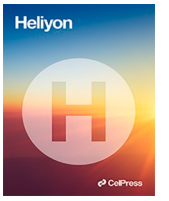
Heliyon 10 (2024) e35553

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Construction and validation of a ubiquitination-related prognostic risk score signature in breast cancer

Kexin Feng [a,1], Xin He [a,1], Ling Qin [a,1], Zihuan Ma [b], Siyao Liu [b], Ziqi Jia [a], Fei Ren [a], Heng Cao [a], Jiang Wu [a], Dongxu Ma [a], Xiang Wang [a,**], Zeyu Xing [a,*]

[a] *Department of Breast Surgical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China*
[b] *Beijing ChosenMed Clinical Laboratory Co. Ltd., Jinghai Industrial Park, Economic and Technological Development Area, Beijing, 100176, China*

ARTICLE INFO

Keywords:
Breast cancer
Biomarker
Risk score
Prognosis
Ubiquitination

ABSTRACT

Background: Breast cancer (BC) is a highly common form of cancer that occurs in many parts of the world. However, early -stage BC is curable. Many patients with BC have poor prognostic outcomes owing to ineffective diagnostic and therapeutic tools. The ubiquitination system and associated proteins were found influencing the outcome of individuals with cancer. Therefore, developing a biomarker associated with ubiquitination genes to forecast BC patient outcomes is a feasible strategy.

Objective: The primary goal of this work was to develop a novel risk score signature capable of accurately estimate the future outcome of patients with BC by targeting ubiquitinated genes.

Methods: Univariate Cox regression analysis was conducted utilizing the E1, E2, and E3 ubiquitination-related genes in the GSE20685 dataset. Genes with $p < 0.01$ were screened again using the Non-negative Matrix Factorization (NMF) algorithm, and the resulting hub genes were composed of a risk score signature. Patients were categorized into two risk groups, and the predictive effect was tested using Kaplan-Meier (KM) and Receiver Operating Characteristic (ROC) curves. This risk score signature was later validated using multiple external datasets, namely TCGA-BRAC, GSE1456, GSE16446, GSE20711, GSE58812 and GSE96058. Immuno-microenvironmental, single-cell, and microbial analyses were also performed.

Results: The selected gene signature comprising six ubiquitination-related genes (*ATG5*, *FBXL20*, *DTX4*, *BIRC3*, *TRIM45*, and *WDR78*) showed good prognostic power in patients with BC. It was validated using multiple externally validated datasets, with KM curves showing significant differences in survival ($p < 0.05$). The KM curves also demonstrated superior predictive ability compared to traditional clinical indicators. Single-cell analysis revealed that Vd2 gd T cells were less abundantin the low-risk group, whereas patients in the high-risk group lacked myeloid dendritic cells. Tumor microbiological analysis revealed a notable variation in microorganism diversity between the high- and low-risk groups.

Conclusion: This study established an risk score signature consisting of six ubiquitination genes, that can accurately forecast the outcome of patients with BC using multiple datasets. It can provide personalized and targeted assistance to provide the evaluation and therapy of individuals having BC.

* Corresponding author.
** Corresponding author.
E-mail addresses: xiangw@vip.sina.com (X. Wang), dr.xing@cicams.ac.cn (Z. Xing).
[1] These authors have contributed equally to this work and share first authorship.

https://doi.org/10.1016/j.heliyon.2024.e35553
Received 10 May 2024; Received in revised form 30 July 2024; Accepted 31 July 2024
Available online 2 August 2024

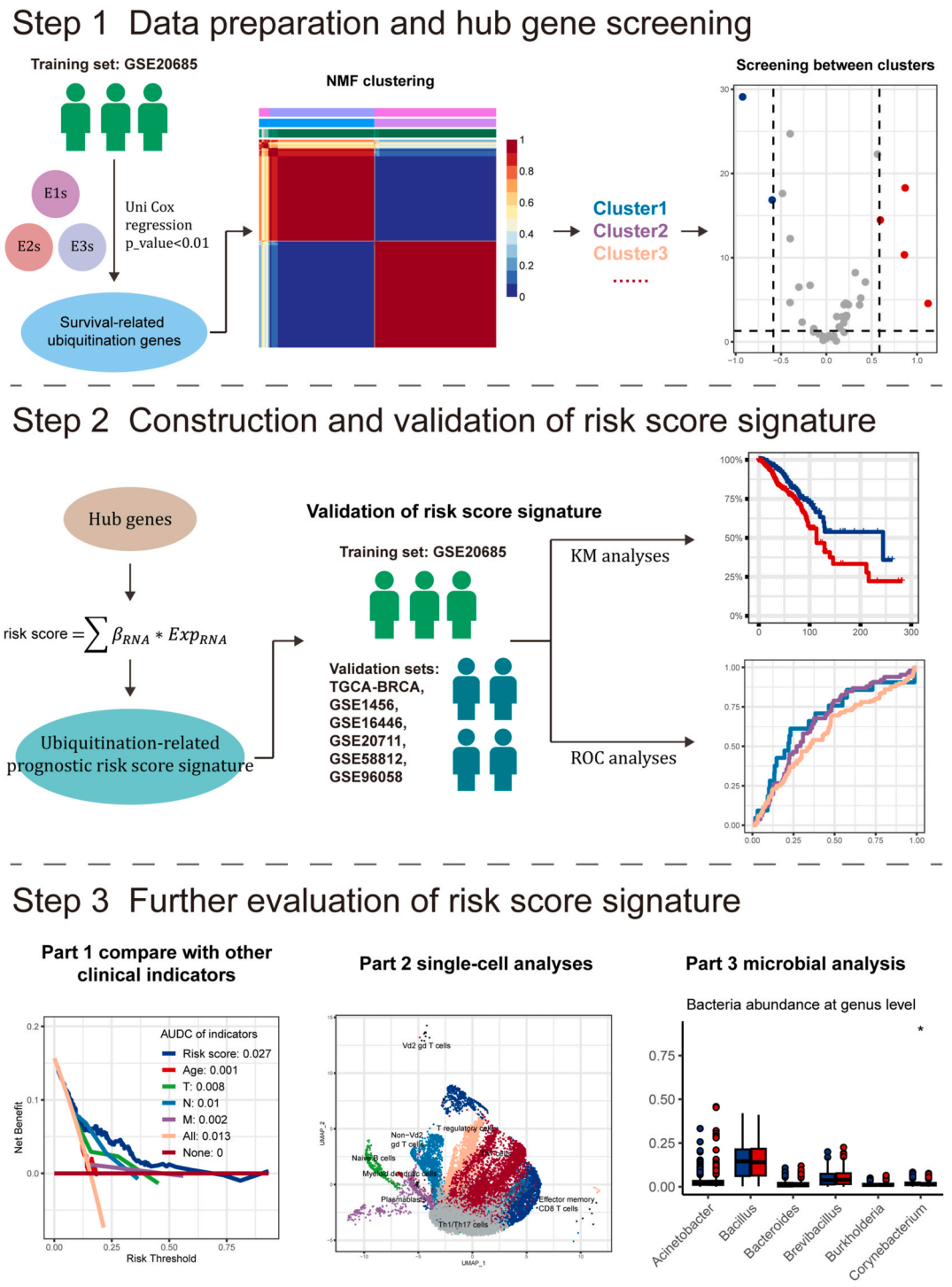


Fig. 1. The workflow of this study.

1. Introduction

The ubiquitin-proteasome system serves as a cellular protein degradation pathway widely present in eukaryotic organisms. It notably affects crucial cellular processes such as cell signaling, cell cycle regulation, receptor trafficking, and immune responses [1]. The ubiquitin-proteasome system consists primarily of ubiquitin, ubiquitin-activating enzymes, ubiquitin-conjugating enzymes, ubiquitin ligases, proteasomes, and substrate proteins. Ubiquitin-activating enzymes, ubiquitin-conjugating enzymes, and ubiquitin ligases attach multiple ubiquitin molecules to the substrate protein to form a ubiquitin chain. Subsequently, the corresponding receptors on the proteasome interact with the ubiquitin chain on the substrate, allowing the protein substrate to unfold and enter the proteasome for degradation [2,3]. Studies have confirmed that ubiquitination regulates metabolic processes in tumor cells, leading to metabolic reprogramming in various cancers [4,5]. Drugs targeting the ubiquitin-proteasome system are also considered potential novel cancer treatment strategies [6].

Breast cancer (BC) without metastasis is regarded as a curable disease, challenges persist owing to the substantial number of cases and limitations in advanced diagnosis and treatment, leading to delayed detection in some patients [7–9]. Moreover, more than 80 % of the BC cases in China are not detected at an early stage [10]. Discovering important indicators to distinguish BC prognosis has considerable potential for improving medical evaluation and therapy effectiveness [11], thereby decreasing treatment-related side effects and negative health outcomes. Therefore, the identification of biomarkers capable of effectively predicting BC prognosis is important.

Currently, the linked protein molecules were revealed to affect the ability to survive of individuals with cancer [12,13]. Investigating the changes with the expression of certain genes may provide an improved comprehension of the mechanistic involvement of ubiquitination in diseases as well as help with clinical decision-making. A recent study focused on the ubiquitination system in 1086 tumor cell lines. These results reveal that the ubiquitin ligase complex composed of UBA6, BIRC6, KCMF1, and UBR4 is essential for the survival of a highly aneuploid subpopulation of epithelial tumors, which may offer a therapeutic opportunity to selectively eliminate these cancer cells [14]. Another study focused on BC and ovarian cancer and found that CUL3, a kind of E3s, can interact with BECN1, facilitating K48-linked ubiquitination and degradation of BECN1, which could inhibit cellular autophagy and promote tumor development [15]. In summary, these studies revealed the value of investigating ubiquitination gene expression in clinical decision-making for BC. Moreover, researchers have identified that bacteria may influence the ubiquitination process of the host cell, thereby impacting the autophagy and immune processes of the host cell [16,17], suggesting the potential for microorganisms present in the tumor microenvironment to influence tumor initiation and progression.

In this study, we collected datasets from diverse populations, including those from Eastern and Western countries, to comprehensively evaluate the association among ubiquitination-related genes and prognosis of individuals with BC. Simultaneously, we identified hub genes to construct ubiquitination-related prognostic signatures and assessed their predictive capabilities for outcomes of patients with BC. Additionally, we conducted single-cell, tumor immune microenvironment, and tumor microbial composition analyses to provide further insights and assistance for clinical diagnosis and therapeutic decisions.

2. Methods

2.1. Data preparation

Fig. 1 illustrates the method of present research.

The gene list of ubiquitin-activating enzymes (E1s), ubiquitin-conjugating enzymes (E2s), and ubiquitin-protein ligases (E3s) was obtained from iUUCD 2.0 database (http://iuucd.biocuckoo.org/).

Eightdatasets were included in this study: a training dataset, GSE20685; a single-cell dataset, GSE176068; and six validation datasets. The RNA expression profiles and relevant clinical data from GSE1456, GSE16446, GSE20685, GSE20711, GSE58812, and GSE96058. Gene expression data and related clinical data of TCGA-BRCA were obtained. The analysis also incorporated single-cell RNA sequencing (scRNA-seq) data and corresponding bulk gene expression data from GSE176078. Detailed information about the data sources is provided in Table S1.

Owing to missing matched bulk RNA-seq data, two patients from single-cell datasets were excluded from the analysis. Available clinical indicators were included as ordered categorical variables. Age was included as a continuous variable in the analysis.

2.2. Identification of patients with different ubiquitination characteristics

The ubiquitination genes were prepared for subsequent non-negative matrix factorization (NMF). Before achieving NMF, a filtering procedure was applied. Univariate Cox regression analysis was implemented to examine the relationship between E1s, E2s, and E3s genes and overall survival (OS) as well as OS time in the training dataset [18]. Eventually, genes with a p-value <0.01 were used for patient separation.

The NMF method was done using the R package “NMF” and the “nmf” function, with the nrun set as 50. The cophenetic value was used to calculate the appropriate number of clusters, where the clustering number K was set to 2–10, and changes in the cophenetic values were observed. The “consensusmap” function was used to visualize the consensus matrix to show how well the samples are classified.

Once the optimal number of clusters was determined, we tested whether patient characteristics among the clusters differed significantly. A list of known immune checkpoints was acquired from the ImmPort Portal website (https://www.immport.org/home).

The distinction in prognosis among the patients was tested using Kaplan-Meier (KM) estimates for survival probabilities. The expression of recognized immune checkpoints was investigated and the infiltration of immune cells was analyzed using the CIBERSORT method.

2.3. Classification of ubiquitination-related hub genes and protein expression analysis

After obtaining patient populations with different characteristics, we further analyzed the differentially expressed genes between these patient populations using Mann–Whitney U and fold change (FC) tests. The survival-related ubiquitination genes ($p < 0.01$ and $|FC|>1.5$) were treated as ubiquitination-related hub genes. A volcano plot was constructed to display the screening results for the hub genes.

Next, We investigated the protein expression profiles of the hub genes using immunohistochemistry (IHC) results from the Human Protein Atlas (HPA) database. We collected and presented the expression abundance and spatial localization of these hub genes in the pathological sections of patients of patients of BC.

2.4. Ubiquitination-related prognostic risk score signature

To establish a prognostic risk score (RS) signature related to ubiquitination, we initially evaluated the risk score for every individual in the training dataset using Formula (1) [19].

$$Risk\ score = \sum \beta_{RNA} \times Exp_{RNA} \quad (1)$$

in this formula, β_{RNA} represents the coefficient in the multivariate Cox proportional hazard regression analysis of the ubiquitination-related hub genes, and Exp_{RNA} represents the expression of the hub genes in the RNA expression profile.

After calculating the risk score for each individual in the GSE20685 training set, which originates from an Eastern population, they were classified into high- and low-risk groups based on the median risk score. A scatter plot was then created to visually represent the distribution of risk scores, survival, and death events between these groups. Additionally, a heatmap [20] has been created to visually represent the expression of hub genes in those two different risk groups.

KM estimates for survival probabilities was utilized to assess the disparities in patient survival between two risk groups. Receiver Operating Characteristic (ROC) curves were plotted at the time points of 1, 3, and 5 years to validate the potential of the risk score signature for predicting patient survival.

The ubiquitination-related prognostic risk score signature was validated using six additional datasets from Western populations. Among these six validation sets, three were from European populations, and the remaining were from North American populations. Equation (1) was used to calculate risk score for every individual within each validation set. Individuals were then stratified into two different risk groups within each dataset using the same method. The prognostic value of the risk score signature in each dataset was verified by KM estimates for survival probabilities and ROC analyses.

2.5. Evaluation of the ubiquitination-related prognostic risk score signature

We conducted decision curve analysis (DCA) to compare the predictive abilities of risk score with other clinical indicators. The DCA was performed using the R package "ggDCA" and the function "dca." We calculated the net benefit brought about by the risk score TNM stage, and age, at 1, 3, and 5 years in clinical decision-making for patients. We also used the "AUDC" function in the package to calculate the Area Under the Decision Curve (AUDC) for each indicator, representing the magnitude of the overall net benefit.

Similarly, we analyzed immune cell infiltration and the expression of immune checkpoint genes to explore the differences in immune characteristics between the high and low-risk groups. Gene Ontology (GO) enrichment analysis was performed using ClusterProfiler package (version4.8.3). The analysis was conducted with a significance threshold of p.adj <0.05, and the Benjamini-Hochberg method was used for multiple testing corrections.

2.6. Analysis of single-cell RNA-seq data

The Seurat R package (version 4.3.0) was used for the single-cell RNA-seq data. First, we determined the distribution of mitochondrial and total gene counts. Considering the high heterogeneity of tumor cells, we retained all cells at this stage and did not perform filtering. After data normalization and principal component analysis, the top 20 principal components were included for further analysis. The "FindClusters" function clustered the single-cell data from 24 BC patients with corresponding bulk RNA-seq data at a 0.3 resolution. The UMAP method, applied via the "RunUMAP" function, reduced dimensionality and visualized cell clustering.

The SingleR R package (version 2.0.0) was employed to annotate cell types. A reference dataset of human primary cells was obtained using the Human Primary Cell Atlas Data" function. Then the major cell types were annotated with the "SingleR" function. Subsequently, we extracted the cell clusters annotated as lymphocyte components and used the "Immundata" function to acquire the reference dataset of human immune cells for a more detailed cell annotation.

Matched bulk RNA-seq data were used to obtain the risk score for 24 individuals with BC. Differences in cellular composition between patients with distinct risk score were observed. We carried out parallel analyses to assess the lymphocyte compositions of the tumors.

2.7. Analysis of microbial composition within the tumor microenvironment

Recent research has indicated that the microbiota within tumors may influence the ubiquitination process in host cells; therefore, we utilized the BIC database to explore the microbial composition within the tumor microenvironment validation set of patients.

The BIC database is a newly established database created by Chen et al. in January 2023 that specifically focuses on cancer-associated bacteria. As mentioned in the publication, the database can provide cancer-associated bacterial information, including the relative abundance of bacteria, bacterial diversity, associations with clinical relevance, co-expression networks of bacteria and human genes, and their associated biological functions.

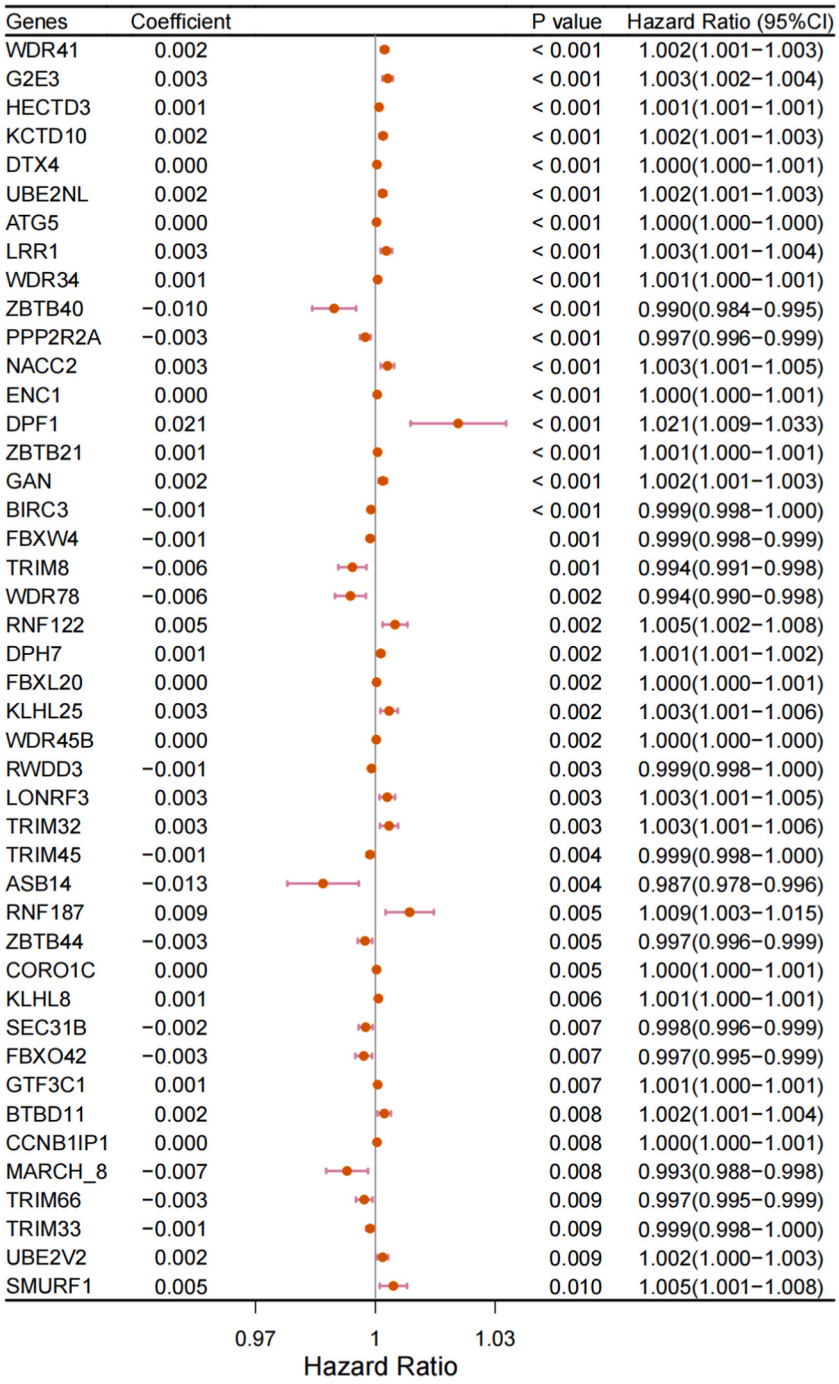


Fig. 2. The results of univariate Cox regression analysis.

In this study, we analyzed the distribution of the 15 most abundant microbial taxa at the tumor tissue genus, family, order, and class levels. After distinguishing between the high- and low-risk groups, we further analyzed the microbial abundance among patients with BC in these categories to determine any significant differences.

2.8. Statistical methods

The major analysis was performed using the R software (version 4.1.2). KM estimates of survival probabilities were utilized to compare survival rates between groups, while ROC curves assessed model accuracy. NMF, DCA, single-cell data, and microbial composition analysis methods have been described previously [21].

All statistical tests were two-sided. During the gene analysis phase, $p \leq 0.01$ was the standard, whereas in other analyses, $p \leq 0.05$ was considered statistically significant.

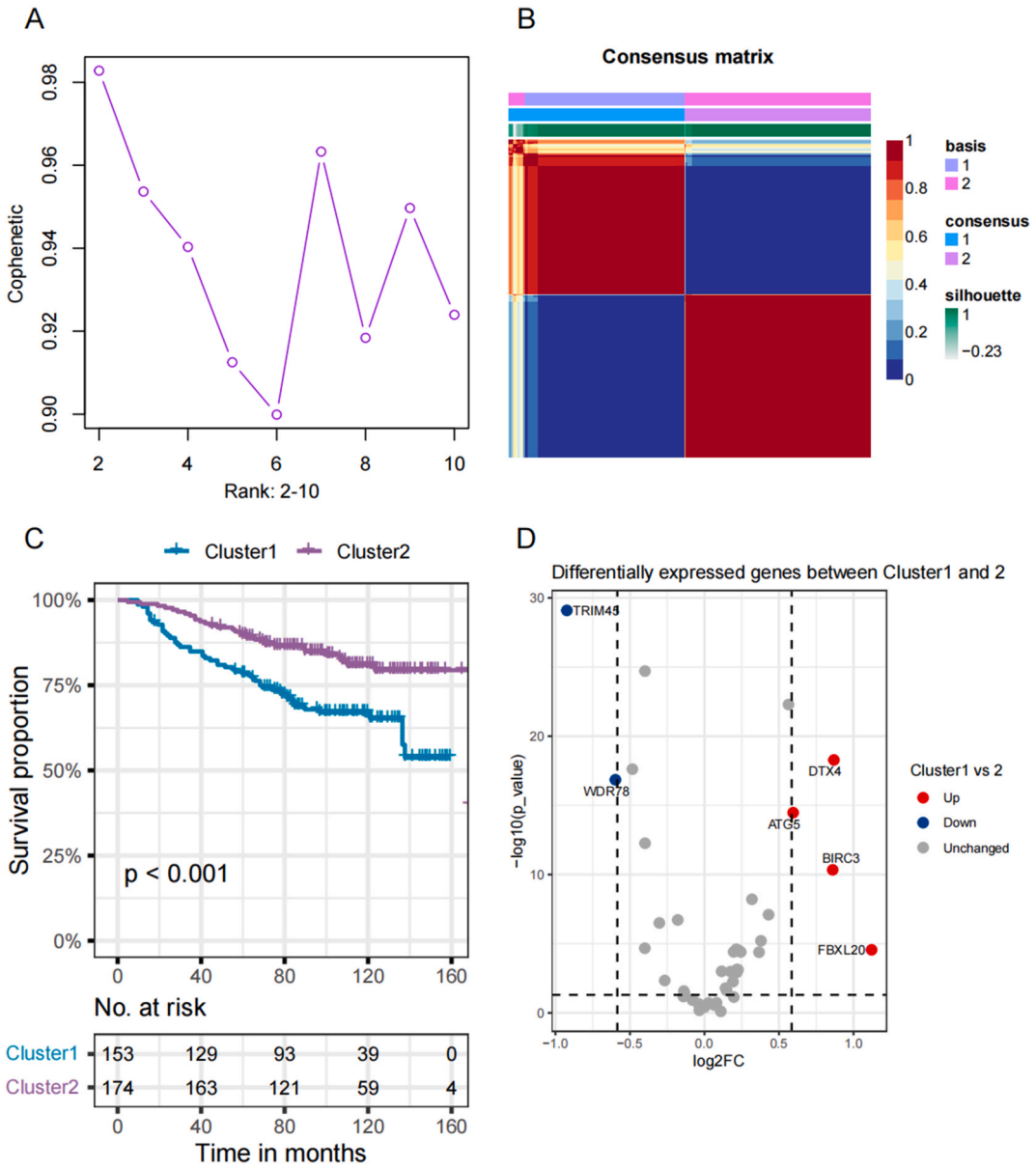


Fig. 3. The results of NMF. (A) Cophenetic correlation coefficient of NMF clustering analysis. (B) Consensus matrix of the two clusters. (C) The K-M plot of the two clusters. (D) The volcano plot of the differentially expressed genes between the two clusters.

3. Results

3.1. Identification of patients with different ubiquitination characteristics

Per the results of univariate Cox regression analysis, 44 prognostic ubiquitination-related genes were identified ($p < 0.01$). Among these, 15 genes showed a negative correlation with patient risk of death, whereas 29 genes showed a positive correlation. The univariate Cox regression analysis results are presented as a forest plot in Fig. 2.

Total 44 genes in the expression matrix of the training set were subjected to NMF analysis. Fig. 3A shows the corresponding cophenetic values for K = 2–10. The highest cophenetic value was obtained at K = 2, with a subsequent decrease in cophenetic value. Therefore, K = 2 was considered the optimal number for classifying patients into two clusters. The consensus matrix also demonstrated good classification performance for the samples at K = 2 (Fig. 3B). Thus, the training set preliminarily identified two distinct clusters of patients with BC with different ubiquitination features.

3.2. Screening for differentially expressed ubiquitination-related hub genes

After identifying two clusters of patients with BC with different ubiquitination features, we first confirmed the differences in other characteristics between these two clusters. KM estimates for survival probabilities revealed a significant difference in survival outcomes (Fig. 3C, $p < 0.001$), whereas CIBERSORT immune cell infiltration and immune checkpoint expression analyses revealed distinct variations in the tumor immune milieu among the clusters of individuals with differing ubiquitination characteristics (Fig. S1).

Subsequently, we performed Mann–Whitney U tests and FC analysis to identify survival-related ubiquitin-related genes that were

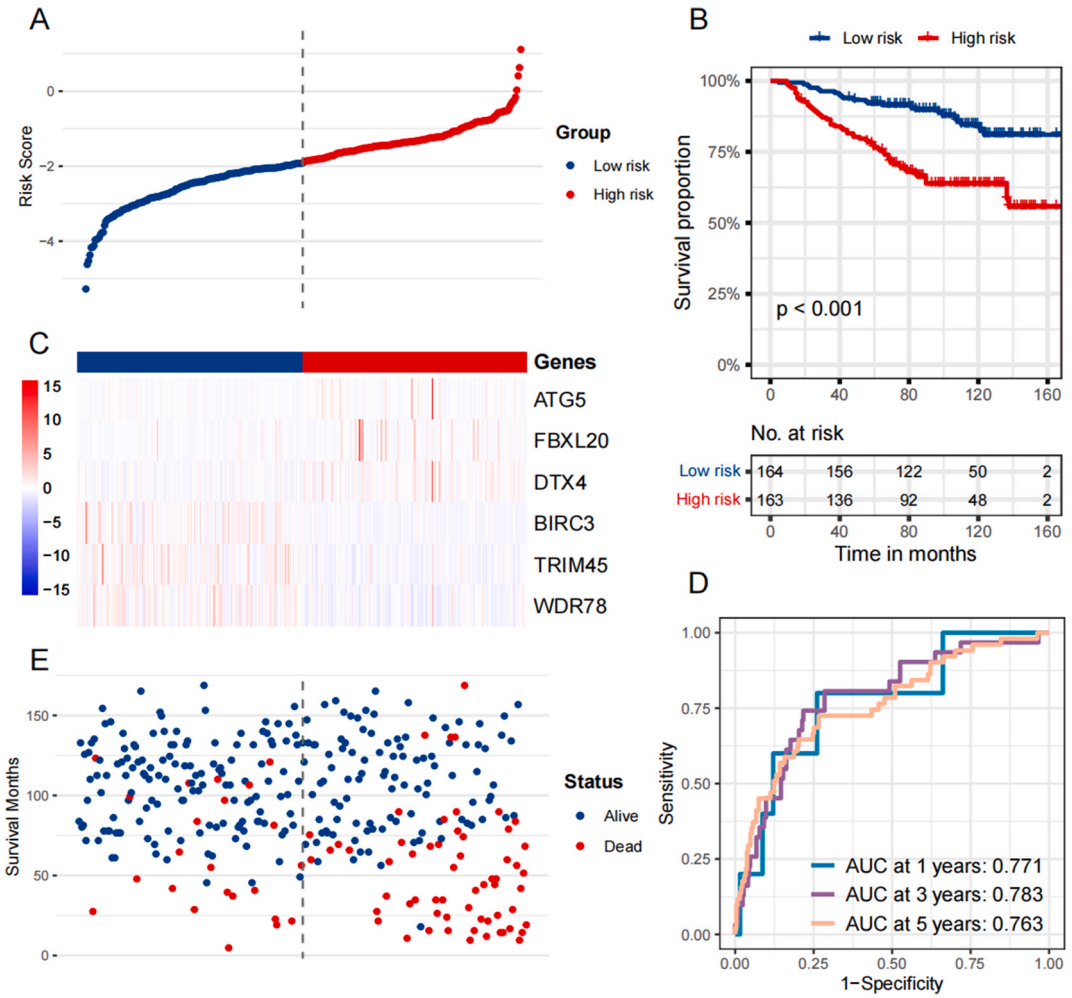


Fig. 4. Grouping of the training dataset. (A) Distribution of risk scores for patients in the high and low risk groups in the training dataset. (B) The K-M plot of the high and low risk groups. (C) The heatmap of the expression of six hub genes. (D) The ROC curves of the high and low risk groups. (E) Distribution of survival for patients in the high and low risk groups.

differentially expressd between the two patient clusters. Six genes: *ATG5, FBXL20, DTX4, BIRC3, TRIM45*, and *WDR78*, were identified. A volcano plot depicting these genes is presented in Fig. 3D, where *ATG5, FBXL20, DTX4,* and *BIRC3* exhibited higher expression levels in cluster1, which had a worse prognosis than the patients in cluster2. Three genes, *ATG5, FBXL20,* and *DTX4*, were positively correlated with patient risk of death in univariate analysis (Fig. 2).

The protein expression profiles of the six genes were obtained from the HPA database. Among them, results for *TRIM45* and *WDR78* did not find any relevant immunohistochemistry analysis. Immunohistochemical analyses of the remaining four genes, including *ATG5, FBXL20, DTX4,* and *BIRC3* are shown in Fig. S2. Immunohistochemical analysis from the HPA database revealed that the protein products of *ATG5, FBXL20,* and *DTX4*, which were significantly associated alongside the risk of death of patients, showed moderate expression levels in the tumor tissue of patients with BC and wereprimarily localized in the cytoplasmic or membranous regions.

3.3. Ubiquitination-related prognostic risk score signature

The risk score for each patient was calculated using the following formula: risk score = $(0.000196 \times \mathrm{Exp}_{ATG5}) + (0.000153 \times \mathrm{Exp}_{FBXL20}) + (0.000308 \times \mathrm{Exp}_{DTX4}) + (-0.00152 \times \mathrm{Exp}_{BIRC3}) + (-0.00122 \times \mathrm{Exp}_{TRIM45}) + (-0.00523 \times \mathrm{Exp}_{WDR78})$. Based on the median value, patients with BC in the training set were further classified into high- and low-risk groups. Fig. 4A shows the distribution of risk scores among all patients in the training set, with the median (lower quartile, upper quartile) range being −1.92 (−2.44, 1.34).

Fig. 4C illustrates the distribution of hub genes in patients from different risk groups. Three genes that were positively associated with the risk of death of patients (*ATG5*, *FBXL20*, and *DTX4*) exhibited higher expression levels in the high-risk group. In contrast, three genes negatively associated with the risk of death of patients (*BIRC3*, *TRIM45*, and *WDR78*) showed higher expression levels within low-risk patients. Fig. 4E depicts outcomes across individuals from the different risk groups.

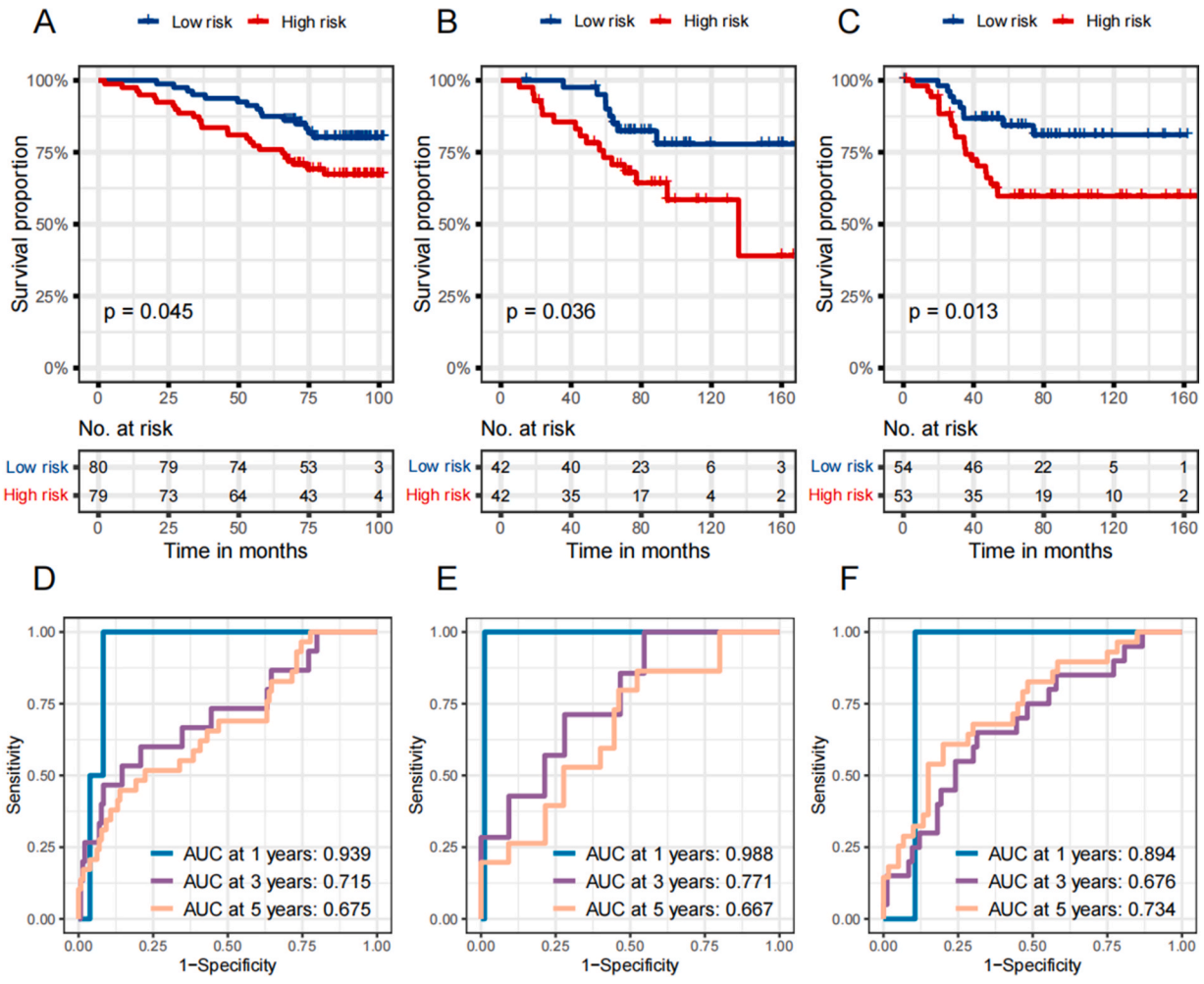


Fig. 5. The results of the validation datasets analysis. (A) The K-M plot of the validation set GSE1456. (B) The K-M plot of the validation set GSE20711. (C) The K-M plot of the validation set GSE58812. (D) The ROC curves of the validation set GSE1456. (E) The ROC curves of the validation set GSE20711. (F) The ROC curves of the validation set GSE58812.

Furthermore, we performed KM estimates for survival probabilities based on the risk score groups and generated receiver operating characteristic (ROC) curves for the risk score of all patients in the training set. In the KM estimates for survival probabilities, the high-risk group demonstrated a significantly lower survival probability than the low-risk group ($p < 0.001$, Fig. 4B). In the ROC curves, risk score exhibited a high AUC at 1, 3, and 5 years ($AUC_{min}>0.750$; Fig. 4D).

Using the same method, the predictive ability of the ubiquitination-related prognostic risk score signature for patient survival in the validation sets consisting of populations from other regions were examined.

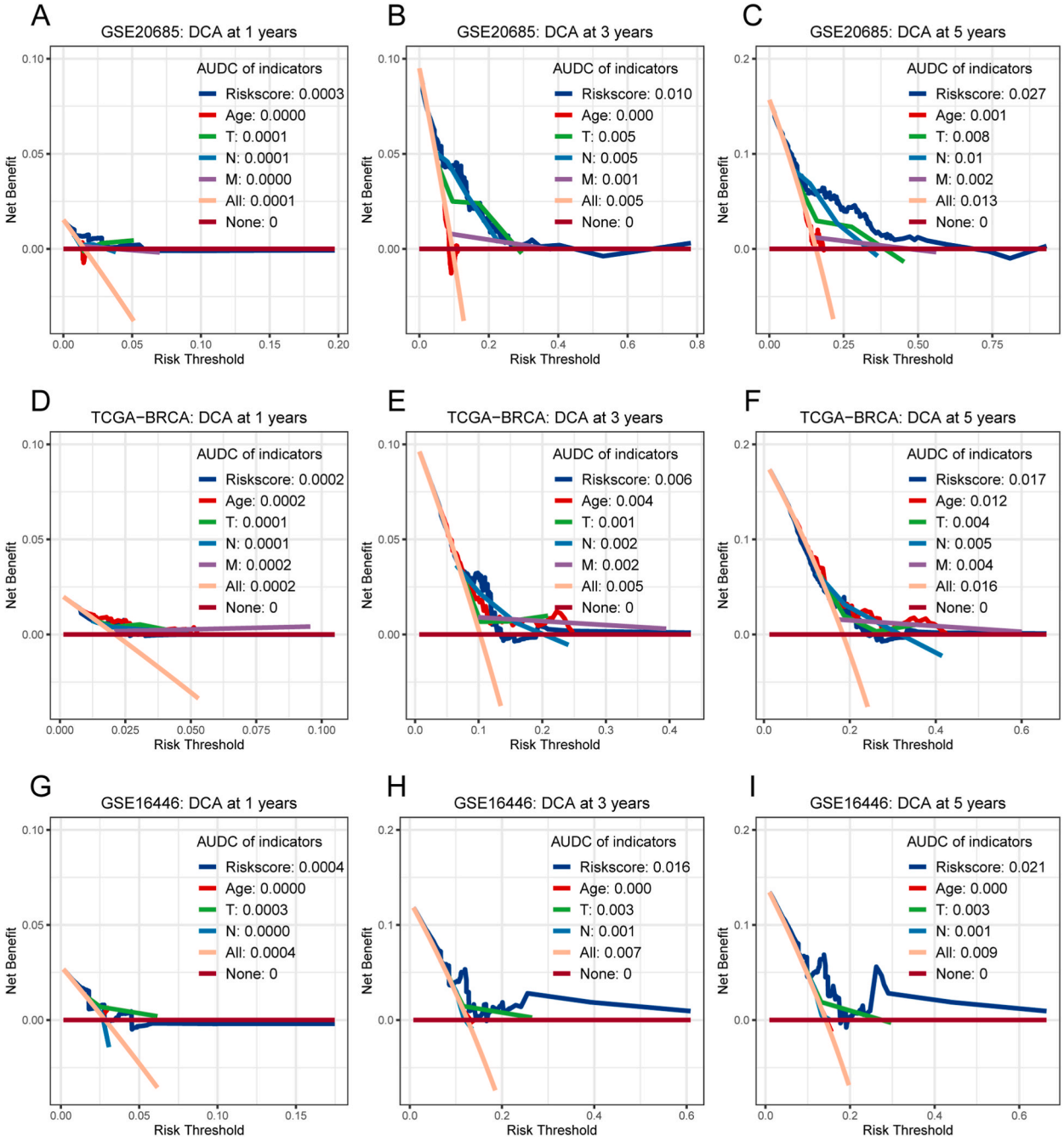


Fig. 6. The DCA plots of three datasets. (A) DCA plots of 1-year survival for patients in the GSE20685 dataset. (B) DCA plots of 3-year survival for patients in the GSE20685 dataset. (C) DCA plots of 5-year survival for patients in the GSE20685 dataset. (D) DCA plots of 1-year survival for patients in the TCGA-BRCA dataset. (E) DCA plots of 3-year survival for patients in the TCGA-BRCA dataset. (F) DCA plots of 5-year survival for patients in the TCGA-BRCA dataset. (G) DCA plots of 1-year survival for patients in the GSE16446 dataset. (H) DCA plots of 3-year survival for patients in the GSE16446 dataset. (I) DCA plots of 5-year survival for patients in the GSE16446 dataset.

The ubiquitination-related prognostic risk score signature performed well in the validation sets GSE1456, GSE20711, and GSE58812 (Fig. 5). In the KM estimates for survival probabilities, survival rates varied significantly between the patients with high- and low-risk BC (Fig. 5A–C; GSE1456: $p = 0.045$, GSE20711: $p = 0.036$, and GSE58812: $p = 0.013$). In the ROC curves of the three datasets, risk score achieved a relatively high AUC close to 0.67 at 1, 3, and 5 years (Fig. 5D–F), indicating the good predictive ability of the ubiquitination-related prognostic risk score signature for patient prognosis.

In the additional validation sets TCGA-BRCA, GSE16446, and GSE96058, the performance of the ubiquitination-related prognostic risk score signature may have been better. However, good discrimination of patient survival was observed in KM estimates for survival probabilities (Figs. S3A–S3C; TCGA-BRCA: $p < 0.001$, GSE16446: $p = 0.045$, GSE96058: $p < 0.001$). In the ROC analysis, risk score showed slightly lower performance at some time points; however overall, risk score still had a certain predictive value (Figs. S3D–S3F).

3.4. Evaluation of the predictive performance of the ubiquitination-related prognostic risk score signature

We used DCA to evaluate the risk score signature along withvarious clinical markers for patient benefit in decision-making. In the datasets used in this study, datasets (GSE20685, TCGA-BRCA, and GSE16446) provided comprehensive clinical indicator information that could be used for this comparison.

The results demonstrated that in the training set GSE20685, the AUDC values of risk score were one, three, and the data showed that over a span of five years, risk score had a greater impact on patients than other clinical indications, demonstrating its ability to

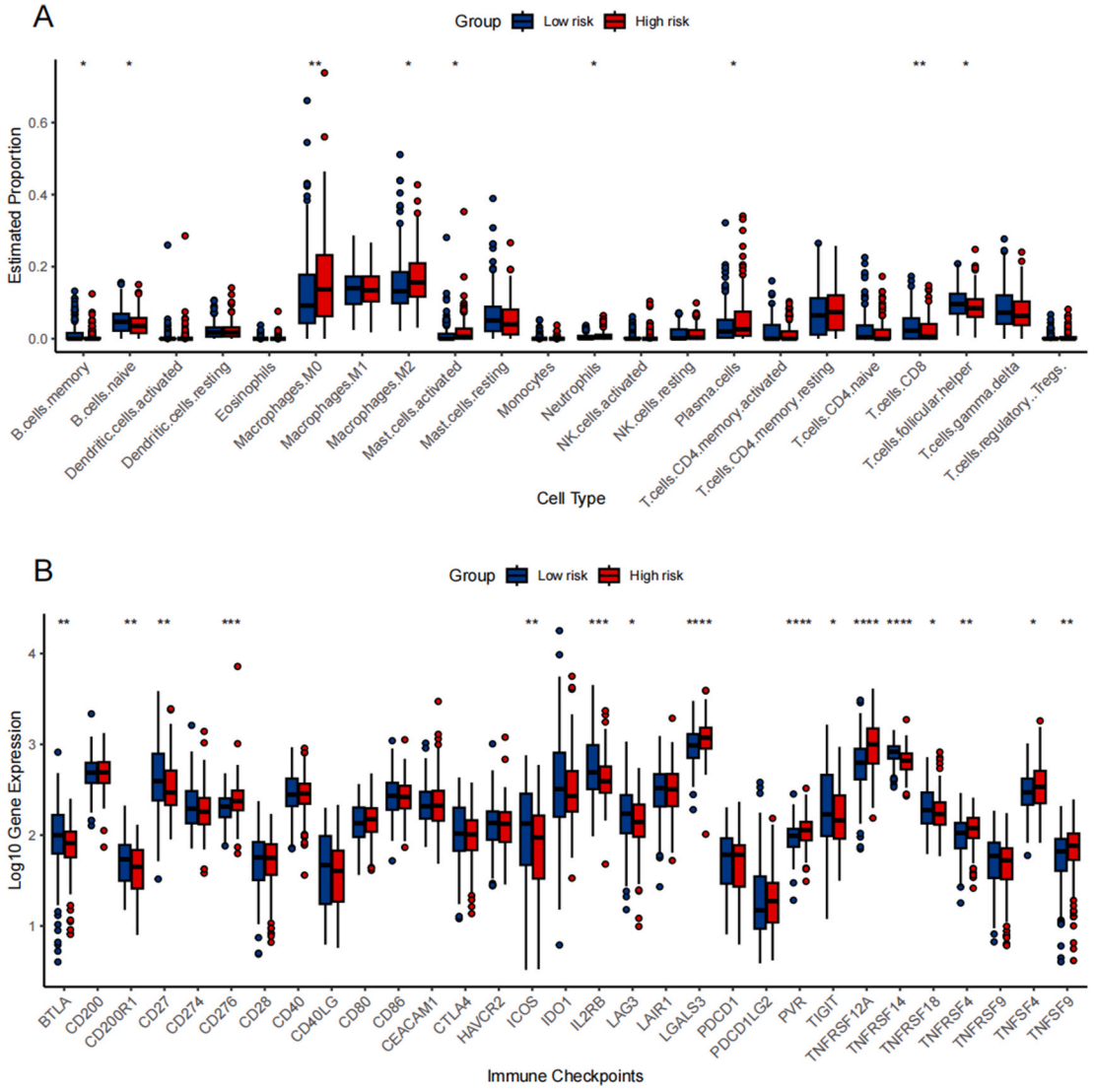


Fig. 7. The results of TME analysis. (A) the immune cell infiltration in the high-risk and low-risk groups of patients in the training set. (B) The expression of immune checkpoint-related genes in the high-risk and low-risk groups of patients in the training set.

provide the greatest overall benefit (Fig. 6A–C). In the TCGA-BRCA validation set, although the net benefit of risk score at one year was comparable to that of age, it was higher than that of other clinical indicators at three and five years (Fig. 6D–F). In the validation set GSE16446, the AUDC values of risk score at one, three, and five years were all higher than those of the other clinical indicators, indicating the highest net benefit it could bring to patients (Fig. 6G–I).

Overall, the DCA results from the training and validation sets indicated that our ubiquitination-related prognostic risk score

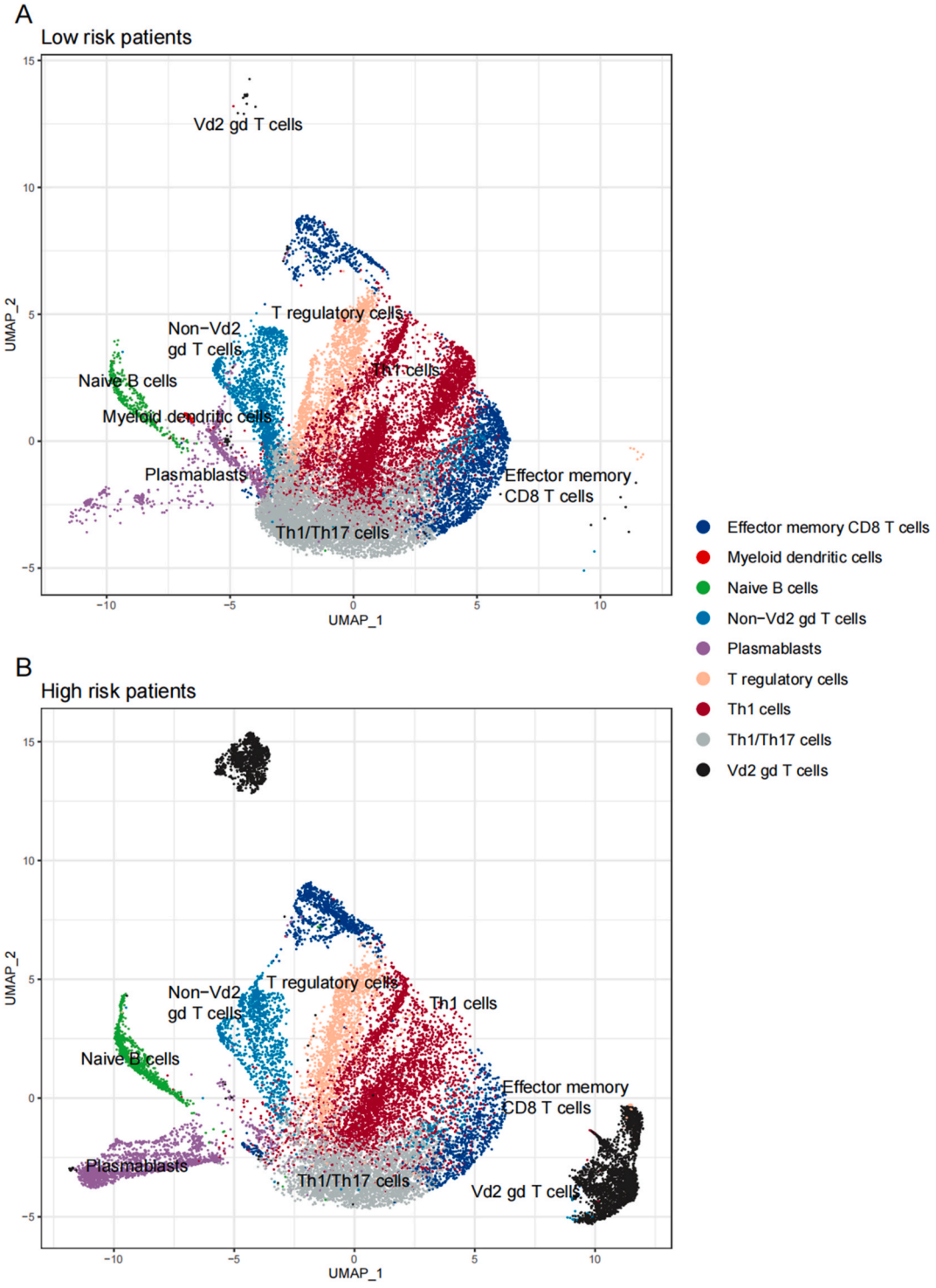


Fig. 8. The annotation results for lymphocyte of (A) low-risk group and (B) high-risk group.

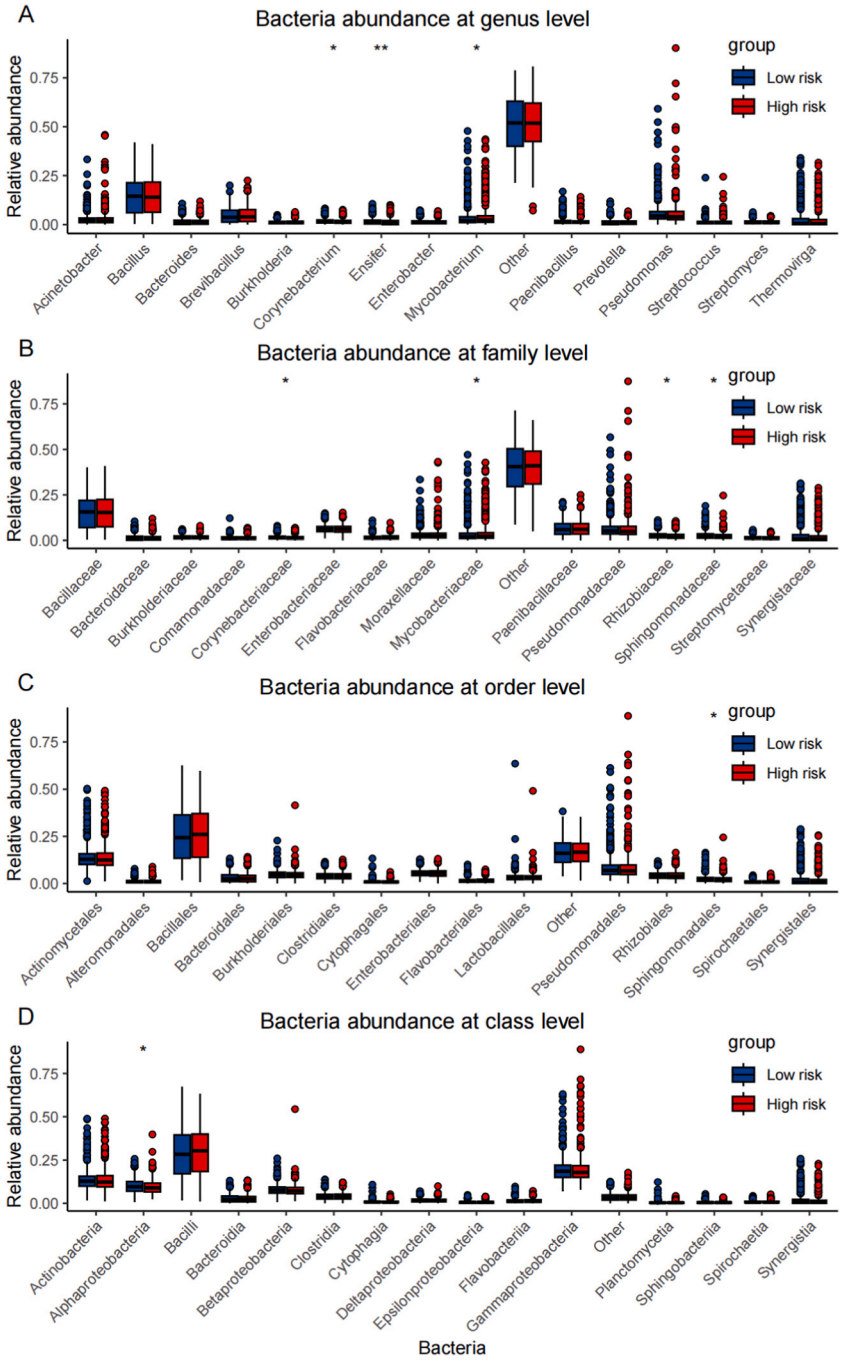


Fig. 9. (A) The distribution of the top 15 most abundant microbial taxa at the genus levels in the tumor tissue of high-risk and low-risk groups of patients in the validation set TCGA-BRCA. (B) The distribution of the top 15 most abundant microbial taxa at the family levels in the tumor tissue of high-risk and low-risk groups of patients in the validation set TCGA-BRCA. (C) The distribution of the top 15 most abundant microbial taxa at the order levels in the tumor tissue of high-risk and low-risk groups of patients in the validation set TCGA-BRCA. (D) The distribution of the top 15 most abundant microbial taxa at the class levels in the tumor tissue of high-risk and low-risk groups of patients in the validation set TCGA-BRCA.

signature provided a higher net benefit to patients than other clinical indicators when used in clinical decision-making. The findings from the GO analysis are presented in Table S2.

3.5. Tumor immune microenvironment and single-cell data analyses

CIBERSORT were employed for estimate immune cell infiltration in patients with BC in the training set. We compared immune cell infiltration and the expression of immune checkpoint-related genes between the high- and low-risk groups. The results are shown in Fig. 7.

Fig. 7A illustrates immune cell infiltration in different risk groups of patients in the training set. The infiltration levels of M0 and M2 macrophages, activated mast cells, neutrophils, and plasma cells were significantly higher in the high-risk group than that in the low-risk group. Infiltration levels of memory B cells, naïve B cells, CD8 + T cells, and follicular helper T cells were considerably greater in the low-risk group.

Fig. 7B illustrates the expression of immune checkpoint-related genes in high- and low-risk patient groups in the training set. The expression levels of *CD276*, *LGALS3*, *PVR*, *TNFRSF12A*, *TNFRSF4*, *TNFSF4*, and in the group at high risk, *TNFSF9* expression levels were greater. Simultaneously, *BTLA*, *CD200R1*, *CD27*, *ICOS*, *IL2RB*, *LAG3*, *TIGIT*, *TNFRSF14*, and *TNFRSF18* showed higher expression levels in the low-risk group.

We further compared the cell types and distributions in tumors from patients with different risks in the single-cell dataset, GSE176068, to better understand the immune microenvironment status in high- and low-risk groups. We directly calculated the risk score of the patients and classified them into risk groups using bulk RNA-seq data.

After excluding two patients without matched data, we annotated the cell types in the single-cell RNA-seq data from 24 patients with BC. The cell annotation results for the high- and low-risk groups (Fig. S4). By comparing cell types in those two risk groups, we observed a lack of macrophages and mesenchymal stem cells (MSC) in the tumor tissues of patients with high-risk.

Next, we selected lymphocyte subsets, including B and T cells, from previously annotated cells for lymphocyte subgroup annotation analysis. The annotation results for the lymphocyte subgroups are presented in Fig. 8. At this stage, we observed that only a small population of cells annotated as Vd2 gd T cells was present in the low-risk group (Fig. 8A). Conversely, the high-risk group lacked the cells annotated as myeloid dendritic cells (Fig. 8B). These results suggest differences in the tumor immune microenvironment between patients with different risks.

3.6. Tumor microbial composition analysis

Using the BIC database, we analyzed the distribution of the top 15 most abundant microbial taxa at the genus, family, order, and class levels in the tumor tissues of patients in TCGA-BRCA validation set (Fig. 9). At each level, microbial taxa belonging to the genera Bacillus, family Bacillaceae, order Bacillales, and class Bacilli were the most abundant.

At the genus level, significant differences in abundance were found between the high- and low-risk groups for microbial taxa, such as Corynebacterium, Ensifer, and Mycobacterium (Fig. 9A). At the family level, significant differences in abundance were observed between the high- and low-risk groups for microbial taxa such as Corynebacteriaceae, Mycobacteriaceae, Rhizoblaceae, and Sphingomonadaceae (Fig. 9B). At the order level, only the abundance of Sphingomonadales differed significantly among those two risk groups (Fig. 9C). At class level, only Alphaproteobacteria differed significantly between the groups (Fig. 9D).

4. Discussion

This study established a six-gene risk score signature encompassing *ATG5*, *FBXL20*, *DTX4*, *BIRC3*, *TRIM45*, and *WDR78*, which effectively grouped patients with BC according to their prognosis and was more reliable than traditional clinical indicators. The risk score signature showed good predictive ability in different datasets, indicating that this risk score characteristic has an opportunity to function as a prognostic factor in patients with BC. In contrast to previous studies that aimed to discover possible indicators for predicting the future outcome of individuals with BC using genes associated with ubiquitination, the risk score signature in this investigation demonstrated superior predictive outcomes to a certain degree. Jun et al. used E2-related genes to construct a model for predicting DFS in patients with BC, which possessed a good predictive effect with an AUC of 0.870, hoever, was not validated in an external dataset [22]. Zhang et al. constructed a prediction model containing four ubiquitination-related genes and demonstrated good prediction ability in the training dataset. However, the AUC in the test dataset was only 0.659 [23], which is lower than the results of this study.

The ubiquitin-proteasome system (UPS) is the predominant pathway for intracellular protein degradation and is essential for regulating normal and cancer-related cellular processes. Thus, defects in the UPS may lead to aberrant protein expression, interactions, and cellular localization [24–26]. The UPS has been shown to target proteins, including a wide range of oncogenic gene products and tumor suppressors, which are associated with various diseases, such as cancer [27]. Six E3-related genes obtained from the screening comprised the risk score signature, and all E1-and E2-related genes were screened. The ubiquitination process is executed by the concurrent activity of these enzymes and E3 serves a crucial function in recognizing substrates [28,29]. E3 ubiquitin ligases inhibit tumorigenesis and tumor progression by targeting and degrading tumor promoters or suppressors in malignant tumors [30–32]. However, mutations in E3 ubiquitin ligases result in dysregulated protein hydrolysis, which impairs the degradation of oncogenes and tumor suppressor genes [33]. Therefore, in this study, we started with E1-, E2-, and E3-related ubiquitination genes and obtained six E3-related genes by univariate Cox regression analysis of GSE20685 public RNA-Seq data as well as NMF screening, The risk score

signature consists of a set of characteristics that may accurately predict the prognosis of patients across various datasets. Notable disparities in survival were observed among the grouped individuals in each of those seven datasets (Figs. 4A and. 5A–C, and Figs. S2A–C). Of these six genes, *FBXL20* is assumed to be involved in regulating the activity of various proteins, leading to the suppression of drug effects, thereby promoting BC tumor growth [34]. Liu et al. suggested that overexpression of *ATG5* promotes prostate cancer development while silencing *ATG5* acts as an inhibitor [35]. *DTX4* has been shown to be correlated with cancers, such as colorectal cancer, liver cancer, and melanoma, where it plays a pro-cancer role through interactions with carcinogens [36–38]. *BIRC3* combines with CXCL13 as a marker for immune checkpoint therapy in patients with BC [39]. Li et al. found that *TRIM45* was more lowly expressed in tumor tissues in breast cancer [40]. A study by Liang et al. found that *WDR78* expression was significantly elevated in patients with early-stage cancer compared to those with advanced-stage cancer, and its downregulation was associated with tumor progression. [41], which is also consistent with the results of this study. In summary, all six hub genes were previously linked to different types of cancer, which further supports the predictive capability of the risk score signature for forecasting outcomes in BC patients.

We explored the disparities in the immune microenvironment among patients categorized into different risk categories. In the immune infiltration analysis (Fig. 7A), macrophages, which have a crucial part in tumor immunity, were found to have infiltrated much more within high risk patients Research has indicated a negative correlation between a large influx of macrophages and decreased patient survival [42]. B-cell infiltration was also considerably greater in low risk patients. This finding is widely linked to a positive outcome in those with cancer [42]. In the immune checkpoint analysis (Fig. 7B), low-risk individuals exhibited notably elevated expression levels of crucial tumor immune checkpoints, including *TIGIT*, *BTLA*, *CD27*, and *CD276*. These findings indicate that low risk individuals may experience better outcome when treated with immunotherapy.

Conversely, *TNFSF4* and *TNFRSF4* were considerably increased in the high-risk group. Together, *TNFSF4* and *TNFRSF4* promote T-cell proliferation and are involved in cytokine production. *TNFSF4* overexpression has been observed in a wide range of BCs and a positive correlation has been discovered between shortened survival and high TNFSF4 expression [43]. In the single-cell analysis in the training dataset (Fig. 8), high-risk patients exhibited greater ratio of Vd2 gd T cells, which belong to one of the subsets of γδ T cells. The γδ T cells are assumed to contribute to tumor progression in BC, resulting in an unfavorable outcomes for patients [44]. Substantial variations were observed in Corynebacterium when comparing microbial taxa between the two groups of patients. (Fig. 9A). Corynebacterium exhibits high signal intensity in patients with both triple-positive and Her2 negative BC, suggesting that the risk score signature can effectively predict patient prognosis.

Despite the promising results, this our research has several drawbacks. The datasets utilized had been predominantly sourced from publicly available repositories, which may have biases and may not completely represent the broader patient population with BC. Therefore, validation using more diverse and independent datasets is required. Potential confounding factors that were not accounted for may have influenced the risk score signature. Focusing solely on ubiquitination-related genes may overlook other crucial factors that affect BC prognosis. Finally, the findings on immune microenvironment and tumor microbiology differences require further experimental validation to confirm their clinical relevance Recent research in the field of cellular memory has highlighted the intricate mechanisms by which cells retain information about past events, influencing future gene expression and interactions. Key findings suggest that epigenetic modifications are essential for regulating gene over time. Additionally, the stability of gene-gene interactions is governed by robust regulatory networks and feedback loops, that can sustain these interactions over varying periods, from transient to long-term [45,46]. This understanding of cellular memory underscores the complexity of cellular regulation and the potential for long-lasting effects on cell behavior [47], highlighting the need for further research.

5. Conclusion

This study established a risk score signature consisting of six ubiquitination genes (*ATG5*, *FBXL20*, *DTX4*, *BIRC3*, *TRIM45*, and *WDR78*) that can efficiently estimate the future outcomes of patients with BC in multiple datasets. It can provide personalized and targeted assistance for the diagnosis and treatment of patients with BC.

Funding

This study did not receive any funding.

Data availability statement

The data used to support the findings of this study are available at UCSC (https://xenabrowser.net/datapages/) and GEO (http://www.ncbi.nlm.nih.gov/geo/) databases. The name of the repository and the accession number: GSE20685 Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/); GSE176068 Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/); GSE1456 Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/); GSE16446 Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/); GSE20685 Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/); GSE20711 Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/); GSE58812Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/); GSE96058 Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/); TCGA-BRCA University of California Santa Cruz (UCSC) database (https://xenabrowser.net/datapages/); GSE176078 Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/).

Ethical statement

The present study was conducted based on publicly available datasets. No human or animal subjects were involved in this study.

CRediT authorship contribution statement

Kexin Feng: Writing – original draft, Visualization, Validation, Project administration, Methodology, Conceptualization. **Xin He:** Writing – original draft, Visualization, Validation, Methodology, Data curation. **Ling Qin:** Writing – original draft, Visualization, Methodology, Formal analysis. **Zihuan Ma:** Writing – original draft, Visualization, Validation, Resources, Formal analysis. **Siyao Liu:** Writing – original draft, Visualization, Software, Funding acquisition, Data curation. **Ziqi Jia:** Writing – review & editing, Software, Resources, Investigation, Formal analysis. **Fei Ren:** Writing – review & editing, Software, Investigation. **Heng Cao:** Writing – review & editing, Software, Investigation. **Jiang Wu:** Writing – review & editing, Resources, Methodology. **Dongxu Ma:** Writing – review & editing, Visualization, Resources. **Xiang Wang:** Writing – review & editing, Project administration, Conceptualization. **Zeyu Xing:** Writing – original draft, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e35553.

References

[1] J.O. Jin, et al., The ubiquitin system: an emerging therapeutic target for lung cancer, Int. J. Mol. Sci. 22 (17) (2021).
[2] C. Pohl, I. Dikic, Cellular quality control by the ubiquitin-proteasome system and autophagy, Science 366 (6467) (2019) 818–822.
[3] G. Cetin, et al., The ubiquitin-proteasome system in immune cells, Biomolecules 11 (1) (2021).
[4] T. Sun, Z. Liu, Q. Yang, The role of ubiquitination and deubiquitination in cancer metabolism, Mol. Cancer 19 (1) (2020) 146.
[5] A.M. Antao, et al., Advances in deubiquitinating enzyme inhibition and applications in cancer therapeutics, Cancers 12 (6) (2020).
[6] J. Park, J. Cho, E.J. Song, Ubiquitin-proteasome system (UPS) as a target for anticancer treatment, Arch Pharm. Res. (Seoul) 43 (11) (2020) 1144–1161.
[7] N. Harbeck, M. Gnant, Breast cancer, Lancet 389 (10074) (2017) 1134–1150.
[8] H. Sung, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA A Cancer J. Clin. 71 (3) (2021) 209–249.
[9] A.F. Dibha, et al., Utilization of secondary metabolites in algae kappaphycus alvarezii as a breast cancer drug with a computational method, Phcog. J. 14 (3) (2022).
[10] L. Fan, et al., Breast cancer in China, Lancet Oncol. 15 (7) (2014) e279–e289.
[11] M. Herdiansyah, et al., in: Silico Study of Cladosporol and its Acyl Derivatives as Anti-breast Cancer against Alpha-Estrogen Receptor, vol. 16, 2024, pp. 142–154.
[12] P.E. Cockram, et al., Ubiquitination in the regulation of inflammatory cell death and cancer, Cell Death Differ. 28 (2) (2021) 591–605.
[13] S. Han, et al., The role of ubiquitination and deubiquitination in tumor invasion and metastasis, Int. J. Biol. Sci. 18 (6) (2022) 2292–2303.
[14] L.D. Cervia, et al., A ubiquitination cascade regulating the integrated stress response and survival in carcinomas, Cancer Discov. 13 (3) (2023) 766–795.
[15] X. Li, et al., CUL3 (cullin 3)-mediated ubiquitination and degradation of BECN1 (beclin 1) inhibit autophagy and promote tumor progression, Autophagy 17 (12) (2021) 4323–4340.
[16] E.G. Otten, et al., Ubiquitylation of lipopolysaccharide by RNF213 during bacterial infection, Nature 594 (7861) (2021) 111–116.
[17] M. Shariq, et al., The exploitation of host autophagy and ubiquitin machinery by Mycobacterium tuberculosis in shaping immune responses and host defense during infection, Autophagy 19 (1) (2023) 3–23.
[18] T. Emura, S. Matsui, H.Y. Chen, compound.Cox: univariate feature selection and compound covariate for predicting survival, Comput. Methods Progr. Biomed. 168 (2019) 21–37.
[19] C.T. Yeh, G.Y. Liao, T. Emura, Sensitivity analysis for survival prognostic prediction with gene selection: a copula method for dependent censoring, Biomedicines 11 (3) (2023).
[20] M.B. Eisen, et al., Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. U. S. A. 95 (25) (1998) 14863–14868.
[21] S.Y. Wang, et al., Construction and validation of a prognostic prediction model for gastric cancer using a series of genes related to lactate metabolism, Heliyon 9 (5) (2023) e16157.
[22] J. Shen, et al., Validation of a disease-free survival prediction model using UBE2C and clinical indicators in breast cancer patients, Breast Cancer 15 (2023) 295–310.
[23] Y. Zheng, et al., Identification of a novel ubiquitination related gene signature for patients with breast cancer, Medicine (Baltim.) 101 (37) (2022) e30598.
[24] A. Ciechanover, Y.T. Kwon, Degradation of misfolded proteins in neurodegenerative diseases: therapeutic targets and strategies, Exp. Mol. Med. 47 (3) (2015) e147.
[25] C.M. Pickart, Mechanisms underlying ubiquitination, Annu. Rev. Biochem. 70 (2001) 503–533.
[26] D. Kornitzer, A. Ciechanover, Modes of regulation of ubiquitin-mediated protein degradation, J. Cell. Physiol. 182 (1) (2000) 1–11.
[27] V.M. Díaz, A.G. de Herreros, F-box proteins: keeping the epithelial-to-mesenchymal transition (EMT) in check, Semin. Cancer Biol. 36 (2016) 71–79.
[28] C.E. Berndsen, C. Wolberger, New insights into ubiquitin E3 ligase mechanism, Nat. Struct. Mol. Biol. 21 (4) (2014) 301–307.
[29] J.S. Brown, S.P. Jackson, Ubiquitylation, neddylation and the DNA damage response, Open Biol 5 (4) (2015) 150018.
[30] D. Wang, et al., E3 ubiquitin ligases in cancer and implications for therapies, Cancer Metastasis Rev. 36 (4) (2017) 683–702.
[31] Z. Cai, et al., The Skp2 pathway: a critical target for cancer therapy, Semin. Cancer Biol. 67 (Pt 2) (2020) 16–33.
[32] J. Cheng, et al., The emerging role for Cullin 4 family of E3 ligases in tumorigenesis, Biochim. Biophys. Acta Rev. Canc 1871 (1) (2019) 138–159.

[33] D. Senft, J. Qi, Z.A. Ronai, Ubiquitin ligases in oncogenic transformation and cancer therapy, Nat. Rev. Cancer 18 (2) (2018) 69–88.
[34] R.K. Manne, et al., FBXL20 promotes breast cancer malignancy by inhibiting apoptosis through degradation of PUMA and BAX, J. Biol. Chem. 297 (4) (2021) 101253.
[35] X. Liu, et al., Cancer-associated fibroblasts promote malignant phenotypes of prostate cancer cells via autophagy : cancer-associated fibroblasts promote prostate cancer development, Apoptosis 28 (5–6) (2023) 881–891.
[36] X. Lin, et al., Notch4+ cancer stem-like cells promote the metastatic and invasive ability of melanoma, Cancer Sci. 107 (8) (2016) 1079–1091.
[37] W.M. Liu, et al., A microarray study of altered gene expression in colorectal cancer cells after treatment with immunomodulatory drugs: differences in action in vivo and in vitro, Mol. Biol. Rep. 37 (4) (2010) 1801–1814.
[38] P. Viatour, et al., Notch signaling inhibits hepatocellular carcinoma following inactivation of the RB pathway, J. Exp. Med. 208 (10) (2011) 1963–1976.
[39] Z.A. Xia, et al., The expression profiles of signature genes from CD103(+)LAG3(+) tumour-infiltrating lymphocyte subsets predict breast cancer survival, BMC Med. 21 (1) (2023) 268.
[40] J. Li, et al., Cuproptosis/ferroptosis-related gene signature is correlated with immune infiltration and predict the prognosis for patients with breast cancer, Front. Pharmacol. 14 (2023) 1192434.
[41] W. Liang, et al., Microarray and bioinformatic analysis reveal the parental genes of m6A modified circRNAs as novel prognostic signatures in colorectal cancer, Front. Oncol. 12 (2022) 939790.
[42] Y. Ino, et al., Immune cell infiltration as an indicator of the immune microenvironment of pancreatic cancer, Br. J. Cancer 108 (4) (2013) 914–923.
[43] K. Li, et al., The immunotherapy candidate TNFSF4 may help the induction of a promising immunological response in breast carcinomas, Sci. Rep. 11 (1) (2021) 18587.
[44] F. Nezhad Shamohammadi, et al., Controversial role of γδ T cells in pancreatic cancer, Int. Immunopharm. 108 (2022) 108895.
[45] Z. Malek-Esfandiari, et al., Molecular dynamics and multi-spectroscopic of the interaction behavior between bladder cancer cells and calf thymus DNA with rebeccamycin: apoptosis through the down regulation of PI3K/AKT signaling pathway, J. Fluoresc. 33 (4) (2023) 1537–1557.
[46] T. Zohoorian-Abootorabi, et al., Separate and simultaneous binding effects through a non-cooperative behavior between cyclophosphamide hydrochloride and fluoxymesterone upon interaction with human serum albumin: multi-spectroscopic and molecular modeling approaches, Spectrochim. Acta Mol. Biomol. Spectrosc. 88 (2012) 177–191.
[47] M. Kaffash, et al., Spectroscopy and molecular simulation on the interaction of Nano-Kaempferol prepared by oil-in-water with two carrier proteins: an investigation of protein–protein interaction, Spectrochim. Acta Mol. Biomol. Spectrosc. 309 (2024) 123815.