JOURNAL OF
EMERGING INVESTIGATORS

Article

Integrated expression, mutation, and survival analysis of 17 key genes in breast cancer using TCGA-BRCA data

Sophia Wang[1], Christopher Wang[2], Haiyan Lei[1]
[1] Richard Montgomery County High School, Rockville, Maryland
[2] Montgomery Blair High School, Silver Spring, Maryland

SUMMARY

Breast cancer is the most common type of cancer in women around the world, and some forms of breast cancer have a strong genetic component. We examined 17 clinically significant genes to assess their potential impact on breast cancer. These included oncogenes (such as *ERBB2* and *CCND1*), tumor suppressors (such as *TP53* and *BRCA1/2*), and hormone receptors (such as *ESR1* and *PGR*). We analyzed gene expression, mutation status, and patient survival using data from the TCGA-BRCA cohort. We hypothesized that breast tumors would exhibit transcriptional dysregulation and mutations in key genes such as *BRCA1, BRCA2,* and *TP53*, and that these molecular changes would be associated with differences in patient survival. We conducted differential expression analysis, heatmap clustering, survival modeling, and mutation analysis. Our results demonstrated substantial transcriptional differences between tumor and normal tissues. The upregulation of *ERBB2, ESR1,* and *CCND1* in malignancies exemplified this phenomenon. Notably, *MYC* and *RB1* showed reduced mRNA expression, while tumor suppressors *BRCA2* and *CHEK2* were increased. Survival analysis suggested possible associations between increased expression of *TP53, BRCA1,* and *BRCA2* with patient outcomes; however, these findings were not statistically significant. Mutation analysis identified *TP53* as the most frequently mutated gene, whereas mutations in *BRCA1* and *BRCA2* were infrequent and not associated with significant changes in transcript abundance. Overall, our findings highlight the complexity of breast cancer biology and emphasize the value of integrated genomic analyses for understanding disease mechanisms and therapeutic implications.

INTRODUCTION

Breast cancer is the most common type of cancer in women around the world and remains a major cause of cancer-related deaths, despite advances in screening and treatment (1,2). Breast cancer consists of several molecular subtypes, including luminal, HER2-enriched, and basal-like tumors, which differ in gene expression patterns and clinical behavior (3). This diversity highlights the importance of understanding the genetic factors that influence tumor development and progression (3). Understanding the functions of critical oncogenes, tumor suppressor genes, and hormone receptor genes in disease progression may provide significant insights into prognosis and therapeutic targeting.

This study focused on 17 genes with clearly delineated roles in breast cancer biology (**Table 1**). These included two hormone receptor genes (*ESR1*, *PGR*), the transcription factor gene *GATA3*, eight tumor suppressor genes (*BRCA1*, *BRCA2*, *TP53*, *PTEN*, *RB1*, *CHEK2*, *CDH1*, *MAP3K1*), and six oncogenes (*PIK3CA*, *AKT1*, *ERBB2*, *MYC*, *CCND1*, *FGFR1*). These genes were selected based on their established involvement in cancer development, cell growth, and hormone signaling (3–8). Oncogenes promote cell growth and survival, tumor suppressor genes help maintain and regulate the cell cycle, and hormone receptor genes mediate cellular responses to hormones (3-8).

Cancer risk is often associated with mutated or dysfunctional forms of these genes rather than their normal functions. For example, *BRCA1* and *BRCA2* are involved in DNA repair through homologous recombination. By the time a person turns 80, germline pathogenic mutations in each gene raise the risk of breast cancer by about 70% (9). Some of these genes are also important for treatment decisions. Oncogenic drivers such as *ERBB2* and *PIK3CA* have become important therapeutic targets in breast cancer. Amplification of *ERBB2* can predict response to HER2-targeted therapies such as trastuzumab, while mutations in *PIK3CA* may indicate responsiveness to PI3K inhibitors in hormone receptor–positive, HER2-negative breast cancer (10,11). Similarly, mutations in *TP53*, a gene that plays a key role in DNA damage response and cell cycle regulation, are found in approximately 25–30% of breast cancers and are associated with more aggressive disease (12). Hormone receptor genes *ESR1* and *PGR*, together with the transcription factor gene *GATA3*, define luminal breast cancer subtypes and influence response to endocrine therapy (7,8).

We employed data from The Cancer Genome Atlas breast cancer cohort (TCGA-BRCA) to examine the transcriptional and mutational characteristics of these 17 genes. This cohort contains extensive genomic and clinical information across multiple breast cancer subtypes (13). We employed DESeq2 to normalize and analyze RNA-seq data (14). Given their well-established relevance to breast cancer pathogenesis, *BRCA1*, *BRCA2*, and *TP53* were examined in greater detail. Based on prior knowledge of their roles in tumor suppression and DNA repair, we hypothesized that breast tumors would exhibit transcriptional dysregulation and mutations in key genes such as *BRCA1*, *BRCA2*, and *TP53*, and that these molecular changes would be associated with differences in patient survival. In this study, we identified transcriptional differences between tumor and normal tissues, observed frequent *TP53* mutations, and found no statistically significant associations between gene expression and patient survival.

Gene	Gene Category	Primary Biological Role	Relevance to Breast Cancer	Key Reference(s)
BRCA1	Tumor suppressor	DNA damage repair via homologous recombination; genomic stability	Pathogenic variants increase hereditary breast cancer risk and impair DNA repair	(9)
BRCA2	Tumor suppressor	Homologous recombination–mediated DNA repair	Germline mutations confer high breast cancer risk and promote genomic instability	(9)
TP53	Tumor suppressor	Regulation of cell cycle arrest, apoptosis, and DNA damage response	Frequently mutated in breast cancer; associated with aggressive subtypes and poor prognosis	(10)
PTEN	Tumor suppressor	Negative regulation of PI3K/AKT signaling	Loss of function promotes tumor growth and survival	(8)
RB1	Tumor suppressor	Cell cycle regulation through inhibition of E2F transcription factors	Dysregulation contributes to uncontrolled cell proliferation	(3)
CHEK2	Tumor suppressor	DNA damage checkpoint signaling	Pathogenic variants increase susceptibility to breast cancer	(6)
CDH1	Tumor suppressor	Cell–cell adhesion and maintenance of tissue architecture	Loss of function linked to invasive and lobular breast cancer	(8)
MAP3K1	Tumor suppressor	Regulation of MAPK signaling pathways	Frequently mutated in luminal breast cancer subtypes	(8)
PIK3CA	Oncogene	Activation of PI3K/AKT signaling pathway	Commonly mutated oncogene and therapeutic target in breast cancer	(12)
AKT1	Oncogene	Cell survival and proliferation signaling	Activating mutations promote tumor growth	(8)
ERBB2	Oncogene	Growth factor receptor (HER2) signaling	Amplification defines HER2-positive breast cancer; target of trastuzumab	(5, 11)
MYC	Oncogene	Transcriptional regulation of cell growth and metabolism	Frequently amplified; oncogenic effects often occur at DNA or protein level	(8)
CCND1	Oncogene	Cell cycle progression (G1/S transition)	Overexpression promotes uncontrolled proliferation	(4)
FGFR1	Oncogene	Growth factor receptor signaling	Amplification associated with poor prognosis and therapy resistance	(8)
ESR1	Hormone receptor	Estrogen signaling and transcriptional regulation	Defines hormone receptor–positive breast cancer; endocrine therapy target	(7)
PGR	Hormone receptor	Progesterone signaling	Marker of functional estrogen receptor signaling and luminal subtype	(3)
GATA3	Transcription factor	Regulation of luminal cell differentiation	Frequently mutated in luminal breast cancer; subtype marker	(3)

Table 1. Summary of the 17 genes analyzed in this study

https://doi.org/10.59720/25-295

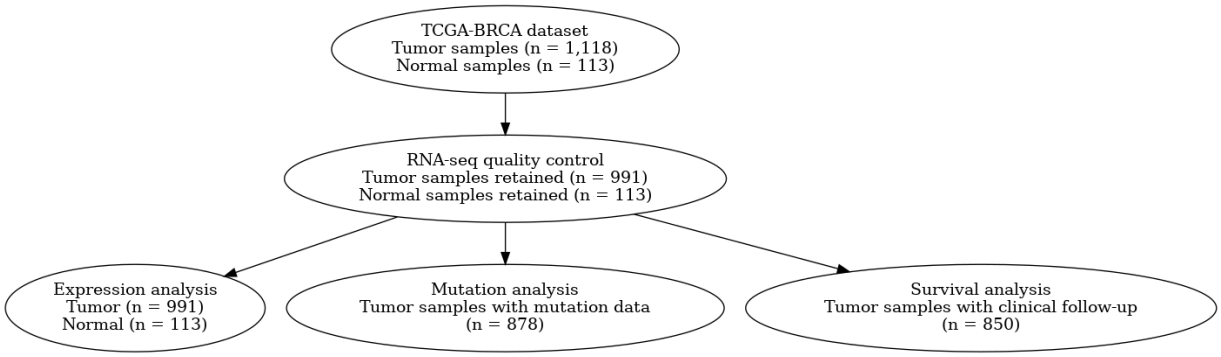


Figure 1. TCGA-BRCA data selection and analysis workflow. RNA-seq gene expression, mutation, and clinical data were obtained from the TCGA-BRCA cohort. Of the 1,118 tumor samples and 113 normal samples initially available, 991 of the tumor samples and all 113 of the normal samples passed RNA-seq quality control and were used for expression analysis. Of the 991 tumor samples retained after quality control, 878 had available somatic mutation data, and 850 had clinical follow-up information for survival analysis.

These findings highlight the complexity of breast cancer biology.

RESULTS

Overview of TCGA-BRCA Sample Selection and Analysis Workflow

We analyzed 1,118 breast tumor samples and 113 normal tissue samples sourced from the TCGA-BRCA dataset (13). After preprocessing and quality control using DESeq2, 991 tumor samples and all 113 normal samples were retained for expression analysis (14). There were 850 tumor samples with available clinical survival follow-up data and 878 tumor samples with somatic mutation data in Mutation Annotation Format (MAF), a standardized file format used to store and summarize somatic mutation data. These sample sizes reflect the practical limitations of working with large public datasets, as not all samples contain complete clinical or genomic information. The analytical workflow included gene expression analysis, survival analysis, and mutation profiling (**Figure 1**).

Differential Expression Analysis Identifies Transcriptional Differences between Tumor and Normal Tissues

To identify transcriptional differences between tumor and normal tissues, we performed differential expression analysis

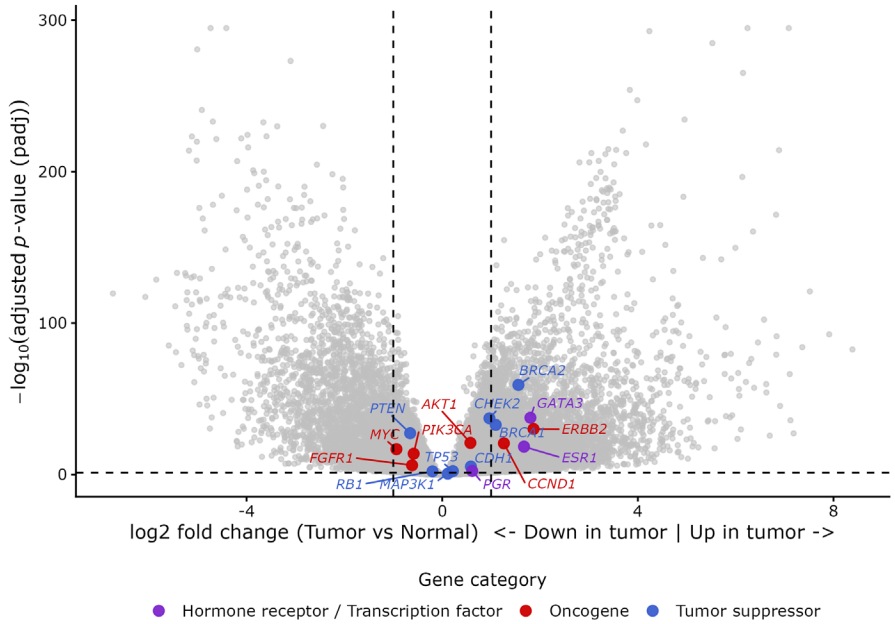


Figure 2. Differential expression analysis of breast tumors versus normal tissue in TCGA-BRCA. A volcano plot showing $\log_2$ fold change in gene expression (tumor vs. normal) plotted against $-\log_{10}$(adjusted p-value). Each point represents a gene analyzed using DESeq2. Dashed vertical lines indicate $\log_2$ fold change thresholds of ±1, and the dashed horizontal line indicates an adjusted *p*-value threshold of 0.05. The 17-breast cancer–related genes analyzed in this study are highlighted and color-coded by functional category: oncogenes (red), tumor suppressor genes (blue), and hormone receptor–related genes (purple). Genes with positive $\log_2$ fold change values indicate higher expression in tumor samples relative to normal tissue, whereas negative values indicate lower expression in tumor samples.

https://doi.org/10.59720/25-295

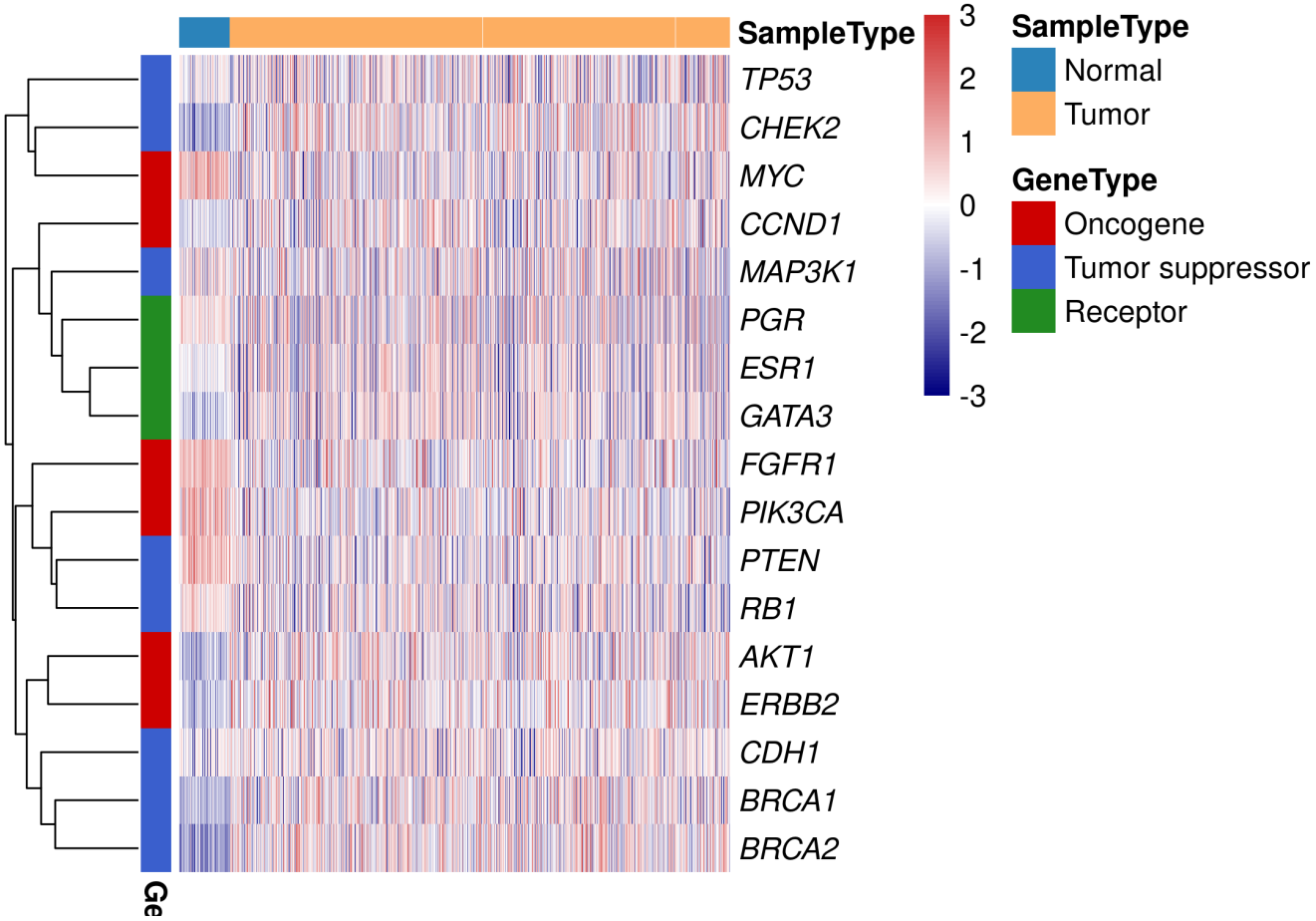


Figure 3. Heatmap of expression patterns for 17 selected breast cancer genes. The heatmap displays row-wise Z-score–standardized gene expression values across tumor and normal samples. Z-scores were clipped to the range −3 to +3 to improve visual discrimination of expression differences while preserving standardized scaling across genes (blue = lower relative expression, white = near the gene mean, red = higher relative expression). Columns are grouped by sample type (Normal and Tumor), and row annotations indicate gene functional category (Receptor, Oncogene, Tumor suppressor). Hierarchical clustering was applied to genes only.

using the DESeq2 Wald test. The global volcano plot (**Figure 2**) shows widespread transcriptional changes, indicating extensive transcriptomic remodeling in breast tumors.

We selected 17 breast cancer–related genes for further analysis based on their established roles in tumor development and progression. Several oncogenes, including *ERBB2*, *CCND1*, and *AKT1*, exhibited higher expression in tumor samples compared to normal tissue, consistent with their previously reported roles in breast cancer biology (4,5). The hormone receptor genes *ESR1* and *PGR* also showed increased expression in tumor samples.

BRCA2 ($\log_2$FC = 1.56, adjusted $p < 0.001$) and *CHEK2* ($\log_2$FC = 0.97, adjusted $p < 0.001$) exhibited higher transcript abundance in tumors, whereas *RB1* ($\log_2$FC = –0.20, adjusted $p < 0.01$) showed reduced expression. *MYC* was downregulated at the mRNA level ($\log_2$FC = –0.94, adjusted $p < 0.001$), suggesting that its oncogenic activity may occur through mechanisms other than transcript overexpression, such as copy number amplification or protein-level regulation. In contrast, *MAP3K1* showed no significant difference in expression between tumor and normal samples ($\log_2$FC ≈ 0.11, adjusted $p = 0.27$).

Heatmap Clustering Distinguishes Tumor and Normal Samples Based on Gene Expression Patterns

A clustered heatmap was generated to visualize expression patterns of the 17 selected genes across tumor and normal samples (**Figure 3**). Tumor and normal samples formed distinct clusters based on gene expression profiles. Genes identified as upregulated in tumors by differential expression analysis (e.g., *ERBB2*, *CCND1*, *ESR1*, *PGR*, *BRCA2*, and *CHEK2*) also exhibited higher relative expression in tumor clusters. Unsurprisingly, genes with lower expression in tumors (e.g., *RB1* and *MYC*) showed lower expression in tumor clusters. *MAP3K1* displayed minimal variation between groups, consistent with the differential expression results. The heatmap also provided a visualization of coordinated expression patterns and sample-level clustering, which are not apparent in the volcano plot. Together, these results demonstrate concordance between independent analytical

https://doi.org/10.59720/25-295

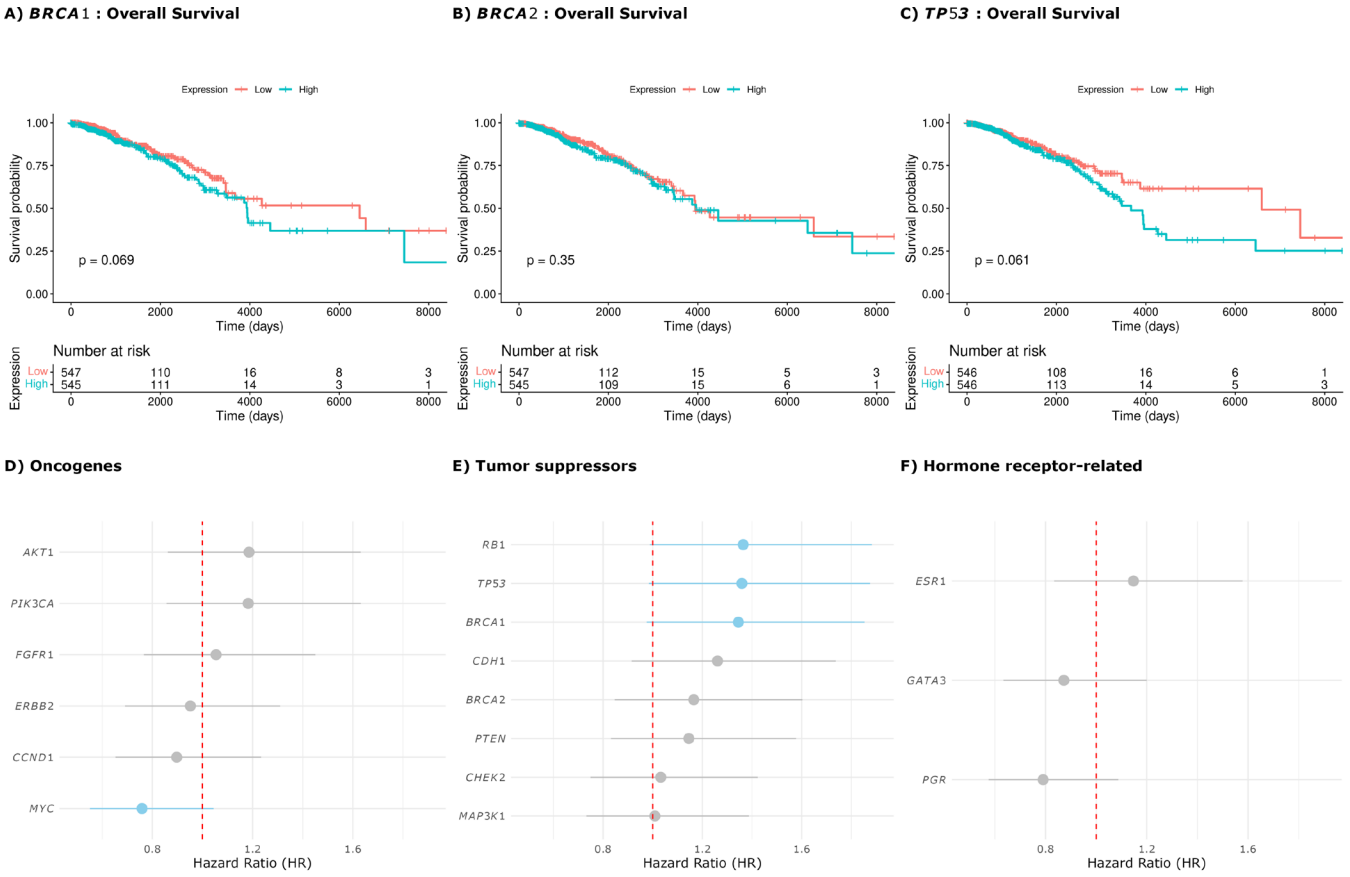


Figure 4: Survival analysis of selected breast cancer genes in TCGA-BRCA. (A–C) Kaplan–Meier survival curves comparing overall survival between patients with high and low expression of *BRCA1, BRCA2,* and *TP53*, stratified by median gene expression. The x-axis represents time in days, and the y-axis represents overall survival probability. Differences between groups were evaluated using the log-rank test. (D-F) Forest plot summarizing Cox proportional hazards regression results for oncogenes (D), tumor suppressor genes (E), and hormone receptor–related genes (F). Points represent hazard ratios (HRs), and horizontal lines indicate 95% confidence intervals. Blue points indicate borderline associations (p < 0.10), whereas gray points indicate non-significant associations. The dashed vertical line indicates HR = 1.

approaches and reinforce the robustness of the observed transcriptional differences between tumor and normal tissues.

Survival Analysis Reveals Non-Significant Associations between Gene Expression and Overall Survival

Survival analyses were conducted to evaluate associations between gene expression levels and overall survival. Patients were stratified into high- and low-expression groups based on median expression values. Because these analyses were based solely on transcript abundance, Kaplan–Meier survival curves were generated for *BRCA1, BRCA2*, and *TP53*, as these genes had sufficient sample sizes for reliable comparison (**Figure 4A–C**).

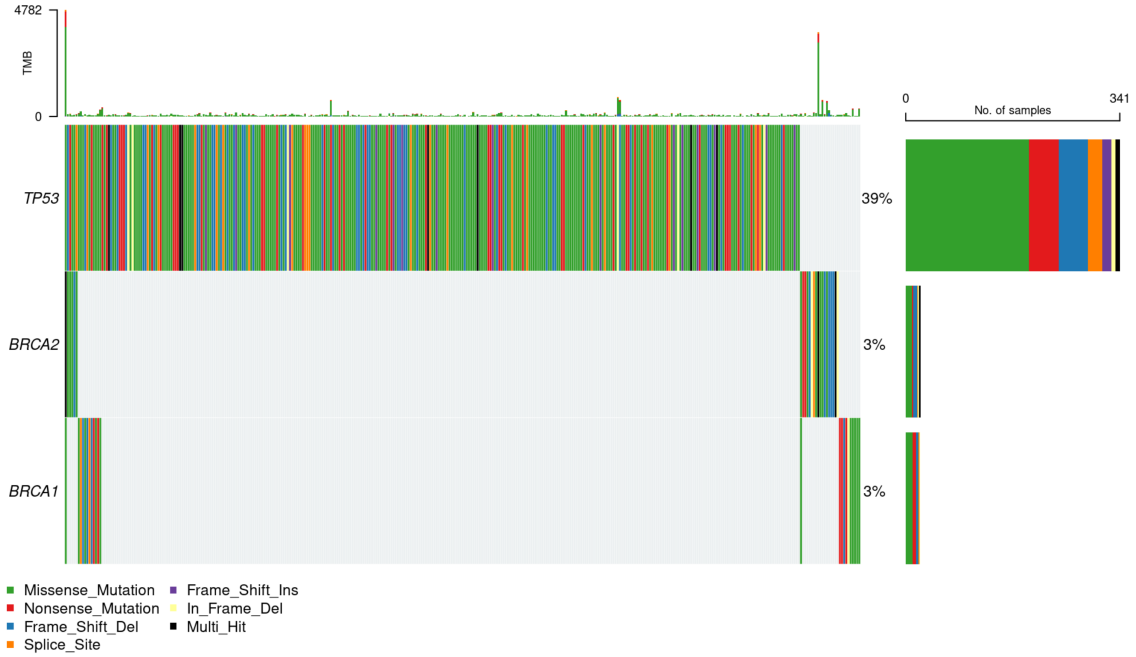
Log-rank tests and Cox proportional hazards regression models were used to assess survival differences. Higher expression levels of *BRCA1*, *BRCA2*, and *TP53* were associated with trends toward poorer survival outcomes; however, none of these associations reached statistical significance ($p > 0.05$) (**Figure 4D-F**). *BRCA1* showed a modest trend (HR = 1.35, 95% CI = 0.98–1.85, $p = 0.07$), and *TP53* showed a similar pattern (HR = 1.36, 95% CI = 0.99–1.88, $p = 0.06$). A forest plot summarizing Cox regression results for all 17 genes is shown (**Figure 4D-F**). Hazard ratios varied across genes, but none reached statistical significance.

Mutation Analysis Shows Frequent *TP53* Alterations without Major Effects on Transcript Abundance

Somatic mutation analysis focused on *TP53*, *BRCA1*, and *BRCA2*. *TP53* mutations were identified in 321 of 878 tumor samples (36.6%), whereas mutations in *BRCA1* and

https://doi.org/10.59720/25-295

A) Mutation landscape of *TP53*, *BRCA1* and *BRCA2*



B) Expression: Mutated vs WT (Wilcoxon test)

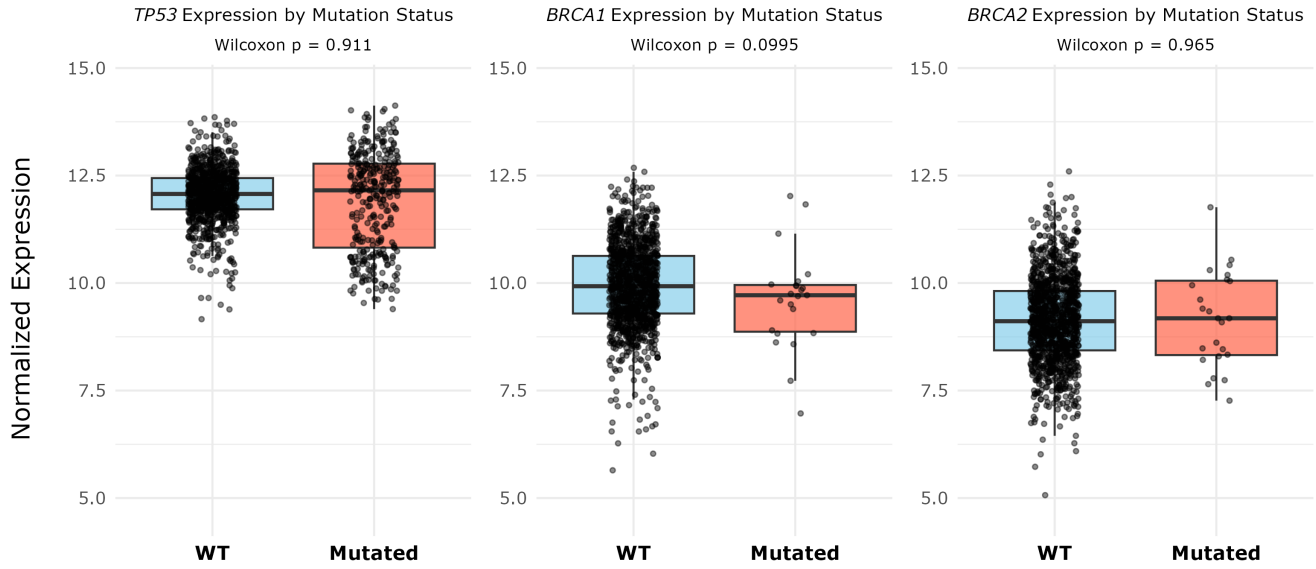


Figure 5: Mutation landscape and expression analysis of TP53, BRCA1, and BRCA2 in TCGA-BRCA. (A) Oncoplot showing the frequency and distribution of somatic mutations in TP53, BRCA1, and BRCA2 across 878 tumor samples with available mutation data from the TCGA-BRCA cohort. A total of 369 tumors (42.0%) harbored at least one mutation among these three genes. (B) Boxplots comparing RNA-seq–derived transcript abundance between wild-type and mutant tumor samples for each gene using the Wilcoxon rank-sum test. Wild-type (WT) samples are defined as tumors without detected somatic mutations in the respective gene, whereas mutant samples harbor at least one somatic mutation in that gene.

https://doi.org/10.59720/25-295

BRCA2 were less frequent, occurring in 12 (1.4%) and 36 (4.1%) tumors, respectively (**Figure 5A**). Overall, 369 tumors (42.0%) harbored at least one mutation in these three genes, with some tumors containing mutations in more than one gene.

Expression comparisons between wild-type and mutant tumor samples were conducted using RNA-seq–derived mRNA transcript abundance values. Wilcoxon rank-sum tests showed no significant differences in expression between wild-type and mutant tumors for *TP53* (p = 0.911), *BRCA1* (p = 0.100), or *BRCA2* (p = 0.965) (**Figure 5B**). These results indicate that mutation status was not associated with substantial changes in mRNA expression levels for these genes within this cohort.

DISCUSSION

Our methodology demonstrates the use of a stepwise analytical framework to examine large-scale cancer genomics datasets. After quality control and filtering, analyses were conducted on subsets of samples with available expression (n = 991 tumors, 113 normal), survival (n = 850 tumors), and mutation data (n = 878 tumors). Variation in sample availability reflects a common limitation of large public datasets and highlights the importance of clearly reporting sample sizes.

The differential expression analysis revealed widespread transcriptional differences between normal and tumor samples, consistent with extensive transcriptomic remodeling in breast cancer. Several oncogenes, including *ERBB2*, *CCND1*, and *AKT1,* exhibited increased expression in tumors, consistent with their roles in promoting tumor growth and proliferation (4,5). *ESR1* and *PGR* were also upregulated, aligning with their association with hormone receptor-positive breast cancer subtypes (7,8).

In contrast, some tumor suppressor genes, including *BRCA2* and *CHEK2*, exhibited increased mRNA expression in tumor samples compared to normal tissue. This finding suggests that elevated transcript levels do not necessarily reflect functional tumor suppression, as gene function may be compromised by mutations or post-transcriptional regulation. In contrast, *MYC* showed reduced mRNA expression in tumors, despite its well-established oncogenic role. This may reflect alternative mechanisms of activation, such as copy number amplification or post-transcriptional regulation (8). Copy-number variation, particularly gene amplification, is a well-established mechanism of *MYC* activation in breast cancer and may result in increased oncogenic signaling even when steady-state mRNA levels do not appear markedly elevated (8).

Heatmap analysis further illustrated distinct clustering between normal and tumor samples, supporting the differential expression results. Tumor samples displayed higher expression of oncogenes such as *ERBB2* and *CCND1*, while tumor suppressor genes showed heterogeneous expression patterns. These findings reinforce the complexity of transcriptional regulation in breast cancer rather than indicating uniform activation or suppression of specific gene categories.

Survival analysis showed no statistically significant associations between gene expression and overall survival. Higher expression of *BRCA1, BRCA2,* and *TP53* showed weak associations with poorer survival outcomes, which may reflect compensatory transcriptional responses in tumors with underlying genomic instability rather than direct effects on patient prognosis. Gene expression measured by bulk RNA-seq does not necessarily reflect protein function or pathway activity. This discrepancy may arise because mRNA abundance may not correspond to functional protein levels. Protein activity may be altered by post-transcriptional regulation, translational efficiency, post-translational modifications, protein degradation, or loss-of-function mutations that impair protein structure without substantially affecting transcript abundance. This is especially relevant for tumor suppressor genes, where inactivating mutations may disrupt protein function even when mRNA expression remains elevated.

Mutation analysis demonstrated that *TP53* was the most frequently mutated gene in this cohort, occurring in 36.6% of tumor samples (**Figure 5A**), consistent with previous reports (10). In contrast, *BRCA1* and *BRCA2* mutations were less frequent, consistent with their established roles in hereditary breast cancer (9,11). Importantly, mutation status was not associated with significant differences in mRNA expression levels, suggesting that the functional consequences of these mutations likely occur at the protein or pathway level rather than through altered transcript abundance, as described in prior studies (10).

This study has several limitations. First, the incomplete availability of mutation and survival data reduced the effective sample size for some analyses. In addition, although the TCGA-BRCA cohort contains detailed clinical metadata (e.g., age, tumor stage, and molecular subtype), these patient-level variables were not incorporated into the present analysis, which may limit the clinical interpretation of these findings. Because breast cancer is a heterogeneous disease with distinct subtypes, analyses across the combined cohort may obscure subtype-specific differences in gene expression, mutation frequency, and survival associations. Second, RNA-seq–based expression measurements do not capture protein abundance or functional activity. As a result, gene expression levels may not fully reflect the biological effects of these genes or their roles in disease progression. Third, the analysis focused on a predefined set of 17 genes and did not capture the full genomic complexity of breast cancer. Although additional genes showed differential expression, they were not further investigated in the present study to maintain focus on clinically established genes. Future studies could incorporate systematic analysis of the most strongly differentially expressed genes to identify additional candidate biomarkers or therapeutic targets. Future studies integrating subtype-specific analyses, clinical characteristics, and additional data types such as proteomics or copy number variation to provide deeper insights into gene function and clinical outcomes.

In summary, our study used an integrated bioinformatics approach to analyze gene expression, mutation, and survival patterns of 17 clinically relevant genes in the TCGA-BRCA dataset. Our findings highlight the complexity of breast cancer biology and demonstrate that gene expression changes, mutation status, and clinical outcomes are not always directly aligned. Overall, our findings partially support our original

https://doi.org/10.59720/25-295

hypothesis. Breast tumors displayed clear transcriptional dysregulation in several clinically relevant genes, and *TP53* mutations were frequent in the TCGA-BRCA cohort. However, although differences in gene expression were observed between tumor and normal tissues, the expected associations between gene expression levels and patient survival were not statistically significant. These findings underscore the value of multidimensional analyses while emphasizing the need for cautious interpretation of transcript-level data in the context of tumor biology, as transcript abundance alone may not fully capture the functional impact of these genes in breast cancer progression.

MATERIALS AND METHODS

Data Source

Data were obtained from the publicly available Cancer Genome Atlas (TCGA) breast cancer cohort (TCGA-BRCA) (13). This dataset includes RNA-seq gene expression data, somatic mutation annotation files, clinical follow-up information, and matched normal tissue samples. Initially, 1,118 tumor samples and 113 normal samples were available for analysis. The TCGA-BRCA cohort includes tumors from multiple molecular subtypes of breast cancer (e.g., luminal, HER2-enriched, and basal-like), and subtype-specific analyses were not performed in this study.

RNA-seq Preprocessing and Differential Expression Analysis

Raw RNA-seq count data were retrieved using the TCGAbiolinks R package (13). Gene-level raw counts were normalized and analyzed using DESeq2, which accounts for library size and dispersion estimation (14). Genes with low read counts across samples were filtered prior to differential expression analysis to reduce noise and improve statistical robustness, as recommended in standard RNA-seq workflows (14). After filtering, 991 tumor samples and all 113 normal samples were retained for expression analysis.

Differential expression analysis between normal and tumor samples was performed using the DESeq2 Wald test. Genes with an absolute $\log_2$ fold change > 1 and an adjusted p-value < 0.05 were considered statistically significant.

Volcano plots were generated using the EnhancedVolcano and ggplot2 packages, and heatmaps were generated using the pheatmap package. For visualization, gene expression values were standardized using row-wise Z-score normalization across samples. To enhance visual contrast while maintaining standardized scaling, Z-scores were clipped to the range −3 to +3 for heatmap display.

Gene Selection

Seventeen genes previously implicated in breast cancer were selected for analysis based on published literature (3-8). These included oncogenes (*PIK3CA*, *AKT1*, *ERBB2*, *MYC*, *CCND1*, *FGFR1*), tumor suppressor genes (*BRCA1*, *BRCA2*, *TP53*, *PTEN*, *RB1*, *CHEK2*, *CDH1*, *MAP3K1*), and hormone receptor–related genes (*ESR1*, *PGR*), and the transcription factor gene *GATA3*.

Survival Analysis

Clinical data, including vital status, days to death, and days to last follow-up, were obtained from TCGA and used to calculate overall survival. Patients were classified into high- and low-expression groups based on the median expression value of each gene, such that individuals with expression values above the median were assigned to the high-expression group and those below the median were assigned to the low-expression group. Kaplan–Meier survival curves were generated, and differences between groups were evaluated using the log-rank test. Cox proportional hazards regression models were used to estimate hazard ratios and 95% confidence intervals. Survival analyses were conducted using the survival and survminer R packages. Only genes with sufficient sample sizes in both expression groups were included in Kaplan–Meier analyses, while Cox regression was performed for all 17 genes.

Mutation Analysis

Somatic mutation data in Mutation Annotation Format (MAF) were obtained from TCGA and analyzed using the maftools package. Mutation frequency and distribution were assessed for *TP53*, *BRCA1*, and *BRCA2*. Gene expression levels were compared between wild-type and mutant tumors using RNA-seq–derived transcript abundance values. Differences were evaluated using the Wilcoxon rank-sum test.

Software

All analyses were performed in R (version 4.5.1) using Bioconductor and CRAN packages. TCGAbiolinks was used to retrieve and preprocess TCGA genomic and clinical data (13). DESeq2 was applied for normalization and differential gene expression analysis (14). The maftools package was used to analyze and visualize somatic mutation data in Mutation Annotation Format (MAF) (15). Survival and survminer packages were used to generate Kaplan–Meier curves and perform Cox proportional hazards regression analyses (16, 17). Visualization packages included ggplot2, EnhancedVolcano, and pheatmap (18-20).

Data Availability

All data analyzed in this study are publicly available through The Cancer Genome Atlas (TCGA) via the Genomic Data Commons (GDC) portal (13). Processed datasets generated during this study are available from the corresponding author upon reasonable request.

Received: Septemeber 15, 2025
Accepted: April 9, 2026
Published: August 6, 2026

REFERENCES

1. Sung, Hyuna, et al. “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.” *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, 2021, pp. 209–49. https://doi.org/10.3322/caac.21660.
2. Siegel, Rebecca L., et al. “Cancer Statistics, 2022.” *CA: A Cancer Journal for Clinicians*, vol. 72, no. 1, 2022, pp. 7–33. https://doi.org/10.3322/caac.21708.
3. Perou, Charles M., et al. “Molecular Portraits of Human

https://doi.org/10.59720/25-295

Breast Tumours." *Nature*, vol. 406, no. 6797, Aug. 2000, pp. 747–52. https://doi.org/10.1038/35021093.
4. Sorlie, Therese, et al. "Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications." *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, Sept. 2001, pp. 10869–74. https://doi.org/10.1073/pnas.191367098.
5. Slamon, Dennis J., et al. "Human Breast Cancer: Correlation of Relapse and Survival with Amplification of the HER-2/*neu* Oncogene." *Science*, vol. 235, no. 4785, Jan. 1987, pp. 177–82. https://doi.org/10.1126/science.3798106.
6. Nik-Zainal, Serena, et al. "Landscape of Somatic Mutations in 560 Breast Cancer Whole-Genome Sequences." *Nature*, vol. 534, no. 7605, June 2016, pp. 47–54. https://doi.org/10.1038/nature17676.
7. Jeselsohn, Rinath, et al. "*ESR1* Mutations—a Mechanism for Acquired Endocrine Resistance in Breast Cancer." *Nature Reviews Clinical Oncology*, vol. 12, no. 10, Oct. 2015, pp. 573–83. https://doi.org/10.1038/nrclinonc.2015.117.
8. Curtis, Christina, et al. "The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups." *Nature*, vol. 486, no. 7403, June 2012, pp. 346–52. https://doi.org/10.1038/nature10983.
9. Kuchenbaecker, Karoline B., et al. "Risks of Breast, Ovarian, and Contralateral Breast Cancer for *BRCA1* and *BRCA2* Mutation Carriers." *JAMA*, vol. 317, no. 23, June 2017, pp. 2402–16. https://doi.org/10.1001/jama.2017.7112.
10. Turner, Nicholas, et al. "Hallmarks of 'BRCAness' in Sporadic Cancers." *Nature Reviews Cancer*, vol. 4, no. 10, Oct. 2004, pp. 814–19. https://doi.org/10.1038/nrc1457.
11. André, Fabrice, et al. "Alpelisib for *PIK3CA*-Mutated, Hormone Receptor–Positive Advanced Breast Cancer." *New England Journal of Medicine*, vol. 380, no. 20, May 2019, pp. 1929–40. https://doi.org/10.1056/NEJMoa1813904.
12. Silwal-Pandit, Laxmi, et al. "*TP53* Mutation Spectrum in Breast Cancer Is Subtype Specific and Has Distinct Prognostic Relevance." *Clinical Cancer Research*, vol. 20, no. 13, June 2014, pp. 3569–80. https://doi.org/10.1158/1078-0432.CCR-13-2943.
13. Koboldt, Daniel C., et al. "Comprehensive Molecular Portraits of Human Breast Tumours." *Nature*, vol. 490, no. 7418, Oct. 2012, pp. 61–70. https://doi.org/10.1038/nature11412.
14. Love, Michael I., et al. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology*, vol. 15, no. 12, Dec. 2014, p. 550. https://doi.org/10.1186/s13059-014-0550-8.
15. Mayakonda, Anand, et al. "Maftools: Efficient and Comprehensive Analysis of Somatic Variants in Cancer." *Genome Research*, vol. 28, no. 11, 2018, pp. 1747–56. https://doi.org/10.1101/gr.239244.118.
16. Therneau, Terry M. "A Package for Survival Analysis in R." *R Package Version 3.5-7*, 2023.
17. Kassambara, Alboukadel, and Marcin Kosinski. "Survminer: Drawing Survival Curves Using 'ggplot2'." *R Package Version 0.4.9*, 2021.
18. Wickham, Hadley. *ggplot2: Elegant Graphics for Data Analysis.* Springer, 2016. https://doi.org/10.1007/978-3-319-24277-4.
19. Blighe, Kevin, et al. "EnhancedVolcano: Publication-Ready Volcano Plots with Enhanced Colouring and Labeling." *R Package Version 1.20.0*, 2024.
20. Kolde, Raivo. "Pheatmap: Pretty Heatmaps." *R Package Version 1.0.12*, 2019.