RNA Seq Analysis Of Triple Negative Breast Cancer Cells Lines For The Identification of Common Therapeutic Targets

By

Afifa Jamil

MSBI-2019-RCMS

Supervisor:

Dr. Rehan Zafar Paracha

Research Centre for Modelling and Simulation

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

October 2021

This thesis is dedicated to *my beloved parents*

Declaration

I, *Afifa Jamil* declare that this thesis titled "RNA Seq Analysis Of Triple Negative Breast Cancer Cells Lines For The Identification of Common Therapeutic Targets" and The work in it is entirely mine, and it was created as a consequence of my own original study.

I confirm that:

1. This work was done wholly or mainly while in candidature for a Master of Science degree at NUST
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at NUST or any other institution, this has been clearly stated
3. Where I have consulted the published work of others, this is always clearly attributed
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work
5. I have acknowledged all main sources of help
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself

Afifa Jamil,
MS-BI-2019

Copyright Notice

Acknowledgments

Fisrtly, I give thanks to Almighty Allah, the most merciful and greatest, the creator of all known and unknown worlds, for providing me with the chance, strength, and drive to finish my research. Infinite salutations to the Holy Prophet (pbuh). I pay my due respect to my supervisor Dr Rehan Zafar Paracha who motivated me to do my research with hard work and dedication. I appreciate his efforts at solving all my queries, reminding my of my abilities and providing confidence to carry out my work. I would also like to thank my senior Iqra Arshad for answering any of my problems and teaching me ways to counter them. I would also like to thank my friends for their moral support. Last but not least, I am grateful for the support of my parents who constantly motivated and guided me.

Abstract

Triple Negative Breast cancer (TNBC) is a heterogeneous disease that is oestrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 negative based on immunohistochemistry. TNBC is distinguished by a specific molecular profile, aggressiveness, and a lack of targeted treatment. Some of the patients that have undergone treatment tend to relapse after a certain time period due to distant metastasis. Therapies already being used to treat TNBC include chemotherapy, targeted therapy, surgery, and radiation therapy. However, TNBC develops resistance to certain therapies over time and results in ineffective treatment. Therefore, no proper treatment is available and research needs to be done on finding new targets for drug manufacturing. In this study, we have performed RNA Sequencing analysis through Galaxy and Ballgown on two data sets of TNBC in order to determine differentially expressed genes that could be used as targets for the cancer. Then, we performed a comparative analysis to determine the common DEGs among the data sets. After this we performed pathway analysis of these DEGs using reactome, to discover targetable pathways. We found seven differentially expressed genes and ten targettable pathways. The common genes that can be targetted to treat the TNBC cancer included DDIT4, DYRK3, ELF3, H2BC21, JUN, LEPROT,and RPL21P16. The most enriched and significant pathways include Pre-NOTCH Transcription, Translation, Expression and Processing, Signaling by NOTCH, programmed cell death, Senescence-Associated Secretory Phenotype (SASP), Oxidative Stress Induced Senescence,HOX gene activation in Hindbrain development and differentiation, transcription and estrogen dependant gene expression.

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

TNBC	Triple Negative Breast Cancer
NGS	Next Generation Sequencing
WHO	World Health Organization
ER	Estrogen receptor
BMI	Body Mass Index
HRT	Hormone Replacement Therapy
ESMO	European Society for Medical Oncology
PFS	Progression Free Survival
PARP	Poly ADP Ribose Polymerase
VEGF	Vascular Endothelial Growth Factor
EGFR	Epidermal Growth Factor inhibitor
TK	Tirosine Kinase
SNP	Single Nucleotide Polymorphism
PD-L1	Programmed Death ligand 1
ADC	Antibody Drug Conjugate

LV	Ladiratuzumab Vedotin
CHD4	Chromodomain-helicase DNA binding protein 4
IPA	Ingenuity Pathway Analysis
NAC	Neoadjuvant Chemotherapy
PDGFR	Platelet Derived Growth Factor
SNEA	Subnetwork Gene Enrichment Analysis
GEO	Gene Expression Omnibus
HER2	Human Epidermal Growth Factor Receptor 2
GFM	Graph FM Index
FPKM	Fragments per kilo-base per million reads

lstlistings titlesec

Chapter 1

Introduction

1.1 Triple Negative Breast Cancer

Cancer is the abnormal proliferation of cells in the body. Cancer cells either remain in one specific location in the body, resulting in a benign tumor, or may spread to distant organs causing metastasis. Although there are many therapies available for cancer, however, many challenges are faced and more research for cancer treatment needs to be carried out [2] Carcinoma is one of the top common malignant cancers and may develop in epithelial tissues. The mutation occurs in three stages:

1. Genetically altered cell

   This involves the stage in which one cell is mutated among a population of normal cells, this leads to an abnormal cell proliferation, even at the time of rest.

2. **Hyperplasia**

   In this stage, the mutated cell and its descendants visually seem normal, however, since they are mutated, they tend to divide abnormally, which is referred to as Hyperplasia. After some time, one cell among these undergoes another mutation, that disrupts cell function further.

3. **Dysplasia**

   After further proliferation, the cells appear abnormal in shape and orientation under the microscope, this stage is known as Dysplasia. After this, further mutations may take place over time.

4. **In situ cancer**

   At this stage, the mutated cells become more abnormal in appearance and proliferation. As the tumor remains inside the epithelial tissue and does not break any boundaries or enter blood vessels, it is known as in situ cancer.

5. **Invasive Cancer**

   As the cells acquire more mutations, some may invade surrounding tissues by breaking the boundaries of the epithelial tissue. This form of tumor is known as malignant or invasive cancer. They can damage different organs in the body by invading them and spreading the mutation. [3]

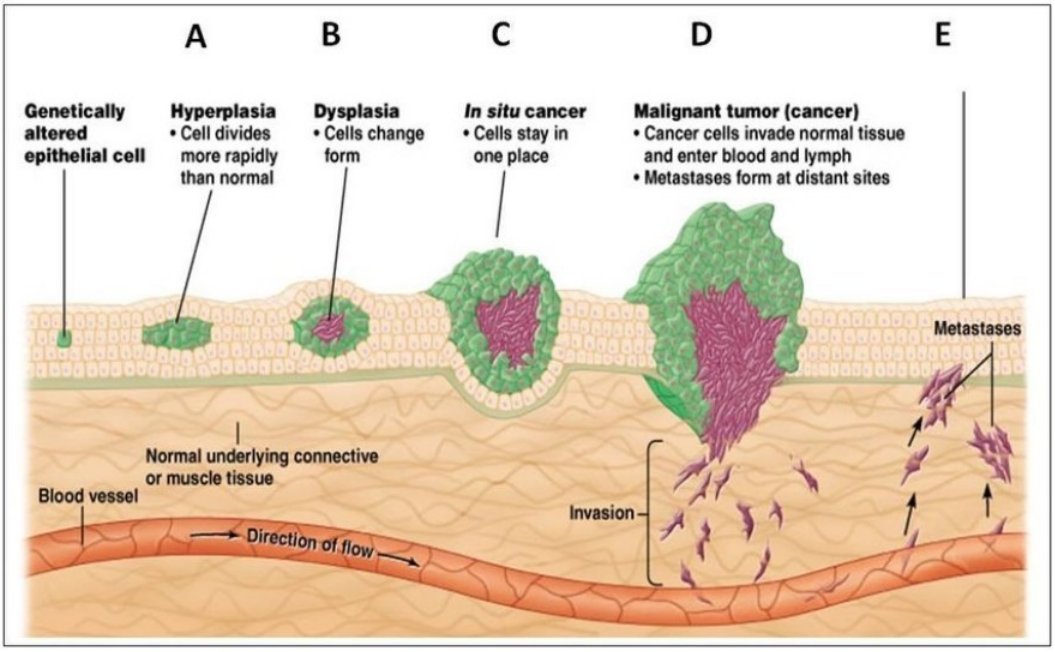


Figure 1.1: Different stages of cancer

1.1.1 Breast Cancer

Breast cancer is tumor that arises in the epithelial cells present in the milk ducts or lobules of the breast.[4] Breast cancer includes two main sub types that include non-invasive (in situ) carcinoma and invasive carcinoma. Breast cancer is one of the most common cancers found in women and is the number one cause of death in women cancer

patients. Noninvasive cancer tends to stay at its original place, whereas invasive cancer can spread out to different regions in the body from the area of origin of the cancer.[1] These two types of cancer can further be divided into subtypes which include ductal, lobular, and medullary etc.

1.1.2 Breast Cancer in Pakistan

1 in 9 women in Pakistan are prone to acquiring breast cancer. Breast cancer has caused many social and financial issues among Pakistani women. Due to Pakistan being a third world country, most of the people of Pakistan are residing in rural areas which lack proper health services such as hospitals and doctors to treat the breast cancers. In Pakistan, due to a low socioeconomic status of women, treatment of breast cancer is difficult.[5] Breast cancer consists of 36.8 percent of all female cancer cases in Pakistan. Breast cancer in Pakistan has an annual incidence rate of 43.9 percent and mortality rate of 23.2 per 100,000 people.[6] Genetic mutation in breast cancer susceptible genes such as BRCA1, BRCA2, TP53, CHEK2, RAD51C, and PALB2 account for 27 of breast cancers in Pakistan. [7] The table below shows the ranking of Pakistan in the world according to the form of cancer, that have been approved by the World Health Organization.[6]

Type	Rate	World Rank
Breast Cancer	19.33	58
Oral Cancer	9.40	10
Lung Cancer	7.81	121
Stomach Cancer	6.66	97
Cervical Cancer	6.56	78
Esophagus Cancer	6.17	42
Lymphomas	4.85	106
Colon-Rectum Cancer	4.08	149
Luekemia	3.61	112
Bladder Cancer	2.86	80
Liver Cancer	2.69	172
Prostate Cancer	2.53	164
Ovary Cancer	2.36	81
Other Neoplasm's	1.57	147
Pancreas Cancer	0.90	172
Uterin Cancer	0.51	142
Skin Cancer	0.31	170

Figure 1.2: World Ranking of Pakistan according to cancer type

In Asia Pakistan has the highest rate of breast cancer.[8] Mammogram is an effective

screening procedure of breast cancer that includes the Xray of the breast, however, in Pakistan most people can are unable to afford this program due to it being expensive, therefore doctors encourage the women to perform self-breast examination techniques and teach this technique so that the women can perform it monthly and keep a track.

1.1.3 Breast Cancer Classification

Breast Cancer Classification is an ongoing and tedious process that is undergoing more development with the increase in research. The variations in breast cancer are mostly evident in the structure of cells as seen under the microscope, hence these cancers are studied at histological and molecular levels. They can be sub typed with molecular profiling, deferentially expressed genes and many more.[9] The two main types of breast cancer classification are mentioned below.

Histopathological Classification

The histopathological tumors are classified on the basis of the diversity of their morphological features. In a study by WHO 2003, the histopathological classification includes 20 major tumors types and 18 minor subtypes.[10] Breast cancer can be classified into invasive and in situ cancers. In situ carcinoma can be further divided into ductal or lobule related carcinoma. Ductal carcinoma has a higher percentage of occurrence than lobular carcinoma. Ductal carcinoma is further divided into five sub types including Comedo, Cribiform, Micropapillary, Solid and Papillary. Invasive Carcinoma is also heterogenous and includes subtypes such as invasive lobular , infiltrating ductal, ductal/lobular, mucinous (colloid), tubular, medullary and papillary carcinomas. (fig1) Among these, invasive Ductal Carcinoma is the most common type of cancer, attributing to 70-80 of invasive cancers. [6] Invasive Ductal Carcinoma is further sub typed as either type 1 (highly differentiated), type 2(averagely differentiated) or type 3 (low differentiated) depending on the stage of cancer growth. [1]

Molecular Classification of Breast Cancer

Due to a lack of molecular insight in histopathological classification of breast cancer, the effect of targeted therapies cannot be properly studied. Hence molecular classification is necessary to classify the cancers in a more concise manner. These molecular classifications were generated using micro array-based analysis and hierarchical clustering. This classification divides breast cancer sub types into ER+ and ER-. According to molecular

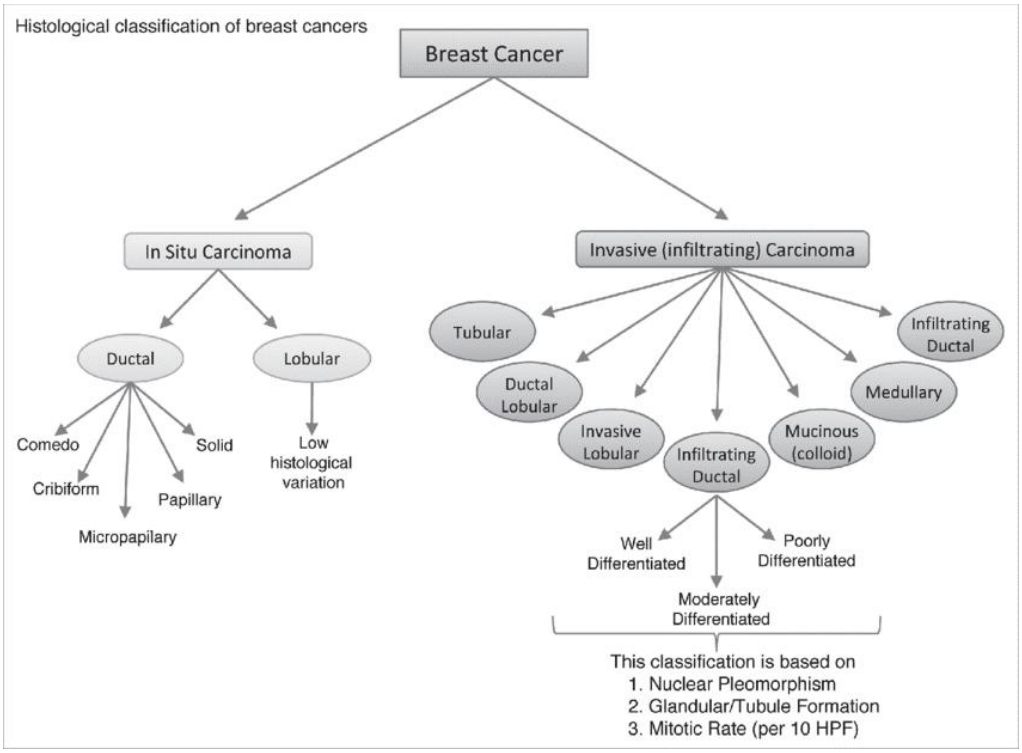


Figure 1.3: Different stages of cancer[1]

sub typing, the breast cancer is classified into three E negative known as Claudin Low, Basal Like, HER2 enriched, Normal Breast like, and two ER positive sub types known as Luminal A and Luminal B. [11]

Triple Negative Breast Cancer

Triple negative breast cancer is a hostile type of breast cancer that is associated with the lack of important receptors, namely estrogen, progesterone, and human epidermal growth factor receptor. [12] Out of 1 million patients of breast cancer, about 170,000 are estimated to be diagnosed with triple negative breast cancer.[13] TNBC is associated with a lower surviving rate compared to other sub types of breast cancer within the first 3 to 5 years after diagnosis. [14] Recent studies have show that TNBC may have a higher occurrence of visceral metastasis, involving brain metastasis as well.[15] Over time the approach to divide breast cancer into its subtypes. Two main approaches include molecular subtyping and the newly developed approach of gene expression analysis.[11] Among the different subtypes of breast cancer, basal like breast cancer holds the most common characteristics to TNBC including a low expression of HER2, ER and PR genes, depicting the TNBC phenotype.[16] Triple negative tumors are associated with

characteristics such as its onset in young people and a large mean tumor size.[17]

Risk Factors for TNBC

Triple negative breast cancer may arise in patients due to several risk factors such as age, gender, environmental factors, and genetic background etc. Hence people with such risk factors are more prone to acquire triple negative breast cancer in their lives. There are various risk factors which include the following:

Age

Triple negative breast cancer occurrence varies with different age groups. The chances of TNBC occurring in younger women aged below age 30 are very low. [18] About 85 percent of TNBC patients are women that are 50 year old and above. 5 percent of these cases are of women below 40.[19]

Lifestyle and Environmental Factors

These are risk factors that are present because of the individual's own choice. These include alcohol consumption and cigarette smoking.[20] Over consumption of alcohol can increase the chances of the individual acquiring TNBC as it results in decrease of DNA repair efficiency or nutrient uptake. It was also discovered that the effect of alcohol consumption on TNBC occurrence would increase in combination with other reasons such as a previous history of breast cancer among relatives [21] or a high BMI (Body Mass Index).[22] Another factor is Hormone Replacement Therapy (HRT) which is a therapy undergone by women who are in post-menopausal phase in order to prevent the negative effects of menopause on their body.[23]

1.1.4 Treatments of TNBC

TNBC is treated similarly as other breast cancer forms, including surgery and chemotherapy.[24] TNBC cannot be treated by using hormone receptor or HER2 based therapies. A combination of surgery and chemotherapy is the most ideal approach against TNBC as it is sensitive to many chemotoxic compounds such as anthracyclines and taxanes.[25]

Cytotoxic Chemotherapy

According to the European Society for Medical Oncology (ESMO) cytotoxic chemotherapy is one of the standard approaches used against TNBC. The mode of treatment used should be done after the evaluation of disease and patient factors.[26] There are several

compounds used in cytotoxic chemotherapy mentioned below:

Microtubule Stabilizers

These are potent tubulin polymerizers that can be used for treating TNBC. Microtubule Stabilizers such as aclitaxel, docetaxel, and cabazitaxel are more effective against TNBC than receptor positive breast cancers. [27] Another microtubule stabilizer is Ixabepilone that is often used to overcome resistance to anthracyclines, taxanes or capecitabine. A combination therapy including both Ixabepilone and capecitabine can help in treating TNBC which is resistant to anthracylcines, taxanes and capecitabine providing a higher progression-free survival (PFS).[28]

Anthracyclines

These are one of the most active compounds used against TNBC. However, the effectiveness of athracyclines against TNBC has remained disputable. [29]

Platinum Agents

Platinum compounds have a tendency to be used against TNBC due to its molecular and phenotypic similarity to BRCA-mutant breast cancer. The platinum generates extra DNA which has two strands. This prevents the generation of replication forks and causes cell death of BRCA mutated cancer cells.[30]

TARGETED therapies

PARP inhibitors

Poly ADP Ribose Polymerase (PARP) is a protein that is important for the repair of single-stranded DNA breaks through base excision repair.[31] Due to the BRCA mutation in TNBC cells, the cancer cells are unable to repair the DNA damage. PARP inhibitors are used against TNBC as they inhibit the repair function of PARP and cause TNBC cells to die, due to their BRCA mutation and lack of DNA repairing ability. PARP inhibitors are also capable of magnifying the anti-cancer effect of treatments such as ionizing radiation, DNA methylators, and platinum when used in combination.[32]

Angiogenesis inhibitors

Vascular endothelial growth factor (VEGF) is more expressed among TNBC patients compared to non TNBC patients.[33] Bevacizumab is an anti-VEGF monocolonal antibody that shows a positive response when used together with chemotherapy against

TNBC.[34]

EGFR inhibitors

TNBC constitutes a highly up-regulated expression of Epidermal Growth Factor Receptor (EGFR) inhibitors resulting in lower response to chemotherapy.[35] This has resulted in the production of anti-EGFR agents like cetuximab and lapatinib that inhibit the function of EGFR by binding to it.[36]

TK inhibitors

Tyrosine kinases are overexpressed in TNBC and result in growth of the cancer. Hence many TK inhibiotrs such as gefitinib, lapatinib, dasatinib, and pazopanib are used to target the phosphorylation of the receptor.[37]

mTOR inhibitors

the mTOR pathway is one of the highly significant pathways that are functioning in providing chemoresistance to TNBC (16) siRNA or rosiglitazone, which is a small-molecule anti-diabetic drug, can be used as an inhibitor against the compounds involved in the mTOR pathway.[38]

Chemotherapy Resistance of Triple Negative Breast Cancer

Although a lot of drugs have been discovered having anti-tumor effect on TNBC, however, they often face resistance by the tumor cells and are ineffective in eliminating the cancer. The following are the elements that help generate TNBC resistance: -

1. ABC transporters

ATP binding cassette transporters use ATP to transport different chemical compounds outside the cell through the cell membrane. They are normally present in large intestine, small intestine, and pancreas. They are involved in efflux of anti-cancer drugs used in therapy.[39] Examples of drug resistance ABC transporters include multidrug-resistant protein-1 and 8, and breast cancer resistant protein. [40]

2. Cancer Stem Cells

Cancer stems cells are able to renew a cancer due to their self-renewal properties. The examination of cancer stem cells in tumor before and after a neoadjuvant chemotherapy showed that these stem cells are not effected by the therapy and remain in the body even after the therapy, after which they replicate and divide to form cancer cells.[41]

3. TGF- Pathway

It part of a cytokine family that contains above 30 growth factors. TGF- pathway is involved in metastasis, cancer proliferation, and chemotherapy resistance. It is involved in the formation of cancer stem cells which can cause chemotherapy resistance.[42]

4. Notch Pathway

The notch pathway contains four receptors on the cell membrane and five ligands.Notch 1-4 signaling is involved in the growth of cancer stem cells, and chemo resistance.[43]

5. Hypoxia

This is used to describe the insufficient supply of oxygen in tissue. As the tumors grow due to cell proliferation, this results in the expansion of blood vessels, which eventually break. These in turn can cause low oxygen supply to the tissue. As the vessels are damaged, it can result in lack of drug transport to the tumor, or ineffectiveness of drugs that require oxygen, hence resulting in chemo resistance.[44]

Next Generation Sequencing

This is used for the sequencing of DNA, enabling the research of high through put data to broaden its spectrum and develop. These methods are faster and more cost effective when compared with sanger sequencing. The sequencing of small fragments of DNA is done in parallel.

Next Generation Sequencing (NGS) can be used for various applications which include:

- Constructing a new genome from unknown organisms. This process is known as De novo assembly and utilizes an 'assembler' to carry out the process. In this case no reference sequence is available and hence the new genome is referred to as novel.

- To identify and analyze genetic variation. This is done by comparison of the sequence to a reference genome by mapping the sequence onto the reference. This can help researchers identify mutations such as Single nucleotide polymorphism (SNPs), copy number variations and other types of mutations. Compared to sanger sequencing, NGS captures a much broader spectrum of variations.

- Perform RNA seq Analysis, this helps identify gene splicing, mutations and differentially expressed genes.

- Interrogate genomes without any bias.

Pathway analysis

This is a process that allows the analysis of pathways that contain a high expression of a specific gene list, identifying pathways that are significantly impacted in a specific condition. Pathway analysis constitutes three major steps which are;

1. Defining the gene names from genetic data
2. Identification of related pathways using pathway analysis tools
3. Visualization and evaluation of results[45]

Pathway analysis provides depth to scientific research by allowing in depth interpretation and hypothesis formation for high throughput biological data. It helps identify the biological role of candidate genes.

Chapter 2

Literature Review

2.1 Treatment of TNBC

A highly invasive cancer called as Triple Negative Breast cancer is known to have a challenging drug treatment due to the lack of three important receptors including estrogen, progesterone, and ER receptors. There is lack of an efficient treatment available for this cancer, and in most cases it results in resistance or resurgence of the tumor after a specific time period. Different treatments have been studied and used to combat TNBC tumour, including targeting of signalling pathways, neoadjuvant and adjuvant therapy, surgery, radiotherapy and targeted treatments using RNA Seq Analysis. Out of all these treatments, the best chance of recovery is provided by RNA Sequencing as it directly effects the target and usually no side effects on the surrounding area.

2.1.1 Therapeutic agents in TNBC

TNBC requires a standard approach of treatment and care which is difficult to achieve because of its short response duration to drug therapies and low survival rates. New possible targets are being investigated for TNBC using genetic analysis through next generation sequencing and potential therapeutic compounds are being tested against these targets. Clinical trials are currently being carried out to discover targets through PARP inhibition, immuno-therapy using methodology of checkpoint inhibitors, antibody-drug conjugates and molecular targeting.[46]

Immunotherapy

Tumor-infiltrating-lymphocytes presence has been shown in breast cancers and have a positive association in early and late stage TNBC. Check point inhibition is a form of immunotherapy that targets immune checkpoints in order to control the immune system and restrict cancer growth. (PD-L1) and cytotoxic T lymphocyte antigen-4 are two different types of immune checkpoint inhibitors which have shown to be helpful in curing different forms of cancer.

In a study by S Adams, a checkpoint inhibitor known as pembrolizumab was used for monotherapy of previously treated TNBC patients showing resurgence of the cancer.Pembrolizumab is shown to be good for anti-cancer therapy of previously treated patients of TNBC showing resurgence of cancer according to the research. This study has shown that it created no toxicities in the patients and the responses to this treatment were quite durable. However, this study only shows the effect of pembrolizumab on previously treated patients and the therapeutic effect on TNBC patients with no previous treatment still need to be researched on.[47] y

Atezolizumab is another immune checkpoint inhibitor. It targets PD-L1 in order to prevent and reverse T cell suppression caused by TNBC.A study was conducted to understand the combined effect of Atezolizumab and chemotherapy agent nab-paclitaxel on patients of TNBC. These patients were randomly selected to recieve a combination of Atezolizumab and nab-paclitaxel or of a placebo and nab-paclitaxel. A longer survival rate was monitored in patients who underwent the chemo-immunotherapy combination of nab-paclitaxel and Atezolizumab compared to the ones who took the placebo and nab-paclitaxel combination. Hence, using Atezolizumab with chemotherapy is an effective way to control TNBC, however, this drug showed only a few positive results and is not strong enough, requiring further research and clinical trials. [48]

PARP Inhibitors

Poly polymerase enzymes are required for DNA repair in normal functioning cells. In cancers that have mutated and defective BRCA1 and BRCA2 genes, the PARP inhibitors targets the PARP enzyme to end the DNA repair function in cancer cells. 20 percent of TNBC consists of BRCA mutation. Several PARP inhibitors under study include olaparib, veliparib, niraparib, rucaparib and talazoparib. [46]

Olaparib is a PARP inhibitor that has potential anti-TNBC effects and can be used for treatment. In a study by Eikesdal, the effect of Olaprib on previously untreated

patients of TNBC was checked through several clinical trials and the genetic expression was analyzed through DNA sequencing. Olaparib yielded a positive response rate for treating TNBC and hence it can be used to treat un-selected TNBC patients with a germ line or HR deficiency, which results in better action of the PARP inhibitor.However, chemotherapy needs to be applied after treatment with olaparib as the drug alone is not strong enough to completely treat the cancer.[49]

Talazoparib is one of the most potent PARP inhibitors which has anti-cancer properties. It has been used for treatment of different cancers such as cancers of lung, osteosarcoma, ovarian, prostate, pancreas, and chronic lymphocytic leukemia. In a study by Gamze, the effect of Talazoparib was tested on TNBC cell lines of different genetic backgrounds. Talazoparib showed to be effective against TNBC cell lines with a BRCA1 mutation, however it was not as effective for wild-type TNBC and more research is required for treatment of such types of cancers. [50]

Antibody drug conjugates

Bio pharmaceutical drug known as Antibody-drug conjugates are used for targeted therapy of cancer. They aim to kill tumour cells without harming the healthy cells through the use of a cancer specific antibody.

Sacituzumab govitecan is an ADC which combines the IgG1 antibody to the tumor associated trophoblast and the tumour inhibitory drug called SN-38. In a research, a phase 3 trial of Sacituzumab Govitecan was conducted in 468 patients of relapse metastatic TNBC and was compared to the effect of single-agent chemotherapy of the physician's choice. Sacituzumab govitecan showed longer progression free survival and hence can be used for treatment, however it has side effects such as diarrhea. [51]

Ladiratuzumab Vedotin (LV) is an ADC that consists of a humanized IgG1 and monoclonal antibody effecting LIV-1 and the MMAE which is a microtubule inhibitor. A phase 1 study of LV monotherapy showed that LV has anti-cancer effects, it can be used with pembrolizumab to effectively treat metastitic triple negative breast cancer. [52]

2.1.2 TNBC and RNA Sequencing

In order to find potential treatments for TNBC and develop drugs for it, a proper analysis of targets involved in this cancer needs to be done. RNA sequencing involves

finding potential targets through analysis of RNA present inside the samples of human DNA through next generation sequencing, it also helps in genetic profiling, detection of epigenetic changes and molecular analysis. Several researches have been done in which targets of TNBC were discovered using RNA sequencing, however drugs for these are still being developed.

In a research by Fou ou Yang, next generation sequencing through cuffdiff was applied to 4 TNBC cell lines to determine the potential of chromodomain-helicase-DNA-binding protein 4 (CHD4) as a target, discover its link to B1 integrin and identify pathways involved. Ingenuity Pathway Analysis (IPA) was used to identify pathways involved in these genes. Immuno Histochemistry staining was also performed after which scoring and statistical analysis such as chi-square and t test were applied. According to this study, 540 significant differentially expressed genes (DEG) were discovered and CHD4 was found to be linked to the integrin signalling pathway. PCR and western blotting showed that CDH4 knockdown effects expression of B1 integrin which inhibits tumor suppressors. Hence this study reveals that CDH4 can be targeted to control the cancerous effects of the B integrin, however the genes and compounds involved in the integrin pathway are yet to be studied in detail in order to develop a potential drug. [53]

In another study, the therapeutic effect of PK[11000, 11007, 11010] against the target p53 gene were studied. IC50 of proliferation by these compounds were checked in a group of 17 breast cancer cell lines. After this RNA sequencing analysis and gene ontology analysis were also checked in order to determine the change in gene expression after treatment. PK11007 turned out to be the compound with most therapeutic properties as it had lowest IC50 values. It showed to up-regulate genes involved in apoptosis and cell death, it also reactivated p53 and related pathways. Hence, PK11007 showed to be a potential drug for targeting p53. However, the exposure time of cell lines to the drug was only 12 hours, which resulted in lack of proper results. PK11007 is however known to cause toxicity and off target cell damage, hence it is not the ideal drug to treat TNBC. [54]

Nine genes including ADCY5, AFF3, FSD1, FSIP1, CMTM5, HMSD, C11orf86, CYP2A7, and ATP1A2 were shown to be potential therapeutic targets and involved in hormone related pathways in a study by Fei Chen. RNA sequencing analysis was performed on 165 TNBC samples and normal samples, breast cancers such as ER+ and HER2+ to

determine genes which are expressed differentially. Pathway analysis was analyzed using clusterprofiler package in language R. These genes effect TNBC by altering the steroid growth hormone pathways. [55]

Variants of canonical miRNAs known as isomiRNAs are also known to be potential targets for treating TNBC. In a study by Anajan Bhardwaj, RNA sequencing of a TNBC cell showing early stages of the cancer were performed. The gene targets of IsomiRNA-140-3p were identified and drugs can be formed against them to treat the cancer. These targets include HMG-CoA reductase (HMGCR) and HMG-CoA synthase 1, which are enzymes involved in the biosynthesis pathway of cholesterol. Statin and aspirin treatment on these targets showed to have a positive effect on the treatment for TNBC. However, aspirin is usually used by patients with cardiovascular diseases and hence the treatment is depicted to work on patients with both TNBC and cardiovascular diseases, and effects on only TNBC patients are yet to be researched on. [56]

Two data sets compromising of transcriptome profiles of TNBC tumours were analysed using RNA-Seq. TNBC metastasis was linked to a collection of genes involved in cell-cell adhesion, immunological regulation, and the Wnt/-catenin pathways, according to whole transcriptome sequencing. They then chose the SHISA3 gene and conducted a series of in vitro/vivo tests to investigate its biological importance. SHISA3 is a tumour suppressor gene which linked to a variety of cancers. Overexpression of SHISA3 suppresses TNBC cell proliferation, colony formation, and metastasis, as well as TNBC growth in xenografts.

SHISA3 was discovered to be a tumour and metastasis suppressor gene in TNBC, making it a promising therapeutic target for metastatic TNBC in the future. Low SHAISA3 expedites tumour aggressiveness. Modulating the canonical Wnt signalling pathway could help, at least in part. Targeted medicines that restore SHISA3 expression appear to be a potential treatment option for the most aggressive TNBC. [57]

The evolution of tumour cells in patients of TNBC in reaction to Neo-Adjuvant Chemotherapy (NAC) was studied in another research, and two unique types of clonal dynamics were discovered: extinction and persistence. After treatment, NAC eradicated the tumour cells, resulting in cell types of only normally diploid , including numerous fibroblasts and immunological cells, in the clonal extinction patients.

The clonal persistence patients, on the other hand, had a substantial number of remain-

ing tumour cells with altered phenotypes/genotypes in reaction to NAC. They further did a detailed investigation of eight patients through single cell sequencing methods of DNA and RNA, which revealed that the CNAs were selected adaptively and already existing. The findings corroborate a chemoresistance hypothesis in which two mechanisms of adaptive evolution and acquired evolution were at work to create the resistant tumour clump.[58]

In another study, allelic prevalence with variations and combinations through RNA sequencing of TNBC and non-TBNC were analyzed. The findings population datasets to show the most common variant alleles, and confirmed a handful of novel variations in tumour genes including TPP1, MAD2L1BP, GLUD2, ESRP2, and SLC30A8. It was shown that an uncommon mutation in ESRP2 inhibits its capacity in connecting to FGFR2 to further validate the results. [59]

Although estrogen receptor alpha is not expressed in TNBC positive patients, estrogen receptor beta is known to exist in 50 to 90 percent of estrogen alpha receptor negative cancers and is found in TNBC patients as well. According to a study, estrogen beta receptor is a potential target for TNBC. In this research a TNBC cell line was engineered with inducable expression on estrogen receptor beta and then effects on tumor cycles were checked. RNA sequencing was performed to determine target genes of the estrogen receptor. These target genes are linked to the Wnt/-catenin pathway and can be targeted through ligands to treat TNBC. However, the ligands will be cell context specific and require further research for drug development. [60]

2.2 TNBC Pathway Analysis

Pathway Analysis is a method to analyze data including omics, transcriptomics, proteomics, and metabolomic data. For this purpose different tools are used such as Enrichr, Ingenuity Pathway Analysis and Reactome etc, libraries containing usable genomic data such as KEGG and GenBank are used by these tools to analyze potential pathways of the target genes.

2.2.1 Previous studies on TNBC pathways

In a study the major cellular pathways that are altered in TNBC were studied using a pathway analysis tool known as sub network enrichment analysis (SNEA) which uses global literature to generate list of genes involved in different pathways and can be analyzed to determine pathways of DEG. Changes in the biosynthesis of pyrimidine or purine, production of energy, and metabolic nicotinamide of biochemical pathways were observed for TNBC genomic data. The repair synthesis of a DNA gene in DNA repair pathways exhibits an upregulation in TNBC data and hence these can be considered as therapeutic targets for further research to develop potential drugs.

In another research by Safia Shaheen, the RNA Sequencing analysis and pathway analysis of two tnbc resistant cell lines treated with two different drugs including JQ1 and Dexamethasone, this was done in order to find out effects of treatment and what pathways should be targeted to avoid the resistance of drug. The cytokine receptor interaction pathway was shown to be a good potential target to be studied for treatment of TNBC. Cytokine signalling is suspected for causing tumor proliferation and causes growth of blood vessels which further promotes cancer growth. [61]

On the basis of gene expression analysis, a study was conducted to determine TNBC's molecular signature. Comparing the sample with the normal breast tissue reveals that TNBC has molecular networks, as seen by changes in the cell cycle, inflammatory response, metabolic pathway, NF-B signalling, repairing of DNA and angiogenesis. TNBC was also classified as a cancer with mixed phenotypes, implying that it might be considered as a distinct subordinate of breast cancer. The findings shed light on the biology of TNBC for the first time. In this research Pathway Studio 7 was used for sub-network gene enrichment analysis (SNEA) of TNBC cell lines to determine important pathways and changes of gene expression in cancer. In TNBC, analysis of gene expression also shows that repair pathways of DNA, like mismatch repair, homologous recombination, and repair synthesis of DNA genes were upregulated. Five genes known as KRT17, KRT5, SFRP1, BCL2, and KRT14 were found to be down regulated in TNBC, and hence can be affected for TNBC. [62]

Through silico technique, Yin and Yang, the two opposite effects in cancer molecules dictate the the upcoming response of cancer cells. Identifying these opposing constituents could help choose possible treatment targets and further characterise the diverse TNBC

subtype. Here, Yin pathways which are more active in cancer and yang pathways which are more active in normal were identified using Gene Set Enrichment Analysis (GSEA). TNBC's pathways were compared to those of other forms of breast tumours using clustering analysis. In TNBC, the FOMX1 pathway was found to be the top upregulated, whereas the PPAR pathway was shown to be the most downregulated. Further research could attack both of these mechanisms at the same time. [63]

In another research the goal of the study was to determine the progressive mechanisms of TNBC by identifying important pathways and genes (TNBC). Methods. Gene Expression Omnibus (GEO) datasets were used to obtain GSE76275 gene expression profiles. They worked on a microarray Super-Series set that contained data from two hundred and sixty five samples, sixty seven of which were non-TNBC while one hundred and ninety eight of which were TNBC. Then, with a p0.01 and a fold change of 1.5 or -1.5, all DEGs were recognized. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes pathway (KEGG) enrichment studies were done on 56 upregulated and 151 downregulated genes. Out of these important DEG that could be used as targets included C9orf152, AR, SOX8, RAB30, and NRK. [64]

Because of its aggressiveness and poor prognosis, TNBC has been in discussion compared to other types of breast cancer. In a research microarray analysis of samples was done to identify DEG and their pathways. Gene Expression Omnibus yielded six microarray data sets containing four hundred and sixty three non-TNBC and four hundred and five TNBC samples. Meta-analysis was used to combine the data sets, resulting in the discovery of 1075 DEG.

A protein interconnection 486 nodes and 1932 edges was created, with 29 hub genes identified using metrics of high topology. On the training data set, 16 characteristics (hub genes) were identified, 12 of which were upregulated (CDC20, NUP88, EGFR, CCNB2, AURKB, PLK1, ENO1, DDX18, MYC, PML, POLR2F and SKP2), and four of which were downregulated (TGFB3, CCND1, SKP1, TGFB3, and GLI3). [65]

2.2.2 Aims and Objectives of our Research

As TNBC is a highly complex breast cancer and lacks three receptors including human epidermal growth factor receptor 2 (HER2), progesterone receptor (PR) and estrogen receptor (ER) it is difficult to treat this cancer. The different drugs already available to

TNBC have cytotoxic side effects, show resistance or are not strong enough to completely inhibit the disease. Due to lack of proper targets that are specific to TNBC, there is no proper treatment available.

Inter and intra humoral heterogeneity of this type of cancer is hard to identify targetable pathways. Hence a thorough analysis of the differentially expressed genes at a molecular level needs to be carried out and the pathway analysis of these genes should be done to analyze the common pathways between these genes and other diseases.The combined studies will provide insights to the relation of common differentially expressed genes between different data sets and help to determine the therapeutic targets.

These common genes will allow us to map a pathway in which these genes are involved. This pathway can further be studied and researched on, so that we can use it as a target for furture therapies to treat TNBC.

Objectives of this study include

1. Identification of DEG in the two data sets through RNA Sequencing Analysis.
2. Comparison and identification of the common DEG present between the two data sets through a comparitive analysis.
3. Identification of common pathways related to the differentially expressed genes through pathway analysis using Reactome.
4. Identification of targets to triple negative breast cancer.

Carrying out these procedures will help us identify differentially expressed genes, possible novel DEGs as welland identify pathways. This will enable us to discover targetable pathways and genes, which can later be researched on to discover possible drugs to treat the TNBC patients.

Chapter 3

Methodology

3.1 Next Generation Sequencing Data Analysis

For the processing of NGS datasets, an online web server known as Galaxy is often used. It is an open source and web based platform used for biomedical analysis. It includes many bioinformatics tools that can help in data analysis of biologists. Galaxy provides 250 GB memory to every registered member. Being an online platform, it is easier to store and process data. The workflow of RNA Sequencing is mentioned in the diagram 3.1.

3.2 Data Collection

For the RNA Seq Analysis data was acquired using a publicly accessible repository known as Array Express. For the selection of a good quality data set, the following points were kept in mind;

- data sets of homo sapiens were only included
- both data sets included cancer cell lines of triple negative cancer
- data sets not involving any form of treatment were utilized
- if the data set included any other form of disease besides triple negative breast cancer, they were removed.

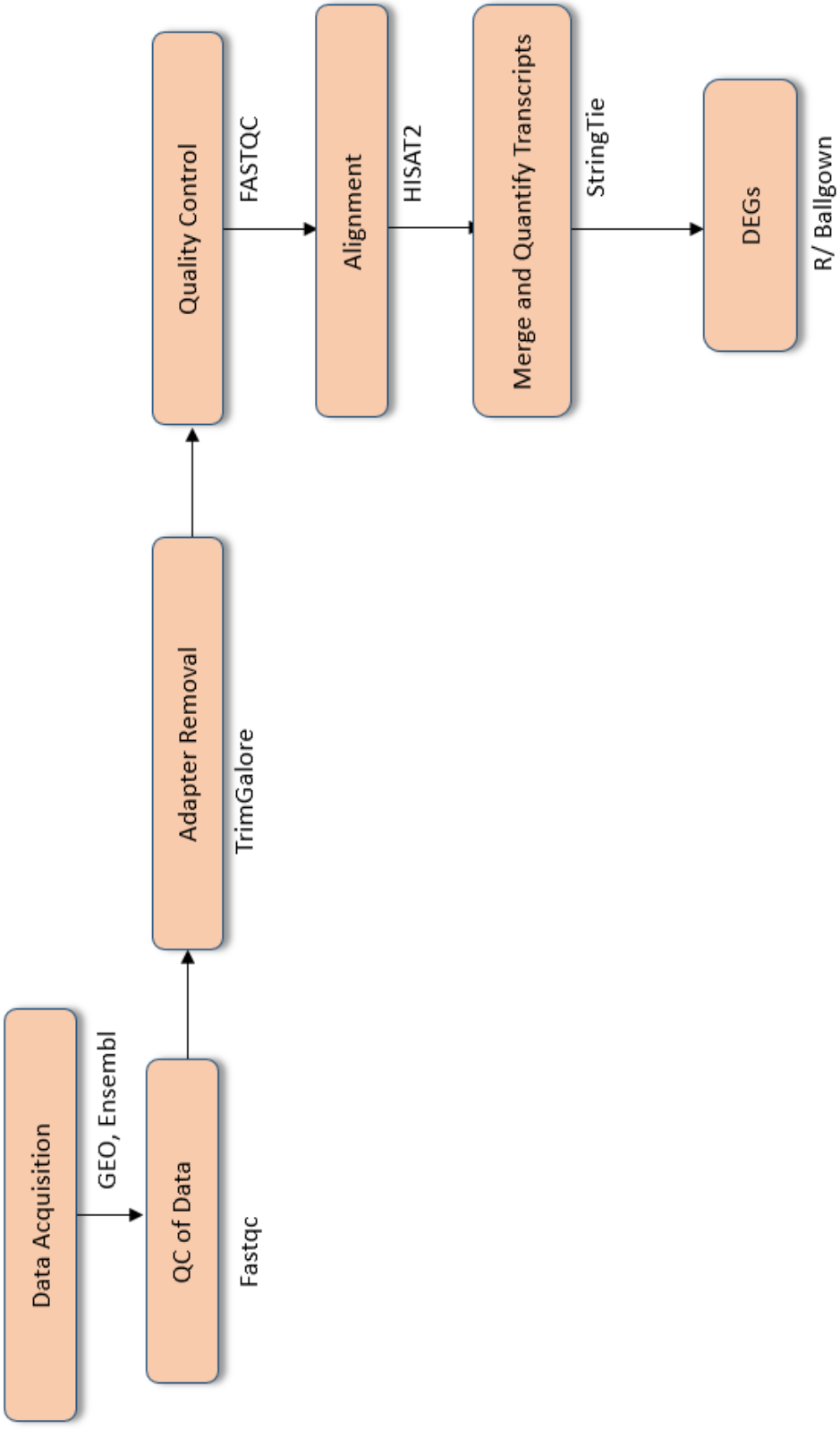


Figure 3.1: RNA Sequencing workflow

3.3 Data Import

Galaxy is a tool available on the online web used to perform biological data analysis and involves graphical user interface, hence it is used globally by the scientific community. It is a data integration platform on which the user can upload the data through their computer, URL, or through many other online sources of biological data. Galaxy is used for proteomics, epigenetics, gene expression, transcript omics, and many other types of studies. Galaxy uses numerous tools for the analysis of biological data uploaded by the user, in order to generate new pipelines with a better output.

The data from the different data sets was imported and loaded from the ENA database on to the galaxy platform. The files were in the form of FastQ files. There are two data sets of breast cancer cell lines, both of them are paired end.

3.4 RNA Seq Analysis

RNA Seq Analysis was then executed on these data through several galaxy tools. The main tools used in the RNA sequencing process include HISAT2, Stringtie and Ballgown. These tools run fast while requiring less memory, they provide better results compared to many other tools. RNA Sequencing analysis was done using galaxy as the domain.

3.5 Quality Assessment and Preprocessing of Data

The high throughput sequencing data may have errors such as adapter contamination, biases in base content or over represented sequences. Hence the quality of the data needs to be checked and any artifacts or low quality data has to be removed so that the research results do not have any error inside them. For this several processes were used which are used for preprocessing.

A) Quality Check of Raw data

For assessment of quality of the data, a tool of galaxy known as FASTQC is applied on the data and the results were analyzed by checking the quality control report of FASTQC. Each FastQC module has its own result part in the report. A sign indicating the passing, rejection and warning is shown in addition to the graphical or list data it

provides. When evaluating sequence data, researchers should be very cautious about relying on these indicators. The criteria utilized to assign these flags are based on a very particular set of assumptions that apply to a very specific set of circumstances. The quality control report includes the analysis results of the following:

1. "Basic data statistics including file name, file type, total sequencing, GC content."

2. "Per base sequence quality that is described by a quality score."

3. "Per sequence quality score which includes a graph comparing the total number of readings vs. the average quality score for that read over its whole length."

4. "Per base sequence content which includes the percentage of every nucleotide that is called throughout the length of the sequence."

5. "Per sequence of GC content"

6. "Per base of N content (nitrogen content)"

7. "Sequence length distribution"

8. "Sequence duplication level"

9. "Over represented sequences"

10. "Adaptor content"

B) "Quality Control"

In order to have good quality data we have to eliminate the errors and bad quality reported through FASTQC. This can be done through performing quality control and preprocessing of the data before applying RNA Seq tools. Here we used Fastp (version:0.19.7) developed in 2018. It is faster than other preprocessing tools and uses C++ language. Fastp includes almost all the functions of preprocessing including sequence trimming and formatting. The raw FASTQC files collection was used an input to Fastp tool which resulted in two outputs:

1) an html quality control report

2) fastqsanger file

Default parameters were used in fastp. Quality filtering options were set according to the requirements.

C) Second Quality Check of pre-processed data

The pre-processed reads were again put through quality check using FASTQC in order to determine if their quality has improved. After approval of quality, these reads are subjected to further analysis.

3.6 Alignment of processed reads to the genome

For the alignment of the processed files, HISAT2 version 2.2.1 was used. HISAT2 can perform fast alignment and it utilizes Burrows-Wheeler transform (BWT) to make graphs, a graph FM index (GFM), with a Hierarchical Graph Index (HGFM) to perform its tasks. Hierarchical Graph FM index is an indexing approach that includes a global GFM index that shows a detail of human genome and also a small list of GFM indexes which help in efficient alignment. The reference used for alignment was the human genome (hg38).

3.7 Post Alignment Processing

After performing alignment, the aligned BAM files were checked again for their quality using FASTQC and duplicate reads were removed in the following ways:

3.8 Read Duplication

Read duplication (RSeQC) version 2.6.4 is used to identify mapping and sequence based duplicate reads. This tool consists of two modules to analyze data that include basic and RNA seq specific modules. The aligned BAM file was then analyzed using the read duplication tool to identify Principle Component Analysis (PCR) duplicates. The upper limit of plotted duplicate times was set to 500, while the minimum mapping quality was set to 30. The output R script was set to 'No'. This tool helps us understand the level of duplicate reads present in our data due to PCR replicates. The result of the tool is visualized graphically through RSEQC plots.

3.9 Mark Duplicates and RmDUp

The Mark Duplicates tool is used to locate and mark the duplicate reads in the output files of Read Duplication tool. This tool ranks reads by the sum of their base quality score and can identify the duplicate reads from the original sequence. It works by comparing the sequences in 5 prime positions of the read pairs. Mark Duplicate tool is used to identify and duplicate the reads in our data present in the form of BAM file. Then the RmDup tool version 2.0.1 is used for duplicate deletion that involves removal of duplicates as reads with the lowest mapping quality are removed.

3.10 StringTie

The StringTie tool version 1.3.4 is a tool available in the galaxy server which helps in estimating the gene and transcript expression levels for aligned data files by assembling them into potential transcripts. A reference annotation file of the human genome known as hg38 is used in GTF format for the assembly of transcripts.

The following input was provided to StringTie:

- "Are there any long reads in the input?" No
- "Specify strand information:" Forward
- "Is it possible to use a reference file to assist in assembly?" Use reference GTF/GFF3
- "Reference file 1105": human reference.gtf
- "Use Reference transcripts only?" No
- "Output files for performing differential expression?" Ballgown
- "Output coverage file?" No
- "Output gene abundance estimation file?" Yes
- "Default parameters were set for remaining input"

This results in creation of multiple isoforms. StringTie is performed on the output files of RmDup. It resulted in output files that include the assembled transcripts, intron, exon and transcript level expression measurements, exon to transcript mapping and intron to transcript mapping files.

3.11 Ballgown Analysis

This tool is used to perform differential expression analysis in order to determine differentially expressed genes of the data. It performs statistical analysis of the assembled transcripts. The output files of StringTie are further processed by Ballgown. It provides graphs and tables of the DEGS and transcripts present in the data in the form of excel and pdf files.

A) Visualization and Analysis of differentialy expresssed genes

Differentially expressed genes can be visualized through the enhanced volcano plot formed through R. The x axis of the graph represents the log2 fold change of the genes while the y axis represents the –log10 p value. This plot represents the over and under expressed genes of the data set and can help identify the significant differentially expressed genes as it also shows Log FC.

3.12 Comparison of differentially expressed genes

The differentially expressed genes of the three types of data analysis including triple negative vs normal, normal vs uninvolved, and uninvolved vs triple are compared in order to determine the common differentially expressed genes. This comparison can be done using online list comparison tools. The results determine the differentially expressed genes of the overall dataset.

3.13 Pathway Analysis

Biological pathways are regulatory mechanisms and processes that help regulate the internal system by activating other mechanisms or synthesize important substances. Gene expression can be used to identify which pathway is related to a specific disease. If a certain gene is active or inactive in that disease it can help identify targets for the disease. Pathway analysis can be done through online tools such as Reactome and Enrichr that use existing databases to determine the pathways a gene is involved in.

The common differentially expressed genes can be analyzed through pathway analysis using reactome. Reactome is freely available and reviewed pathway database. It helps

in the interpretation and visualization of pathways. uses several databases including Ensembl, Uniprot and KEGG etc. Hence it uses these data bases to identify pathways activated or inhibited in our disease.

Chapter 4

Results

4.1 Next Generation Sequencing Data Analysis

4.1.1 GSE52194

(A) Preprocessing of Transcript Data

Analysis of the transcript data obtained from galaxy is done using Ballgown Analysis. This generates different forms of graphs that are shown in the form of figures. In figure 4.1 (a) a boxplot of log2 transformed FPKM (fragments per kilo base of exon model per million reads mapped) is shown .

It shows that the median of each sample coincides around 0.5 of log2 transformed FPKM values. The whiskers represent the spread of the data while the dots represent the outliers of the data.

The distribution of transcript count per gene is presented in 4.1 (b). Here transcript per gene is plotted on x axis while frequency of transcript is shown on the y axis. Gene consisting of exactly one transcript are 55000 while genes that are present in more than one transcript are shown to be 10000 in number. The maximum number of transcript for one gene is 5.

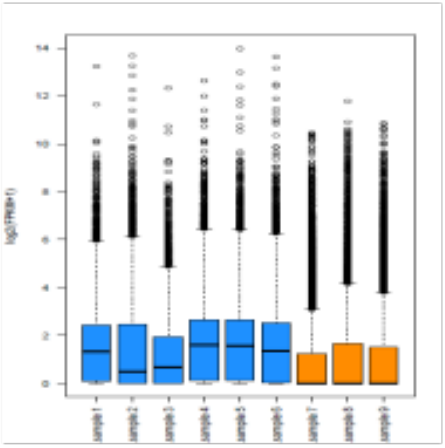
In figure 4.1 (c) The distribution of transcript length is shown. Here the transcript length in base pairs is represented in at the x-axis while the frequency is shown over the y axis. The distribution is skewed positively, with right handled long tail. The majority of the transcript are present with a length of 2000 bp.

Lastly, in 4.1 (d) figure, the frequency distribution of log2 fold change value of all genes

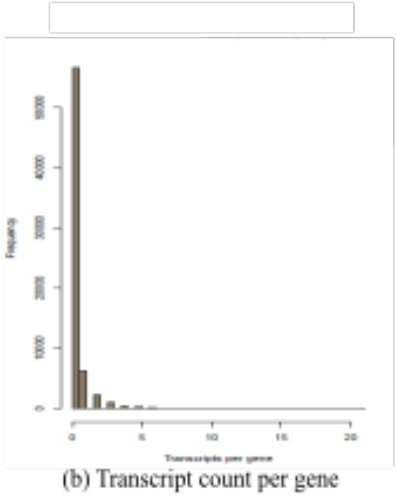
are represented. Fold change values are present on the x-axis while the frequency of the differential expression values is shown on the y-axis. A cutoff value of ±0.5 is chosen in order to filter out the biologically significant genes.

4.1.2 Visualization of DEGs

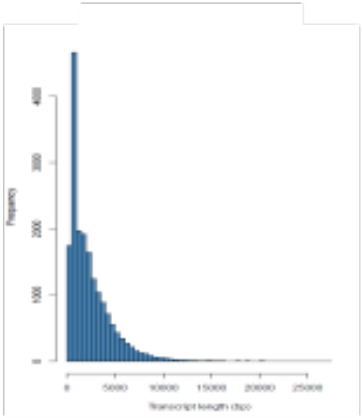
Enhanced volcano plot in Figure (e) represents the visualization of the DEGs through their p value with the log2 of fold change which represents overall magnitude of change. Log2 fold change is shown over the x axis while the log p value is plotted on the y-axis. The plot represents the significant genes to have a p value lower than 0.05 and a log2 fold change that is above ±0.5. Each gene is represented by a single dot on this plot. The Red dotted genes are biologically and statistically significant as they pass the criteria of significance with a p-value lower than 0.05 and a fold change above ±0.5. The ten most significant differentially expressed genes found in this data are shown in the table below.



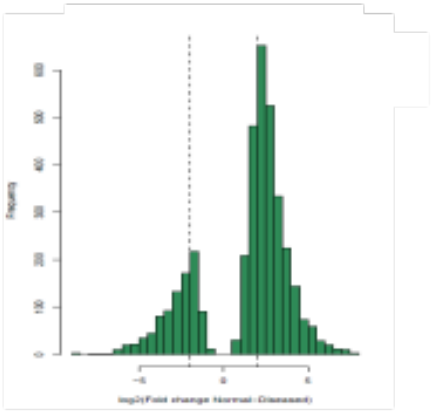
(a) Distribution of FPKMs



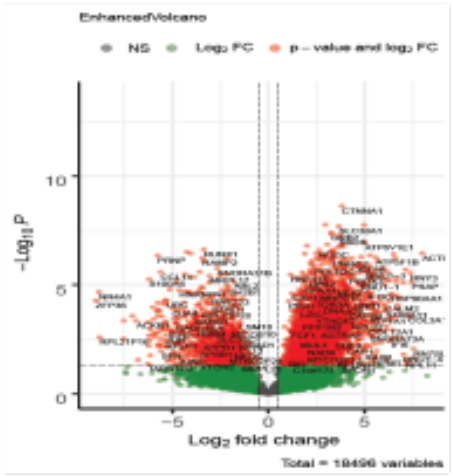
(b) Transcript count per gene



(c) Distribution of transcript lengths



(d) Distribution of D.E valuese



(e) Enhanced Volcano plot

Figure 4.1: GSE52194: (a) This shows the distribution of FPKM. (b) The number of transcripts produced by each gene.(c) The length of transcripts for each gene. (d) Frequency and Log2 fold change values of the genes. (e) Volcano plot showing the significant DEGs.

Gene name	Fold Change	Differential Expression	pval
GRB2	9.576300719	3.259468457	4.00E-08
FKBP8	12.97010439	3.697118186	4.89E-08
SLC38A1	12.63829094	3.659729478	2.03E-08
FAM32A	7.77234207	2.958349397	1.73E-08
CALM1	31.50327355	4.977429844	1.80E-08
ARCN1	12.4128674	3.633764515	1.82E-07
ATP6V1E1	32.40727148	5.018245653	1.15E-07
RUNX1	0.097049314	-3.365138177	2.37E-07
NUDC	6.244007897	2.642472362	2.55E-07
HACD3	13.57508602	3.762889435	1.53E-07

Table 4.1: Top 10 differentially expressed genes GSE2194

4.1.3 E-GEOD-58135

In this data set, there were a total of three types of data, including data of triple negative breast cancer, normal and un-involved breast tissue. RNA Sequencing analysis of all three types of data were done against each other and then compared. The results are shown below:

(1) Normal vs TNBC

(A) Preprocessing of Transcript Data Analysis

The results of the ballgown analysis of Normal vs TNBC cell lines is shown on figure 4.2. In (a) A boxplot showing the distribution of FPKM values is shown. The median of the box plots coincides around 0.5 of log2 transformed FPKM. The whiskers of the data spread upwards with outliers ranking upto 16 of log2 FPKM.

The distribution of transcript count per gene is presented in (b). Genes consisting of exactly one transcript are almost 55000 while genes that are present in more than one transcript are shown to be about 10000 in number. The maximum number of transcript for one gene is 4.

The distribution of transcript length is represented in figure (c).The distribution is skewed positively. The transcript lengths reside from 0 to 5000 bp, with most of them having a transcript length of 1000 bp.

Lastly, in (d) figure, the frequency distribution of log2 fold change value of all genes are represented. Fold change values are present oVer the x-axis while the frequency of the differential expression values is shown at the y-axis. A cutoff value of ± 0.5 is chosen in order to filter out the biologically significant genes.

4.1.4 Visualization of DEGs

Enhanced volcano plot in Figure (e) shows the visualization of the DEGs through their p value with log2 of fold change which represents the magnitude of change.The genes represented inside the box formed around 0 fold change are considered insignificant and hence are not considered in our research. Most of the genes on the plot are residing at the left hand side of the plot, which means that they are down regulated genes and are less expressed in the cancer than in normal cell lines. The up-regulated genes located at the right side of the plot include TNSRSF25, MIIP and TARDBP etc. These up

regulated genes have a p value below 0.05 and a log2 fold change which is above + 0.5. Hence these genes are biologically and statistically significant as they pass the criteria of significance with a p-value below 0.05 and a fold change bigger than ±0.5. The top ten differentially expressed genes found in this data include shown in the table below.

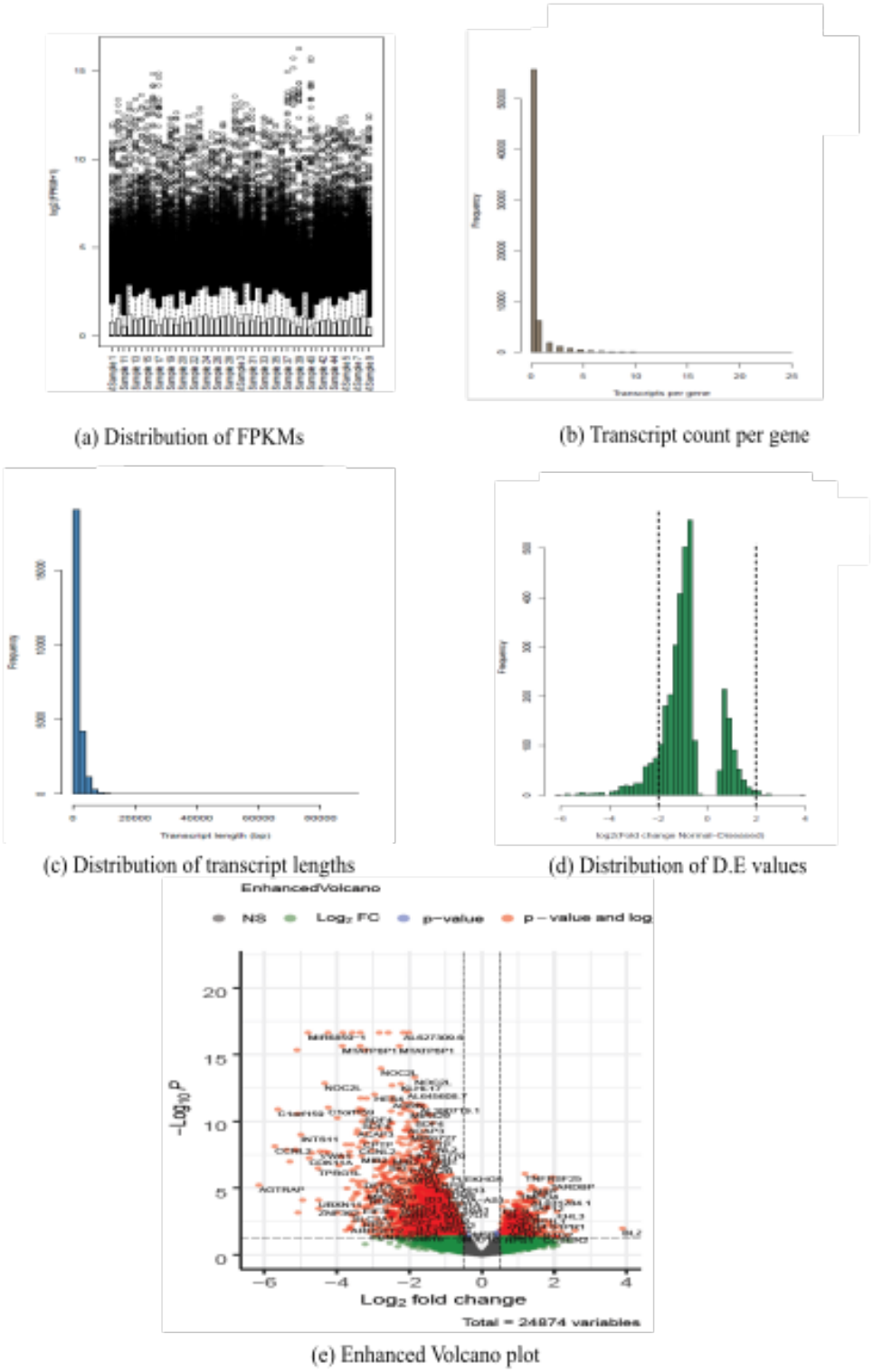


Figure 4.2: TNBC vs Normal: (a) The distribution of the FPKMs of every sample (b) Number of transcripts mapped by each gene (c) Transcript length of genes (d) Distribution of gene frequency according to fold change (e) Enhance Volcano plot showing significant DEGs

Gene name	Fold Change	Differential Expression	pval
BLZF1	14.72	3.87	0.010705454
AL591846.1	5.933060602	2.568776519	0.01555983
KDF1	5.305260405	2.407423567	9.15E-05
JMJD1C	5.053058416	2.337156858	0.031834653
MIA3	4.537616914	2.181934815	0.017695331
ECHS1	4.334115557	2.11573762	0.046222317
FHL3	4.217192864	2.076283001	0.000580588
HDGF	4.181649964	2.064072302	0.008117528
TIAL1	4.122798198	2.043623846	1.17E-05
RPL12P14	4.099829423	2.035563886	0.00087725

Table 4.2: Top 10 differentially expressed genes TNBC vs Normal

(2) Triple Negative Breast Cancer vs Uninvolved

(A) Preprocessing of Transcript Data Analysis

Analysis of the transcript data obtained from galaxy is done using Ballgown Analysis. This generates different forms of graphs that are shown in the form of figures. In figure 4.3 (a) a box plot of log2 transformed FPKM is shown.

It shows the Distribution of FPKM values of a total of 53 samples, with 40 samples of TNBC and 13 samples of uninvolved cancers. It shows that the median of each sample coincides around 0.5 of log2 transformed FPKM values. The whiskers show the distribution of the FPKM values while the dots represent the outliers of our dataset.

Overall distribution of transcript count per gene is shown in figure 4.3 (b). Here transcript per gene is plotted on x axis while frequency of transcript is shown on the y axis. Gene consisting of exactly one transcript are 6249 while genes that are present in more than one transcript are shown to be 5246 in number. The maximum number of transcript for a single gene is 25.

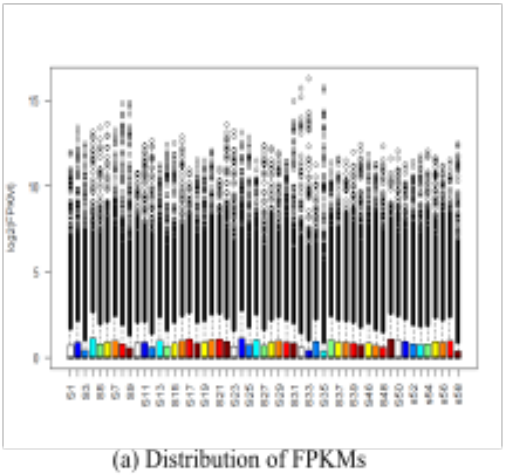
The distribution of transcript length is represented in figure 4.3 (c). Here the transcript length in base pairs is shown in at the x-axis while the frequency is shown at the y axis. The length of transcripts vary from 0 to 10000 bp. Most of the transcripts (upto 20,000) have a length of upto 2000 bp. The distribution is skewed positively with a right handed long tail.

In figure 4.3 (d) , the frequency distribution of log2 fold change value of all genes are represented. Most of the genes are showing to have a negative fold change up to -2, with about 700 genes having a fold change between 0 to -1. In the right side of the graph, the genes with a positive fold change are shown. Upto 500 genes are showing a fold change between 0 to +1. Hence this shows that over all mostly the differentially expressed genes have more expression in uninvolved cell lines rather than in the TNBC disease.

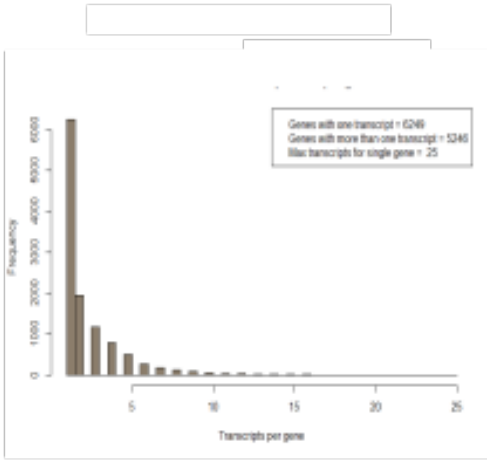
4.1.5 Visualization of DEGs

Finally, the Enhanced volcano plot in represented in Figure 4.3 (e) shows the visualization of the DEGs through their p value and FC value. The plot represents the significant genes to have a p value much below 0.05 with a log2 fold change which is above ± 0.5.

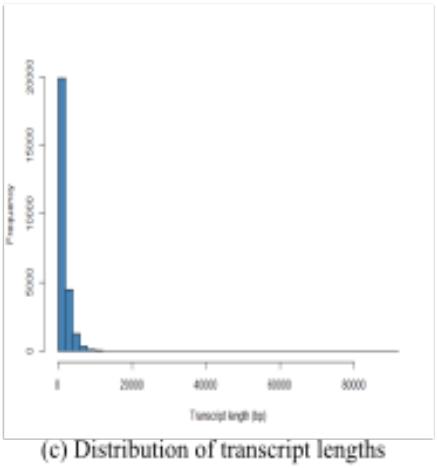
Most of the genes are scattered at the left hand side of the plot, having a negative fold change, which shows that these genes are more expressed in uninvolved cell lines rather than in TNBC cell lines. The top ten expressed genes found in this data include shown in the table 4.2 below.



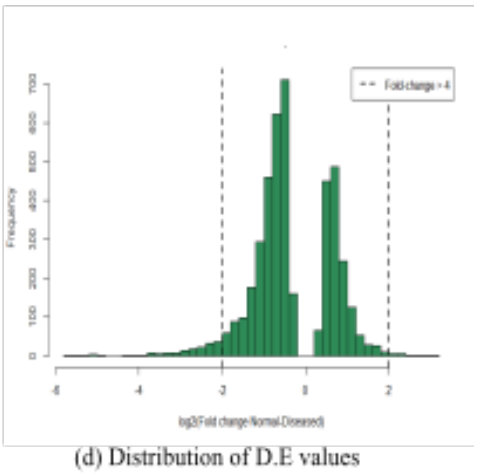
(a) Distribution of FPKMs



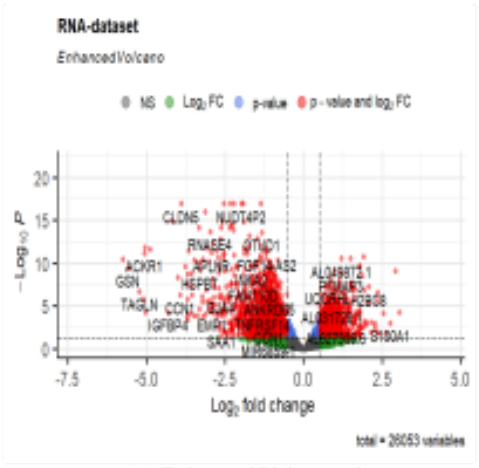
(b) Transcript count per gene



(c) Distribution of transcript lengths



(d) Distribution of D.E values



(e) Enhanced Volcano plot

Figure 4.3: TNBC vs Uninvolved: (a) The distribution of the FPKMs of the samples (b) Number of transcripts per gene (c) Transcript length of genes (d) Distribution of gene frequency according to fold change (e) Enhance Volcano plot showing significant DEGs

Gene name	Fold Change	Differential Expression	pval
ISG15	5.076760238	2.343908127	0.011150153
S100A1	6.752165464	2.755350258	0.000981534
RAB25	4.815146152	2.267579588	0.00337427
MEX3A	5.176973039	2.372108805	2.79E-07
CLDN9	3.19933529	1.677772195	5.80E-05
MIR3648-1	3.108647272	1.636286928	0.000191662
RASD2	3.798777145	1.925535079	0.000627586
RPL29	3.502438636	1.808359774	0.035277879
RNA5SP202	3.726484045	1.897815083	0.003192961
MIR6845	3.019645696	1.594379284	0.00107259

Table 4.3: Top 10 differentially expressed genes TNBC vs Uninvolved

(3) Normal vs Uninvolved

(A) Preprocessing of Transcript Data Analysis

Analysis of the transcript data obtained from galaxy is done using Ballgown Analysis. This generates different forms of graphs that are shown in the form of figures. In figure (a) a box plot of log2 transformed FPKM is shown .

This graph represents the log2 transformed FPKM of a total of 18 samples, including 5 normal and 13 uninvolved samples. The median of the box plot each sample coincides around 0.5 of log2 transformed FPKM values. The whiskers are presenting the distribution of the FPKM expression while the dots represent the outliers of the data, which spread above 12 log2 FPKM.

The distribution of transcript count per gene is presented in (b). The transcript counts vary from 0 to 10. About 5100 genes have a single transcript. While about 5020 genes have more than 1 transcript per gene is plotted on x axis while frequency of transcript is shown on the y axis.

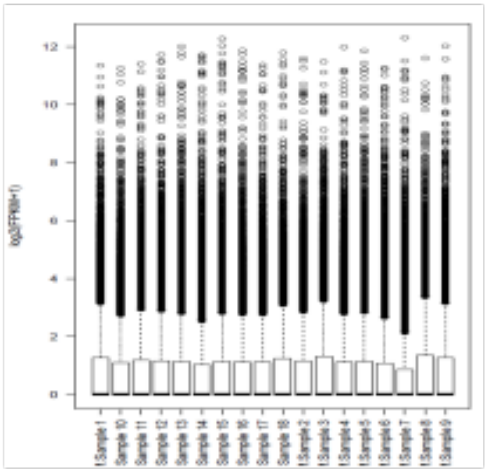
In figure (c) The distribution of transcript length is shown. On the x-axis, the transcript length is indicated in base pairs, while the frequency is shown on the y axis. The distribution is skewed positively with a right handed long tail. The majority of the transcript are present with a length of 1000 bp.

The frequency distribution of log2 fold change values for all genes is depicted in the (d) figure. X-axis displays fold change values, while the y-axis shows the frequency of differential expression values. A cutoff value of ± 0.5 is chosen in order to filter out the biologically significant genes. The graph is more concentrated on the left hand side, indicating a higher number of genes having a negative fold change and hence being more expressed in normal while being down-regulated in uninvolved sample.

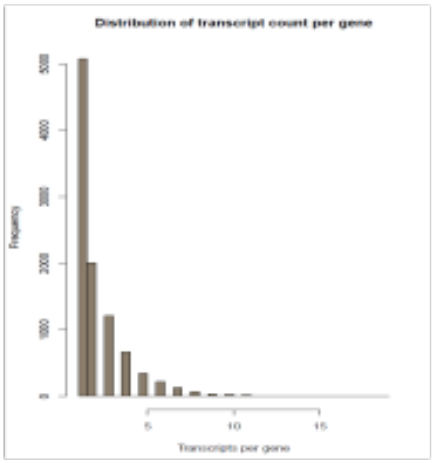
4.1.6 Visualization of DEGs

Finally in figure (d) the Enhanced volcano representing the visualization of the DEGs through their p value and log2 fold change has been presented. Following scatter of the dots represented by the genes is almost equal in both the right and left side of the plot. Some of the gene that are up regulated include DPM3, MYADM, and FOS etc. These genes are expressed more in uninvolved cell lines than in the normal ones. The

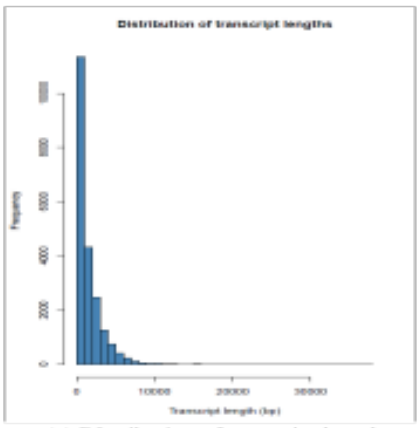
genes shown inside the box around zero fold change are not significant as they have a p value lower than 0.05. The plot represents the significant genes to have a p value lower than 0.05 and a log2 fold change which is above ±0.5. Each gene is represented by a single dot on this plot. The red dotted genes are biologically and statistically significant because its p-value is below 0.05 and the fold change is just above 0.5. The top ten differentially expressed genes found in this data include shown in the table 4.1.



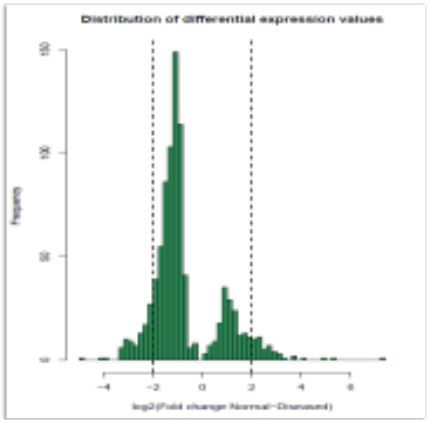
(a) Log2 transformed FPKM



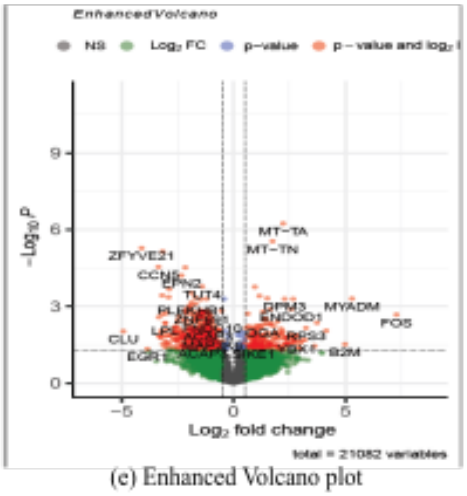
(b) Transcript count per gene



(c) DIstribution of transcript length



(d) Frequency Distribution of log2 fold change



(e) Enhanced Volcano plot

Figure 4.4: Normal vs Uninvolved: (a) The distribution of the FPKMs of every sample (b) Number of transcripts mapped by each gene (c) Transcript length of genes (d) Distribution of gene frequency according to fold change (e) Enhance Volcano plot showing significant DEGs

Gene name	Fold Change	Differential Expression	pval
PTPRU	5.418923218	2.438006205	0.002465315
NDUFS5	9.284802216	3.214871178	0.01554031
LEPROT	5.029774894	2.330493834	0.035249372
S100A11	8.880266354	3.150602949	0.002026757
CRABP2	9.573716679	3.259079112	0.016545027
TAGLN2	7.112173233	2.830290465	0.013768814
IRF2BP2	4.779505962	2.256861501	0.013625134
DEPP1	4.908911086	2.295403036	0.044556667
DDIT4	5.021569211	2.328138268	0.009180036
LDHA	6.26997556	2.648459819	0.017392411

Table 4.4: Top 10 differentially expressed genes Normal vs Uninvolved

4.1.7 Comparative Analysis

The differentially expressed genes of Data 1 and Data 2 were compared and the common genes were identified. Out of 7801 genes of Data 2 and 3741 genes from data 1, about 7 genes were found common. These are shown in the table below:

Gene name	geneIDs	feature
DDIT4	ENSG00000168209.6	transcript
DYRK3	ENSG00000143479.17	transcript
ELF3	ENSG00000163435.16	transcript
H2BC21	ENSG00000184678.10	transcript
JUN	ENSG00000177606.8	transcript
LEPROT	ENSG00000213625.9	transcript
RPL21P16	ENSG00000220842.6	transcript

Table 4.5: Common DEGs among GSE52194 and E-GEOD-58135

Figure 4.5: Common differentially expressed genes among Data 1 and Data

These genes were further subjected to a pathway analysis using Reactome Pathway Browser. This resulted in several pathways out of which we selected the top 10 pathways with the most significant P values. The pathways were enriched in the genes including JUN, DYRK3, H2BC21, DDIT4, and ELF3. These pathways are mentioned below:

Table 4.6: Top 10 Target Pathways

Pathway identifier	Pathway name	Entities p Value	Submitted entities found
R-HSA-1912408	Pre-NOTCH Transcription and Translation	1.72E-07	JUN;ELF3;H2BC21
R-HSA-1912422	Pre-NOTCH Expression and Processing	4.45E-07	JUN;ELF3;H2BC21
R-HSA-157118	Signaling by NOTCH	4.45E-07	JUN;ELF3;H2BC21
R-HSA-9645723	Diseases of programmed cell death	1.94E-05	JUN;H2BC21
R-HSA-2559582	Senescence-Associated Secretory Phenotype (SASP)	2.00E-05	JUN;H2BC21
R-HSA-2559580	Oxidative Stress Induced Senescence	3.91E-05	JUN;H2BC21
R-HSA-5617472	Activation of anterior HOX genes in hind brain development during early embryogenesis	4.11E-05	JUN;H2BC21
R-HSA-5619507	Activation of HOX genes during differentiation	4.11E-05	JUN;H2BC21
R-HSA-212436	Generic Transcription Pathway	1555	JUN;DYRK3;H2BC21;DDIT4
R-HSA-9018519	Estrogen-dependent gene expression	9.51E-05	JUN;H2BC21

Figure 4.6: Top 10 Target Pathways

Chapter 5

Discussion

In our research we conducted a thorough RNA Sequencing analysis of two datasets including data of TNBC, normal and uninvolved cells, in order to discover differentially expressed genes. We further compared the RNA Seq results of the two datasets and found the common differentially expressed genes among them. Subsequently, we carried out a pathway analysis on the following genes using Reactome tool to identify the target pathways. This resulted in the 7 gene targets and 10 targetable pathways. The comparative analysis resulted in finding seven differentially expressed genes namely; DDIT4, DYRK3, ELF3, H2BC21, JUN, LEPROT, and RPL21P16. DNA-damage-inducible transcript, (DDIT4) also called REDD1 (regulated in development and DNA damage responses 1) is activated in stress situations through various stressors such as hypoxia, methyl methane sulfonate, ionizing radiation (IR), energy depletion and heat shock.. According to studies, DDIT4 is involved in the inhibition of mammalian target of rapamycin (mTOR) and is normally downregulated in human cells. DDIT4 over expression can lead to several diseases such as neurodegenerative disorders, preeclampsia, diabetes, and ischemic proliferative retinopathy, and cancer. DDIT4 expression is shown to be usually upregulated in aggressive tumors such as TNBC as they consist of hypoxia, this can inhibit the mTOR pathway and cause growth of cancer stem cells, increasing tumorigenesis. [66]

The next gene target found is DYRK3 (Dual Specificity Tyrosine Phosphorylation Regulated Kinase 3). These genes play important functions in cell growth, development, and survival. [67] DYRK3 is often found to be active in erythroid cells, adult kidney, testis, and liver. The mutation of this gene has been found to be a causative agent

in many diseases such as lung cancer, neuroblastoma, and gastric cancer. However, no proper description of this gene has been found to be linked to Triple negative breast cancer, hence, this gene can be considered as our unique discovery. [68]

Another gene target found is ELF3 (Transcription factor E74-like factor 3). It is a part of the Ets transcription factor family and is present on chromosome 1q32.1 and forms a 371 aa protein. It is found in the large and small intestine, esophagus, the stomach, salivary gland, liver, bladder and prostate. This gene's expression is usually inactivated or amplified in cancers and may causes several types of cancers such as those of biliary tract, colorectal, stomach, gastric, breast, liver and colon.[69]

In a research the effect of ELF3 expression on TNBC cell lines was studied using wet lab techniques such as western blot, soft agar colony formation, and immunofluorescence assays. The results showed that although ELF3 over expression is found in lethal and aggressive TNBC, however, according to the study, it showed to control the metastatic ability of the TNBC cancer[70]

H2BC21 is a gene which expresses for the protein Histone H2B type 2-E which is an essential part of a nucleosome and is required for managing the structure of DNA. Histone proteins are generally required for transcription regulation, DNA repair, chromosomal stability and replication of DNA. Although a similar histone producing gene known as H2BC20P has shown to cause resistance in TNBC, however, not much data is available on the effect of H2BC21 on TNBC, hence this is another novel gene in our results.[71]

LEPROT gene, also known as LEPR, is associated to several types of cancers such as breast, colon, gastric and esophageal cancer. [72] LEPROT is known to be a gene effecting obesity, which is suspected to cause resistance to chemotherapy in TNBC patients. [73] However, no proper treatment has been developed against this gene target yet. RPL21P16 gene is known to be associated with breast cancer, however, it does not have any link to TNBC while other members of the ribosome protein family still have some link to the disease. Hence further studies need to be done on this gene's effect on TNBC. [74]

In the top ten up-regulated pathways, three of the pathways were linked to the Notch receptor. The Notch pathways help in signaling for transcription and translation in the cell. Notch signaling has shown to be oncogenic and promotes cell proliferation in different cancers. It does this by activating tumor promoting genes and anti-apoptotic

pathways. It also promotes growth of stem cells. [75] The Notch signalling pathways are activated in breast, non-small lung, and blood cancers, while they are inactive in small lung cancer and squamous cell cancers. [76] In TNBC Notch signalling pathway causes the aggressive phenotype of the cancer. It also promotes growth of stem cells and cancer inducing cells. [77]

Another upregulated pathway is of the programmed cell death. The disruption of this pathway is responsible for diseases such as amyotrophic lateral sclerosis Parkinson's disease, and other neural diseases. It also causes cancer and autoimmunity. According to researches, programmed cell death can be targeted to treat TNBC, however, further research needs to be done in order to discover potential drugs to be used against it. [78]

Senescence-Associated Secretory Phenotype (SASP) and Oxidative Stress Induced Senescence pathway promote cancer growth through cell proliferation. For TNBC a common senolytic drug often used is GX15-070 . Some drugs such as ABT-737, BET inhibitors are being used to target these pathways, however, the off target effects and side effects of these drugs still need to be studied in detail. [78] Such senescence pathways can develop resistance in cancer.

Activation of anterior HOX genes in hindbrain development during early embryogenesis is not known to be linked to TNBC and hence may be one of the novel target pathways detected in our research. They have been found in other types of cancers such as colorectal. [79]

Generic Transcription Pathway is shown to have some effect on TNBC, however proper research still needs to be carried out. The Estrogen-dependent gene expression is found to have effect on TNBC as well and has been controlled through certain drugs. [80]

Chapter 6

Conclusion

In this research, we determined the potential therapeutic targets genes and pathways among two data sets containing a variety of tissue types. We performed RNA sequencing to determine the differentially expressed genes and did a comparative analysis of the results of both data sets to obtain seven common differentially expressed genes. We then carried out a pathway analysis of these resultant genes using Reactome and obtained ten up regulated pathways. The seven target genes obtained were DDIT4, DYRK3, ELF3, H2BC21, JUN, LEPROT, and RPL21P16. Out of these genes five genes are known to have some link to TNBC according to literature, however, genes DYRK3 and RPL21P16 do not have any literature regarding their effect or connection to TNBC, and hence are a unique finding of our research. Among the ten identified target pathways, the notch signalling pathways were discovered. The Notch signalling pathways are previously known to be linked to TNBC through literature available online. Hence, this research further adds to the confirmation of its connection to TNBC. One pathway known as Activation of anterior HOX genes in hind brain development during early embryo-genesis has no previous literature showing its connection to TNBC.Other pathways such as Generic Transcription Pathway, Estrogen-dependent gene expression pathway, Senescence-Associated Secretory Phenotype (SASP), Oxidative Stress Induced Senescence pathway and programmed cell death have literature where they have been studied to determine their effect on TNBC. Therefore, its is a novel pathway in our research. The unique genes and pathway can be further studied to determine therapeutic targets against them in order to treat TNBC.

In the future, we would like to confirm our results by performing wet lab analysis on

similar type of data. We will also perform further drug analysis to determine good therapeutic drugs for these targets, especially the newly discovered ones. We would also study the pathways in more detail to determine how they link with each other and other diseases besides TNBC, so that we can determine common drugs that can be used for TNBC as well.

Chapter 7

Appendix

7.1 Source code for Analysis of DEGs using Ballgown

if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")

BiocManager::install("ballgown")

install.packages("plyr")

library(ballgown)

library(genefilter)

library(plyr)

library(devtools)

library(ballgown)

setwd("F:/tpn data 2/uinvolved")

dwe<- "F:/tpn data 2/uinvolved"

pheno$_{d}ata = read.csv(file = "table.csv", header = TRUE, sep = ",")$

pheno = read.csv("pheno.csv")

pheno

sort1.pheno <-pheno[order(pheno$Sample),]

bg = ballgown(dataDir= dwe, samplePattern='Sample', meas='all', pData=sort1.pheno)

bg

save(bg, file='E:/data/bg.rda')

structure(bg)$exon$

bg

$\text{bg}_filt = subset(bg, "rowVars(texpr(bg)) > 1", genomesubset = TRUE)$

$\text{bg}_table = texpr(bg_filt, ' all')$

$\text{bg}_gene_names = unique(bg_table[, 9 : 10])$

$\text{transcript}_expression = as.data.frame(texpr(bg_filt))$

head($\text{transcript}_expression$)

row.names($\text{transcript}_expression$)

$\text{results}_transcripts = stattest(bg_filt, feature = "transcript", covariate = "Phenotype", adjustvars = NULL, getFC = TRUE, meas = "FPKM")$

$\text{results}_genes = stattest(bg_filt, feature = "gene", covariate = "Phenotype", adjustvars = NULL, getFC = TRUE, meas = "FPKM")$

files

$\text{write.csv}(\text{bg}_gene_names, "E : /data/bg_gene_names.csv")$

$\text{write.csv}(\text{bg}_table, "E : /data/bg_table.csv")$

$\text{write.csv}(\text{results}_genes, "E : /data/results_genes.csv")$

$\text{write.csv}(\text{results}_transcripts, "E : /data/transcripts.csv")$

transcripts analysis

$\text{results}_transcriptsmer =$

$\text{data.frame}(\text{geneNames=ballgown::geneNames}(\text{bg}_filt), geneIDs = ballgown :: geneIDs(bg_filt), transc$
$ballgown :: transcriptNames(bg_filt), \text{results}_transcripts)$

$\text{results}_transcripts = arrange(\text{results}_transcripts, pval)$

$\text{results}_genes = arrange(\text{results}_genes, pval)$

$\text{write.csv}(\text{results}_transcripts, "E : /data/chrX_transcript_results.csv", row.names = FALSE)$

$\text{write.csv}(\text{results}_genes, "E : /data/chrX_gene_results.csv", row.names = FALSE)$

tra <- $\text{subset}(\text{results}_transcripts, \text{results}_transcripts\text{pval}<0.05)$

gen <- $\text{subset}(\text{results}_genes, \text{results}_genes\text{pval}<0.05)$

write.csv(tra, "E:/data/filtered transcripts.csv")

write.csv(gen, "E:/data/filtered genes.csv")

Box plot tropical= c('darkorange', 'dodgerblue', 'hotpink', 'limegreen', 'yellow')

palette(tropical)

fpkm = texpr(bg$_f ilt, meas = "FPKM")$

fpkm = log2(fpkm+1)

boxplot (fpkm,col=as.numeric(pheno$phenotype), las = 2, ylab =' log2(FPKM + 1)')$

pdf("boxplot.pdf", height=8, width=8)

boxplot(mtcars$mpg)$

dev.off()

transcript$_g ene_t able = indexes(bg_f ilt)$t2g

head(transcript$_g ene_t able)$

Each row of data represents a transcript. Many of these transcripts represent the same gene. Determine the numbers of transcripts and unique genes

length(row.names(transcript$_g ene_t able))$

length(unique(transcript$_g ene_t able[, "g_i d"]))$

counts=table(transcript$_g ene_t able[, "g_i d"])$

c$_o ne = length(which(counts == 1))$

c$_m ore_t han_o ne = length(which(counts > 1))$

c$_m ax = max(counts)$

hist(counts, breaks=50, col="bisque4", xlab="Transcripts per gene", main="Distribution of transcript count per gene")

legend$_t ext = c(paste("Geneswithonetranscript = ", c_o ne), paste("Geneswithmorethanonetranscript = ", c_m ore_t han_o ne), paste("Maxtranscriptsforsinglegene = ", c_m ax))$

legend("topright", legend$_t ext, lty = NULL)$

Plot 2 - the distribution of transcript sizes as a histogram full$_t able <- -texpr(bg_f ilt,' all')$

hist(full$_t able$length, breaks=50, xlab="Transcript length (bp)", main="Distribution of transcript lengths", col="steelblue")

$\text{data}_colors = (c(\text{"white"}, \text{"blue"}, \text{"007FFF"}, \text{"cyan"}, \text{"7FFF7F"}, \text{"yellow"}, \text{"FF7F00"}, \text{"red"}, \text{"7F0000"}$

$\min_nonzero = 1$

Set the columns for finding FPKM and create shorter names for figures

$\text{data}_columns = c(1:18)$

$\text{short}_names = c(\text{"S1"}, \text{"S2"}, \text{"S3"}, \text{"S4"}, \text{"S5"}, \text{"S6"}, \text{"S7"}, \text{"S8"}, \text{"S9"}, \text{"S10"}, \text{"S11"}, \text{"S12"}, \text{"S13"}, \text{"S14}$

Plot 3 - View the range of values and general distribution of FPKM values for all libraries
Create boxplots for this purpose Display on a log2 scale and add the minimum non-zero value to avoid log2(0)

$\text{boxplot(log2(transcript}_expression[, data_columns]+min_nonzero), col = data_colors, names = short_names, las = 2, ylab = \text{"log2}(FPKM)\text{"}, main = \text{"Distributionof}FPKMs\text{"})$

colors = colorRampPalette(c("blue", "blue", "007FFF", "cyan","7FFF7F", "yellow", "FF7F00", "red", "7F0000")) $\text{smoothScatter(x=log2(x+min}_nonzero), xlab = \text{"}FPKM(SRR218_N, Replicate1)\text{"}, yl$ $\text{"}FPKM(SRR219_N, Replicate2)\text{"}, main = \text{"Comparisonofexpressionvaluesforapairofreplicates"}, c$ $colors, nbin = 200)$

Compare the correlation 'distance' between all replicates

$\text{transcript}_expression[, \text{"sum"}] = apply(transcript_expression[, data_columns], 1, sum)$

Identify the genes with a grand sum FPKM of at least 5 - we will filter out the genes with very low expression across the board $\text{i = which(transcript}_expression[, \text{"sum"}] > 5)Calculatethecorrelationbetweenallpairsofdatar = cor(transcript_expression[i, data_columns], use = \text{"pairwise.complete.obs"}, method = \text{"pearson"})$

r

Plot 8 - Convert correlation to 'distance', and use 'multi-dimensional scaling' to display the relative differences between libraries

d=1-r

$\text{data}_columns = c(1:32)$

mds=cmdscale(d, k=2, eig=TRUE)

par(mfrow=c(1,1))

$\text{plot(mds}points, type = \text{"n"}, xlab = \text{""}, ylab = \text{""}, main = \text{"}MDSdistanceplot\text{"}, xlim = c(-0.25, 0.25), ylim = c(-0.25, 0.25))$

points(mds$points[,1], mds$points[,2], col="grey", cex=2, pch=16) text(mds$points[,1], mds$points[,2], short_names, col = data_colors)

sig=which(results_transcriptsmer$pval<0.05)

results_transcriptsmer[,"de"] =

log2(results_transcriptsmer[,"fc"])

write.csv(results_transcriptsmer,"E:/data/fil3de.csv")

hist(results_transcriptsmer[sig,"de"], breaks = 50, col = "seagreen", xlab = "log2(FoldchangeNormal Diseased)", main = "Distributionofdifferentialexpressionvalues")

abline(v=-2, col="black", lwd=2, lty=2)

abline(v=2, col="black", lwd=2, lty=2)

legend("topright", "Fold-change > 4", lwd=2, lty=2)

enhanced volcano

BiocManager::install("EnhancedVolcano")

library(EnhancedVolcano)

EnhancedVolcano(results_transcriptsmer, lab = results_transcriptsmer$geneNames, x = "de", y = "pval", pCutoff = 0.05, FCcutoff = 0.5, title = "RNA-dataset")

trans <- subset(results_transcriptsmer, pval < 0.05)

trans <- subset(results_transcriptsmer, fc < −0.5|fc > 0.5)

trans <- subset(results_transcriptsmer, pval < 0.05)

trans <- subset(trans, de < -0.5 | de > 0.5)

write.csv(trans,"E:/data/results_transcriptsmertwithcutoffvolcano.csv")

volcano BiocManager::install("Volcanoplot")

library(Volcanoplot)

volcano_names <− ifelse((tra$coefficients)>=1, tra$id, NA)

volcanoplot(tra, coef = 1L, style = "p-value", highlight = 100, names = volcano_names, xlab = "Log2FoldChange", ylab = NULL, pch = 16, cex = 0.35)

References

[1] Gautam K Malhotra, Xiangshan Zhao, Hamid Band, and Vimla Band. Histological, molecular and functional subtypes of breast cancers. *Cancer biology & therapy*, 10 (10):955–960, 2010.

[2] Freddie Bray, Ahmedin Jemal, Nathan Grey, Jacques Ferlay, and David Forman. Global cancer transitions according to the human development index (2008–2030): a population-based study. *The lancet oncology*, 13(8):790–801, 2012.

[3] Robert A Weinberg. How cancer arises. *Scientific American*, 275(3):62–70, 1996.

[4] Elizabeth Meade and Maura Dowling. Early breast cancer: diagnosis, treatment and survivorship. *British Journal of Nursing*, 21(Sup17):S4–S7, 2012.

[5] Saba Sohail and Shams Nadeem Alam. Breast cancer in pakistan-awareness and early detection. 2007.

[6] Rashid Menhas and Shumaila Umer. Breast cancer among pakistani women. *Iranian journal of public health*, pages 586–7, 2015.

[7] Muhammad U Rashid, Sidra Gull, Kashif Asghar, Noor Muhammad, Asim Amin, and Ute Hamann. Prevalence of tp53 germ line mutations in young pakistani breast cancer patients. *Familial cancer*, 11(2):307–311, 2012.

[8] Farhana Badar, Imran Moid, Farah Waheed, Anbreen Zaidi, Bilal Naqvi, and Shakeeb Yunus. Variables associated with recurrence in breast cancer patients-the shaukat khanum memorial experience. *Asian Pac J Cancer Prev*, 6:54–7, 2005.

[9] Giuseppe Viale. The current state of breast cancer classification. *Annals of Oncology*, 23:x207–x210, 2012.

[10] Andrew M Hanby et al. Tavassoli fa, devilee p: Pathology and genetics: Tumours of the breast and female genital organs. who classification of tumours series-volume iv. lyon, france: Iarc press. *Breast Cancer Research*, 6(3):1–2, 2004.

[11] Charles M Perou, Therese Sørlie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *nature*, 406 (6797):747–752, 2000.

[12] Darrel W Cleere. Triple-negative breast cancer: a clinical update. *Community Oncology*, 5(7):203–211, 2010.

[13] Steven S Coughlin and Donatus U Ekwueme. Breast cancer as a global health concern. *Cancer epidemiology*, 33(5):315–318, 2009.

[14] William D Foulkes, Ingunn M Stefansson, Pierre O Chappuis, Louis R Bégin, John R Goffin, Nora Wong, Michel Trudel, and Lars A Akslen. Germline brca1 mutations and a basal epithelial phenotype in breast cancer. *JNCI: Journal of the National Cancer Institute*, 95(19):1482–1485, 2003.

[15] F Heitz, P Harter, A Traut, HJ Lueck, B Beutel, and A du Bois. Cerebral metastases (cm) in breast cancer (bc) with focus on triple-negative tumors. *Journal of Clinical Oncology*, 26(15_suppl):1010–1010, 2008.

[16] Carey Anders and Lisa A Carey. Understanding and treating triple-negative breast cancer. *Oncology (Williston Park, NY)*, 22(11):1233, 2008.

[17] Lisa A Carey, Charles M Perou, Chad A Livasy, Lynn G Dressler, David Cowan, Kathleen Conway, Gamze Karaca, Melissa A Troester, Chiu Kit Tse, Sharon Edmiston, et al. Race, breast cancer subtypes, and survival in the carolina breast cancer study. *Jama*, 295(21):2492–2502, 2006.

[18] Jessica M Dolle, Janet R Daling, Emily White, Louise A Brinton, David R Doody, Peggy L Porter, and Kathleen E Malone. Risk factors for triple-negative breast cancer in women under the age of 45 years. *Cancer Epidemiology and Prevention Biomarkers*, 18(4):1157–1166, 2009.

[19] Graham A Colditz, Bernard A Rosner, Wendy Y Chen, Michelle D Holmes, and Susan E Hankinson. Risk factors for breast cancer according to estrogen and pro-

gesterone receptor status. *Journal of the national cancer institute*, 96(3):218–228, 2004.

[20] S Eva Singletary. Rating the risk factors for breast cancer. *Annals of surgery*, 237 (4):474, 2003.

[21] Anthony Howell, Annie S Anderson, Robert B Clarke, Stephen W Duffy, D Gareth Evans, Montserat Garcia-Closas, Andy J Gescher, Timothy J Key, John M Saxton, and Michelle N Harvie. Risk determination and prevention of breast cancer. *Breast Cancer Research*, 16(5):1–19, 2014.

[22] Steinar Tretli. Height and weight in relation to breast cancer morbidity and mortality. a prospective study of 570,000 women in norway. *International Journal of Cancer*, 44(1):23–30, 1989.

[23] V Beral. Breast cancer and hormone replacement therapy (hrt): collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. In *Breast Cancer Research*, volume 4, pages 1–1. Springer, 2002.

[24] Sharon Nofech-Mozes, Maureen Trudeau, Harriet K Kahn, Rebecca Dent, Ellen Rawlinson, Ping Sun, Steven A Narod, and Wedad M Hanna. Patterns of recurrence in the basal and non-basal subtypes of triple-negative breast cancers. *Breast cancer research and treatment*, 118(1):131–137, 2009.

[25] Roohi Ismail-Khan and Marilyn M Bui. A review of triple-negative breast cancer. *Cancer control*, 17(3):173–176, 2010.

[26] Hui Yao, Guangchun He, Shichao Yan, Chao Chen, Liujiang Song, Thomas J Rosol, and Xiyun Deng. Triple-negative breast cancer: is there a treatment on the horizon? *Oncotarget*, 8(1):1913, 2017.

[27] Miguel Martín, Álvaro Rodríguez-Lescure, Amparo Ruiz, Emilio Alba, Lourdes Calvo, Manuel Ruiz-Borrego, Ana Santaballa, César A Rodríguez, Carmen Crespo, Mar Abad, et al. Molecular predictors of efficacy of adjuvant weekly paclitaxel in early breast cancer. *Breast cancer research and treatment*, 123(1):149–157, 2010.

[28] Xavier B Pivot, Rubi K Li, Eva S Thomas, Hyun-Cheol Chung, Luis E Fein, Valorie F Chan, Jacek Jassem, Fernando Hurtado de Mendoza, Pralay Mukhopadyay,

and Henri H Roché. Activity of ixabepilone in oestrogen receptor-negative and oestrogen receptor-progesterone receptor-human epidermal growth factor receptor 2-negative metastatic breast cancer. *European Journal of Cancer*, 45(17):2940–2946, 2009.

[29] Cornelia Liedtke, Chafika Mazouni, Kenneth R Hess, Fabrice André, Attila Tordai, Jaime A Mejia, W Fraser Symmans, Ana M Gonzalez-Angulo, Bryan Hennessy, Marjorie Green, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *Journal of clinical oncology*, 26(8): 1275–1281, 2008.

[30] L Staudacher, PH Cottu, V Dieras, A Vincent-Salomon, MN Guilhaume, L Escalup, T Dorval, P Beuzeboc, L Mignot, and JY Pierga. Platinum-based chemotherapy in metastatic triple-negative breast cancer: the institut curie experience. *Annals of oncology*, 22(4):848–856, 2011.

[31] Françoise Dantzer, Guadelupe de la Rubia, Josiane Ménissier-de Murcia, Zdenek Hostomsky, Gilbert de Murcia, and Valérie Schreiber. Base excision repair is impaired in mammalian cells lacking poly (adp-ribose) polymerase-1. *Biochemistry*, 39(25):7559–7569, 2000.

[32] Kedar Hastak, Elizabeth Alli, and James M Ford. Synergistic chemosensitivity of triple-negative breast cancer cell lines to poly (adp-ribose) polymerase inhibition, gemcitabine, and cisplatin. *Cancer research*, 70(20):7970–7980, 2010.

[33] BK Linderholm, H Hellborg, U Johansson, G Elmberger, L Skoog, J Lehtiö, and R Lewensohn. Significantly higher levels of vascular endothelial growth factor (vegf) and shorter survival times for patients with primary operable triple-negative breast cancer. *Annals of oncology*, 20(10):1639–1646, 2009.

[34] DW Miles, V Diéras, J Cortés, A-A Duenne, J Yi, and J O'shaughnessy. First-line bevacizumab in combination with chemotherapy for her2-negative metastatic breast cancer: pooled and subgroup analyses of data from 2447 patients. *Annals of oncology*, 24(11):2773–2780, 2013.

[35] Giuseppe Viale, Nicole Rotmensz, Patrick Maisonneuve, Luca Bottiglieri, Emilia Montagna, Alberto Luini, Paolo Veronesi, Mattia Intra, Rosalba Torrisi, Anna

Cardillo, et al. Invasive ductal carcinoma of the breast with the "triple-negative" phenotype: prognostic implications of egfr immunoreactivity. *Breast cancer research and treatment*, 116(2):317–328, 2009.

[36] Roy S Herbst and Dong M Shin. Monoclonal antibodies to target epidermal growth factor receptor–positive tumors: a new paradigm for cancer therapy. *Cancer*, 94 (5):1593–1611, 2002.

[37] Budhi Singh Yadav, Suresh C Sharma, Priyanka Chanana, and Swaty Jhamb. Systemic treatment strategies for triple-negative breast cancer. *World journal of clinical oncology*, 5(2):125, 2014.

[38] Ulises D Orlando, Ana F Castillo, Melina A Dattilo, Angela R Solano, Paula M Maloberti, and Ernesto J Podesta. Acyl-coa synthetase-4, a new regulator of mtor and a potential therapeutic target for enhanced estrogen receptor function in receptor-positive and-negative breast cancer. *Oncotarget*, 6(40):42632, 2015.

[39] Tristan M Sissung, Caitlin E Baum, C Tyler Kirkland, Rui Gao, Erin R Gardner, and William D Figg. Pharmacogenetics of membrane transporters: an update on current approaches. *Molecular biotechnology*, 44(2):152–167, 2010.

[40] Akimitsu Yamada, Takashi Ishikawa, Ikuko Ota, Mariko Kimura, Daisuke Shimizu, Mikiko Tanabe, Takashi Chishima, Takeshi Sasaki, Yasushi Ichikawa, Satoshi Morita, et al. High expression of atp-binding cassette transporter abcc11 in breast tumors is associated with aggressive subtypes and low disease-free survival. *Breast cancer research and treatment*, 137(3):773–782, 2013.

[41] Chad J Creighton, Xiaoxian Li, Melissa Landis, J Michael Dixon, Veronique M Neumeister, Ashley Sjolund, David L Rimm, Helen Wong, Angel Rodriguez, Jason I Herschkowitz, et al. Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proceedings of the National Academy of Sciences*, 106(33):13820–13825, 2009.

[42] Cindy Neuzillet, Annemilaï Tijeras-Raballand, Romain Cohen, Jérôme Cros, Sandrine Faivre, Eric Raymond, and Armand de Gramont. Targeting the tgfβ pathway for cancer therapy. *Pharmacology & therapeutics*, 147:22–31, 2015.

[43] Jon C Aster, Warren S Pear, and Stephen C Blacklow. The varied roles of notch in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 12:245–275, 2017.

[44] Leo E Gerweck, Shashirekha Vijayappa, and Sergey Kozin. Tumor ph controls the in vivo efficacy of weak acid and base chemotherapeutics. *Molecular cancer therapeutics*, 5(5):1275–1279, 2006.

[45] Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, 14(2):482–517, 2019.

[46] Tomas G Lyons. Targeted therapies for triple-negative breast cancer. *Current treatment options in oncology*, 20(11):1–13, 2019.

[47] S Adams, P Schmid, HS Rugo, EP Winer, D Loirat, A Awada, DW Cescon, H Iwata, M Campone, R Nanda, et al. Pembrolizumab monotherapy for previously treated metastatic triple-negative breast cancer: cohort a of the phase ii keynote-086 study. *Annals of Oncology*, 30(3):397–404, 2019.

[48] Peter Schmid, Sylvia Adams, Hope S Rugo, Andreas Schneeweiss, Carlos H Barrios, Hiroji Iwata, Véronique Diéras, Roberto Hegg, Seock-Ah Im, Gail Shaw Wright, et al. Atezolizumab and nab-paclitaxel in advanced triple-negative breast cancer. *New England Journal of Medicine*, 379(22):2108–2121, 2018.

[49] Hans Petter Eikesdal, Synnøve Yndestad, Asmaa Elzawahry, Alba Llop-Guevara, Bjørnar Gilje, Egil Støre Blix, Helge Espelid, Steinar Lundgren, Jürgen Geisler, Geirfinn Vagstad, et al. Olaparib monotherapy as primary treatment in unselected triple negative breast cancer. *Annals of Oncology*, 32(2):240–249, 2021.

[50] Gamze Guney Eskiler, Gulsah Cecener, Unal Egeli, and Berrin Tunca. Bmn 673 (talazoparib): A potent parp inhibitor for triple negative breast cancer with different genetic profile. *Journal of biochemical and molecular toxicology*, 33(5):e22286, 2019.

[51] Aditya Bardia, Sara A Hurvitz, Sara M Tolaney, Delphine Loirat, Kevin Punie, Mafalda Oliveira, Adam Brufsky, Sagar D Sardesai, Kevin Kalinsky, Amelia B

Zelnak, et al. Sacituzumab govitecan in metastatic triple-negative breast cancer. *New England Journal of Medicine*, 384(16):1529–1541, 2021.

[52] Hyo S Han, Carlos A Alemany, Ursa Abigail Brown-Glaberman, Timothy J Pluard, Rajni Sinha, Danielle Sterrenberg, Kathy S Albain, Reva K Basho, David Biggs, Valentina Boni, et al. Sgnlva-002: Single-arm, open label phase ib/ii study of ladiratuzumab vedotin (lv) in combination with pembrolizumab for first-line treatment of patients with unresectable locally advanced or metastatic triple-negative breast cancer., 2019.

[53] Zheng Dan, He Xiujing, Luo Ting, Zhong Xiaorong, Zheng Hong, Yang Jiqiao, Li Yanchu, and Jing Jing. Long non-coding rna btg3-7: 1 and jund co-regulate c21orf91 to promote triple-negative breast cancer progress. *Frontiers in Molecular Biosciences*, 7:380, 2021.

[54] Vitaly Chasov, Regina Mirgayazova, Ekaterina Zmievskaya, Raniya Khadiullina, Aygul Valiullina, Joseph Stephenson Clarke, Albert Rizvanov, Matthias GJ Baud, and Emil Bulatov. Key players in the mutant p53 team: small molecules, gene editing, immunotherapy. *Frontiers in Oncology*, 10, 2020.

[55] Fei Chen, Yuancheng Li, Na Qin, Fengliang Wang, Jiangbo Du, Cheng Wang, Fangzhi Du, Tao Jiang, Yue Jiang, Juncheng Dai, et al. Rna-seq analysis identified hormone-related genes associated with prognosis of triple negative breast cancer. *Journal of biomedical research*, 34(2):129, 2020.

[56] Anjana Bhardwaj, Harpreet Singh, Kimal Rajapakshe, Kazunoshin Tachibana, Nivetha Ganesan, Yinghong Pan, Preethi H Gunaratne, Cristian Coarfa, and Isabelle Bedrosian. Regulation of mirna-29c and its downstream pathways in preneoplastic progression of triple-negative breast cancer. *Oncotarget*, 8(12):19645, 2017.

[57] Noura Khaled, Nicolas Sonnier, Ioana Molnar, Flora Ponelle-Chachuat, Myriam Kossai, Nina Radosevic-Robin, Maud Privat, and Yannick Bidet. Rna sequencing reveals the differential expression profiles of rna in metastatic triple negative breast cancer and identifies shisa3 as an efficient tumor suppressor gene. *American Journal of Cancer Research*, 11(9):4568, 2021.

[58] Charissa Kim, Ruli Gao, Emi Sei, Rachel Brandt, Johan Hartman, Thomas Hatschek, Nicola Crosetto, Theodoros Foukakis, and Nicholas E Navin. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell*, 173(4):879–893, 2018.

[59] Anelia Horvath, Suresh Babu Pakala, Prakriti Mudvari, Sirigiri Divijendra Natha Reddy, Kazufumi Ohshiro, Sandra Casimiro, Ricardo Pires, Suzanne AW Fuqua, Masakazu Toi, Luis Costa, et al. Novel insights into breast cancer genetic variance through rna sequencing. *Scientific reports*, 3(1):1–13, 2013.

[60] Elena Alexandrova, Jessica Lamberti, Pasquale Saggese, Giovanni Pecoraro, Domenico Memoli, Valeria Mirici Cappa, Maria Ravo, Roberta Iorio, Roberta Tarallo, Francesca Rizzo, et al. Small non-coding rna profiling identifies mir-181a-5p as a mediator of estrogen receptor beta-induced inhibition of cholesterol biosynthesis in triple-negative breast cancer. *Cells*, 9(4):874, 2020.

[61] Safa Shaheen, Febin Fawaz, Shaheen Shah, and Dietrich Büsselberg. Differential expression and pathway analysis in drug-resistant triple-negative breast cancer cell lines using rnaseq analysis. *International journal of molecular sciences*, 19(6):1810, 2018.

[62] Valeria Ossovskaya, Yipeng Wang, Adam Budoff, Qiang Xu, Alexander Lituev, Olga Potapova, Gordon Vansant, Joseph Monforte, and Nikolai Daraselia. Exploring molecular pathways of triple-negative breast cancer. *Genes & cancer*, 2(9): 870–879, 2011.

[63] Shavira Narrandes, Shujun Huang, Leigh Murphy, and Wayne Xu. The exploration of contrasting pathways in triple negative breast cancer (tnbc). *BMC cancer*, 18 (1):1–8, 2018.

[64] Pengzhi Dong, Bing Yu, Lanlan Pan, Xiaoxuan Tian, and Fangfang Liu. Identification of key genes and pathways in triple-negative breast cancer by integrated bioinformatics analysis. *BioMed research international*, 2018, 2018.

[65] Maria Karmella Apaya, Jeng-Yuan Shiau, Guo-Shiou Liao, Yu-Jen Liang, Chia-Wei Chen, Hsin-Chou Yang, Chi-Hong Chu, Jyh-Cherng Yu, and Lie-Fen Shyur. Integrated omics-based pathway analyses uncover cyp epoxygenase-associated net-

works as theranostic targets for metastatic triple negative breast cancer. *Journal of Experimental & Clinical Cancer Research*, 38(1):187, 2019.

[66] Fahimeh Fattahi, Leili Saeednejad Zanjani, Zohreh Habibi Shams, Jafar Kiani, Mitra Mehrazma, Mohammad Najafi, and Zahra Madjd. High expression of dna damage-inducible transcript 4 (ddit4) is associated with advanced pathological features in the patients with colorectal cancer. *Scientific Reports*, 11(1):1–17, 2021.

[67] Diya Zhang, Ke Li, Connie L Erickson-Miller, Mitch Weiss, and Don M Wojchowski. Dyrk gene structure and erythroid-restricted features of dyrk3 gene expression. *Genomics*, 85(1):117–130, 2005.

[68] Ekaterina Ivanova, Shrey Dharamvir Sharma, Anna Brichkina, Petra Pfefferle, Ursula Keber, Axel Pagenstecher, and Matthias Lauth. Dyrk3 contributes to differentiation and hypoxic control in neuroblastoma. *Biochemical and Biophysical Research Communications*, 567:215–221, 2021.

[69] Ian Y Luk, Camilla M Reehorst, and John M Mariadason. Elf3, elf5, ehf and spdef transcription factors in tissue homeostasis and cancer. *Molecules*, 23(9):2191, 2018.

[70] S Park and Y Kwon. The role of elf3 in metastasis of triple negative breast cancer. *Annals of Oncology*, 28:x34, 2017.

[71] Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395, 2011.

[72] He Huang, Jun Zhang, Fei Ling, Yuhong Huang, Min Yang, Yao Zhang, Yuanyi Wei, Qingqing Zhang, Honghai Wang, Lin Song, et al. Leptin receptor (lepr) promotes proliferation, migration, and invasion and inhibits apoptosis in hepatocellular carcinoma by regulating anxa7. *Cancer Cell International*, 21(1):1–9, 2021.

[73] Laura W Bowers, Joseph Gung, Claire G Lineberger, and Stephen D Hursting. Abstract sy28-04: Obesity-associated leptin signaling promotes chemotherapy resistance in triple-negative breast cancer: The role of tumor-initiating cell enrichment, 2019.

[74] Ziyue Lin, Rui Peng, Yan Sun, Luyu Zhang, and Zheng Zhang. Identification of ribosomal protein family in triple-negative breast cancer by bioinformatics analysis. *Bioscience Reports*, 41(1):BSR20200869, 2021.

[75] Vandana Venkatesh, Raghu Nataraj, Gopenath S Thangaraj, Murugesan Karthikeyan, Ashok Gnanasekaran, Shanmukhappa B Kaginelli, Gobianand Kuppanna, Chandrashekrappa Gowdru Kallappa, and Kanthesh M Basalingappa. Targeting notch signalling pathway of cancer stem cells. *Stem cell investigation*, 5, 2018.

[76] Claudio Talora, Antonio F Campese, Diana Bellavia, Maria Pia Felli, Alessandra Vacca, Alberto Gulino, and Isabella Screpanti. Notch signaling and diseases: an evolutionary journey from a simple beginning to complex outcomes. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1782(9):489–497, 2008.

[77] MV Giuli, E Giuliani, I Screpanti, D Bellavia, and S Checquolo. Notch signaling activation as a hallmark for triple-negative breast cancer subtype. *Journal of oncology*, 2019, 2019.

[78] Akhil Chawla, Anne V Philips, Gheath Alatrash, and Elizabeth Mittendorf. Immune checkpoints: A therapeutic target in triple negative breast cancer. *Oncoimmunology*, 3(4):e28325, 2014.

[79] Tze Mun Loo, Kenichi Miyata, Yoko Tanaka, and Akiko Takahashi. Cellular senescence and senescence-associated secretory phenotype via the cgas-sting signaling pathway in cancer. *Cancer science*, 111(2):304–311, 2020.

[80] Carlotta Penzo, Laura Arnoldo, Silvia Pegoraro, Sara Petrosino, Gloria Ros, Rossella Zanin, Jacek R Wiśniewski, Guidalberto Manfioletti, and Riccardo Sgarra. Hmga1 modulates gene transcription sustaining a tumor signalling pathway acting on the epigenetic status of triple-negative breast cancer cells. *Cancers*, 11(8):1105, 2019.