IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Uncovering Subtype-Specific Biomarkers in Breast Cancer through Bioinformatics Analysis

Mihika Deora[1], Nirupma Singh[2]

1) Oberoi International School, Mumbai, Maharashtra, India
2) Department of Biological Sciences and Engineering, Netaji Subhas University of Technology (Formerly NSIT), Dwarka, Delhi, India; nirupmajadaun21@gmail.com

ABSTRACT: Breast cancer is the most common cancer in the world, accounting for over 28.2% of all female cancers. However, there are still no effective subtype-specific biomarkers to help in diagnosis and more targeted therapies for patients with breast cancer. This study uses advanced bioinformatics approaches to identify subtype-specific biomarkers for *four* molecular subtypes and analyse their potential role in treatment processes. To accomplish this objective, differential gene expression analysis (DGE) was conducted using the GEO2R (Gene Expression Omnibus) tool to gain different data sets. Using the data, a network was constructed in the database STRING, which was then analysed using Cytoscape to identify the topological parameters. Pathway analysis was conducted in the Reactome database to determine the top-enriched pathways in which the significant hub genes for breast cancer are present. The study identified the top significant genes and hub genes in breast cancer subtypes, assessing their ability as biomarkers for more personalised treatments through detailed DGE, network, and pathways analysis. Notably, *RPS27A* emerged as the top significant gene in all the subtypes, with its presence in the EML4 and NUDC in the mitotic spindle formation pathway for all 4 subtypes showing its potential for therapy. These findings will enhance understanding of the treatment processes of breast cancer and aim for more targeted therapies for different subtypes.

KEYWORDS: Computational Biology and Bioinformatics, Computational Biomodelling, Cancer Biology Analysis, Network Biology, Pathway Analysis.

Introduction

Breast cancer (BC), the most commonly diagnosed cancer in women worldwide, leads to significant morbidity and mortality, placing a considerable strain on healthcare systems. Breast cancer affects millions of women globally, with approximately 1.5 million new cases annually, making it a leading cause of cancer-related deaths.[1] In India, some foundations such as the ICGA (Indian Cancer Genome Atlas) are developing technologies to identify the genetic basis of cancer in the Indian population and genetic biomarkers that will improve the rate of detection and better-targeted therapies.[2]

Breast cancer can be classified into several subtypes, which are grouped according to the immunohistochemical expression of hormone receptors. Luminal A is characterised by the presence of the ER and PR receptors and the absence of the HER2 (human epidermal growth factor receptor 2) receptor. Clinically, this subtype grows at a slower rate, has a lower chance of relapse, and has a higher survival rate compared to others. It presents a positive and faster response to hormone therapy in comparison to chemotherapy.[3] According to the European Society for Medical Oncology, genetic platforms identify the preferred treatment for the patient based on the severity, risk of relapse, and survival rate.[4] Luminal B grows faster and is harder to predict than Luminal A, but is also characterised by the presence of PR+ and sometimes PR- receptors.[5] Hormonal therapy, along with chemotherapy, can be beneficial to it. The presence of HER2 expression characterises HER2 and causes it to grow at a faster rate compared to the luminal types. The prediction has improved since the introduction of more HER2-targeted therapies, specifically directed drugs, and a high response to chemotherapy. Triple-negative (TNBC) has ER-, PR-, and HER2- receptors, which cause it to have highly aggressive behaviour, early relapses, a higher proliferation rate, changes within the repair genes, and genomic stability. *BRCA1* mutation carriers often have the basal-like subtype, which is comparable to TNBC but has different genetic markers.[6]

By identifying new biomarkers and the genetic basis of the disease, its risk and progression can be monitored and better understood. Studies have used different bioinformatic approaches to focus on the molecular heterogeneity of breast cancer progression. The National Institute of Biomedical Genomics (NIBMG) uses biomedical genomics to identify the genetic markers associated with the disease. Although various biomarkers have been proposed, the severity of breast cancer requires more efficient data methods, such as bioinformatics approaches that can help bring data from diverse sources together and offer a more holistic view of the disease. Research to identify molecular biomarkers that would be more efficient for therapies has been done, which has helped improve the progression of the disease.[7] By using gene expression profiling to uncover intrinsic subtypes, researchers carried out groundbreaking research on the molecular classification of breast cancer, which has since impacted therapeutic approaches.[8] To gain a better understanding of tumor heterogeneity, this work was extended by dividing breast cancer into ten different subgroups using integrative genomic analysis.[9]

DOI: 10.36838/v8i3.71

While there has been significant progress in breast cancer research, it lacks deeper bioinformatic analysis. These studies have relied on genomic data, overlooking the proteomic and transcriptomic factors, which are not able to capture tumor heterogeneity. Subtype-specific biomarker identification is required for deeper analysis. To find new biomarkers and treatment targets, advanced computational techniques are needed due to the complexity of breast cancer subtypes. Bioinformatics tools such as STRING, CYTOSCAPE, and REACTOME will be employed in the research to perform differential gene expression analysis, pathway analysis, and network analysis, allowing researchers to explore the disease-related pathways in the body and genetic mutations that are related to the disease. The use of such bioinformatic tools allows vast datasets to be analysed together quickly and efficiently, providing a deeper understanding of the progression and the possible treatments. Furthermore, research has looked at the mutational signatures of breast cancer; however, further research is required to determine the functional implications of these mutations and the effect they have on the body's response to the treatment.[10] Additionally, there is still research on triple-negative breast cancer, mainly reliable therapeutic targets and personalised treatments for this subtype's aggressiveness.

However, previous research studies have primarily relied on genomic data and ignored the incorporation of proteomic and transcriptomic data, which could provide a more in-depth understanding of the disease. This gap in the research shows that a more comprehensive approach with more in-depth biological research would be beneficial. For example, whereas genetic mutations have been researched in great detail, little is known about how they affect the function and patterns of protein production. Filling in these gaps will improve our knowledge of the illness and result in better methods for diagnosis and treatment. Thus, this paper argues that the leverage of bioinformatic tools that incorporate multi-omics data will help identify new biomarkers, creating better therapeutic targets, personalised treatments, and earlier detection of the disease. This study intends to identify formerly unknown molecular patterns that might be useful treatment targets by examining a variety of datasets. By enabling more accurate disease characterisation and customised treatments, the discoveries will support the expanding field of personalised therapies.

Methods

Data Collection:

The gene expression data were retrieved from the NCBI GEO platform. Specifically, the subtypes, Luminal A, Luminal B, HER2-Positive, and Triple negative RNA-Seq data are extracted for analysis. *GSE233242 -'Tumor circadian clock strength influences metastatic potential and predicts patient prognosis in Luminal A breast cancer*' investigates the circadian clocks in human breast tumors by conducting an expression profiling using high-throughput sequencing. *GSE214344 - 'A genome-wide cell-free DNA methylation analysis identifies an episignature associated with metastatic luminal B breast cancer.*' aims to discover non-invasive biomarkers of the disease using an epigenomic approach. *GSE52194 - 'mRNA-sequencing of breast cancer subtypes and normal tissue*' uses RNA sequencing technology to identify the digital transcriptome. *GSE167152 - 'Comparative Characterisation of 3D Chromatin Organisation in Triple-Negative Breast Cancers [RNA-seq]'* detected CTCF-dependent TNBC-susceptible loss/gain of 3D chromatin organisations using expression profiling by high-throughput sequencing.

Differential Gene Expression:

Differential gene expression analysis was conducted using the GEO2R analysis tool for the RNA-Seq datasets obtained for each subtype. GEO2R applies a default normalisation to each of the datasets (log2 transformation and quantile normalisation for microarray data, variance-stabilising transformation for RNA-seq) before it undergoes differential expression analysis. This allows for fewer technical biases and comparability across samples, which increases the reliability of the identified DEGs. For the RNA-Seq data of subtype Luminal A, 29 tumor samples and 42 normal samples were assigned as test and control groups, respectively. The second type, Luminal B, had 7 cell-free DNA samples from luminal B patients for the test group and 5 normal cell-free DNA samples for the control group. The third subtype, TNBC, had 18 triple-negative breast cancer cell samples and 2 normal samples, which were assigned as the test and control groups, respectively. The last subtype, HER2+, had 5 tumor samples and 3 that matched normal samples. The raw data were normalised using DESeq2 (Differential Expression analysis based on the Negative Binomial distribution), and the DEGs (Differentially Expressed Genes) were selected based on an adjusted p-value < 0.05 and a fold change > 2. To ensure robustness and reduce the possibility of background noise, a threshold of FC > 2 was applied to identify those genes with significant, biologically meaningful expression changes together with an adjusted p-value < 0.05.

Network Construction and Analysis:

A network was constructed using the STRING database by identifying gene-gene interactions for the significant genes of each subtype. A list of the top 2000 differentially expressed genes was used as input to generate a network. The interactions with a high confidence of 0.7 were retained. The network was exported as a short tabular text output to visualise in Cytoscape to analyse the topological parameters and visualise the network. The software provided gene interactions and positive topological parameters, which were downloaded for further analysis.

Pathway Analysis:

From the topological parameters file created in Cytoscape, the degree was set in a descending manner to identify the top genes. Pathway enrichment analysis was performed using Reactome for the significant DEGs (differentially expressed genes) to identify biological pathways significantly affected by the differentially expressed genes. The top enriched pathways with a p-value < 0.05 were considered significant. The pathway analysis results adopted from Reactome were then used to

search for the top 10 hub genes of the network to check their presence in the top enriched pathways.

Results and Discussion

This section represents the results obtained from the bioinformatics analysis carried out for different subtypes of breast cancer and the top genes and pathways identified to gain insights into personalised treatment approaches. The focus is on understanding the specific biological processes, genes, and pathways associated with the subtypes of breast cancer. After performing DGE analysis for Luminal A - GSE233242, 13450 significant genes were identified, for Luminal B - GSE270967, 2088 significant genes were found, for TNBC - GSE167152, 4546 significant genes were identified, and for HER2+ - GSE52194, 4575 significant genes were found. The overexpressed and underexpressed genes were visualised in the form of volcano plots for each subtype, as shown in Figure 1. The analysis revealed a total of 22,572 genes associated with different subtypes of breast cancer. The key findings also included the identification of different genes and pathways that play a role in breast cancer.

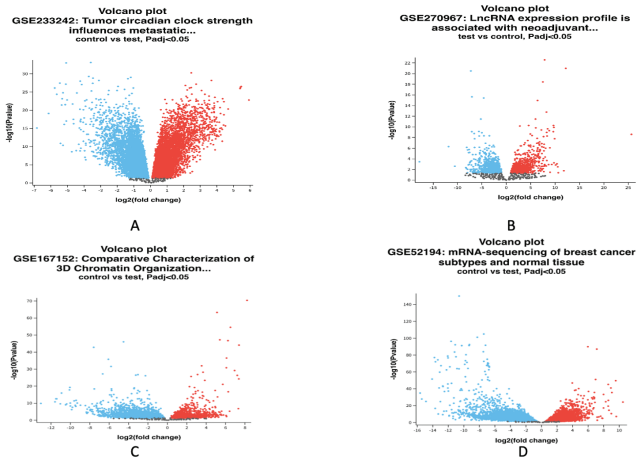

Figure 1: Volcano plots for the significant genes identified from differential gene expression analysis by GEO2R tool. (A) This plot shows the significant DE genes for Luminal A subtype, (B) This plot shows the significant DE genes for Luminal B subtype, (C) This plot shows the significant DE genes for TNBC subtype, (D) This plot shows the significant DE genes for HER2+ subtype. This analysis helped in identifying the differentially expressed genes for each subtype of breast cancer as compared to the normal individuals. The x-axis shows the fold change of the genes and y-axis represents the adjusted p-value.

The network construction for gene interactions of DE genes in STRING showed densely connected networks for each subtype with high-confidence interactions of >0.7, as shown in Figure 2. For Luminal A, the gene interaction network visualised in Cytoscape had 1567 nodes and 2117 edges. The network for the Luminal B subtype had 1564 nodes and 1844 edges. For TNBC, the network constructed and visualised included 1432 nodes and 4122 edges. For the last subtype, HER2+, the network had a total of 1398 nodes and 2059 edges. After performing network analysis in Cytoscape, four different topological measures, namely, degree, betweenness centrality, clustering coefficient, and closeness centrality, helped in identifying top hub genes of the network for each subtype. The top 10 hub genes identified for each breast cancer subtype are listed in Table 1.

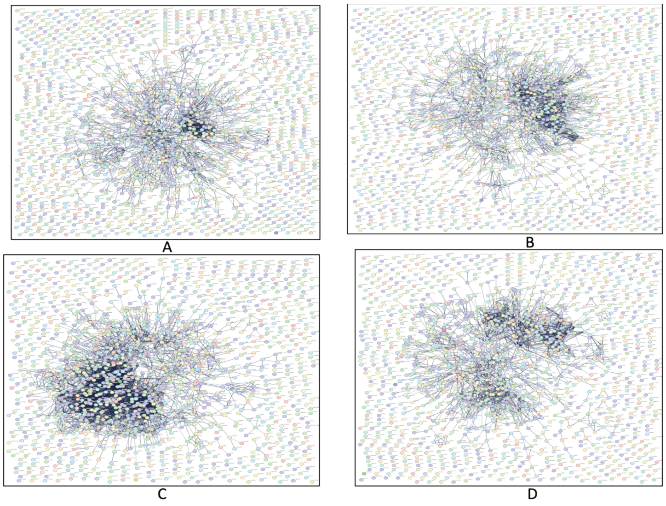

Figure 2: Networks constructed in STRING and visualized in Cytoscape for each of the four subtypes using the significant genes from the DGE analysis. (A) Gene-interaction network of Luminal A subtype, (B) Gene-interaction network of Luminal B subtype, (C) Gene-interaction network of TNBC subtype, (D) Gene-interaction network of HER2+ subtype. Network analysis was carried out to identify the most significant genes of each subtype of breast cancer.

Table 1: A list of top 10 hub genes for each subtype of breast cancer computed on the basis of topological parameters of gene-interaction network.

S.No.	Luminal A	Luminal B	TNBC	HER2+
1.	*RPS27A*	*IL1B*	*RPS27A*	*RPS27A*
2.	*HDAC1*	*IFNG*	*H3C13*	*IL1B*
3.	*RPL11*	*IL10*	*H3C12*	*H3C13*
4.	*RPL5*	*CXCL8*	*CENPA*	*PTPRC*
5.	*MRPL24*	*CXCL10*	*CCNA2*	*RPS8*
6.	*RPS8*	*HSP90AA1*	*H4C5*	*BUB1*
7.	*RPS7*	*CCL2*	*H4C8*	*RPS27*
8.	*CENPA*	*CALML5*	*CDC20*	*RPL5*
9.	*MRPS5*	*ESR1*	*CCNB1*	*FN1*
10.	*RPS27*	*TLR2*	*BUB1*	*CDC20*

The pathway analysis highlighted the top enriched, over-represented Reactome pathways for each subtype of breast cancer. The top 5 enriched pathways for subtype Luminal A were Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal, Amplification of signal from the kinetochores, Chromatin modifying enzymes, Chromatin organisation, and Cytokine Signalling in the Immune system. The top 5 enriched pathways for subtype Luminal B were Interleukin-10 signalling, Signalling by Interleukins, Chemokine receptors bind chemokines, Peptide ligand-binding receptors, and Cytokine Signalling in the Immune system. The top 5 enriched pathways for subtype HER2+ were EML4 and NUDC in mitotic spindle formation, Cytokine Signalling in the Immune system, RHO GTPases Activate Formins, Amplification of signal from the kinetochores, and Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal. The top 5 enriched pathways for subtype TNBC

DOI: 10.36838/v8i3.71

were Cell Cycle, Mitotic, Cell Cycle, Cell Cycle Checkpoints, M Phase, and G2/M Checkpoints.

The top 25 hub genes of the gene-interaction network were checked for their presence in the top 25 pathways. For the Luminal A subtype, of the top 25 genes, the top 13 were found in the pathway, and 12 were not in the enriched pathways. For Luminal B, 15 of the top 25 genes were found in the pathways, and 10 were not found. In the triple-negative (TNBC) subtype, of the top 25 genes, 24 were found, and 1 was not found in the top 25 pathways. Lastly, for the HER2+ subtype, 16 genes were found, and 9 genes were not visible in the top pathway.

This study is about different subtypes of breast cancer and makes use of different bioinformatic approaches, such as DGE analysis, network analysis, and pathway analysis, to identify key biomarkers related to breast cancer. The primary objective of this paper was to perform a comprehensive set of analyses to identify subtype-specific biomarkers and determine their potential role in treatment processes. The results indicated that the genes found were significantly upregulated in individuals with breast cancer. Gene enrichment analysis revealed that 'Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal' is heavily involved in the progression of luminal A breast cancer. 'Cell Cycle, Mitotic' is most involved for triple-negative, and 'EML4 and NUDC in mitotic spindle formation' for HER2+ breast cancer. These results provided insights into more personalised treatments and targeted therapies for patients.

In previous studies conducted by ASC Omega suggested that immune-related gene expression has a pivotal role in the progression of this disease. The identification of key prognostic gene signatures, including those derived from WGCNA and LASSO analysis, serves as a critical biomarker for breast cancer, which adds to the already existing evidence that suggests there is a genetic predisposition to breast cancer. In the study, a similar approach was used as this research paper, which involved reliance on TCGA datasets contributing to the growing knowledge of subtype-specific biomarkers' relevance in treatments.

The identification of the subtype-specific biomarkers as diagnostic markers suggests their role in developing more personalised treatments for breast cancer subtypes. Additionally, these findings could also lead to the discovery of more therapeutic targets for patients with this disease. Conducting pathway and network analysis provided an overall view of the systems biology behind different subtypes of breast cancer.

A major strength of this study is the comprehensive approach and use of multiple bioinformatic analyses and tools than other literature, which allowed for deeper analysis of the genomic data while focusing on each breast cancer subtype. One limitation could be the use of datasets that are available to anybody, as they may not fully represent the diversity of the global population and may not be classified. The biomarkers identified in this study can further be subjected to the identification of the lead compounds by using more advanced bioinformatic approaches, such as molecular docking and molecular dynamics simulation. The viability of these biomarkers can be further experimentally validated for their suitability in the clinical setting. Additionally, expanding the dataset to include a more diverse population could enhance the generalizability of the findings.

In order to enhance the therapeutic relevance of these results, wet-lab confirmation of the hub genes and pathways would be crucial. In order to precisely quantify changes in gene expression, methods like quantitative polymerase chain reaction (qPCR) might be employed to confirm the differential expression of candidate genes at the mRNA level between samples of breast cancer tissue and matched controls. Additionally, patient-derived tumor sections may be subjected to immunohistochemistry (IHC) to verify the protein-level expression of these biomarkers, enabling the spatial localization of the proteins within the tissue microenvironment. To guarantee that the discovered biomarkers are not only statistically significant but also functionally confirmed for their potential in diagnosis and treatment, these techniques would close the gap between in silico predictions and biological relevance.

■ Conclusion

In conclusion, this study identifies the biomarkers for each subtype of breast cancer by making use of bioinformatic approaches such as differential gene expression analysis, network analysis, and pathway analysis, which could further be experimentally validated for their potential in personalised treatments. *RPS27A* was identified as a key biomarker across 3 subtypes (Luminal A, HER2+, and TNBC), determining its potential as a therapeutic target. *RPS27A*'s recurring occurrence across several subtypes has biological significance because, despite being traditionally thought of as a housekeeping gene necessary for ribosome function, new research indicates it also plays oncogenic roles, including promoting proliferation and altering the ubiquitin–proteasome pathway. Reactome pathways showed enrichment of pathways related to mitotic spindle formation, cytokine signalling, chromatin organisation, and cell cycle regulation. Hub network genes, *HDAC1*, *IFNG*, *H3C13*, and *IL1B*, were identified as unique biomarkers for each subtype, Luminal A, Luminal B, HER2+, and TNBC, respectively. Overall, this study paves the way for therapeutic intervention of key biomarkers for each specific subtype of breast cancer that can help in personalised treatments.

■ Acknowledgments

MD would like to acknowledge Aashna Saraf, Founder of CreatED, for providing valuable feedback and guidance throughout the project.

■ References

1. Wilkinson, L.; Gathani, T., Understanding breast cancer as a global health concern. *British Journal of Radiology* **2022**, *95* (1130), 20211033.
2. Dhup, S.; Ramamurthy, R., A genome atlas mapping cancers in India. *Nature India* **2024**.
3. Eroles, P.; Bosch, A.; Alejandro Pérez-Fidalgo, J.; Lluch, A., Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. *Cancer Treatment Reviews* **2012**, *38* (6), 698-707.

4. Zhou, G. Q.; Lv, J. W.; Tang, L. L.; Mao, Y. P.; Guo, R.; Ma, J.; Sun, Y., Evaluation of the National Comprehensive Cancer Network and European Society for Medical Oncology Nasopharyngeal Carcinoma Surveillance Guidelines. *Front Oncol* **2020**, *10*, 119.
5. Gajbe, B.; Das Kurmi, B.; Kenwat, R.; Paliwal, R.; Paliwal, S. R., Chapter 1 - Breast cancer: introduction. In *Targeted Nanomedicine for Breast Cancer Therapy*, Paliwal, S. R.; Paliwal, R., Eds. Academic Press: 2022; pp 3-26.
6. Derakhshan, F.; Reis-Filho, J. S., Pathogenesis of Triple-Negative Breast Cancer. *Annu Rev Pathol* **2022**, *17*, 181-204.
7. Alam, M. S.; Sultana, A.; Reza, M. S.; Amanullah, M.; Kabir, S. R.; Mollah, M. N. H., Integrated bioinformatics and statistical approaches to explore molecular biomarkers for breast cancer diagnosis, prognosis and therapies. *PLoS One* **2022**, *17* (5), e0268967.
8. Zhang, X., Molecular Classification of Breast Cancer: Relevance and Challenges. *Archives of Pathology & Laboratory Medicine* **2022**, *147* (1), 46-51.
9. Guo, L.; Kong, D.; Liu, J.; Zhan, L.; Luo, L.; Zheng, W.; Zheng, Q.; Chen, C.; Sun, S., Breast cancer heterogeneity and its implication in personalized precision therapy. *Exp Hematol Oncol* **2023**, *12* (1), 3.
10. Jiménez-Santos, M. J.; García-Martín, S.; Fustero-Torre, C.; Di Domenico, T.; Gómez-López, G.; Al-Shahrour, F., Bioinformatics roadmap for therapy selection in cancer genomics. *Mol Oncol* **2022**, *16* (21), 3881-3908.

Authors

Mihika Deora is presently a Grade 11 student at Oberoi International School, OGC. She has a keen interest in biology and chemistry and has completed two internships and courses in the field. Additionally, her academic curiosity extends to areas such as neuroscience, genetics, and biochemistry, reflecting her passion for interdisciplinary learning and research.

Nirupma Singh is a Bioinformatics Scientist with a doctorate in Biotechnology and Bioinformatics from the University of Delhi, with six years of hands-on research and development experience. Her journey is marked by a robust foundation in Machine Learning and Python, with five years of expertise. Proficient in Linux and cloud servers like AWS. She excels in structural biology, systems biology, protein/gene network analysis, data mining, and computational genomics.

DOI: 10.36838/v8i3.71