pathways that underline cancer.

1.2 Breast Cancer:

Breast cancer is complex and has significant health issues that impacts people worldwide. This uncontrolled division of cancer cells in the breast tissue leads to the development of tumors in the breast. This is causing the mark of this illness. Understanding the risk factors, early detection methods, and available treatment choices are essential for managing breast cancer and improving the results for patients (American Cancer Society, 2021).

Breast cancer is a complicated disease that is caused by several risk factors like Age, family background, specific genetic variants like BRCA1 and BRCA2, physiological factors like early menstruation and late menopause, and lifestyle factors like obesity, inactivity, and alcohol intake are some of the risk factors that are responsible for this so it's not a single factor disease. so, early diagnosis improving survival chances the for-breast cancer patient. Breast cancer is more treatable if it is examined at an early stage by help of clinical and self-examine this examination might be crucial for the cure of this complicated disease. Once the stage and features of the tumor are detected then the treatment options for breast cancer are surgery, radiation therapy, chemotherapy, hormone therapy, and targeted therapy.(Sun et al., 2017)

Breast cancer is a type of cancer that can spread to various parts of the body. When cancer cells split from the original breast tumor and go to different tissues or organs via the lymphatic system or circulation, this is known as metastasis. These cells can develop into metastatic cancers, also known as secondary tumors, in other organs such the brain, liver, lungs, or bones. Metastatic breast cancer is the term used to describe breast cancer that has spread to lymph nodes and other bodily organs. Many people believe that this stage is more advanced and more difficult to cure(Martínez et al., 2010).

1.3 Types of Breast Cancer:

Breast cancer is a complicated disease with many forms and subtypes. Breast cancer is often categorized according to specific tumor characteristics and the specific

breast cells that are damaged. Some common types of breast cancer are listed below(Fournier et al., 2005).

- **Ductal Carcinoma in Situ (DCIS):** breast cancer that stays inside the milk ducts is called DCIS. Outside the ducts, the breast tissue has not appeared.
- **Invasive Ductal Carcinoma (IDC):** is the most common type of breast cancer. that spread in the tissue around the breast after beginning in the milk ducts. It has potentially spread to other parts of the body.
- **Lobular Carcinoma in Situ (LCIS:** is the most communal form of breast cancer. It starts in the milk ducts before moving on to the breast tissue. It can spread to other places of the body.
- **Invasive Lobular Carcinoma (ILC:** begins in the milk-producing lobules and spreads to the neighboring breast tissue. Additionally, it could spread to other parts of the body.
- **Medullary Carcinoma:** Medullary carcinoma is an special type of invasive breast cancer. Under a microscope, it seems to be abundant with lymphocytes.
- **Tubular Carcinoma:** Tubular carcinoma is an uncommon type of invasive breast cancer. In this case tumor consists of small, tube-like structures and has a usually optimistic prognosis.
- **Mucinous Carcinoma (Colloid Carcinoma):** also known as colloid carcinoma, is categorized by the presence of mucin, a sticky material, inside the tumor. It usually has a better outcome than other types of breast cancer.
- **Papillary Carcinoma:** This type of cancer develops finger-like growths inside the tumor. Typically, it is not
- as aggressive of breast cancer.
- **Metaplastic Carcinoma:** This is a serious form of breast cancer and is known as metaplastic carcinoma. It is differentiated by the presence of both non-glandular and hormonal (ductal or lobular) parts.

- **Triple-Negative Breast Cancer (TNBC**): TNBC does not express the human epidermal growth factor receptor 2 (HER2), the estrogen receptor (ER), or the progesterone receptor (PR). It is typically harder to cure and more aggressive(De Laurentiis et al., 2010).(Tan et al., 2018).

There are three subtypes of molecular breast cancer that are involved in TNBC**.**

1. **Estrogen Receptor:** estrogen receptor-negative breast cancer, does not have estrogen receptors, it cannot be treated with hormone therapy. Because ER2 patients cannot benefit from treatments that target estrogen pathways, that use alternative forms of therapy. (Alves & Ditzel, 2023).
2. **HER2:** Cancer with Human Epidermal Growth Factor Receptor 2 (HER2) that is overexpression of the HER2 protein, which promotes fast and aggressive tumor development in breast. Targeted treatments that directly block HER2 signaling pathways, such as a drug known as are effective against this subtype of breast cancer.(Chen & Russo, 2009).
3. **Progesterone Receptor:** Progesterone receptor-positive (PR+) cancer is a type of hormone receptor-positive cancer that links breast cancer, where the tumor cells have receptors that bind to the hormone progesterone. This binding can raise the growth of cancer cells. PR+ cancers are usually identified by immunohistochemical analysis, which helps to classify the breast cancer subtypes and guide treatment decisions. These cancers are often less aggressive and may respond well to hormone therapies, such as selective progesterone receptor modulators or aromatase inhibitors, which aim to block the hormone's action or reduce its production. (Beltjens et al., 2021a).

1.4 Mainly affected by Breast Cancer:

Breast cancer is a fatal disease that has significant negative effects especially on women. According to the World Health Organization (WHO), breast cancer contributes to a significant number of deaths caused by cancer and is the most common disease in women globally. This condition includes consequences on the body in addition to psychological and emotional impacts on them. Although this

disease is typically found in women, it may occur to men. About 2,600 men in the US are affected by breast cancer each year, making up less than 1% percent all cases. Additionally, transgender women have a higher risk of developing breast cancer than cisgender males (Edwards et al., 2013).

1.5 Early signs of breast cancer:

Although every individual's first signs of breast cancer are, the following are some typical signs and indicators to be careful of (Park et al., 2018).

1. **Nipple discharge:** The initial sign of breast cancer is the appearance of a growth in the breast tissue. Although most of these lumps are not painful, and not all breast lumps are malignant.
2. **Changes in breast size or shape:** Inspect for any apparent variations in the breast's parameters, such as unevenness, shrinking, or swelling in breast.
3. **Skin changes:** Keep an eye out for any modifications to the nipples' appearance, such as inverting, scaling, or dimpling of the skin around them.
4. **Breast pain or tenderness:** Breast pain although cause discomfort in the breasts is not normally an early indication of breast cancer, it is nevertheless important to get it tested if you have continuous, mysterious breast pain or sensitivity as well.
5. **Swollen lymph nodes:** Swollen lymph node in breast spreads to the armpit or the area surrounding the collarbone, inflammation of the lymph nodes may result in breast cancer.

It's important to note that experiencing the one or more symptoms does not necessarily mean you have breast cancer. Many of these signs can be caused by noncancerous conditions. However, if you notice any of these changes, it is advisable to consult a healthcare professional for proper diagnosis. Regular breast self-exams, clinical breast exams, and mammograms can help with early detection of breast cancer.(Beltjens et al., 2021b).

1.6 Causes of breast cancer:

Breast cancer risk is significantly increased by gene mutations in hereditary like BRCA1 and BRCA2. A relatively insignificant percentage of breast cancer cases are carried on by these mutations(Hulka & Stark, 1995).

- **Family History:** the probability of Breast cancer can be increased by a family history, especially in first-degree relatives mean (parents, siblings, and children). many affected family members, have early diagnosis, and male had breast cancer all increase the risk.
- **Hormonal Factors:** The development of breast cancer is also prejudiced by hormones. In the case of a woman's lifetime, increased exposure to estrogen and progesterone may increase the risk of breast cancer. Early menstruation (before age 12), late menopause (beyond age 55), and never having children or having children after the age of 30 are only a few examples of factors that can raise the chance of breast cancer.
- **Age and Gender:** The risk of breast cancer rise with aging, and this is more common in women. Although breast cancer can occur in men, it is significantly less common in men as compared to women.
- **Personal History:** If an individual has earlier had breast cancer, they have more chance of developing breast cancer.
- **Dense Breast Tissue:** Women with dense breast tissue have a more chance of developing breast cancer. But Dense breast tissue can also make it more challenging to detect tumors during mammography.
- **Lifestyle Factors:** Certain lifestyle and habits also contribute to an increased risk of breast cancer.
- **Excessive Alcohol Consumption:** taking heavy alcohol regularly can increase risk of breast cancer.
- **Lack of Physical Activity:** an inactive lifestyle and not engaging in regular physical activity may also increase the risk of breast cancer.

- **Obesity:** overweight or obese, especially after menopause associated with a higher risk of breast cancer.
- **Hormone Replacement Therapy (HRT):** hormone replacement therapy, which includes both estrogen and progesterone, has been linked to an increased risk of breast cancer.
- **Environmental Factors:** Environmental factors on breast cancer risk is not fully understood, some potential factors being studied include in it.
- **Exposure to Endocrine Disrupting Chemicals (EDCs):** Prolonged exposure to plastics cosmetics, pesticides chemicals, and other products may have an impact on breast cancer risk.
- **Radiation Exposure:** radiation treatments, such as those used for previous cancer treatments, may increase the risk of developing breast cancer.
- **Shift Work:** There are few evidence strongly indicating that long-term night shift work may slightly increase the risk of breast cancer.

It is significant to remember that a person's presence of one or more of these risk factors does not guarantee that they will develop breast cancer. Even if there are no proven risk factors, the disease can nonetheless affect a lot of people. Mammograms, clinical breast exams, and self-examinations of the breast on a regular basis are crucial for early identification and better treatment outcomes.(Dogan & Turnbull, 2012).

1.7 Stages of Breast cancer:

Breast cancer is a complicated disease that develops through different stages. Knowing about these stages of breast cancer is important for accurate diagnosis and treatment procedure. This section provides an overview of the breast cancer stages, highlighting their characteristics.(Li et al., 2017).

Stage 0: Ductal Carcinoma in Situ (DCIS)

Stage 0, referred to as DCIS, is the earliest stage of breast cancer. At this point, abnormal cells have been identified in the milk duct lining, but they have not yet spread to the tissue surrounding it. The tumor is non-invasive and has not migrated to the lymph nodes or other parts of the body. Treatment options for DCIS include hormone therapy, radiation therapy, and surgery.

Stage I: Early-Stage Breast Cancer

IA and IB are two subtypes of stage I breast cancer.

a. **Stage IA:** It refers to a tumor that is up to two centimeters in size and has not spread to the lymph nodes. It is not invasive and has not proliferated to other organs. For stage IA breast cancer, hormone treatment and radiation therapy are commonly used in combination with surgery, such as lumpectomy.
b. **Stage IB:** There are two categories for breast cancer. Nevertheless, cancer cells ranging in size from 0.2 to 2 mm can still be found in lymph nodes. First, a breast tumor is not possible. However, the tumor can grow up to 2 cm and may or may not develop metastases to lymph nodes. Treatment options for stage IB breast cancer include hormone treatment, chemotherapy, and radiation therapy in addition to surgery.

Stage II: Locally Advanced Breast Cancer

Stage II breast cancer is further divided into subcategories known as IIA and IIB.

a. **Stage IIA:** There are three ways that breast cancer can appear. First, one to three axillary lymph nodes or lymph nodes close to the breastbone have cancer cells, but there is no tumor in the breast. Second, the tumor has migrated to the axillary lymph nodes and can reach a size of up to 2 centimeters. Finally, the tumor has not yet reached the axillary lymph nodes and may be more than 2 cm but not more than 5 cm. Surgery, radiation treatment, chemotherapy, and hormone therapy may all be used to treat stage IIA breast cancer.

b. **Stage IIB:** Breast cancer can be distinguished by a tumor that is larger than 2 cm but smaller than 5 cm in circumference. Furthermore, the cancer cells bigger than 0.2 millimeters but not larger than 2 millimeters are found in the lymph nodes, or the tumor has spread to one to three axillary lymph nodes or lymph nodes around the breastbone. Surgery, radiation therapy, chemotherapy, and hormone therapy are all options for treating stage IIB breast cancer, much as stage IIA.

Stage III: Locally Advanced Breast Cancer:

Stage III breast cancer is divided into subcategories known as IIIA, IIIB, and IIIC.

a. **Stage IIIA:** Breast cancer can be classified into three groups. In the first place, there may be no breast tumor, a little tumor, or a large tumor together with malignancy found in four to nine lymph nodes in the axilla or around the breastbone. Second there are little clusters of breast cancer cells in the lymph node and the tumor is larger than 5 cm. Additionally, the tumor has grown to a size of more than 5 cm, and it may have spread to one to three lymph nodes in the axilla or near the breastbone. Treatments for stage IIIA breast cancer typically involve hormone therapy, chemotherapy, radiation therapy, and surgery.
b. **Stage IIIB:** Any size of tumor that has progressed to the breast skin or chest wall is a hallmark of breast cancer. Furthermore, up to nine lymph nodes in the axilla or adjacent to the breastbone could have been affected. Treatment options for stage IIIB breast cancer include hormone therapy, chemotherapy, radiation therapy, and surgery.
c. **Stage IIIC:** A tumor of any size that has spread to the breast's skin and/or chest wall, the absence of a breast tumor, or both are characteristics of breast cancer. Furthermore, the cancer's spread has impacted at least ten axillary lymph nodes, lymph nodes above or below the collarbone, axillary lymph nodes, or lymph nodes around the breastbone. The treatments available for

stage IIIC breast cancer may involve hormone therapy, chemotherapy, radiation therapy, and surgery.

Stage IV breast cancer:

Stage 4 is the most advanced stage of breast cancer. This stage referred to the metastatic stage of breast cancer. In addition to the breast and its associated lymph nodes, cancer cells have now spread to the brain, liver, lungs, bones, and lymph nodes. When it comes to stage 4 breast cancer, the size of the tumor and the number of affected lymph nodes are not the primary reasons for concern. To reduce symptoms, improve survival, and maintain the maximum possible level of living, focus is instead placed on determining the level of metastasis and treating the disease. Treatment for stage 4 breast cancer seeks to improve overall health, reduce symptoms, and stop the spread of the cancer. Combining systemic medications including immunotherapy, hormone therapy, targeted therapy, and chemotherapy is a typical therapeutic method. These treatments target cancer cells throughout the body, prevent their spread, or even eradicate, them.

Additional supportive therapies including radiation therapy, surgery, or rest home treatment may also be used to manage symptoms or outcomes associated with stage 4 breast cancer. Hospice treatment places a high priority on managing pain, treating symptoms, and attending to patients' emotional and psychological needs.

In conclusion, breast cancer progresses through several stages, beginning with the early DCIS stage and ending with locally advanced stages such as IIIA, IIIB, and IIIC. Every stage has its own characteristics and treatments. Understanding the stages of breast cancer is crucial for a proper diagnosis, treatment decisions, and prognosis assessment.(Sharma, 2018).

1.8 Prevalence:

Breast cancer is a prominent form of cancer that affects millions of individuals worldwide. It is one of most frequently diagnosed cancers in women, and its frequency has caused alarm. The accurate number of cases of breast cancer differs

depending on the group and area. According to global cancer statistics, breast cancer is the most common kind of cancer diagnosed and the leading cause of cancerrelated deaths among women. It's important to keep in mind that although it occurs far less commonly, males can still develop breast cancer. Because breast cancer is increasingly common as people age, age is a significant risk factor for the disease delayed or no childbearing, early menstruation and delayed menopause. Obesity, inactivity, excessive alcohol use, and smoking are a few lifestyle choices that have been found to be modifiable risk factors that may contribute to the incidence of breast cancer. Ionizing radiation and specific environmental factors can also cause breast cancer. The incidence of breast cancer is influenced by a wide range of healthcare and socioeconomic variables. Because it allows for the adoption of early detection methods like mammography screening programs, early breast cancer detection is essential to improving treatment choices and consequences. Socioeconomic inequalities, limited access to medical facilities, and misunderstanding about breast cancer screening and symptoms are the main causes of higher incidence rates in particular regions.

In many countries, efforts to increase awareness of breast cancer encourage early detection and increase screening rates have resulted in higher screening rates as well as higher rates of survival. And improve current therapies. Advances in medical research and treatment methods, such as targeted therapy and personalized medicine, have also improved outcomes for breast cancer patients. Educating individuals about the value of routine screenings, early detection, and breast cancer prevention is the aim of healthcare efforts, advocacy campaigns, and educational programs. It is believed that by taking these steps, more lives will be saved, and the occurrence of breast cancer will decline. In conclusion, breast cancer is a major public health concern due to its high occurrence in both men and women. Breast cancer prevalence is influenced by several factors, including age, genetics, lifestyle choices, and access to healthcare. Early detection, improved treatment options, ongoing awareness efforts, and the encouragement of preventative actions are all

necessary to fight breast cancer and decrease its consequences on individuals and communities.(Sandhu et al., 2016).

1.9 RNA Sequencing:

RNA sequencing (RNA-Seq) is an advanced method that is widely used for analyzing an organism's transcriptome. Analyzing the whole set of RNA molecules present in a cell or tissue sample can provide information regarding alternative splicing, post-transcriptional changes, and gene expression level. RNA-Sequencing involves several important steps. In First step the RNA in the sample is separated and converted to complementary DNA (cDNA) via reverse transcription. The cDNA is then segmented and sequenced using high-throughput sequencing methods. The produced short reads are then mapped to a reference transcriptome or genome to determine their origin. A range of computer tools and procedures are then used to analyze the data and quantify the levels of gene expression (Wang et al., 2009; Oshlack et al., 2010). RNA-Seq has several advantages over previous techniques, such as its ability to discover alternative splicing processes, detect new transcripts, and more precisely measure gene expression level (Wang et al., 2009). It also makes it possible for researchers to examine non-coding RNAs and examine RNA modifications such as alternative splicing and RNA editing (Wang et al., 2009). RNA-Seq experiment data may be further examined to identify functional gene networks, differentially expressed genes, and potential biomarkers (Wang et al., 2009; Oshlack et al., 2010). Several bioinformatics tools and pipelines, including quantification tools like Cufflinks, HTSeq, and Salmon and alignment tools like TopHat, STAR, and HISAT2, have been developed to process and interpret RNASeq data. The database used for the retrieval of the three different datasets is Array Express (Trapnell et al., 2013; Patro et al., 2017).

- **Array Express**

  Array Express is a database that is used publicly and contains high quality gene expression data. It gives genomics researchers a useful tool by enabling them to

save, exchange, and examine gene expression data produced by different experimental methods. Array Express provides a variety of data, such as highthroughput sequencing and microarray studies, along with other kinds of functional genomics data. An essential component of the larger European Molecular Biology Laboratory (EMBL) resources, the database is managed by the European Bioinformatics Institute (EMBL-EBI). It is a useful tool for the scientific community since it supports data sharing and maintains the open access requirements. By providing thorough information for every experiment and enforcing submission guidelines, Array Express guarantees the quality of the data. To locate pertinent gene expression data for their investigations, researchers can browse and make use of Array Express. Users may filter and retrieve datasets according to certain criteria thanks to the database's powerful data retrieval features. Additionally, Array Express provides a thorough compilation of gene expression data by integrating with other bioinformatics sites, like the Gene Expression Omnibus (GEO). Array Express has been widely used in several studies and has contributed to the advancement of genomic research. It’s a open source and standardized data format that makes it an important resource for scientists to share and explore gene expression data, enabling discoveries and insights into the complicated workings of genes and their regulation.

1.10 Pathway Analysis:

Pathway analysis is a important that helps in identifying biological pathways significantly enriched in a given set of genes. The Database for Annotation, Visualization, and Integrated Discovery (DAVID) is a widely used tool for this

purpose. DAVID integrates functional annotation tools to analyze large gene lists, providing insights into biological themes such as molecular pathways, cellular functions, and disease associations. By mapping genes to pathways like those in KEGG. DAVID enables researchers to interpret complex datasets and uncover mechanisms underlying diseases. This approach is particularly useful in identifying key pathways involved in conditions like cancer, aiding hypothesis generation and the development of targeted therapeutic strategies.

1.11 Network Analysis

Network analysis is performed by using STRING Database.

STRING Database:

A string database is a specific type of database made primarily to store and work with character strings. It is widely used throughout several fields, including as information retrieval, natural language processing, and bioinformatics. Strings are regarded as basic entities in a string database, and they may be subjected to several operations, including comparison, pattern matching, indexing, and search. To maximize string processing processes, string databases use certain data structures and algorithms. These include inverted indexes, suffix trees, and tri structures, which allow for quick retrieval and search based on precise matches or substring patterns. Furthermore, methods like indexing and compression may be used to improve query performance and reduce storage needs. The GenBank database, a comprehensive collection of DNA and protein sequences, is a well-known example of a string database. Innovations in molecular biology and genomics are made possible by its ability to store, search, and retrieve genetic information. With its extensive collection of scientific papers, the PubMed database is another important database. It helps effective information retrieval in the field of biomedical research by enabling researchers to search for papers using keywords, author names, and other criteria. To sum up, string databases are crucial resources for effectively

processing and storing character strings. They are essential in many fields because they make it possible for practitioners and academics to effectively organize, search, and retrieve information that is written. The creation of specialized algorithms and data structures has made a major contribution to this.

1.12: RNA Seq Analysis with Python:

RNA-seq analysis is an important step to understanding gene expression and identifying differentially expressed genes (DEGs). Using Python to find DEGs and analyzed them successfully through statistical approaches like DESeq2, which provides comprehensive outputs including log2 fold changes, p-values, and adjusted p-values for each gene. These results offer comprehensions into genes that are upregulated or downregulated under specific conditions, helping to classify potential biomarkers or pathways associated with diseases. Python libraries such as Pandas and NumPy facilitate data manipulation, while visualization tools like Matplotlib and Seaborn can help present the findings graphically. This statistical approach to DEG analysis serves as an initial step for further biological analysis or downstream analysis, such as pathway enrichment or gene ontology studies.

1.13 Problem Statement:

The number of people dying from breast cancer is rising daily. Its significant incidence in underdeveloped countries demands rapid diagnosis and treatment. However, in most cases, we are still reluctant to detect it immediately, the rate of false positive or negative findings are produced, leading to a delayed diagnosis and, ultimately, a delayed course of therapy. Therefore, it is imperative to create a highquality breast cancer diagnostic approach to reduce the number of false positive and delayed diagnoses, which will ultimately result in early diagnosis and progressive treatment. And aid in preventing or lowering the death rate from breast cancer

1.14 Objective:

The primary objective of this research is to identify genes that are directly involved in ER+ and HER2+ breast cancer subtypes and to find out which of these genes are upregulated and downregulated. Furthermore, the study purposes to explore the pathways linked with these genes using tools like string for interaction analysis and g: Profiler for functional enrichment. By understanding the connections and biological significance of these genes, this research seeks to provide insights into the molecular mechanisms underlying ER+ and HER2+ breast cancer, contributing to the development of targeted therapies and precision medicine approaches.

CHAPTER 2

MATERIALS AND METHODS

In this we will discuss the dataset that is used for this project and methodology performed on those datasets to find the DEG's and eventually the effective method for breast cancer diagnosis.

2.1 Dataset Description:

In this project RNA Seq DATASETS are used taken from Array Express. Array Express is a public repository for storing and sharing gene expression data. It is a database that allows researchers to submit, search, and download various types of high-throughput functional genomics data.

S.NO	Dataset Description	Accession ID
1	An Integrated Model of the Transcriptome Landscape of HER2-Positive Breast Cancer.	E-GEOD-45419

Table 1: Dataset Descriptions

2.2 Obtaining Datasets:

There is one dataset that is taken from Array Express. That dataset is about a whole transcriptome profiling of triple negative breast cancer tumors in which one factor is ER+ which has 8 samples, and the second factor is estrogen positive (HER2+) which has 8 samples so in total there are 16 samples in this dataset. The first factor is about the Estrogen receptor role in tumor suppression of triple negative breast cancer and the second factor is about the Estrogen receptor (HER2) role in tumor suppression of triple negative breast cancer. Both factors are involved in TNBC so in total there are 16 samples of ER+ and HER2+ of triple negative breast cancer.

2.3 Data pre-processing

Data preprocessing was done using R at the DESeq2 step, and further analysis was performed using Python. The workflow was adjusted accordingly, R languages provide a user-interface, libraries and packages for bioinformatics analysis, enabling researchers to perform complex data analysis tasks and accelerate their research in various biological domains. So, I used these languages on all the three datasets that are discussed briefly below:

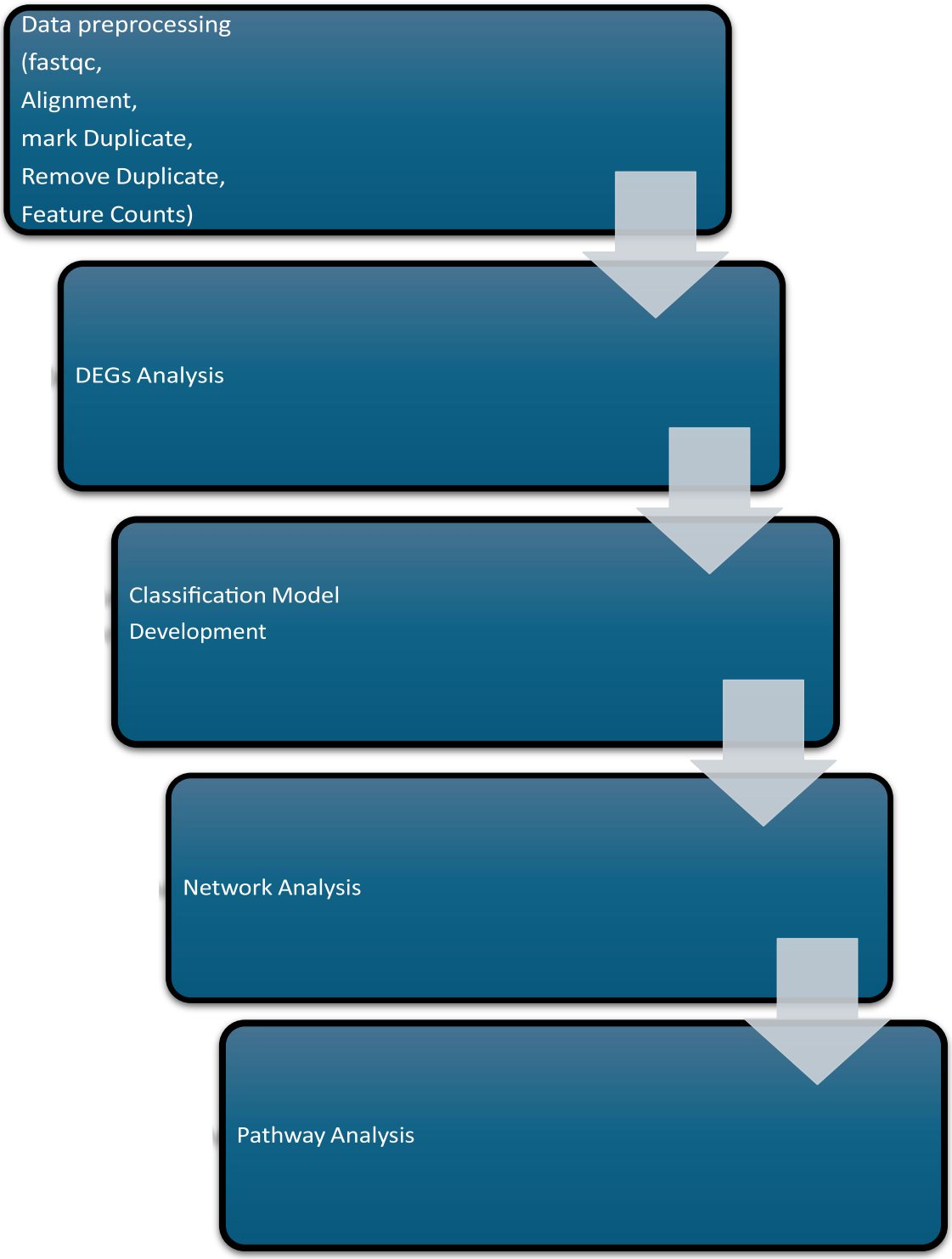


Figure1: Workflow

2.3.1 FastQc:

FastQC is a basic step that uses bioinformatics for quality control of highthroughput sequencing data. It gives us a comprehensive assessment of the quality and feature of sequencing reads in a fast and efficient manner. The primary purpose is to identify potential problem and biases in sequencing data that could affect further analysis.so fastqc give us a quality control report that allows users to evaluate various metrics and visualizations related to the sequencing data.

2.3.2 Alignment:

HiSAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts 2) is a widely used bioinformatics tool for aligning high-throughput RNA Sequence data to a reference genome. It is mainly designed for mapping RNA-Sequence reads that may contain spliced alignments due to the presence of introns in eukaryotic genomes.

2.3.3 Mark duplicates:

MarkDuplicates is a step commonly used for identifying and marking duplicate reads in high-throughput sequencing data. The MarkDuplicates tool identifies duplicate reads based on their alignment position or sequencing characteristics. E.g.: identical start positions, orientation, and fragment lengths. It is particularly useful for removing PCR or optical duplicates that can occur during library preparation and sequencing. The RemoveDuplicate is another step in RNA data preprocessing a command-line utility used to remove duplicate reads from sequencing data files. It is commonly used in bioinformatics workflows to process high-throughput sequencing data, such as data generated from next-generation sequencing platforms in Galaxy tool. The RemoveDuplicate tool typically takes a sorted input file containing aligned sequencing reads and identifies and removes duplicate reads based on their alignment coordinates or other characteristics this is important step in data pre-processing because duplicate gene give error for further analysis results.

2.3.4 Feature counts:

The Feature Counts are commonly used for quantifying the number of sequencings reads that align to genomic features, such as genes or exons. It is typically used in the analysis of RNA-Sequence data to determine gene expression levels. Feature Counts takes as input aligned sequencing reads in BAM format and a set of genomic features such as a GTF or GFF file that defines the regions of interest, such as genes or exons. The tool assigns each aligned read to the corresponding genomic feature based on its alignment position. It then counts the number of reads assigned to each feature, providing a measure of the expression level of that feature.

2.3.5 DeSeq2:

When we do RNA-Sequence analysis by using python and RYP module that work just like R, then it generates DESeq2 file direct in bioinformatics research for differential gene expression analysis give the Result of this Deseq2 file that is further used to analyze RNA-Seq data and identify genes that are differentially expressed between different experimental conditions or groups.DESeq2 typically requires sample as a input files in the form of count matrices, where each row represents a gene, and each column represents a sample. These count matrices can be generated from RNA-Seq data using tools like Feature Counts, which quantify the number of reads aligned to each gene. The output of the DESeq2 tool in R includes statistical measures such as fold change, p-values, and adjusted p-values. Then it goes for further analysis in for getting the visualization of the results, such as generating heatmaps, volcano plots, and MA plots, to help interpret the differential expression analysis. By getting the DESeq2 result through python, researchers can perform comprehensive and statistically rigorous differential expression analysis on their RNA-Seq data, enabling them to gain insights into gene regulation and biological processes associated with different experimental conditions or groups.

2.4 Enrichment Analysis:

Enrichment analysis was done by using DAVID and visualization of GO terms including biological processes, molecular functions and cell components was done by g: Profiler.

2.5 DAVID:

This site facilitates Annotation, Visualization, and Integrated Discovery Database supplies to researchers with a complete suite of functional annotation tools to help them comprehend the biological significance of lengthy gene lists. These technologies are supported by the extensive DAVID which combines functional annotations from many sources and is based on the DAVID Gene concept. DAVID tools can do the following for any given gene list:

- Find enriched functionally related gene groups of Gene List
- Cluster redundant annotation terms of Gene List
- KEGG pathway maps of Gene List
- Molecular function Gene List

2.5 Cytoscape:

Cytoscape is a powerful open-source platform used for visualizing and analyzing protein-protein interactions and other types of biological networks. It enables the integration of experimental data, annotations, and molecular interaction networks, offering a wide range of features for network analysis. The tool supports the visualization of direct (physical) and indirect (functional) interactions, as well as the ability to integrate data from multiple sources, including computational predictions and curated databases.

CHAPTER 3

RESULTS

3.1 Distribution of Log2 Fold:

The histogram shows the spreading of Log2 Fold Change values attained from the RNA-Seq analysis of ER+ and HER2+ cancer data. The x-axis denotes the Log2 Fold Change, while the y-axis indicates the frequency of genes corresponding to these values.

- The plot displays a symmetrical bell-shaped distribution centered around 0, revealing that most genes exhibit little to no significant fold change.
- A large fraction of the data points lie between -1 and +1 Log2 Fold Change, indicating minimal expression changes for the majority of genes.
- The distributions of extend tails of both ends, showing the fewer genes with higher upregulation mean (positive values) or downregulation mean (negative values).
- This distribution implies that while some genes are differentially expressed, the majority show no significant change in expression levels between the compared conditions.

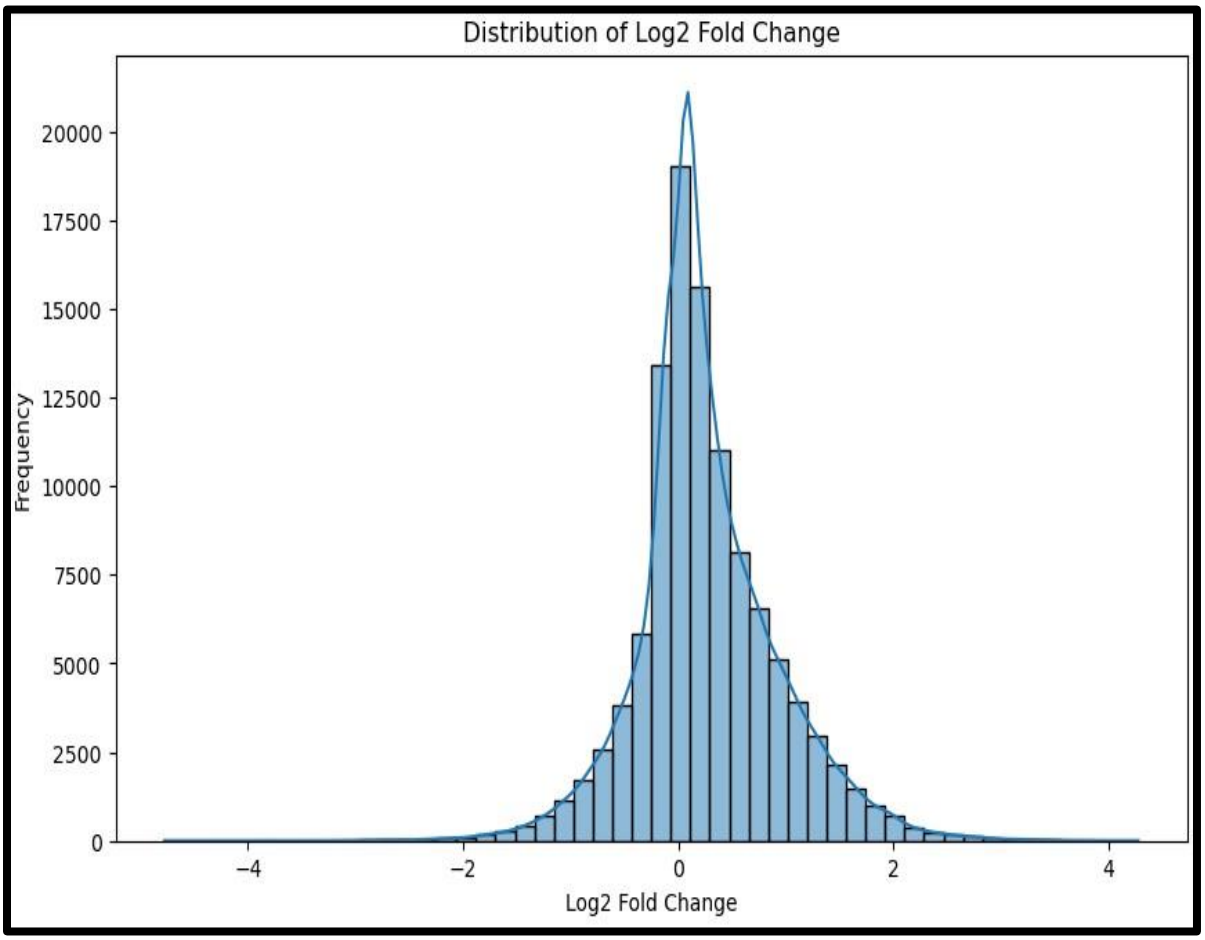
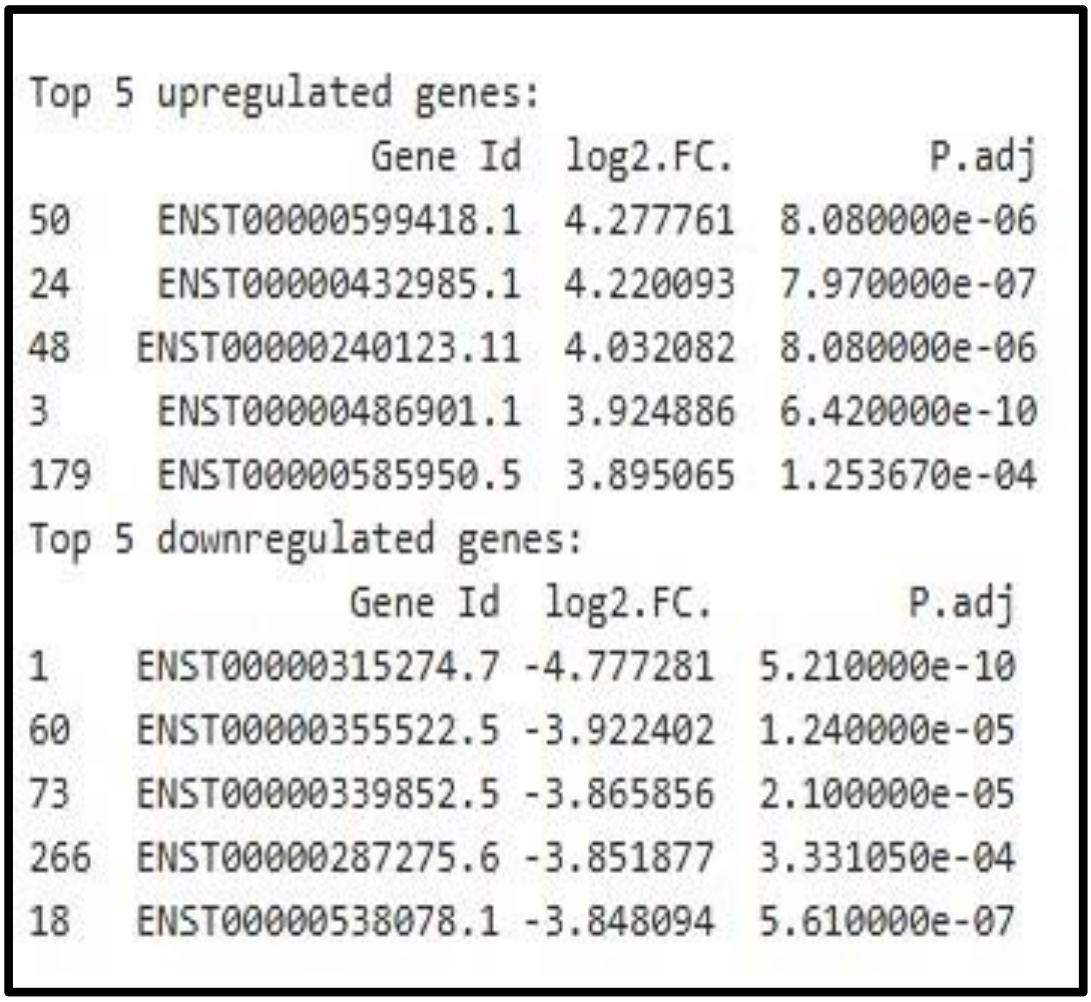


Figure 2: Log2 fold Change with Genes Frequency

3.2 Downregulated RNA-Seq Analysis:

This table highlights the top 5 upregulated and 5 downregulated genes in the RNASeq analysis of ER+ and HER2+ cancer data, based on Log2 Fold Change (log2.FC) and adjusted p-values (P.adj) according to the Data. The upregulated genes indicate positive Log2 Fold Change values ranging from 3.89 to 4.28, indicating that expression levels increasing with gene IDs such as ENST00000599418.1 and ENST00000432985.1 that proving the highest upregulation and adjusted p-values between 10^{-6} and 10^{-7}. Similarly for the downregulated genes exhibit negative Log2 Fold Change values from -3.84 to -4.77, indicating decreased expression levels, with ENST00000315274.7 and ENST00000355522.5 showing the most significant downregulation and adjusted p-values as low as 10^{-10}.



Top 5 upregulated genes:

	Gene Id	log2.FC.	P.adj
50	ENST00000599418.1	4.277761	8.080000e-06
24	ENST00000432985.1	4.220093	7.970000e-07
48	ENST00000240123.11	4.032082	8.080000e-06
3	ENST00000486901.1	3.924886	6.420000e-10
179	ENST00000585950.5	3.895065	1.253670e-04

Top 5 downregulated genes:

	Gene Id	log2.FC.	P.adj
1	ENST00000315274.7	-4.777281	5.210000e-10
60	ENST00000355522.5	-3.922402	1.240000e-05
73	ENST00000339852.5	-3.865856	2.100000e-05
266	ENST00000287275.6	-3.851877	3.331050e-04
18	ENST00000538078.1	-3.848094	5.610000e-07

Figure 3: Up and Down Regulated Gene

3.3 Volcano Plot:

The Volcano plot demonstrates differential gene expressions between ER+ and HER2+ cancer samples based on RNA-seq data, here each dot represents a gene. The x-axis shows the log2 fold change, and the y-axis represents the -log10 p-value. Those Genes that are above the horizontal line (p-value cutoff) and outside the vertical lines (fold change cutoff) are statistically significant, indicating substantial expression differences between the two cancer types. Positive values correspond to upregulated genes and negative values represent downregulated genes. Genes that are in the central region are not significantly expressed genes.

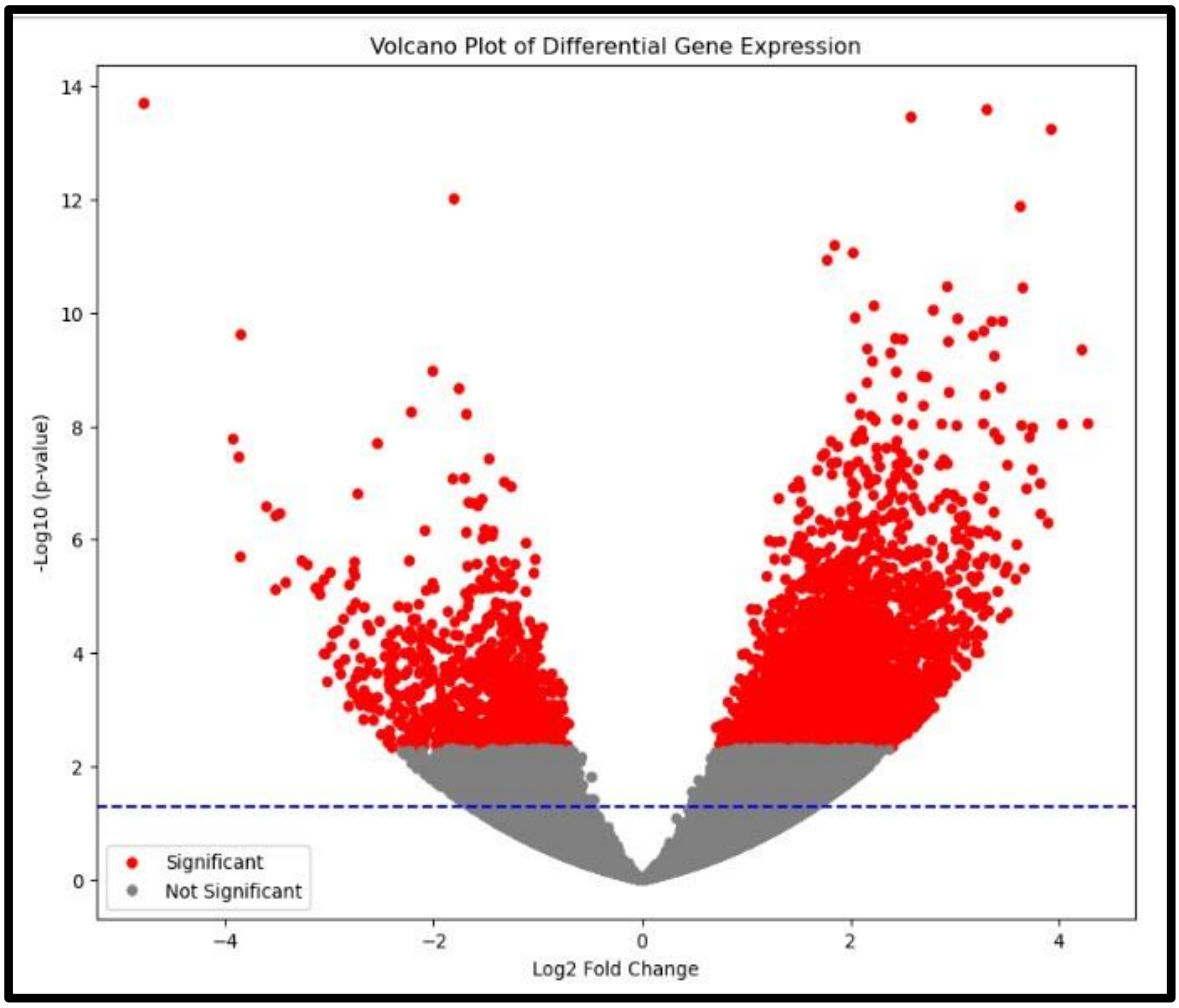


Figure 4: Volcano Plot

3.4 MA Plot:

MA plot visualizes differential gene expressions between ER+ and HER2+ cancer samples from RNA-seq data, with each dot representing a gene. The x-axis shows the average expression level, while the y-axis represents the log2 fold change. Genes that are above and below the horizontal blue line are statistically significant, with upregulated genes appearing above the line and while downregulated genes below line. Genes near the blue line are not significantly.

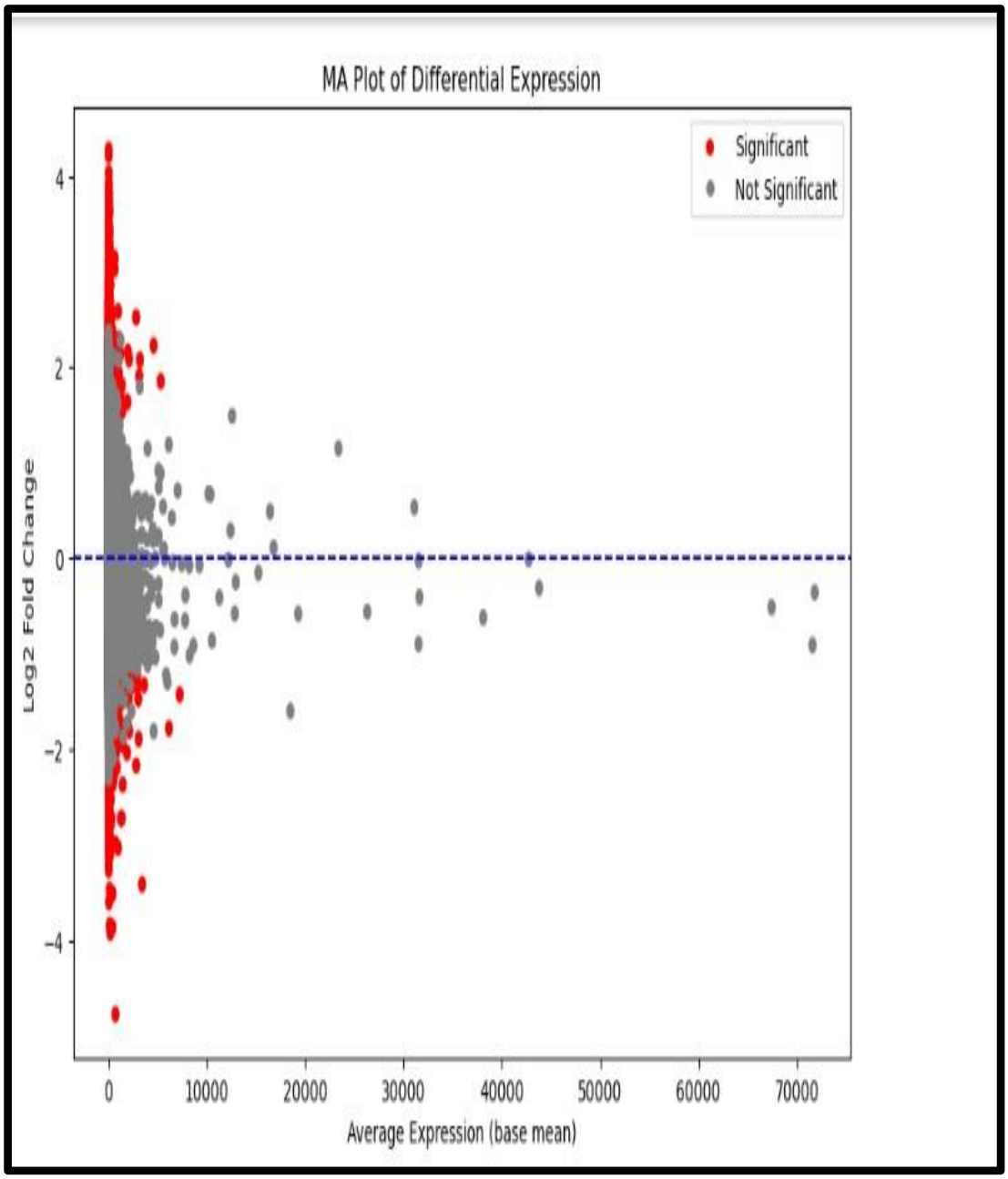


Figure 5: MA Plot

3.5 Correlation Heatmap of Gene Expression Metrics:

Heatmap visualizes correlations between various gene expression metrics from RNA-seq data comparing ER+ and HER2+ breast cancers. A strong positive correlation is observed between log2 fold change and Wald statistics as the Wald statistics determines the significance of fold changes. Here Base mean expressions indicate that a moderate positive correlation with log2.FC and Wald.Stats, indicating that genes with higher average expression levels tend to exhibit larger fold changes and greater significance. Conversely, p-value and adjusted p-value shows a negative correlation with log2.FC and Wald.Stats.

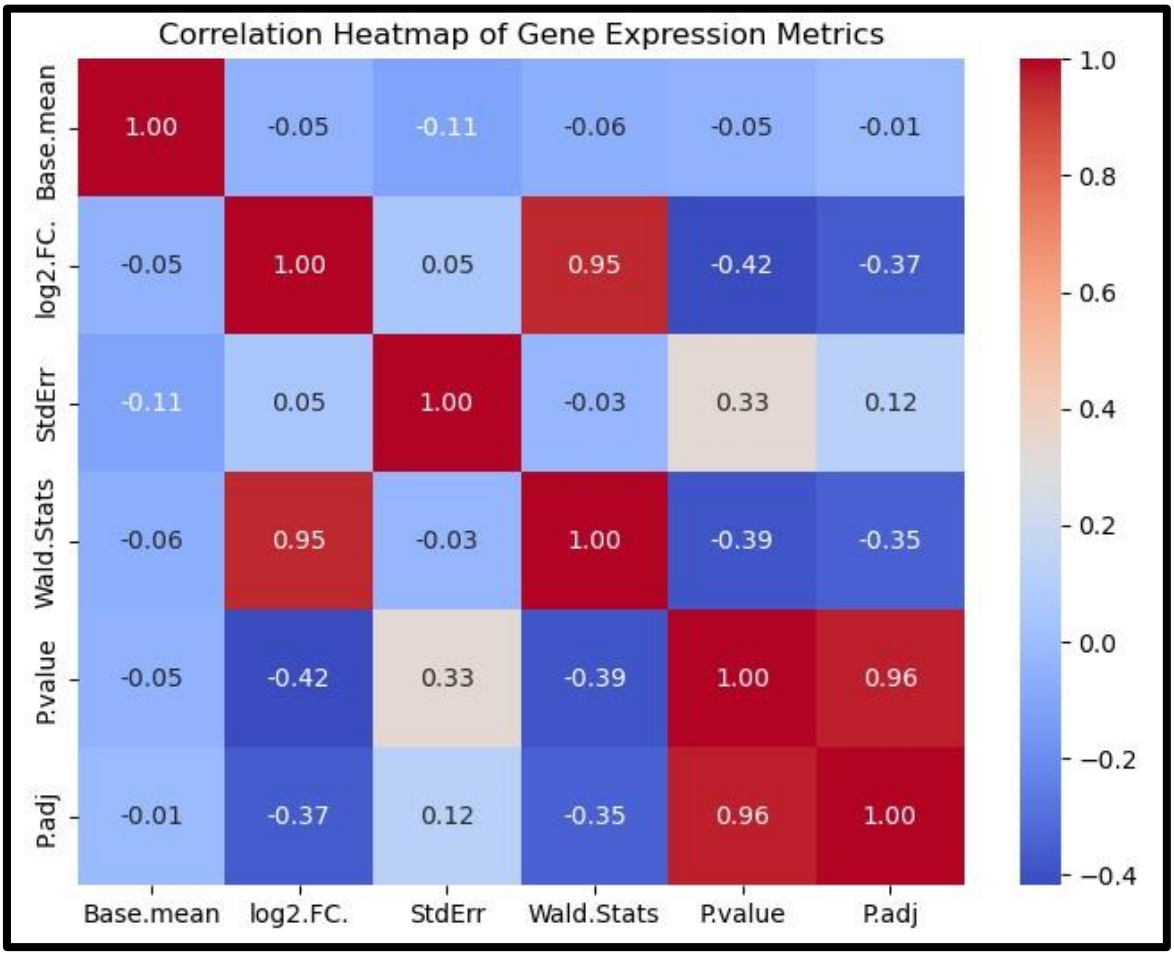


Figure 6: Heatmap

3.6 Distribution of Log2 Fold Change by Gene Significance:

Boxplots describe the distribution of log2 fold change values for genes classified as "Significant" and "Not Significant" based on their differential expression between ER+ and HER2+ breast cancer samples. Significant genes show a wider distribution of log2 fold change values, with a median around 1.5, that show substantial differences in expression between the cancer subtypes. On the other hand, nonsignificant genes have a narrower distribution centered around 0, that reflect the minimal changes in expression levels.

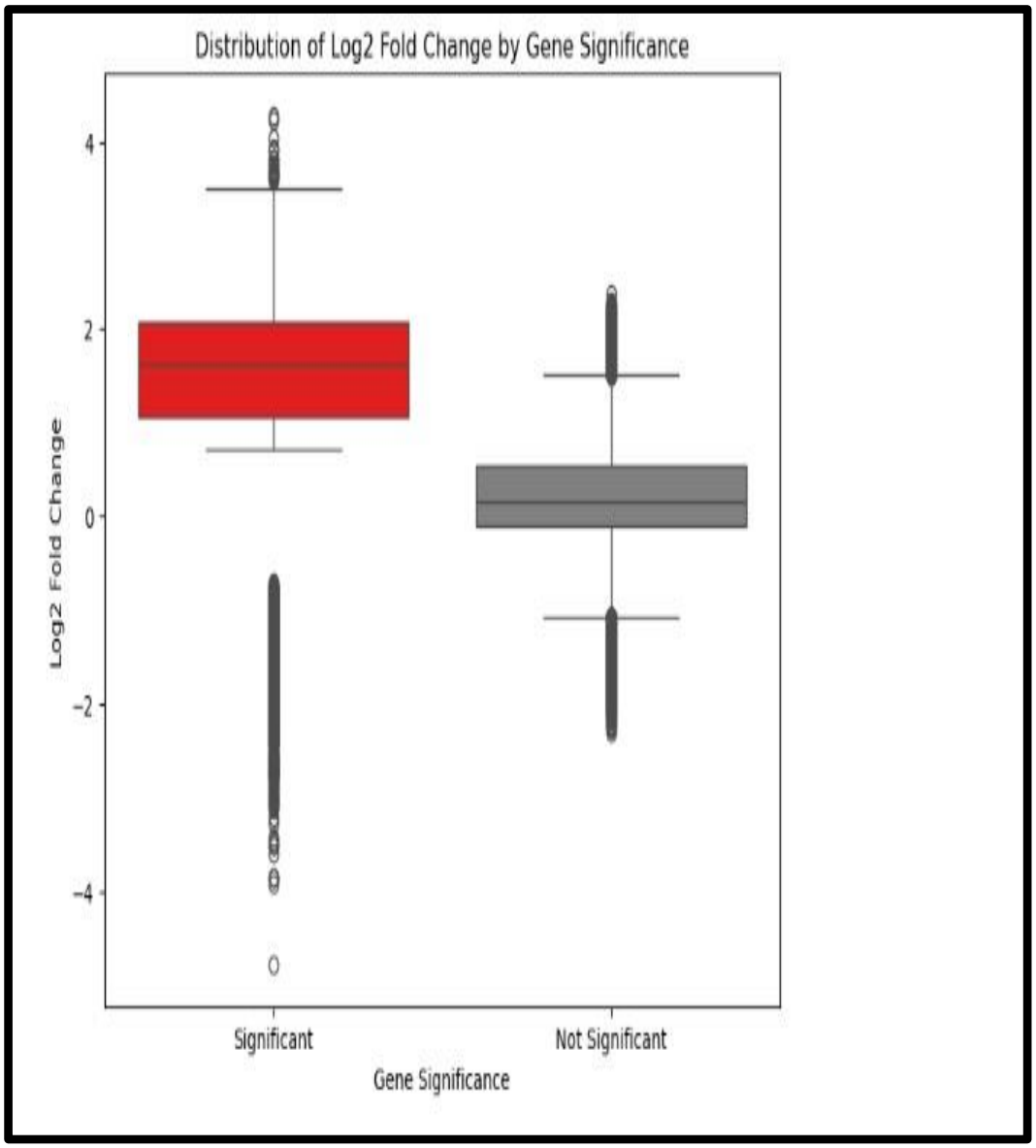


Figure 7: Distribution of Significance Gene

3.7 Hierarchical Clustering Dendrogram of Top 1000 Genes:

Dendrogram describes the hierarchical clustering of the top 1000 genes based on their expression patterns across ER+ and HER2+ breast cancer samples. Genes are classified into clusters based on the similarity of their expression levels, closely related genes forming tighter clusters and diverse genes separated by longer branches. The lengths of these branches represent the distance between clusters, while longer branches showing the greater dissimilarity.

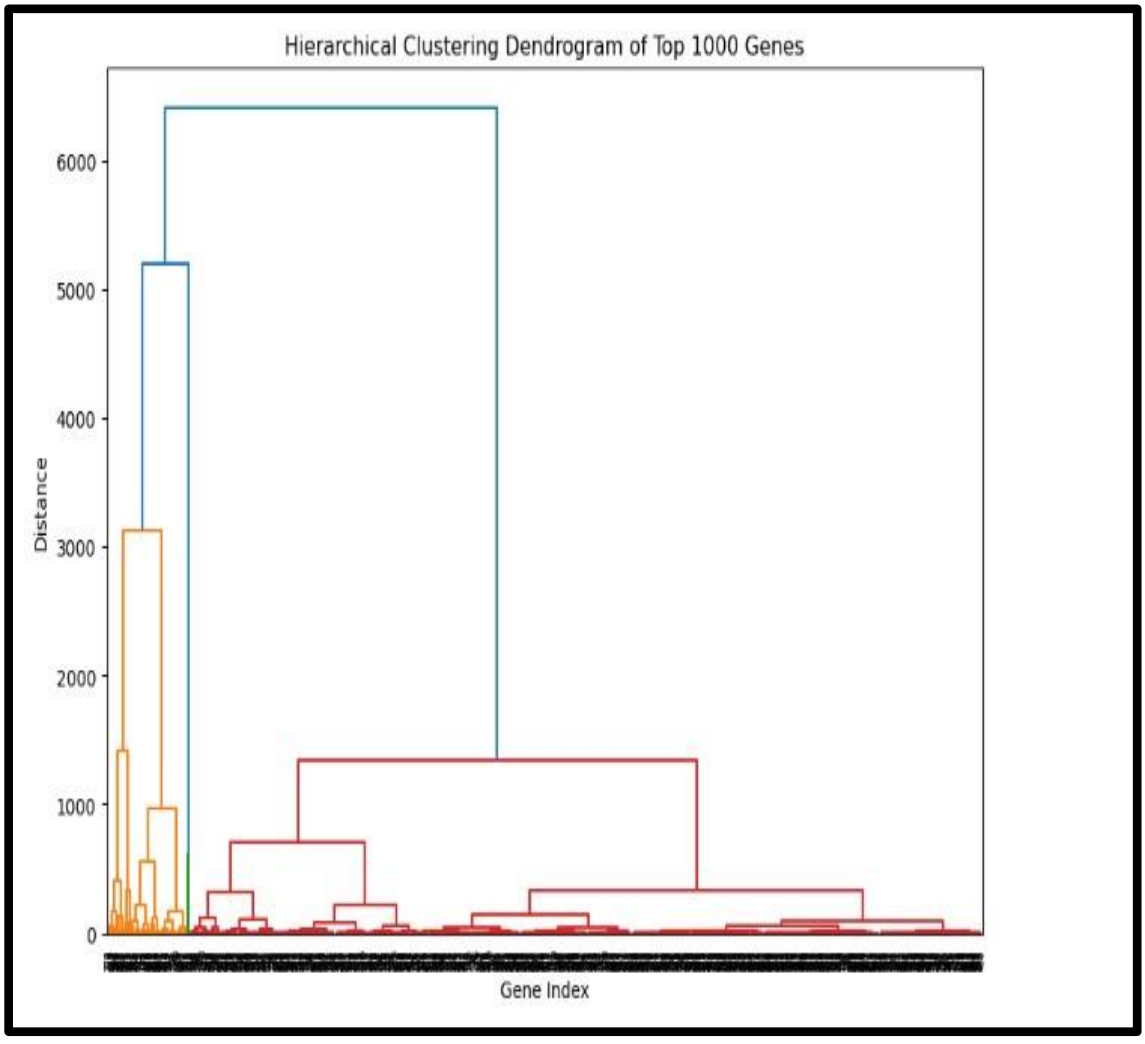


Figure 8: Hierarchical clustering of genes

3.8 Go profiler

G. Profiler is a important tool that is widely used in bioinformatics. That tool used for functional enrichment analysis and gene annotation. It helps us to identify the biological significance of a given genes list, proteins, or other genomic features that may be mapping them to known biological pathways. It also gives us gene ontology (GO) terms with Gene Id and its description. It synchronizes that data with multiple databases, such as KEGG, Reactome, and TRANSFAC, to provide insights into the functional roles and relationships of genes in a biological context.

In a graph, the colors of the nodes typically represent the significance level of the enriched terms or pathways. For example:

- **Red or darker colors**: Indicate higher significance means lower p-value.
- **Lighter colors or yellow**: Indicate lower significance means higher p-value.
- **Lower p**-value indicates higher statistical significance and Higher P-value show the lower statistical significance.

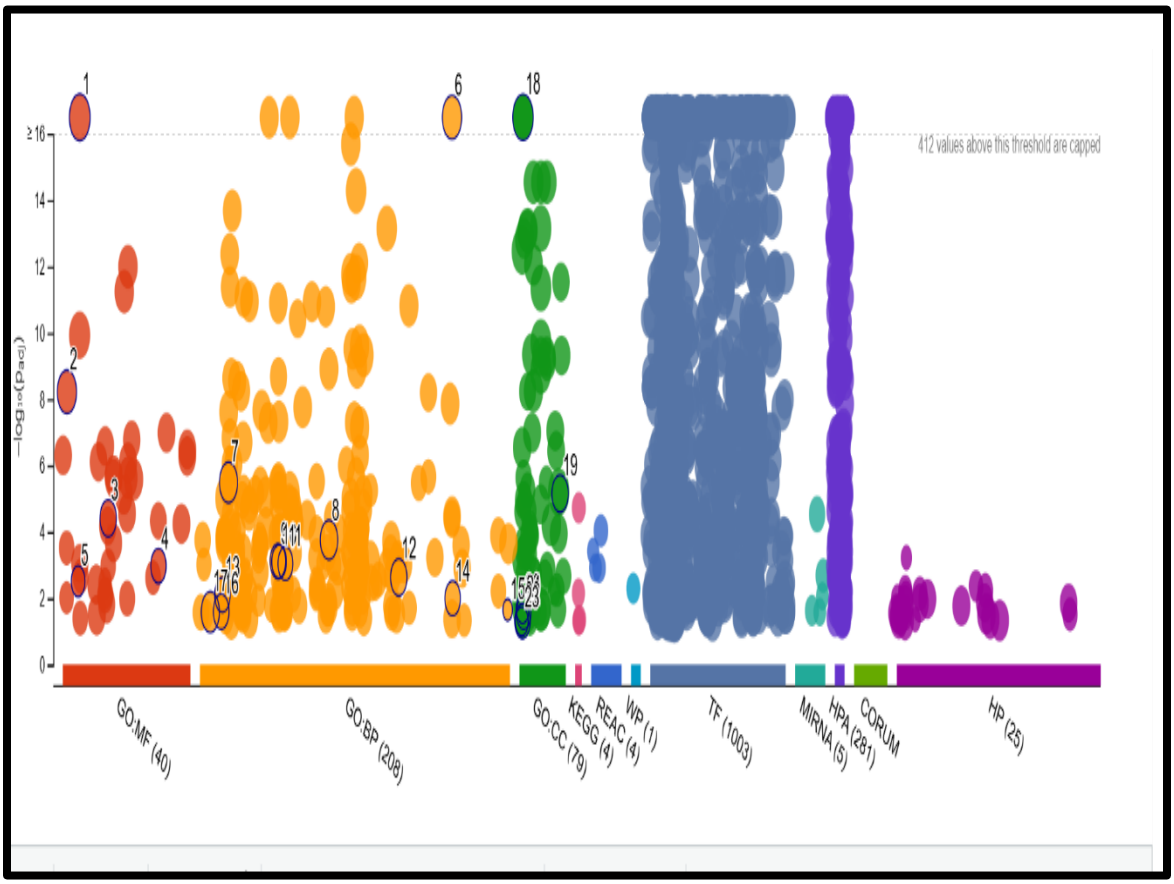


Figure 9: Graphical Representation of g: Profiler

Gene ID	Gene names	Description
ENST00000525250	PNPLA2	patatin like phospholipase domain containing 2
ENST00000315274	MMP1	matrix metallopeptidase 1
ENST00000546227	RHOF	ras homolog family member F, filopodia associated
ENST00000486901	HSPG2	heparan sulfate proteoglycan 2
ENST00000547726	HNRNPA3P10	heterogeneous nuclear ribonucleoprotein A3 pseudogene 10
ENST00000503548	UVSSA	UV stimulated scaffold protein A
ENST00000647157	SERPINB6	serpin family B member 6
ENST00000450036	NME4	NME/NM23 nucleoside diphosphate kinase 4
ENST00000586944	PTBP1	polypyrimidine tract binding protein 1

Table2: Gene Description against Ensemble Id's

Category	Term	Count
Molecular Functions	GO:0005515~protein binding	8228
Molecular Functions	GO:0003723~RNA binding	1034
Molecular Functions	GO:0046872~metal ion binding	1844
Molecular Functions	GO:0000978~RNA polymerase II cis-regulatory region sequence-specific DNA binding	882
Molecular Functions	GO:0003677~DNA binding	788
Molecular Functions	GO:0000981~DNA-binding transcription factor activity, RNA polymerase II-specific	904
Molecular Functions	GO:0004842~ubiquitin-protein transferase activity	180
Molecular Functions	GO:0031267~small GTPase binding	221
Molecular Functions	GO:0001227~DNA-binding transcription repressor activity, RNA polymerase II-specific	265

Table 3: Molecular Function

Category	Term	Count
Biological Process	GO:0006357~regulation of transcription by RNA polymerase II	1115
Biological Process	GO:0045944~positive regulation of transcription by RNA polymerase II	851
Biological Process	GO:0045893~positive regulation of DNAtemplated transcription	502
Biological Process	GO:0006355~regulation of DNA-templated transcription	583
Biological Process	GO:0000122~negative regulation of transcription by RNA polymerase II	647
Biological Process	GO:0015031~protein transport	304
Biological Process	GO:0001525~angiogenesis	195
Biological Process	GO:0016477~cell migration	209
Biological Process	GO:0006974~DNA damage response	223

Table 4: Biological process

Category	Term	Count
Cellular Components	GO:0005654~nucleoplasm	2852
Cellular Components	GO:0005829~cytosol	3751
Cellular Components	GO:0005634~nucleus	3990
Cellular Components	GO:0005737~cytoplasm	3609
Cellular Components	GO:0005794~Golgi apparatus	**757**
Cellular Components	GO:0016020~membrane	3068
Cellular Components	GO:0070062~extracellular exosome	1453
Cellular Components	GO:0005739~mitochondrion	985
Cellular Components	GO:0005925~focal adhesion	327

Table 5: Cellular Components

Category	Term	Count
KEGG_PATHWAY	hsa05168: Herpes simplex virus 1 infection	379
KEGG_PATHWAY	hsa04510:Focal adhesion	154
KEGG_PATHWAY	hsa04144: Endocytosis	185
KEGG_PATHWAY	hsa05205: Proteoglycans in cancer	151
KEGG_PATHWAY	hsa04071: Sphingolipid signaling pathway	95
KEGG_PATHWAY	hsa04350: TGF-beta signaling pathway	84
KEGG_PATHWAY	hsa04070: Phosphatidylinositol signaling system	77
KEGG_PATHWAY	hsa04360: Axon guidance	134
KEGG_PATHWAY	hsa05131: Shigellosis	176s

Table 6: KEGG Pathway

3.15 Network Analysis

Network analysis is performed using Cytoscape. The interactions for all the differentially expressed genes are shown in the figure. We have *4303* similar genes across three datasets that show various types of interactions, including strong, weak, and neutral interactions. Different lines represent different types of interactions among genes: a dark line indicates a strong interaction, a light line indicates a weak interaction, and a line with neither dark nor light shading represents a neutral interaction among genes.

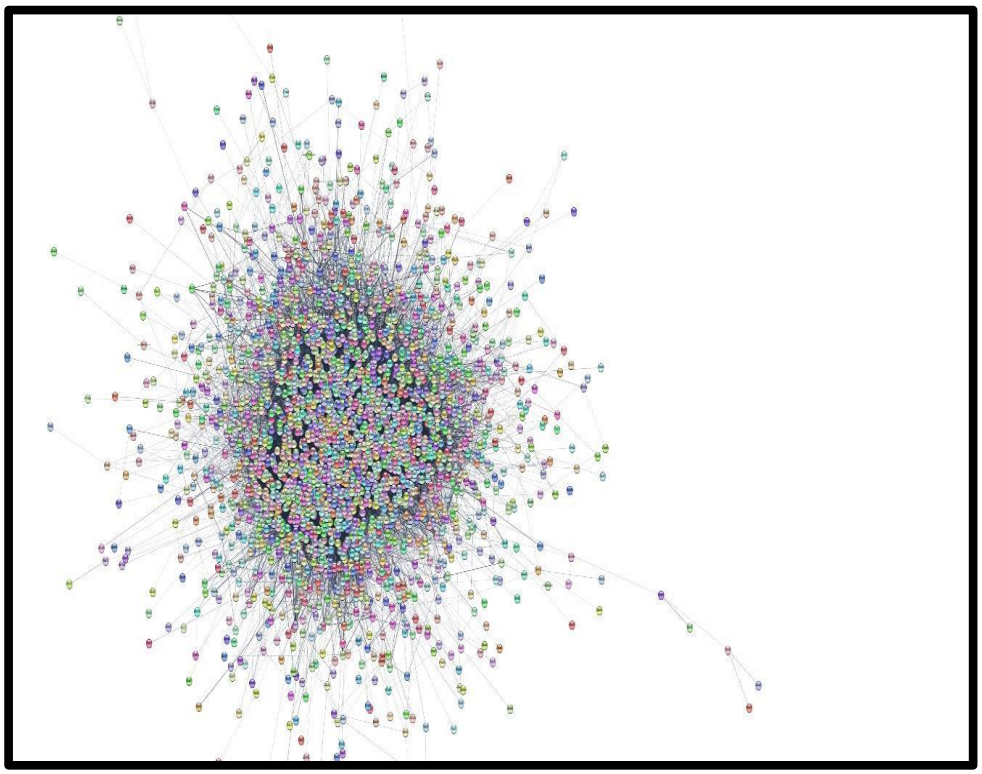


Figure 10: Network Analysis for all genes

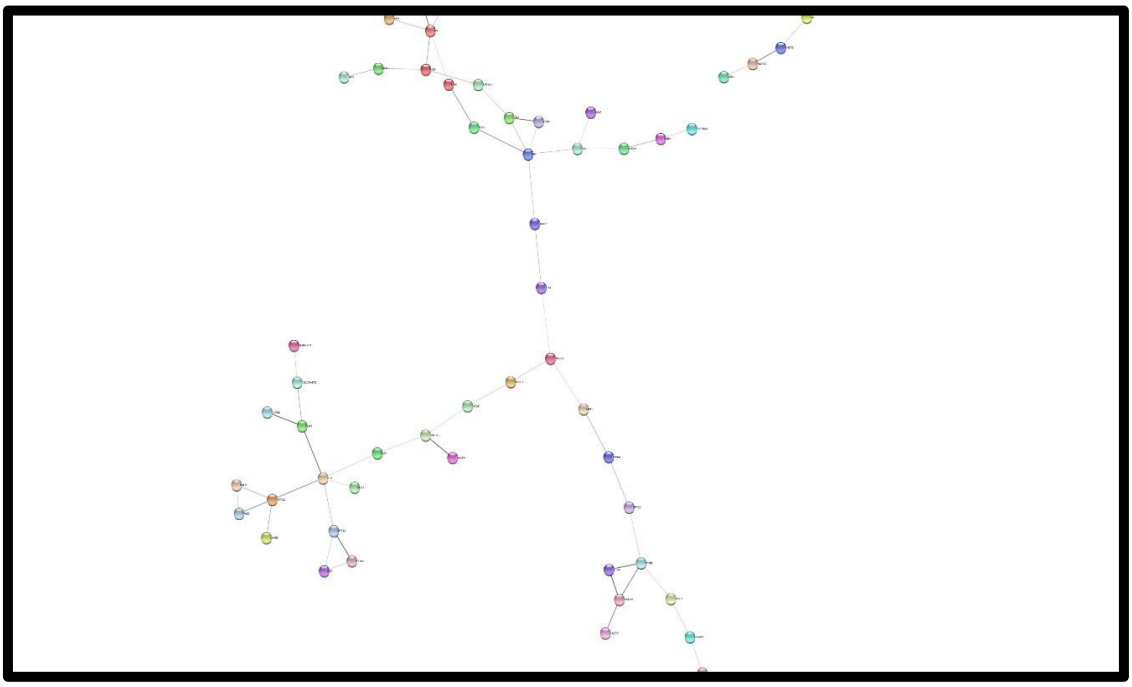


Figure 11: Top 150 genes

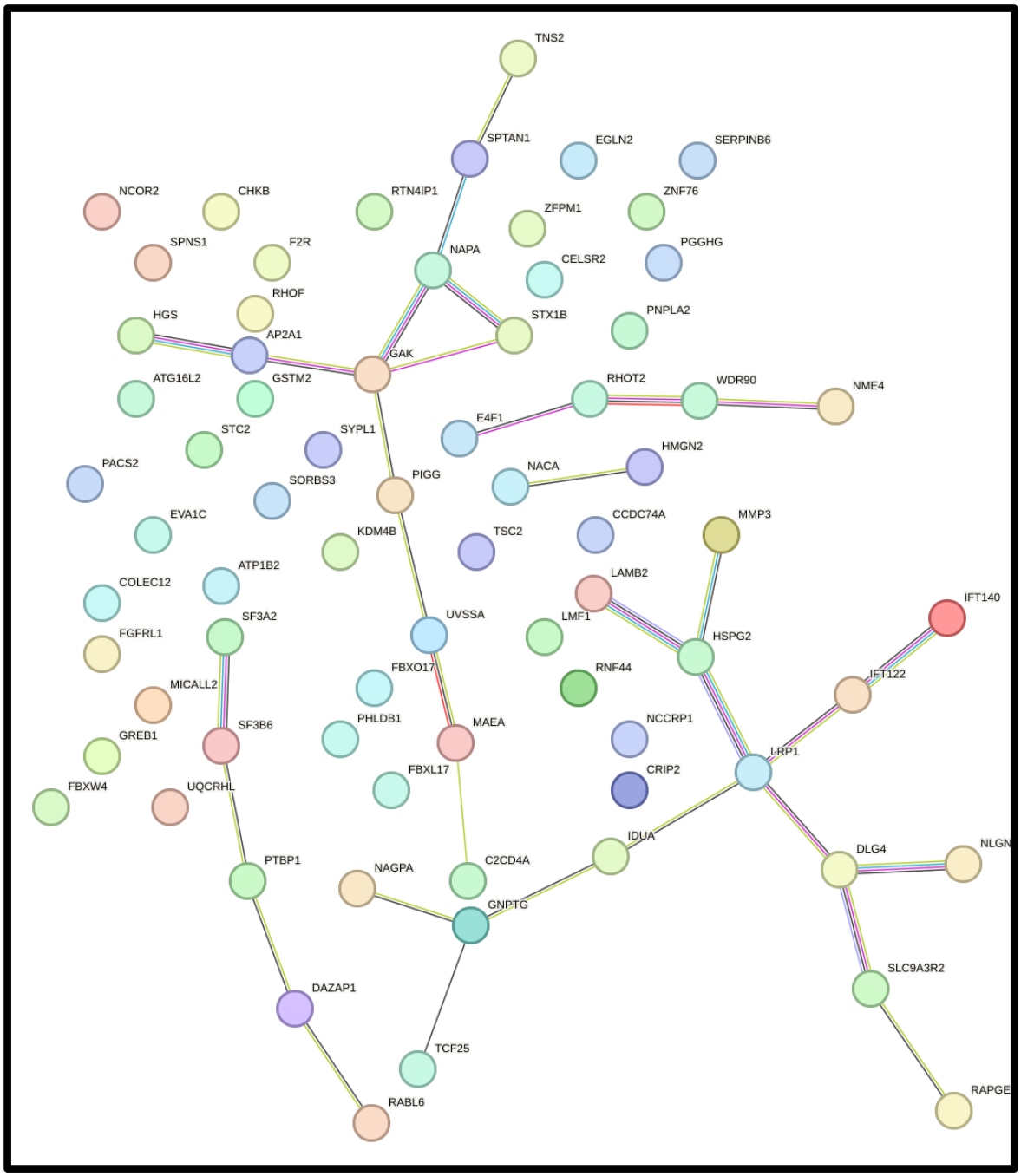


Figure 12: Top 100 Gene

CHAPTER 4

DISCUSSION

4. Discussion

As explained in previous chapter that three datasets are taken to perform this research. The main purpose of this research is to find out the crucial hub genes contributing in the triple negative breast cancer. These hub genes will eventually help us to generate the reliable and quick diagnostic method for the breast cancer. This will lead to the early treatment of the disease and prevention from the false positive or false negative results of breast cancer. For this study the Python used to analyze the datasets which are obtained from Arrayexpress. In Arrayexpress we analyzed the one dataset in detail and obtained MA plot, p-value Histogram, PC plot of all the top differential expressed genes. All these types of plots, graphs, illustrations and representations help us to differentiate and analyze both groups (ER+ and HER2+) datasets are in one group. The result of the Cytoscape was used for network analysis is in the form of a graph containing nodes which are genes and interactions between them is shown with the help of vertices lines. The thick lines show the strong interaction between those nodes and thin lines shows the weak interaction between the nodes (genes).

4.1 Conclusions

In this research, 4309 DEG's were identified. Data pre-processing of samples was done on python which includes different sequential steps like fastqc, alignment, Mark duplicate, Rm duplicate, feature count and lastly finding DEG's by PyDeseq2 library. Pathway analysis was done using DAVID and Network analysis was done by using Cytoscape. This study will help us to develop the reliable and quick diagnostic method for breast cancer to increase the survival rate of breast cancer patients.

CHAPTER 5

REFERENCES

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), 394424.

Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. CA: a cancer journal for clinicians, 70(1), 7-30.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell, 144(5), 646-674.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. Nature, 458(7239), 719-724.

Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. Nature Medicine, 10(8), 789–799.

King, M. C., Marks, J. H., & Mandell, J. B. (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science, 302(5645), 643–646.

Stratton, M. R. (2011). Exploring the genomes of cancer cells: progress and promise. Science, 331(6024), 1553–1558.

Negrini, S., Gorgoulis, V. G., & Halazonetis, T. D. (2010). Genomic instability — an evolving hallmark of cancer. Nature Reviews Molecular Cell Biology, 11(3), 220–228.

Feinberg, A. P., Koldobskiy, M. A., & Göndör, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nature Reviews Genetics, 17(5), 284–299.

Metastatic breast cancer: What is it, symptoms, and more (2023) National Breast Cancer Foundation. Available at: https://www.nationalbreastcancer.org/metastatic-breast-cancer/ (Accessed: 15 June 2023).

Gupta, G. P., & Massagué, J. (2006). Cancer metastasis: building a framework. Cell, 127(4), 679-695. doi: 10.1016/j.cell.2006.11.001

Wang Z et al. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10(1):57-63.

Oshlack A et al. (2010). RNA-Seq analysis: a practical workflow and software toolkit. Methods. 48(3):249-62.

Trapnell C et al. (2013). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7(3):562-78.

Patro R et al. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 14(4):417-419.

Brazma A et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 2003 Jan 1;31(1):68-71.

Parkinson H et al. ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 2007 Jan 1;35(Database issue):D747-50.

Kolesnikov N et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res. 2015 Jan 28;43(Database issue): D1113-6.

Benson DA, Cavanaugh M, Clark K, et al. GenBank. Nucleic Acids Res. 2013;41(Database issue):D36-D42. doi:10.1093/nar/gks1195

Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2012;40(Database issue):D13-D25. doi:10.1093/nar/gkr1184

Unger R, Sussman JL. Genomic methods for protein structure approximation. Curr Opin Struct Biol. 1996;6(2):210-216. doi:10.1016/s0959-440x(96)80065-4

PubMed Central. National Center for Biotechnology Information. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/

Alves, C. L., & Ditzel, H. J. (2023). Drugging the PI3K/AKT/mTOR Pathway in ER+ Breast Cancer. *International Journal of Molecular Sciences*, *24*(5), 4522. https://doi.org/10.3390/ijms24054522

Beltjens, F., Molly, D., Bertaut, A., Richard, C., Desmoulins, I., Loustalot, C., Charon-Barra, C., Courcet, E., Bergeron, A., Ladoire, S., Jankowski, C., Boidot, R., & Arnould, L. (2021a). <scp>ER</scp> −/ <scp>PR</scp> + breast cancer: A distinct entity, which is morphologically and molecularly close to triple-negative breast cancer. *International Journal of Cancer*, *149*(1), 200–213. https://doi.org/10.1002/ijc.33539

Beltjens, F., Molly, D., Bertaut, A., Richard, C., Desmoulins, I., Loustalot, C., Charon-Barra, C., Courcet, E., Bergeron, A., Ladoire, S., Jankowski, C., Boidot, R., & Arnould, L. (2021b). <scp>ER</scp> −/ <scp>PR</scp> + breast cancer: A distinct entity, which is morphologically and molecularly close to triple-negative breast cancer. *International Journal of Cancer*, *149*(1), 200–213. https://doi.org/10.1002/ijc.33539

Chen, J.-Q., & Russo, J. (2009). ERα-negative and triple negative breast cancer: Molecular features and potential therapeutic approaches. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1796*(2), 162–175. https://doi.org/10.1016/j.bbcan.2009.06.003

De Laurentiis, M., Cianniello, D., Caputo, R., Stanzione, B., Arpino, G., Cinieri, S., Lorusso, V., & De Placido, S. (2010). Treatment of triple negative breast cancer (TNBC): current options and future perspectives. *Cancer Treatment Reviews*, *36*, S80–S86. https://doi.org/10.1016/S0305-7372(10)70025-6

Dogan, B. E., & Turnbull, L. W. (2012). Imaging of triple-negative breast cancer. *Annals of Oncology*, *23*, vi23–vi29. https://doi.org/10.1093/annonc/mds191

Fournier, A., Berrino, F., Riboli, E., Avenel, V., & Clavel-Chapelon, F. (2005). Breast cancer risk in relation to different types of hormone replacement therapy in the E3N-EPIC cohort. *International Journal of Cancer*, *114*(3), 448–454. https://doi.org/10.1002/ijc.20710

Hausman, D. M. (2019). What Is Cancer? *Perspectives in Biology and Medicine*, *62*(4), 778–784. https://doi.org/10.1353/pbm.2019.0046

Hulka, B. S., & Stark, A. T. (1995). Breast cancer: cause and prevention. *The Lancet*, *346*(8979), 883–887. https://doi.org/10.1016/S0140-6736(95)92713-1

Li, X., Yang, J., Peng, L., Sahin, A. A., Huo, L., Ward, K. C., O'Regan, R., Torres, M. A., & Meisel, J. L. (2017). Triple-negative breast cancer has worse overall survival and cause-specific survival than non-triple-negative breast cancer. *Breast Cancer Research and Treatment*, *161*(2), 279–287. https://doi.org/10.1007/s10549-016-4059-6

Martínez, M. E., Cruz, G. I., Brewster, A. M., Bondy, M. L., & Thompson, P. A. (2010). What Can We Learn about Disease Etiology from Case-Case Analyses? Lessons from Breast Cancer. *Cancer Epidemiology, Biomarkers & Prevention*, *19*(11), 2710–2714. https://doi.org/10.1158/1055-9965.EPI-10-0742

Park, J. H., Ahn, J.-H., & Kim, S.-B. (2018). How shall we treat early triple-negative breast cancer (TNBC): from the current standard to upcoming immuno-molecular strategies. *ESMO Open*, *3*, e000357. https://doi.org/10.1136/esmoopen-2018-000357

Sandhu, G. S., Erqou, S., Patterson, H., & Mathew, A. (2016). Prevalence of Triple-Negative Breast Cancer in India: Systematic Review and Meta-Analysis. *Journal of Global Oncology*, *2*(6), 412–421. https://doi.org/10.1200/JGO.2016.005397

Sharma, P. (2018). Update on the Treatment of Early-Stage Triple-Negative Breast Cancer. *Current Treatment Options in Oncology*, *19*(5), 22. https://doi.org/10.1007/s11864-018-0539-8

Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., Shi, W., Jiang, J., Yao, P.-P., & Zhu, H.-P. (2017). Risk Factors and Preventions of Breast Cancer. *International Journal of Biological Sciences*, *13*(11), 1387–1397. https://doi.org/10.7150/ijbs.21635

Tan, T. J., Chan, J. J., Kamis, S., & Dent, R. A. (2018). What is the role of immunotherapy in breast cancer? *Chinese Clinical Oncology*, *7*(2), 13–13. https://doi.org/10.21037/cco.2018.04.01