Integrated Network Analysis and Clinical Validation of Hub Genes Associated with Relapse-Free Survival in Breast Carcinoma

Sounish Singha

Indian Institute of Science Education and Research (IISER) Berhampur, Odisha, India

Email: 25011@iiserbpr.ac.in
ORCID: https://orcid.org/0009-0007-5450-8335

Abstract:

Breast carcinoma remains a leading cause of cancer-related mortality globally. The identification of robust molecular biomarkers is critical for early diagnosis and improving patient prognosis. This study utilized the gene expression dataset GSE10810 from the NCBI Gene Expression Omnibus (GEO). A subset of 10 samples (5 tumor, 5 normal) was analyzed to identify Differentially Expressed Genes (DEGs) using GEO2R with cut-off criteria of $|logFC| > 1.5$ and adjusted $P < 0.05$. Protein-protein interaction (PPI) networks were constructed using the STRING database and visualized in Cytoscape. The CytoHubba plugin (MCC algorithm) was employed to screen for the top 10 hub genes. Clinical relevance was validated via Relapse-Free Survival (RFS) analysis using the Kaplan-Meier Plotter. Analysis identified a core signature of 10 hub genes: BIRC5, UBE2C, CCNB2, MAD2L1, SPAG5, NUSAP1, MCM4, RRM2, BUB1B, and PBK. These genes are primarily enriched in mitotic cell cycle regulation and chromosomal segregation. Survival analysis revealed that elevated expression of these hub genes is significantly associated with poor Relapse-Free Survival ($P < 0.05$) in breast cancer patients. Specifically, BIRC5 and UBE2C demonstrated the highest prognostic risk potential. This study highlights a significant 10-gene signature associated with tumor proliferation and poor prognosis. These findings suggest that BIRC5 and UBE2C may serve as promising diagnostic biomarkers and therapeutic targets for breast carcinoma.

Keywords: Breast Cancer, Bioinformatics, Hub Genes, KM Plotter, GSE10810, BIRC5.

1. Introduction:

Breast cancer is the most frequently diagnosed malignancy and the leading cause of cancer-related death among women worldwide. Despite significant advancements in therapeutic strategies, including surgery, chemotherapy, and endocrine therapy, the prognosis for patients with advanced or metastatic disease remains poor [1]. The clinical behavior of breast cancer is highly heterogeneous, meaning that patients with the same histological stage often exhibit vastly different responses to treatment and survival outcomes [2]. Therefore, the identification of robust molecular biomarkers is urgently needed to improve early diagnosis, refine prognostic stratification, and develop more effective targeted therapies.

In recent years, the rapid development of high-throughput microarray technology and bioinformatics analysis has revolutionized cancer research. Gene expression profiling allows researchers to monitor the expression levels of thousands of genes simultaneously, providing a comprehensive view of the transcriptomic landscape of tumorigenesis [3]. By integrating data from public repositories such as the Gene Expression Omnibus (GEO), computational biologists can identify Differentially Expressed Genes (DEGs) that drive cancer progression and metastasis [4].

However, a single gene often provides limited diagnostic value due to the complex interaction networks within cells. Consequently, network-centric approaches, such as Protein-Protein Interaction (PPI) network analysis, have become essential for identifying "hub genes"–key nodes that regulate critical biological pathways [5-6]. Hub genes are often functionally more significant than non-hub genes and have been shown to serve as more stable diagnostic and prognostic markers [7].

In this study, we employed an integrated bioinformatics approach to identify key candidate genes involved in breast carcinoma. We analyzed the gene expression dataset GSE10810 to screen for significant DEGs between tumor and normal tissues. Subsequently, we constructed a PPI network to identify core hub genes and validated their prognostic value using the Kaplan-Meier Plotter database. The identification of these molecular signatures may provide new insights into the mechanisms of breast cancer tumorigenesis and highlight potential targets for precision medicine.

2. Materials and Methods:

Microarray Data Acquisition

The gene expression profile dataset GSE10810 was obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). The platform used was the [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array (GPL570). A subset of 10 samples was selected for this study to ensure balanced grouping, comprising 5 breast cancer tissue samples and 5 matched normal tissue samples.

Identification of Differentially Expressed Genes (DEGs)

Differential expression analysis was performed using the GEO2R interactive tool, which utilizes the limma (Linear Models for Microarray Data) R package to identify significant changes in gene expression between tumor and control groups[8-9]. The cut-off criteria for screening DEGs were set as an adjusted P-value < 0.05 and a |logFC| (log fold change) > 1.5. Genes with logFC > 1.5 were classified as upregulated, while those with logFC < -1.5 were classified as downregulated.

Protein-Protein Interaction (PPI) Network Construction

To evaluate the functional interactions among the DEGs, a PPI network was constructed using the STRING (Search Tool for the Retrieval of Interacting Genes) database (version 12.0; https://string-db.org/) [10]. The organism was set to Homo sapiens, and a minimum required interaction score of 0.4 (medium confidence) was applied. Disconnected nodes were hidden to simplify the network structure.

Hub Gene Identification and Visualization

The PPI network data was exported and visualized using Cytoscape software (version 3.9.1) [11]. The CytoHubba plugin was employed to identify key hub genes within the network [12]. The Maximal Clique Centrality (MCC) algorithm, which has been shown to be one of the most effective methods for finding essential nodes in biological networks, was used to rank the top 10 genes.

Survival Analysis

The prognostic value of the identified hub genes was validated using the Kaplan-Meier Plotter database (http://kmplot.com/analysis/), which integrates gene expression data and clinical information from breast cancer patients [13]. We analyzed the correlation between gene expression levels and Relapse-Free Survival (RFS). Patients were divided into high- and low-expression groups based on the "auto select best cutoff" algorithm. The Hazard Ratio (HR) with 95% confidence intervals and log-rank P-values were calculated, with $P < 0.05$ considered statistically significant.

3. Results:

Identification of DEGs in Breast Cancer

Based on the analysis of the GSE10810 dataset using GEO2R, a total of 5062 differentially expressed genes (DEGs) were identified using the criteria of $|logFC| > 1.5$ and adjusted P-value < 0.05. Among these, 1196 genes were significantly upregulated, while 492 genes were downregulated. The distribution of these genes is visualized in the volcano plot (Figure 1), which illustrates the magnitude of fold change versus statistical significance.

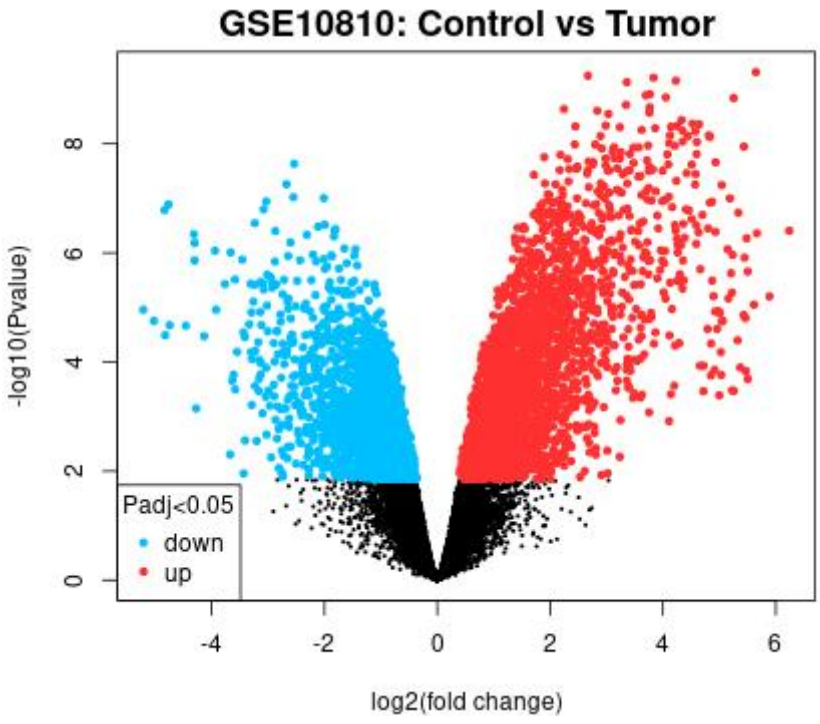

Figure 1. Identification of Differentially Expressed Genes (DEGs) in breast carcinoma. *Volcano plot visualizing the gene expression variation between breast cancer tissues and normal control samples from the GSE10810 dataset. The x-axis represents the log2 Fold Change (logFC), and the y-axis represents the negative log10 of the adjusted P-value. The horizontal line indicates the significance threshold ($P < 0.05$), and the vertical lines indicate the fold-change threshold ($|logFC| > 1.5$). Red points represent significantly upregulated genes, while blue points represent significantly downregulated genes.*

PPI Network Construction and Hub Gene Analysis

The identified DEGs were mapped to the STRING database to construct a protein-protein interaction (PPI) network. The resulting network (Figure 2) exhibited complex interconnections, representing the functional landscape of the cancer-associated proteome.

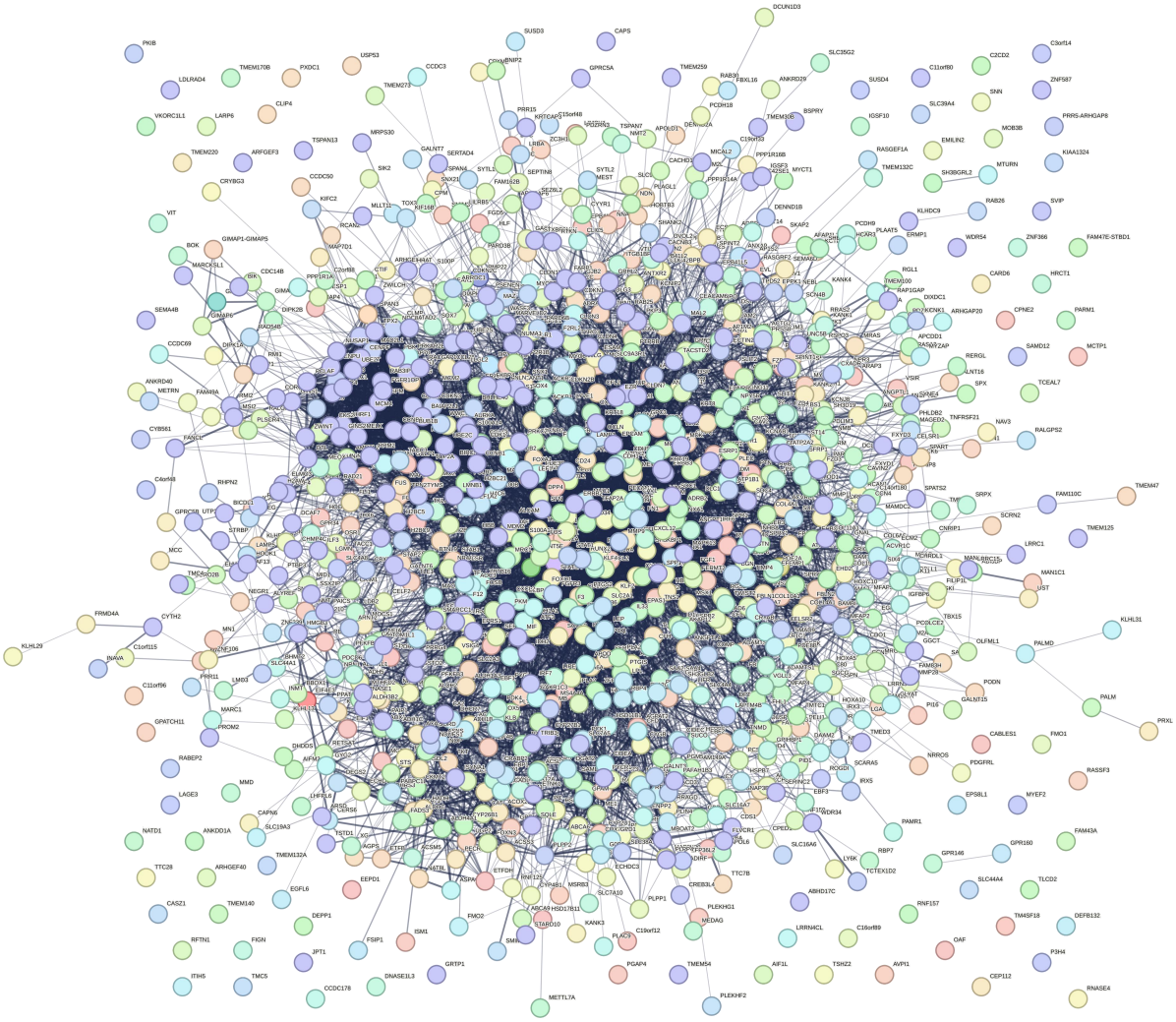

Figure 2. Protein-Protein Interaction (PPI) network construction. *The global interaction network of the identified DEGs was constructed using the STRING database (version 12.0). The network includes upregulated and downregulated genes with a minimum interaction confidence score of 0.4 (medium confidence). Each node*

represents a protein, and each edge represents a predicted functional association. Disconnected nodes were hidden to simplify the network visualization.

To identify the most critical nodes within this network, the CytoHubba plugin in Cytoscape was utilized. Based on the Maximal Clique Centrality (MCC) algorithm, the top 10 hub genes were identified. These genes include: MCM4, CCNB2, MAD2L1, SPAG5, NUSAP1, UBE2C, BIRC5, RRM2, BUB1B, and PBK. The localized network of these top 10 hub genes is shown in Figure 3, where the red color indicates a higher MCC score.

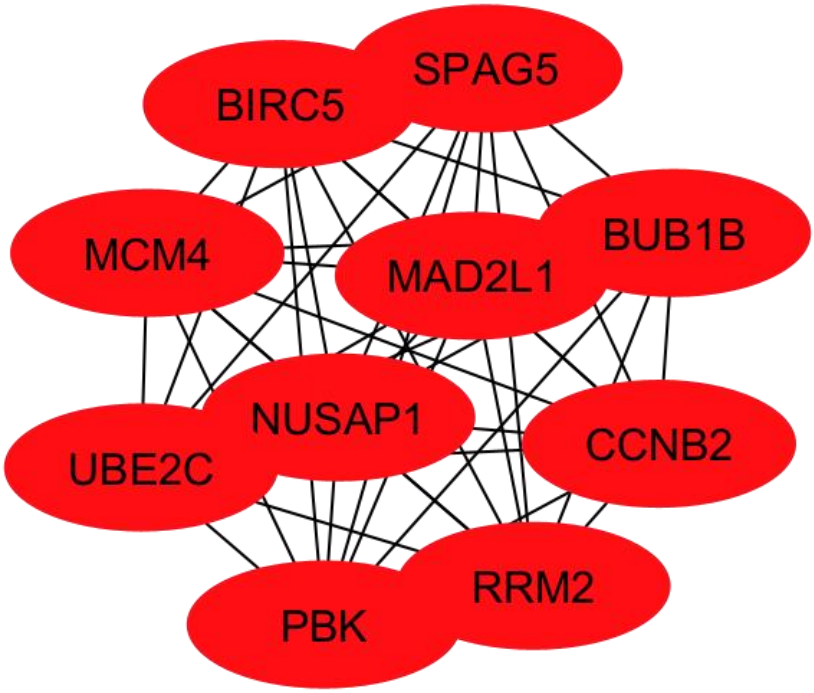

Figure 3. Identification of the top 10 Hub Genes. *The sub-network of the top 10 hub genes was identified using the CytoHubba plugin in Cytoscape. The nodes were ranked based on the Maximal Clique Centrality (MCC) algorithm. The color intensity corresponds to the MCC score, with red indicating the highest ranking. The identified hub genes are MCM4, CCNB2, MAD2L1, SPAG5, NUSAP1, UBE2C, BIRC5, RRM2, BUB1B, and PBK.*

Survival Analysis and Clinical Validation

To evaluate the clinical significance of the hub genes, we performed survival analysis using the Kaplan-Meier Plotter database. The results revealed that all 10 hub genes were significantly associated with patient outcomes. High expression of BIRC5 (HR = 1.59, P = 2.3e-14), UBE2C (HR = 1.84, P = <1e-16), and CCNB2 (HR = 2.25, P = <1e-16) was strongly correlated with reduced Relapse-Free Survival (RFS) in breast cancer patients

(Figure 4). These findings suggest that the identified hub genes are not only central to the biological network but also serve as potent prognostic indicators in a clinical setting.

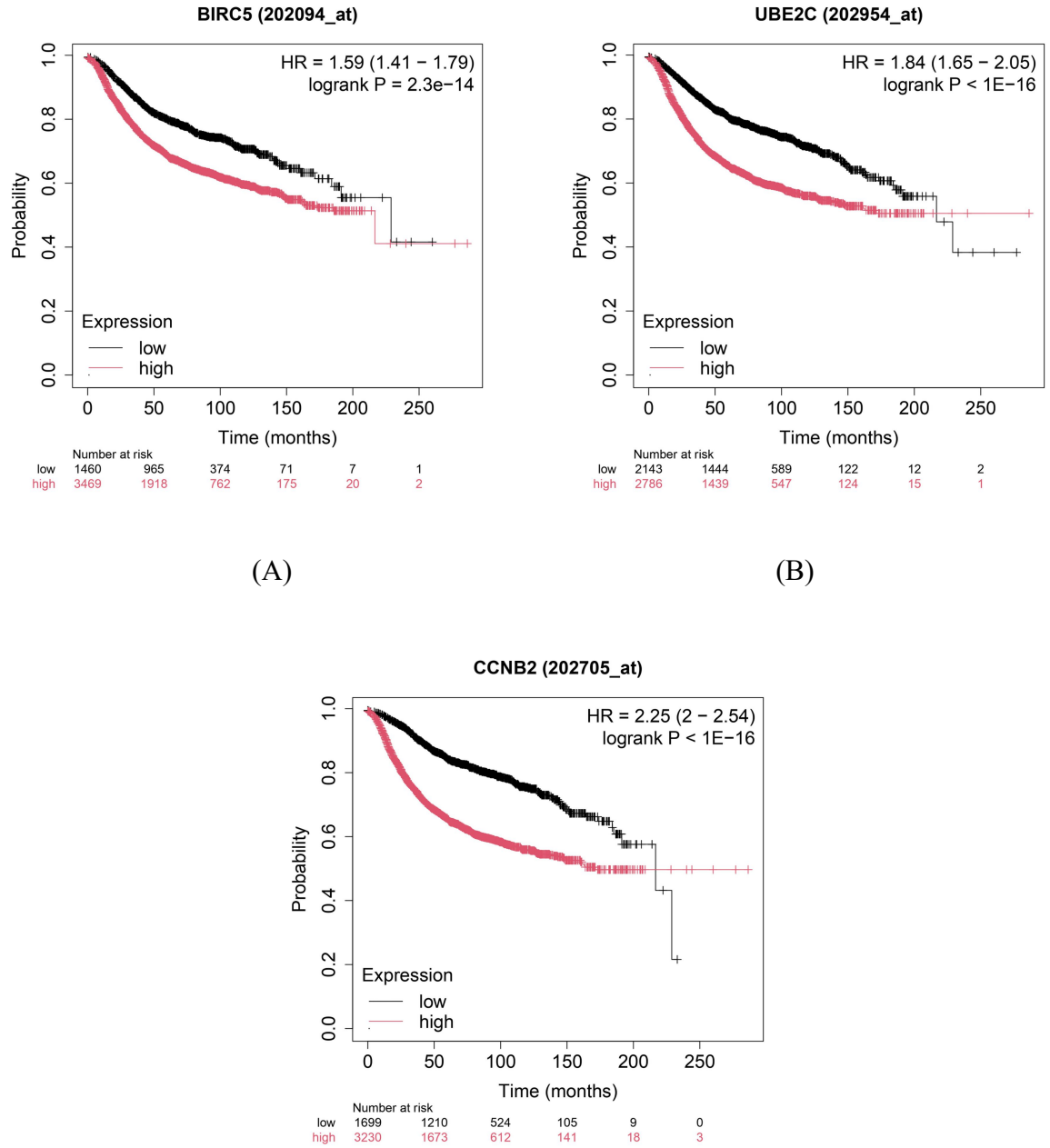

Figure 4. Prognostic value of key hub genes in breast cancer. *Kaplan-Meier survival curves analyzing the correlation between hub gene expression and Relapse-Free Survival (RFS) in breast cancer patients (n=4,929). Patients were stratified into high-expression (red line) and low-expression (black line) groups based on auto-selected cutoff values. (A) BIRC5 (Survivin), (B) UBE2C, and (C) CCNB2. P-values were calculated using the log-rank test, and Hazard Ratios (HR) are displayed to indicate relative risk. All three*

genes show a statistically significant correlation (P < 0.05) between high expression and reduced survival time.

4. Discussion:

The identification of robust molecular signatures is a critical step in advancing the diagnosis and treatment of breast carcinoma. In this study, we utilized an integrated bioinformatics workflow to analyze the transcriptomic profile of breast cancer samples from the GSE10810 dataset. By combining differential expression analysis with protein-protein interaction (PPI) network modeling, we successfully identified a core signature of 10 hub genes–BIRC5, UBE2C, CCNB2, MAD2L1, SPAG5, NUSAP1, MCM4, RRM2, BUB1B, and PBK–that are significantly upregulated in tumor tissues and strongly associated with poor patient prognosis.

Functional analysis of these hub genes reveals a distinct convergence on biological processes related to cell cycle regulation, mitotic spindle assembly, and chromosomal segregation. This suggests that the primary mechanism driving tumorigenesis in this cohort is aberrant cell proliferation and genomic instability.

Among the top-ranked genes, BIRC5 (Survivin) stands out as a critical node. As a member of the inhibitor of apoptosis (IAP) family, BIRC5 functions to prevent cell death and regulate the mitotic spindle checkpoint. Our survival analysis confirms that high BIRC5 expression is strongly correlated with reduced Relapse-Free Survival (RFS). This aligns with previous studies indicating that BIRC5 overexpression is a hallmark of aggressive breast cancer phenotypes and resistance to chemotherapy.

Similarly, UBE2C and CCNB2 were identified as central regulators in our network. UBE2C is a ubiquitin-conjugating enzyme that targets cell cycle proteins for degradation, thereby driving the progression from metaphase to anaphase. Its overexpression has been linked to the failure of the spindle assembly checkpoint, leading to aneuploidy and tumor progression. CCNB2 (Cyclin B2), a key regulator of the G2/M transition, was also found to be upregulated. The co-expression of these genes highlights a disrupted cell cycle machinery that allows cancer cells to bypass normal growth checkpoints.

We also observed significant involvement of spindle assembly checkpoint (SAC) components, specifically MAD2L1 and BUB1B. While these genes are essential for normal cell division, their overexpression in cancer cells can paradoxically promote chromosomal instability (CIN),

a driver of tumor heterogeneity and metastasis. The identification of these genes in our network suggests that the tumor samples analyzed in GSE10810 likely possess a high degree of proliferative activity and genomic instability.

Limitations

It is important to acknowledge the limitations of this study. First, the sample size was restricted to 10 samples (5 tumor, 5 normal), which may introduce selection bias. However, the high concordance of our findings with large-scale databases (like KM Plotter, n=4,929) supports the validity of these hub genes. Second, this study relies solely on computational analysis. While the prognostic value of these genes was validated in silico, further experimental verification using qPCR or Western Blotting in clinical samples is required to confirm their biological function.

Conclusion

In conclusion, this study identifies a 10-gene signature centered around BIRC5 and UBE2C as a potential prognostic biomarker for breast carcinoma. These genes play pivotal roles in the dysregulation of the cell cycle and are predictive of poor survival outcomes. Future studies should focus on validating this signature as a target for novel therapeutic interventions.

References:

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024 May-Jun;74(3):229-263. doi: 10.3322/caac.21834. Epub 2024 Apr 4. PMID: 38572751.

[2] Zardavas D, Irrthum A, Swanton C, Piccart M. Clinical management of breast cancer heterogeneity. Nat Rev Clin Oncol. 2015 Jul;12(7):381-94. doi: 10.1038/nrclinonc.2015.73. Epub 2015 Apr 21. PMID: 25895611.

[3] Tao Z, Shi A, Li R, Wang Y, Wang X, Zhao J. Microarray bioinformatics in cancer- a review. J BUON. 2017 Jul-Aug;22(4):838-843. PMID: 29155508.

[4] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10. doi: 10.1093/nar/30.1.207. PMID: 11752295; PMCID: PMC99122.

[5] Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004 Feb;5(2):101-13. doi: 10.1038/nrg1272. PMID: 14735121.

[6] Wang S, Wu R, Lu J, Jiang Y, Huang T, Cai YD. Protein-protein interaction networks as miners of biological discovery. Proteomics. 2022 Aug;22(15-16):e2100190. doi: 10.1002/pmic.202100190. Epub 2022 May 24. PMID: 35567424.

[7] He X, Zhang J. Why do hubs tend to be essential in protein networks? PLoS Genet. 2006 Jun 2;2(6):e88. doi: 10.1371/journal.pgen.0020088. Epub 2006 Apr 26. PMID: 16751849; PMCID: PMC1473040.

[8] Clough E, Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Lee H, Zhang N, Serova N, Wagner L, Zalunin V, Kochergin A, Soboleva A. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. Nucleic Acids Res. 2024 Jan 5;52(D1):D138-D144. doi: 10.1093/nar/gkad965. PMID: 37933855; PMCID: PMC10767856.

[9] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47. doi: 10.1093/nar/gkv007. Epub 2015 Jan 20. PMID: 25605792; PMCID: PMC4402510.

[10] Szklarczyk D, Nastou K, Koutrouli M, Kirsch R, Mehryary F, Hachilif R, Hu D, Peluso ME, Huang Q, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C. The STRING database in 2025: protein networks with directionality of regulation. Nucleic Acids Res. 2025 Jan 6;53(D1):D730-D737. doi: 10.1093/nar/gkae1113. PMID: 39558183; PMCID: PMC11701646.

[11] Kohl M, Wiese S, Warscheid B. Cytoscape: software for visualization and analysis of biological networks. Methods Mol Biol. 2011;696:291-303. doi: 10.1007/978-1-60761-987-1_18. PMID: 21063955.

[12] Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol. 2014;8 Suppl 4(Suppl 4):S11. doi: 10.1186/1752-0509-8-S4-S11. Epub 2014 Dec 8. PMID: 25521941; PMCID: PMC4290687.

[13] Győrffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, Szallasi Z. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Res Treat. 2010 Oct;123(3):725-31. doi: 10.1007/s10549-009-0674-9. Epub 2009 Dec 18. PMID: 20020197.