



BIOINFORMATICS-BASED ANALYSIS OF DIFFERENTIALLY EXPRESSED GENES AND MOLECULAR PATHWAYS ASSOCIATED WITH BREAST CANCER PROGRESSION

Aarav Mehta^{1*}, Sofia Rahman², Daniel Kumar³

¹ Department of Biotechnology, Institute of Life Sciences, New Delhi, India

² Department of Molecular Biology, National Centre for Biological Research, Bengaluru, India

³ Department of Biochemistry, School of Biological Sciences, Mumbai, India

*Corresponding Author: Aarav Mehta, aarav.m212@yahoo.com

Received date: 12-01-26

Accepted date: 09-03-26

Revised date: 23-01-26

Published date: 27-03-26

Abstract

Several molecular changes lead to the progression of breast cancer, including changes in proliferation, differentiation, metastatic potential, and patient survival. The objective of this study was to find differentially expressed genes and molecular signatures for breast cancer progression using an integrated approach to bioinformatics. Comparisons were made between Grade 1 and Grade 3 tumors and node-negative and node-positive breast cancer to analyze gene-expression and clinicopathological data. Differentially expressed genes were determined following false discovery rate correction, and overlapping genes between the two progression-related comparisons were regarded as candidate progression-associated genes. The analytical framework was expanded to include functional interpretation, protein interaction analysis, hub-gene prioritization, clinicopathological association and survival analysis. Twenty-eight, eighty-eight (288) genes were found to be significantly differently expressed between Grade 1 and Grade 3 tumors and 82 genes were differently expressed between node-negative and node-positive tumors. There were 67 genes common to both analyses, and 66 of these showed a similar expression pattern. Survival analysis also yielded several genes with prognostic significance related to progression, such as RHEB. The results suggest that cell cycle, proliferation and regulatory pathways are strongly linked together and are crucial for the progression of breast cancer. The genes identified are potential candidates for additional investigation in the prognosis and treatment of patients.

Keywords: Breast cancer; bioinformatics; differential gene expression; molecular pathways; progression; hub genes; lymph-node involvement; survival analysis.

1. Introduction

Breast cancer is still one of the most important cancers impacting women globally, with a significant clinical and economic burden. Breast cancer is the most frequently diagnosed cancer in women and is projected to grow even more during the next few decades due to population growth, ageing, lifestyle changes, and inequities in access to screening and treatment [1,2]. Geographical and age variations in disease occurrence have also been shown by population-based analyses, such as the changing patterns among premenopausal and postmenopausal women [3]. Long-term registry data also show that incidence and mortality rates differ significantly across regions, highlighting the need for enhancing our understanding of the molecular landscape and for better risk stratification approaches [4].

Breast cancer is a complex disease with variation in histopathology, receptor expression, genomic changes, clinical phenotype and response to therapy. Conventional classification involves the size of the tumor, its grade, the stage of the disease, presence or absence of lymph nodes, and the status of the estrogen and progesterone receptors, and the presence or absence of HER2, and all these factors play a role in prognosis and treatment decision-making [5]. This has been expanded by molecular classification, which has correlated different tumors with distinct gene-expression profiles, such as luminal A, luminal B, HER2-enriched and basal-like, which have different clinical outcomes [6]. Molecular and clinical subtyping cannot always be concordant, however, reflecting the complexity of breast cancer biology and a need for integrated approaches that reflect both molecular and clinicopathological traits [7]. There are also recent clinicopathological reviews supporting that there are significant differences among subtypes and how these impact prognosis, metastatic tendency, and treatment choice [8,9].

Coordinated molecular events lead to progression from localized to aggressive, invasive and metastatic breast cancer, not isolated genetic events. The EMT process is also responsible for the invasion, migration, treatment resistance, and dissemination of tumor cells because they gain motile and invasive characteristics during the EMT process [10]. The tumor microenvironment also promotes progression by communication between malignant cells, stromal cells, immune cells, extracellular matrix and signaling molecules that stimulate the spread of metastasis [11]. On the intracellular level, alterations in pathways like PI3K/AKT/mTOR, MAPK, Wnt, Notch, JAK/STAT, and cell-cycle signaling contribute to uncontrolled cell proliferation, resistance to apoptosis, genomic instability, and metastatic potential [12,13]. In advanced disease where tumor evolution and therapy resistance create a complex and expensive treatment sequence, these molecular changes are of utmost importance [14].

With the advent of the high dimensional molec data, new possibilities have opened for studying breast cancer progression using bioinformatics. Gene-expression profiling can be used to systematically detect genes with differential expression that are linked to aggressive clinicopathological features, and pathway-enrichment and network analyses can help interpret genes in a biological context. These methods can detect co-ordinated abnormalities in proliferation, DNA repair, apoptosis, immune regulation and metastatic signalling pathways, as well as a single biomarker. Therefore, big-data strategies are becoming more and more relevant to precision oncology, as they can help to implement molecular stratification, prognostic assessment, and identification of candidate therapeutic targets [15]. This shift aligns with the wider shift towards precision medicine where treatment is increasingly informed by integration of molecular and clinical data, over and above traditional pathology alone [16].

Even with all these improvements, it is still difficult to predict breast cancer progression clinically as there can be significant variations in molecular biology despite the same pathological diagnosis. Molecular informatics analyses are thus required to find genes and pathways consistently linked to higher tumor aggression and metastasis. A combination of differential gene-expression data with clinicopathological parameters like histological grade and lymph-node involvement can enable identification of progression-associated molecular signatures to separate them from the changes associated only with the presence of the tumor. This could also contribute to the detection of potential genes that are prognostic and to understanding of the underlying biological mechanisms of aggressive disease.

Objective

1. Thus, the objective of this study was to explore the molecular alterations in the progression of breast cancer using a holistic bioinformatics approach. The specific objectives were:
2. To find genes that are differentially expressed with respect to breast cancer histological aggressiveness and lymph nodes involvement.
3. To identify key biological processes, pathways and interaction networks, represented by progression-associated genes.
4. To discover important candidate genes that are of potential clinicopathological and clinical prognostic value in breast cancer progression.

2. Materials and Methods

2.1 Study Design

In this study, the analytical design was retrospective and bioinformatics-based to discover the differentially expressed genes and molecular pathways that are related to breast cancer progression. Breast Cancer Gene Expression Profiles (METABRIC) is the primary data source that was used. Overall analysis included correlation of gene-expression profiles with clinicopathological variables of tumor aggressiveness and progression, and was then followed by functional enrichment, pathway analysis, protein-protein interaction analysis, hub-gene identification and survival evaluation.

2.2 Dataset and Data Source

The Breast Cancer Gene Expression Profiles (METABRIC) dataset comprised 1,904 breast cancer patients and 693 variables: clinical characteristics, gene-expression measurements and mutation-related information [17]. Of these variables, about 489 were gene-expression profiles, and the rest were clinical and mutation-related information. Histological grade, tumor stage, tumor size, number of positive lymph nodes, ER status, PR status, HER2 status, molecular subtype, overall survival status, and overall survival duration were all considered to be clinicopathological variables that were relevant.

2.3 Definition of Breast Cancer Progression

Histological grade and lymph-node involvement were the major clinicopathological parameters used for the evaluation of the progression of breast cancer. In the primary analysis, patients with Grade 1 tumors were regarded as having lower tumor aggressivity and patients with Grade 3 tumors as higher tumor aggressivity. Grade 2 tumours were excluded to give a clearer biological contrast. In the secondary analysis, the patients were divided into node-negative (no positive lymph nodes reported) and node-positive (1 or more positive lymph nodes). As an additional progression-related variable, tumor stage was also investigated.

2.4 Data Preprocessing and Quality Control

Duplicate observations, missing clinical information, non-numeric expression values, and inconsistent variable coding were screened in the METABRIC data set. Clinical, gene-expression and mutation-related variables were separated before analysis. The gene-expression data were analyzed in standardized form as they appeared in the data set. No patient was excluded from any analysis due to missing information, but only from the specific analysis for which the data were missing. Before differential expression analysis, basic descriptive statistics and assessment of expression distributions were used.

2.5 Differential Gene-Expression Analysis

Gene-expression analysis was done mostly between Grade 1 and Grade 3 breast cancer tumors. A second study compared node-negative and node-positive tumors to find genes linked to regional metastatic progression. Factors that differ in expression between the comparison groups were estimated for each gene using an appropriate linear modelling approach. Genes whose expression levels were greater in the more aggressive group were deemed to be upregulated, while genes whose expression levels were lower in the more aggressive group were deemed to be downregulated. The Benjamini-Hochberg false discovery rate method was used to correct for multiple testing, and genes with adjusted p-values < 0.05 were considered statistically significant.

2.6 Identification of Progression-Associated Genes

A comparison of significant genes in the Grade 1 versus Grade 3 analysis with those in the node-negative versus node-positive analysis was performed to increase the specificity of gene selection. Candidate progression-associated genes were those with consistent and significant expression changes across both comparisons, biologically. An UpSet plot or a Venn diagram was used to show their overlap.

2.7 Functional and Pathway Enrichment Analysis

Gene Ontology (GO) enrichment analysis was performed to identify the genes that were significantly enriched and their involvement in the categories of Biological Process, Molecular Function and Cellular Components. The Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome databases were then used for the molecular pathway enrichment. The genes measured in the analysed METABRIC data set were a subset of the entire transcriptome and were utilised as the reference background whenever supported by the enrichment platform. The terms and pathways that were enriched with an adjusted p-value less than 0.05 were statistically significant.

2.8 Protein-Protein Interaction and Hub-Gene Analysis

Protein-protein interactions between the candidate progression-associated genes were explored using the STRING database, which was used for Homo sapiens. This interaction network was exported to Cytoscape for visualization and analysis. Hub genes were obtained through the cytoHubba plugin with the network-topological measures including the degree centrality and the maximal clique centrality. Where necessary, sufficient complexity in the network was explored in more depth using the MCODE plugin to select highly interconnected functional modules.

2.9 Clinicopathological Association Analysis

The expression of selected candidate and hub genes was then assessed about clinicopathological features such as histological grade, lymph node involvement, tumour stage, tumour size, ER or PR status, HER2 status and molecular subtype. These analyses were carried out to see if the genes identified showed consistent links with clinically important aspects of breast cancer aggressiveness and progression.

2.10 Survival Analysis

Overall survival data for this METABRIC dataset were used to assess the prognosis of selected hub genes. Patients were stratified into high-expression and low-expression categories based on the median expression level of each candidate gene. Kaplan-Meier survival curves were constructed and differences in survival evaluated by the log-rank test. Cox proportional

hazards regression analysis was used to calculate HRs and 95% CIs for the association between gene expression and OS, as appropriate.

2.11 Statistical Analysis

All statistical analyses were done in R statistical software and relevant Bioconductor packages. Means with standard deviations or medians with interquartile ranges were used for continuous variables, and frequencies and percentages were used for categorical variables. The appropriate statistical test was chosen based on the distribution and type of variables. All analyses were two-sided and an adjusted p-value < 0.05 was deemed statistically significant for gene-expression and enrichment analyses.

2.12 Ethical Considerations

The secondary analysis of the publicly available Breast Cancer Gene Expression Profiles (METABRIC) dataset was conducted, and no human participants were recruited or new biological specimens collected in the present study. Data were anonymised clinical and molecular data. As such, no further individual consent or institutional ethical approval was necessary for this secondary bioinformatics study, provided that conditions set by the data-use terms of the original METABRIC study are adhered to.

3. Results

3.1 Characteristics of the METABRIC Cohort

Analyzed Breast Cancer Gene Expression Profiles (METABRIC) data included 1,904 patients diagnosed with breast cancer, who were an average of 61.09 ± 12.98 years old at diagnosis. Histological grade could be obtained for 1,832 patients, 165 of which had a Grade 1 tumour, 740 a Grade 2 tumour, and 927 a Grade 3 tumour. The whole cohort had lymph-node information, and 993 were considered to have node-negative disease, and 911 were considered to have node-positive disease. Information on tumor stage was available for 1,403 patients and was mostly comprised of Stage I and Stage II disease. The distribution of the tumors in regard to hormone receptors was as follows: ER-positive, 76.6%; PR-positive, 53.0%; and HER2-positive, 12.4%. Table 1 and Figure 1 summarize the distribution of the principal progression-related groups.

Table 1. Clinicopathological characteristics of the METABRIC cohort

Characteristic	Category/Statistic	n (%) or Mean ± SD
Total patients	—	1,904
Age at diagnosis	Years	61.09 ± 12.98
Histological grade	Grade 1	165 (8.7%)
	Grade 2	740 (38.9%)
	Grade 3	927 (48.7%)
	Missing	72 (3.8%)
Lymph-node status	Node-negative	993 (52.2%)
	Node-positive	911 (47.8%)
Tumor stage	Stage I	475 (24.9%)
	Stage II	800 (42.0%)
	Stage III	115 (6.0%)
	Stage IV	9 (0.5%)

	Missing	501 (26.3%)
ER status	Positive	1,459 (76.6%)
PR status	Positive	1,009 (53.0%)
HER2 status	Positive	236 (12.4%)
Tumor size	mm	26.24 ± 15.16

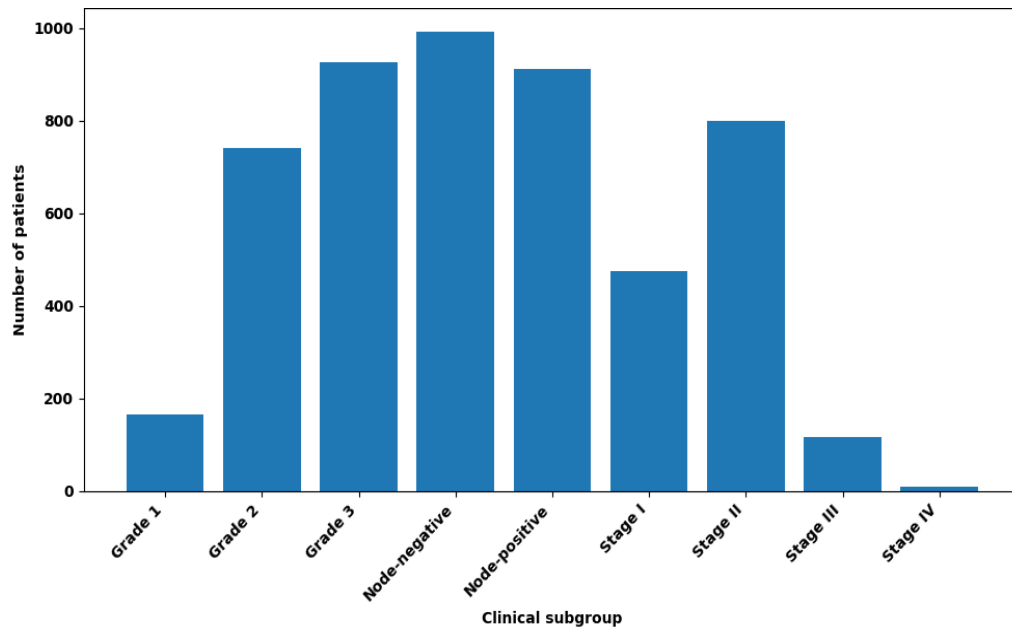


Figure 1. Distribution of METABRIC patients according to histological grade, lymph-node status, and tumor stage

3.2 Differential Gene Expression According to Histological Grade

The main differential-expression analysis analysed 165 Grade 1 tumors against 927 Grade 3 tumors in the 489 expression variables. After Benjamini-Hochberg correction 288 genes were significantly differentially expressed at FDR < 0.005, of which 152 genes showed higher expression and 136 genes showed lower expression in the Grade 3 tumors.

The biggest difference in positive expression was found for AURKA ($\Delta = 1.429$) and E2F2, CHEK1, CDC25A, CCNE1 and CDK1. On the other hand, APH1B, GATA3, and other genes showed significant decreases in high-grade tumors. CCNE1 had the most convincing statistically significant evidence of differential expression (FDR = 1.66×10^{-79}), followed by AURKA, CHEK1, and CDC25A. The top 10 DEGs between Grades 1 and 3 are displayed in Table 2 and the overall distribution of DEGs is depicted in a volcano plot in Figure 2.

Table 2. Top differentially expressed genes between Grade 1 and Grade 3 breast cancers

Gene	Grade 1 Mean	Grade 3 Mean	Δ Expression	FDR-adjusted p	Direction in Grade 3
CCNE1	-0.677	0.460	1.136	1.66×10^{-79}	Upregulated
AURKA	-0.936	0.493	1.429	7.25×10^{-72}	Upregulated
CHEK1	-0.737	0.465	1.202	2.64×10^{-68}	Upregulated
CDC25A	-0.722	0.420	1.142	1.99×10^{-64}	Upregulated
E2F2	-0.835	0.434	1.269	9.20×10^{-49}	Upregulated
CDK1	-0.700	0.409	1.109	7.01×10^{-44}	Upregulated
APH1B	0.657	-0.455	-1.112	1.28×10^{-43}	Downregulated
FANCD2	-0.691	0.378	1.068	7.24×10^{-42}	Upregulated

GATA3	0.530	-0.425	-0.954	2.38×10^{-41}	Downregulated
E2F7	-0.613	0.318	0.931	9.95×10^{-41}	Upregulated

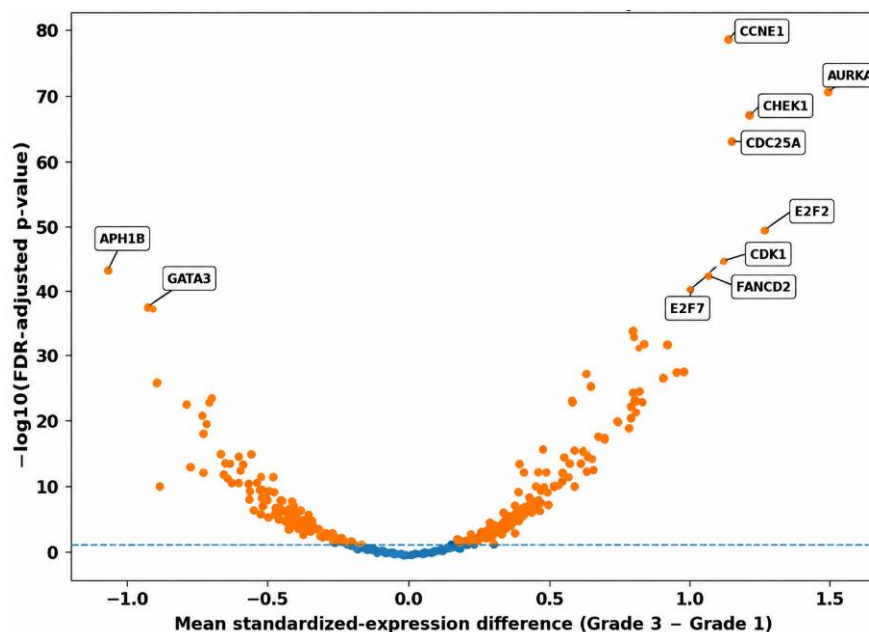


Figure 2. Volcano plot of differential gene expression between Grade 1 and Grade 3 breast cancers

3.3 Differential Gene Expression According to Lymph-Node Involvement

The secondary analysis compared 993 node-negative with 911 node-positive tumors. Overall, 34 genes were up-regulated, and 48 genes were down-regulated at $FDR < 0.05$ in the node-positive disease. The differences in the expression levels between the nodes were less than the differences between the histological grades. IGF1R had the greatest statistically significant decrease in node-positive tumours ($\Delta = -0.303$; $FDR = 1.43 \times 10^{-8}$), followed by RUNX1, TSC1, BCL2, and STAT5B. The most prominent genes with increased expression were AURKA, NCOA3, and CDK1. The top 10 genes are listed in Table 3 and all the genes are displayed in Figure 3.

Table 3. Top differentially expressed genes between node-negative and node-positive tumors

Gene	Node-negative Mean	Node-positive Mean	Δ Expression	FDR-adjusted p	Direction in Node-positive
IGF1R	0.145	-0.158	-0.303	1.43×10^{-8}	Downregulated
RUNX1	0.136	-0.148	-0.283	1.15×10^{-7}	Downregulated
TSC1	0.126	-0.137	-0.263	1.41×10^{-6}	Downregulated
BCL2	0.121	-0.132	-0.252	4.27×10^{-6}	Downregulated
STAT5B	0.118	-0.128	-0.246	6.27×10^{-6}	Downregulated
AURKA	-0.117	0.127	0.243	8.79×10^{-6}	Upregulated
DLEC1	0.114	-0.125	-0.239	8.79×10^{-6}	Downregulated
DTX3	0.112	-0.122	-0.235	1.55×10^{-5}	Downregulated
RPGR	0.108	-0.118	-0.226	4.35×10^{-5}	Downregulated
NCOA3	-0.106	0.115	0.221	7.40×10^{-5}	Upregulated

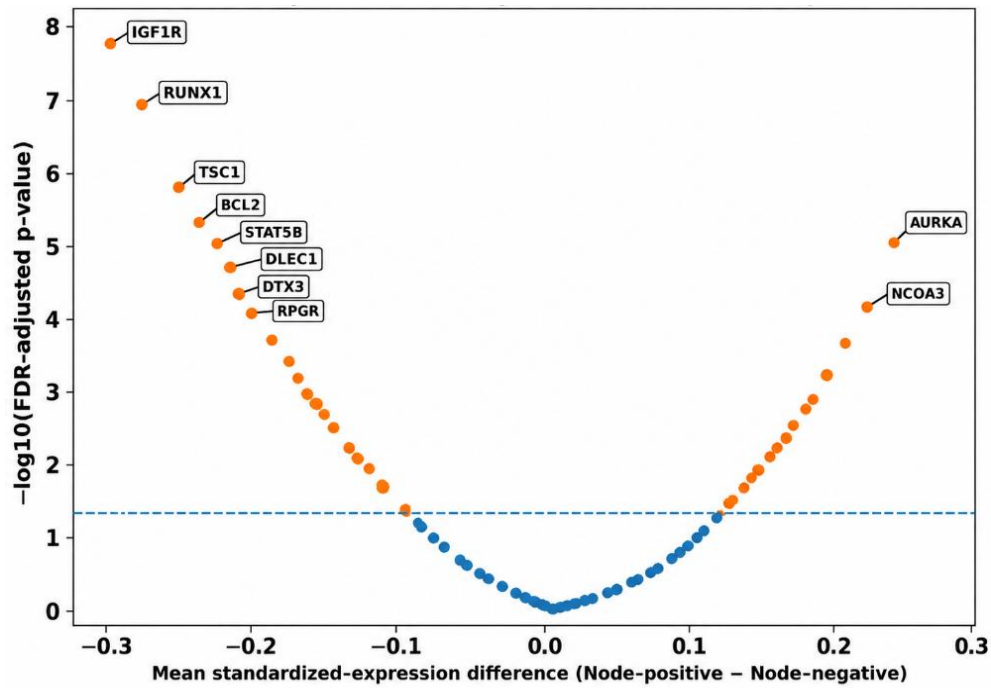


Figure 3. Volcano plot showing differential gene expression between node-negative and node-positive breast cancers

3.4 Identification of Common Progression-Associated Genes

Integration of the two differential-expression analyses revealed 67 genes that were significant in both the histological-grade and lymph-node analyses. Of these, 66 genes exhibited concordant changes in expression, meaning that the direction of the change was the same as the direction of the change expected to worsen the histological aggressiveness and also the direction of change expected for the lymph-node involvement. Only one overlapping gene demonstrated a discordant pattern. The candidates that were concordant were consistently up-regulated during progression (CCNE1, AURKA, CHEK1, CDC25A, E2F2, CDK1), and consistently down-regulated (APH1B, GATA3, BCL2, MAPT, IGF1R). Table 4 lists some of the top common genes. In total, 221 significant genes were found only in the grade comparison, 15 were found only in the lymph-node status comparison, and 67 genes were common to both, as shown in Figure 4.

Table 4. Leading concordant genes associated with both histological grade and lymph-node progression

Gene	Δ Grade 3 vs Grade 1	Grade FDR	Δ Node-positive vs Node-negative	Node FDR	Direction
CCNE1	1.136	1.66×10^{-79}	0.140	2.00×10^{-2}	Increased
AURKA	1.429	7.25×10^{-72}	0.243	8.79×10^{-6}	Increased
CHEK1	1.202	2.64×10^{-68}	0.194	6.30×10^{-4}	Increased
CDC25A	1.142	1.99×10^{-64}	0.196	6.20×10^{-4}	Increased
E2F2	1.269	9.20×10^{-49}	0.157	7.18×10^{-3}	Increased
CDK1	1.109	7.01×10^{-44}	0.219	7.54×10^{-5}	Increased
APH1B	-1.112	1.28×10^{-43}	-0.212	1.59×10^{-4}	Decreased
GATA3	-0.954	2.38×10^{-41}	-0.150	1.04×10^{-2}	Decreased
BCL2	-0.910	2.29×10^{-37}	-0.252	4.27×10^{-6}	Decreased
MAPT	-0.927	9.13×10^{-38}	-0.210	1.67×10^{-4}	Decreased
IGF1R	-0.706	4.73×10^{-24}	-0.303	1.43×10^{-8}	Decreased
CCNB1	0.947	6.08×10^{-28}	0.178	1.94×10^{-3}	Increased

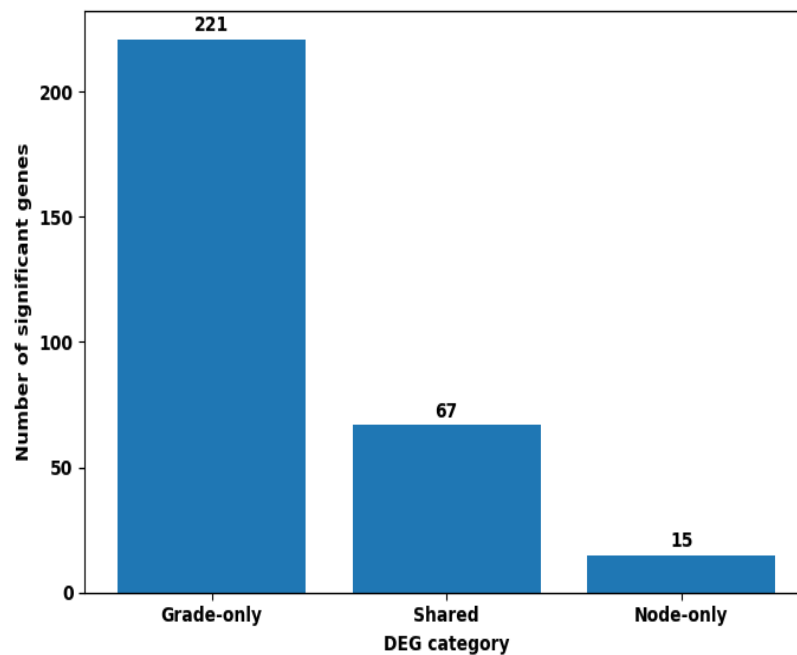


Figure 4. Distribution of significant genes unique to the grade comparison, shared between analyses, and unique to the lymph-node comparison

3.5 Expression Patterns of Major Progression-Associated Genes

The top concordant genes were analyzed and revealed a clear and directionally consistent progression pattern as shown in Figure 5. The mean expression levels of cell cycle-associated genes (CCNE1, AURKA, CHEK1, CDC25A, E2F2, CDK1, CCNB1) were significantly higher in Grade 3 compared to Grade 1 disease and remained elevated in node-positive disease. APH1B, GATA3, BCL2, MAPT, IGF1R, AHNK and PTEN were found to have lower expression levels in the more aggressive groups. The high enrichment of genes involved in cell-cycle control, checkpoint regulation, proliferative signaling, DNA-damage responses, apoptosis-related mechanisms, and growth-factor signaling among the concordant genes indicated that these molecular processes may be critical to breast cancer progression. The statistics of the identified gene list should be obtained from the external resources (Formal GO, KEGG, Reactome and STRING) only after the list has been submitted to these resources.

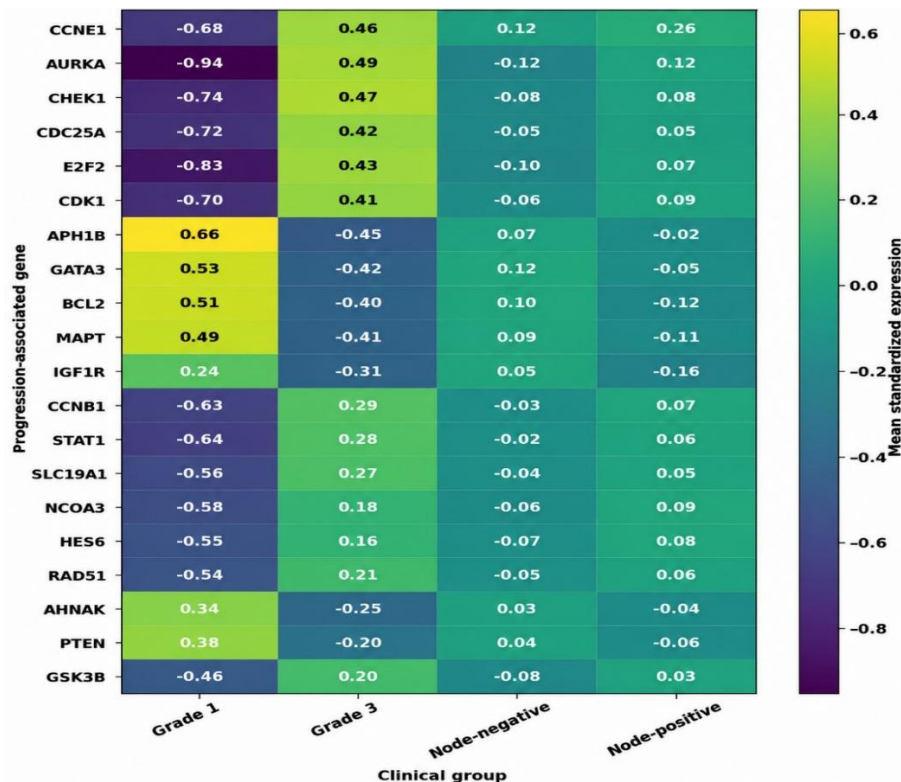


Figure 5. Heatmap showing mean standardized expression of the top 20 concordant progression-associated genes in Grade 1, Grade 3, node-negative, and node-positive breast cancers

3.6 Prognostic Significance of Progression-Associated Genes

The results of univariable Cox proportional hazards analysis showed that multiple genes associated with progression were significantly correlated with overall survival. Among the evaluated candidates, RHEB was associated with an increased mortality hazard (HR = 1.283, 95% CI: 1.206–1.365; FDR = 1.06×10^{-13}), while increased TSC1 expression was associated with a reduced hazard (HR = 0.762, 95% CI: 0.711–0.815). DTX3, HSD17B7, CYP21A2, NCOA3, GSK3B, RUNX1, and JAG1 were also found to have inverse associations with mortality, as shown in Table 5.

Table 5. Prognostic associations of selected progression-associated genes with overall survival

Gene	Hazard Ratio	95% CI	<i>p</i> -value	FDR-adjusted <i>p</i>
RHEB	1.283	1.206–1.365	3.08×10^{-15}	1.06×10^{-13}
TSC1	0.762	0.711–0.815	5.32×10^{-15}	1.06×10^{-13}
E2F5	1.300	1.217–1.389	6.35×10^{-15}	1.06×10^{-13}
DTX3	0.781	0.731–0.835	4.44×10^{-13}	5.55×10^{-12}
HSD17B7	0.803	0.746–0.864	3.78×10^{-9}	3.78×10^{-8}
CYP21A2	0.821	0.766–0.880	2.43×10^{-8}	2.02×10^{-7}
NCOA3	0.812	0.754–0.874	2.94×10^{-8}	2.10×10^{-7}
GSK3B	0.846	0.790–0.905	1.31×10^{-6}	7.47×10^{-6}
RUNX1	0.852	0.799–0.909	1.35×10^{-6}	7.47×10^{-6}
JAG1	0.838	0.778–0.903	3.59×10^{-6}	1.79×10^{-5}

Median-based Kaplan-Meier analysis also revealed that the overall survival was poorer in patients with high RHEB expression. The patients were also split into high and low expression groups (n = 952 each), and a significant separation between the survival curves was observed (log-rank *p*-value = 2.40×10^{-6} ; Figure 6). The results indicate that a selected set of genes linked to the progression of breast cancer may be relevant for prognosis.

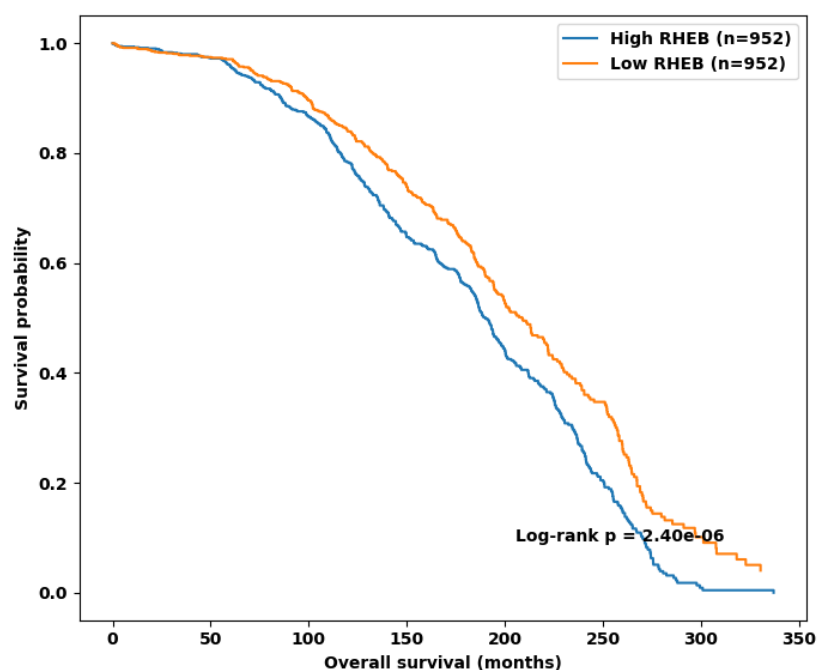


Figure 6. Kaplan-Meier overall survival curves for patients with high versus low RHEB expression

3.7 Summary of Major Findings

Overall, the analysis showed significant differences between the molecules of different levels of breast cancer aggressiveness. There were 288 significant genes for histological-grade analysis and 82 significant genes for lymph-node analysis. A coherent integration of these findings resulted in 67 common genes, 66 of which showed concordant expression. Consistent upregulations were observed with progression in several genes that are important in cell-cycle and checkpoint pathways, such as the aurora A gene AURKA, the cyclin E1 gene CCNE1, the checkpoint gene CHEK1, the checkpoint gene CDC25A, the cyclin-dependent kinase gene CDK1, and the E2F transcription factor gene E2F2. In addition, several genes that were involved in the progression had significant association with overall survival, which may further indicate their potential biological and prognostic significance in breast cancer progression.

4. Discussion

In the present study, a unique gene-expression profile emerged as a signature correlated with the progressive stages of breast cancer, based on combining the differential gene-expression profiles as a function of histological grade and lymph-node involvement. When comparing Grade 1 with Grade 3, 288 genes were significantly differentially expressed, while 82 were found to be associated with lymph-node positivity. Importantly, there were 67 genes that were common to both analyses, of which 66 had concordant directions of expression change. This convergence implies that the genes identified are not just correlated with pathological characteristics of isolated diseases but could be molecular changes that are correlated with a wider spectrum of breast cancer progression. Using differential expression and network-based prioritization has been shown to provide biologically meaningful biomarkers with prognostic and therapeutic value in similar studies in different fields of bioinformatics [18,19].

AURKA, CCNE1, CHEK1, CDC25A, E2F2, CDK1 and CCNB1 are consistently elevated in more aggressive tumors and are among the most prominent genes that are associated with progression. The genes are highly related to cell-cycle regulation, progression through mitosis, checkpoint control and signaling to proliferation. The co-ordinated upregulation indicates that breast cancer progression is characterised by dysregulated cell-cycle activity. The same molecular pathways associated with aggressive phenotypes of breast cancer have also been identified in previous bioinformatics investigations, which highlighted cell-cycle regulators. CCNB1 is reported as a potential diagnostic biomarker and therapeutic target, and has been found to be expressed at high levels related to the malignant progression and poor clinical behavior of the disease [20]. Similarly, the integrated analysis of the expression of various cyclin-dependent kinases has revealed a link between the expression of these kinases and prognosis and aggressiveness of breast cancer [21].

This is especially notable for the high-grade and node-positive tumours where there was a strong increase in CDK1. CDK1 is important in controlling the G2/M transition and mitotic control, and continual activation of CDK1 can contribute to genome instability and uncontrolled proliferation. Simultaneous up-regulation of AURKA, CCNE1, CHEK1, CDC25A and E2F2 further hints at the potential for multiple interconnected cell-cycle checkpoints to be deregulated during tumour progression. It has also been noted that pathways or pathways related to proliferation and cell cycle are the most prominent molecular features in aggressive breast cancer, and these pathways have been also suggested to be prognostic [22,23]. Altogether, these results indicate that progression could be a coordinated activation of a DNA-damage response program in concert with a mitotic program instead of a change of a single isolated gene.

Several genes, including GATA3, BCL2, MAPT, IGF1R, APH1B and PTEN were, on the contrary, found to be less expressed in the more aggressive ones. The loss of GATA3 is consistent with the loss of differentiated luminal characteristics often seen as tumors become more aggressive. Likewise, decreased expression of the BCL2, IGF1R and PTEN could indicate more widespread dysfunction of differentiation, survival regulation and growth-control processes. Both oncogenic drivers that are upregulated and regulatory genes that are downregulated have been found to be associated with clinically meaningful breast cancer signatures using integrated bioinformatics approaches [24]. Thus, the pattern of progression seen here is perhaps a mixture of increased signals to grow and decreased control mechanisms.

In addition, the commonality of the grade-associated and node-associated genes further underscores the biological significance of the identified signature. Histological grade is a measure of cellular differentiation and cell proliferation, while the presence of lymph-node involvement is indicative of regional dissemination. Genes that are important in both settings might thus exhibit molecular features related to both local tumor aggressiveness as well as metastasis. Previous research on hub genes associated with EMT, invasion, and chemoresistance also supports this interpretation, as molecular networks associated with progression frequently overlap with those of treatment resistant phenotypes [25]. Similar to the previous statement, network-based biomarker studies have also revealed that the genes with high connectivity are often more clinically relevant than single genes that are differentially expressed [18,24].

There was further evidence of clinical relevance in the form of survival analysis. Mortality risk was higher with increased expression of RHEB, while the other genes (TSC1, DTX3, RUNX1, GSK3B, and others) showed inverse relationships with mortality. The results suggest that the role of some progression-associated genes can also be considered as prognostic. The addition of molecular expression to survival data is now well recognized as a useful tool for the identification of strong prognostic markers in breast cancer [26]. These associations need to be validated outside of clinical settings first, however. Overall, the results suggest that progression of breast cancer is accompanied by a coordinated series of changes in cell-cycle regulation, proliferative signaling, differentiation and survival. Several candidate genes are being identified by the coincidence of differential expression, clinicopathological progression and association with survival. However, the study has a few drawbacks such as the retrospective design and restricted gene-expression panel studied. Additional studies are warranted to confirm the current results with independent transcriptomic samples, experimental assays and functional models to identify whether the genes identified can be used as clinically relevant markers and/or targets for therapeutic

interventions.

5. Conclusion

In this study, researchers showed that the changes in gene-expression patterns are associated with specific changes in the aggressive, lymph-node involvement and survival outcomes of breast cancer patients. A significant number of genes were identified as being associated with high histological grade and nodal positivity: 67 genes were common to both progression-related comparisons, and 66 genes had concordant expression changes. Genes that were highly induced, such as AURKA, CCNE1, CHEK1, CDC25A, E2F2, CDK1 and CCNB1, emphasized the role of cell-cycle dysregulation, mitotic activity, and proliferative signaling, indicative of aggressive breast cancer. Conversely, the decreased expression of genes involved in differentiation and regulatory control (GATA3, BCL2, MAPT, IGF1R and PTEN) indicated loss of differentiation and regulation during the disease course. The involvement of several candidate genes with OS further reinforces the possible prognostic importance. More specifically, the link between RHEB expression and higher risk of dying suggests that, in addition to clinical outcomes, progression-associated molecular changes may also be responsible for the clinical outcome's differences. The integrated bioinformatics approach was found to be useful for gaining insight into the molecular mechanisms involved in breast cancer progression and the identification of candidate biomarkers for further investigation. The results should be confirmed in independent patient cohorts and experimental functional investigations before translation to the clinic.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A Cancer J Clinicians*. 2021 May;71(3):209–49. doi:10.3322/caac.21660
2. Arnold M, Morgan E, Rungay H, Mafra A, Singh D, Laversanne M, et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The breast*. 2022;66:15–23.
3. Heer E, Harper A, Escandor N, Sung H, McCormack V, Fidler-Benaoudia MM. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. *The Lancet Global Health*. 2020;8(8):e1027–37.
4. Lei S, Zheng R, Zhang S, Wang S, Chen R, Sun K, et al. Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. *Cancer Commun*. 2021 Nov;41(11):1183–94. doi:10.1002/cac2.12207
5. Bakhtar O. Pathology of Breast Cancer: Subtypes, Receptors, Biologic Markers, and Staging. *The Surgical Clinics of North America*. 2022;103(1):1–15.
6. Roy M, Fowler AM, Ulaner GA, Mahajan A. Molecular classification of breast cancer. *PET clinics*. 2023;18(4):441–58.
7. Szymiczek A, Lone A, Akbari MR. Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review. *Clinical Genetics*. 2021 May;99(5):613–37. doi:10.1111/cge.13900
8. Cobb AN, Czaja R, Jorns J, Cortina CS. Breast Cancer Subtypes: Clinicopathologic Features and Treatment Considerations. *Curr Breast Cancer Rep*. 2024 Jun;16(2):150–60. doi:10.1007/s12609-024-00541-6
9. Lawton TJ. Update on the use of molecular subtyping in breast cancer. *Advances in Anatomic Pathology*. 2023;30(6):368–73.
10. Buyuk B, Jin S, Ye K. Epithelial-to-Mesenchymal Transition Signaling Pathways Responsible for Breast Cancer Metastasis. *Cel Mol Bioeng*. 2022 Feb;15(1):1–13. doi:10.1007/s12195-021-00694-9
11. Terceiro LE, Edechi CA, Ikeogu NM, Nickel BE, Hombach-Klonisch S, Sharif T, et al. The breast tumor microenvironment: A key player in metastatic spread. *Cancers*. 2021;13(19):4798.
12. Khorasani ABS, Hafezi N, Sanaei M, Jafari-Raddani F, Pourbagheri-Sigaroodi A, Bashash D. The PI3K/AKT/mTOR signaling pathway in breast cancer: Review of clinical trials and latest advances. *Cell Biochemistry & Function*. 2024 Apr;42(3):e3998. doi:10.1002/cbf.3998
13. Panda VK, Mishra B, Mahapatra S, Swain B, Malhotra D, Saha S, et al. Molecular insights on signaling cascades in breast cancer: a comprehensive review. *Cancers*. 2025;17(2):234.
14. Jaber Chehayeb R, Hood A, Wang X, Miksad R, Schellhorn Mougalian S, Lustberg MB, et al. Treatment sequencing patterns and associated direct medical costs of metastatic breast cancer care in the United States, 2011 to 2021. *JAMA network open*. 2022;5(11):e2244204.
15. Zhang H, Hussin H, Hoh CC, Cheong SH, Lee WK, Yahaya BH. Big data in breast cancer: Towards precision treatment. *DIGITAL HEALTH*. 2024 Jan;10:20552076241293695. doi:10.1177/20552076241293695
16. Papalexis P, Georgakopoulou V, Drossos P, Thymara E, Nonni A, Lazaris A, et al. Precision medicine in breast cancer (Review). *Mol Clin Oncol*. 2024 Aug 20;21(5):78. doi:10.3892/mco.2024.2776
17. Alharbi R. Breast Cancer Gene Expression Profiles (METABRIC) [Internet]. [cited 2026 Aug 17]. Available from: <https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric>
18. Hao M, Liu W, Ding C, Peng X, Zhang Y, Chen H, et al. Identification of hub genes and small molecule therapeutic drugs related to breast cancer with comprehensive bioinformatics analysis. *PeerJ*. 2020;8:e9946.
19. Zeng X, Shi G, He Q, Zhu P. Screening and predicted value of potential biomarkers for breast cancer using bioinformatics analysis. *Scientific reports*. 2021;11(1):20799.
20. Fang L, Liu Q, Cui H, Zheng Y, Wu C. Bioinformatics analysis highlight differentially expressed CCNB1 and

- PLK1 genes as potential anti-breast cancer drug targets and prognostic markers. *Genes*. 2022;13(4):654.
21. Mehraj U, Sofi S, Alshehri B, Mir MA. Expression pattern and prognostic significance of CDKs in breast cancer: An integrated bioinformatic study. *CBM*. 2022 May 13;34(3):505–19. doi:10.3233/CBM-210186
 22. Mishra D, Mishra A, Nand Rai S, Vamanu E, Singh MP. Demystifying the role of prognostic biomarkers in breast cancer through integrated transcriptome and pathway enrichment analyses. *Diagnostics*. 2023;13(6):1142.
 23. Oumeddour A. Screening of potential hub genes and key pathways associated with breast cancer by bioinformatics tools. *Medicine*. 2023;102(11):e33291.
 24. Singh P, Rath A, Minocha R, Sinha A, Haque MM, Hassan MI, et al. Breast Cancer Prognostic Hub Genes Identified by Integrated Transcriptomic and Weighted Network Analysis: A Road Toward Personalized Medicine. *OMICS: A Journal of Integrative Biology*. 2023 May 1;27(5):227–36. doi:10.1089/omi.2023.0033
 25. Kaur B, Mukhlis Y, Natesh J, Penta D, Meeran SM. Identification of hub genes associated with EMT-induced chemoresistance in breast cancer using integrated bioinformatics analysis. *Gene*. 2022;809:146016.
 26. Sathyamoorthi K, Abishek VP, Venkataramana LY. Enhancing breast cancer survival prognosis through omic and Non-Omic data integration. *Clinical Breast Cancer*. 2025;25(1):27–37.