Identification of a Novel Gene Model-Based Homologous Recombination Deficiency Score to Improve Survival Prediction of TNBC.

Wenxiang Zhang
Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Bolun Ai
Department of Breast Surgery, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Xiangyi Kong
Department of Breast Surgery, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Xiangyu Wang
Department of Breast Surgery, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Jie Zhai
Department of Breast Surgery, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Ran Gao
Department of Breast Surgery, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Yihang Qi
Department of Breast Surgery, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Qiang Liu
Department of Breast Surgery, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Mengliu Zhu
Department of Breast Surgery, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Yingpeng Ren
Department of Breast Surgery, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
Yi Fang

Department of Breast Surgery, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College

Jing Wang (wangjing@cicams.ac.cn)

Chinese Academy of Medical Sciences and Peking Union Medical College https://orcid.org/0000-0002-3224-3993

Primary research

Keywords: triple-negative breast cancer, homologous recombination deficiency, prognosis, overall survival, TCGA

Posted Date: May 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-472194/v1

Abstract

Background

Triple-negative breast cancer (TNBC) is a specific histological type of breast cancer with a poor prognosis, early recurrence, which lacks durable chemotherapy responses and effective targeted therapies. We aimed to construct an accurate prognostic risk model based on homologous recombination deficiency (HRD) - gene expression profiles for improving prognosis prediction of TNBC.

Methods

Triple-negative breast cancer RNA sequencing data and sample clinical information were downloaded from the breast invasive carcinoma (BRCA) cohort in the Cancer Genome Atlas (TCGA) database. Combined with the HRD database, tumor samples were divided into two sets. We screened differentially expressed genes (DEGs) and then identified HRD-related prognostic genes using weighted gene co-expression network analysis (WGCNA) and Cox regression analysis. The least absolute shrinkage and selection operator (LASSO) and multivariate Cox regression analysis were used to identifying key prognostic genes. Risk scores were calculated and compared with HRD score, Kaplan–Meier (KM) survival analysis were used to assess its prognostic power. GSE103091 dataset from GEO (Gene Expression Omnibus) database was used to validate the signature. Univariate and multivariate Cox regression were performed to independently verify the prognosis of the risk score. A nomogram was constructed and revealed by time-dependent ROC curves to guide clinical practice.

Results

We found that HRD tumor samples (HRD score > = 42) in TNBC patients were associated with poor overall survival (p = 0.027). We identified a total of 147 differential genes including 203 up-regulated and 213 down-regulated genes, among which 29 were prognosis-related genes. Through the LASSO method, 6 key prognostic genes ((MUCL1, IVL, FAM46C, CHI3L1, PRR15L, and CLEC3A) were selected and a 6-gene risk score was constructed. We found risk score was negatively associated with homologous recombination deficiency (HRD) scores (r = -0.22, p = 0.019). Compared with the low-risk group, Kaplan-Meier survival analysis shows that the high-risk group has an obvious poorer prognosis (P < 0.0001). Finally, we integrated the risk score model and clinical factors of TNBC (AJCC-stage, HRD score, T stage, and N stage) to construct a compound nomogram. Time-dependent ROC curves showed the risk score performed better in 1-, 3- and 5-year survival predictions compared with AJCC-stage.

Conclusions

Based on HRD gene expression data, our six HRD-related gene signature and nomogram could be practical and reliable tools for predicting OS in patients with TNBC.

Introduction

Triple-negative breast cancer (TNBC) is an aggressive subtype of breast cancer, which is characterized by the loss of the expression of estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) and accounts for 10–20% of all types of breast cancer [1]. Compared with non-TNBC patients, TNBC patients have the characteristics of younger age, higher histological grade, larger tumor size, higher positive rate of lymph nodes and easier metastasis to lung, brain and other parts [1, 2]. Secondly, TNBC cannot benefit from endocrine therapy and anti-HER-2 targeted therapy, chemotherapy has become the main adjuvant therapy for TNBC patients. However, with the increase of drug resistance, the effective chemotherapy of TNBC is limited, even if active treatment is carried out, the median overall survival of patients with metastatic, triple-negative breast cancer is less than one year [2]. TNBC also has obvious heterogeneity. There are survival differences between different subtypes. Not all patients have a poor prognosis. Traditional clinicopathological indicators and single molecular markers have obvious limitations in predicting prognosis.

Homologous recombination repair (HRR) is an important pathway signal pathway for several cellular processes including the error-free repair of DNA double-strand breaks (DSB) and the recovery of stalled DNA replication forks [3], in which the key proteins are BRCA1 and BRCA2. BRCA1 or BRCA2 loss of function leads to homologous recombination deficiency (HRD). However, germline and somatic alteration of other HRR-related genes, such as PALB2, CDK12, RAD51, CHEK2, ATM or BRCA1 gene promoter methylation can lead to HRD in sporadic cancers, broadly termed BRCAness [4, 5]. HRD has a high incidence of ovarian cancer, prostate cancer, breast cancer and pancreatic cancer [6, 7]. some clinical studies had found that HRD status is highly correlated with the sensitivity of platinum-based chemotherapeutics and PARP inhibitors and is a key indicator of treatment options and prognosis for a variety of tumors[8]. To this end, the development and clinical evaluation of platforms to identify HR defects have recently been a subject of intense investigation, especially in TNBC, as this subtype is considered to be enriched for the HR pathway deficiency. Approximately 40% and 70% of TNBC tumors have a presumed HRD phenotype[9, 10], HRD score based on NGS assay was defined as the unweighted sum of the loss of heterozygosity (LOH), telomeric allelic imbalance (TAI), and large-scale state transitions (LOS) scores [11]. High HRD scores have shown to be significantly associated with sensitivity to neoadjuvant platinum-based chemotherapy in TNBC [10, 12]. However, most of the work carried out around HRD is to study the mutations of HRD-related genes, and the accuracy of genomics to identify HRD. there is no transcriptomics research is involved.

In the present study, we gathered information about the clinical features and RNA sequencing data of 123 TNBC tumor samples from the TCGA database and identified HRD grouping based on the genome. we then construct a prognostic model of the HRD transcriptome and compare it with the genome HRD score.

Methods

Data acquisition and preprocessing

The TCGA (The Cancer Genome Atlas) BC mRNA expression profile (The TCGA-BRCA cohort) and the associated clinical information were downloaded from Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). Use R (version 3.6.0) software to standardize and process data. The HRD information of tumor samples can be obtained from HRD related data (HRD score see: PMID: 29617664, TCGA_DDR_Data_Resources.xlsx form: DDR footprints), the tumor samples were divided into HRD tumor samples (HRD score ≥ 42) and non-HRD tumor samples (HRD score <42) according to the HRD score. GSE103091 data set was obtained from Gene Expression Omnibus (GEO, https ://www.ncbi.nlm.nih.gov/geo/) for validation.

Identification and enrichment analysis of HRD-related differentially expressed genes in TNBC

To acquire differentially expressed genes (DEGs) in TNBC (between the HRD and non-HRD samples). The R language limma package (http://www.r-project.org/) was used to analyze for differentially expressed genes between HRD and non-HRD samples, the threshold for a significant difference was designated a P-value of <0.005. The R package “cluster profile” was used to carry out Gene Ontology (GO) enrichment analysis including biological process (BP), cell components (CC) and molecular functions (MF) for the differentially expressed HRD-related genes. The same tool is also used for the enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis. The weighted gene co-expression network analysis (WGCNA) package was used to identify key modules associated with prognosis based on differentially expressed HRD-related genes.

Identification and verification of prognostic gene signatures

We next performed univariate Cox survival analysis on the 199 key module genes to screen cancer-related prognostic factors. Variables with a P value less than 0.1 in the univariate analysis were screed to identify independent prognostic factors. The 29 prognosis-related genes identified were included in the LASSO regression analysis by using the R package “glmnet” to screen the genes. Then, the multivariable Cox proportional risk regression analysis was carried out to establish the prognosis model of TNBC.

The risk score of each patient was calculated according to the mRNA expression level of each risk gene and the risk coefficient, we used the following formula to calculate the risk score of each patient.

$$\text{Risk score} = \sum_{i=1}^{6} \beta_i * \text{Exp}_i.$$

where Exp_i is the expression level of each prognostic gene, and β_i is the regression coefficient of it.

patients were divided into the high-risk group and low-risk group with the median risk score as the threshold. K-M curve was used to compare the survival difference between the above two groups.

To validate the prognostic value of our risk model, we obtained the GSEGSE103091 data set from the Gene Expression Omnibus (GEO) database, the risk score of each patient was calculated using the coefficients of 6 genes above. Then the patients were stratified into high-risk and low-risk groups by the median risk score. The correlation between risk score and HRD score was evaluated using Spearman's rank correlation analysis. The KM survival analysis used to validate the multi-gene prognostic signature.

In order to analyze the stability of the risk prediction model in different levels of other clinical prognostic parameters, the KM curves were used to compare the difference of subgroups of AJCC-stage. In addition, in order to independently verify the prognosis of the risk score. single variable and multivariate Cox regression analyses were conducted. Risk score, T stage, N stage, AJCC -stage and HRD score were used as covariates.

Construction of gene prognostic nomogram

A composite nomogram was constructed based on all independent prognostic parameters screened by univariate and multivariate Cox proportional hazards regression analysis above and compared with AJCC staging systems in the clinical cohort. we used the “rms” and “survival” packages in R to predict the probability of 1-year, 3-year and 5-year OS.

Results

Identification of DEGs and Functional Analysis

The TCGA-BRCA cohort consisted of 1104 tumor samples (123 TNBC and 981 non-TNBC samples) and 114 normal samples. Combined with the HRD database, we screened out 110 samples with HRD scores in TNBC samples and divided them into HRD tumor samples and non-HRD tumor samples according to the HRD score (>=42 for HRD samples). Kaplan-Meier survival analysis showed that HRD tumor samples in TNBC patients were associated with poor overall survival **(Figure 1)**. The limma package of R was used for detecting the DEGs between HRD samples and non-HRD samples. There were 417 differential genes including 203 up-regulated and 214 down-regulated genes with significant differences ($P < 0.05$) **(Figure 2A-B)**. full result can be obtained in the **supplementary Table1** and **Supplementary Table2**. all the differentially expressed genes (DEGs) were further analyzed by carrying out gene ontology and KEGG enrichment, The DEGs based on the HRD score, were mainly enriched in BP related to digestion and inner ear development, CC was associated with extracellular matrix and collagen−containing extracellular matrix, and MF terms were associated with ion channel activity **(Figure 2C-E)**. The KEGG pathway identified was protein digestion and absorption pathway **(Figure 2F)**.

Identification of Key Prognostic Genes in TCGA -TNBC dataset

To evaluate the prognostic effect of DEGs, the weighted gene co-expression network analysis (WGCNA) package was used to identify key modules associated with prognosis. Finally, the key finding of our study is that brown and blue modules containing 99 genes and 100 genes respectively were found to have a

stronger correlation with overall survival **(Figure 3)**. 199 key module genes were analyzed by single variable Cox regression (**Supplementary Table3)**. We identified a set of 29 genes whose P values were less than 0.1, part of the results is shown in **Table 1**. At the same time, we selected the genes included in the subsequent risk model to draw the Kaplan-Meier curve, and the result is shown in **Figure 4**. To further identify the 29 candidate genes that were significantly correlated with the prognosis of TNBC patients, LASSO regression was performed, which was related with 6 genes (MUCL1, IVL, FAM46C, CHI3L1, PRR15L and CLEC3A) in DEGs that significantly associated with OS **(Figure 5)**.

Construction and Estimation of a 6-Gene Risk Score

According to the gene expression level and the risk coefficient of each gene, the risk score of each patient was calculated **(Supplementary Table4)**, the median risk score of each model was used as the threshold to divide the samples into high- and low-risk groups to draw the Kaplan-Meier (KM) curve and the risk factor linkage diagram of model evaluation **(Figure 6)**, compared with the low-risk group, Kaplan-Meier survival analysis shows that the high-risk group has an obvious poorer prognosis (P<0.0001) **(Figure 6A)**. Besides, in order to analyze the stability of the risk prediction model, Kaplan-Meier survival curve analyses showed that the low-risk group was significantly correlated with better OS in N0 stage (P = 0.0032), N1-N3 stage (P = 0.00033), stage I + stage II stage (P = 0.00011) and T1-T2 (P = 0.00015) patients **(Figure7)**.

Interanion and Validation of the 6-gene risk score

Firstly, we conducted a Spearman's correlation test to evaluate the correlation between the HRD score and the 6-gene risk score. Scatter plot of the 6-gene risk score and HRD score showing the negative linear relationship between the two variables (Pearson correlation coefficient = -0.22) **(Figure 8)**. In additional, to verify the predictive value of the six-gene prognostic signature, we conducted internal and external verification. The risk score in univariate analysis was significantly correlated with overall survival (OS) (HR = 0.074, 95% CI = 0.017–0.032, P =0.00056) **(Figure 9A)**. Multivariate analysis showed that the risk score was an independent prognostic indicator (HR = 39.373, 95% CI = 7.059–219.624, P < 0.001) **(Figure 9B)**. Besides, we obtained the GSEGSE103091 data set from the Gene Expression Omnibus (GEO) database and used our risk scoring model for survival analysis. Finally, the results were consistent with the training group **(Figure 10)**. All in all, the above data showed that 6-gene risk score was an independent risk factor of patients with TNBC.

Construction of nomogram

To better predict patients' prognosis and guide clinical practice, we integrated the risk score model and clinical factors of TNBC (AJCC-stage, HRD score, T stage, and N stage) to construct a compound nomogram **(Figure.11A)**. The calibration curve of the nomogram was abnormal, so we used 1-year, 3-year, 5-year ROC curves instead. The 1-,3-and 5-year time-dependent ROC curves were used to evaluate the predictive ability of the risk score (AUC = 0.785 of 1 year, 0.710 of 3 years, and 0.847 of 5 years), and its predictive ability was found to be better than AJCC-stage **(Figure.11B-D)**.

Discussion

Breast cancer is the most frequently diagnosed cancer with an incidence rate of 11.7% of newly diagnosed female cancers and the first most common cause of cancer mortality in women worldwide [13]. Triple-negative breast cancer (TNBC) accounts for 20% of all molecular subtypes of breast cancer [14]. Although significant benefits of new targeted therapy and immunotherapy have been reported in the past few decades. Due to the high heterogeneity and few genetic targets, these tumors have the worst prognosis among all of the breast cancer subtypes and the overall survival rate within five years is less than 70% [15]. Traditional clinicopathological parameters, tumor-node-metastasis (TNM) staging system and single molecular markers have obvious limitations in predicting prognosis. It is necessary to identify the effective prognostic biomarkers of TNBC and establish the relevant prognostic risk prediction model. Multiple studies have reported that mutations in HRD-related genes are associated with the prognosis of multiple tumors, as well as the efficacy of PARP inhibitors. However, there is no research involving any transcriptomics about HRD. In this study, we construct a prognostic risk model based on transcriptome data of HRD. This signature could be used to efficiently determine the overall survival time of TNBC patients.

In this study, we screened out 110 TNBC samples with HRD scores according to the TCGA breast cancer dataset and HRD database. the DEGs were identified from HRD tumor samples and non-HRD tumor samples in TNBC patients. After WGCNA analyses, univariate, LASSO and multivariate Cox regression analyses, MUCL1, IVL, FAM46C, CHI3L1, PRR15L and CLEC3A were screened out as prognostic genes ultimately to develop the prognostic model. These genes contained in the signature have previously been reported to be associated with different cancer in various ways.

Mucin-like 1 (MUCL1) is a gene encoding a low molecular weight glycoprotein with high similarity to sialomucins, which was only expressed in salivary glands and breast tissues. It was identified as a breast-specific gene for breast cancer micrometastasi [16]. Most recently, Liu Liang et al. used IHC technology to detect the expression level of MUCL1 in paraffin-embedded tissues of 89 triple negative breast cancer patients and found that high MUCL1 expression is significantly correlated with high recurrence and death rates in triple negative breast cancer patients [17]. In additional, several studies have shown that MUCL1 expression strongly correlates with clinical stage of TNM and the status of axillary lymph node metastasis[17]. Involucrin (IVL), a component of keratinocyte crosslinked envelope, is found in the cytoplasm and crosslinked with membrane proteins by transglutaminase. This gene is mapped to 1q21, among calpactin I light chain, trichohyalin, profillaggrin, loricrin, and calcyclin. Recently, IHL has been identified as a novel hub gene that shows a significant up-regulation in colon adenocarcinoma as compared to normal tissue [18]. So far, there is little research on IVL in TNBC. only one study reported that 6-mRNA signature including IVL may act as a potential prognostic biomarker in patients with TNBC[19]. Family with sequence similarity 46, member C (FAM46C) is a member of the FAM46 family, it is located on chromosome 1p12 and seems to play a role in the regulation of translation by acting as an mRNA stabilizing factor. its abnormal deletions in tumor tissues were confirmed in multiple myeloma[20] and gastric cancer[21]. Zhang, et al reported that FAM46C was downregulated in

hepatocellular carcinoma (HCC)and induced cell apoptosis through regulating Ras/MEK/ERK pathway [22]. In additional, FAM46C was downregulated in prostate cancer to inhibit cell proliferation and cell cycle progression and promote apoptosis through PTEN/AKT signaling pathway[23]. However, there is no research on FAM46C in TNBC. CHI3L1, on human chromosome 1q32.1, encodes a secreted glycoprotein called YKL-40, which plays an important role in inflammation, angiogenesis, radioresistance, and cancer progression. Overexpression of CHI3L1 have been de described in various types of cancer, including oligodendroglia, glioblastoma, osteosarcoma, breast, and small-cell lung cancers[24, 25]. YKL-40 expression was significantly upregulated in NSCLC tissues, and associated with poor prognosis and shorter survival [25]. PRR15L, also known as ATAD4, which encodes a protein of unknown function, to date, no report has been ascribed to this function of this gene. C-type lectin domain family 3 member A (CLEC3A), belonging to the superfamily of C-type lectins, is known to associate with cell adhesion which influenced results in tumor cell proliferation and metastasis[26, 27]. It was reported that CLEC3A expressed initially in cartilage and was associated with osteoarthritis. Recently, Ni, J et al [28] figured out that high CLEC3A expression significantly correlated with poor prognosis in IDC patients and promoted invasion and metastasis of breast cancer through activating PI3K/AKT signaling. Our findings suggest that these genes may be acted as important biomarkers to predict survival outcomes in patients with TNBC. If we can explore their specific mechanisms of action in triple-negative breast cancer more extensively and in-depth, it is likely that they can be used as new cancer biomarkers.

After identifying the six prognostic genes, the risk score model of HRD signature was developed and investigated for its prognostic value in TNBC patients, a clear separation was observed in the survival curve between patients in high-risk and low-risk subgroups, which was evaluated as a category variable (divided by median cutoff). We also found that the low-risk group had a very low proportion of deaths. Furthermore, we performed a stratification analysis, and the results suggested that 6-gene risk score in clinical subgroups (N0 stage, N1 + N2 + N3 stage, stage I + stage II, T1 + T2 stage) could still better predict the prognosis of TNBC. Then the univariate and multivariate Cox regression analysis showed that 6-gene risk score could be an independent factor to evaluate the prognosis. Finally, we developed a nomogram to guide clinical practice including AJCC-stage, HRD score, T stage, and N stage, and risk score to construct a nomogram to predict the 3-year and 5-year survival of TNBC patients. when compared with the TNM stage, 6-gene risk score showed the even better predictive ability in the ROC analysis.

However, we acknowledge several limitations in our study. Firstly, our study only focused on the large-scale mRNA sequencing data from the TCGA platform. the results report may be biased. Secondly, we searched the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) for external validation, but many data sets have no prognostic OS information. Thirdly, in this study, the gene function and participation mechanism of six-gene models have not been clarified, and the relationship with the occurrence and development of breast cancer needs to be further confirmed by research under in vitro and in vivo conditions.

Conclusion

In summary, our study developed a six HRD-related gene risk score for the prognostic prediction of TNBC based on samples deposited in the TCGA database. which can reduce the waste of medical resources and contribute to personalized treatment decisions.

Declarations

Acknowledgements

Not applicable.

Funding

This study was supported by the National Natural Science Foundation of China (No.81872160). The above funders had no further role in the study design; collection, analysis, and interpretation of data; writing of the manuscript; or decision to submit this manuscript for publication

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in The Cancer Genome Atlas database and additional files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Authors'information

Wenxiang Zhang and BoLun Ai contributed equally to this paper.

Affiliations

National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China.

Wenxiang Zhang, MD, Bolun Ai, MD, Xiangyi Kong, MD, Xiangyu Wang, MD, Jie Zhai, MD, Ran Gao, MD, Yihang Qi, MD, Qiang Liu, MD, Mengliu Zhu, MD, Yingpeng Ren, MD, Yi Fang, MD * Jing Wang, MD*

Contributions

WJ and FY designed the overall study and revised the paper, ZWX and ABL performed public data interpretation and drafted the manuscript. WXY, KXY, ZJ and LQ participated in data collection, RYP, QYH, GR and ZML contributed to data analysis. All authors read and approved the final manuscript.

Corresponding Authors

Correspondence to Yi Fang and Jing Wang.

References

1. Brown M, Tsodikov A, Bauer KR, Parise CA, Caggiano V: **The role of human epidermal growth factor receptor 2 in the survival of women with estrogen and progesterone receptor-negative, invasive breast cancer: the California Cancer Registry, 1999-2004**. *Cancer* 2008, **112**(4):737-747.
2. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, Lickley LA, Rawlinson E, Sun P, Narod SA: **Triple-negative breast cancer: clinical features and patterns of recurrence**. *Clin Cancer Res* 2007, **13**(15 Pt 1):4429-4434.
3. Pfaffle HN, Wang M, Gheorghiu L, Ferraiolo N, Greninger P, Borgmann K, Settleman J, Benes CH, Sequist LV, Zou L *et al*: **EGFR-activating mutations correlate with a Fanconi anemia-like cellular phenotype that includes PARP inhibitor sensitivity**. *Cancer Res* 2013, **73**(20):6254-6263.
4. Turner N, Tutt A, Ashworth A: **Hallmarks of 'BRCAness' in sporadic cancers**. *Nat Rev Cancer* 2004, **4**(10):814-819.
5. Hoppe MM, Sundar R, Tan DSP, Jeyasekharan AD: **Biomarkers for Homologous Recombination Deficiency in Cancer**. *J Natl Cancer Inst* 2018, **110**(7):704-713.
6. Nielsen FC, van Overeem Hansen T, Sorensen CS: **Hereditary breast and ovarian cancer: new genes in confined pathways**. *Nat Rev Cancer* 2016, **16**(9):599-612.
7. Nguyen L, J WMM, Van Hoeck A, Cuppen E: **Pan-cancer landscape of homologous recombination deficiency**. *Nat Commun* 2020, **11**(1):5584.
8. Suh DH, Kim M, Lee KH, Eom KY, Kjeldsen MK, Mirza MR, Kim JW: **Major clinical research advances in gynecologic cancer in 2017**. *J Gynecol Oncol* 2018, **29**(2):e31.
9. Akashi-Tanaka S, Watanabe C, Takamaru T, Kuwayama T, Ikeda M, Ohyama H, Mori M, Yoshida R, Hashimoto R, Terumasa S *et al*: **BRCAness predicts resistance to taxane-containing regimens in triple negative breast cancer during neoadjuvant chemotherapy**. *Clin Breast Cancer* 2015, **15**(1):80-85.
10. Telli ML, Timms KM, Reid J, Hennessy B, Mills GB, Jensen KC, Szallasi Z, Barry WT, Winer EP, Tung NM *et al*: **Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer**. *Clin Cancer Res* 2016, **22**(15):3764-3773.
11. Moes-Sosnowska J, Rzepecka IK, Chodzynska J, Dansonka-Mieszkowska A, Szafron LM, Balabas A, Lotocka R, Sobiczewski P, Kupryjanczyk J: **Clinical importance of FANCD2, BRIP1, BRCA1, BRCA2 and FANCF expression in ovarian carcinomas**. *Cancer Biol Ther* 2019, **20**(6):843-854.

12. Timms KM, Abkevich V, Hughes E, Neff C, Reid J, Morris B, Kalva S, Potter J, Tran TV, Chen J *et al*: **Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes**. *Breast Cancer Res* 2014, **16**(6):475.
13. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F: **Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries**. *CA Cancer J Clin* 2021.
14. Podo F, Buydens LM, Degani H, Hilhorst R, Klipp E, Gribbestad IS, Van Huffel S, van Laarhoven HW, Luts J, Monleon D *et al*: **Triple-negative breast cancer: present challenges and new perspectives**. *Mol Oncol* 2010, **4**(3):209-229.
15. Chalakur-Ramireddy NKR, Pakala SB: **Combined drug therapeutic strategies for the effective treatment of Triple Negative Breast Cancer**. *Biosci Rep* 2018, **38**(1).
16. Liu ZZ, Xie XD, Qu SX, Zheng ZD, Wang YK: **Small breast epithelial mucin (SBEM) has the potential to be a marker for predicting hematogenous micrometastasis and response to neoadjuvant chemotherapy in breast cancer**. *Clin Exp Metastasis* 2010, **27**(4):251-259.
17. Liu L, Liu Z, Qu S, Zheng Z, Liu Y, Xie X, Song F: **Small breast epithelial mucin tumor tissue expression is associated with increased risk of recurrence and death in triple-negative breast cancer patients**. *Diagn Pathol* 2013, **8**:71.
18. Wang H, Liu J, Li J, Zang D, Wang X, Chen Y, Gu T, Su W, Song N: **Identification of gene modules and hub genes in colon adenocarcinoma associated with pathological stage based on WGCNA analysis**. *Cancer Genet* 2020, **242**:1-7.
19. Lv X, He M, Zhao Y, Zhang L, Zhu W, Jiang L, Yan Y, Fan Y, Zhao H, Zhou S *et al*: **Identification of potential key genes and pathways predicting pathogenesis and prognosis for triple-negative breast cancer**. *Cancer Cell Int* 2019, **19**:172.
20. Manfrini N, Mancino M, Miluzio A, Oliveto S, Balestra M, Calamita P, Alfieri R, Rossi RL, Sassoe-Pognetto M, Salio C *et al*: **FAM46C and FNDC3A Are Multiple Myeloma Tumor Suppressors That Act in Concert to Impair Clearing of Protein Aggregates and Autophagy**. *Cancer Res* 2020, **80**(21):4693-4706.
21. Tanaka H, Kanda M, Shimizu D, Tanaka C, Kobayashi D, Hayashi M, Iwata N, Yamada S, Fujii T, Nakayama G *et al*: **FAM46C Serves as a Predictor of Hepatic Recurrence in Patients with Resectable Gastric Cancer**. *Ann Surg Oncol* 2017, **24**(11):3438-3445.
22. Zhang QY, Yue XQ, Jiang YP, Han T, Xin HL: **FAM46C is critical for the anti-proliferation and pro-apoptotic effects of norcantharidin in hepatocellular carcinoma cells**. *Sci Rep* 2017, **7**(1):396.
23. Ma L, He H, Jiang K, Jiang P, He H, Feng S, Chen K, Shao J, Deng G: **FAM46C inhibits cell proliferation and cell cycle progression and promotes apoptosis through PTEN/AKT signaling pathway and is associated with chemosensitivity in prostate cancer**. *Aging (Albany NY)* 2020, **12**(7):6352-6369.
24. Bergmann OJ, Johansen JS, Klausen TW, Mylin AK, Kristensen JS, Kjeldsen E, Johnsen HE: **High serum concentration of YKL-40 is associated with short survival in patients with acute myeloid leukemia**. *Clin Cancer Res* 2005, **11**(24 Pt 1):8644-8652.

25. Wang XW, Cai CL, Xu JM, Jin H, Xu ZY: **Increased expression of chitinase 3-like 1 is a prognosis marker for non-small cell lung cancer correlated with tumor angiogenesis**. *Tumour Biol* 2015, **36**(2):901-907.
26. Tsunezumi J, Higashi S, Miyazaki K: **Matrilysin (MMP-7) cleaves C-type lectin domain family 3 member A (CLEC3A) on tumor cell surface and modulates its cell adhesion activity**. *J Cell Biochem* 2009, **106**(4):693-702.
27. Boguslawska J, Rodzik K, Poplawski P, Kedzierska H, Rybicka B, Sokol E, Tanski Z, Piekielko-Witkowska A: **TGF-beta1 targets a microRNA network that regulates cellular adhesion and migration in renal cancer**. *Cancer Lett* 2018, **412**:155-169.
28. Ni J, Peng Y, Yang FL, Xi X, Huang XW, He C: **Overexpression of CLEC3A promotes tumor progression and poor prognosis in breast invasive ductal cancer**. *Onco Targets Ther* 2018, **11**:3303-3312.

Tables

Table 1
Results of single variable Cox regression of some genes.

Gene	HR (95% CI)	*P*-value	Cutoff
CHI3L1	0.78 (0.64–0.96)	0.018	9.241414
CLEC3A	1.3 (1.1–1.6)	0.0023	1.876011
FAM46C	0.69 (0.47-1)	0.057	7.500264
IVL	1.2 (1-1.5)	0.035	3.095136
MUCL1	1.3 (1.1–1.5)	0.0028	4.78792
PRR15L	1.7 (1.1–2.5)	0.0095	7.098303
Note: Cutoff, the threshold to distinguish the high and low expression of the gene.			

Figures

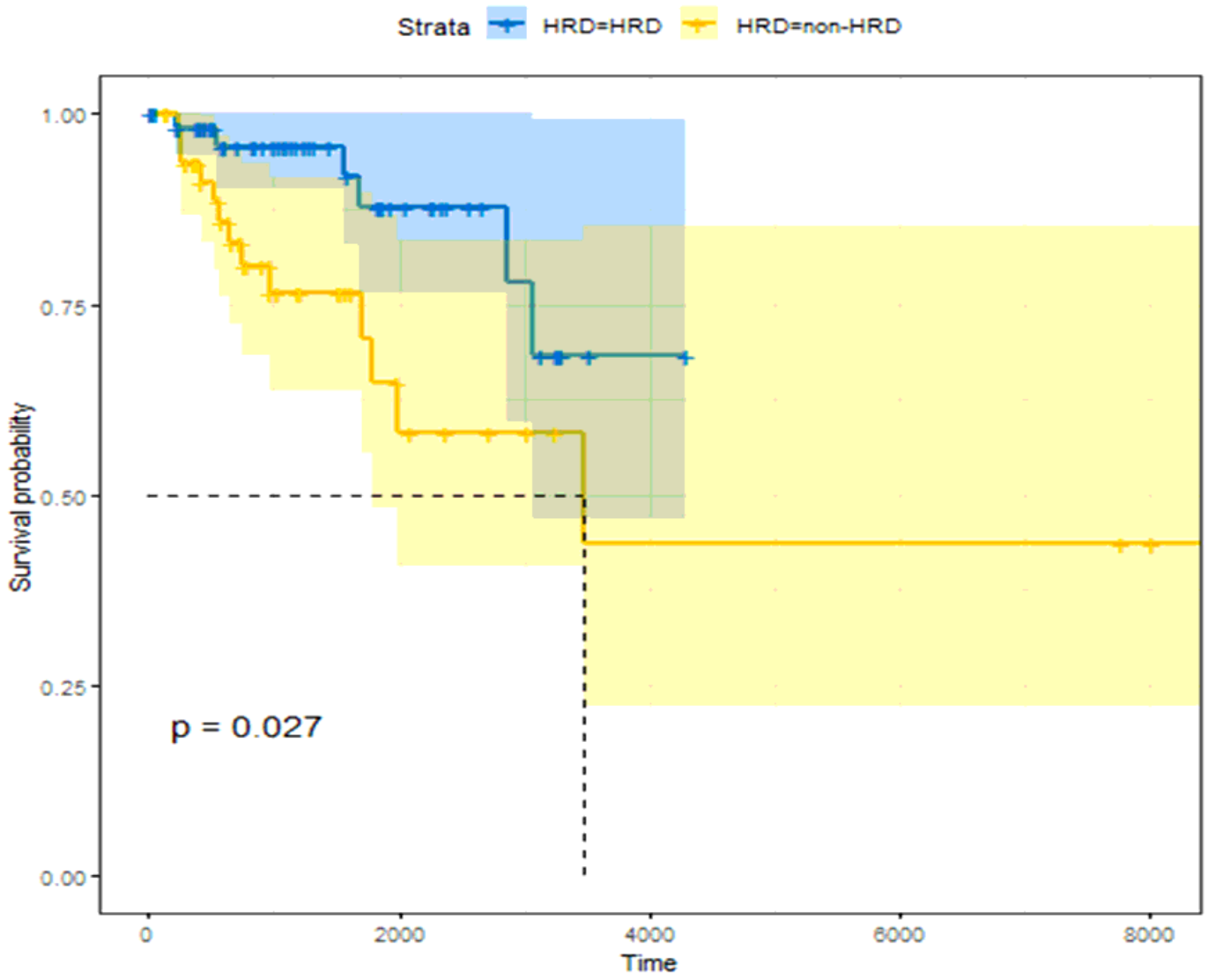

Figure 1

Kaplan-Meier curves of HRD samples with overall survival time.

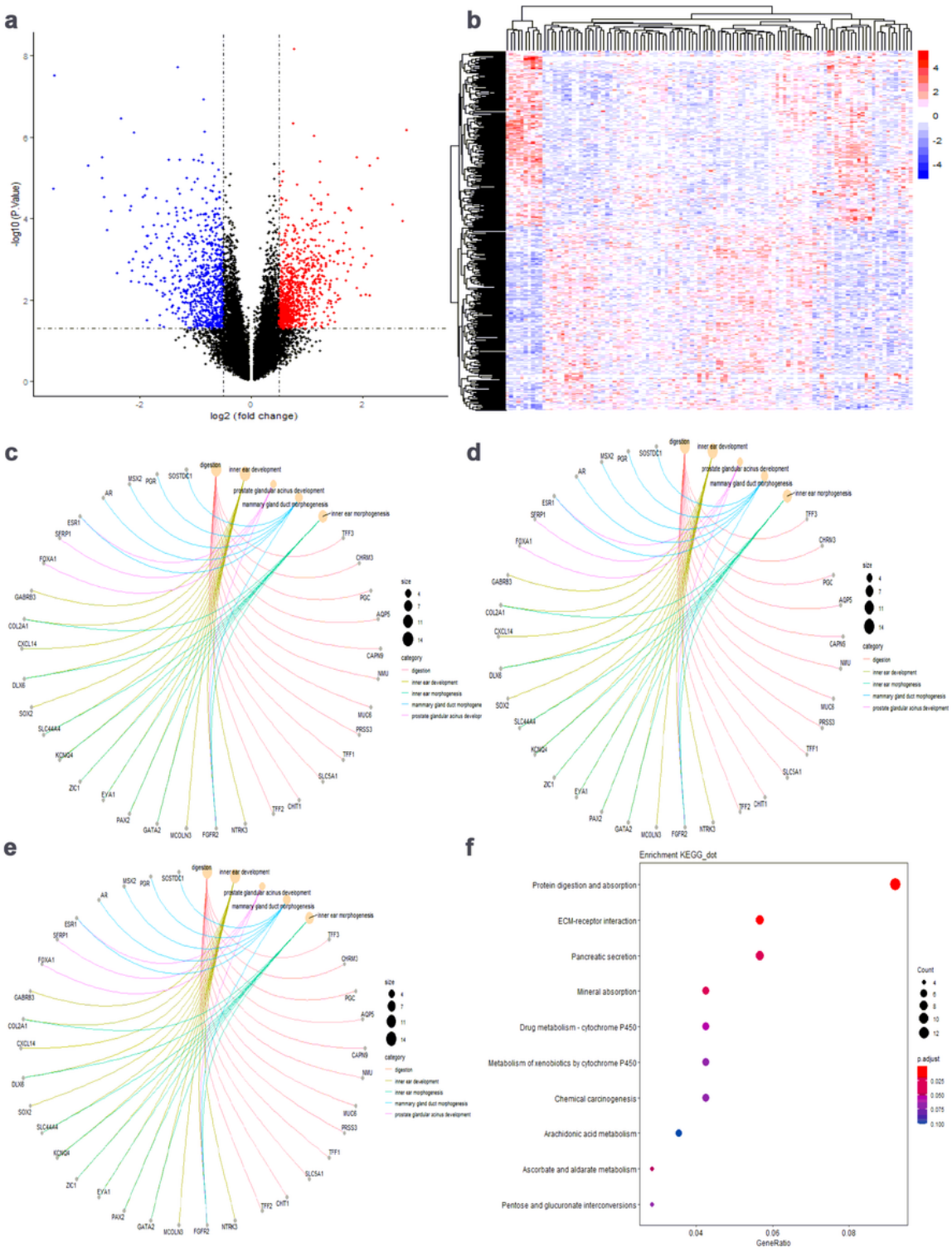

Figure 2

Identification of differentially expressed genes (DEGs) and enrichment analysis of DEGs in triple-negative breast cancer (TNBC). a Volcano plots showing the DEGs between the HRD and non-HRD samples. The red, blue and black dots represent genes with upregulated, downregulated, and unchanged expression, respectively. b Heat map showing the differential gene expression between the HRD and non-HRD samples. c-e GO enrichment contains three categories including biological process, cellular component

and molecular function. f KEGG enrichment analysis of DEGs. The depth of color corresponds to the enrichment significance, the size of the circle indicates the enriched gene count.

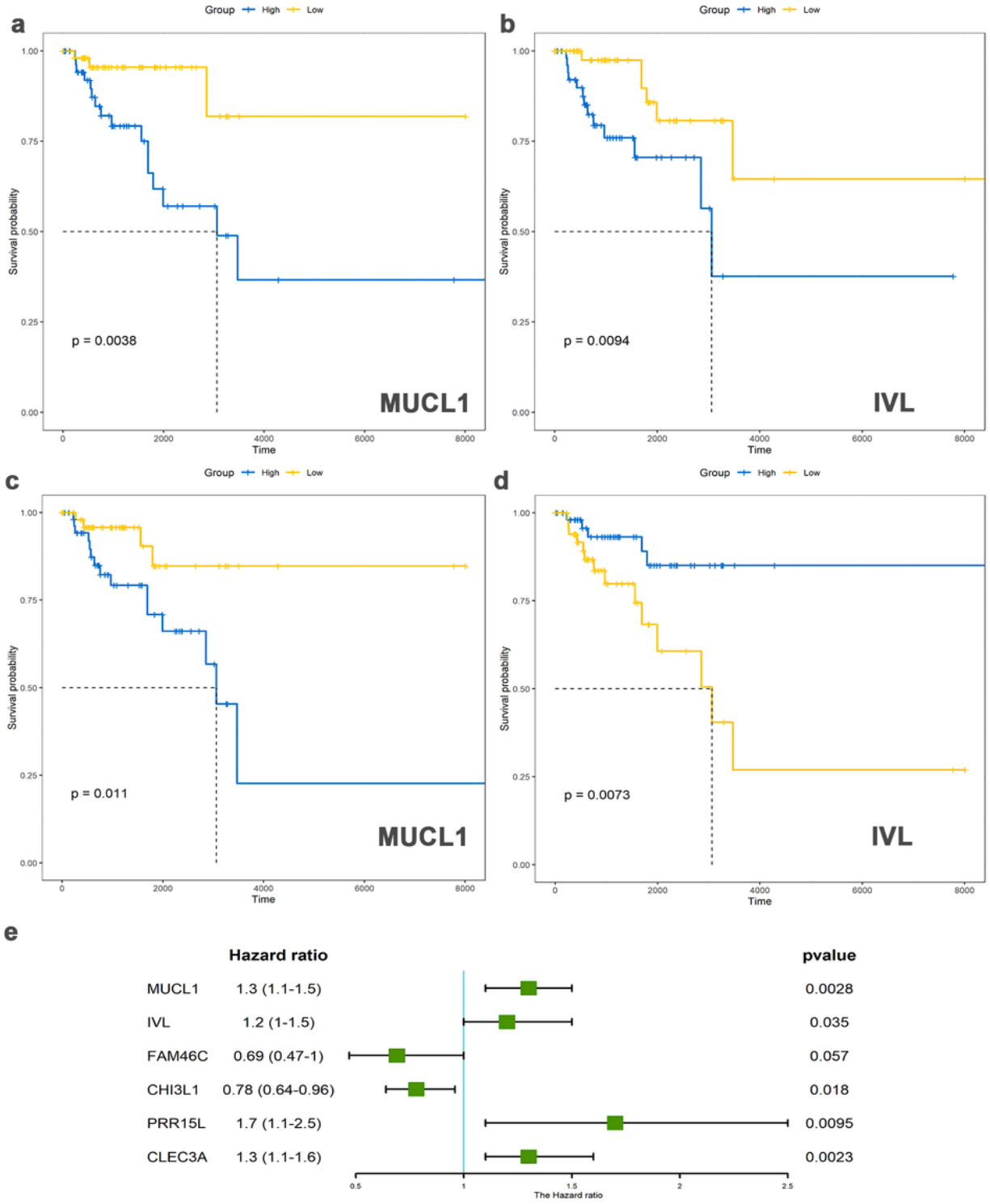

Figure 3

WGCNA analysis of DEGs. a Analysis of topology for various soft-thresholding powers. The left panel shows the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis). The right panel displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-

axis). In all, 4 was the fittest power value. b Clustering dendrograms of genes with dissimilarity based on the topological overlap, together with assigned module colors. As a result, 5 coexpression modules were constructed and are shown in different colors. c The degree of association between DEGs in the module and Metastasis (gene trait significance, GS). d Module-trait relationships. Heatmap of the correlation between module eigengenes and clinical characteristics of TNBC, the table is color-coded by correlation according to the color legend.

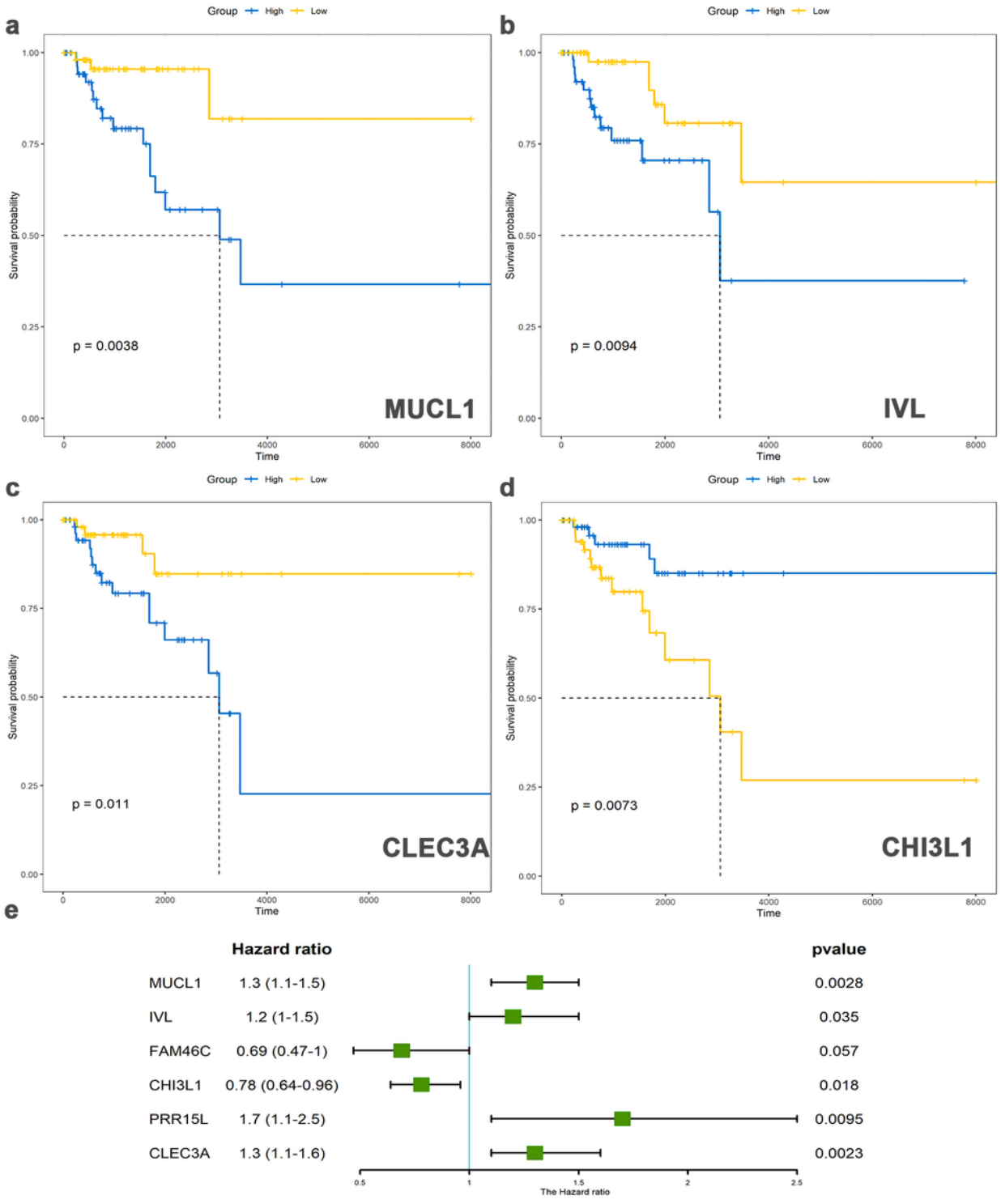

Figure 4

Kaplan–Meier survival analysis and forest map of some prognostic genes. a-d Kaplan–Meier survival analysis of MUCL1, IVL, CLEC3A, CHI3L1. e Forest map based on the univariate COX regressions of some prognostic genes (MUCL1, IVL, FAM46C, CHI3L1, PRR15L, and CLEC3A).

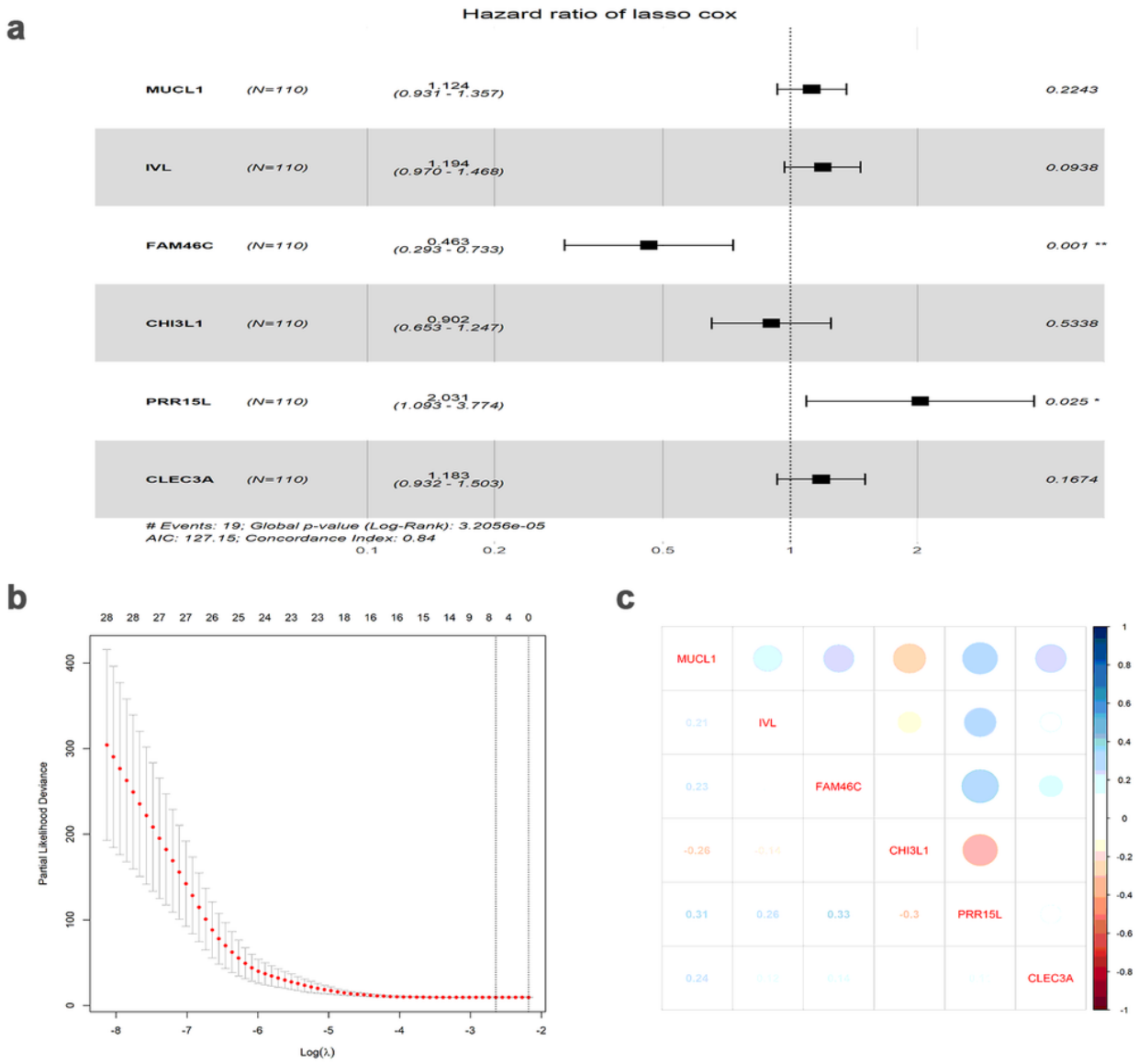

Figure 5

Identification of 6 significantly prognostic genes and their expression data in TNBC. a Multivariate Cox regression analysis of 6 prognostic genes in TCGA-BRCA cohort. b Partial likelihood deviation map, LASSO regression with fivefold cross-validation obtained 6 prognostic genes using minimum lambda value. c

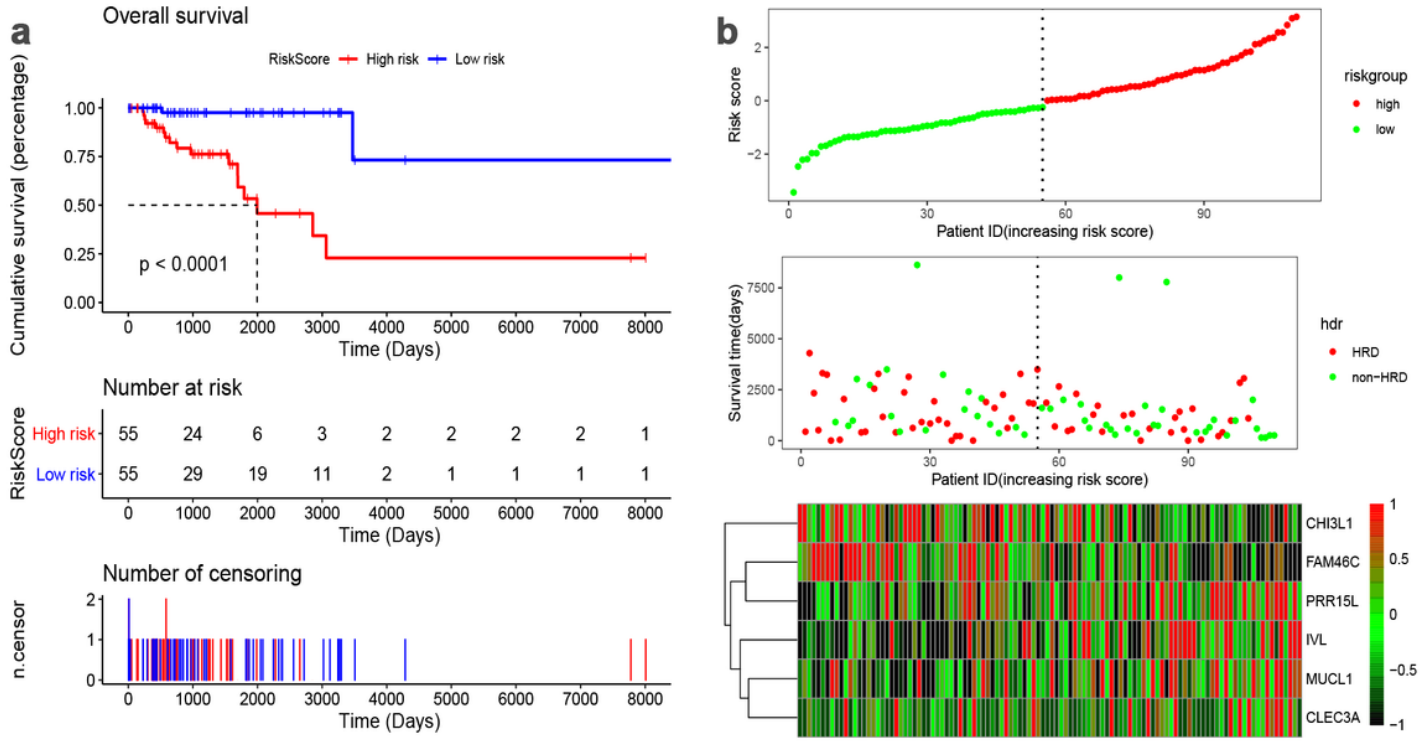

Figure 6

Prognostic analysis of six-gene signature. a Kaplan–Meier survival analysis of the six-gene signature, the upper part shows the Kaplan-Meier curves for the high- and low-risk groups; the middle shows the number at risk with time in the high- and low-risk groups; the bottom shows the number of censoring variations with time in the high- and low-risk groups. b The risk factor linkage diagram of model evaluation, the upper part shows the curve of risk score; the middle shows the number of living patient variations with time in the high- and low-risk groups; the bottom shows heatmap of the expression profiles of the six prognostic genes in low- and high-risk group.

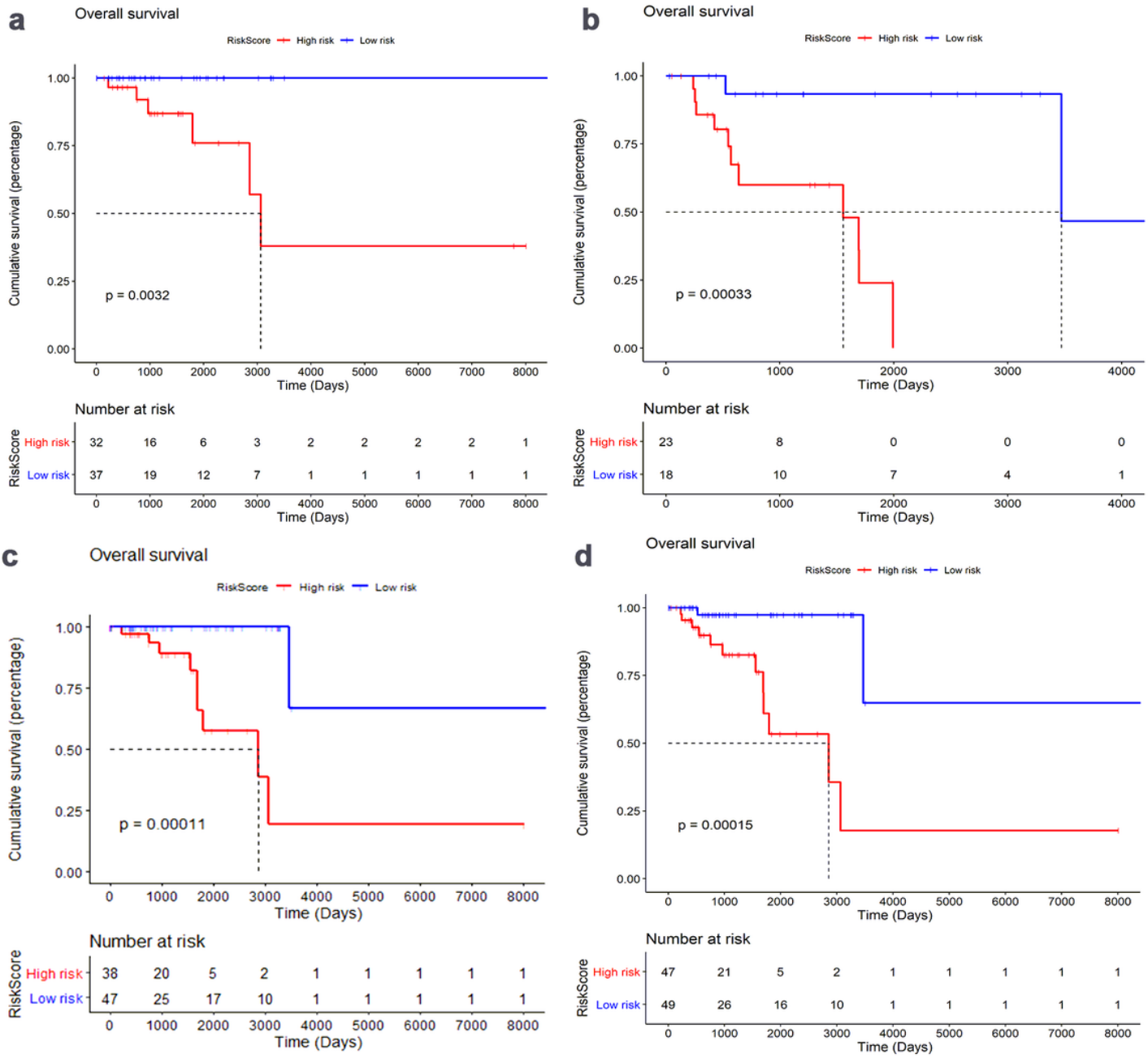

Figure 7

Validation of the stability of the 6-gene risk score. Kaplan–Meier survival for OS in subgroups stratified by N0 stage, N1-N3 stage, stage I + stage II stage, and T1-T2 patients.

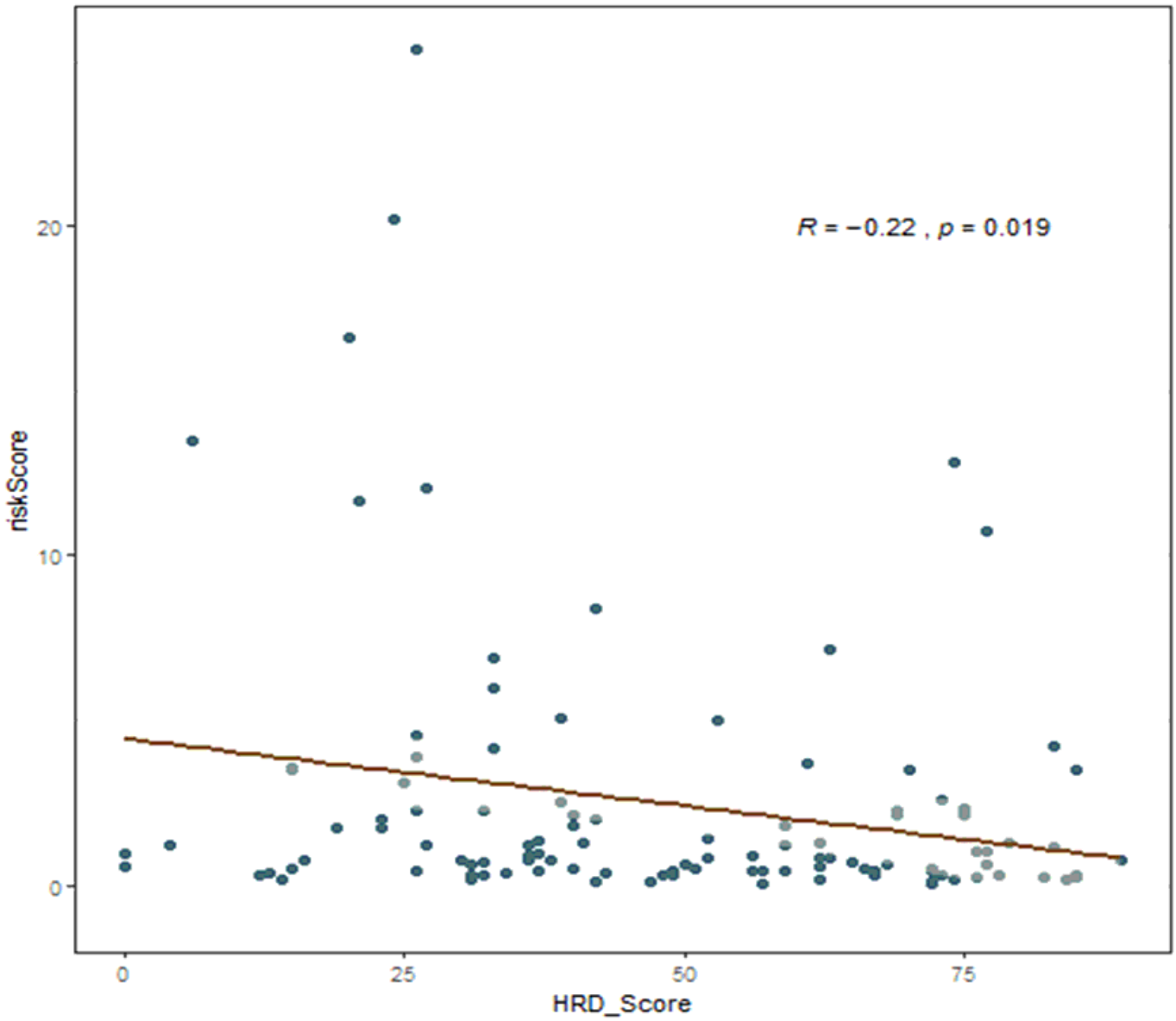

Figure 8

The correlation between the 6-gene score and HRD scores.

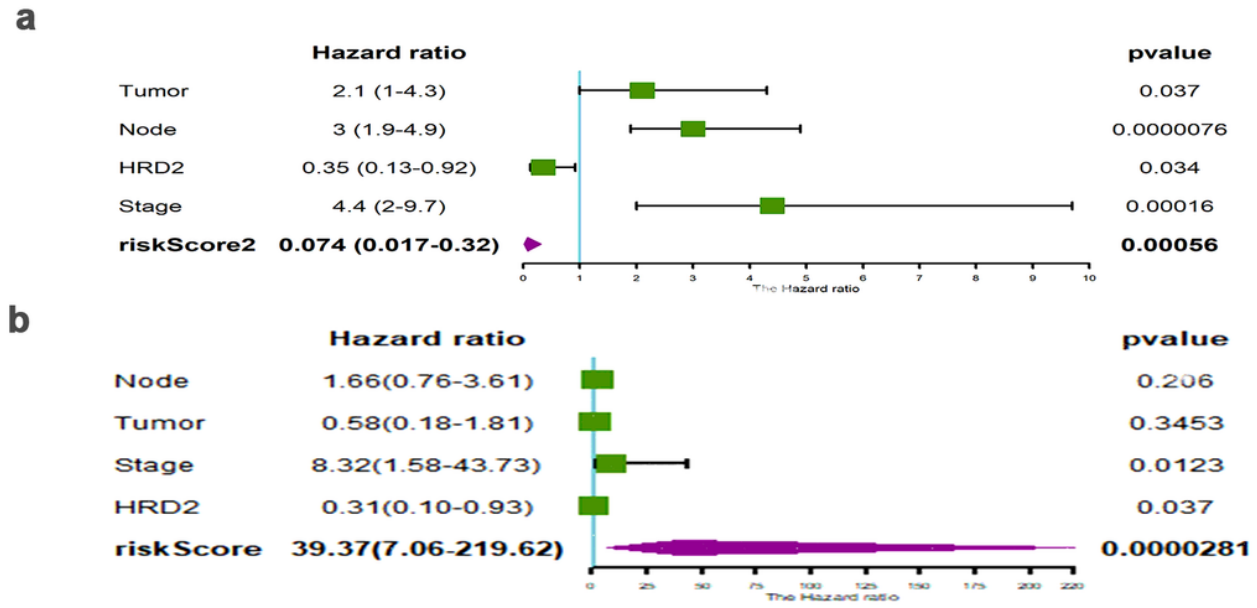

Figure 9

Identifying the independent prognostic parameters, univariate and multivariate Cox regression analyses of clinical factors associated with overall survival. a Forrest plot of univariate Cox regression analysis in TNBC. b Forrest plot of multivariate Cox regression analysis in TNBC.

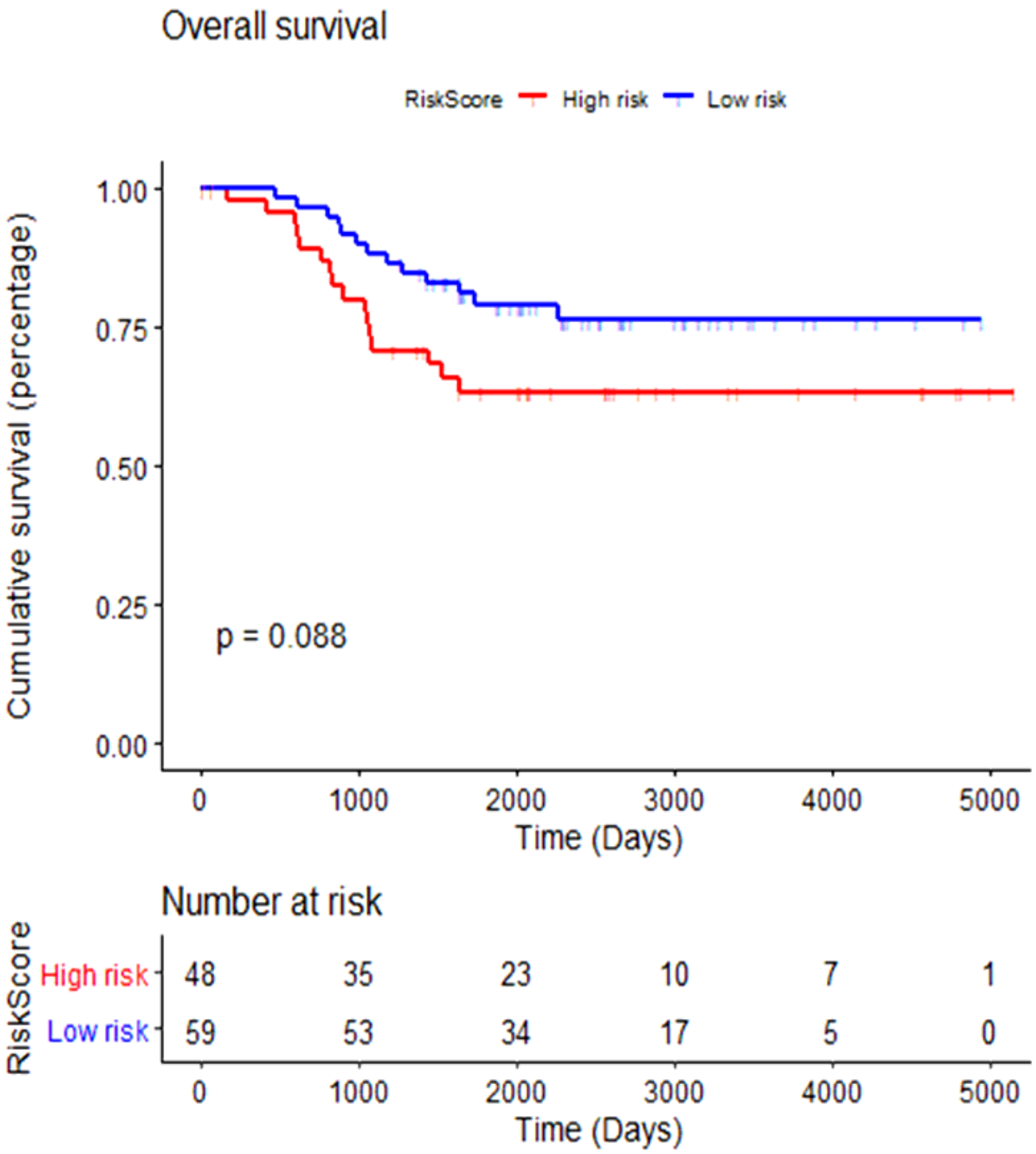

Figure 10

Validation of the 6-gene signature. GSE103091 was regarded as the external validation set. Kaplan–Meier survival analysis of the 6-gene signature in external validation set.

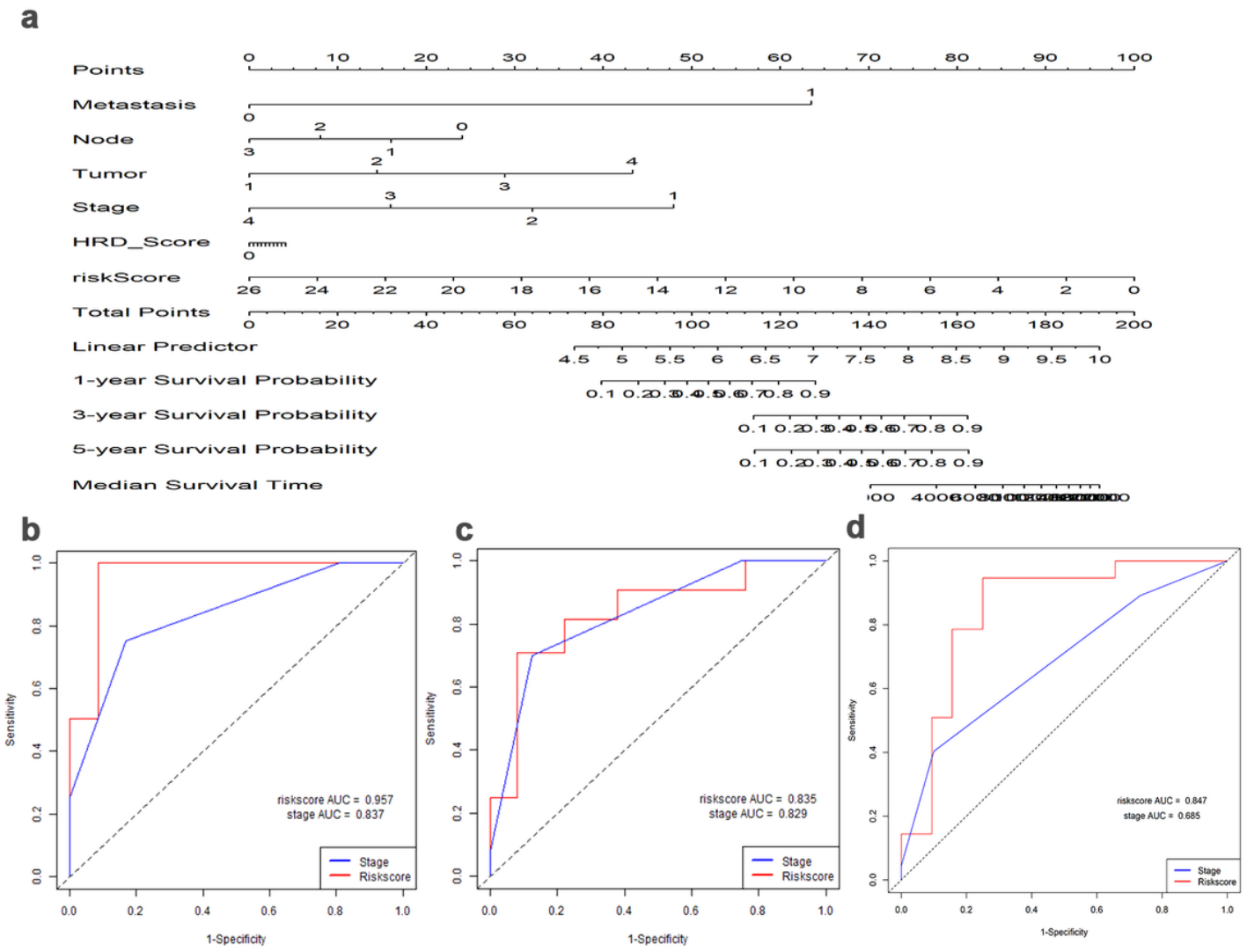

Figure 11

Construction of a nomogram for overall survival prediction in TNBC. a the nomogram consists of AJCC-stage, HRD score, T stage, N stage and the risk score based on the six-gene signature. b-d ROC curves for predicting 1, 3 and 5-year overall survival between 6-gene score and AJCC-stage.

Supplementary Files

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

- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable4.xlsx