A six-mRNA signature-based model for the prognosis prediction of breast cancer

Yan Huang (✉ huangyan86789@126.com)

  The People's Hospital of Nanchuan, Chongqing Medical University

Xia Wan

  The People's Hospital of Nanchuan, Chongqing Medical University.

Shixiang Wang

  The People's Hospital of Nanchuan, Chongqing Medical University.ngqing Medical of

Research

Abstract

Background

Due to the lack of predictors, high mortality and poor prognosis have always been a serious threat to the breast cancer (BC) patients. Accumulating studies have shown that molecular markers which affect the tumor microenvironment (TME) play an important role in the development of cancer. Here, we aim to use machine learning to identify a new prognostic gene signature and independent prognostic factors related to BC survival through comprehensive bioinformatics analysis.

Results

This 6-mRNA signature-based model is not only an important indicator to predict the prognosis and survival of BC, but also a potential indicator to monitor the clinical therapeutic effect with certain clinical significance.

Conclusions

A new 6-gene prognostic signature was discovered and it is a promising independent predictor. The 6 feature genes can function as important biomarkers for BC clinical treatment. At the same time, our study also provides a new insight to explore the molecular mechanism of TME and immune cells that influence BC progress.

1. Background

According to the latest statistics from *A Cancer Journal for Clinicians*, the three most common malignant tumors among women are breast cancer (BC), uterine cancer and colorectal cancer[1]. BC as a complex heterogeneous disease is considered to be the main cause of cancer death, and about 30% of BC is heritable[2-4]. Although systematic treatment of BC has made great progress, the increasing incidence and the high specific mortality caused by tumor recurrence and metastasis still pose a serious threat to the life health of women worldwide[5,6]. The invasion and metastasis of BC can be influenced by tumor microenvironment (TME)[7]. The matrix containing fibroblasts and immune cells in TME is often altered as the tumor progresses, and the molecular markers involved in this process are also important[8]. It is a good starting point to study the interaction between BC and these TME-related molecular markers.

In recent years, many studies have shown that prognostic markers can predict possible outcomes associated with cancer progression, recurrence, or death[9]. At the same time, multi-gene signatures for prognosis prediction have also been studied, providing a reliable prognostic tool for individualized treatment of BC patients[10,11]. Renhua Li *et al.* have identified a signature-based model consisting of 16 master regulatory genes that can be used to identify the molecular subtypes of BC and is significantly associated with the prognosis and survival rate of BC[12]. Therefore, construction of effective prognostic gene signatures will bring more hope for BC patients.

In this study, immune and stromal cells of BC samples were scored and used to analyze the influence of TME on the prognosis of BC. Then, differentially expressed genes (DEGs) were screened out. DEGs is one of the research hotspots of BC. Maria Libera Ascierto *et al.* have identified that CD96, CD1d, CRTAM, LFA-1 and DNAM1 as DEGs participate in the interaction between natural killer cells and tumor cells, and their expression is increased in BC patients with good prognosis[13]. Bao *et al.* suggest that the interaction pathway of DEGs and extracellular matrix-receptor (ECM-receptor) may perform an important role in BC[14]. Studies have also discovered that some BC-related DEGs can be used as prognostic biomarkers. For example, Rodrigues-Ferreira *et al.* have found that EB1 and ATIP 3 can be used as biomarkers to improve the prognosis of BC[15]. These studies suggest that it is feasible to mine TME-related DEGs as prognostic markers of BC to improve the survival status.

Recent advances in bioinformatics and the availability of RNA-seq transcriptome data in public databases have contributed a lot to mine biomarkers for diagnosis and prognosis[16]. Machine learning analysis can be used for survival calculation, risk classification and prognosis prediction of a variety of cancers including BC[17,18]. With the application of machine learning analysis, the accuracy of tumor outcome prediction has increased by 15%~20%[19]. In this study, we used machine learning analysis, including univariate Cox regression, LASSO regression and multivariate Cox regression analysis, to construct a gene signature-based model for prognosis prediction of BC patients based on TME-associated DEGs. 6 feature genes related to tumor immunity were identified and evaluated for their value as prognostic indicators. Our findings provide a new prognostic gene signature and independent prognostic factors for BC, which have certain clinical significance.

2. Results

2.1. Screening of TME-related DEGs in BC

Based on the FPKM data of TCGA-BRCA mRNA, the ESTIMATE prediction software was used to score the immune and stromal cells in each sample (Table S1). Survival analysis was performed based on the immune score and stromal score combined with the clinical information of downloaded TCGA-BRCA samples (Table S2). It was found that the survival time of patients with high immune cell infiltration was significantly longer than that of patients with low immune cell infiltration, indicating that the infiltration degree of immune cells remarkably affected the prognosis of BC, while the infiltration degree of stromal cells had no significant impact on the prognosis (Fig. 1A). Thereafter, all patients were grouped into high-score and low-score groups according to the median value of the immune score. Wilcox test was used for differential analysis and a total of 829 TME-related DEGs (Table S3) were obtained, including 726 up-regulated genes and 103 down-regulated genes (Fig. 1B). Furthermore, the R package "clusterProfiler" was used to carry out GO and KEGG pathway enrichment analysis on these obtained DEGs, and it was observed that most DEGs were related to the proliferation and activation of immune cells (Fig. 1C). These results suggested that the TME-related DEGs influenced the prognosis of BC by affecting tumor immunity.

2.2. Construction of a gene signature-based model for BC prognosis prediction

Univariate Cox, LASSO and multivariate Cox regression analysis were performed on the obtained TME-associated DEGs to construct a multivariate Cox risk regression model with variable gene expression. Firstly, univariate Cox regression analysis was carried out on the 829 DEGs to screen out survival-related DEGs, among which 154 DEGs with significantly statistical significance ($p < 0.01$) were selected (**Table S4**). LASSO regression analysis was used to establish a prediction model based on the 154 survival-related DGEs. In order to avoid overfitting of the prediction model, 10-fold cross validation method was used to choose tuning parameter (lambda) as 1000, and genes with higher correlation were removed for reducing model complexity. Finally, 13 relatively independent feature genes were acquired for subsequent analysis (Fig. 2A-B). Afterwards, multivariate Cox regression analysis was performed on the 13 feature genes, and a multivariate Cox risk regression model was constructed based on the survival time and survival status of patients, so as to screen the independent factors that dramatically influenced the prognosis of BC patients. The formula of the risk regression model was as follow: $Riskscore = \sum_{i=1}^{n} (Coef_i * x_i)$, where $Coef_i$ is the coefficient, and x_i is the z-score-transformed relative expression value of each selected gene. 6 relatively independent feature genes (PXDNL, FABP7, STX11, PIGR, SHISAL2A, WDR17) were identified to have the potential as independent prognostic factors associated with BC immunity (Fig. 2C).

2.3. Verification Of The Prognostic Gene Signature

In order to verify the accuracy of the prognostic gene signature containing 6 feature genes, ROC analysis was used to evaluate the predictive efficiency of the model. The ROC curves showed that the AUC value was 0.728, indicating that the model had a high accuracy (Fig. 3A). In addition, the samples were divided into high-risk group and low-risk group based on the median value of the sample risk score, and survival analysis was conducted on the two groups. The results exhibited that the survival time of patients in the high-risk group was significantly shorter than that in the low-risk group (Fig. 3B). In the meantime, higher risk scores were associated with shorter survival and higher mortality (Fig. 3C). These results verified that the 6-gene signature model performed well in predicting BC survival and prognosis with a good accuracy.

2.4 The 6 feature genes are independent prognostic factors affecting BC immunity
In order to analyze whether the 6 feature genes in the 6-gene signature model could be independent prognostic factors of BC, we conducted survival analysis on these feature genes, and found that these 6 feature genes all had

a significant impact on the prognosis of patients (Fig. 4A). FABP7, PIGR, SHISAL2A, STX11 and WDR17 were favorable prognostic factors, and their high expression was associated with better prognosis. PXDNL was an adverse prognostic factor, and patients with low expression of PXDNL had better prognosis. Moreover, basic clinical information (age, gender, stage, T_stage, M_stage and N_stage) was combined to conduct univariate and multivariate Cox regression analysis on these 6 feature genes. It was observed that the 6 feature genes were significantly associated with the prognosis of BC patients, especially showed an intimate correlation with age and stage, as the prognostic model or independent factors ($p <$ 0.001), (Fig. 4B). These results suggested that the 6 feature genes associated with tumor immunity could be used as independent prognostic factors for BC.

3. Discussion

With the development of systemic chemotherapy and radiotherapy in clinic, the survival of BC patients has been greatly improved. However, survival rates of patients with metastasis and recurrence are still unsatisfied. Therefore, identifying new biomarkers to predict the prognosis of BC patients is helpful for the customization of personalized treatment. High-throughput RNA-seq contributes to exploring various biomarkers for diagnosis and prognosis prediction of many cancers, including BC, along with the development of technology[20–23].

In this study, we firstly found that the high infiltration degree of immune cells in TME was beneficial to the prognosis of BC patients. Further, machine learning analysis was used to systematically analyze the potential role of tumor immunity-related DEGs as prognostic predictors, construct a prognostic gene signature-based model and evaluate the value of the feature genes as prognostic markers of BC. In order to solve the "curse of dimensionality" (the combination of small sample and a large number of genes), which is common in high-throughput biological data, we analyzed the expression data using LASSO regression model and cross-validation method. LASSO regression is adept at dealing with high-dimensional regression variables by reducing all regression coefficients to zero to force many regression variables to be completely zero, thus, LASSO regression model has high stability and prediction accuracy[24,25]. The BC survival-related DEGs with a better predictive performance were identified by

LASSO regression model and cross validation in this study, and a 6-mRNA signature model was constructed to predict the prognosis of BC.

There are several important findings during the analysis. Firstly, we found that most DEGs influenced the prognosis of BC patients by affecting the proliferation and activation of immune cells in TME. Secondly, 154 survival-related DEGs were screened using univariate Cox analysis and most of them were risk factors that may cause cancer (97/154). Finally, we constructed a prognostic gene signature model by using multivariate Cox analysis, and screened 6 BC immunity-related feature genes (PXDNL, FABP7, STX11, PIGR, SHISAL2A, WDR17) which could be used as independent prognostic factors. PXDNL (Peroxidasin like), a member of the peroxidase gene family, has a variety of biological functions, including hormone biosynthesis, host defense and cell movement[26]. A recent study by Li *et al.* has reported that PXDNL is closely related to the development of BC, and the high expression of PXDNL is a potential independent prognostic indicator of BC[27]. Pongor L *et al.* have found that PXDNL is a super-tumor suppressor gene in BC after analysis on gene chip data[28]. FABP7 (Fatty acid-binding protein 7) is a member of multi-gene fatty acid binding protein family and has been extensively studied in BC. Alshareeda AT *et al.* have discovered that FABP7 is poorly expressed in BC, and it can inhibit BC cell proliferation and promote differentiation through the JAK/STAT pathway[29]. In the study of Liu RZ *et al.*, FABP7 is significantly up-regulated in triple-negative BC, and increased FABP7 level is associated with poor prognosis, absence of estrogen and progesterone receptors (ER, PR) and HER 2, increased cell proliferation and high tumor grade[30]. The expression of PIGR (Polymeric Immunoglobulin Receptor) is down-regulated in BC tissues[14]. Bin Xiao *et al.* have screened PIGR as a feature gene to construct a risk scoring system related to the prognosis of BC[31]. Through these prior studies, we found that PXDNL, FABP7 and PIGR of the 6 feature genes were closely related to the occurrence and development of BC and they could serve as prognostic markers of BC. While STX11, SHISAL2A and WDR17 had not been reported in BC, indicating that these three genes were newly discovered, and the survival analysis results of our study indicated that these three genes may be favorable factors for the prognosis of BC patients.

4. Conclusions

There are some limitations that cannot be neglected in this study. Our results have not been experimentally verified, but we will study further to make the results more scientifically rigorous. In conclusion, a new 6-gene prognostic signature was discovered and it is a promising independent predictor. The 6 feature genes can function as important biomarkers for BC clinical treatment. At the same time, our study also provides a new insight to explore the molecular mechanism of TME and immune cells that influence BC progress.

5. Methods
5.1. Data acquisition

RNA-seq data from 1,109 patients and clinical data from 1097 patients were obtained from the TCGA-BRCA in the public database TCGA (https://www.cancer.gov/).

5.2. Data Processing

Based on the Fragments per Kilobase Million (FPKM) data of TCGA-BRCA mRNA, immune and stromal cells of each sample were scored by ESTIMATE prediction software[32] to predict tumor purity and analyze the influence of TME on the prognosis of BC. The samples were divided into high-score group and low-score group based on the median value of the immune score. Differential analysis was conducted by using Wilcox test to obtain DEGs affecting tumor immunity.

5.3. Enrichment Analysis

GO and KEGG pathway enrichment analysis were performed on DEGs affecting tumor immunity by using R-package "ClusterProfiler". On the basis of GO analysis, gene functional annotation was conducted from three aspects: biological process (BP), cell component (CC) and molecular function (MF), with FDR < 0.05 as the threshold.

5.4. Construction Of A Prognostic Gene Signature-based Model

Univariate Cox regression analysis was used to screen the DEGs associated with BC survival, and p value less than 0.01 was set as the standard to screen the candidate feature genes for subsequent analysis. LASSO regression analysis was used for further screening. It is an effective method for analysis of high-dimensional predictors and has been widely used in Cox proportional risk regression model for survival analysis[33,34]. 10-fold cross-validation method was used to select the tuning parameter (lambda) to determine the penalty degree and screen out relatively independent feature genes. Furthermore, multivariate Cox regression analysis was performed and feature genes with $p < 0.05$ were used to construct the prognostic signature.

5.5. Validation Of The Prognostic Model And Feature Genes

Patients were divided into high-risk group and low-risk group according to the median risk score, and Kaplan-Meier (K-M) survival analysis was performed in both groups. In addition, receiver operating characteristic (ROC) analysis was used to evaluate the predictive efficiency of the model and the area under the curve (AUC) was calculated. Meanwhile, survival analysis was carried out on the selected prognostic feature genes to evaluate their value as prognostic and survival markers. All of these processes were performed by R software (version 3.5.1).

6. Abbreviations

BC: breast cancer; TME: tumor microenvironment; ECM-receptor: extracellular matrix-receptor; PXDNL: Peroxidasin like; FABP7: Fatty acid-binding protein 7; ER, PR: estrogen and progesterone receptors; PIGR: Polymeric Immunoglobulin Receptor; FPKM: Fragments per Kilobase Million; CC: cell component; MF: molecular function; K-M: Kaplan-Meier; ROC: operating characteristic; AUC: area under the curve;

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests

Funding

None.

Authors' contributions

YH analyzed and interpreted the patient data regarding the hematological disease and the transplant. XW and SXW performed the histological examination of the kidney, and was a major contributor in writing the manuscript. All authors read and approved the final manuscript."

Acknowledgements

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1. Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, et al. Cancer treatment and survivorship statistics, 2019. CA Cancer J Clin. 2019;69:363–85.

2. Ren JF, Hakki AH, Kotler MN, Iskandrian AS. Exercise systolic blood pressure: a powerful determinant of increased left ventricular mass in patients with hypertension. J Am Coll Cardiol. 1985;5:1224–31.

3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394–424.

4. Möller S, Mucci LA, Harris JR, Scheike T, Holst K, Halekoh U, et al. The Heritability of Breast Cancer among Women in the Nordic Twin Study of Cancer. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2016;25:145–150.

5. DeSantis C, Ma J, Bryan L, Jemal A. Breast cancer statistics, 2013. CA Cancer J Clin. 2014;64:52–62.

6. Matsumoto A, Jinno H, Ando T, Fujii T, Nakamura T, Saito J, et al. Biological markers of invasive breast cancer. Jpn J Clin Oncol. 2016;46:99–105.

7. Noël A, Foidart JM. The role of stroma in breast carcinoma growth in vivo. J Mammary Gland Biol Neoplasia. 1998;3:215–25.

8. Tuxhorn JA, McAlhany SJ, Dang TD, Ayala GE, Rowley DR. Stromal cells promote angiogenesis and growth of human prostate tumors in a differential reactive stroma (DRS) xenograft model. Cancer Res. 2002;62:3298–307.

9. Ballman KV. Biomarker: Predictive or Prognostic? Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2015;33:3968–71.

10. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6.

11. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005;365:671–9.

12. Li R, Campos J, Iida J. A Gene Regulatory Program in Human Breast Cancer. Genetics. 2015;201:1341–8.

13. Ascierto ML, Idowu MO, Zhao Y, Khalak H, Payne KK, Wang XY, et al. Molecular signatures mostly associated with NK cells are predictive of relapse free survival in breast cancer patients. Journal of translational medicine. 2013;11:145.

14. Bao Y, Wang L, Shi L, Yun F, Liu X, Chen Y, et al. Transcriptome profiling revealed multiple genes and ECM-receptor interaction pathways that may be associated with breast cancer. Cell Mol Biol Lett. 2019;24:38.

15. Rodrigues-Ferreira S, Nehlig A, Monchecourt C, Nasr S, Fuhrmann L, Lacroix-Triki M, et al. Combinatorial expression of microtubule-associated EB1 and ATIP3 biomarkers improves breast cancer prognosis. Breast cancer research treatment. 2019;173:573–83.

16. He Y, Ma J, Wang A, Wang W, Luo S, Liu Y, et al. A support vector machine and a random forest classifier indicates a 15-miRNA set related to osteosarcoma recurrence. Onco Targets Ther. 2018;11:253–69.

17. Chen YC, Ke WC, Chiu HW. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med. 2014;48:1–7.

18. De Smet F, Pochet NL, Engelen K, Van Gorp T, Van Hummelen P, Marchal K, et al. Predicting the clinical behavior of ovarian cancer from gene expression profiles. International journal of gynecological cancer: official journal of the International Gynecological Cancer Society. 2006;16(Suppl 1):147–51.

19. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer informatics. 2007;2:59–77.

20. Picornell AC, Echavarria I, Alvarez E, López-Tarruella S, Jerez Y, Hoadley K, et al. Breast cancer PAM50 signature: correlation and concordance between RNA-Seq and digital multiplexed gene expression technologies in a triple negative breast cancer series. BMC Genomics. 2019;20:452.

21. Brandão RD, Mensaert K, López-Perolio I, Tserpelis D, Xenakis M, Lattimore V, et al. Targeted RNA-seq successfully identifies normal and pathogenic splicing events in breast/ovarian cancer susceptibility and Lynch syndrome genes. International journal of cancer. 2019;145:401–14.

22. Yang L, Zhang X, Hou Q, Huang M, Zhang H, Jiang Z, et al. Single-cell RNA-seq of esophageal squamous cell carcinoma cell line with fractionated irradiation reveals radioresistant gene expression patterns. BMC Genomics. 2019;20:611.

23. Feng H, Li T, Zhang X. Characterization of kinase gene expression and splicing profile in prostate cancer with RNA-Seq data. BMC Genomics. 2018;19:564.

24. Mao Y, Fu Z, Zhang Y, Dong L, Zhang Y, Zhang Q, et al. A six-microRNA risk score model predicts prognosis in esophageal squamous cell carcinoma. J Cell Physiol. 2019;234:6810–9.

25. Algamal ZY, Lee MH. Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Comput Biol Med. 2015;67:136–45.

26. Gu SQ, Bakthavachalu B, Han J, Patil DP, Otsuka Y, Guda C, et al. Identification of the human PMR1 mRNA endonuclease as an alternatively processed product of the gene for peroxidasin-like protein. 18. New York: RNA; 2012. pp. 1186–96.

27. Li Y, Jiao Y, Luo Z, Li Y, Liu Y. High peroxidasin-like expression is a potential and independent prognostic biomarker in breast cancer. Medicine. 2019;98:e17703.

28. Pongor L, Kormos M, Hatzis C, Pusztai L, Szabó A, Győrffy B. A genome-wide approach to link genotype to clinical outcome by utilizing next generation sequencing and gene chip data of 6,697 breast cancer patients. Genome medicine. 2015;7:104.

29. Alshareeda AT, Rakha EA, Nolan CC, Ellis IO, Green AR. Fatty acid binding protein 7 expression and its sub-cellular localization in breast cancer. Breast cancer research treatment. 2012;134:519–29.

30. Liu RZ, Graham K, Glubrecht DD, Lai R, Mackey JR, Godbout R. A fatty acid-binding protein 7/RXRβ pathway enhances survival and proliferation in triple-negative breast cancer. J Pathol. 2012;228:310–21.

31. Xiao B, Chen L, Ke Y, Hang J, Cao L, Zhang R, et al. Identification of methylation sites and signature genes with prognostic value for luminal breast cancer. BMC Cancer. 2018;18:405.

32. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications. 2013;4:2612.

33. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Statistics in medicine. 2007;26:5512–28.

34. Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in medicine. 1997;16:385–95.

Figures
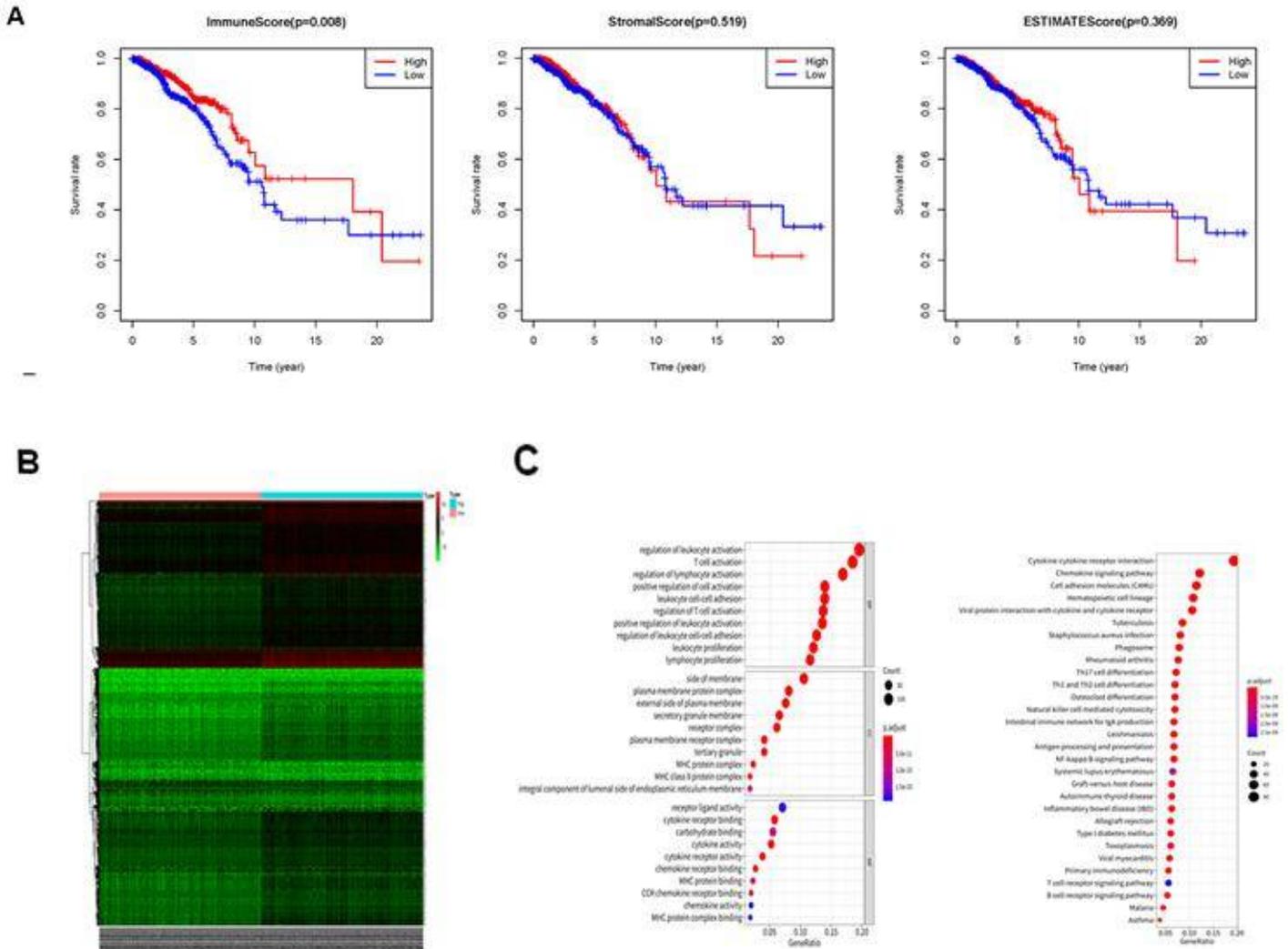
Figure 1

Identification of the DEGs associated with TME in BC (A) Survival analysis of BC patients in the high-score and low-score groups (patients were grouped based on the median value of the immune score, stromal score and ESTIMATE score); (B) Heat map of DEGs in high and low immune score groups; (C) The results of GO and KEGG enrichment analysis of the screened DEGs.
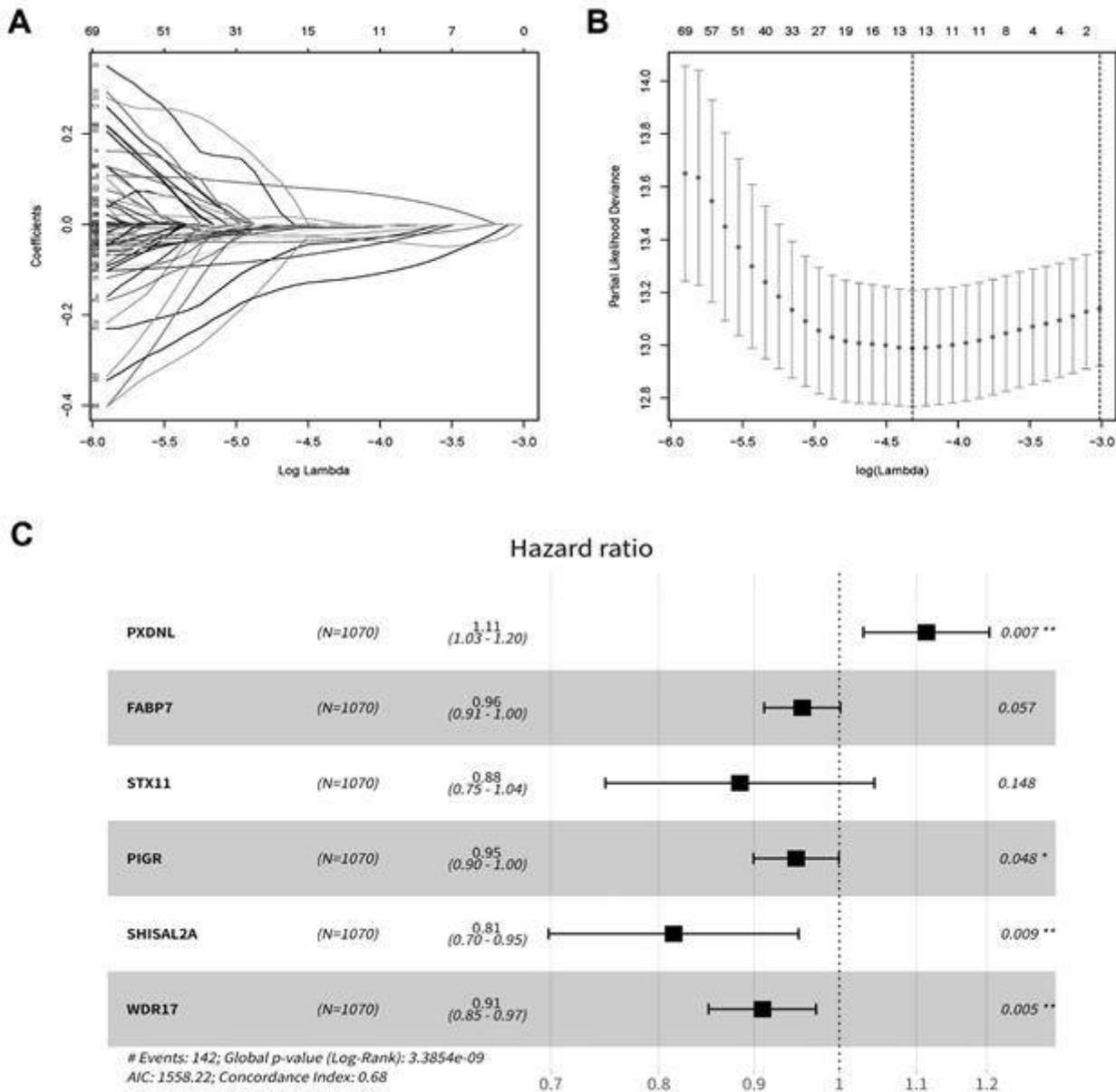
Figure 2

A prognostic gene signature-based model of BC is constructed and feature genes are screened (A) The LASSO coefficient profile of the survival related DEGs, which is generated against the sequence of log lambda; (B) The parameter lambda of LASSO model was selected by using10-fold cross validation via minimum criteria, and the dotted vertical lines were drawn at the optimal value with the minimum criteria and the 1 standard error of the minimum criteria; (C) In the forest map of multivariate Cox regression analysis, the gene with a hazard radio <1 (left side of dotted line) was considered a low-risk factor, while the gene with a hazard radio >1 (right side of the dotted line) was regarded as a high-risk factor.
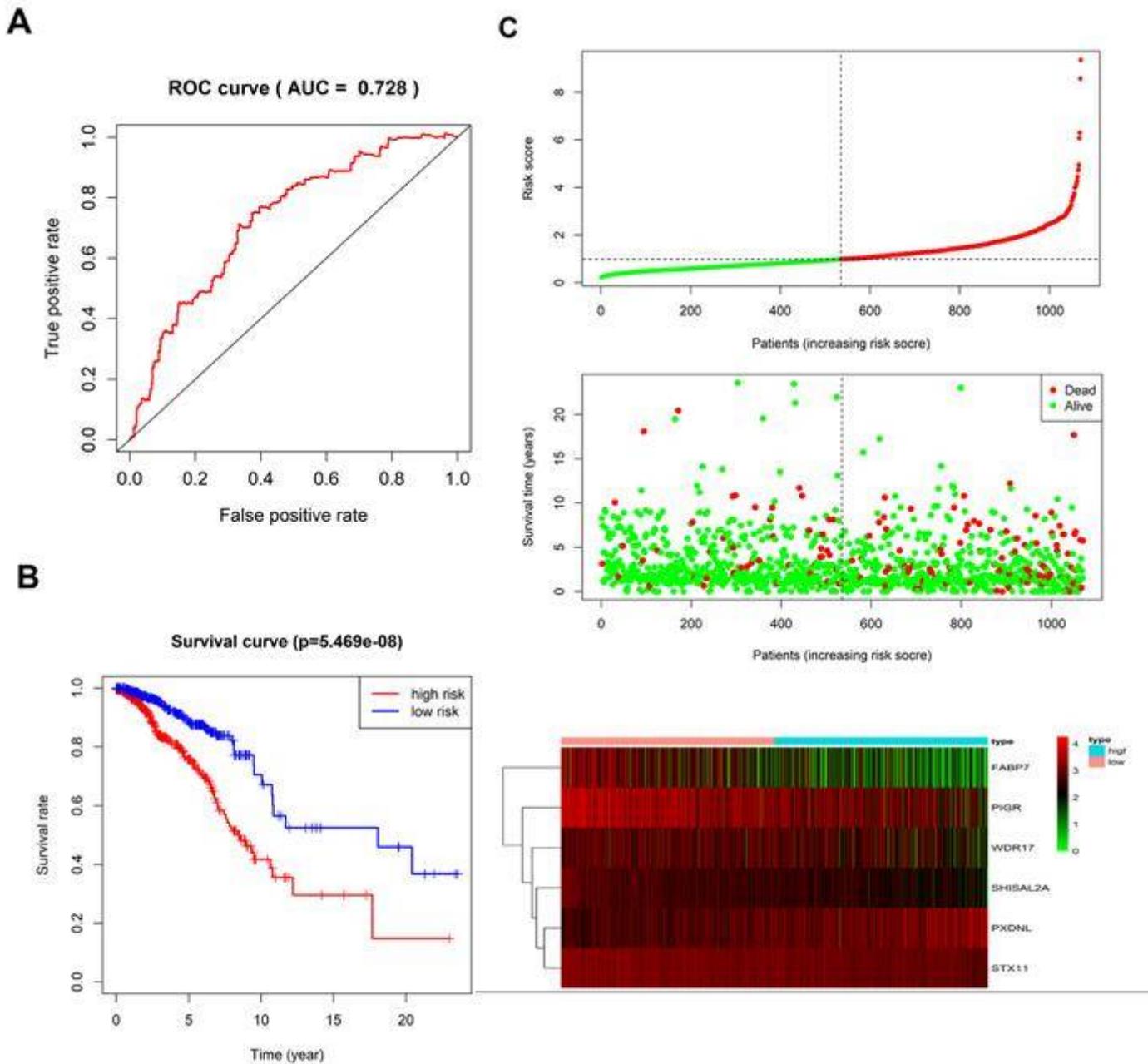
Figure 3

Verification of the 6-gene signature model (A) The ROC curve of the risk model (the Y-axis is true positive rate, and the X-axis is false positive rate); (B) Survival analysis of patients in the high-risk (red) and low-risk (blue) groups; (C) Distribution of the 6-gene risk score, patients' survival status and expression of the 6 prognostic factors. With the increase of risk score, patients' survival time is shortened, death events are increased, and the expression levels of the 6 feature genes are increased or decreased.
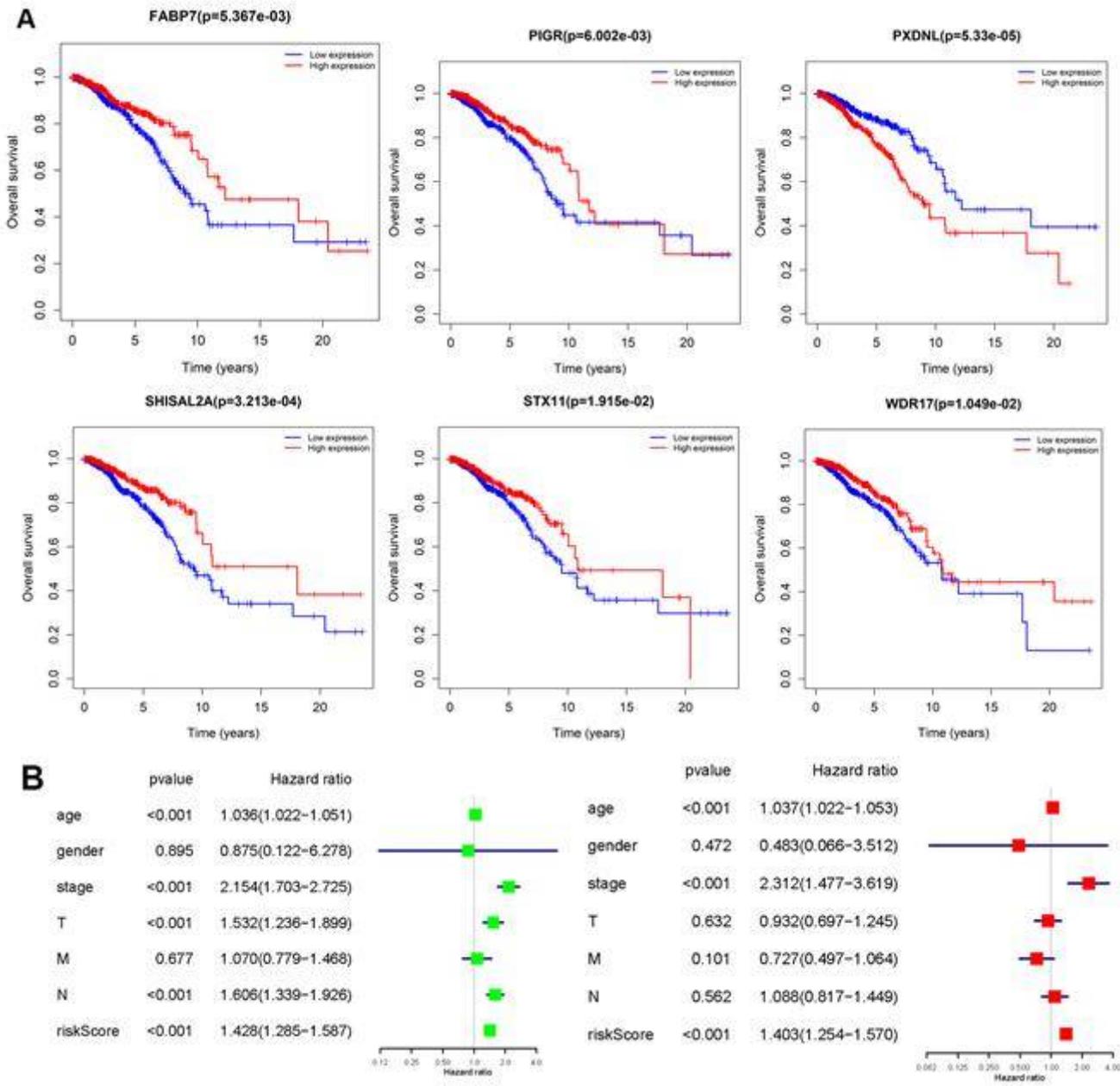
Figure 4

The 6 feature genes are independent prognostic factors affecting BC immunity (A) Survival analysis of patients based on the expression of the 6 feature genes in the risk model. The red line represents the high-expression group, and the blue line represents the low-expression group; (B) The forest map showed the results of univariate and multivariate Cox regression analysis based on the expression of the 6 genes combined with clinical information.

Supplementary Files

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

- S1.xlsx
- S4.xlsx
- S3.xls
- S2.xls