INSIGHTS-JOURNAL OF
HEALTH AND REHABILITATION

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

COMPARATIVE ANALYSIS OF GENE EXPRESSION PROFILES ACROSS BREAST CANCER STAGES USING NEXT GENERATION SEQUENCING DATA

Original Research

Momna Quddus[1], Romaisa Saeed[1], Laiba Sultana[1], Naveen Kaleem[1], Aqsa Naz[1], Waseem Haider[2], Saadia Momal Zafar[2]*, Rafia Anwer[3]*
[1]Faculty of Rehabilitation and Allied Health Sciences (FRAHS), Riphah International University, Islamabad, Pakistan.
[2]COMSATS University Islamabad (CUI), Pakistan.
[3]Faculty of Rehabilitation and Allied Health Sciences (FRAHS), Riphah International University, Faisalabad, Pakistan.
Corresponding Author: Rafia Anwer, Faculty of Rehabilitation and Allied Health Sciences (FRAHS), Riphah International University Faisalabad, Pakistan Email: rafia.anwer@riphahfsd.edu.pk. Saadia Momal Zafar, COMSATS University Islamabad (CUI), Pakistan. Email: saadiamomal98@gmail.com
Acknowledgement: The authors acknowledge the GEO database for providing publicly accessible datasets used in this study.

Conflict of Interest: None

Grant Support & Financial Support: None

ABSTRACT

Background: Breast cancer remains the most frequently diagnosed malignancy among women worldwide, with rising incidence in developing regions including Pakistan. It is a biologically heterogeneous disease influenced by genetic, environmental, and hormonal factors, and is classified into molecular subtypes with distinct clinical behaviors. Advances in RNA sequencing and next-generation sequencing have enabled detailed transcriptomic profiling; however, comparative stage-wise gene expression analysis remains limited, restricting identification of reliable biomarkers and therapeutic targets.

Objective: To analyze differential gene expression patterns between normal breast tissue and tumor samples across stages, identify disrupted molecular pathways, and determine potential diagnostic biomarkers and therapeutic targets in breast cancer.

Methods: A retrospective in silico bioinformatics study was conducted using RNA-Seq datasets retrieved from the Gene Expression Omnibus database. Data preprocessing and normalization were performed, followed by differential expression analysis using R with thresholds of p-value <0.05 and $|\log2 \text{ fold change}| \geq 1.5$. Functional enrichment analyses including Gene Ontology and KEGG pathways were conducted. Protein-protein interaction networks were constructed using STRING and visualized in Cytoscape, where CytoHubba was applied to identify hub genes. Selected genes were validated using the GEPIA platform.

Results: A total of 10,565 genes were analyzed, from which 648 differentially expressed genes were identified. Among these, 387 genes were detected in ductal carcinoma in situ, including 227 upregulated and 160 downregulated genes, while 261 genes were identified in invasive ductal carcinoma, including 115 upregulated and 146 downregulated genes. Functional enrichment revealed significant involvement in cell cycle regulation, mitosis, and chromosome segregation. KEGG analysis highlighted enrichment in platinum drug resistance, TNF signaling, and antifolate resistance pathways. Network analysis identified 10 hub genes, with BRCA1, TOP2A, FOXM1, TYMS, and BIRC5 showing significant upregulation and strong interaction connectivity.

Conclusion: This study demonstrated distinct stage-associated gene expression alterations in breast cancer and identified key molecular pathways and hub genes with potential diagnostic and therapeutic relevance. These findings provide a foundation for future validation studies and support the development of personalized strategies for early detection and targeted treatment.

Keywords: Bioinformatics, Breast Neoplasms, Gene Expression Profiling, Molecular Targeted Therapy, RNA Sequencing, Transcriptome, Tumor Biomarkers

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

INTRODUCTION

Breast cancer is a complex and heterogeneous malignancy arising from the epithelial components of the breast, most commonly the ducts and lobules, within a structurally intricate organ composed of glandular tissue, stroma, vasculature, and lymphatics. Under normal physiological conditions, breast cells undergo tightly regulated cycles of proliferation, differentiation, and apoptosis; however, disruption of these mechanisms may lead to uncontrolled cellular growth and tumor formation. The biological behavior of breast cancer is further influenced by the tumor microenvironment, including stromal interactions and immune cell infiltration, which contribute to tumor progression and metastasis. Despite significant advances in screening, diagnosis, and treatment, breast cancer remains one of the leading causes of cancer-related morbidity and mortality among women worldwide, posing a persistent global health challenge(1, 2). Globally, breast cancer is now the most commonly diagnosed malignancy, accounting for approximately 2.3 million new cases and a substantial proportion of cancer-related deaths annually. The burden of disease continues to rise, with projections suggesting a marked increase in incidence and mortality by 2040. Notably, regional variations exist, with developing countries experiencing increasing incidence rates and limited access to early detection and treatment facilities. In Pakistan, breast cancer represents the most frequent malignancy among women, particularly in urban centers such as Karachi, where it constitutes a significant proportion of female cancers. These epidemiological trends underscore the urgent need for improved diagnostic, prognostic, and therapeutic strategies tailored to diverse populations(3, 4).

Breast cancer exhibits considerable biological diversity, which is reflected in its classification into molecular subtypes based on hormone receptor status and HER2 expression, including luminal A, luminal B, HER2-enriched, and triple-negative breast cancer. These subtypes differ in their clinical behavior, treatment response, and prognosis, highlighting the importance of accurate molecular characterization. Furthermore, disease staging, as defined by the AJCC system, integrates tumor size, nodal involvement, metastatic spread, and molecular biomarkers to provide a comprehensive assessment of disease severity and guide clinical decision-making. However, traditional staging approaches, while clinically useful, may not fully capture the underlying molecular complexity that drives tumor progression across different stages(5, 6). The etiology of breast cancer is multifactorial, involving a combination of genetic, hormonal, environmental, and lifestyle factors. Established risk factors include advancing age, reproductive history, hormonal exposure, obesity, alcohol consumption, and genetic mutations such as BRCA1 and BRCA2. Environmental exposures, including ionizing radiation, chemical carcinogens, and endocrine-disrupting compounds, have also been implicated in disease development. Despite extensive research, the precise mechanisms underlying tumor initiation and progression remain incompletely understood, particularly in relation to the dynamic changes that occur at the molecular level during different stages of the disease(7, 8).

In recent years, advances in high-throughput sequencing technologies, particularly next-generation sequencing (NGS), have revolutionized cancer research by enabling comprehensive analysis of genomic and transcriptomic alterations. NGS allows simultaneous examination of thousands of genes, facilitating the identification of differential gene expression patterns, novel mutations, and molecular pathways involved in tumor development and progression. In breast cancer, RNA sequencing (RNA-Seq) has been instrumental in uncovering tumor heterogeneity, identifying novel biomarkers, and elucidating mechanisms of treatment resistance. Moreover, integrative approaches combining bulk and single-cell sequencing data have provided deeper insights into tumor biology, immune interactions, and metastatic processes(9, 10). Despite these advancements, a critical gap persists in understanding how gene expression profiles vary systematically across different stages of breast cancer and how these variations can be leveraged to improve diagnosis, prognostication, and personalized treatment strategies. Most existing studies focus on specific subtypes or isolated molecular markers, while comprehensive stage-wise comparative analyses remain limited. Addressing this gap is essential for identifying stage-specific biomarkers and therapeutic targets that can enhance clinical outcomes(11, 12).

Therefore, the present study aims to perform a comparative analysis of gene expression profiles across different stages of breast cancer using next-generation sequencing data. It seeks to evaluate differential gene expression patterns, identify key regulatory and hub genes through network analysis, and explore potential biomarkers associated with disease progression. It is hypothesized that significant differences exist in gene expression profiles across breast cancer stages, reflecting underlying biological changes that may have diagnostic and therapeutic relevance, while the null hypothesis assumes no such differences between stages.

METHODS

This study was conducted as a retrospective, comparative, in silico investigation designed to evaluate differences in gene expression profiles between breast cancer samples representing different disease stages and normal breast tissue samples. A case-control analytical framework was adopted, in which normal breast samples served as controls and breast cancer samples served as cases. The overall methodological approach relied on publicly available high-throughput transcriptomic data and integrated bioinformatics analyses to

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

explore differential gene expression, functional enrichment, interaction networks, and potential biomarkers associated with breast cancer progression(13). The study setting was entirely database-based, and the dataset was obtained from the Gene Expression Omnibus (GEO), a publicly accessible repository for gene expression studies. The RNA sequencing dataset used in this work was retrieved from GEO accession number GSE228582. Publicly archived datasets were selected because they provide standardized, reusable transcriptomic data and permit reproducible large-scale molecular analyses without direct patient recruitment or laboratory-based sample processing. The use of an established international repository also strengthened the transparency and traceability of the data source(14).

The study population consisted of samples derived from human breast tissue available within the selected dataset. Samples were considered eligible for analysis if they met the predefined inclusion criteria, namely: human origin, confirmed association with breast cancer or normal breast tissue comparison groups, and availability of RNA-Seq-based gene expression data. Samples were excluded if they were non-human in origin, unrelated to breast cancer, or contained incomplete or unusable gene expression information. Only samples with sufficient expression data and relevant clinical grouping for comparative analysis were retained. Because the study used secondary, de-identified, publicly available data, no direct interaction with participants occurred and no additional biological specimens were collected from patients(15). The sample size was determined by the number of eligible samples available in the selected GEO dataset after screening according to the inclusion and exclusion criteria. Thus, the study did not employ conventional sample size estimation based on power calculations, as the analysis depended on pre-existing online data. Instead, all eligible samples within the dataset were included to maximize the robustness of the comparative transcriptomic analysis. This approach is commonly used in bioinformatics studies based on archived sequencing data, where the available dataset defines the analytic sample(16).

Data collection involved retrieval of processed or count-based RNA-Seq expression data from the selected GEO series. After downloading the dataset, the expression matrix was reviewed and organized for downstream analysis. Initial preprocessing was performed to improve data quality and analytic reliability. Genes with missing values, duplicated entries, or consistently absent expression across the compared groups were removed. Low-information features with negligible expression across samples were also filtered out to reduce noise and improve the sensitivity of differential expression testing. Spreadsheet-based screening was initially used for basic inspection and cleaning, after which formal statistical analysis was carried out in the R environment(17). Differential gene expression analysis was performed using the edgeR package in R, which is specifically designed for count-based RNA-Seq data. Raw or imported count data were first structured into a digital gene expression object and then normalized to account for variation in library size and sequencing depth across samples. Following normalization, statistical modeling was undertaken to compare gene expression levels between normal samples and breast cancer samples across the relevant disease stages. Genes were considered significantly differentially expressed if they met the predefined thresholds of an adjusted p-value below 0.05 and an absolute log fold change greater than 1. Genes with positive log fold change values above this threshold were interpreted as upregulated, whereas genes with negative log fold change values below -1 were interpreted as downregulated. This analytical framework allowed identification of genes that showed meaningful transcriptional changes associated with breast cancer development and stage-wise progression(18).

To better understand the biological significance of the differentially expressed genes, functional enrichment analysis was subsequently undertaken. Gene Ontology (GO) analysis was used to examine enrichment across biological processes, molecular functions, and cellular components, while Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was used to identify the signaling and metabolic pathways in which the significant genes were involved. These analyses were intended to clarify the broader biological context of the detected transcriptional alterations and to highlight molecular mechanisms potentially involved in tumor initiation, progression, invasion, and stage-dependent behavior. By linking differentially expressed genes to functionally enriched pathways, the study aimed to move beyond gene lists and provide more meaningful biological interpretation(19). To explore the interaction patterns among the significant genes, protein-protein interaction network analysis was performed using STRING database-derived interaction information. The resulting interaction data were imported into Cytoscape for network visualization and further topological analysis. Cytoscape was selected because it provides an effective platform for identifying molecular connectivity patterns and graphically examining biologically relevant interactions. Within Cytoscape, the CytoHubba plugin was used to identify hub genes, defined as highly connected genes that may play central regulatory roles within the interaction network. These hub genes were considered candidates of particular biological and potential clinical relevance because central network position is often associated with essential functional importance in disease pathways(20).

To further assess the relevance of the identified hub genes, validation was performed using the Gene Expression Profiling Interactive Analysis (GEPIA) platform. This step enabled external evaluation of gene expression patterns and supported assessment of their consistency in breast cancer-related datasets. Validation was included to strengthen confidence in the final candidate genes and to determine whether the observed expression trends were biologically plausible and potentially clinically meaningful. Through this multistep strategy, the study sought to identify robust biomarkers that may contribute to better understanding of breast cancer biology and may guide future diagnostic or therapeutic research(21). Statistical significance throughout the differential expression workflow

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

was interpreted using adjusted p-values in order to reduce false-positive findings related to multiple testing. Wherever applicable, corrected significance thresholds were preferred over unadjusted values. The analysis therefore combined quantitative statistical testing with bioinformatics-based functional interpretation and network prioritization. This layered approach increased methodological rigor by integrating expression analysis, biological annotation, interaction mapping, and external validation into a coherent analytic pipeline(22).

With regard to ethical considerations, the study used secondary data from a publicly accessible repository containing de-identified molecular datasets. As no direct patient contact, recruitment, intervention, or collection of identifiable personal information occurred, individual informed consent was not obtained specifically for this analysis. No institutional review board or ethical committee approval number was provided in the available study materials. In such circumstances, studies based exclusively on anonymized public-domain datasets are often considered exempt from formal ethical approval; however, final exemption status depends on the policy of the host institution. Therefore, it would be methodologically appropriate to state that the analysis was conducted using publicly available de-identified data and that formal ethical approval was not required or was waived, if confirmed by the relevant institutional policy. If this work was submitted to an academic institution, the authors should verify whether an exemption letter or ethical waiver reference is needed for documentation(23). Overall, the methodology combined retrospective transcriptomic analysis with standardized computational tools to characterize stage-related gene expression changes in breast cancer. By integrating preprocessing, differential expression testing, enrichment analysis, interaction network construction, hub gene identification, and external validation, the study provided a structured framework for identifying biologically meaningful genes and pathways associated with breast cancer progression.

RESULTS

The analysis was conducted to identify differentially expressed genes associated with breast cancer by comparing normal breast samples with tumor samples and by examining the functional pathways linked to the dysregulated genes. The RNA-Seq dataset retrieved from the GEO repository contained expression count data for 10,565 genes. Following screening and study-specific filtering, 60 genes were shortlisted for focused downstream analysis. From these, a smaller subset of 10 genes was finally highlighted during the selection process for detailed assessment. Normal breast samples were used as controls, while tumor samples included both ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC) groups. After data retrieval, differential expression analysis was performed in R using standard RNA-Seq analytical criteria, including log fold change, log counts per million, p value, and false discovery rate. The normalized count distributions of normal and tumor samples were first examined to assess comparability across groups. The box plot profiles demonstrated median-centered distributions across samples, indicating that the expression data had been normalized and were suitable for comparative analysis. A similar pattern was observed when normal samples were compared separately with diseased groups and when normal samples were compared with combined tumor samples.

Density plots of log-transformed expression values showed substantial overlap between normal and tumor samples in both stage-specific and combined comparisons. The expression curves followed broadly comparable distributions across groups, with only modest variation in peak width and height. These findings confirmed that the normalized data were of acceptable quality for downstream differential expression analysis and did not show marked outlier-driven distortion in distribution patterns. The volcano plot analysis demonstrated distinct groups of significantly upregulated and downregulated genes when tumor samples were compared with normal breast tissue. Genes with positive log2 fold change and significant p values were classified as upregulated, whereas genes with negative log2 fold change and significant p values were classified as downregulated. Genes without statistically significant expression differences remained clustered around the nonsignificant region of the plot.

The heatmap-based differential expression analysis showed clear transcriptional differences between normal and malignant samples. In the DCIS comparison, a total of 387 differentially expressed genes were identified. Among these, 227 genes were upregulated and 160 genes were downregulated relative to normal breast tissue. In the IDC comparison, a total of 261 differentially expressed genes were identified, of which 115 were upregulated and 146 were downregulated compared with normal samples. These expression patterns showed that both non-invasive and invasive breast lesions exhibited substantial transcriptional alterations relative to normal tissue, with distinct distributions of upregulated and downregulated genes in each disease stage. Stage-related gene distribution further showed that 60 genes were involved across the two breast cancer stages examined. Of these, 28 genes were associated with DCIS and 32 genes were associated with IDC. The combined heatmap of these genes demonstrated differential clustering across the disease groups, reflecting variation in gene expression profiles between the non-invasive and invasive stages.

Functional enrichment analysis of the differentially expressed genes identified significant Gene Ontology categories across biological processes, molecular functions, and cellular components. The enriched biological terms were predominantly related to cell division, mitosis, nuclear division, and chromosome segregation. The bar-based GO visualization demonstrated that the most enriched terms had higher gene ratios and lower adjusted p values, indicating a greater concentration of differentially expressed genes within these cellular

and proliferative functions. Additional enrichment analysis using Enrichr demonstrated that the dysregulated genes were linked to biologically relevant breast cancer-associated functions. The enriched gene sets were mainly related to cell proliferation, metastatic behavior, and hormone-associated signaling pathways. Across both disease stages, the total panel of 60 selected genes contributed to the enriched signatures identified in the functional annotation output.

KEGG pathway analysis further showed that the differentially expressed genes were significantly enriched in several biologically relevant pathways. The top enriched pathways included platinum drug resistance and antifolate resistance, among others. Pathways with a higher number of mapped genes showed greater overlap with the input dataset, while pathways with lower adjusted p values demonstrated stronger statistical enrichment. These pathway-level findings indicated that the identified genes were distributed across multiple cancer-relevant molecular networks. Protein-protein interaction network analysis was then performed using STRING to examine the connectivity of the differentially expressed genes. The network construction showed that several genes occupied highly connected central positions within the interaction map. Subsequent Cytoscape analysis with the CytoHubba plugin identified the top 10 hub genes as BRCA1, HMMR, MKI67, TYMS, BIRC5, TOP2A, FOXM1, AURKA, EXO1, and UBE2C. These genes were identified as upregulated in breast tumor tissue compared with normal tissue and represented the most connected nodes within the network structure.

Validation analysis was then undertaken for selected hub genes using the GEPIA platform. Expression profiling confirmed increased expression of BRCA1, TYMS, BIRC5, TOP2A, and FOXM1 in tumor tissue compared with normal breast tissue. In the validation plots, tumor samples consistently showed higher expression levels than normal controls, supporting the reproducibility of the upregulated hub gene profile identified in the primary analysis. Overall, the results demonstrated that breast cancer samples exhibited substantial differential gene expression relative to normal breast tissue, with 387 differentially expressed genes identified in DCIS and 261 in IDC. Among the shortlisted genes, 28 were linked with DCIS and 32 with IDC. Functional annotation showed enrichment in cell cycle- and mitosis-related processes, pathway analysis identified significant cancer-associated molecular pathways, and network analysis highlighted 10 highly connected hub genes, several of which were subsequently validated as overexpressed in tumor tissue.

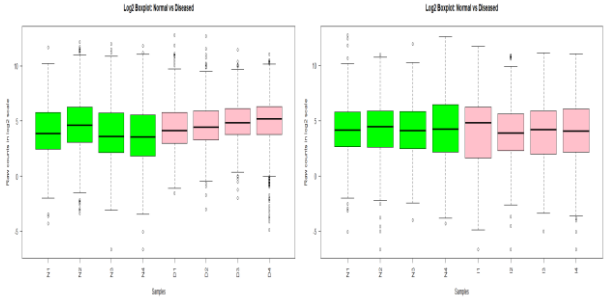


Figure 4.2.1: Boxplot shows normalization of samples (Normal vs DCIS)

Figure 4.2.2: Boxplot shows normalization of samples (Normal vs IDC)

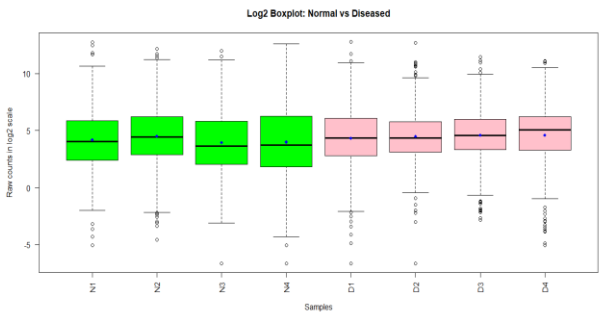


Figure 4.2.3: Box plot shows normalization of samples (Normal vs Combined)

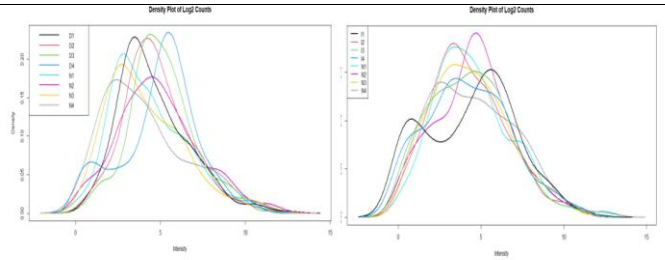


Figure 4.3.1: Density plot shows normalization of samples (Normal vs DCIS)

Figure 4.3.2: Density plot shows normalization of samples (Normal vs IDC)

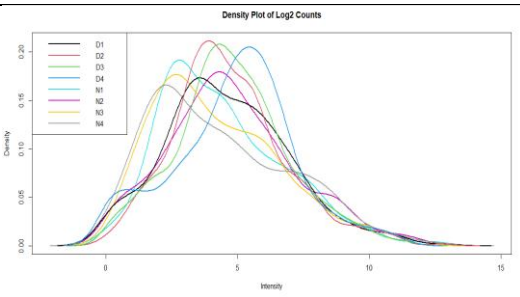


Figure 4.3.3: Density plot shows normalization of samples (Normal vs Tumor)

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

Figure 4.4.1: Volcano plot DCIS

Figure 4.4.2: Volcano plot of IDC

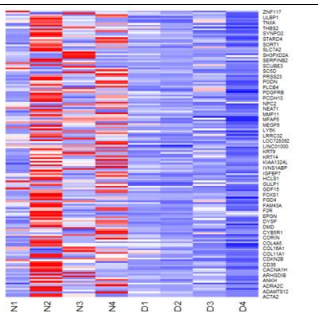


Figure 4.5.1.1: Heatmap of downregulated genes in DCIS

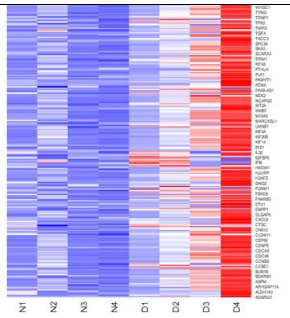


Figure 4.5.1.2: Heatmap of upregulated genes in DCIS

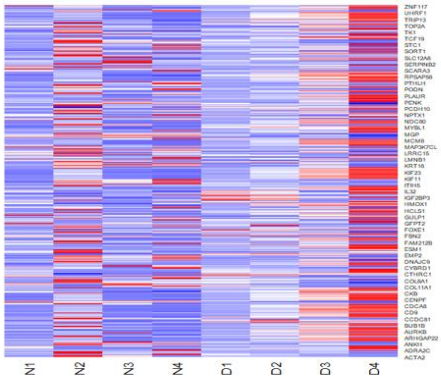


Figure 4.5.1.3: Combined Heatmap of Upregulated and Downregulated genes in DCIS

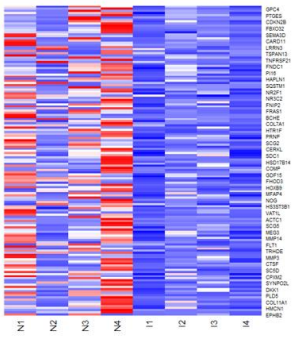


Figure 4.5.2.1: Heatmap of downregulated genes in IDC

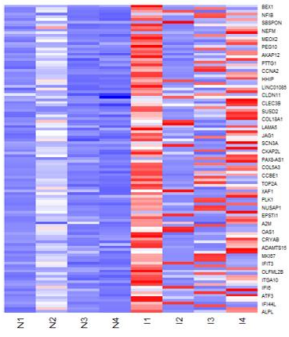


Figure 4.5.2.2: Heatmap of upregulated genes in IDC

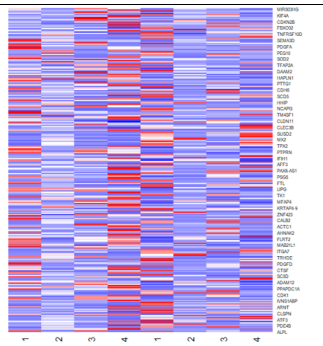


Figure 4.5.2.3: Combined Heatmap of Upregulated and Downregulated genes in IDC

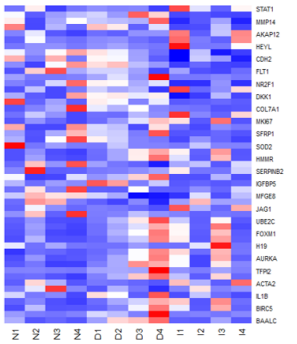


Figure 4.5.3.1: Detailed heatmap of combined genes

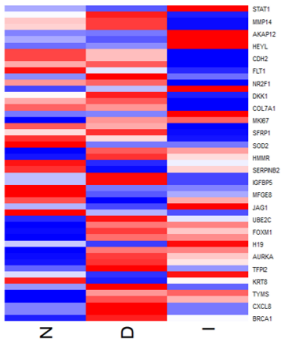


Figure 4.5.3.2: Heatmap of combined genes

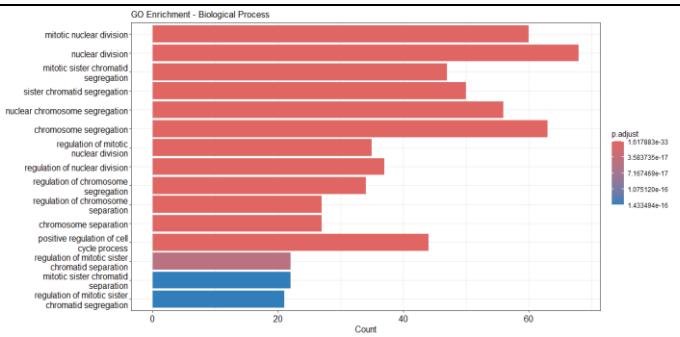


Figure 4.6.1.1: Box plot of Go term analysis DEGs by R

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

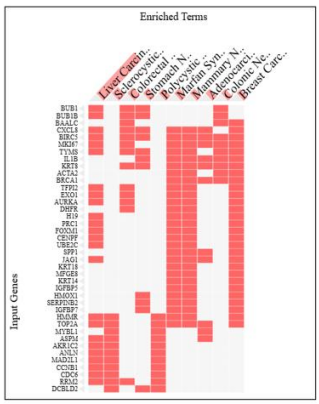


Figure 4.7.1.1: Identification of specific Breast Cancer genes from DEGs in DCIS

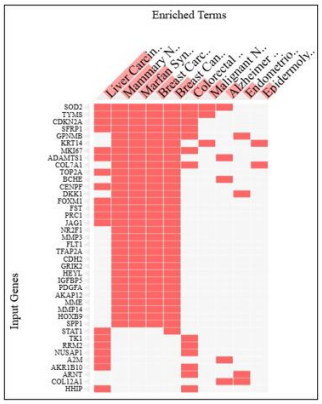


Figure 4.7.1.2: Identification of specific Breast Cancer genes from DEGs in IDC

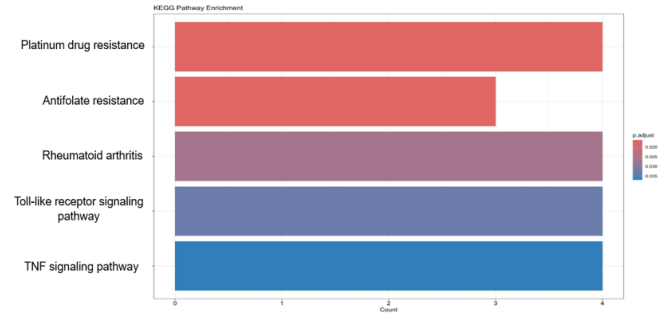


Figure 4.8.1.1: Box plot of KEGG Pathway analysis of DEGs by R

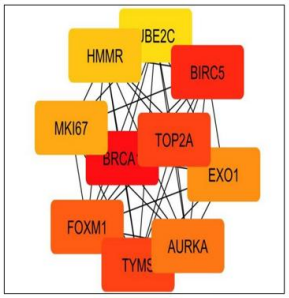


Figure 4.10.1: Identification of Top 10 hub genes using Cytohubba.

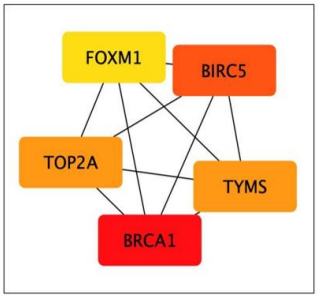


Figure 4.10.2: Identification of Top 5 hub genes using Cytohubba

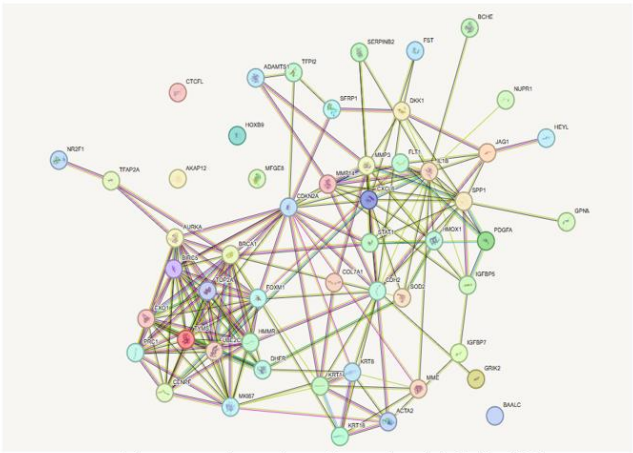
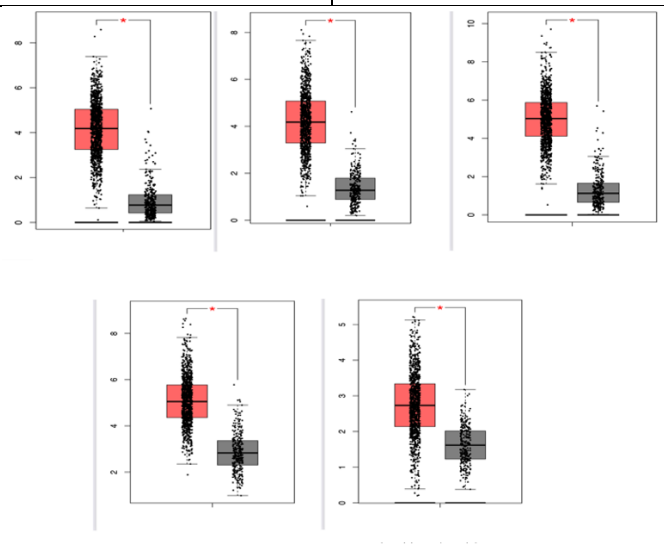


Figure 4.11.1: The protein-protein network analysis for identified DEGs



DISCUSSION

The present study explored stage-related transcriptional alterations in breast cancer by comparing normal breast tissue with tumor samples and by examining the molecular landscape associated with ductal carcinoma in situ and invasive ductal carcinoma. The findings demonstrated substantial dysregulation of gene expression in both disease states, with 387 differentially expressed genes identified in

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

ductal carcinoma in situ and 261 in invasive ductal carcinoma after applying the predefined statistical thresholds. These results supported the central premise of the study that breast cancer progression is accompanied by measurable changes in gene expression and that selected dysregulated genes may serve as biologically relevant markers of disease development and transition. By integrating differential expression analysis with functional enrichment, pathway annotation, and network-based prioritization, the study generated a coherent transcriptomic profile that highlighted genes and pathways potentially involved in proliferation, invasion, treatment resistance, and tumor progression(24). A major finding of the study was the predominance of genes linked to cell-cycle regulation, mitosis, and nuclear division. The Gene Ontology analysis consistently showed enrichment in biological processes related to chromosome segregation, cell division, and proliferative activity. This pattern was biologically plausible because uncontrolled proliferation is a defining feature of malignant transformation. The observed enrichment suggested that the identified differentially expressed genes were not randomly distributed but were concentrated in regulatory systems central to tumor expansion. Such findings aligned with the widely accepted view that breast cancer progression is strongly driven by dysregulated cell-cycle signaling and failure of genomic control mechanisms. The concentration of differentially expressed genes in these pathways also strengthened the internal consistency of the analysis, because these are among the most frequently disturbed biological systems in malignant breast tissue(24, 25).

The KEGG pathway analysis extended this observation by demonstrating enrichment in pathways related to platinum drug resistance, antifolate resistance, tumor necrosis factor signaling, rheumatoid arthritis-associated inflammatory signaling, and toll-like receptor signaling. Although some of these pathways may appear indirect at first glance, together they reflected a biologically meaningful network of cancer survival, treatment adaptation, and inflammatory signaling. Resistance pathways were particularly relevant because breast cancer progression is not solely determined by tumor growth but also by the capacity of malignant cells to survive therapeutic pressure. Likewise, inflammatory and innate immune signaling pathways have increasingly been recognized as important components of the breast tumor microenvironment, influencing proliferation, angiogenesis, invasion, and immune escape. Therefore, the enrichment of these pathways suggested that the transcriptomic changes observed in this study extended beyond proliferation alone and involved broader mechanisms relevant to clinical behavior(5). The identification of hub genes through protein-protein interaction analysis represented another important contribution of the study. The 10 most connected genes—BRCA1, HMMR, MKI67, TYMS, BIRC5, TOP2A, FOXM1, AURKA, EXO1, and UBE2C—formed a biologically credible group with established or emerging relevance in breast cancer. Their central network positions implied that they may function as regulatory nodes rather than merely passive markers of disease. In particular, the validation of BRCA1, TYMS, BIRC5, TOP2A, and FOXM1 in external expression analysis strengthened the likelihood that the identified genes reflected genuine disease-associated signals rather than dataset-specific artifacts. These genes are involved in DNA repair, nucleotide synthesis, mitotic regulation, apoptosis inhibition, and proliferative signaling, all of which are highly relevant to breast tumor biology(10).

Among the identified genes, BRCA1 remained especially important because of its established role in homologous recombination repair, genomic stability, and hereditary breast cancer susceptibility. Tumors associated with BRCA1 dysfunction have often been linked with high histologic grade, aggressive biological behavior, and distinct molecular phenotypes, particularly the triple-negative and basal-like patterns. The finding of BRCA1 upregulation in the present analysis did not necessarily imply preserved tumor-suppressive function, as altered expression in tumor tissue may coexist with defective pathway activity or compensatory molecular responses. This distinction is important because transcript abundance alone does not always capture protein activity or functional integrity. Even so, the emergence of BRCA1 as a hub gene in the interaction network was consistent with its pivotal role in breast cancer biology(15). HMMR, FOXM1, and UBE2C appeared particularly relevant in the context of disease progression and transition toward invasiveness. HMMR is involved in motility-related signaling and cytoskeletal regulation, making it biologically plausible as a contributor to invasive potential. FOXM1 is a transcription factor that promotes proliferation, cell-cycle progression, and epithelial-mesenchymal transition, and its overexpression has repeatedly been implicated in tumor aggressiveness. UBE2C, which supports ubiquitin-mediated cell-cycle progression, has similarly been associated with proliferative activity and poor tumor behavior. Taken together, these genes appeared especially relevant to the biological distinction between non-invasive and invasive lesions. Their prominence in the present analysis suggested that they may be useful candidates for further investigation into stage transition, although the current study design did not provide direct experimental confirmation of causality(20).

The remaining hub genes also showed strong biological relevance. TOP2A encodes topoisomerase IIα, a key enzyme in DNA replication, transcription, chromosomal condensation, and segregation, and has long been associated with rapidly proliferating breast tumors and HER2-amplified disease. TYMS plays a central role in folate metabolism and nucleotide synthesis, and its dysregulation may contribute both to malignant proliferation and resistance to antifolate therapies. AURKA is essential for spindle assembly and chromosomal alignment, and its overexpression has been linked with medication resistance and poor prognosis in breast cancer. EXO1 has been associated with DNA repair, cell-cycle activity, and unfavorable outcomes in breast malignancy. MKI67 remains one of the most established markers of proliferation in clinical pathology and has been widely used to reflect the growth fraction of tumor cells. BIRC5,

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

or survivin, contributes to inhibition of apoptosis and mitotic regulation and has attracted interest as a possible therapeutic target, particularly in aggressive breast cancer subtypes. The fact that many of these genes converged on proliferation, survival, and genomic stability added biological credibility to the overall network model generated in this study(24). Comparison with previous literature showed broad agreement between the present findings and earlier transcriptomic and next-generation sequencing studies in breast cancer. Several previous investigations have reported that molecular profiling can identify clinically relevant genomic or transcriptomic alterations capable of refining tumor classification and improving precision treatment strategies. Other studies have emphasized the value of RNA-based analyses in uncovering subtype-specific signatures, especially in biologically aggressive disease such as triple-negative breast cancer. The current findings were consistent with these broader observations in that they demonstrated the ability of transcriptomic analysis to identify genes associated with tumor aggressiveness, proliferation, and resistance-related pathways. However, the present study differed in its focus on stage-related comparison involving ductal carcinoma in situ and invasive ductal carcinoma, which offered a narrower but clinically meaningful perspective on progression-related biology(25).

At the same time, the findings should be interpreted with appropriate caution. Although the study identified genes that appeared capable of distinguishing tumor from normal tissue and of highlighting biological differences across disease states, the evidence remained bioinformatic and associative rather than mechanistic. The results were useful for prioritizing candidate biomarkers, but they did not establish that these genes independently caused progression from ductal carcinoma in situ to invasive ductal carcinoma. Likewise, the study suggested that genes such as HMMR, FOXM1, and UBE2C may be especially informative in the transition toward invasive behavior, yet this inference would require direct comparative validation in larger stage-specific cohorts and functional studies. Therefore, the current analysis provided a strong hypothesis-generating framework rather than definitive proof of clinical utility(7). The study had several strengths. It used an RNA-Seq-based analytical design, which provided a broader and more sensitive view of transcriptomic alterations than many targeted or low-throughput methods. The integration of differential expression analysis with Gene Ontology, KEGG enrichment, protein interaction mapping, Cytoscape-based hub gene prioritization, and external validation added methodological depth and reduced reliance on a single analytic layer. This multistep approach increased the confidence that the identified genes were biologically meaningful. In addition, the inclusion of both non-invasive and invasive disease states enhanced the relevance of the study to breast cancer progression rather than limiting the analysis to tumor-versus-normal contrasts alone(12).

Despite these strengths, several limitations were present. The dataset was derived from a single public source and represented a restricted geographic population, which may limit generalizability. Public datasets also depend on the quality and completeness of the original submissions, and the present study remained constrained by the sample composition, metadata availability, and sequencing characteristics of the archived dataset. The analysis did not include large multicenter external cohorts for independent validation. Functional experiments, protein-level confirmation, and clinical outcome analyses were also absent, meaning that the biological significance of the identified genes could not be fully resolved at the translational level. In addition, the study did not provide detailed subgroup analyses according to receptor status, molecular subtype, survival outcome, treatment exposure, or nodal and metastatic burden, all of which could influence gene expression profiles in clinically important ways. Computational resource limitations, restricted data access, and time constraints may also have reduced the breadth of the analysis(16). Another important limitation concerned the distinction between disease stages. While the study successfully identified dysregulated genes in ductal carcinoma in situ and invasive ductal carcinoma relative to normal tissue, it did not fully establish a direct comparative framework between the two tumor stages themselves using comprehensive statistical modeling. As a result, some conclusions regarding stage transition were biologically suggestive rather than analytically conclusive. Similarly, although the study aimed to identify diagnostic biomarkers, the results did not include performance-based measures such as receiver operating characteristic analysis, sensitivity, specificity, or predictive modeling. Therefore, the proposed biomarkers should presently be considered candidate molecular markers rather than clinically validated diagnostic tools.

These limitations also pointed directly toward future research priorities. Larger and more diverse transcriptomic datasets should be used to validate the identified biomarkers across different populations and clinical settings. Multi-omics integration involving transcriptomics, genomics, proteomics, and epigenomics would provide a more complete picture of breast cancer biology and help determine whether the observed transcriptional changes correspond to protein-level and pathway-level functional effects. Functional laboratory studies are particularly needed to clarify whether genes such as HMMR, FOXM1, UBE2C, EXO1, and BIRC5 actively drive invasion, treatment resistance, or tumor progression, or whether they primarily reflect downstream consequences of malignant transformation. Longitudinal studies examining how these genes change across tumor development, recurrence, and treatment exposure would also be valuable. In addition, translational work linking these biomarkers to liquid biopsy platforms, targeted therapeutic response, and clinical outcomes may improve their eventual applicability in personalized medicine(1). The present study demonstrated that breast cancer was characterized by marked transcriptomic dysregulation involving proliferation-associated genes, resistance-related pathways, and highly connected hub genes with plausible roles in disease progression. The results were in broad agreement with previous sequencing-based

INSIGHTS-JHR
INSIGHTS-JOURNAL OF HEALTH AND REHABILITATION

breast cancer research and further emphasized the value of next-generation transcriptomic analysis in identifying biologically meaningful candidate biomarkers. Although the study did not establish clinical applicability on its own, it provided a rational and methodologically integrated basis for future validation work. The identified genes, particularly BRCA1, HMMR, FOXM1, TOP2A, TYMS, BIRC5, EXO1, AURKA, MKI67, and UBE2C, merit further investigation as markers of tumor biology, progression, and possible therapeutic vulnerability in breast cancer.

CONCLUSION

This study demonstrated that breast cancer progression is accompanied by distinct alterations in gene expression and associated molecular pathways when normal breast tissue is compared with ductal carcinoma in situ and invasive ductal carcinoma. The integrated transcriptomic and bioinformatics analyses identified biologically meaningful differentially expressed genes, highlighted disruption of cell cycle- and mitosis-related processes, and revealed key pathway enrichment linked to treatment resistance and inflammatory signaling. Network-based analysis further prioritized several hub genes, particularly BRCA1, TOP2A, FOXM1, TYMS, and BIRC5, as potentially important biomarkers with diagnostic and therapeutic relevance, while external validation supported the credibility of these findings. Overall, the study fulfilled its objective by clarifying stage-related molecular differences in breast cancer and providing a focused set of candidate genes and pathways that may support future advances in early detection, prognostic assessment, and more individualized treatment strategies.

AUTHOR CONTRIBUTION

Author	Contribution
Momna Quddus	Conceptualization, Methodology, Formal Analysis, Writing - Original Draft, Validation, Supervision
Romaisa Saeed	Methodology, Investigation, Data Curation, Writing - Review & Editing
Laiba Sultana	Investigation, Data Curation, Formal Analysis, Software
Naveen Kaleem	Software, Validation, Writing - Original Draft
Aqsa Naz	Formal Analysis, Writing - Review & Editing
Waseem Haider	Writing - Review & Editing, Assistance with Data Curation
Saadia Momal Zafar	Supervision, Conceptualization, Review & Editing
Rafia Anwer	Formal Analysis, Writing - Review & Editing

REFERENCES

1. Colomer R, González-Farré B, Ballesteros AI, Peg V, Bermejo B, Pérez-Mies B, et al. Biomarkers in breast cancer 2024: an updated consensus statement by the Spanish Society of Medical Oncology and the Spanish Society of Pathology. Clin Transl Oncol. 2024;26(12):2935-51.

2. Watanabe S, Haratani K, Nakayama T, Takahama T, Iwasa T, Takeda M, et al. Clinical and molecular landscape of metastatic extramammary Paget's disease. Oncologist. 2025;30(9).

3. Cai P, Li J, An M, Li M, Guo J, Li J, et al. Comprehensive analysis of RNA-5-methylcytosine modification in breast cancer brain metastasis. Future Oncol. 2024;20(37):2993-3008.

4. Veiga DFT, Nesta A, Zhao Y, Deslattes Mays A, Huynh R, Rossi R, et al. A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. Sci Adv. 2022;8(3):eabg6711.

5. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021;39(5):599-608.

6. Hlevnjak M, Heublein S, Thewes V, Wagener L, Pixberg C, Fremd C, et al. Delivering precision oncology in metastatic breast cancer: Clinical impact of comprehensive genomic profiling-The CATCH experience. Int J Cancer. 2026;158(6):1675-89.

7. Pilati C, Soulabaille A, Gallois C, Blons H, Cayre A, Sroussi M, et al. ERBB2 Comprehensive Profiling and Prognostication in Stage III Colon Cancer: Findings From PETACC8 and IDEA-France Cohorts. Gastroenterology. 2025;168(4):714-24.e4.

8. Koh MZ, Ho WY, Yeap SK, Ali NM, Yong CY, Boo L, et al. Exosomal-microRNA transcriptome profiling of Parental and CSC-like MDA-MB-231 cells in response to cisplatin treatment. Pathol Res Pract. 2022;233:153854.

9. Guerini-Rocco E, Bellerba F, Concardi A, Taormina SV, Cammarata G, Fumagalli C, et al. Expression of immune-related genes and breast cancer recurrence in women with ductal carcinoma in situ. Eur J Cancer. 2024;203:114063.

10. Aftimos P, Oliveira M, Irrthum A, Fumagalli D, Sotiriou C, Gal-Yam EN, et al. Genomic and Transcriptomic Analyses of Breast Cancer Primaries and Matched Metastases in AURORA, the Breast International Group (BIG) Molecular Screening Initiative. Cancer Discov. 2021;11(11):2796-811.

11. Bansal R, Adeyelu T, Elliott A, Tan AR, Ribeiro JR, Meisel J, et al. Genomic Landscape of Malignant Phyllodes Tumors Identifies Subsets for Targeted Therapy. JCO Precis Oncol. 2024;8:e2400289.

12. Marcinak CT, Murtaza M, Wilke LG. Genomic Profiling and Liquid Biopsies for Breast Cancer. Surg Clin North Am. 2023;103(1):49-61.

13. Nakayama J, Matsunaga H, Arikawa K, Yoda T, Hosokawa M, Takeyama H, et al. Identification of two cancer stem cell-like populations in triple-negative breast cancer xenografts. Dis Model Mech. 2022;15(6).

14. Yoon H, Lee S. Integration of Genomic Profiling and Organoid Development in Precision Oncology. Int J Mol Sci. 2021;23(1).

15. Sideris N, Dama P, Bayraktar S, Stiff T, Castellano L. LncRNAs in breast cancer: a link to future approaches. Cancer Gene Ther. 2022;29(12):1866-77.

16. Patel K, Rao DM, Sundersingh S, Velusami S, Rajkumar T, Nair B, et al. MicroRNA Expression Profile in Early-Stage Breast Cancers. Microrna. 2024;13(1):71-81.

17. Sun L, Wu A, Bean GR, Hagemann IS, Lin CY. Molecular Testing in Breast Cancer: Current Status and Future Directions. J Mol Diagn. 2021;23(11):1422-32.

18. Park AY, Han MR, Seo BK, Ju HY, Son GS, Lee HY, et al. MRI-based breast cancer radiogenomics using RNA profiling: association with subtypes in a single-center prospective study. Breast Cancer Res. 2023;25(1):79.

19. Yang S, Wang X, Zhou X, Hou L, Wu J, Zhang W, et al. ncRNA-mediated ceRNA regulatory network: Transcriptomic insights into breast cancer progression and treatment strategies. Biomed Pharmacother. 2023;162:114698.

20. Rakha EA, Pareja FG. New Advances in Molecular Breast Cancer Pathology. Semin Cancer Biol. 2021;72:102-13.

21. Mohd Zuhdi NF, Siddig A, Mohd Nafi SN, Md Salleh MS, Yahya MM, Wan Zain WZ, et al. Next-generation sequencing in breast cancer: current clinical applications and future directions. Ann Med. 2025;57(1):2569989.

22. Karousi P, Kontos CK, Papakotsi P, Kostakis IK, Skaltsounis AL, Scorilas A. Next-generation sequencing reveals altered gene expression and enriched pathways in triple-negative breast cancer cells treated with oleuropein and oleocanthal. Funct Integr Genomics. 2023;23(4):299.

23. Lin Y, Wang J, Wang K, Bai S, Thennavan A, Wei R, et al. Normal breast tissues harbour rare populations of aneuploid epithelial cells. Nature. 2024;636(8043):663-70.

24. Xu Y, Su GH, Ma D, Xiao Y, Shao ZM, Jiang YZ. Technological advances in cancer immunity: from immunogenomics to single-cell analysis and artificial intelligence. Signal Transduct Target Ther. 2021;6(1):312.

25. Wang F, Xu Y, Wang R, Zhang B, Smith N, Notaro A, et al. TEQUILA-seq: a versatile and low-cost method for targeted long-read RNA sequencing. Nat Commun. 2023;14(1):4760.