Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrative Transcriptomic Analysis Identifies a Novel 22-Gene Signature Driving Breast Cancer Progression via the Proteasome-Chemokine Axis

Md. Shakawat Hossain

shakawathossain96907@gmail.com

Shahjalal University of Science and Technology

Research Article

Keywords: Breast Cancer, Prognostic Signature, LASSO Cox Regression, Proteasome-Chemokine Axis, Precision Oncology, Tumor Immune Microenvironment

Posted Date: February 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8921240/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Breast cancer heterogeneity often limits traditional clinicopathological predictions. Using transcriptomic data from the SCAN-B cohort (N = 1,185), we developed a 22-gene prognostic signature via LASSO Cox regression. This signature successfully stratified patients into distinct risk groups, with the high-risk group showing significantly decreased overall survival ($p < 0.0001$). Multivariate analysis confirmed the risk score as a robust independent predictor (HR = 3.44, 95% CI: 2.10–5.65, $p < 0.001$), surpassing traditional markers like age and ER/PR status. Mechanistically, the signature defines a "Proliferation-Immune Axis": risk correlates with proteasome hyperactivity and proliferation (*TK1, MARCO*), while survival advantage links to chemokine-driven immune recruitment (*CXCL13, TCRVB*). The model demonstrated high accuracy in predicting early recurrence (AUC = 0.847). When integrated into a calibrated nomogram, this signature provides a precise tool for capturing the proteasome-chemokine interplay, enhancing individualized risk assessment and precision oncology decisions in breast cancer.

Introduction

Recent data indicate that breast cancer in women now represents approximately one in every four cases among females and accounts for 11–12% of all new cancer diagnoses globally, overtaking lung cancer as the most frequently diagnosed disease worldwide[1–4].It continues to be the leading cause of cancer-related mortality among women globally, accounting for around 685,000 deaths each year, despite notable improvements in treatment methods[5,6]. Significant variation at the transcriptome, proteomic, and genomic levels characterizes the illness. Even among individuals with comparable diagnoses, clinical outcomes are unexpectedly diverse, even though certain molecular subtypes (such as HER2 + and Triple-Negative) have guided systemic therapy[7,8].

The TNM staging system (Tumor-Node-Metastasis) is a major component of current therapy planning. However, relying solely on TNM categorization often leads to prognostic inaccuracies because these conventional morphological metrics fail to capture the tumor's fundamental molecular complexities. Patients with equivalent stages may have significantly varied survival trajectories, which might lead to overtreatment of indolent illness or undertreatment of aggressive malignancies. This discrepancy poses a serious therapeutic problem. Therefore, to improve risk assessment and enable precision oncology, combining molecular, immunological, and genomic data with traditional staging is crucial[9–11].

Recent developments in high-throughput RNA-seq have revolutionized the search for biomarkers and made it possible to decipher intricate oncogenic signaling networks[12]. Strong genomic signals, including immunological profiles, lncRNAs, and DNA methylation, have demonstrated the capacity to independently predict survival and direct treatment choices[13–15].Nevertheless, there is still little use of these signals in therapeutic settings. There is a gap between statistical discovery and trustworthy bedside tools, since many suggested gene expression profiles are criticized for being overfit, methodologically flawed, or missing thorough validation[16–19].

In order to overcome these obstacles, this study uses the large Sweden Cancerome Analysis Network–Breast (SCAN-B) cohort and a strict bioinformatics framework to create a strong survival-associated gene signature. We went beyond basic data mining to pinpoint a particular "Proliferation-Immune Axis" causing tumor growth by fusing differential expression analysis with genome-wide survival screening. Additionally, we built a repeatable prognostic model using LASSO Cox regression and included it into a clinical nomogram. We provide a fresh scientific justification for risk classification in breast cancer by critically analyzing the biological interaction between chemokine-mediated immune surveillance and proteasome-driven proliferation.

Results

Baseline Clinicopathological and Molecular Characteristics

To ensure accurate prognostic modeling, transcriptome profiles underwent stringent quality control. PCA and expression analysis confirmed that molecular subtypes were stratified to eliminate batch effects while preserving biological variance **(Supplementary Figure S1A, H)**. The validated cohort (N = 1,185, Table 1) primarily consisted of node-negative (63.8%) and Grade 2 (50.2%) tumors, with a mean age of 62.67 ± 13.13 years. Luminal A (52.3%) and Luminal B (23%) were the predominant subtypes, and 83.7% of tumors were PR-positive, while 90% were ER-positive. Although Luminal A was the most common subtype, strong Ki67 expression was observed in 58.3% of the cohort, indicating significant proliferative heterogeneity.

Table 1
Baseline clinicopathological characteristics and molecular stratification of the study cohort.

Variable	Level	Count	Percent
Sample		1185	
Age (years)			**62.67 ± 13.13**
Tumor Size (mm)			**19.95 ± 12.59**
Tumor Grade	G1	188	15.9%
	G2	595	50.2%
	G3	402	33.9%
Lymph Node Status	Node (Negative)	756	63.8%
	Node (Positive)	429	36.2%
ER Status	Negative	119	10%
	Positive	1066	90%
PR Status	Negative	193	16.3%
	Positive	992	83.7%
HER2 Status	Negative	1049	88.5%
	Positive	136	11.5%
Ki67 Status	High	691	58.3%
	Low	494	41.7%
PAM50 subtype	Basal-like	109	9.2%
	HER2-enriched	93	7.8%
	Luminal A	620	52.3%
	Luminal B	272	23%
	Normal-like	91	7.7%

Abbreviations: SD, Standard Deviation; ER, Estrogen Receptor; PR, Progesterone Receptor; HER2, Human Epidermal Growth Factor Receptor 2

Screening and Prioritization of Survival-Associated Differentially Expressed Genes

We conducted a genome-wide univariate Cox proportional hazards regression over the whole transcriptome to methodically find transcripts linked to patient outcomes. A group of genes that were

substantially linked with overall survival was identified by this screening (Fig. 2A, Table 2). The lncRNA *AX746755* (HR = 0.58) and *GREB1* were shown to be substantial protective variables, whereas *MRPL47* (HR = 3.09, $p < 0.001$) and *NDUFB5* (HR = 3.68, $p < 0.001$) emerged as powerful risk factors.

Table 2

Top candidates identified by univariate Cox regression analysis are associated with overall survival—key prognostic candidates identified by the intersection of univariate Cox regression and differential expression analysis. "Risk Type" denotes whether the gene expression correlates with increased mortality (Risk) or better survival (Protective).

Gene Symbol	Hazard Ratio (HR)	95% CI	P-value	Prognostic Type
AX746755	0.58	0.48–0.70	2.23×10^{-8}	Protective
GREB1	0.71	0.63–0.80	2.90×10^{-8}	Protective
MRPL47	3.09	2.04–4.68	9.54×10^{-8}	Risk
IL8	1.29	1.17–1.43	8.23×10^{-7}	Risk
HSPA6	1.53	1.29–1.82	9.92×10^{-7}	Risk
NDUFB5	3.68	2.18–6.23	1.16×10^{-6}	Risk
RRAGD	1.81	1.42–2.31	1.76×10^{-6}	Risk
CD22	0.52	0.40–0.68	2.31×10^{-6}	Protective
ZNF516	0.58	0.46–0.72	2.75×10^{-6}	Protective
FAIM3	0.59	0.47–0.73	3.56×10^{-6}	Protective

Note: CI = Confidence Interval. HR > 1 denotes increased risk of mortality. Full list available in **Table S1(Provided in the separate file: Supplementary_Data.xlsx)**. Survival-relevant DEGs full list available in **Table S2(Provided in the separate file: Supplementary_Data.xlsx).**

[INSERT FIGURE 2 **HERE]**

We combined prognostic information with differential expression profiles obtained from tumor vs normal tissue comparisons to find survival-relevant differentially expressed genes (SR-DEGs) and rank physiologically functioning indicators **(**Fig. 2B**)**. To exclude non-functional passenger variants, we intersected the survival-associated genes with substantial DEGs (|logFC| > 1, adj. $p < 0.05$) **(**Fig. 2D, **Supplementary Table S2)**. Strong stratification power was demonstrated by the ensuing Survival-Relevant DEGs (SR-DEGs) signature. Different expression patterns correlated with clinical risk were identified by hierarchical clustering of the top 20 candidates **(**Fig. 2C**)**. For example, the protective gene *ABAT* was significantly downregulated in high-risk samples, whereas the cell-cycle regulator *UBE2C* was continuously increased.

Construction of the Prognostic Gene Signature

We used the Least Absolute Shrinkage and Selection Operator (LASSO) Cox regression approach to address multicollinearity among the candidate genes and reduce the risk of overfitting. We determined the ideal penalty parameter (λ_{min} = 0.00941) that reduced the partial likelihood deviance using 10-fold cross-validation **(**Fig. 3A**).**

[INSERT FIGURE 3 **HERE]**

The candidate pool was effectively reduced in dimensionality to a stable signature with **22 genes (**Fig. 3B, Table 3**)**. Based on the linear combination of expression levels weighted by their individual LASSO coefficients, we computed a risk score for every patient (full calculation details provided in **Supplementary Table S3**):

Table 3
Key genes constituting the LASSO prognostic signature and their biological relevance.

Gene Symbol	LASSO Coefficient	Prognostic Type	Biological Interpretation
TK1	+ 0.070	Risk	Marker of high proliferation/DNA repair.
MARCO	+ 0.058	Risk	Macrophage receptor (M2-like); immunosuppressive.
CCL8	+ 0.019	Risk	Chemokine involved in metastasis/EMT.
AB306139	-0.060	Protective	LncRNA with potential regulatory function.
TP63	-0.053	Protective	Tumor suppressor, maintains epithelial integrity.
SLC7A2	-0.045	Protective	Amino acid transporter (Arginine).
CXCL13	-0.039	Protective	B-cell attractant / TLS formation.
ELOVL2	-0.030	Protective	Fatty acid elongation/metabolism.

Note: Coefficients were derived from the LASSO Cox regression model using the λmin criterion. Positive coefficients indicate genes associated with high risk (poor survival), while negative coefficients indicate protective genes—abbreviations: EMT, Epithelial-Mesenchymal Transition; TLS, Tertiary Lymphoid Structure; LncRNA, Long non-coding RNA. The complete list of 22 genes is provided in **Table S3**.

Risk score = $\sum_{i=1}^{n} (\mathrm{Expression}(i) \times \mathrm{Coefficient}(i))$

Together with the metastasis-associated chemokine *CCL8* ($\mathrm{coefficient}$ = 0.019), the model revealed *TK1* ($\mathrm{coefficient}$ = 0.070) and *MARCO* ($\mathrm{coefficient}$ = 0.058) as the main risk factors linked to poor survival, as shown in Fig. 3C and Table 2. On the other hand, protective markers, including the tumor suppressor *TP63* ($\mathrm{coefficient}$ = -0.053) and the lncRNA *AB306139* ($\mathrm{coefficient}$ = -0.060) were

elevated in the signature, indicating that immune surveillance and epithelial integrity maintenance provide a survival benefit.

Independence of the Prognostic Model

We performed univariate and multivariate Cox regression analysis using common clinicopathological factors to assess the gene signature's clinical usefulness. Together with factors including age, ER status, and PR status, the Risk Score was significantly correlated with overall survival in the univariate analysis (HR = 4.54, 95% CI: 3.18–6.49, $p < 0.001$) (Table 4).

Table 4
Univariate and Multivariate Cox regression analysis of the gene signature and clinicopathological characteristics

Variable	Univariate HR (95% CI)	P-value	Multivariate HR (95% CI)	P-value
Risk Score	4.54 (3.18–6.49)	< 0.001	3.44 (2.10–5.65)	< 0.001
Age	1.11 (1.08–1.13)	< 0.001	1.09 (1.07–1.12)	< 0.001
ER Status (Pos vs Neg)	0.39 (0.23–0.67)	0.001	1.15 (0.48–2.78)	0.752
PR Status (Pos vs Neg)	0.35 (0.22–0.56)	< 0.001	0.64 (0.32–1.26)	0.196
HER2 Status (Pos vs Neg)	0.75 (0.34–1.62)	0.462	0.77 (0.34–1.74)	0.525
Ki67 Status (Pos vs Neg)	1.40 (0.89–2.22)	0.150	1.03 (0.56–1.89)	0.934
Grade (G2 vs G1)	1.09 (0.54–2.20)	0.813	0.88 (0.41–1.88)	0.746
Grade (G3 vs G1)	1.73 (0.86–3.49)	0.124	0.79 (0.31–2.00)	0.623

[INSERT FIGURE 4 **HERE]**

A multivariate Cox regression model was then created to take confounding variables into consideration. The Risk Score demonstrated its prognostic independence by maintaining its statistical significance and showing a strong predictive value (HR = 3.44, 95% CI: 2.10–5.65, $p < 0.001$). The robustness of the gene signature in capturing tumor aggressiveness beyond standard clinical parameters is highlighted by the fact that, although Age remained a significant factor (HR = 1.09, $p < 0.001$), traditional pathological markers like ER and PR status lost their independent prognostic significance ($p > 0.05$) when the Risk Score was incorporated (Fig. 4).

Prognostic Value and Risk Stratification

Patients were divided into high- and low-risk groups according to the median risk score in order to verify the clinical usefulness of the signature. Risk ratings showed a substantial association with patient

outcomes, as Fig. 5A illustrates, with a considerably higher death rate in the high-risk group. A substantial difference in prognosis was verified by Kaplan-Meier survival analysis **(**Fig. 5B**)**, with the high-risk group exhibiting significantly worse overall survival (OS) than the low-risk group ($p < 0.0001$).

[INSERT FIGURE 5 **HERE]**

Time-dependent ROC analysis was used to assess the prediction accuracy of the model **(**Fig. 5D**)**. With an Area Under the Curve (AUC) of **0.847 at one year** and **0.784 at three**, the signature demonstrated significant predictive efficiency for early-stage outcomes. The molecular signature is most sensitive for predicting early disease development, as seen by the loss in predictive power during longer follow-up periods (5-year AUC = 0.543). Additionally, the biological foundation of the risk score was confirmed by the expression landscape of the hallmark genes **(**Fig. 5C**).** While immune-protective variables like *CXCL13* and *TCRVB* were enriched in the low-risk group, proliferation markers like *TK1* were consistently increased in high-risk individuals.

Functional Enrichment and Biological Insights

Functional enrichment analysis was used to clarify the biological processes behind the prognostic signature. Twenty of the 22 genes in the final prognostic signature were successfully mapped to Entrez Gene IDs (genes *TCRVB* and *AB306139* were not able to be mapped and were not included in the enrichment analysis). Nineteen of these mapped genes were successfully annotated in the database for the Gene Ontology (GO) study.

The signature is functionally biased toward immunological modulation, according to GO enrichment analysis **(**Fig. 6A**)**. Among the Molecular Function (MF) phrases that were most substantially enriched were chemokine activity, G protein-coupled receptor binding, and CCR chemokine receptor binding. This enrichment indicates that the signature's predictive significance is mostly dependent on its capacity to record receptor–ligand interactions in the tumor immunological milieu.

[INSERT FIGURE 6 **HERE]**

A targeted enrichment of the Drug metabolism – other enzymes pathway ($p < 0.05$) was found by complementary KEGG pathway analysis **(**Fig. 6B**)**. Three important biomarkers—*GSTM5, NAT1*, and *TK1*—drive this relationship, suggesting a possible connection between patient survival and xenobiotic detoxifying ability. Additionally, phenotype-level insight was obtained by Gene Set Enrichment Analysis (GSEA) **(**Fig. 6C**)**, which showed a highly significant enrichment of the Proteasome pathway (FDR = 1.10×10^{-8}) in the high-risk group. The vigorous turnover of intracellular proteins necessary for quick tumor growth is correlated with the significant overexpression of proteasomal components.

Development of a Clinical Nomogram

By combining variables such as age, tumor grade, and the Risk Score, a prognostic nomogram was created to evaluate individual risk (Fig. 7A). Each variable is given a point value between 0 and 100, and the total number of points predicts 1-, 3-, and 5-year OS. The multivariate model is shown for clinical usage via the nomogram.

[INSERT FIGURE 7 **HERE]**

Calibration curves were used to assess the model's accuracy (Fig. 7B). Good calibration was shown by the 1-year (Red) and 3-year (Blue) plots' strong concordance with observed survival. Clinical usefulness was evaluated using Decision Curve Analysis (DCA) (Fig. 7C). Across threshold probabilities (0.0 to 0.4), the nomogram (colored line) provided more net benefits than “treat-all” or “treat-none” techniques, enhancing decision-making without needless interventions.

Discussion

In this study, we developed a robust 22-gene prognostic signature that precisely predicts overall survival in breast cancer using the extensive SCAN-B cohort. Recent genomic research has brought to light the limitations of conventional TNM staging in reflecting the genetic heterogeneity of breast cancers[20]. By combining genome-wide survival screening with differential expression analysis, we identified a specific "Proliferation-Immune" axis of tumor progression, moving beyond basic data mining. Our stratification demonstrates that this signature effectively differentiates patients with different clinical outcomes, with the high-risk group showing significantly reduced survival ($P < 0.0001$). Our results are consistent with current multi-omics research that indicates proliferative markers and immune microenvironment characteristics together greatly improve prognosis accuracy when compared to clinical indicators alone[21]. In contrast to conventional indicators such as ER/PR status and Ki67 index, our multivariate study verified that the 22-gene signature represents an independent predictive predictor (HR = 3.44). This outcome is consistent with prior high-impact research where transcriptome markers outperformed conventional pathological staging in multivariate models in terms of predictive usefulness[22].

The Proliferation-Immune Axis: The selected genes represent distinct features of tumor aggressiveness. The high weight of *TK1* (+ 0.070), a key enzyme in DNA salvage, underscores its role as a strong molecular indicator of proliferation. In breast cancer, elevated *TK1* levels have repeatedly been identified as a sign of adverse outcomes and rapid tumor progression[23]. Similarly, *MARCO* (+ 0.058) aligns with its role in defining immunosuppressive tumor-associated macrophages (TAMs), which facilitate immune tolerance[24,25]. On the other hand, the protective factors *CXCL13* (-0.039) and *TCRVB* (-0.017) demonstrate the vital role of the adaptive immune microenvironment. Recent groundbreaking studies published in Nature have shown that the recruitment of B-cells and T-cells is facilitated by the *CXCL13*-driven creation of Tertiary Lymphoid Structures (TLS), a process that is closely associated with positive clinical outcomes and immunotherapy response[26,27].Collectively, these findings support a conceptual model of breast cancer progression called the “Proliferation-Immune Axis,” as shown in Fig. 8.

[Insert Fig. 8 **Here]**

Functional enrichment analyses provided biological insights into why the signature performs well. There is significant enrichment in chemokine-related functions (GO Analysis), especially with *CXCL13* and *CCL8*, emphasizing the importance of immune cell recruitment for survival[28]. The signature detects important metabolic changes in addition to immunity. The "Drug metabolism – other enzymes" pathway, which is driven by *GSTM5* and *NAT1*, was shown to be enriched by the KEGG study. Prior research has connected *NAT1* expression to xenobiotic detoxification and metabolic adaptability in breast cancer, which affects the susceptibility of the malignancy to chemotherapy[29].

The Proteasome pathway (FDR = 1.10×10^{-8}) was shown to be a prominent characteristic of the high-risk phenotype concurrently with the GSEA results. The ubiquitin-proteasome system (UPS) is hyperactivated in aggressive cancers, which allows the cells to handle proteotoxic stress while dividing quickly[30].

We created a composite nomogram that combines the Risk Score with age and tumor grade to close the gap between genetic discovery and clinical practice. As advised by precision oncology guidelines, the therapeutic value of such integrated models is supported by the greater net benefit demonstrated in our Decision Curve Analysis (DCA)[31]. Interestingly, the time-dependent ROC analysis showed that the short-term (1-year AUC = 0.847) predictive accuracy was higher than the long-term (5-year AUC = 0.543). This temporal disparity aligns with the idea of "tumor dormancy," according to which processes less reliant on active cell cycle drive late recurrence, while transcriptome signs of proliferation frequently predict early recurrence[32].

Despite these promising results, our study has limitations. While the SCAN-B cohort provides high-quality sequencing data, the retrospective design requires prospective validation in independent clinical trials. The model's accuracy diminishes over time (5-year AUC = 0.543), indicating a need for more refined models to predict late recurrences. Additionally, although 20 genes were mapped for functional analysis, two genes (*AB306139* and *TCRVB*) lacked annotation in standard databases, possibly missing some biological insights. Finally, external validation with datasets like TCGA and experimental validation (e.g., qRT-PCR or immunohistochemistry) of key drivers *TK1* and *CXCL13* is essential before clinical implementation.

In conclusion, this study identifies a novel 22-gene prognostic signature that stratifies breast cancer patients by capturing the interplay between proteasome-driven tumor proliferation and chemokine-mediated immune surveillance. By integrating this molecular signature with standard clinicopathological features, we constructed a clinically applicable nomogram that improves risk prediction and supports personalized therapeutic decision-making. These findings highlight the proteasome-chemokine axis as a potential therapeutic target and underscore the value of integrating metabolic and immune biomarkers for precision oncology.

Methods

Data Acquisition and Pre-processing

A cohort comprising 1,185 breast cancer patients, each with matched clinical annotations and comprehensive whole-transcriptome profiles, was analyzed. The data were sourced from the Gene Expression Omnibus (GEO) database[33] (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE96058. The dataset was generated by the *Swedish Cancerome Analysis Network - Breast* (SCAN-B) initiative and was originally published[34].

RNA sequencing data were obtained using the Illumina platform. Data processing and statistical analysis were performed using R (version 4.5.2). Quality control measures excluded samples lacking key survival metrics, such as Overall Survival time or status. Raw count data were background-corrected and normalized with the *edgeR*[35] and *limma*[36] packages. Gene expression levels were transformed using the $\log2(x+1)$ function to stabilize variance and mitigate heteroscedasticity. Principal Component Analysis (PCA) was conducted to visualize sample distribution and detect possible batch effects or outliers. Data distribution consistency was evaluated using boxplots and violin plots before further analysis. The overall study design and analytical workflow are summarized in Fig. 1.

[INSERT FIGURE 1 **HERE]**

Screening for Survival-Associated Candidate Genes

Genome-wide Univariate Screening

A dual-screening strategy was implemented to identify reliable prognostic indicators among survival-associated candidate genes and to minimize the false discovery rates common in high-dimensional datasets. Initially, a genome-wide univariate screening was conducted using Cox proportional hazards regression analysis on the global transcriptome, retaining genes significantly associated with overall survival (OS) at $p < 0.05$.

Differential Expression Analysis (DEA)

Subsequently, differential expression analysis (DEA) was conducted to compare gene expression profiles between distinct phenotypic groups, such as high-grade and low-grade tumors, utilizing the *DESeq2* package[37] with a negative binomial distribution. Differentially expressed genes (DEGs) were identified based on an absolute log2 fold change exceeding 1 and adjusted *p*-values below 0.05(Benjamini-Hochberg)[38]. The final candidate set for prognostic modeling comprised the intersection of survival-associated genes and statistically significant DEGs, visualized through a Venn diagram.

Construction of the Prognostic Signature via LASSO

To address multicollinearity among candidate genes and develop a simplified model, the Least Absolute Shrinkage and Selection Operator (LASSO) Cox regression algorithm was utilized through the *glmnet* R package[39]. This method mitigates overfitting by imposing a penalty on the absolute magnitude of regression coefficients[40]. The optimal penalty parameter (λ) was identified via 10-fold cross-validation, specifically choosing the λ value that is within one standard error of the minimum value (λ_{1se}) to ensure the stability of the model. A robust risk score was calculated for each patient using the following formula:

$$\text{Risk Score} = \sum_{i=1}^{n} \beta\ (i) \times\ \text{Expr}\ (\text{i})$$

Where β_i represents the LASSO regression coefficient and Expr_i is the log-transformed expression value of gene i.

Prognostic Validation and Risk Stratification

Patients were stratified into High-Risk and Low-Risk cohorts according to the median risk score. Survival differences between these groups were assessed using Kaplan-Meier curves by *survival* package[41] and the log-rank test. The model's sensitivity and specificity for predicting 1-, 3-, and 5-year survival were evaluated with time-dependent receiver operating characteristic (ROC) curves using the *timeROC* package[42]. A heatmap depicting the genomic landscape of the risk model was generated with the *pheatmap* package, arranging patients by risk score and displaying the expression profiles of signature genes alongside clinical features and survival status. This visualization facilitated assessment of the signature's ability to distinguish between risk groups and the association between gene expression levels and risk classification. Furthermore, multivariate Cox regression analyses were performed to determine whether the risk score served as an independent prognostic factor after adjusting for standard clinicopathological variables, such as age, tumor grade, and estrogen receptor (ER) status.

Functional Enrichment Analysis

To elucidate the molecular mechanisms of the high-risk phenotype, a functional enrichment analysis was conducted using the *clusterProfiler* package[43]. This involved analyzing Gene Ontology (GO) terms and *Kyoto Encyclopedia of Genes and Genomes (KEGG)* pathways[44] to identify enriched biological processes. Additionally, *Gene Set Enrichment Analysis (GSEA)*[45] was performed to identify hallmark pathways (c2.cp.kegg.v7.4.symbols) that were significantly activated in the High-Risk group, using a permutation test with n = 1000 to derive normalized enrichment scores (NES).

Clinical Utility Assessment

Using the *rms* R program[46], A composite Nomogram was created to aid in clinical translation. To forecast individual survival odds, this method combined the genetic Risk Score with important clinical characteristics (age, stage, and grade). Calibration Curves are used to evaluate how well observed survival outcomes match expected probabilities. Decision Curve Analysis (DCA)[47] is used to measure the nomogram's net clinical benefit across a range of threshold probabilities, compared with treat-all or treat-none approaches.

Abbreviations

ABAT:4-Aminobutyrate Aminotransferase **AUC**:Area Under the Curve **CCL8**:C-C Motif Chemokine Ligand 8 **CI**:Confidence Interval **CXCL13**:C-X-C Motif Chemokine Ligand 13 **DCA**:Decision Curve Analysis **DEA**:Differential Expression Analysis **DEGs**:Differentially Expressed Genes **ER**:Estrogen Receptor **GEO**:Gene Expression Omnibus **GO**:Gene Ontology **GSEA**:Gene Set Enrichment Analysis **HER2**:Human Epidermal Growth Factor Receptor 2 **HR**:Hazard Ratio **KEGG**:Kyoto Encyclopedia of Genes and Genomes **LASSO**:Least Absolute Shrinkage and Selection Operator **MARCO**:Macrophage Receptor with Collagenous Structure **NES**:Normalized Enrichment Score **OS**:Overall Survival **PCA**:Principal Component Analysis **PR**:Progesterone Receptor **RNA-seq**:RNA Sequencing **ROC**:Receiver Operating Characteristic **SCAN-B**:Sweden Cancerome Analysis Network–Breast **SR-DEGs**:Survival-Relevant Differentially Expressed Genes **TAMs**:Tumor-Associated Macrophages **TCGA**:The Cancer Genome Atlas **TCRVB**:T-Cell Receptor Variable Beta **TK1**:Thymidine Kinase 1 **TNM**:Tumor–Node–Metastasis

Declarations

Ethics Approval and Consent to Participate

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. This study utilized exclusively publicly available, de-identified data from the Gene Expression Omnibus (GEO).

Consent for Publication

Not applicable, as the dataset contains no identifiable personal information.

Data Availability Statement

Publicly available datasets were analyzed in this study. The data can be found in the Gene Expression Omnibus (GEO) repository under accession number GSE96058 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96058).

Code Availability

The complete R scripts used for the LASSO regression analysis, survival modeling, and figure generation are openly available on GitHub at: https://github.com/shakawat96907/Breast-Cancer-22-Gene-Signature . The repository includes the source code (22_gene_signature_analysis.R) necessary to replicate the

findings. The datasets used, which are publicly accessible, can be found in the Gene Expression Omnibus (GEO) repository under accession number GSE96058.

Competing Interests

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

The author received no specific funding for this work.

Author Contributions

Md. Shakawat Hossain conceived the conceptual design, performed the bioinformatics analysis, interpreted the data, drafted the manuscript, and approved the final version for submission.

Generative AI Declaration

During the preparation of this work, the author used Gemini (Google) and Grammarly in order to improve language clarity, refine sentence structure, and organize the logical flow of the manuscript. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Acknowledgments

The author gratefully acknowledges the **Sweden Cancerome Analysis Network–Breast (SCAN-B)** initiative for making their transcriptomic and clinical data publicly available, without which this study would not have been possible.

Supplementary Information

Supplementary Material.docx: Table S3 (The 22-gene LASSO prognostic signature with biological annotations), **Table S4** (Detailed gene information and model coefficients), and **Figure S1** (Overview of clinicopathological features)

Supplementary Data.xlsx: Table S1 (Genome-wide univariate Cox regression findings), **Table S2** (Survival-relevant DEGs), and **Table S5** (Gene Ontology enrichment analysis results) are among the extensive datasets included in this file.

References

1. Cancer Communications – 2021 - Lei - Global patterns of breast cancer incidence and mortality A population-based cancer.

2. Sung H, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71:209–49.
3. Yi M, et al. Epidemiological trends of women's cancers from 1990 to 2019 at the global, regional, and national levels: a population-based study. Biomark Res. 2021;9:55.
4. Li T, et al. Global status and attributable risk factors of breast, cervical, ovarian, and uterine cancers from 1990 to 2021. J Hematol Oncol J Hematol Oncol. 2025;18:5.
5. Elbasheer MMA, Dodwell D, Gathani T. Understanding global variation in breast cancer mortality.
6. Arnold M. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast* (2022).
7. Schettini F. Dissecting the biological heterogeneity of HER2-positive breast cancer. (2021).
8. Ye Z. Multimodal diagnostic models and subtype analysis for neoadjuvant therapy in breast cancer.
9. Tamm A, et al. Supporting cancer research on real-world data: extracting colorectal cancer status and explicitly written TNM stages from free-text imaging and histopathology reports. BMJ Health Care Inf. 2025;32:e101521.
10. Shen H, et al. Breaking the heterogeneity barrier: a robust prognostic signature for survival stratification and immune profiling in triple-negative breast cancer. Front Immunol. 2025;16:1611917.
11. Wang S, et al. Integrating PANoptosis insights to enhance breast cancer prognosis and therapeutic decision-making. Front Immunol. 2024;15:1359204.
12. Mock A, Braun M, Scholl C, Fröhling S, Erkut C. Transcriptome profiling for precision cancer medicine using shallow nanopore cDNA sequencing. Sci Rep. 2023;13:2378.
13. Rakha EA, et al. Assessment of Predictive Biomarkers in Breast Cancer: Challenges and Updates. Pathobiology. 2022;89:263–77.
14. Xie P, An R, Yu S, He J, Zhang H. A novel immune subtype classification of ER-positive, PR-negative and HER2-negative breast cancer based on the genomic and transcriptomic landscape. J Transl Med. 2021;19:398.
15. Li C, et al. Integrating single-cell and bulk transcriptomic analyses to develop a cancer-associated fibroblast-derived biomarker for predicting prognosis and therapeutic response in breast cancer. Front Immunol. 2024;14:1307588.
16. Bueno-Fortes S, Berral-Gonzalez A, Sánchez-Santos JM, Martin-Merino M, De Las Rivas J. Identification of a gene expression signature associated with breast cancer survival and risk that improves clinical genomic platforms. Bioinforma Adv. 2023;3:vbad037.
17. Qian Y, Itzel T, Ebert M, Teufel A. Deep View of HCC Gene Expression Signatures and Their Comparison with Other Cancers. Cancers. 2022;14:4322.
18. Varnier R, et al. Using Breast Cancer Gene Expression Signatures in Clinical Practice: Unsolved Issues, Ongoing Trials and Future Perspectives. Cancers. 2021;13:4840.

19. Kwa M, Makris A, Esteva FJ. Clinical utility of gene-expression signatures in early stage breast cancer. Nat Rev Clin Oncol. 2017;14:595–610.
20. Wei Z, et al. Deep Learning-Based Multi-Omics Integration Robustly Predicts Relapse in Prostate Cancer. Front Oncol. 2022;12:893424.
21. Sammut S-J, et al. Multi-omic machine learning predictor of breast cancer therapy response. Nature. 2022;601:623–9.
22. Phakathi B, et al. PAM50 intrinsic subtypes, risk of recurrence score and breast cancer survival in HIV-positive and HIV-negative patients—a South African cohort study. Breast Cancer Res Treat. 2023;200:337–46.
23. Alegre MM, Robison RA, O'Neill KL. Thymidine Kinase 1: A Universal Marker for Cancer. Cancer Clin Oncol. 2013;2:159.
24. Cassetta L, Pollard JW. Tumor-associated macrophages. Curr Biol. 2020;30:R246–8.
25. Georgoudaki A-M, et al. Reprogramming Tumor-Associated Macrophages by Antibody Targeting Inhibits Cancer Progression and Metastasis. Cell Rep. 2016;15:2000–11.
26. Cabrita R, et al. Tertiary lymphoid structures improve immunotherapy and survival in melanoma. Nature. 2020;577:561–5.
27. Helmink BA, et al. B cells and tertiary lymphoid structures promote immunotherapy response. Nature. 2020;577:549–55.
28. Nagarsheth N, Wicha MS, Zou W. Chemokines in the cancer microenvironment and their relevance in cancer immunotherapy. Nat Rev Immunol. 2017;17:559–72.
29. Li P, Butcher NJ, Minchin RF, Arylamine. N-Acetyltransferase 1 Regulates Expression of Matrix Metalloproteinase 9 in Breast Cancer Cells: Role of Hypoxia-Inducible Factor 1-α. Mol Pharmacol. 2019;96:573–9.
30. Manasanch EE, Orlowski RZ. Proteasome inhibitors in cancer therapy. Nat Rev Clin Oncol. 2017;14:417–33.
31. written on behalf of AME Big-Data Clinical Trial Collaborative Group. Decision curve analysis: a technical note. Ann Transl Med. 2018;6:308–308.
32. Phan TG, Croucher PI. The dormant cancer cell life cycle. Nat Rev Cancer. 2020;20:398–411.
33. Barrett T, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012;41:D991–5.
34. Brueffer C, et al. Clinical Value of RNA Sequencing–Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network—Breast Initiative. JCO Precis Oncol. 2018;1–18. 10.1200/PO.17.00135.
35. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

36. Ritchie ME, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47–47.
37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
38. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Stat Methodol. 1995;57:289–300.
39. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33:1–22.
40. Daneshvar A, Mousa G. Regression shrinkage and selection via least quantile shrinkage and selection operator. PLoS ONE. 2023;18:e0266267.
41. Therneau TM, survival. Survival Analysis. 3.8-6 https://doi.org/10.32614/CRAN.package.survival (2001).
42. Blanche P, Dartigues J, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013;32:5381–97.
43. Wu T, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov. 2021;2:100141.
44. Kanehisa MKEGG. Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30.
45. Subramanian A et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550 (2005).
46. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, NY: Springer New York; 2001. 10.1007/978-1-4757-3462-1.
47. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Med Decis Mak. 2006;26:565–74.

Figures

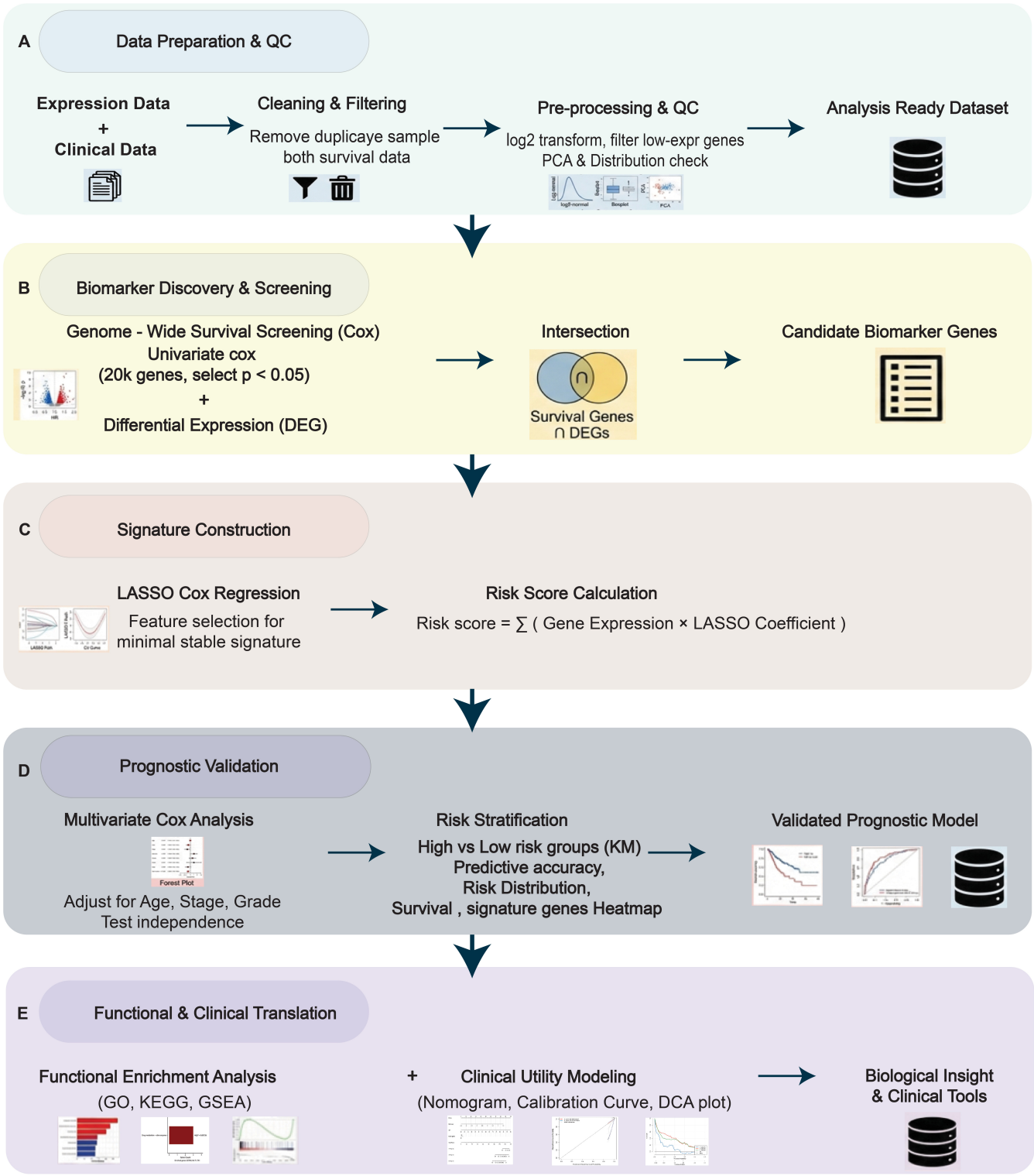


Figure 1

Study Workflow. (A) Pre-processing, standardization, and quality assurance of transcriptomic and clinical datasets. (B) Identification of potential biomarkers by integrating survival-associated genes identified through univariate Cox regression with differentially expressed genes (DEGs). (C) Construction of a prognostic risk signature using LASSO Cox regression analysis. (D) Validation of predictive accuracy using time-dependent ROC curves, multivariate independent testing, and Kaplan-Meier survival analysis.

(E) Development of a clinical nomogram and evaluation of clinical utility through Decision Curve Analysis (DCA) and functional enrichment analysis (GO, KEGG, GSEA).

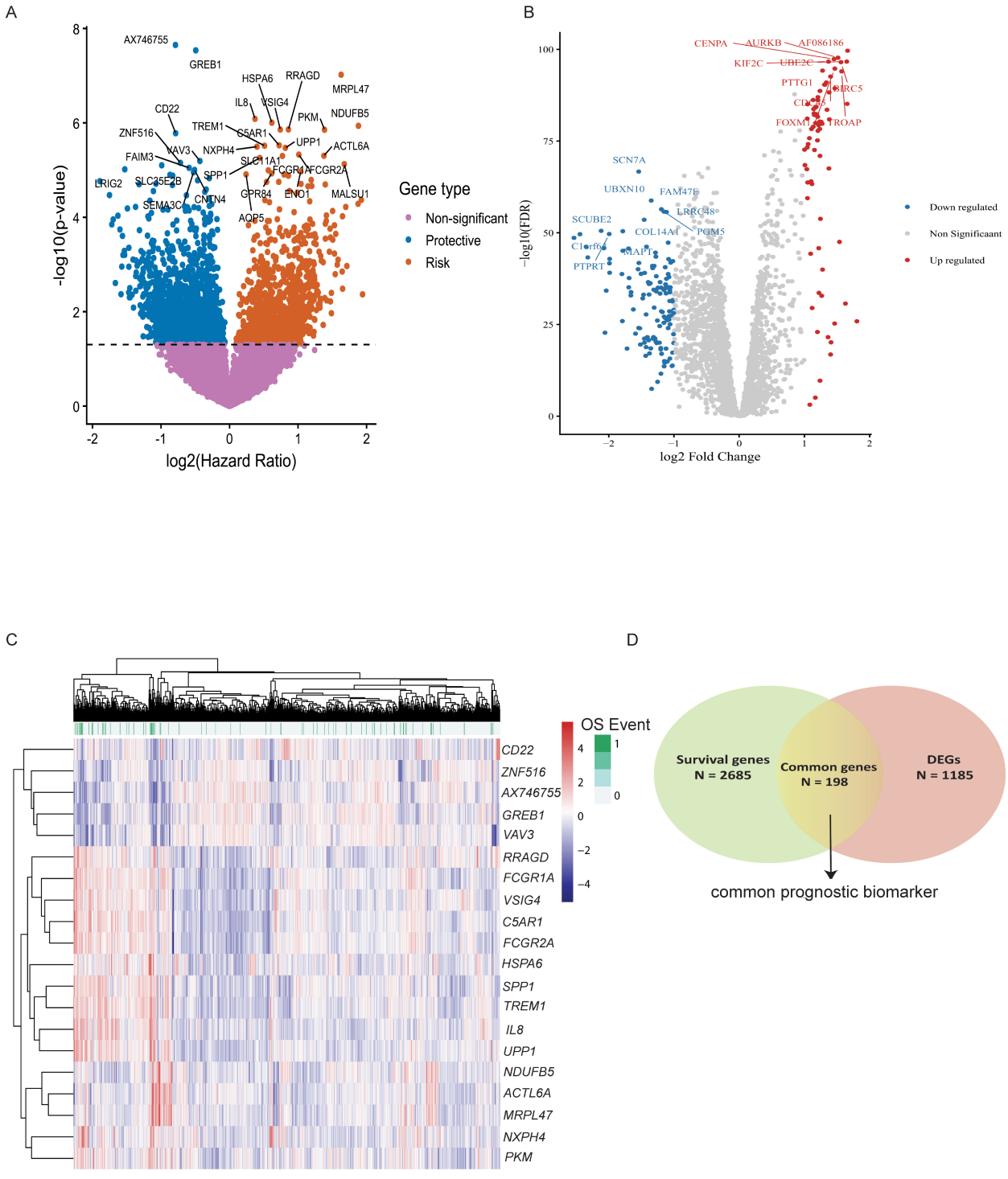


Figure 2

Identification of a Prognostic Gene Signature through Integration of Survival and Differential Expression Analyses. (A) Genome-wide univariate Cox regression volcano plot. The x-axis displays the log2 Hazard Ratio (HR), while the y-axis represents the -log10 (P-value). Genes associated with increased risk (HR > 1,

highlighted in red) and protection (HR < 1, highlighted in blue) are emphasized (P < 0.05). (B) Volcano plot illustrating differentially expressed genes (DEGs) when comparing tumor and normal tissues. Significant DEGs are characterized by |logFC| > 1 and adjusted P-value < 0.05 (upregulated genes shown in red; downregulated in blue). (C) Hierarchical clustering heatmap of the top 20 survival-relevant DEGs (SR-DEGs). Rows denote genes, and columns denote patient samples. The color gradient indicates Z-score normalized expression values, revealing distinct stratification among clinical risk groups. (D) Venn diagram depicting the intersection between significant DEGs and genes associated with survival, thereby identifying the final set of biologically relevant prognostic markers.

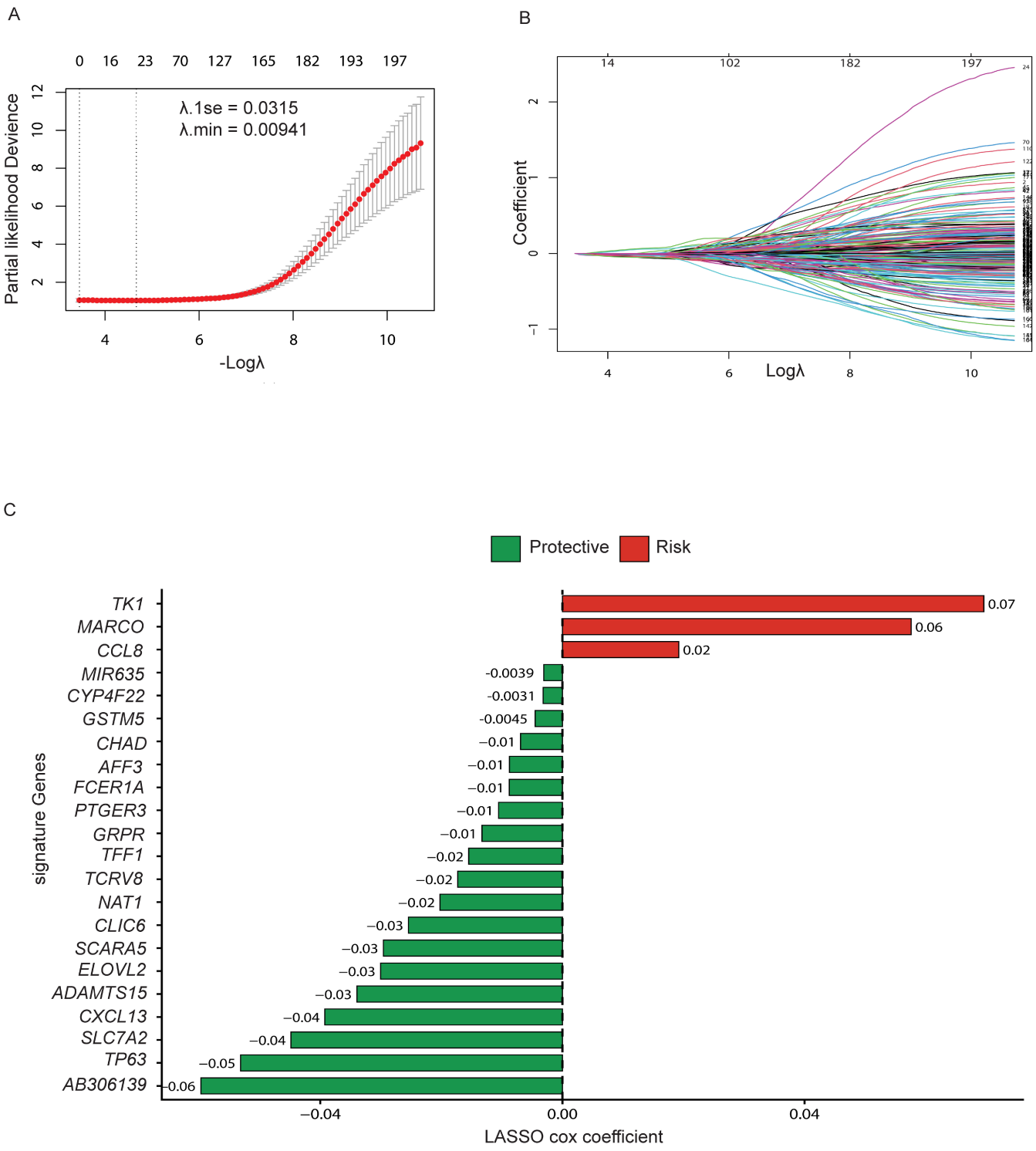


Figure 3

Construction of a prognostic gene signature using LASSO Cox regression. (A) Tuning parameter selection via 10-fold cross-validation, plotting partial likelihood deviance against log(λ). Vertical dotted lines show λmin (minimum deviance) and λ1se (1-standard error). (B) LASSO coefficient profiles for survival-related genes, showing how each gene's coefficient changes with increasing λ. (C) Distribution

of LASSO coefficients for the prognostic signature, with red (positive) indicating risk factors and blue (negative) indicating protective factors based on the λmin model.

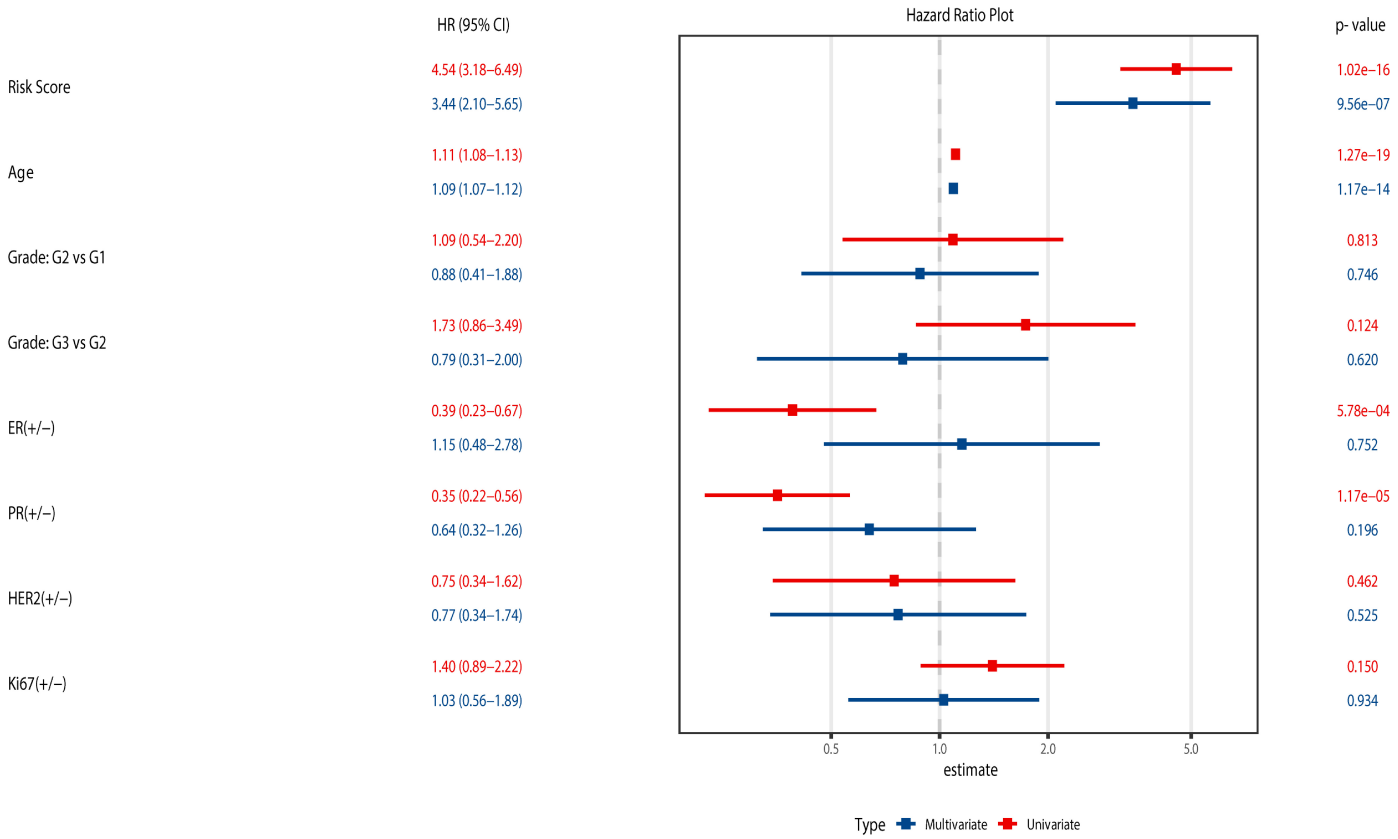


Figure 4

Independent Prognostic Value of the Gene Signature. Forest plot visualizing the hazard ratios (HR) and 95% confidence intervals (CI) derived from multivariate Cox regression analysis. The model adjusts for standard clinicopathological covariates, including Age, Grade, and Receptor status (ER, PR, HER2, Ki67). The Risk Score (Red) remains a significant predictor of overall survival independent of clinical factors.

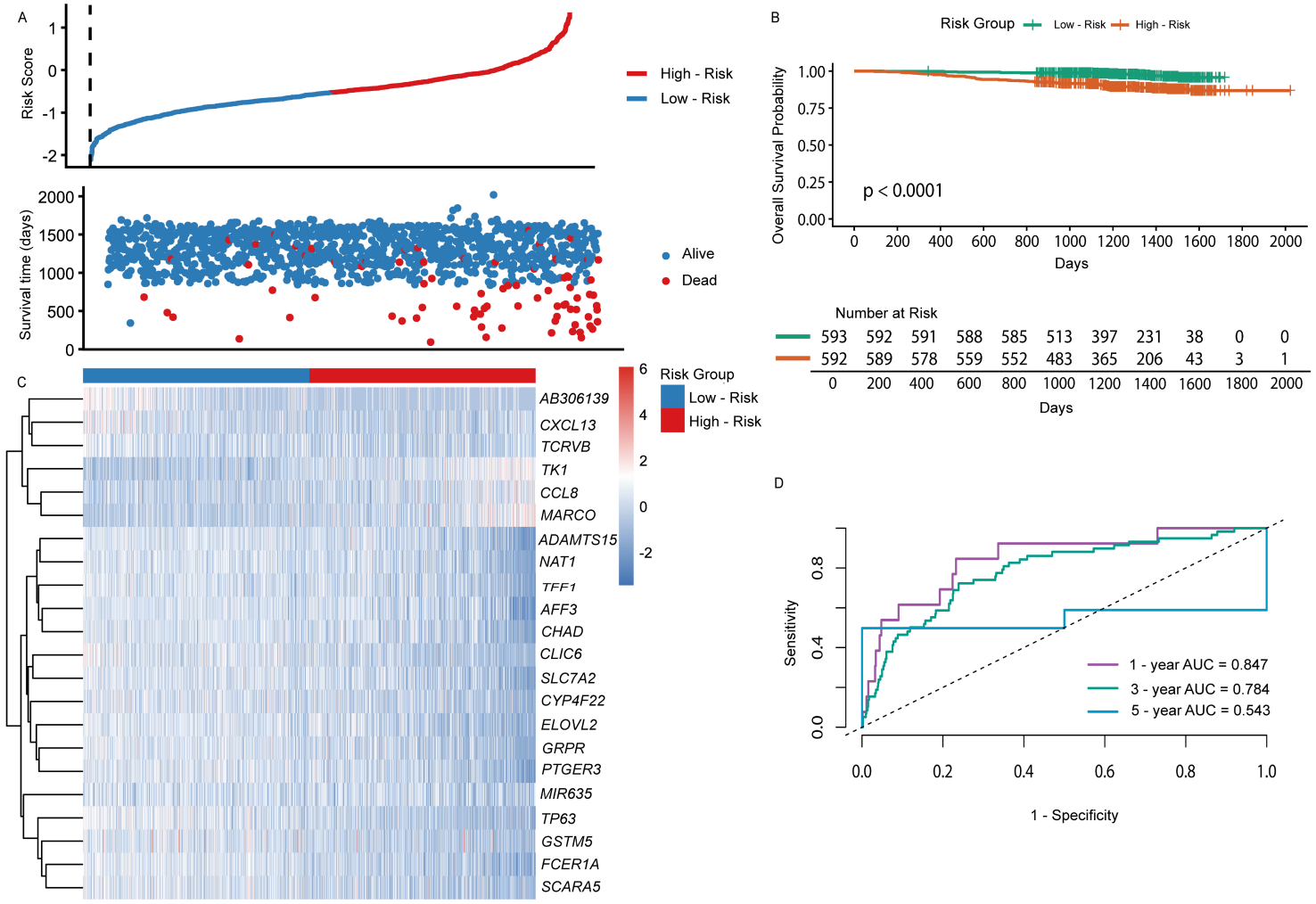


Figure 5

Prognostic Performance of the 22-Gene Signature. (A) Risk scores and survival status: risk scores are shown in ascending order; survival status, with blue for alive and red for deceased, shows more deaths in the high-risk group. (B) Kaplan-Meier analysis: high-risk (orange) had significantly poorer survival ($p < 0.0001$, Log-rank test). (C) Hierarchical heatmap: gene expression differences between low- and high-risk groups, red for upregulation, blue for downregulation. (D) ROC analysis: AUC of 0.847 (1-year), 0.784 (3-year), and 0.543 for 5-year survival.

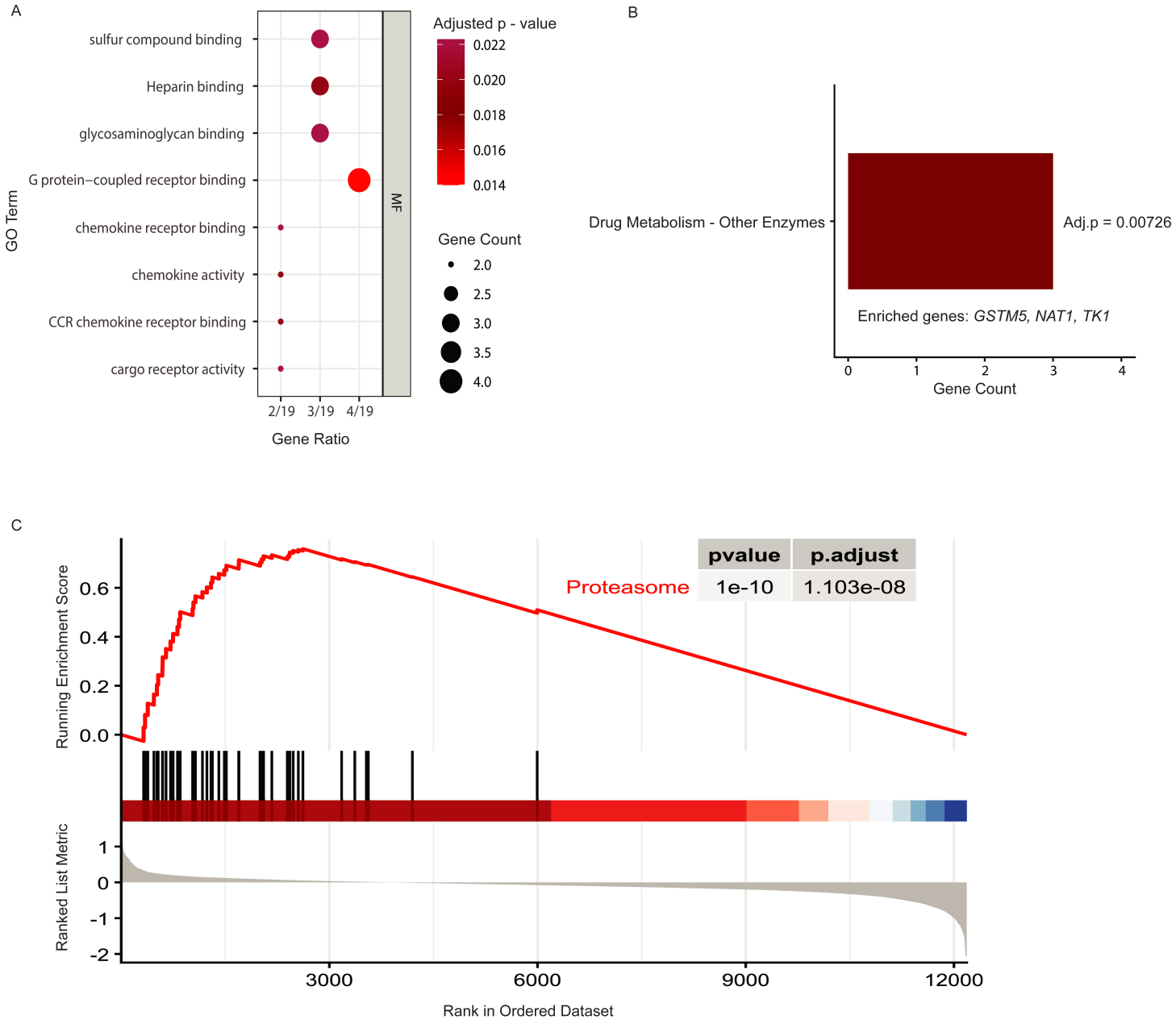


Figure 6

The functional landscape and pathway enrichment of a 22-gene prognostic signature. (A) GO analysis of molecular functions highlights chemokine activity and receptor binding, with dot size indicating gene number and color showing adjusted P-value. (B) KEGG pathway analysis identifies "Drug metabolism - other enzymes" ($P < 0.01$), driven by *GSTM5*, *NAT1*, and *TK1*. (C) GSEA plot reveals significant enrichment of the "Proteasome" pathway (FDR = 1.10×10^{-8}) in high-risk patients, indicating increased protein degradation.

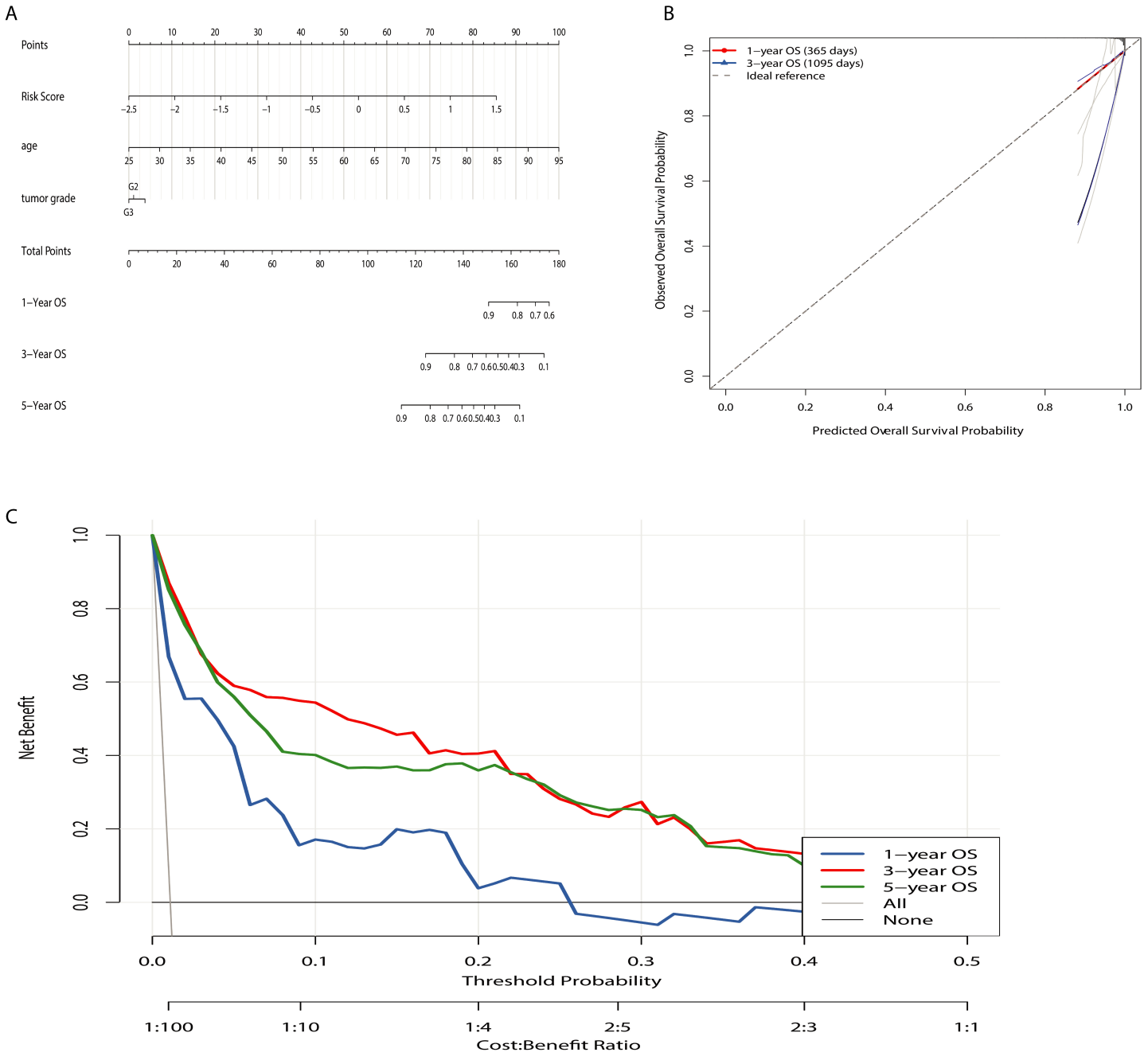


Figure 7

Construction and Evaluation of a Prognostic Nomogram for Overall Survival. (A) Prognostic nomogram for predicting 1-, 3-, and 5-year overall survival (OS). The nomogram integrates age, tumor grade, and Risk Score. To use, locate the patient's value for each variable on the corresponding axis, draw a straight line upward to the "Points" axis to determine the score, sum the points for all variables, and locate the total on the "Total Points" axis to estimate survival probabilities. **(B)** Calibration curves for 1-year (red) and 3-year (blue) OS. The x-axis represents the nomogram-predicted survival probability, and the y-axis represents the observed survival probability. The gray dashed line represents the ideal prediction. **(C)** Decision Curve Analysis (DCA) for the nomogram. The y-axis measures the net benefit. The colored lines

represent the net benefit of the nomogram at 1, 3, and 5 years across different threshold probabilities, compared to the strategies of assuming all patients have the event (gray line) or none do (black line).

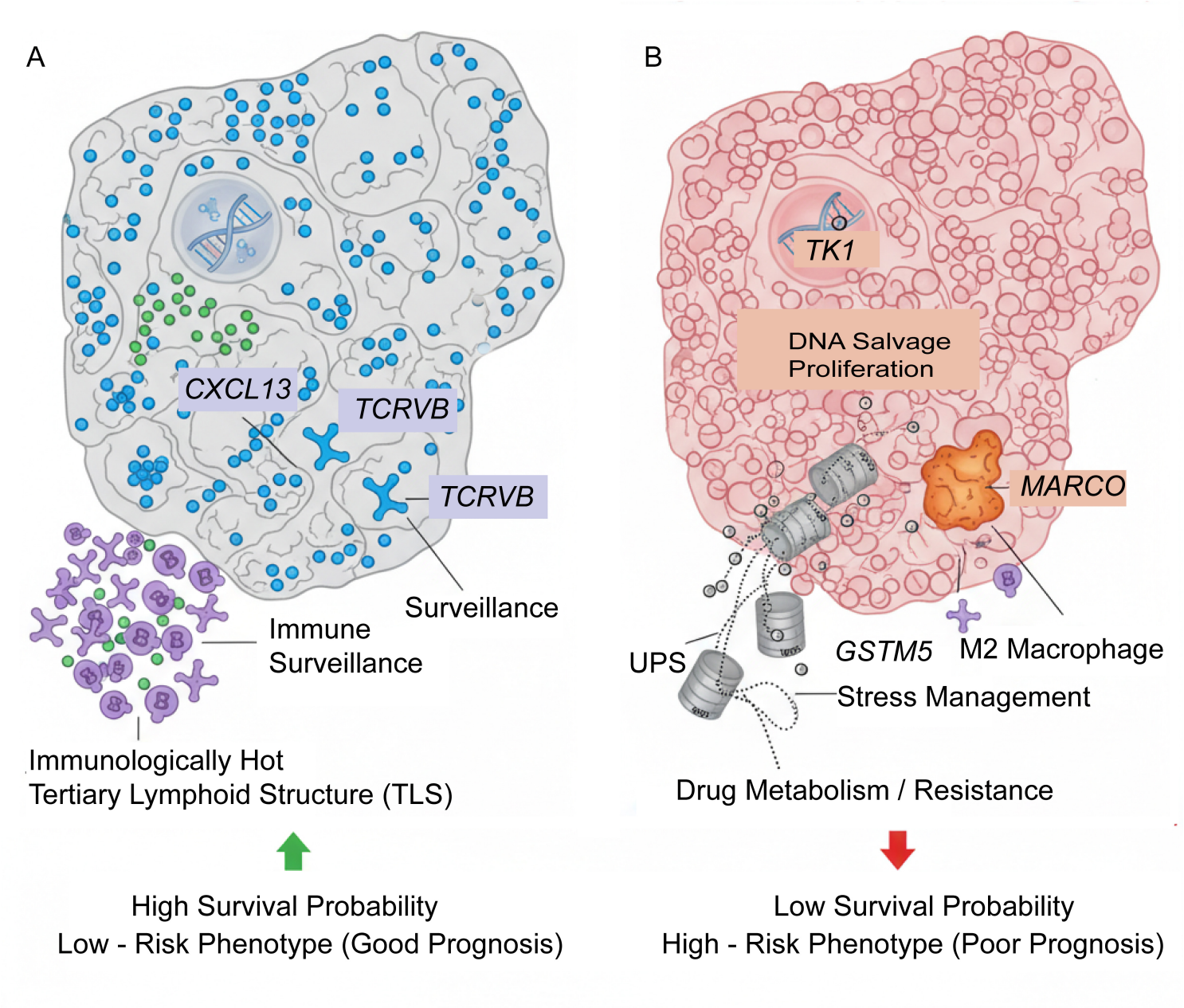


Figure 8

The "Proliferation-Immune Axis" in breast cancer progression. An immunologically "hot" microenvironment is indicative of the **(A)** low-risk phenotype (good prognosis). Effective tumor monitoring is facilitated by the recruitment of adaptive immune cells and the creation of Tertiary Lymphoid Structures (TLS) driven by high expression of *CXCL13* and *TCRVB*. **(B)** Poor Prognosis (High-Risk Phenotype): characterized by immunological evasion and vigorous growth. In order to cope with proteotoxic stress, this phenotype depends on *TK1*-driven DNA synthesis and proteasome (UPS) hyperactivation. *MARCO* upregulation promotes immunosuppressive M2 macrophages, whereas *GSTM5*

helps with drug resistance and detoxification. TLS stands for Tertiary Lymphoid Structure, while UPS stands for Ubiquitin-Proteasome System.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx.docx
- SupplementaryData.xlsx.xlsx