DISCOVERING COMMON PATHOGENIC PROCESSES BETWEEN COVID-19 AND HFRS BY INTEGRATING RNA-SEQ DIFFERENTIAL EXPRESSION ANALYSIS WITH MACHINE LEARNING

By
Muhammad Ahsan Shoaib
2021-GCUF-01873

Thesis submitted in partial fulfillment of
the requirements for the degree of

BACHELOR OF SCIENCE
IN
BIOINFORMATICS

DEPARTMENT OF BIOINFORMATICS AND BIOTECHNOLOGY GC,
UNIVERSITY FAISALABAD

June 2025

DECLARATION

The work reported in this thesis was carried out by me under the supervision of Dr. Zubair, Department of Bioinformatics and Biotechnology, Government College University Faisalabad, Pakistan.

I hereby declare that the title of thesis "Discovering common pathogenic processes between covid-19 and hfrs by integrating rna-seq differential expression analysis with machine learning" and the contents of thesis are the product of my own research and no part has been copied from any published source (except the references, standard mathematical or genetic models/ equations/ formulas/protocols etc.). I further declare that this work has not been submitted for award of any other degree/ diploma. The University may take action if the information provided is found inaccurate at any stage.

Name: Muhammad Ahsan Shoaib
Registration No: 2021-GCUF-01873

Certificate by the Supervisory Committee

We certify that the contents and form of thesis submitted by Mr. Muhammad Ahsan Shoaib, Registration No. 2021-GCUF-01896 has been found satisfactory and according to prescribed format. We recommend it to be processed for the evaluation by the External Examiner for the award of degree.

Contents

Acknowledgments

Abstract

CHAPTER 1

INTRODUCTION

Triple Negative Breast Cancer lacks the expression (<1%) of the estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) as assessed by immunohistochemistry [47]. Out of all the cancer cases, TNBC is the 15% of all cases, being a lethal disease for women. [96]. TNBC is more frequently occured in younger women specifically more prevalent in black women.[68]. Based on the evidence we have till now, TNBC is merely an umbrella term that covers an extensive collection of Breast Cancer. These types vary a lot in their genetic makeup, how they behave, how they look under a microscope, and clinical differences [13]. According to the ICCR Histological grading of breast cancer, Grade 1 is Well differentiated with total score of 3–5, Grade 2 is Moderately differentiated with total score of 6–7, Grade 3 is Poorly differentiated with total score of 8–9. Score is the percentage of tumor area forming tubules [icc]. Grade 3 is referred to as high grade and Grade 1 is low grade. According to various studies, most of these tumors are high-grade invasive ductal carcinomas [13]. Conventional TNBCs are high-grade carcinomas, characterized by complex genomes, high levels of genetic instability, and a high degree of intertumor and intratumor heterogeneity [21]. The natural history, molecular features and optimal therapy of an other subset of TNBCs is low-grade lesions is highly different from those of high-grade TNBCs [21].

Ovarian cancer, known as silent killer, is challenging to diagnose until it is at advanced stage because of its vague symptoms [76]. Each year, 152,000 women die of ovarian cancer which makes it the eight most common and fifth deadliest cancer in women worldwide. Ovarian cancer has three main types: epithelial (most common), germ cell, and sex-cord-stromal, with the latter two comprising only about 5% of all ovarian cancers. [95]. It has nonspecific symptoms like abdominal bloating, abdominal pain, urinary frequency, early satiety or feeling full, or changes in bowel habits [76].

Most TNBCs fall into the basal-like intrinsic subtype on the PAM50 subtyping assay. TNBCs were once considered to be synonymous with basal-like breast cancer but it is currently accepted that TNBCs display a remarkable diversity at the gene expression level as well [14]. [8] proposed that the risk profiles for basal-like breast

cancer and high-grade serous ovarian cancer are strongly correlated. Other than that, several studies show that TNBC and Ovarian cancer shares common risk factors. The study [67] focuses on finding common somatic mutations between TNBC and Ovarian Cancer. [18] explains the interaction of GATA3 with wild type BRCA1 in TNBC, and GATA3 is also present in substantial amounts in Ovarian Cancer associated with a poor prognosis in HGSOC patients. In Saudi women, constitutional BRCA1 promoter methylation and MGMT promoter methylation have been shown to be associated with an increased risk of ovarian cancer and breast cancer [3]. Overall, these studies show the common genetic risk factors of TNBC and Ovarian Cancer. [17] stated the common therapeutic treatments of both cancers. As Cisplatin plays a completely different but important role in the treatment of both female cancer types. These studies explains the common biomarkers and overlapping biological pathways of TNBC and HGSOC.

As both the cancer types have various similarities, however, there is a still need to explore the common genetic risk factors causing these cancers. The comparative study of correction of both the cancers can assist developing common therapeutic treatments.

The transcriptomic data analysis of TNBC and HGSOC will uncover the more similarities between both cancers at genetic level. The valuable insights from these data analysis will better explain the root causes of TNBC and HGSOC. The same methodology is used to find common genetic risk factors of other diseases [56], however, there is a gap between TNBC and HGSOC. To fill this gap, this study, using microarray and bioinformatics approaches, identifies the common genes and pathways among TNBC and HGSOC. Initially, microarray data analysis was performed to prepare the data followed by the identification of Differentially Expressed Genes (DEGs). The common DEGs were then used for Gene Set Expression Analysis (GSEA) to find common biological pathways. Protein Protein Interaction networks were constructed to find hub genes using different methods. Here, for the first time, we have characterized the biological processes and pathways commonly dysregulated in TNBC and HGSOC.

CHAPTER 2

LITERATURE REVIEW

2.1 Cancer

Unlike other diseases, cancer is not just a single disease; rather, it encompasses a vast group of diseases that are all caused by uncontrolled growth and proliferation of abnormal cells. Several factors causes cancer including genetics, environmental factors, and infections. Cancer is becoming one of the main cause of death in the whole world. One out of every 5 deaths is due to cancer [57]. Several factors help researchers understanding the essence of cancer including oncogenes and tumor suppressor genes. Projects like human genomes project and sequencing technologies disclosed the cancer mysteries at genomic level also explaining the heterogeneity between different cancer types. Other technologies like computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) helped better diagnosis and staging of cancers [72] [83].

Cancer Stages, Grades and Classification

Initially, cancer was classified on the basis of origin of cancer. Later on, it was known that different type of cancers can emerge from the same origin. On the basis of cell types, cancers are classified into five classes: (1) carcinoma begins in epithelial cells (2) sarcoma are derived from mesenchymal cells (3) lymphoma, leukemia and myeloma, originating in hematopoietic or blood-forming cells; (4) germ cell tumors, developing from germ cells; and (5) neuroblastoma, glioma, glioblastoma and others derived from cells of the central and peripheral nervous system and denoted as neuroectodermal tumors because of their beginning in the early embryo. [87].

[88] first discover the multistage behavior of cancer. Staging in cancer refers to the spread of cancer cells. Stage I, stage II, stage III, stage IV and stage IV representing the most advanced stage. Grading in cancer is also an important parameter and is independent to the type and stage of cancer. The grading system is (1) G1 (highly differentiated), (2) G2 (moderately differentiated), (3) G3 (poorly differentiated) and

(4) G4 (undifferentiated)[87].

Hallmarks in Cancer

[26] in 2002 organized the vast catalog of cancer into six fundamental principles and capabilities that are shared by all type of cancers. These six principles were the cancer hallmarks. The six biological capabilities were: self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. Each of these six capabilities are acquired by cell during cancer development.

In 2011, these hallmarks were updated with the addition of two more hallmarks and two emerging hallmakrs, respectively: genome instability and mutation, tumor-promoting inflammation, deregulated cellular energetics, and evading immune destruction [27].

In the latest 2022 edition, four more cancer hallmarks were suggested: epigenetic reprogramming, phenotypic plasticity, senescence, and polymorphic microbiomes [25]. These three papers mentioning the cancer hallmarks are the most cited cancer-related publications of the century which reveal the importance of hallmark approach.

Breast Cancer

Breast cancer is one of the most frequently spreading cancer among all the types of cancers. It covers a major ratio of deaths in women diagnosed with cancer. Breast cancer can be divided into different types on the basis of availability of treatments of cellular markers[39] [91]. Repeated exposure of breast cells to circulating ovarian hormones causes BC which makes it a homonally mediated disease [48].

Breast cancer starts in the lining of the tissues of the breast. About 95% of BC are carcinomas [64]. The proliferation of uncontrolled cells in ducts that carry milk from breast to nipple causes Ductal carcinoma. It is the 75% of all invasive breast cancers. Lobular carcinoma begins in the lining of lobules that produce milk. It accounts for 5% to 10% of all invasive breast cancers [31]. In infiltrating or invasive carcinoma the cancer penetrates into the cells of membrane from the epithelial cells. a very small number of breast cancers may arise from the muscle, fat, or connective tissues of breast [49].

Etiology

Understanding the etiology and epidemiology of BC uncovers the causes, risk factors and the disease trend in different ethnicities.

Females experience more alterations of sex hormones through out their life as compared to the males. Each year, 280 000 women and 3000 men are reported as BC patients. The percentage of being a BC patient is 100 times more in women than men [71]. High stimulation of estrogen and progesterone is pronounced prevalence in women.

Increase in BC risk can be influenced by the family history of BC [48]. Most of the BC cases are diagnosed after the age of 50 which clears that increase in age is associated with higher risk of BC [79]. The repeated exposure of breast cells to the ovarian hormones is the cause of BC making is hormonally mediated diseased[30]. Pregnancy related factors and Oral contraceptives also increases the risk for BC [81].

BC is highly dependent on the choice of life style. Various life style choices can increase and decrease the risk of BC. Weight and obesity is one of the promising factor that contribute in BC in postmenopausal women. Interestingly, the same thing obesity decrease the risk of BC in premenopausal women. The consumption of alcohol and smoking cigarette for long time have significant role in causing BC. Intake of 1-2 alcohol drinks per day can increase 30-50% BC risk while smoking for long line can increase approximately 10% BC risk [11]. Secondhand smokers are also prone to BC especially double positive BC. Other life style choices like physical activity, psychological factors, and diet also have significant effect in BC. In context of diet, high intake of carbohydrate is associated with high risk of BC. However, it is modified by age as the risk of BC with high intake of carbohydrate is less in women younger than 50. Researchers have also reported a positive association of breast cancer risk with red meat intake, espe- cially well-done red meat[48].

Some environmental factors including nitrogen dioxide (NO2) and nitrogen oxides (NOx) play significant role in causing BC [81]. Moreover, a group of chemicals, Phthalates are reported as significant risk factors associated with BC. Dioxins, polychlorinated biphenyl (PCB), polycyclic aromatic hydrocarbons (PAH), bisphenol A (BPA), and per- and polyfluoroalkyl substances (PFAS) are all reported as endocrine disrupting chemicals and hypothesized to interfere with estrogen. Out of all these, dioxin and BPA are not associated with BC, while PCB, PFAS, and PAH re observed to be linked with BC [48].

The genes that inherit and cause BC are BRCA1, BRCA2, aur PALB-2. BRCA1 and BRCA2 account for 80% of heredity breast cancer. In general, heredity is consid-

ered as only 5% to 6% of BC. Women with positive BRCA-1 and/or BRCA-2 are 50% to 85% risk of developing BC, and 15% to 65% at risk of developing ovarian cancer [64].

Epidemiology

Epidemiology discuss the trend of disease in different groups, regions and ethinicites which help in understanding the disease. In 2018, around 2.08 million women were diagnosed with BC and it covers the 36% of all the cancer patients [54]. BC cases are increasing in all regions of the world, however, it is especially elevating in industrialized countries. America being one of the biggest industrialized country reported 234,087 breast cancer cases in 2018, 55,439 in the United Kingdom, 56,162 in France and 71,888 in Germany [58]. BC was the most diagnosed cancer type in Poland in women. The highest incidence rate in the world is found in Belgium. South Asia and Africa has the lowest incidences of BC with standardized incidence rate does not exceed 25/105. Overall, BC is the first cause of death from malignant tumors in women. Developing countries has the highest rate of deaths recorded [73].

Subtypes and Classification of Breast Cancer

Breast cancer can be classified into 4 categories on the basis of immunohistochemical expression of 3 hormone receptors: Estrogen Receptor(ER), Progesterone Receptor(PR) and Human Epidermal growth factor receptor(HER2) [59]. The different expression levels of these hormone receptors, the types of breast cancer are: (a) luminal A, (b) luminal B, (c) HER2-Enriched, (d) Basal-Like and Normal-Like breast cancer[91].Triple Negative Breast Cancer (TNBC) is the fifth type that shows no expression of all three receptors [59].

Early-stage BC has 10-year survival rates of over 90%. Metastatic breast cancer (BC) makes up about 6–7% of initial diagnoses and later develops in 30% of early-stage BC cases, with a 5-year relative survival rate of 25% and a median overall survival of $\tilde{2}$ years. Other several classifications of BC are reported as well. Currently, the most widely used classification system is the Tumour Node Metastasis (TNM) system. It classifies on the basis of primary tumour (T) size, nodal (N) involvement, and metastasis (M). This system was developed and is maintained by the Union for International Cancer Control (UICC) and is also used by the American Joint Committee on Cancer (AJCC). The Nottingham Prognostic Index (NPI18) was the first BC prognostic staging system. It is based on the lymph node stage (1–3), histological grade (1–3) and the primary tumour size [63].

TABLE 1 Characteristics of subtypes of breast cancer

	Luminal A	Luminal B	HER2	TNBC	Reference
Frequency (%)	50	15	20	15	16, 17
ER	Yes	Yes	Some cases	No	15
PR	Yes	Some cases	Some cases	No	17
HER	No	No	Yes	No	18
miRNAs	*Let*-7f, *Let*-7c, *miR-10*, *miR-29a*, *miR-181a*, *miR-223* and *miR-652*	*miR*155, *miR-93*, *miR-18a*, *miR-135b*, *miR-718*, *miR-4516*, *miR-210*, and *miR-125b-5p*	*miR*-150 and *miR*-142–3p	*miR*-153, *miR*-10b, *miR*-26a, and *miR*146a	19–22
Ki67	Some cases	Some cases	High	High	23
Mutations	No	BRCA2	p53	p53 and BRCA1	24, 25
Prognosis	Good	Middle	Middle/Bad	Bad	26
Therapy	Hormonal	Hormonal/ Chemo	Hormonal/Chemo/ Herceptin	Chemo/ Experimental	27, 28

[59]

Triple Negative Breast Cancer

In triple negative breast cancer, ER, PR expressions are $< 1\%$ and HER2 is negative detected by immunohistochemistry [39]. TNBC covers 15% to 20% of total breast cancer cases and has poorest prognosis among BC types. TNBC subtype is further classified into: basal-like (BL1 and BL2), claudin-low, mesenchymal (MES), luminal androgen receptor (LAR), and immunomodulatory (IM). About 80% of TNBC are BL and BL has over-expression of keratin 5, 17 and epithelial grown factor receptors (EGFR)-related genes, [59] [91].

Risk Factors and Etiology

In a recent study, a comprehensive meta analysis was performed on thirty three studies to understand the role of different risk factors in TNBC. They concluded that family history, longer use of oral contraceptive and higher breast cancer density were playing a significant role in increasing risk for TNBC. Moreover, factors like later age at menarche, later age at first birth and breastfeeding were associated with TNBC. In black women, higher parity was associated with a higher risk of TNBC

[38]. The incidences of TNBC are higher in younger women. Several studies revealed that it is more common in age from 20 to 39 years. Review of cases estimated that every passing decade decreases the risk of TNBC by 16% [75]. On the basis of data from the National Cancer Database from 2010 to 2011, the rates of TNBC decreased from 23.3% in patients of age 30 to 10% in patients of age more than 70% [62]. In accordance to basal-like TNBC, it is more common in younger women while half of the patients older than 65 years with triple-negative cancers have nonbasal subtypes [22].

The link between obesity and BC is well reported and discussed earlier but it varies by receptor types. Meta-analysis of several studies reported the trends in these receptor types. In women of 50 years or younger, the risk of TNBC was found to be increased with compared with HR+/HER2- cases. While no increase was observed for the same comparison in older women [90]. A separate meta-analysis of nine studies that reported estrogen receptor (ER) and progesterone receptor (PR) status found no elevated risk of ER-/PR- breast cancer among overweight (RR = 1.06; 95% CI: 0.95-1.18) or obese (RR = 0.98; 95% CI: 0.78-1.22) postmenopausal women. In contrast, a clear increased risk was observed for ER+/PR+ breast cancers [53]. In premenopausal women, an increased risk of ER-/PR- breast cancer was observed in those who were overweight (RR = 1.26; 95% CI: 1.07-1.49). However, a pooled analysis of over 700,000 cases found no elevated risk of triple-negative breast cancer in premenopausal women over the age of 25, and even identified a protective effect associated with each 5-unit increase in BMI among women aged 18 to 24 (OR = 0.83; 95% CI: 0.70–0.99) [66].

In somatic genomic alterations, TP53 is the most frequently mutated driver gene (>80%), followed by PIK3CA. PTEN, KMT2C,andRB1 are other genetic alterations with <5% frequency [14]. For the copy number alterations, EGFR and FGFR2 are amplified; and PTEN loss are found in TNBC [70]. Different genes whose amplification is the risk factor for TNBC are MYC, PIK3CA, CDK6, KRAS, FGFR1, IGF1R, CCNE1, CDKN2A/B, EGFR, AKT1 and deletions in BRCA2, PTEN, MDM2, RB1, CCND3, ESR1, CDKN2A/B, SMAD4, NF1, NCOR1,and TP53 [6]. Other than these, [33] a comprehensive analysis of pathogenic variants of 21 genes in 10,901 TNBC patients compared with controls identified an increased risk with mutations in PALB2 (OR, 11.05; 95% CI, 8.14–15.09), BARD1 (OR, 7.35; 95% CI, 4.8–11.17), CDKN2A (OR, 7.71; 95% CI, 2.3–24.10), RAD51D (OR, 4.33; 95% CI, 2.19–8.66), MLH1 (OR, 3.37; 95% CI, 1.05–9.64), RAD51C (OR, 2.97; 95% CI, 1.86–4.75), TP53 (OR, 2.51; 95% CI, 1.21–4.98), and BRIP1 (OR, 2.41; 95% CI, 1.52–3.76) [69].

BRCA1 gene at chromosome 17q21 was discovered to be the genetic factor for causing hereditary BC and ovarian cancer. Out of all these hereditary BC cases, 70%

are the TNBC in premenopausal. BRCA mutations are the positive prognostic factor for TNBC is is found very common in Ashkenazi Jews.

Epidemiology

Basal-Like tumors were thought to be the same as TNBC because it also frequently lacks the expression of ER, PR, and HER2. However, later on it is observed that both are not entirely synonymous. By the PAM50 intrinsic subtype classifier, around 5% to 10% TNBCs are luminal, HER2-enriched, or normal-like [33]. Other than these, 6 more TNBC subtypes were reported by Lehmann and colleagues. The subtypes are basal-like 1, basal-like 2, mesenchymal, luminal androgen receptor, mesenchymal stem–like, and immunomodulatory [40].

The epidemiology of the TNBC is very challenging to study because of it's heterogeneity. It is not a single disease rather a group of diseases. Studying epidemiology of Basal-Like uncovers that such tumors occurs in African American (AA) women. This is associated with multiparity, earlier menarche, low rates of breastfeeding, elevated waist-to-hip ratio, and high adiposity [7]. Epidemiologic studies reported ER and PR expression but many epidemiologic studies do not mention the expression of HER2. Thus the enough data is unavailable about epidemiology of TNBC.

Society differences may contribute to the elevated risk. BC is observed 2 to 3 fold increased risk in black American women [51]. 15.4 to 18.4 per 100,000 mortality rate was reported in Africa in 2018. Europe and North America has double and triple incidence rate but the mortality rates are lesser in these regions as compared to Africa. There is considerable variation in triple-negative breast cancer rates among different groups of Black women. The highest rates are observed in African American women (23.7% of cases) and West African women (24.1%), while lower rates are seen in East African (11.6%) and Caribbean women (21.2%) [78]. Hispanic and non-Hispanic Whites show the similar rate of TNBC diseases. They show high heterogeneity with varying rates of breast cancer among Cubans, Puerto Ricans, and Mexicans. In Asian women, the overall rates of TNBC are low. For premenopausal TNBC, the rates are significantly low. However, with the increase in age the number of incidences becomes similar [43].

Ovarian Cancer

Because of vague symptoms, Ovarian Cancer does not diagnose until it is on advance stage making it a silent killer for women. Rather than originating from ovaries, ovarian cancer, one of the gynecological cancer, starts from fallopian tube [76]. Worldwide, it was seventh most common cancer type as of 2018. The rate of new cases of ovarian

cancer was 10.3 per 100,000 women per year. The death rate was 5.9 per 100,000 women per year. These rates are age-adjusted and based on 2018–2022 cases and 2019–2023 deaths [can]. The median age at diagnosis of OC is around 63 years in most developed countries [4].

Sub-Types on the basis of Histopathology

We can divide ovarian cancer into three primary forms: (a) epithelial (the most prevalent), (b) germ cell, and (c) sex-cord-stromal. Germ cell and sex-cord-stromal make up only the 5% of all ovarian cancer. For epithilial ovarian cancer, (a) Serous, (b) Endometrioid, (c) mucinous, and (d) clear cell are the four major common histological subtypes. So we can say that, ovarian cancer is also a type of carcinomas. Furthermore, High-grade serous carcinomas (HGSC) and low-grade severe carcinomas are the two main types of serous tumors (LGSC). HGSCs make up 70% to 80% of all epithelial categories, about 90% of ovarian cancers in the Western are surface epithelial carcinomas [60]. These LGSC and HGSC are also known as Type 1 and Type 2 ovarian carcinomas.

Type 1 includes endometrioid, serous type of low-grade, clear-cell, and mucinous cancers. Seromucous and Brenner tumors are the rarest sub-types of Type 1 OC. Type 1 are diagnosed at early stages and caused by atypical proliferative (borderline) tumors. Moreover, type 1 have the low grade as well. As mentioned, type 2 are the HGSC being more fatal as compared to type 1. Type 2 includes high-grade carcinoma of serous type, carcinosarcoma, and carcinom. HGSC are diagnosed at critical stages and they have extreme rate of proliferation with a fast and aggressive progress rate and instability of a very high degree in chromosomes. [65].

The most common type of ovarian carcinomas is serous type. It usually is seen as low grade (10% of every tumor of serous sub-type) or as high-grade cancer (90% of every tumor of serous sub-type). Low grade is mostly detected in younger age while HGSC is more likely to be diagnosed in older age group of people. Moreover, LGSC has a good prognosis and HGSC has a 10-year rate of morality of 70% [52]. Clear-cell ovarian cancers are less common and make only 5% of ovarian carcinomas.

WHO Classification

World Health Organization (WHO) also categorized the ovarian cancer into several types. The WHO classes of OC are: 1) Epithelial tumors, 2) Mesenchymal tumors, 3) Mixed epithelial and mesenchymal tumors, 4) Sex-cord stromal tumors, 5) Germ cell tumors, 6) Monodermal teratoma and somatic type tumors arising in dermoid cyst, 7) Miscellaneous tumors, 8) Mesothelial tumors, 9) Soft tissue tumors, 10) Tumor-like lesions, 11) Lymphoid/myeloid tumors, and 12) Secondary tumors.

Furthermore, epithelial tumors are classified into: 1) Serous tumors (cystadenoma, adenofibroma, surface papilloma, and borderline tumor), 2) Mucinous tumors (cystadenoma, adenofibroma, borderline tumor, and carcinoma), 3) Endometriod tumors (cystadenoma, adenofibroma, borderline tumor, and carcinoma), 4) Clear-cell tumors (cystadenoma, adenofibroma, borderline tumor, and carcinoma), 5) Seromucinous tumors (cystadenoma, adenofibroma, and borderline tumor), and 6) Brenner tumors (borderline and malignant [65]. As epithelial tumors are further classified into 6 sub-types, Mesenchymal tumors are also sub-divided into following sub-types: 1) Endometriod stromal sarcoma (low grade and high grade), 2) Smooth muscle tumors (leiomyoma and leiomyosarcoma), and 3) Ovarian myxoma. Mixed epithelial and mesenchymal tumors are further classified into adenocarcinoma [92].

Sex cord stromal tumors are further classified into: 1) Pure stromal tumors (fibroma, thecoma, sclerosing tumor, microcystic tumor, signet ring tumor, leydig cell tumor, steroid cell tumor, and fibrosarcoma), 2) Pure sex cord tumors (adult granulosa cell tumor, juvenile granulosa cell tumor, sertoli cell tumor, and sex cord tumor with annular tubules), and 3) Mixed sex cord stromal tumors (sertoli-leydig cell tumor, sex cord stromal tumor, and gynandroblastoma) [7]. Germ cell tumors can be further classified into: 1) Benign teratoma, 2) Immature tertoma, 3) Dysgerminoma, 4) Yolk sac tumor, 5) Embryonal carcinoma, 6) Choriocarcinoma, 7) Mixed cell tumor, 8) Monodermal teratoma and somatic type tumors arising from a dermoid cyst, and 9) Germ cell sex cord stromal tumors [65].

Miscellaneous tumors can be further classified into: 1) Rete cystadenoma/adenoma, 2) Wolffian tumor, 3) Solid pseudopapillary tumor, 4) Small cell carcinoma of the ovary, and 5) Wilms tumor. Tumor-like lesions can be further classified into: 1) Follicle cyst, 2) Corpus luteum cyst, 3) Large solitary luteinized follicle cyst, 4) Hyperreactio luteinalis, 5) Pregnancy luteoma, 6) Stromal hyperplasia and hyperthecosis, 6) Fibromatosis and massive edema, and 7) Leydig cell hyperplasia [92].

Risk Factors and Etiology

OC is the cancer of postmenopausal women and is rare in women below the age of 40 years. In younger women, the risk of development of Endometriosis-associated epithelial ovarian cancers is higher. The late age of natural menopause is associated in elevating the risk of OC. Late age of menopause starts at 55 years of age or beyond [4]. Studies clearly says that current smokers has a double of risk of getting an OC. As compared to never smokers, current smokers have a 83–125% higher risk of ovarian mucinous borderline malignant tumours [42].

In the context of heredity and family history, just as other cancers, the family

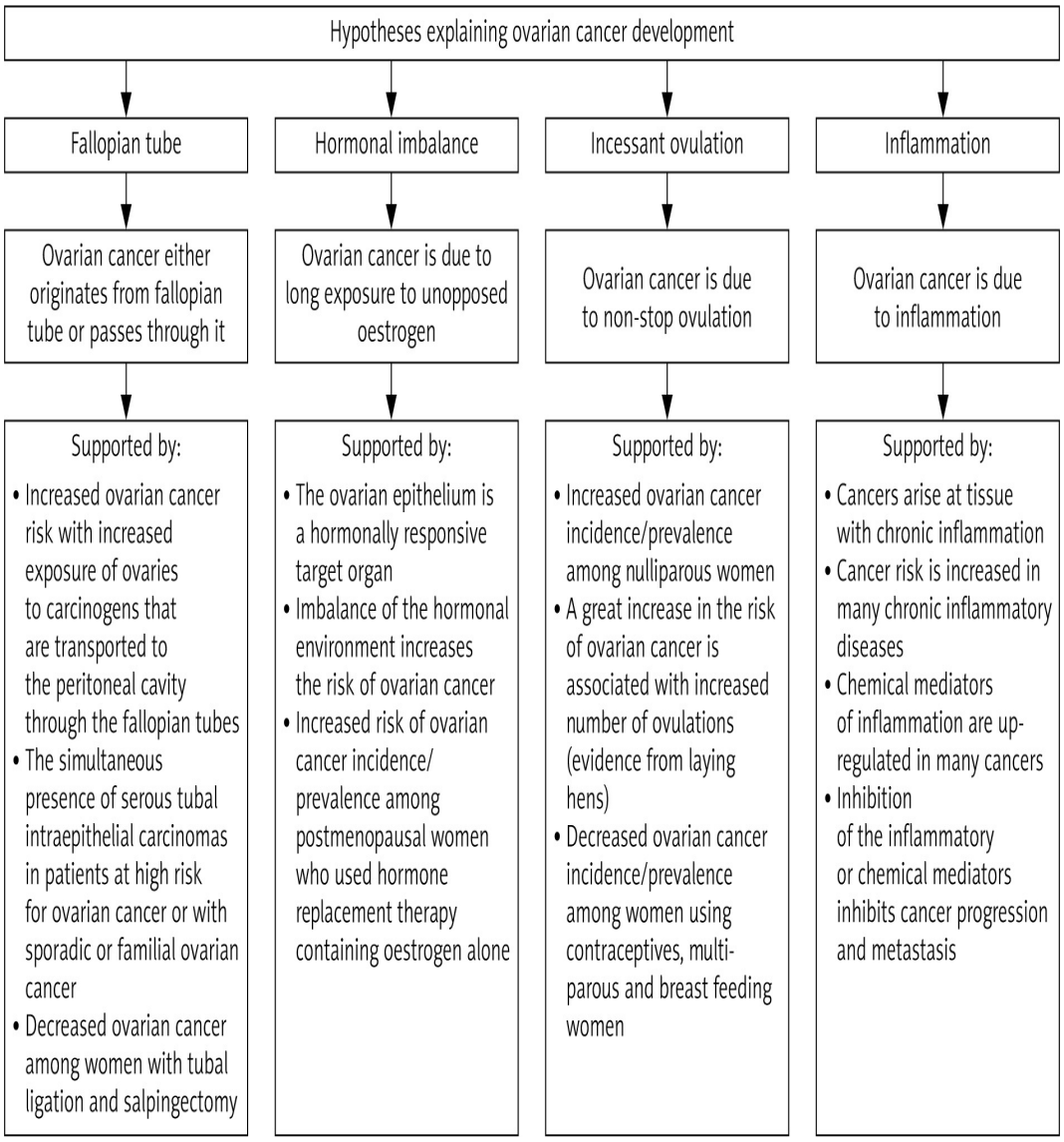


Figure 2.1: Ovarian Cancer types
[4]

history of OC increases the risk of OC. Genetics is the cause of OC in 10 to 15% of the cases. The mutations in BRCA1, BRCA2 and MMR gene are the direct genetic risk factors of OC and can increase the risk factor from 1.6% to 40%, 18%, and 10%, respectively [76]. Overall, it is a very complex disease which can develop from different types of cells and we discussed its various sub-types in previous section.

There is a genetic link between OC and BC known as hereditary breast and ovarian cancer syndrome (HBOC). Family history of BC also increases the risk factor of OC and vice versa [80]. Dietary factors like low levels of Vitamin D; and certain ethnicity backgrounds specifically Jewish, French Canadian and Dutch are also associated with increase in risk factor for OC. Approximately 90 % of OC incidences are developed from ovarian surface epithelial cells [84]. High intake of fatty acids and dense caloric foods is highly associated to a number of cancers. As, over weight and obesity is the significant risk factor in TNBC it also contribute in causing OC [4]. Moreover, women in developed countries are less prone to breastfeeding. Breastfeeding is reported to has a protective effect against OC [28]. BRCA mutations is the main cause of OC which is is common in White (2.9%) than Black (1.4%) cases and in Jew ish (10.2%) vs. non-Jewish (2.0%) cases. The mortality rate in White American women is 4% of the total death rate due to cancers [4].

Epidemiology

OC is the one of the most fatal gynaecologic that causes two thirds of all deaths due to gynaecologic cancers. OC is spread in Western Europe at the highest rate, intermediate in Southern and Eastern Europe and South America and lowest in the Middle East and Asia. The more ovulatory cycles the higher the risk of OC which shows a direct association of ovulation with OC [84]. One of the reason of prevalence of OC in developed countries as compared to developing countries is the significant decrease in the fertility rate. Developing countries have larger families while in developed countries people prefer small or no families. Even in America, African American women have 40% less chances to have OC as compared to the White American Women.

Comonalities in both cancers.

Common things in breast cancer and ovarian cancer. [4] has the details in genetics section.

Microarray Data Analysis

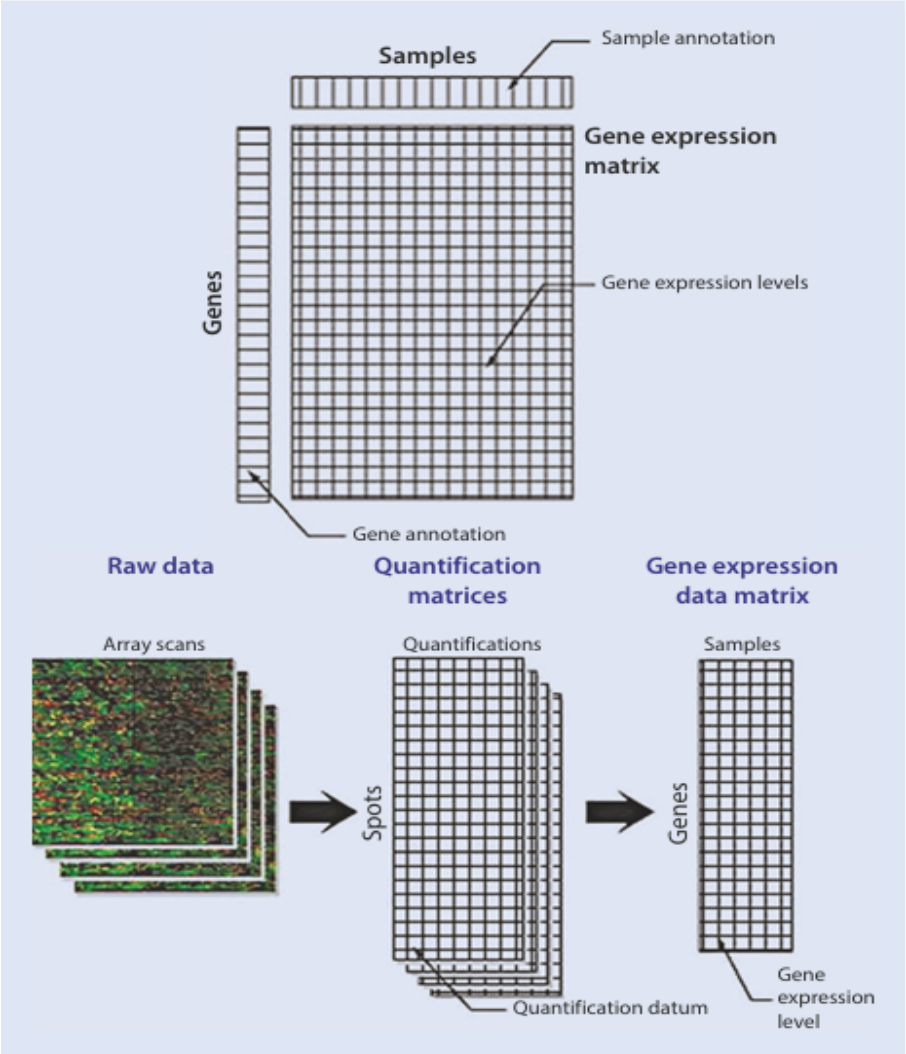
Microarray Experiment

The DNA microarray experiment can be done in 5 steps. Step 1 is making a DNA biochip. In step 2, a specimen is prepared which separates mRNA from the

experimental and control specimens individually. Step 3 involves a hybridization step, in which experimental specimen is stained red and the control specimen is stained blue. Step 4 quantifies of the brightness of the fluorescence of each probe that is hybridized with fluorescent dyes with the laser fluorescent scanner. I last and 5th step, numerical value for the expression of each gene are analyzed [36].

Structure and Normalization of Microarray data

The data from the microarray experiment is indicated in gene expression matrix. It is divided intro three parts: (a) a list and annotation of genes, (b) a list and annotation of specimen, (c) numerical value of the gene expression which is actually the elements of the matrix. [36]



Data normalization is a paramount step in microarray data analysis. Normally, because the biological data are simple and doesnt require data normalization or pre-processing before quantitative analysis. But Microarray experiments use large-scale data normalization and relative values5 of the intensity of light, and each chip varies

during the manufacturing process. Moreover, the red and green dyes have different DNA binding and signal detecting efficiencies, requiring data normalization and preprocessing [36].

For single channel expression array data, Quantile normalization is a global mean or median technique used for the normalization [9]. First it arranges the expression values in an order, then takes the average of probes. Minus the probe intensity with average values and rearrange the values to their original order [29]. Robust Multi-chip Average (RMA) package in R is commonly used to make expression matrix from affymetrix data. It first normalize the data, then perform background correction and calculation og gene expression values [34]. For illumina data, Robust Spline Normalization (RSN) is used for Quantile normalization [16].

There are a number of methods for the normalization of gene expression data. We have discussed quantile normalization, while other methods are also available. Loess normalization perform the local regression data. Log-transformation is the simplest and very common normalization technique. Standardization is also a data normalization technique which doesn't bind values to a specific range. In standardization, Z-score is the most common method used for standardization of data [9].

Feature Selection & Feature Extraction

The expression data we obtain from biochips is of very high dimension. High dimensional data is very sparse and not suitable for prediction or analysis. For that reason, this high dimensional data undergoes a process called dimensionality reduction which literally means reducing the dimensionality of the data. After dimensionality reduction, now the data is ready for further analysis. Dimensionality reduction is carried out either by extracting important features form the data or by removing irrelevant or unimportant features. This is called featuring engineering [9]. Feature extraction or selection techniques can be divided into four categoreis: (i) filter methods, (ii) wrapper methods, (iii) embedded methods, and (iv) hybrid methods [82], [15], [5].

In filter approach, the selection of features is carried out on the basis of fold change or p-value. Statistical test like ANOVA, that are often used for classification are also employed for feature selection purposes. Wrapper approach is based on the learning algorithms in which in which subset of features are selected and evaluated. It overcomes the limitations of filter method. As the name mentioned, embedded methods uses both wrapper and filter and approaches for feature selection. A common embedded technique is LASSO (Least Absolute Shrinkage and Selection Operator). Hybrid approach integrate the two or more than two wrapper and filter approaches. The main challegen hybrid approach overcomes is the overfitting. Recursive Feature

Elimination with a linear SVM (SVM-RFE) is a well-known feature selection hybrid approach [23]. The drawback of hybrid approach is its high computational power usage and also it ignores the correction between the features [9].

Differential Gene Expression (DEGs)

DEGs analysis is a bioinformatics' technique which is used to compare two or more than two groups of samples. These samples come from different sequencing lab techniques such as RNA seq and microarray. The groups could be diseased vs healthy or response of genes towards different treatment drugs [50]. It is the most common analysis in gene expression analysis. DEG analysis can be performed by employing different statistical techniques, for example t-test, ANOVA, chi-squared etc [9]. The analysis mainly identify the genes that show different levels of expression as compared to the normal genes [50]. By this, we can find specific genes involved in causing a disease or genes that show certain response behavior to treatments or drugs [35]. In R language, commonly used packages for DEG analysis are limma, affy and oligo etc [9].

Steps

In DEG analysis, we initially perform normalization on our data and preprocess it. Noice is reduced in the data to resolve technical issues that are generated by the machines [45].

After Cleaning the data, a model is selected on the basis of distribution of data [44]. The data following a specific statistical distribution is gone through Parametric methods while for other types of data with complex data distribution Non parametric tests are preferred [74].

The third step identify the significant differences in gene expressions of processed data. The differentially expressed genes (DEGs) are identified based on an absolute log2(fold change) > 1 and an FDR (padj) <0.01 [41]. Disease mechanisms, role of specific drug, effect of different conditions or environments and other information about pathways involved in a mechanism can be extracted from these differentially expressed genes [44]. The final step is the visualization of these DEGs which allows researchers to observe trends, patterns, and anomalies in the data [86].

Gene Set Enrichment Analysis

PPI Network Analysis

CHAPTER 3

2.2 MATERIALS AND METHODS

Two microarray datasets were downloaded from ArrayExpress [60] database. GSE76250 [46] was TNBC and GSE26712 [10] was Ovarian Cancer. GSE76250 contains 198 total samples out which 33 were paired normal breast tissues and 165 were TNBC samples. We divided the dataset on the basis of Cancer Grade and made 3 categories i.e. Grade 0, Grade 2 and Grade 3. Grade 0, Grade 2, and Grade 3 datasets had 19, 35, 69 number of samples, respectively. In Ovarian Cancer dataset, we tool 60 samples, 10 normal and 50 diseased.

2.2.1 Dataset Preprocessing

Data was imported and read in R studio. No null values were found in the data. We calculated data summary of all the 4 datasets; and plotted Boxplot and histogram to analyze the distribution of dataset. RMA function of oligo package normalized the dataset, performed background correction and expression value calculation. RMA automatically apply log2 transformation on the data as well during normalization process. PCA was performed on expression data. Then probe annotation gave the gene ids and against gene symbols using hta20transcriptcluster database. Low expressed genes were removed with the threshold 5. Now, the filtered dataset was ready for differential gene expression analysis.

2.2.2 Differential Gene Expression Analysis and Common Genes

Limma package of R extracted the differentially expressed genes in all the datasets. Genes were filtered with the threshold of fold change >1.0 and value of $p < 0.05$. Volcano plot visualized the DEGs and Heatmap visualized their expression. After that, common genes in TNBC and Ovarian Cancer were extracted by drawing Venn diagram using python. As a result, we obtained 3 sets of common genes for each grade, i.e. common genes between ovarian cancer and TNBC grade 0, common genes between ovarian cancer and TNBC grade 2, common genes between ovarian cancer and TNBC grade 3,

2.2.3 Gene Set Enrichment Analysis

Before GSEA, we removed duplicate genes. Filtered genes were sorted by decreasing order. ClusterProfiler package gave the Gene Ontology and KEGG pathway of each set of common genes. For Gene Ontology parameters ont was set to All, minGSSize to 3, maxGSSize 100, verbose to TRUE and pvalueCutoff = 0.05. For KEGG, organism set to hsa, keyType to ncbi-geneid, exponent to 1, minGSSize to 3, maxGSSize to 100, pvalueCutoff to 0.05, verbose to TRUE and pAdjustMethod to BH.

2.2.4 PPI Network Analysis and Identification of Hub Genes

Genemani was used to construct the PPI network of all set of genes through Cytoscape. Using those networks, hub genes were identified for grade 0, grade 2 and grade 3, separately. The 12 methods to calculate the hub genes are: Betweeness, BottleNeck, Closeness, Clustering Coefficient, Degree, DMNC, EcCentricity, EPC, MCC, MNC, Radiality, and Stress. The complete methodology is given in Figure **??**

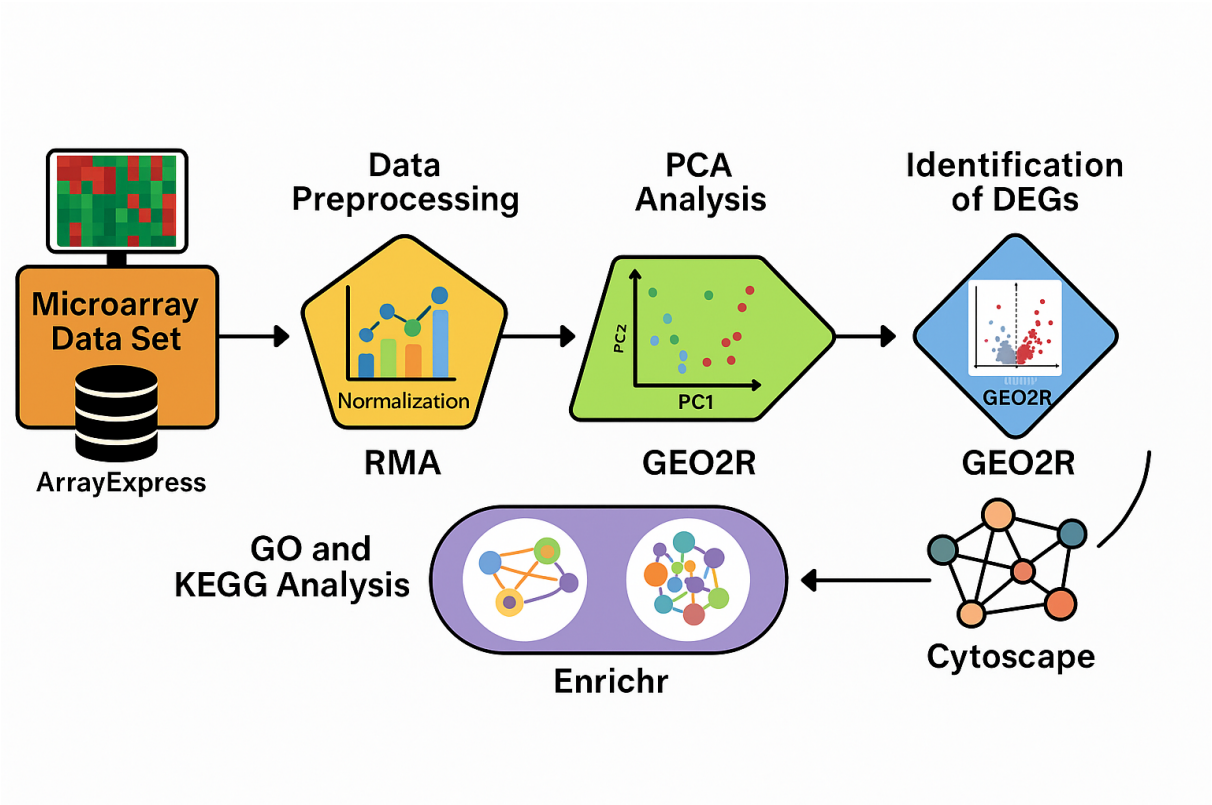


Figure 2.2: methodology

2.3 RESULTS

2.3.1 DEGs Screening of Ovarian Cancer and TNBC and their Common Genes

We obtained two datasets GSE76250 and GSE26712 for TNBC and ovarian cancer. Both datasets contained normal and diseased samples. TNBC dataset was divided into grade 0, grade 2 and grade 3 sub-datasets. Before DEG analysis, the data was normalized. Background correction, normalization and calculation was performed using RMA (Figure 2.3). PCA was applied on data and plotted(Figure **??**). Microarray data analysis of these sub-datasets revealed 201 DEGs for grade 0, 165 for grade 2 and 192 for grade 3. For ovarian cancer dataset, 1757 DEGs were found.

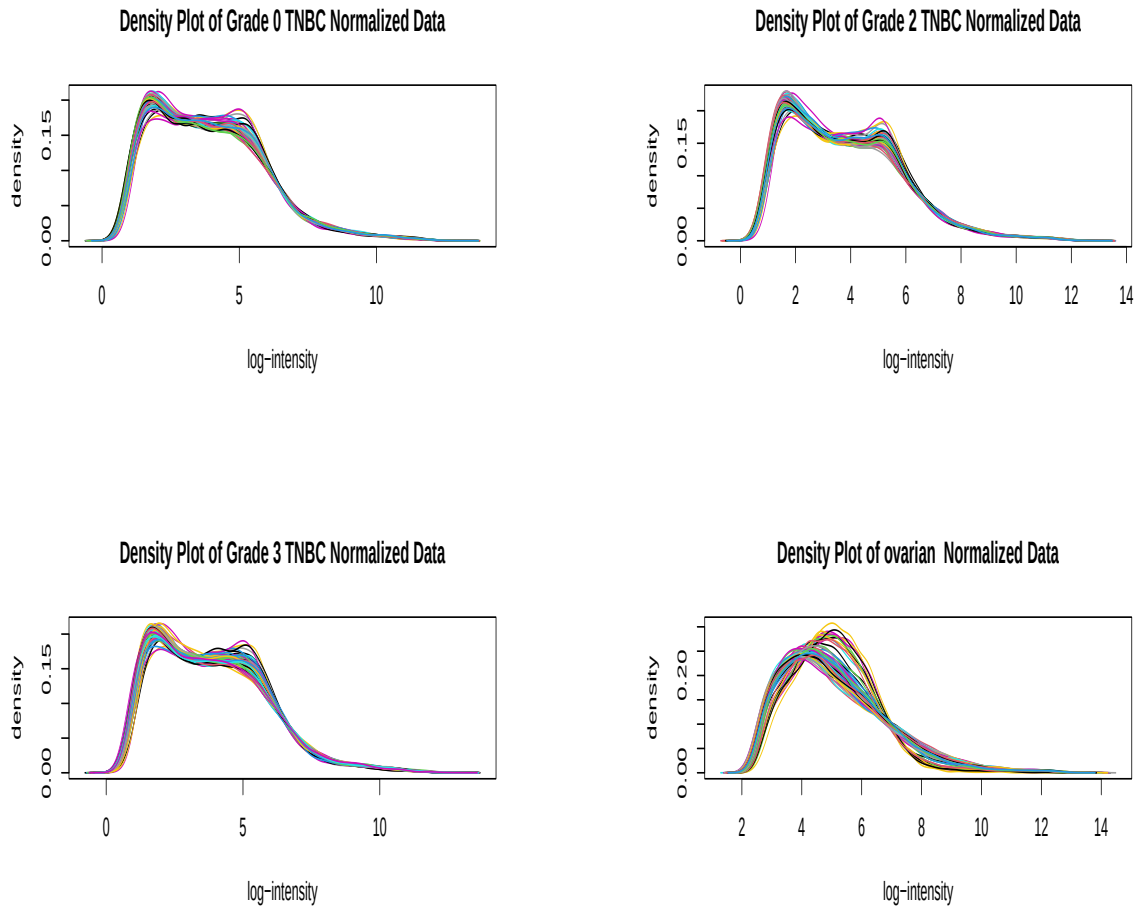


Figure 2.3: Normalized Distributions

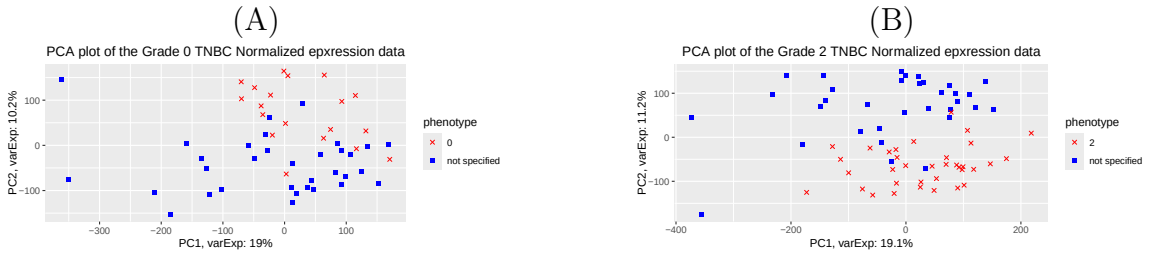


Figure 2.4: PCA plots of (A) Grade 0 normalized data and (B) Grade 2 normalized data.

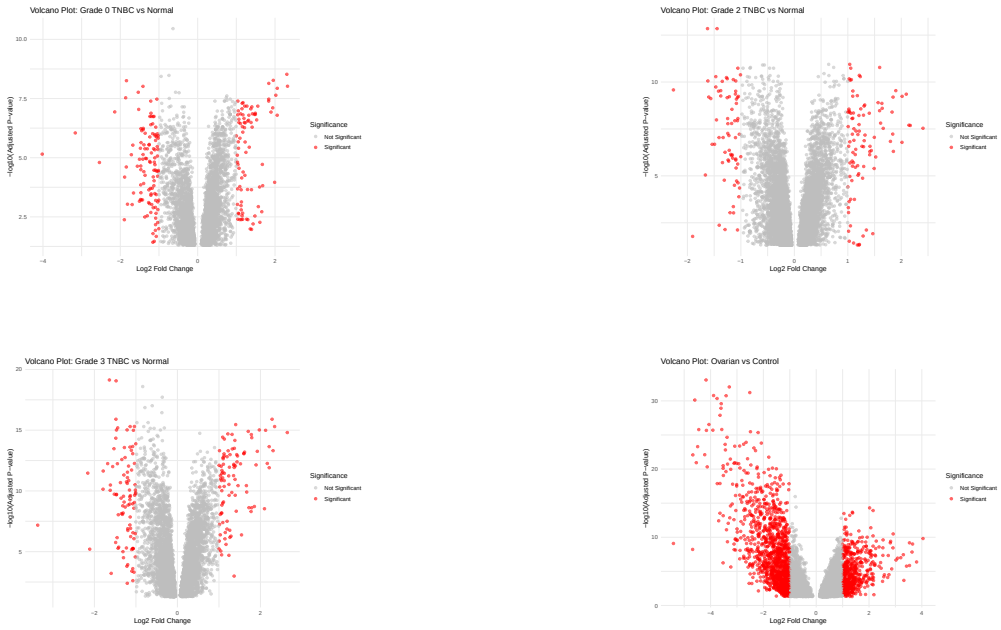


Figure 2.5: Volcano plots for differential gene expression in various cancer grades and ovarian cancer

2.3.2 Identification of common transcriptional signatures between TNBC and Ovarian Cancer

A Venn diagram was constructed between the sub-datasets of TNBC and ovarian cancer after differential gene expression analysis. 22 genes were common in grade 0 and ovarian cancer dataset. Similarly, 20 and 21 common genes for grade 2 and grade 3, respectively. These show higher similarity between TNBC and OC. The overlapping genes among the grades were 12. the details of all the genes are given in table 1, table 2, table 3 and table 4.

2.3.3 Gene Set Expression Analysis

GO

KEGG and GO analysis was performed on these common genes to analyze the common pathways between grades of TNBC and Ovarian Cancer. In grade 0, GO analysis

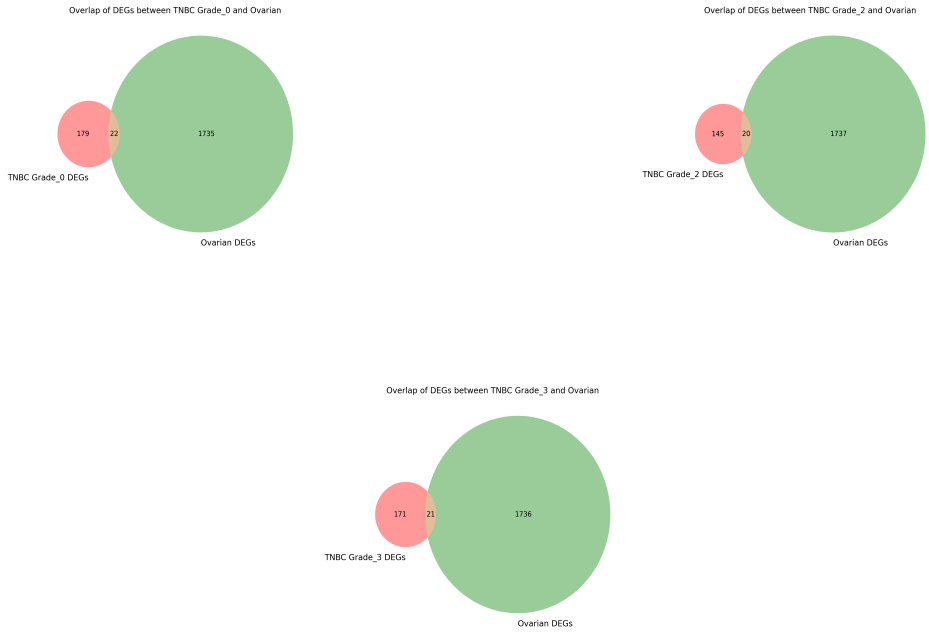


Figure 2.6: Venn diagrams showing DEGs overlaps in Grade 0, Grade 2, and Grade 3

reveals that membrane-enclosed lumen, organelle lumen intracellular organelle lumen and nucleus are the Cellular Components (CC) with nucleus being the most important CC with set size 11 while others having set size 6. Positive regulation of metabolic process and organelle organization are the two biological processes in which genes were involved. No Molecular Function (MF) is identified.

For grade 2, intracellular membrane-bounded organelle and nucleus are the major CC with highest set size of 13 and 8, respectively. In terms of Molecular Function, genes were involved int organic cyclic compound binding. Biosynthetic process, cellular metabolic process, negative regulation of biological process and negative regulation of cellular process are the major biological process in which genes are taking part.

In the end, grade 3, intracellular anatomical structure and nucleus are the major CC, organic cyclic compound binding, binding and nucleic acid binding are the MF involved with binding having the highest 18 set size. In terms of biological processes, genes are involved in biosynthetic process, cellular metabolic process, regulation of biosynthetic process and regulation of cellular metabolic process.

All the three datasets were sharing nucleus in terms of CC. Grade 0 has no MF identified while organic cyclic compound binding is common in grade 2 and grade 3. The shared effected BP between Grade 2 and Grade 3 were biosynthetic process, cellular metabolic process, positive regulation of molecular function, animal organ development and regulation of molecular function. There were 3 BPs in grade 3 that

Sr. No.	Gene Names
1.	TDP2
2.	CDH10
3.	B2M
4.	ZNF331
5.	MFAP3
6.	CIDEA
7.	AKAP6
8.	TAAR2
9.	WDR1
10.	CLEC3B
11.	LDAH
12.	MXRA5

Table 2.1: List of Grades' overlapping Genes

were not shared with grade 2. Those are regulation of biosynthetic process, regulation of cellular metabolic process and positive regulation of cellular metabolic process. In Grade 3, a total of 3 MFs were found, 1 in Grade 2 and 0 in Grade 0 which clearly shows that in grade 3 genes were involved in more MFs.

Just like overlapping BPs, MFs and CCs; Grades have 12 overlapping genes. Out of 12 common genes, few were repeating numerous times in GO analysis. Out of 12, we pick the top 3 repeating genes, that are B2M, RHOA and AKAP6. Following is the list of all the genes and their occurring frequencies. .

KEGG

After GO, KEGG pathway analysis was performed to identify the underlying pathways in which these genes are involved. Although, no KEGG pathway fulfill the criteria of p-value > 0.05, however, some strong pathways can be considered. The details of some of the KEGG pathways is given in bar.

2.3.4 The PPI Network

The ppi network was constructed using Genemania [85] in Cytoscape. Significant interactions were observed in almost all of the genes. There are 12 methods to find hub genes in the ppi network. Top 10 hub genes was calculated for all of the methods. In grade 0, IFI27, B2M, OASL and STAT1 occurred in 10, 9, 9, 9, 9 methods, respectively out of 12 methods. For grade 2, RHOA occurred in 10 methods, B2M in 8 and SELENOF in 8 methods as well. AP2M1, TMSB4X and RHOA were most significant hub genes found in grade 3 with the frequency of 10, 10, and 9 among all the 12 methods, respectively.

Among the grades, B2M occurred in grade 0 and grade 2; RHOA occurred in

Table 2.2: Gene Frequency Table in GO Analysis

Gene	Frequency
B2M	34
RHOA	30
AKAP6	21
TDP2	20
CLEC3B	18
ZNF331	12
STAU2	10
SLC5A3	10
KYAT3	10
KHDRBS1	10
NOP16	10
MXI1	7
KDM5B	4
SELENOF	3
RASL11B	3
MOXD1	3
KIFBP	3
DET1	3
IFI27	2
CIDEA	2
TINAGL1	1
ANXA9	1
TMEM121	1
CDH10	1
WDR1	1

grade 2 and grade 3. IFI27, OASL, and STAT1 was unique in grade 0; SELENOF was unique in grade 2 and AP2M1 and TMSB4X was unique in grade 3.

2.4 DISCUSSION

TNBC depends upon the expression levels of ER, PR and HER2 [39]. It's majorly the genetic alterations that cause all of the cancers, TP53 being somatic alteration for TNBC followed by PIK3CA. PTEN, KMT2C,andRB1 are the other genetic mutations [14]. Amplification of some genes like MYC, PIK3CA, CDK6 [6] and CNV alterations like EGFR and FGFR2 are the genetic basis of TNBC [70]. Like TNBC and other cancers, OC has genetic basis with vague symptoms that make it silent killer [76]. Alterations in BRCA1, BRCA2 and MMR gene are the major genetic causative risk factors of OC [76]. Some of the studies like [80] clearly explain the link between OC and BC that family history of one cancer increase the risk of other cancer. Other than that, there are multiple risk factors that are common in TNBC and OC. For

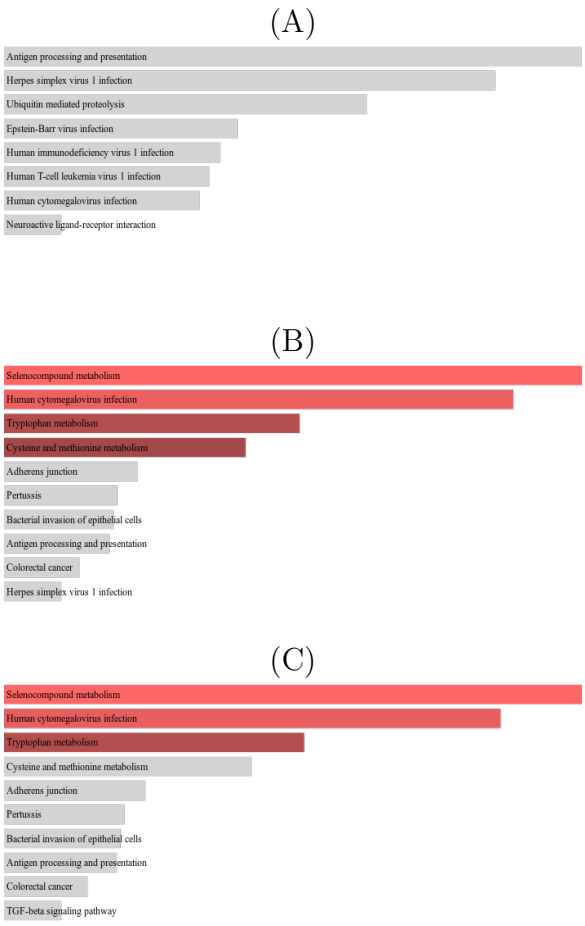
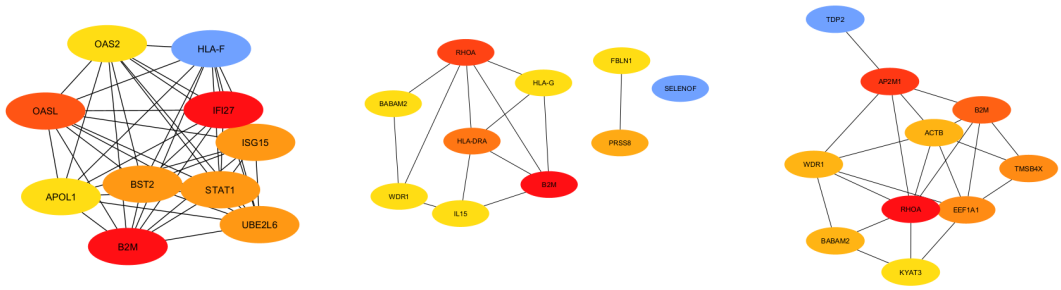


Figure 2.7: KEGG 2021 Human Pathways bar graphs for (A) Grade 0, (B) Grade 2, and (C) Grade 3.

example BRCA1 and BRCA2 mutations are the significant causative agents for both the cancers.

As, both the cancers are sharing commonalities between them and increase the risk of one another, there are still many more to explore. Whole genome transcriptomic analysis has been used to find the potential therapeutic targets, and to explore many diseases including cancers. For early diagnosis it is necessary to identify the sharing potential biomarkers of both the cancers as therapeutic targets. Our study, utilized bioinformatics approach on microarray data of patients of TNBC and OC to find common biomarkers between OC and TNBC. The study categorized the TNBC dataset according to the cancer grading system to facilitate a comprehensive analysis. It find 4 common hub genes between TNBC grade 0 and OC, 3 for grade 2 and 3 for grade 3. These were the most significant and top common hub genes.

The methodology we applied has been in use for the identification of various potential biomarkers among a variety of genetic diseases. As, [56] used the same methodology with the addition of Machine Learning validation on COVID-19 and HFRS. They



(a) Grade 0 (b) Grade 2 (c) Grade 3

Figure 2.8: Network degree analysis for the top 10 nodes in (a) Grade 0, (b) Grade 2, and (c) Grade 3.

identify 6 common hub genes involved in both diseases and they can be taken for further analysis. [12] aimed to identify the key genes involved in TNBC. They took 116 diseased samples and 113 normal samples from The Cancer Genome Atlas (TCGA) database. They performed weighted gene co-expression network analysis (WGCNA) as well on the data. On performing GO and KEGG, they analyzed that 147 genes were enriched in nuclear division, chromosomal region, ATPase activity, and cell cycle signaling. After PPI network construction, total of 15 hub genes were selected. In the study [93], 361 genes were found that were involved in oncogene-induced cell senescence, cyclin B1-CDK1 complex, protein kinase A catalytic subunit binding, cell cycle, and p53 signaling pathway. 10 hub genes were further identified out of these 361 genes. Then the overall survival and progression-free survival (PFS) of these 10 hub genes were evaluated in the Kaplan-Meier-plotter database. The individual studies were present that were aim to investigate the DEGs relevant to prognosis of OC and TNBC. However, a comparative approach to study two cancers, TNBC and OC, side by side was missing. Our study filled this gap and utilizes the opportunity to find common therapeutic targets for TNBC and OC.

There were total of 8 hub genes in all three grades. B2M gene was common in grade 0 and grade 2. Studies like [89], using SEREX analysis identified B2M gene as the potential target molecules of functional antibodies. It also tells that the expression of B2M was down-regulated by TGF-beta1 in OC cells. Beta-2-microglobulin (B2M) is a housekeeping protein and light chain subunit of the major histocompatibility complex (MHC) class I complex. It has a clear role in tumor immunity. Being a pleiotropic signaling molecule, it moduleate epithelial-to-mesenchymal transition (EMT) through iron-responsive pathways [55]. [77] discussed the expression of B2M in normal vs cancerous ovarian cells. High levels of B2M is detected in breast and Ovarian cancer both, but not specifically for TNBC [37]. TGF-beta1 has a clear prognostic value in TNBC [94]. The study explains that high expression of cytoplasmic

TGF-beta1 was significantly associated with higher histologic tumor grade and lymph node status of TNBC is consistant with several prior single center retrospective studies. Our KEGG pathways analysis of TNBC grade 2 shows the effected beta-signaling pathways. Moreover, in OC, B2M is regulated by TGF-beta signaling pathway. This concluded that B2M has a clear role in causing TNBC while it is already reported as a potential biomarker for OC. Talking about RHOA gene which was present in TNBC grade 2 and grade 3 hub genes, a RhoA-dependent, proteolytic-independent invasion mechanism is exploit by TMX2-28 breast cancer cells. Targeting the RhoA pathway in triple-negative, basal-like breast cancers that have a proteolytic-independent invasion mechanism may provide therapeutic strategies for the treatment of patients with increased risk of metastasis [19]. Increased level of Rhoa gene expression is also reported in advanced ovarian carcinomas [32].

KEGG pathway analysis of common genes resulted that the effected pathways include the crucial immune system process as antigen processing and presentation pathway was common in all the grades. This immune system process is reported as affected pathway in both TNBC [61] and OC [24]. B2M gene has a direct link with process as antigen processing and presentation pathway [20]. KEGG adherens junction pathway explains how cell adhesion and communication are important for tissues and their development. The main function of GO binding is to highlight how these protein connections are involved with the formation and signaling of junctions in the body. RHOA was discovered to be involved in adherens junction (KEGG hsa04530) and bind (GO:0005488), thus impacting how cells move through their cytoskeleton and stick to other cells. While AP2M1, TMSB4X, B2M, OASL and STAT1 are involved in binding, SELENOF seems to be part of the process as well and RHOA is the only gene directly supporting the adherens junction pathways.

Concluding our study, it uses a integrative bioinformatics' approach to identify the common pathways and genes involved in TNBC and OC. Further analysis on the identified genes can be carried out for their conformation as a potential therapeutic biomarkers. Some other factors, such as clinical trials, animal studies, and epidemiological data, also need to be considered before any therapies can be developed and implemented. Additionally, the differences and similarities between TNBC and OC in terms of genetic factors, clinical manifestations, and disease progression need to be further investigated to develop targeted and effective treatments.

Bibliography

[can] Cancer of the Ovary - Cancer Stat Facts — seer.cancer.gov. https://seer.cancer.gov/statfacts/html/ovary.html. [Accessed 12-05-2025].

[icc] iccr-cancer.org. https://www.iccr-cancer.org/docs/ICCR-Invasive-HistoTG.pdf. [Accessed 08-05-2025].

[3] Al-Moghrabi, N., Al-Showimi, M., Alqahtani, A., Almalik, O., Alhusaini, H., Almalki, G., Saad, A., and Alsunayi, E. (2024). Constitutional brca1 and mgmt methylation are significant risk factors for triple-negative breast cancer and high-grade serous ovarian cancer in saudi women. *International Journal of Molecular Sciences*, 25(6):3108.

[4] Ali, A. T., Al-Ani, O., and Al-Ani, F. (2023). Epidemiology and risk factors for ovarian cancer. *Menopause Review/Przegląd Menopauzalny*, 22(2):93–104.

[5] Almugren, N. and Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE access*, 7:78533–78548.

[6] Bareche, Y., Venet, D., Ignatiadis, M., Aftimos, P., Piccart, M., Rothe, F., and Sotiriou, C. (2018). Unravelling triple-negative breast cancer molecular heterogeneity using an integrative multiomic analysis. *Annals of oncology*, 29(4):895–902.

[7] Bauer, K. R., Brown, M., Cress, R. D., Parise, C. A., and Caggiano, V. (2007). Descriptive analysis of estrogen receptor (er)-negative, progesterone receptor (pr)-negative, and her2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the california cancer registry. *cancer*, 109(9):1721–1728.

[8] Begg, C. B., Rice, M. S., Zabor, E. C., and Tworoger, S. S. (2017). Examining the common aetiology of serous ovarian cancers and basal-like breast cancers using double primaries. *British Journal of Cancer*, 116(8):1088–1091.

[9] Bhandari, N., Walambe, R., Kotecha, K., and Khare, S. P. (2022). A comprehensive survey on computational learning methods for analysis of gene expression data. *Frontiers in Molecular Biosciences*, 9:907150.

[10] Bonome, T., Levine, D. A., Shih, J., Randonovich, M., Pise-Masison, C. A., Bogomolniy, F., Ozbun, L., Brady, J., Barrett, J. C., Boyd, J., et al. (2008). A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer research*, 68(13):5478–5486.

[11] Brooks, P. J. and Zakhari, S. (2013). Moderate alcohol consumption and breast cancer in women: from epidemiology to mechanisms and interventions. *Alcoholism: Clinical and Experimental Research*, 37(1):23–30.

[12] Chen, D.-L., Cai, J.-H., and Wang, C. C. (2022). Identification of key prognostic genes of triple negative breast cancer by lasso-based machine learning and bioinformatics analysis. *Genes*, 13(5):902.

[13] Derakhshan, F. and Reis-Filho, J. S. (2022a). Pathogenesis of triple-negative breast cancer. *Annu. Rev. Pathol.*, 17(1):181–204.

[14] Derakhshan, F. and Reis-Filho, J. S. (2022b). Pathogenesis of triple-negative breast cancer. *Annual Review of Pathology: Mechanisms of Disease*, 17(1):181–204.

[15] Dhote, Y., Agrawal, S., and Deen, A. J. (2015). A survey on feature selection techniques for internet traffic classification. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 1375–1380. IEEE.

[16] Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing illumina microarray. *Bioinformatics*, 24(13):1547–1548.

[17] Eckstein, N. (2011). Platinum resistance in breast and ovarian cancer cell lines. *Journal of Experimental & Clinical Cancer Research*, 30:1–11.

[18] El-Arabey, A. A. and Abdalla, M. (2022). The role of gata3 in the metastasis of triple-negative breast cancer and high-grade serous ovarian cancer. *Human cell*, 35(4):1298–1300.

[19] Fagan-Solis, K. D., Schneider, S. S., Pentecost, B. T., Bentley, B. A., Otis, C. N., Gierthy, J. F., and Arcaro, K. F. (2013). The rhoa pathway mediates mmp-2 and mmp-9-independent invasive behavior in a triple-negative breast cancer cell line. *Journal of cellular biochemistry*, 114(6):1385–1394.

[20] Gettinger, S., Choi, J., Hastings, K., Truini, A., Datar, I., Sowell, R., Wurtz, A., Dong, W., Cai, G., Melnick, M. A., et al. (2017). Impaired hla class i antigen processing and presentation as a mechanism of acquired resistance to immune checkpoint inhibitors in lung cancer. *Cancer discovery*, 7(12):1420–1435.

[21] Geyer, F. C., Pareja, F., Weigelt, B., Rakha, E., Ellis, I. O., Schnitt, S. J., and Reis-Filho, J. S. (2017). The spectrum of triple-negative breast disease: high-and low-grade lesions. *The American journal of pathology*, 187(10):2139–2151.

[22] Gulbahce, H. E., Bernard, P. S., Weltzien, E. K., Factor, R. E., Kushi, L. H., Caan, B. J., and Sweeney, C. (2018). Differences in molecular features of triple-negative breast cancers based on the age at diagnosis. *Cancer*, 124(24):4676–4684.

[23] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46:389–422.

[24] Han, L. Y., Fletcher, M. S., Urbauer, D. L., Mueller, P., Landen, C. N., Kamat, A. A., Lin, Y. G., Merritt, W. M., Spannuth, W. A., Deavers, M. T., et al. (2008). Hla class i antigen processing machinery component expression and intratumoral t-cell infiltrate as independent prognostic markers in ovarian carcinoma. *Clinical cancer research*, 14(11):3372–3379.

[25] Hanahan, D. (2022). Hallmarks of cancer: new dimensions. *Cancer discovery*, 12(1):31–46.

[26] Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *cell*, 100(1):57–70.

[27] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

[28] Hanna, L. and Adams, M. (2006). Prevention of ovarian cancer. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 20(2):339–362.

[29] Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in rna-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216.

[30] Henderson, B. E. and Feigelson, H. S. (2000). Hormonal carcinogenesis. *Carcinogenesis*, 21(3):427–433.

[31] Hirshaut, Y. and Pressman, P. (2009). *Breast cancer: The complete guide*. Bantam.

[32] Horiuchi, A., Imai, T., Wang, C., Ohira, S., Feng, Y., Nikaido, T., and Konishi, I. (2003). Up-regulation of small gtpases, rhoa and rhoc, is associated with tumor progression in ovarian carcinoma. *Laboratory investigation*, 83(6):861–870.

[33] Howard, F. M. and Olopade, O. I. (2021). Epidemiology of triple-negative breast cancer: a review. *The Cancer Journal*, 27(1):8–16.

[34] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

[35] Kakati, T., Bhattacharyya, D. K., Barah, P., and Kalita, J. K. (2019). Comparison of methods for differential co-expression analysis for disease biomarker prediction. *Computers in biology and medicine*, 113:103380.

[36] Kim, J. H. (2019). Advanced microarray data analysis. In *Genome Data Analysis*, pages 79–93. Springer Singapore, Singapore.

[37] Klein, B., Levin, I., Lurie, H., Nyska, A., Shapira, J., Kravits, M., and Klein, T. (1994). Elevated pretreatment levels of soluble cd8 and beta-2-microglobulin as indicators of relapse in breast-cancer patients. *International Journal of Oncology*, 4(2):471–474.

[38] Kumar, N., Ehsan, S., Banerjee, S., Fernandez Perez, C., Lhuilier, I., Neuner, J., Friebel-Klingner, T., Fayanju, O. M., Nair, B., Niinuma, S. A., et al. (2024). The unique risk factor profile of triple-negative breast cancer: a comprehensive meta-analysis. *JNCI: Journal of the National Cancer Institute*, 116(8):1210–1219.

[39] Kumar, P. and Aggarwal, R. (2016). An overview of triple-negative breast cancer. *Archives of gynecology and obstetrics*, 293:247–269.

[40] Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., Pietenpol, J. A., et al. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation*, 121(7):2750–2767.

[41] Li, J., Guo, H., Lou, Q., Zeng, Y., Guo, Z., Xu, P., Gu, Y., Gao, S., Xu, B., Han, S., et al. (2025). Natural variation of indels in the ctb3 promoter confers cold tolerance in japonica rice. *Nature Communications*, 16(1):1613.

[42] Li, K., Hüsing, A., Fortner, R. T., Tjønneland, A., Hansen, L., Dossus, L., Chang-Claude, J., Bergmann, M., Steffen, A., Bamia, C., et al. (2015). An epidemiologic risk prediction model for ovarian cancer in europe: the epic study. *British journal of cancer*, 112(7):1257–1265.

[43] Lin, C.-H., Yap, Y. S., Lee, K.-H., Im, S.-A., Naito, Y., Yeo, W., Ueno, T., Kwong, A., Li, H., Huang, S.-M., et al. (2019). Contrasting epidemiology and clinicopathology of female breast cancer in asians vs the us population. *JNCI: Journal of the National Cancer Institute*, 111(12):1298–1306.

[44] Lindholm Carlström, E., Niazi, A., Etemadikhah, M., Halvardson, J., Enroth, S., Stockmeier, C. A., Rajkowska, G., Nilsson, B., and Feuk, L. (2021). Transcriptome analysis of post-mortem brain tissue reveals up-regulation of the complement cascade in a subgroup of schizophrenia patients. *Genes*, 12(8):1242.

[45] Liu, S., Wang, Z., Zhu, R., Wang, F., Cheng, Y., and Liu, Y. (2021). Three differential expression analysis methods for rna sequencing: limma, edger, deseq2. *Journal of Visualized Experiments (JoVE)*, (175):e62528.

[46] Liu, Y.-R., Jiang, Y.-Z., Xu, X.-E., Hu, X., Yu, K.-D., and Shao, Z.-M. (2016). Comprehensive transcriptome profiling reveals multigene signatures in triple-negative breast cancer. *Clinical cancer research*, 22(7):1653–1662.

[47] Loizides, S. and Constantinidou, A. (2023). Triple negative breast cancer: Immunogenicity, tumor microenvironment, and immunotherapy. *Frontiers in genetics*, 13:1095839.

[48] Luo, J., Craver, A., Moore, K., Stepniak, L., King, J., Herbert, J., and Aschebrook-Kilfoy, B. (2022). Etiology of breast cancer: a perspective from epidemiologic studies. *Journal of the National Cancer Center*, 2(4):195–197.

[49] Manoharan, S. and Pugalendhi, P. (2010). Breast cancer: An overview. *Journal of Cell & Tissue Research*, 10(3).

[50] McDermaid, A., Monier, B., Zhao, J., Liu, B., and Ma, Q. (2019). Interpretation of differential gene expression results of rna-seq data: review and integration. *Briefings in bioinformatics*, 20(6):2044–2054.

[51] Millikan, R. C., Newman, B., Tse, C.-K., Moorman, P. G., Conway, K., Smith, L. V., Labbok, M. H., Geradts, J., Bensen, J. T., Jackson, S., et al. (2008). Epidemiology of basal-like breast cancer. *Breast cancer research and treatment*, 109:123–139.

[52] Moreno-Bueno, G., Gamallo, C., Pérez-Gallego, L., de Mora, J. C., Suárez, A., and Palacios, J. (2001). β-catenin expression pattern, β-catenin gene mutations, and microsatellite instability in endometrioid ovarian carcinomas and synchronous endometrial carcinomas. *Diagnostic molecular pathology*, 10(2):116–122.

[53] Munsell, M. F., Sprague, B. L., Berry, D. A., Chisholm, G., and Trentham-Dietz, A. (2014). Body mass index and breast cancer risk according to postmenopausal estrogen-progestin use and hormone receptor status. *Epidemiologic reviews*, 36(1):114–136.

[54] Nardin, S., Mora, E., Varughese, F. M., D'Avanzo, F., Vachanaram, A. R., Rossi, V., Saggia, C., Rubinelli, S., and Gennari, A. (2020). Breast cancer survivorship, quality of life, and late toxicities. *Frontiers in oncology*, 10:864.

[55] Nomura, T., Huang, W.-C., E. Zhau, H., Josson, S., Mimata, H., and WK Chung, L. (2014). β2-microglobulin-mediated signaling as a target for cancer therapy. *Anti-Cancer Agents in Medicinal Chemistry-Anti-Cancer Agents)*, 14(3):343–352.

[56] Noor, F., Ashfaq, U. A., Bakar, A., ul Haq, W., Allemailem, K. S., Alharbi, B. F., Al-Megrin, W. A. I., and Tahir ul Qamar, M. (2023). Discovering common pathogenic processes between covid-19 and hfrs by integrating rna-seq differential expression analysis with machine learning. *Frontiers in Microbiology*, 14:1175844.

[57] Nwosu, N. (2024). Cancer: A disease of modern times? *Cureus*, 16(11).

[58] Organization, W. H. et al. (2020). Global cancer observatory. international agency for research on cancer. *World Health Organization*.

[59] Orrantia-Borunda, E., Anchondo-Nuñez, P., Acuña-Aguilar, L. E., Gómez-Valles, F. O., and Ramírez-Valdespino, C. A. (2022). Subtypes of breast cancer. *Breast Cancer [Internet]*.

[60] Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., et al. (2007). Arrayexpress—a public database of microarray experiments and gene expression profiles. *Nucleic acids research*, 35(suppl_1):D747–D750.

[61] Pedersen, M. H., Hood, B. L., Beck, H. C., Conrads, T. P., Ditzel, H. J., and Leth-Larsen, R. (2017). Downregulation of antigen presentation-associated pathway proteins is linked to poor outcome in triple-negative breast cancer patient tumors. *Oncoimmunology*, 6(5):e1305531.

[62] Plasilova, M. L., Hayse, B., Killelea, B. K., Horowitz, N. R., Chagpar, A. B., and Lannin, D. R. (2016). Features of triple-negative breast cancer: Analysis of 38,813 cases from the national cancer database. *Medicine*, 95(35):e4614.

[63] Rakha, E. A., Tse, G. M., and Quinn, C. M. (2023). An update on the pathological classification of breast cancer. *Histopathology*, 82(1):5–16.

[64] Richie, R. C. and Swanson, J. O. (2003). Breast cancer: a review of the literature. *JOURNAL OF INSURANCE MEDICINE-NEW YORK THEN DENVER–*, 35(2):85–101.

[65] Rosen, D. G., Yang, G., Liu, G., Mercado-Uribe, I., Chang, B., Xiao, X. S., Zheng, J., Xue, F.-X., and Liu, J. (2009). Ovarian cancer: pathology, biology, and disease models. *Frontiers in bioscience: a journal and virtual library*, 14:2089.

[66] Schoemaker, M. J., Nichols, H. B., Wright, L. B., Brook, M. N., Jones, M. E., O'Brien, K. M., Adami, H.-O., Baglietto, L., Bernstein, L., Bertrand, K. A., et al. (2018). Association of body mass index and age with subsequent breast cancer risk in premenopausal women. *JAMA oncology*, 4(11):e181771–e181771.

[67] Serio, P. A. d. M. P., de Lima Pereira, G. F., Katayama, M. L. H., Roela, R. A., Maistro, S., and Folgueira, M. A. A. K. (2021). Somatic mutational profile of high-grade serous ovarian carcinoma and triple-negative breast carcinoma in young and elderly patients: similarities and divergences. *Cells*, 10(12):3586.

[68] Sharma, P. (2016). Biology and management of patients with triple-negative breast cancer. *The oncologist*, 21(9):1050–1062.

[69] Shimelis, H., LaDuca, H., Hu, C., Hart, S. N., Na, J., Thomas, A., Akinhanmi, M., Moore, R. M., Brauch, H., Cox, A., et al. (2018). Triple-negative breast cancer risk genes identified by multigene hereditary cancer panel testing. *JNCI: Journal of the National Cancer Institute*, 110(8):855–862.

[70] Shiu, K.-K., Natrajan, R., Geyer, F. C., Ashworth, A., and Reis-Filho, J. S. (2010). Dna amplifications in breast cancer: genotypic–phenotypic correlations. *Future Oncology*, 6(6):967–984.

[71] Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2022). Cancer statistics, 2022. *CA: a cancer journal for clinicians*, 72(1):7–33.

[72] Singh, H., Kumar, R., Singh, A. P., Malhotra, M., Rani, R., and Singh, A. P. (2024). Cancer: A review. *Int. J. Med. Phar. Drug Re*, 8:2.

[73] Smolarz, B., Nowak, A. Z., and Romanowicz, H. (2022). Breast cancer—epidemiology, classification, pathogenesis and treatment (review of literature). *Cancers*, 14(10):2569.

[74] Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14:1–18.

[75] Stark, A., Schultz, D., Kapke, A., Nadkarni, P., Burke, M., Linden, M., and Raju, U. (2009). Obesity and risk of the less commonly diagnosed subtypes of breast cancer. *European Journal of Surgical Oncology (EJSO)*, 35(9):928–935.

[76] Stewart, C., Ralyea, C., and Lockwood, S. (2019). Ovarian cancer: an integrated review. In *Seminars in oncology nursing*, volume 35, pages 151–156. Elsevier.

[77] Sun, W., Gui, L., Zuo, X., Zhang, L., Zhou, D., Duan, X., Ren, W., and Xu, G. (2016). Human epithelial-type ovarian tumour marker beta-2-microglobulin is regulated by the tgf-β signaling pathway. *Journal of translational medicine*, 14:1–13.

[78] Sung, H., DeSantis, C. E., Fedewa, S. A., Kantelhardt, E. J., and Jemal, A. (2019). Breast cancer subtypes among eastern-african–born black women and other black women in the united states. *Cancer*, 125(19):3401–3411.

[79] Sweeney, C., Blair, C. K., Anderson, K. E., Lazovich, D., and Folsom, A. R. (2004). Risk factors for breast cancer in elderly women. *American Journal of Epidemiology*, 160(9):868–875.

[80] Tschernichovsky, R. and Goodman, A. (2017). Risk-reducing strategies for ovarian cancer in brca mutation carriers: a balancing act. *The oncologist*, 22(4):450–459.

[81] Turner, M. C., Andersen, Z. J., Baccarelli, A., Diver, W. R., Gapstur, S. M., Pope III, C. A., Prada, D., Samet, J., Thurston, G., and Cohen, A. (2020). Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. *CA: a cancer journal for clinicians*, 70(6):460–479.

[82] Tyagi, V. and Mishra, A. (2013). A survey on different feature selection methods for microarray data analysis. *International Journal of Computer Applications*, 67(16).

[83] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *science*, 339(6127):1546–1558.

[84] Walker, J. L., Powell, C. B., Chen, L.-m., Carter, J., Bae Jump, V. L., Parker, L. P., Borowsky, M. E., and Gibb, R. K. (2015). Society of gynecologic oncology recommendations for the prevention of ovarian cancer. *Cancer*, 121(13):2108–2120.

[85] Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., et al. (2010). The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl_2):W214–W220.

[86] Wodrich, M. D., Sawatlon, B., Busch, M., and Corminboeuf, C. (2021). The genesis of molecular volcano plots. *Accounts of chemical research*, 54(5):1107–1117.

[87] Xu, Y., Cui, J., and Puett, D. (2014). *Cancer Classification and Molecular Signature Identification*, pages 65–87. Springer New York, New York, NY.

[88] Yamagiwa, K. and Ichikawa, K. (1918). Experimental study of the pathogenesis of carcinoma. *The Journal of Cancer Research*, 3(1):1–29.

[89] Yang, H.-S., Li, Y., Deng, H. X., and Peng, F. (2009). Identification of beta2-microglobulin as a potential target for ovarian cancer. *Cancer Biology & Therapy*, 8(24):2323–2328.

[90] Yang, X. R., Chang-Claude, J., Goode, E. L., Couch, F. J., Nevanlinna, H., Milne, R. L., Gaudet, M., Schmidt, M. K., Broeks, A., Cox, A., et al. (2011). Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the breast cancer association consortium studies. *Journal of the National Cancer Institute*, 103(3):250–263.

[91] Zagami, P. and Carey, L. A. (2022). Triple negative breast cancer: Pitfalls and progress. *NPJ breast cancer*, 8(1):95.

[92] Zamwar, U. M., Anjankar, A. P., and Anjankar, A. (2022). Aetiology, epidemiology, histopathology, classification, detailed evaluation, and treatment of ovarian cancer. *Cureus*, 14(10).

[93] Zhang, L., Sun, L., Zhang, B., and Chen, L. (2019). Identification of differentially expressed genes (degs) relevant to prognosis of ovarian cancer by use of integrated bioinformatics analysis and validation by immunohistochemistry assay. *Medical science monitor: international medical journal of experimental and clinical research*, 25:9902.

[94] Zhang, M., Wu, J., Mao, K., Deng, H., Yang, Y., Zhou, E., and Liu, J. (2017). Role of transforming growth factor-β1 in triple negative breast cancer patients. *International Journal of Surgery*, 45:72–76.

[95] Zhang, R., Siu, M. K., Ngan, H. Y., and Chan, K. K. (2022). Molecular biomarkers for the early detection of ovarian cancer. *International Journal of Molecular Sciences*, 23(19):12041.

[96] Zhang, Z., Zhang, R., and Li, D. (2023). Molecular biology mechanisms and emerging therapeutics of triple-negative breast cancer. *Biologics: Targets and Therapy*, pages 113–128.