Am J Transl Res 2022;14(7):4515-4531
www.ajtr.org /ISSN:1943-8141/AJTR0142415

Original Article

Constructing a novel prognostic signature of tumor driver genes for breast cancer

Liqiang Zhou[1*], Yali Yi[2*], Chuan Liu[3], Zhiqing Chen[3]

[1]Department of General Surgery, The Second Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi, China; [2]Department of Oncology, The Second Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi, China; [3]Key Laboratory of Molecular Medicine of Jiangxi Province, The Second Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi, China. [*]Equal contributors.

Received February 17, 2022; Accepted May 27, 2022; Epub July 15, 2022; Published July 30, 2022

Abstract: Objectives: To systematically explore the function and prognostic ability of tumor-driver genes (TDGs) in breast carcinoma (BRCA). Methods: Functional enrichment analysis of BRCA differentially expressed TDGs was assesed. We used univariate Cox, lasso, and multivariate Cox regression to identify the independent prognostic TDGs of BRCA. Then we constructed a prognostic signature and verified its predictive performance. Gene set enrichment analysis of the signal pathway revealed the differences between the prognostic signature high- and low-risk groups. Finally, a nomogram related to the prognostic model was established and verified. Results: A total of 595 differentially expressed TDGs were identified, which are related to various molecular mechanisms of BRCA progression. We identified 8 independent prognostic TDGs for BRCA and validated their expression and prognosis with public data and clinical samples. The BRCA cohort was divided into training and validation cohorts, and prognostic signatures were constructed separately. The log-rank test showed that the survival rate of the high-risk group was significantly lower than that of the low-risk group in the prognostic signature ($P<0.001$); the AUC in the three cohorts were 0.805, 0.712, and 0.760, respectively; the nomogram also showed better predictive performance. Analyzing the difference between the two risk subtypes, the high-risk group is mainly enriched in angiogenesis, MTORC1, epithelial-mesenchymal transition and glycolysis, which means it is highly malignant. Conclusions: The prognostic signature and nomogram was confirmed to accurately predict the prognosis of patients with BRCA and we validated the hub genes, suggesting their potential as future therapeutic targets.

Keywords: Breast cancer, tumor drive gene, prognostic signature, nomogram

Introduction

Breast carcinoma (BRCA) is the most common malignant tumor in women and one of the main causes of tumor deaths in women. According to statistics for 2020, there are approximately 2.26 million new cases of BRCA in women worldwide and 680,000 deaths, far exceeding other cancer types in women, replacing lung cancer, thus becoming the main type of cancer in the world [1]. Therefore, it is crucial to achieve early diagnosis and treatment for BRCA. In recent years, massive research on oncogenes and tumor suppressor genes, various heterologous proteins and tumor antigens, has made certain progress in the development and application of drugs for specific tumor markers [2]. A series of evidence has indicated that molecular targeted therapy is a promising research direction for cancer treatments [3]. High-throughput sequencing combined with bioinformatics to analyze genomics data contributed to the exploration of markers that are related to the diagnosis, treatment and prognosis of malignant tumors from a molecular perspective [4]. The prediction of patients' survival rate through molecular markers is helpful to provide individualized decision-making for patients is the clinic.

With the deepening of exploration, researchers gradually define tumors as "genomic diseases", that is, tumors are the result of the continuous accumulation of mutations in the tumor cell's genome [5]. Among tumor cell mutations, only a small part plays an important role in tumor

occurrence and progression. These mutations are called driver mutations, and these expressed genes that are affected by driver mutations are called driver genes [6]. Among them, Wang et al. identified FAM83H-AS1 as a driver gene for lung adenocarcinoma and also as a target for the treatment of lung cancer [7]. In addition, Qian et al. showed that FMR1 Autosomal Homolog 1 is a novel driver gene in lung cancer and can help to predict the prognosis of patients with lung cancer [8]. Cancer driver genes are involved in the regulation of multiple biological processes such as cell growth, cell cycle and DNA replication [9]. Chromodomain Helicase DNA Binding Protein 1 Like was shown to prevent lipopolysaccharide-induced hepatocellular carcinoma cell death [10]. Paired box 4 can inhibit the expression of A disintegrin and metalloproteinase (ADAMs) to regulate epithelial cell carcinoma metastasis [11]. Fibroblast growth factor 19 participates in the self-renewal of liver cancer stem cells and promotes the progression of liver cancer cells [12]. However, in current breast cancer research, the function and prognostic power of driver genes has not been systematically analyzed.

This study integrated the data of Genotype-Tissue Expression (GTEx) database [13] and The Cancer Genome Atlas (TCGA) databases [14], analyzed the expression characteristics of driver genes in breast cancer, and their potential molecular biological functions, and also identified the key genes for independent prognosis. Based on the vital tumor driver genes obtained, a prognostic signature was constructed, and a variety of methods were used to analyze the accuracy of its prognosis. According to the signature, BRCA patients can be divided into two subtypes. We analyzed the clinicopathological characteristics and the differences among the signal pathways between the two subtypes. In addition, we applied the nomogram to visualize the prognostic signature, which can intuitively help clinicians make precise and individualized treatment decisions.

Materials and methods

Identify differentially expressed tumor driver genes in BRCA

Transcript per million (TPM) data of normal breast tissue gene expression from GTEx was downloaded. The Fragments Per Kilobase Million (FPKM) sequencing data of BRCA and normal breast tissue in TCGA was downloaded and converted into TPM data. It was then run in the "sva" R package to integrate the two datasets, remove batch effects and perform background correction, and merge them into one data set. The relevant information of tumor driver genes was obtained from the network of cancer genes home (NCG, http://ncg.kcl.ac.uk/) [15], and we extracted the expression of driver genes in each sample from the fusion dataset. Subsequently, using |log2 Fold Change (FC)|>1, and False Discovery Rate (FDR)<0.05 as the screening condition, the "limma" R package [16] identified differentially expressed tumor driver genes (DETDEs) in BRCA. The "ggplot2" R package [17] was used to draw volcano plots and heat maps for visualization.

Gene ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG) function analysis

GO annotations provide a consistent description of gene function, help to develop a controllable vocabulary, and gives information that are non-species specific. It includes cellular component (CC), molecular function (MF) and biological process (BP). The KEGG is a comprehensive database that integrates genomic, chemical, and systemic functional information. To understand the function of DETDEs in BRCA and the molecular mechanisms involved, we performed GO and KEGG function enrichment analysis. With P<0.05 and FDR<0.05 as the screening conditions, the "clusterProfiler" R package [18] was used for enrichment analysis of this DETDEs, and the "GOplot" R packages [19] to visualize the obtained TOP 10 items for each section.

Identify independent prognostic hub tumor drive genes

We first used the "survival" R package to perform univariate Cox regression analysis of DETDEs in the TCGA-BRCA cohort. With P<0.05 set to consider the prognostic-related tumor driver genes. Then we used the "glmnet" package to perform Least Absolute Shrinkage and Selection Operator (Lasso) regression to eliminate the multicollinearity between the prognostic-related tumor driver genes and obtained tumor driver genes that are significantly related

to the prognosis [20]. In addition, we randomly divided all TCGA-BRCA samples with clinical information into a training cohort and testing cohort. We used the "survival" R package to perform multivariate Cox regression analysis on the tumor driver genes obtained in the previous step in the training cohorts and obtained hub tumor driver genes that independently predict the prognosis of breast carcinoma.

Establishment and validation of prognostic signatures of tumor driver genes

In the training cohort, we established a novel prognostic model based on the obtained independence prognostic tumor driver genes. Using multivariate Cox regression analysis, by combining the coefficient (β) and expression (EXP) of each gene, according to the formula: Risksocres = β1 * EXP1 + β2 * EXP2 + ... + βi * EXPi, the risk-score of each patient was calculated separately. According to the median of the scores, we divided all the patients in the training cohort into high- and low-risk groups. Further, we drew the Kaplan Meier (K-M) curve and performed Log-Rank test to analyze the differences in survival rate between the high and low risk groups. The 5-year receiver operating curve (ROC) was used to assess the accuracy of the prognostic model for predicting the 5-year survival rate of breast cancer patients. Combining clinicopathological information and risk scores, univariate and multivariate Cox regression analysis verified the independent prognostic ability of risk-scores for breast cancer. For the verification cohort, we used the same method as the training cohort to construct and verify the prognostic signature. We also integrated the training and verification cohorts to form a complete cohort for re-verification.

Clinicopathological characteristics and molecular mechanism analysis of different risk groups

To understand the potential differences between the two risk groups, we conducted further analysis in the TCGA-BRCA complete cohort. First, we performed a dimensionality reduction analysis of the entire cohort using principal component analysis and analyzed differences between the two risk groups. Subsequently, we performed a chi-square test to find the differences in clinicopathological characteristics between the two risk groups. Furthermore, we used gene set enrichment analysis (GSEA) [21] to enrich the high- and low-risk groups and analyzed the signal pathways that promote tumor progression in the high-risk group. Hallmark 7.4 was chosen as the reference gene set, and $P<0.05$ and $FDR<0.05$ are regarded as significant.

Construction and verification based on tumor driver gene nomogram

We drew a nomogram based on the independent prognostic factors identified in the complete cohort in the previous step ($P<0.05$) to help clinicians make accurate decisions about BRCA patients. By calculating the risk scores of BRCA patients and the corresponding age, pathological stage and pharmaceutical status, the corresponding scores were obtained, and the survival rate of the patients was calculated via the total score. Then we drew 5-year and 10-year calibration curves to evaluate the predictive performance of the nomogram. Among them, the slope of the curve tends close to 1 is considered to an excellent predictive ability.

Hub tumor driver gene expression and prognostic verification in public database

For the obtained hub-independent prognosis tumor driver genes, we used multiple public databases to verify their expression and prognosis respectively. We first used the immunohistochemical data of hub tumor driver genes in the human protein atlas (HPA) database (https://www.proteinatlas.org) [22] to verify their protein expression differences in BRCA. Subsequently, we obtained the Paul A. Northcott dataset from the oncomine database (https://www.oncomine.org) [23], and extracted the expression data of hub tumor driver genes to verify the expression of RNA. Kaplan-Meier Plotter (http://kmplot.com) [24] includes prognostic information of multiple breast cancers, which contribute to elaborate on the relationship between genes and prognosis and draw Kaplan-Meier curves to verify the prognostic ability of hub tumor driver genes.

Cell culture and rt-qPCR detection

To verify the expression differences of hub tumor driver genes in BRCA, we carried out experimental verification. Human normal mam-

Table 1. Primer sequences used in this study

ID	Forward primer (5'-3')	Reverse primer (5'-3')
ACTB	CACCATTGGCAATGAGCGGTTC	AGGTCTTTGCGGATGTCCACGT
CHST1	ATTGATCTCGGGGTCCATCTG	GTCCTGCAATCACACACAGAG
LEF1	TGCCAAATATGAATAACGACCCA	GAGAAAAGTGCTCGTCACTGT
LCP1	GATCAGTGTCCGATGAGGAAATG	CCAGATCACCTGTAGCCATCA
FLT3	CTGAATTGCCAGCCACATTTTG	GGAACGCTCTCAGATATGCAG
SAV1	GTGCTCCTAGTGTACCTCGGT	CTCGTGCGTAAACCTGAAGC
EZR	ACCAATCAATGTCCGAGTTACC	GCCGATAGTCTTTACCACCTGA
ABCC9	TCAACCTGGTCCCTCATGTCT	CAGGAGAGCGAATGTAAGAATCC
MERTK	CTCTGGCGTAGAGCTATCACT	AGGCTGGGTTGGTGAAAACA

mary epithelial cell line MCF10A and breast cancer cell line MCF7 were obtained from the Chinese Academy of Sciences, and were cultured in DMEM medium (Gibico, USA) containing 10% fetal bovine serum (Gibico, USA) at 37°C and 5% CO_2. We extracted total RNA from MCF10A and MCF7 cell lines according to the method of Trizol reagent (Thermo Fisher, USA). Subsequently, according to the instruction manual of the RR047A reverse transcription kit (Takara, Japan), the gDNA of the total RNA was removed and the RNA was reverse transcribed into cDNA. Finally, we prepared the reaction system according to the instructions of the real-time quantitative qPCR kit RR820A (Takara, Japan), and used β-Actin as the internal reference gene to detect the difference in the expression of hub tumor driver genes in breast cancer in the 7900HT system (Applied Biosystems, USA). The primers used were synthesized by Sangon Biotech (Shanghai) Co., Ltd., Shanghai, China, and the primer sequences were shown in **Table 1**.

Results

A total of 595 DETDGs were identified in breast cancer

We downloaded 459 normal breast tissue samples from GTEx and obtained data from 113 normal breast tissues and 1,109 breast carcinoma tissues from TCGA. After data conversion, background removal, batch effect removal, and background correction, we obtained a sequencing data matrix containing information from 572 normal breasts and 1,109 breast cancers. After the tumor gene names were obtained from NCG, we extracted the expression data of 2,595 tumor-drive genes from the integrated data matrix. According to the screening conditions of "limma", we identified 595 dysregulated tumor driver genes of breast carcinoma, including 268 that were up-regulated and 327 that were down-regulated (**Figure 1A**). We visualized the expression of DETDEGs in both normal tissue and breast cancer with a heat map (**Figure 1B**).

The functional enrichment analysis of GO and KEGG of DETDGs

We conducted enrichment analysis on the obtained DETDGs to explore their functions and the molecular mechanisms involved. Among them, the BP of GO is mainly enriched in: gland development, protein kinase B signaling (AKT), peptidyl-tyrosine modification, peptidyl-tyrosine phosphorylation, urogenital system development, epithelial tube morphogenesis, inositol lipid-mediated signaling, phosphatidylinositol-mediated signaling, phosphatidylinositol 3-kinase signaling (PI3K), and morphogenesis of a branching epithelium (**Figure 2A**). The CC of GO is mainly enriched in: banded collagen fibril, fibrillar collagen trimer, collagen-containing extracellular matrix, membrane region, membrane microdomain, membrane raft, cell-substrate junction, cell-cell junction, complex of collagen trimers, and condensed chromosome (**Figure 2B**). The MF of GO is mainly enriched in: growth factor binding, protein tyrosine kinase activity, transmembrane receptor protein tyrosine kinase activity, transmembrane receptor protein kinase activity, extracellular matrix structural constituent, protein phosphatase binding, cytokine binding, phosphatase binding, and platelet-derived growth factor binding (**Figure 2C**). For KEGG, the main enriched ones were Rap1 signaling pathway, Ras signaling pathway, Transcriptional misregulation in cancer, Proteoglycans in cancer, mitogen-activated protein kinase (MAPK) signaling pathway, and PI3K/AKT signaling pathway (**Figure 2D**).

We identified 8 independent prognostic hub tumor driver genes

In TCGA-BRCA, a total of 1090 samples had overall survival information. Through univariate Cox regression analysis, we identified 28 tumor driver genes associated with prognosis from

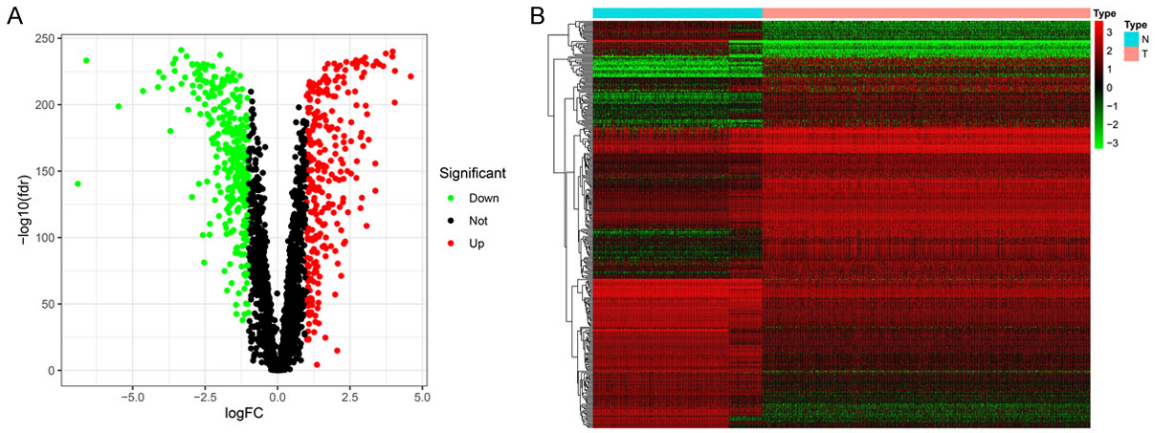


Figure 1. Identification of 595 differentially expressed drive genes in breast cancer. A. Volcano plots show differentially expressed driver genes, with green dots representing 327 down-regulated genes and red dots representing 268 up-regulated genes. B. Volcano plot showing the 535 differentially expressed tumor driver gene expression.

596 DETDEs. Among them, there were 9 tumor driver genes with hazard ratios (HR)>1, and 19 driver genes with HR<1 (**Figure 3A**). Then, we performed Lasso regression analysis to remove the collinearity of each tumor driver gene, from which 21 candidate genes were identified (**Figure 3B**). Then, we randomly divided breast cancer samples with prognostic information into a training cohort (n=546) and a validation cohort (n=544). In the training cohort, multivariate Cox regression analysis identified 8 hub independent prognostic tumor driver genes from the candidate genes. Among them (**Figure 3C**), we found a HR>1 for MER Receptor Tyrosine Kinase (MERTK), ATP Binding Cassette Subfamily C Member 9 (ABCC9), Carbohydrate Sulfotransferase 1 (CHST1), and Ezrin (EZR); and a HR<1 for Salvador Family WW Domain Containing Protein 1 (SAV1), Fms Related Receptor Tyrosine Kinase 3 (FLT3), Lymphocyte Cytosolic Protein 1 (LCP1), and Lymphoid Enhancer Binding Factor 1 (LEF1).

We established and verified the prognostic signature of tumor driver genes

Based on the obtained 8 hub tumor driver genes, we performed multivariate Cox regression analysis to establish a prognostic model. According to the formula: Risk-score = (0.376 * EXP_{MERTK}) + (0.393 * EXP_{ABCC9}) + (-0.691 * EXP_{SAV1}) + (0.153 * EXP_{CHST1}) + (0.399 * EXP_{EZR}) + (-0.210 * EXP_{FLT3}) + (-0.271 * EXP_{LCP1}) + (-0.333 * EXP_{LEF1}), we calculated the risk-score of each sample. In terms of the median risk-score of the training cohort, we divided the samples into high- and low-risk groups. In the validation cohort, risk grouping is evaluated through the same method. We merged the training and validation cohort to form a TCGA-BRCA complete cohort for revalidation. First, the K-M curves in the three cohorts indicated that there are significant differences in survival rates between different risk groups, and the low-risk group reflects a better survival status (**Figure 4A**, **4E**, **4I**, P<0.0001). The high-risk group had higher mortality and shorter survival time in the complete cohort (**Figure 4K**). The area under curve (AUC) of the five-year ROC curves calculated for the three cohorts were 0.805, 0.712, and 0.760, respectively, which reflected the intermediate prognostic ability of the prognostic signature (**Figure 4B**, **4F**, **4J**). Univariate and multivariate Cox regression analysis determined that risk-score in the three cohorts is an independent prognostic factor for BRCA patients (**Figure 4C**, **4D**, **4G**, **4H**, **4L**, **4M**).

High-risk BRCA samples are more malignant

We performed a 3D principal component analysis for dimensionality reduction of the two risk groups in the complete cohort. The results are shown in **Figure 5A**, the high and low risk groups have significant differences, which was regarded as different subtypes of BRCA. Subsequently, we analyzed the association of the two risk groups with the clinicopathological features of BRCA. The analysis results of the chi-square test showed (**Figure 5B**) that risk grouping was correlated with N stage, T stage, patho-

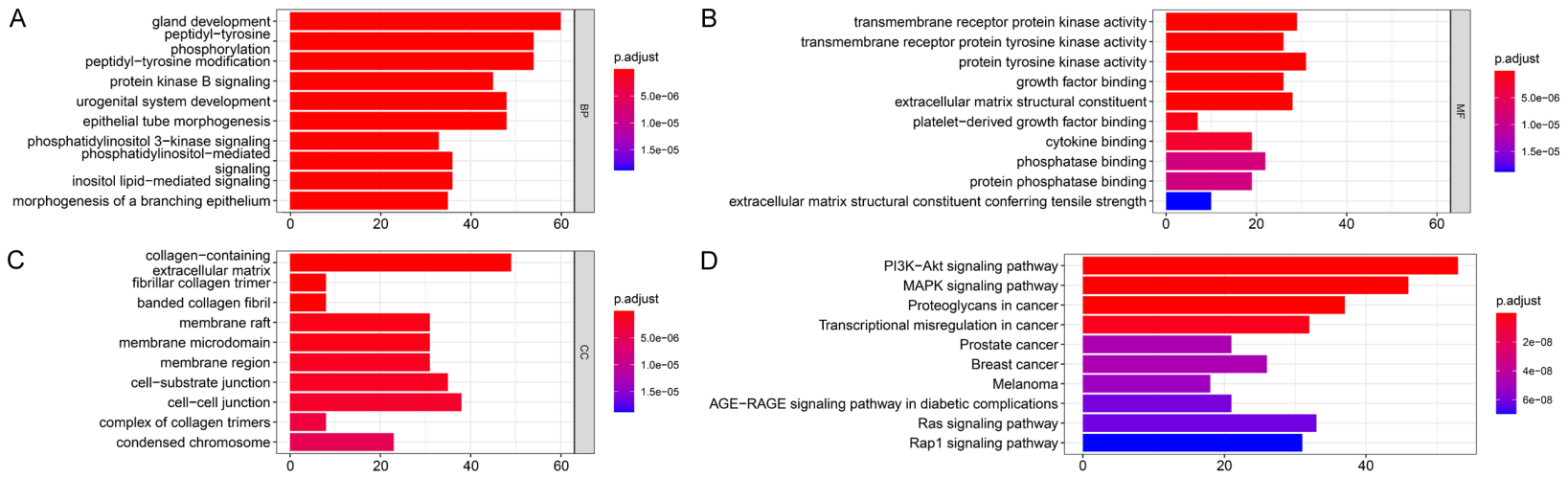


Figure 2. Functional enrichment analysis of the 535 differentially expressed driver genes, Top 10 results for each section. A. Cellular components of gene ontology. B. Molecular function of gene ontology. C. Biological process of gene ontology. D. Kyoto encyclopedia of genes and genomes function enrichment analysis.

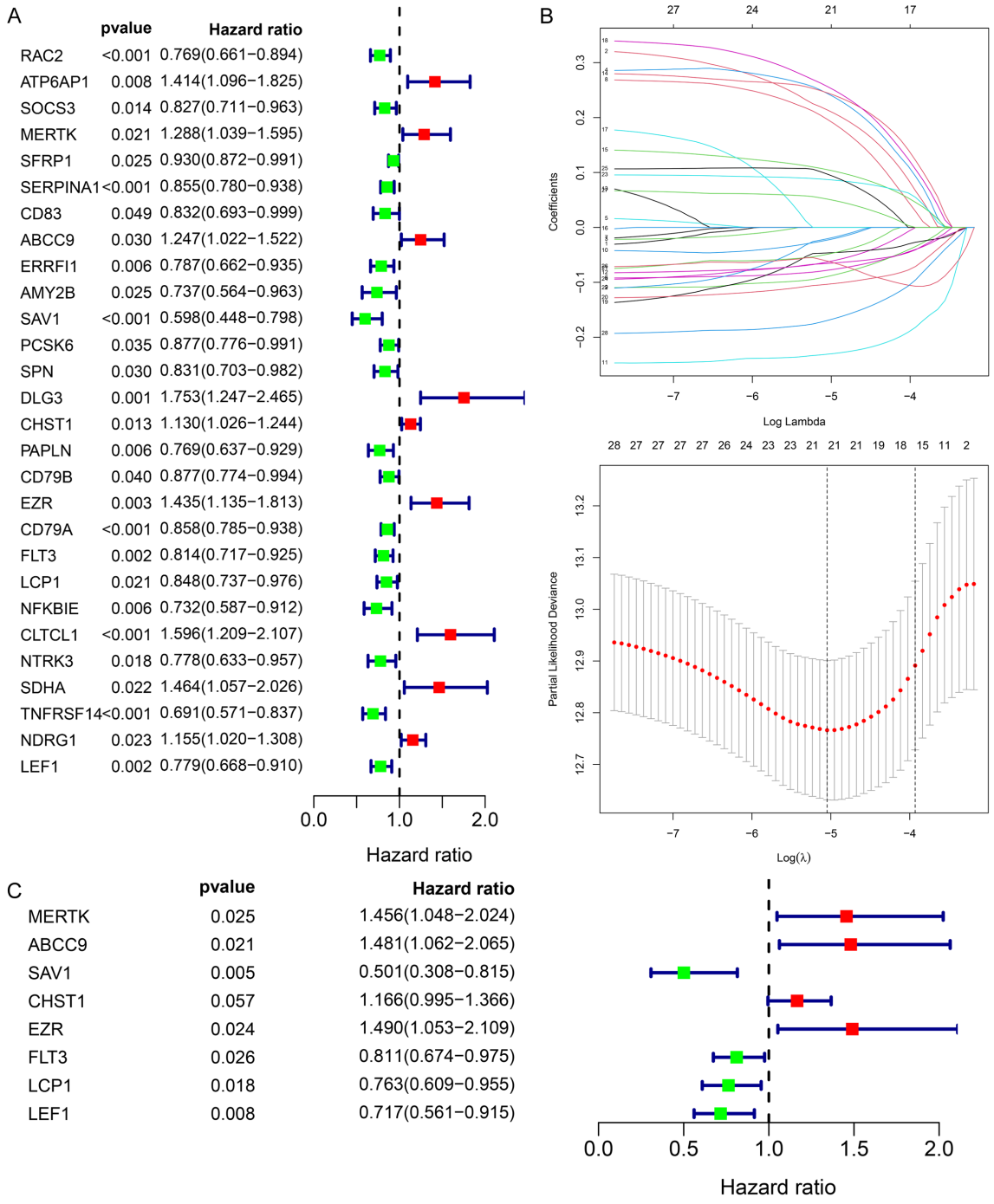
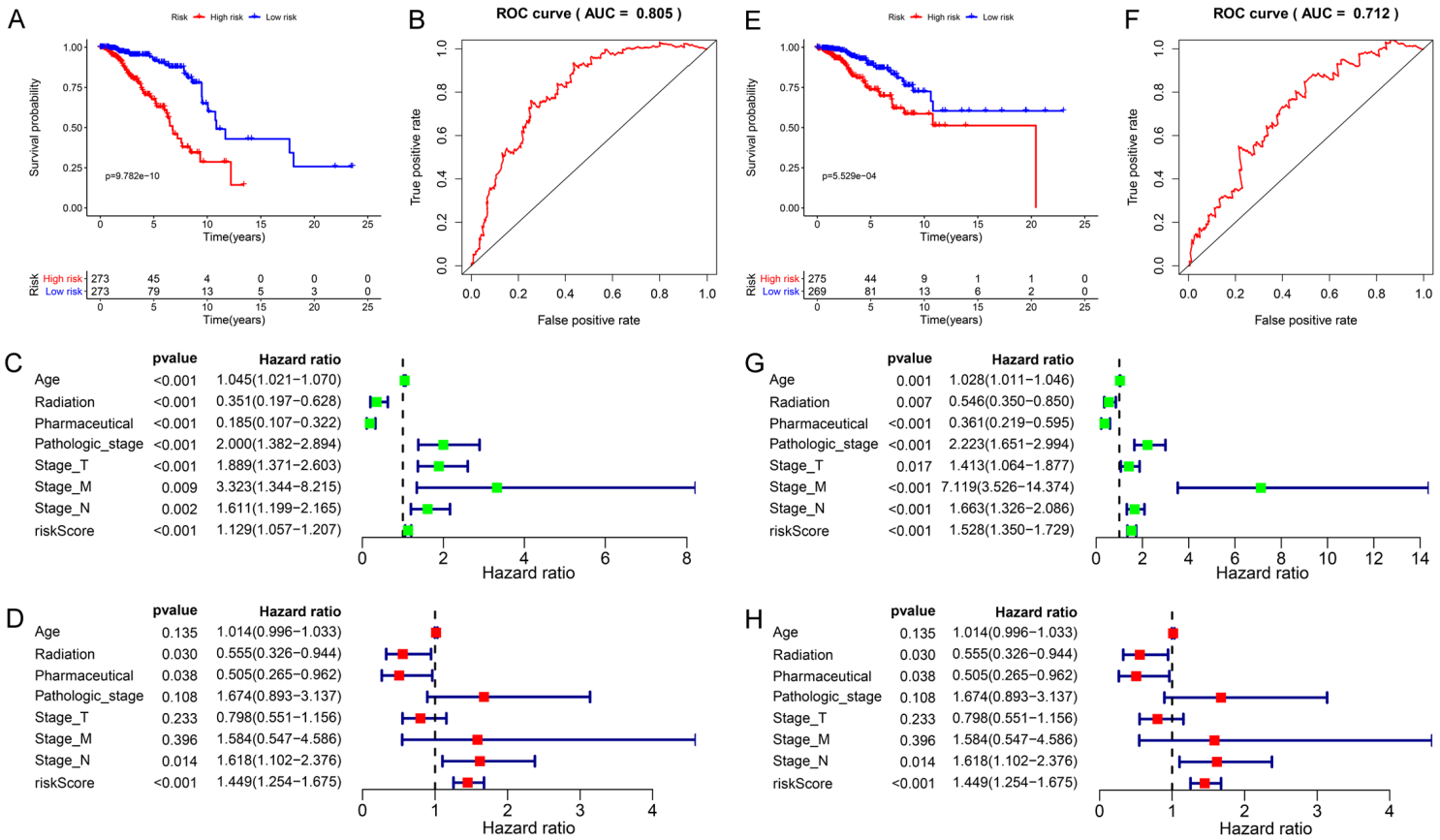
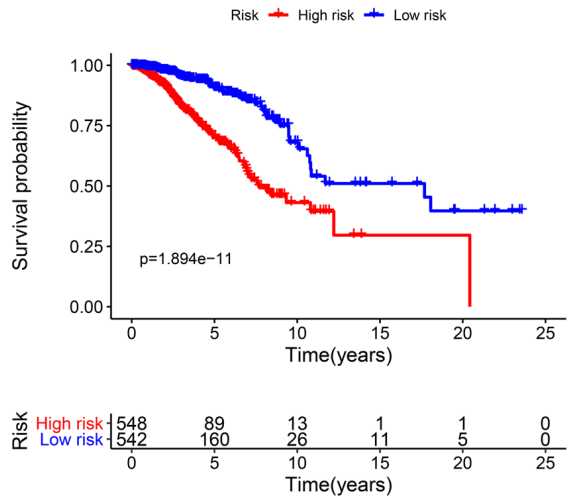


Figure 3. Identification of the 8 breast cancer hub independent prognostic drive genes. A. Univariate Cox regression analysis screened 28 prognostic related tumor driver genes. B. Lasso regression analysis identified 21 candidate genes from 28 prognostic related genes. C. Multivariate Cox regression analysis identified 8 hub-independent prognostic breast cancer driver genes from candidate genes.

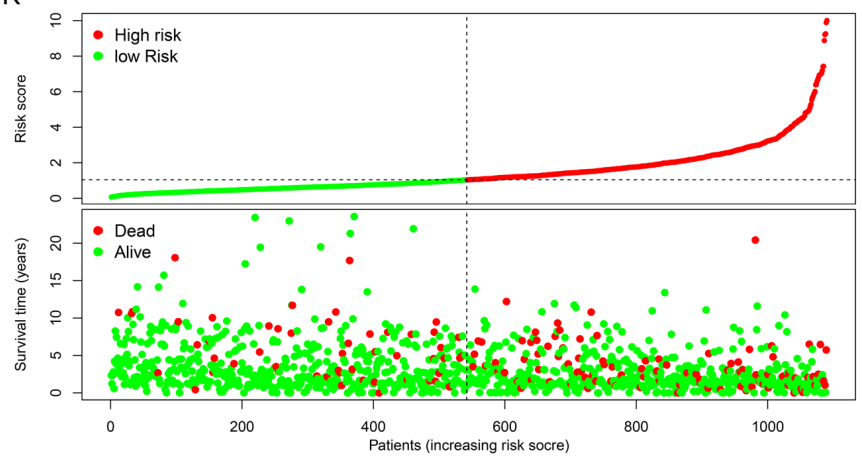
logical stage, and age (P<0.01). We performed gene set enrichment analysis to find the differences in molecular mechanisms between high and low risk groups. The high-risk groups are mainly enriched in angiogenesis, glycolysis, Mammalian target of rapamycin complex 1 (MTORC1), PI3K/AKT/MTOR, epithelial mesenchymal transition (EMT) signaling pathway (**Figure 5C**). The above signals are closely related to malignant biological processes such as


A
Risk
High risk
Low risk
Survival probability
p=9.782e−10
Time(years)
High risk 273 45 4 0 0 0
Low risk 273 79 13 5 3 0
B
ROC curve (AUC = 0.805)
True positive rate
False positive rate
E
Risk
High risk
Low risk
Survival probability
p=5.529e−04
Time(years)
High risk 275 44 9 1 1 0
Low risk 269 81 13 6 2 0
F
ROC curve (AUC = 0.712)
True positive rate
False positive rate
C
pvalue Hazard ratio
Age <0.001 1.045(1.021−1.070)
Radiation <0.001 0.351(0.197−0.628)
Pharmaceutical <0.001 0.185(0.107−0.322)
Pathologic_stage <0.001 2.000(1.382−2.894)
Stage_T <0.001 1.889(1.371−2.603)
Stage_M 0.009 3.323(1.344−8.215)
Stage_N 0.002 1.611(1.199−2.165)
riskScore <0.001 1.129(1.057−1.207)
Hazard ratio
G
pvalue Hazard ratio
Age 0.001 1.028(1.011−1.046)
Radiation 0.007 0.546(0.350−0.850)
Pharmaceutical <0.001 0.361(0.219−0.595)
Pathologic_stage <0.001 2.223(1.651−2.994)
Stage_T 0.017 1.413(1.064−1.877)
Stage_M <0.001 7.119(3.526−14.374)
Stage_N <0.001 1.663(1.326−2.086)
riskScore <0.001 1.528(1.350−1.729)
Hazard ratio
D
pvalue Hazard ratio
Age 0.135 1.014(0.996−1.033)
Radiation 0.030 0.555(0.326−0.944)
Pharmaceutical 0.038 0.505(0.265−0.962)
Pathologic_stage 0.108 1.674(0.893−3.137)
Stage_T 0.233 0.798(0.551−1.156)
Stage_M 0.396 1.584(0.547−4.586)
Stage_N 0.014 1.618(1.102−2.376)
riskScore <0.001 1.449(1.254−1.675)
Hazard ratio
H
pvalue Hazard ratio
Age 0.135 1.014(0.996−1.033)
Radiation 0.030 0.555(0.326−0.944)
Pharmaceutical 0.038 0.505(0.265−0.962)
Pathologic_stage 0.108 1.674(0.893−3.137)
Stage_T 0.233 0.798(0.551−1.156)
Stage_M 0.396 1.584(0.547−4.586)
Stage_N 0.014 1.618(1.102−2.376)
riskScore <0.001 1.449(1.254−1.675)
Hazard ratio

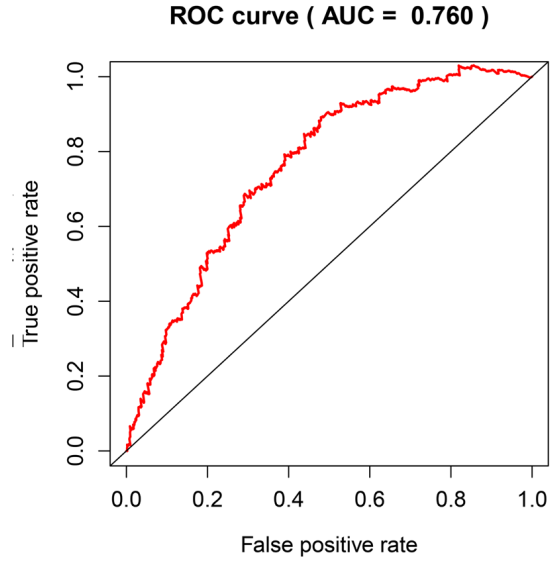
I


Risk
High risk
Low risk
Survival probability
1.00
0.75
0.50
0.25
0.00
p=1.894e−11
0
5
10
15
20
25
Time(years)
Risk
High risk 548 89 13 1 1 0
Low risk 542 160 26 11 5 0
0
5
10
15
20
25
Time(years)

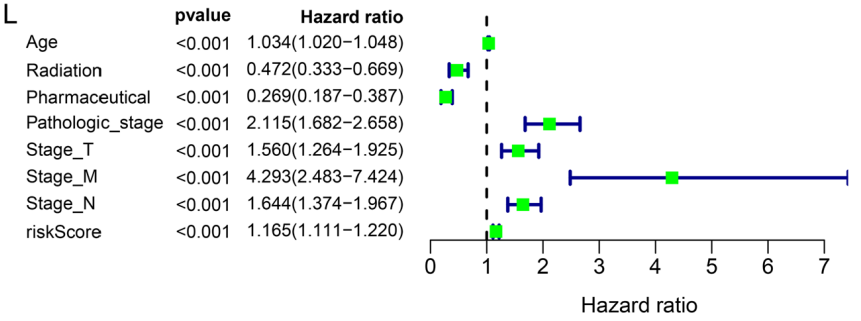
K


High risk
low Risk
Risk score
0
2
4
6
8
10
Dead
Alive
Survival time (years)
0
5
10
15
20
0
200
400
600
800
1000
Patients (increasing risk socre)

J


ROC curve (AUC = 0.760)
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

L


pvalue
Hazard ratio
Age <0.001 1.034(1.020−1.048)
Radiation <0.001 0.472(0.333−0.669)
Pharmaceutical <0.001 0.269(0.187−0.387)
Pathologic_stage <0.001 2.115(1.682−2.658)
Stage_T <0.001 1.560(1.264−1.925)
Stage_M <0.001 4.293(2.483−7.424)
Stage_N <0.001 1.644(1.374−1.967)
riskScore <0.001 1.165(1.111−1.220)
0
1
2
3
4
5
6
7
Hazard ratio

M

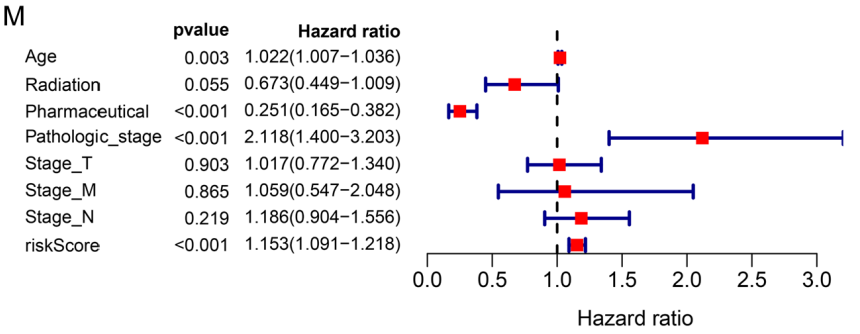

pvalue
Hazard ratio
Age 0.003 1.022(1.007−1.036)
Radiation 0.055 0.673(0.449−1.009)
Pharmaceutical <0.001 0.251(0.165−0.382)
Pathologic_stage <0.001 2.118(1.400−3.203)
Stage_T 0.903 1.017(0.772−1.340)
Stage_M 0.865 1.059(0.547−2.048)
Stage_N 0.219 1.186(0.904−1.556)
riskScore <0.001 1.153(1.091−1.218)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Hazard ratio

Figure 4. Identification and construction of tumor driver gene-related prognostic signatures. A, E, I. Kaplan-Meier curve showed that the survival rate of the high-risk group was significantly lower than that of the low-risk group. B, F, J. The area under the 5-year subject-operable curve was calculated to assess the predictive performance of the prognostic signature. C, G, L. Univariate Cox regression analysis identifies clinicopathological characteristics, risk scores and the prognosis of breast cancer patients. D, H, M. Univariate Cox regression analysis identified clinicopathological characteristics with risk score independent prognostic performance. K. Distribution map of survival status in the complete cohort.

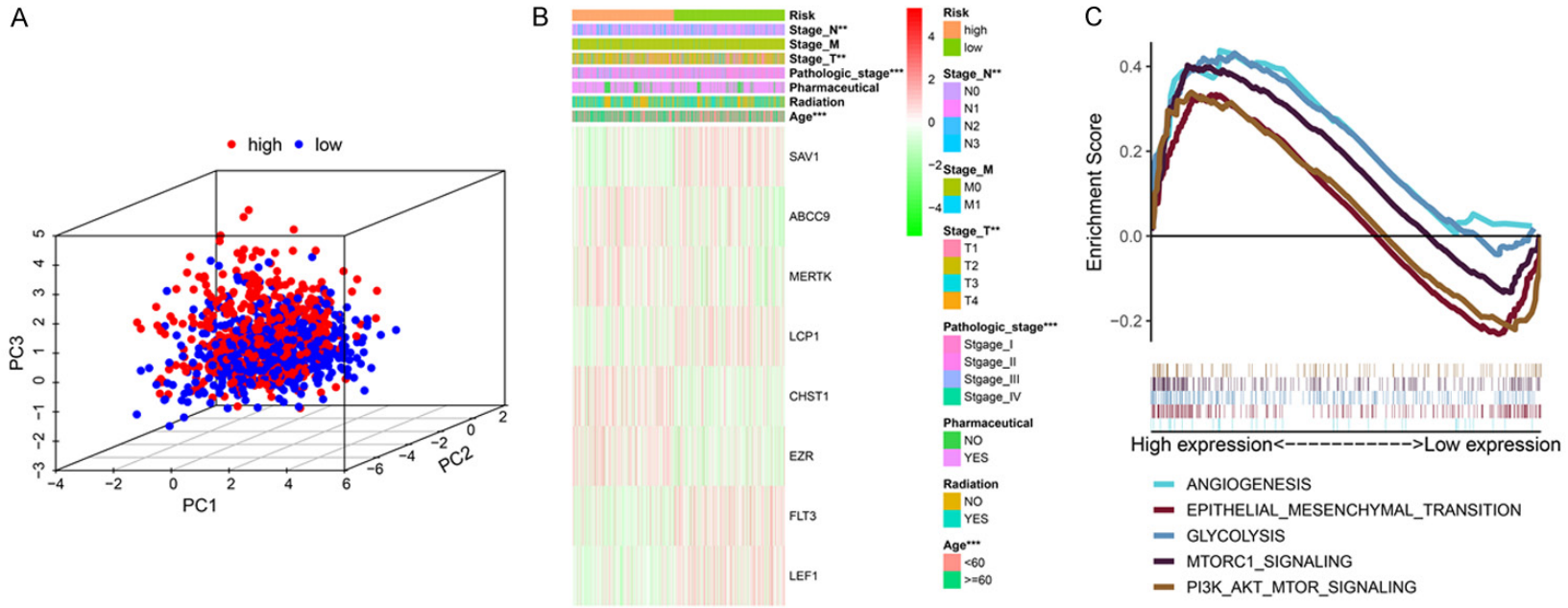


Figure 5. Analysis of the difference between the high and low risk groups. A. Principal component analysis showed significant differences between the two risk groups. B. Risk grouping is related to Nstage, T stage, pathological analysis, and age. C. Gene Set Enrichment Analysis showed that the high-risk group was enriched in angiogenesis, epithelial-mesenchymal transition, glycolysis, MTORC1 signaling, and PI3K/Akt/MTOR signaling.

tumor proliferation, invasion, and metastasis, suggesting that the high-risk group is more malignant.

Nomogram help clinical decision-making

The nomogram is made to visualize the constructed prognostic signature, and the survival rate of COAD patients can be calculated more intuitively. Integrating multiple predictors is in favor of better predictive power. We identified independent prognostic factors from the multivariate Cox regression analysis in the complete cohort, and constructed a nomogram based on this. We included age, pharmaceutical, pathologic stage, and risk scores, and assigned corresponding scores according to each patient's situation. Therefore, the total score we obtained can directly and accurately predict the 1-5 year survival rate of BRCA patients (**Figure 6A**). Subsequently, we plotted 5- and 10-year calibration curves to understand the predictive power of the nomogram. The results showed that the actual predicted survival rates of the two calibration curves were almost identical to the ideal values, demonstrating the excellent predictive performance of the nomogram (**Figure 6B**).

Hub-driver gene expression and prognostic verification

From our analysis in BRCA, a total of 8 hub-independent tumor driver genes were identified, which are potential targets for BRCA treatment. According to the immunohistochemical data from the HPA database, EZR was highly expressed in BRCA, while METRTK, ABCC9, and SAV1 were down-regulation (**Figure 7A**). In the oncomine database, we analyzed the mRNA expression of hub tumor driver genes. The results showed that LEF1, LCP1, FLT3, EZR, and CHST1 were up-regulated in BRCA, while the expression of SAV1, ABCC9, and MERTK were down-regulated (**Figure 7B**). In addition, we analyzed the expression of hub tumor genes in human normal mammary epithelial cell line MCF10A and breast cancer cell line MCF7 by rt-qPCR. The results of the *in vitro* experiments were consistent with those of oncomine database (**Figure 7C**). Subsequently, the relationship between hub-driven gene expression and prognosis was verified. High expression of LCP1, LEF1, SAV1, FLT3 was associated with a better prognosis, and high expression of METRK, ABCC9, CHST1, and EZR were associated with a poor prognosis (**Figure 7D**).

Discussion

In this study, we integrated the data of TCGA and GTEx to identify the differentially expressed tumor driver genes of BRCA. Firstly, we explored the potential molecular biological functions of these genes; the results are mainly enriched in the Ras, MAPK, PI3K/Akt and Rap1 signal pathways. These signals mutually regulate crosstalk and promote tumor progression. Among them, the Ras gene of Ras signaling is a classic tumor driver gene, which is activated to form an oncogene with oncogenic activity, causing cells to proliferate uncontrollably and become malignant [25]. Furthermore, Rap1 is a member of the Ras small GTPases family, which activates extracellular regulated protein kinases independent of Ras in an environment-dependent manner, thereby playing an important role in tumor EMT and metabolic reprogramming [26]. In addition, The N-terminus of the Ras protein can be combined with Raf and is a serine/threonine protein kinase (MAPKKK) that is activated during this period, and further transduces and activates MAPK into the nucleus, which activates various transcription factors [27]. The MAPK signaling pathway consists of four distinct cascades, including extracellular signal-related kinases (ERK1/2), Jun N-terminal kinases (JNK1/2/3), p38-MAPK and ERK5, the above signal activation is related to tumor cell differentiation, migration, senescence and apoptosis [28]. By directly activating p110α and p110δ of PI3K, Ras mediates tumor cell growth, autophagy, and triggers downstream signaling events including Akt [29]. Taken together these signals suggest that these differentially expressed tumor driver genes play crucial roles in BRCA progression.

Subsequently, we used a variety of statistical methods and constructed a prognostic signature based on 8 hubs driver genes. We drew K-M and ROC curves to identify the excellent predictive performance. According to the prognostic model, BRCA samples were divided into high- and low-risk groups. We analyzed the differences in clinical pathological characteristics and molecular pathways. Among them, the risk-group is related to the age, TMN stage and pathological stage of BRCA patients. GSEA showed that high-risk patients are mainly en-

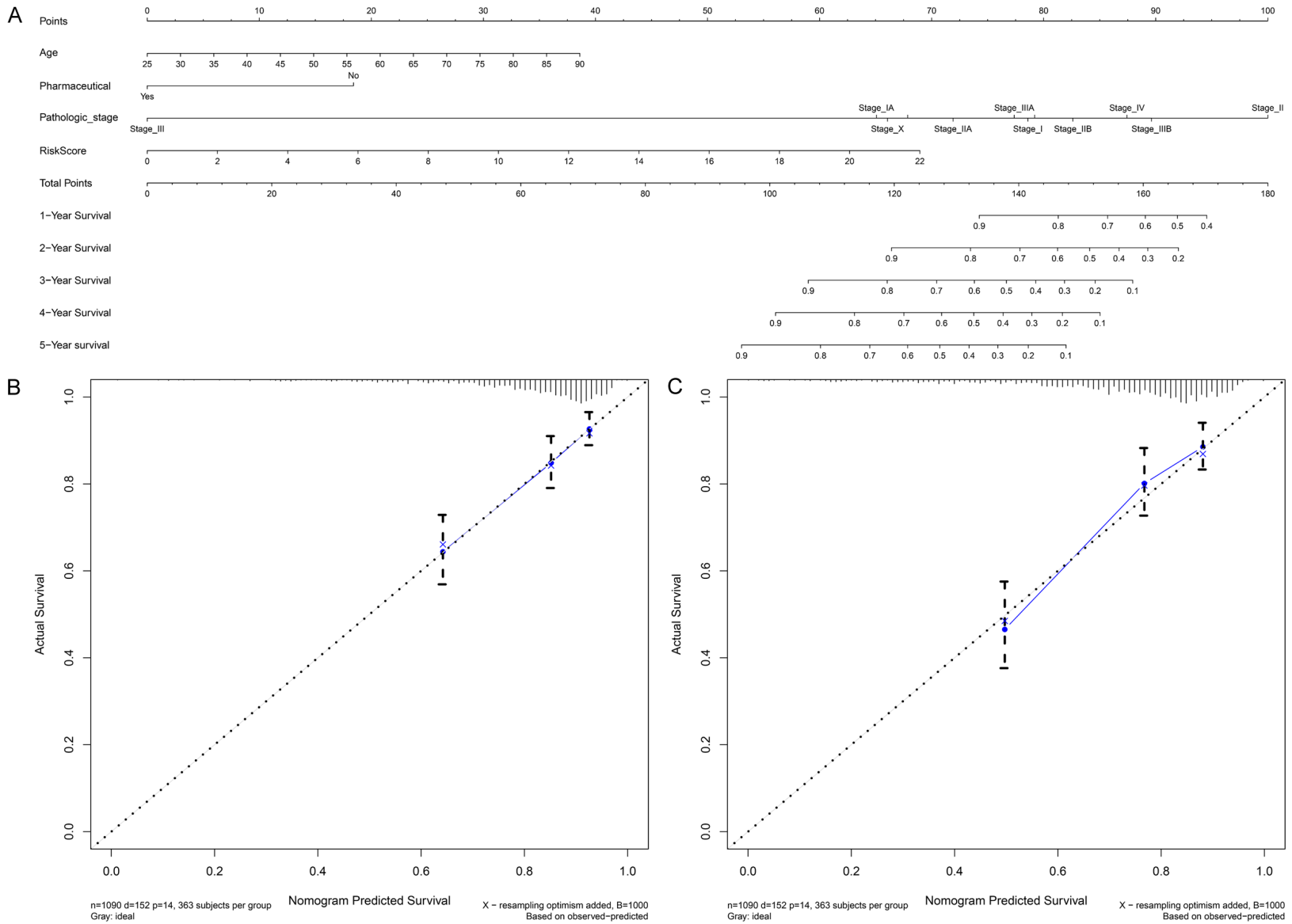


Figure 6. Nomogram construction and verification. A. A nomogram based on the prognostic signature of the driver gene and clinicopathological characteristics. B. 5-year calibration curve validated nomogram. C. 10-year calibration curve validated nomogram.

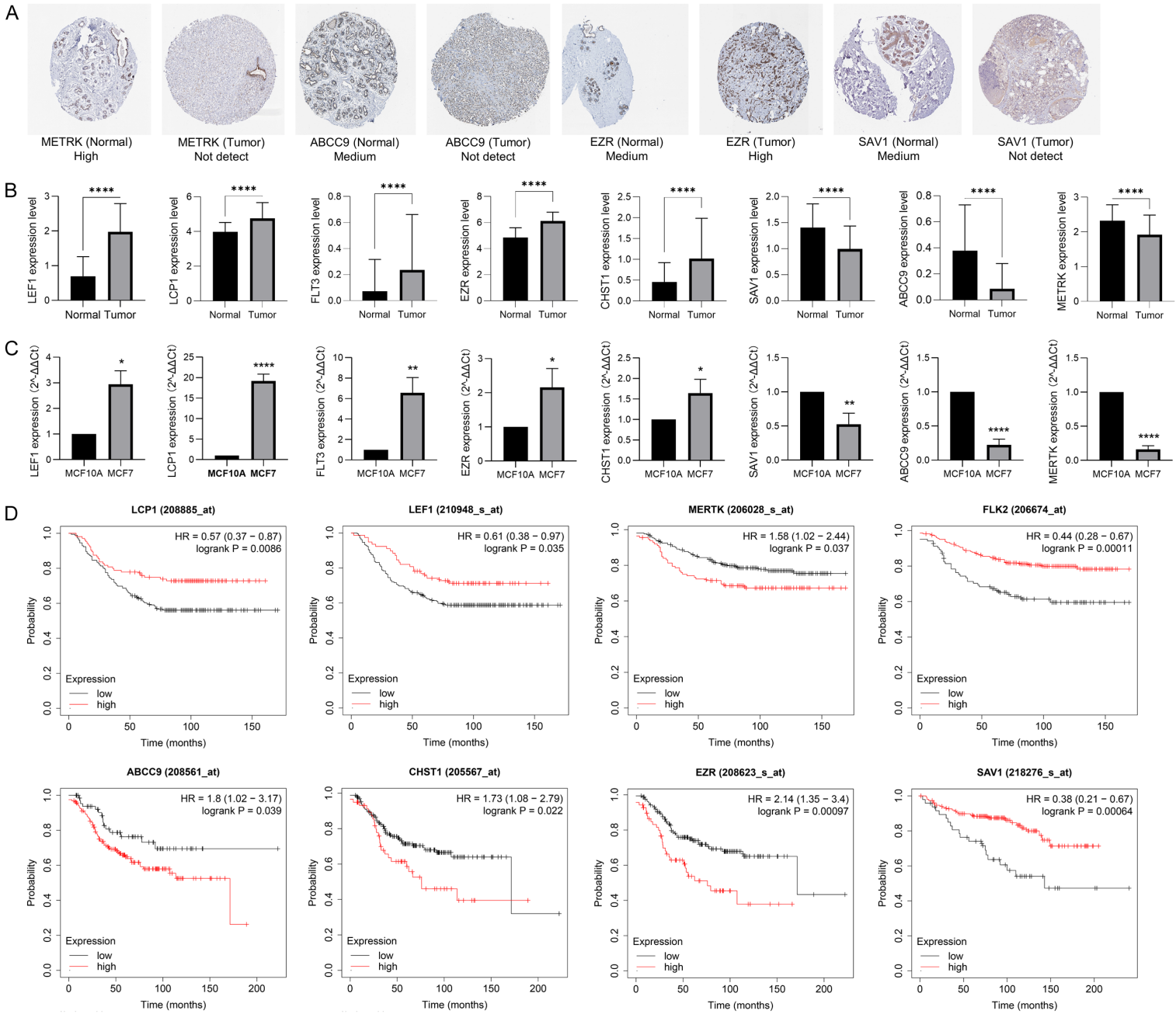

A
METRK (Normal)
High
METRK (Tumor)
Not detect
ABCC9 (Normal)
Medium
ABCC9 (Tumor)
Not detect
EZR (Normal)
Medium
EZR (Tumor)
High
SAV1 (Normal)
Medium
SAV1 (Tumor)
Not detect
B
LEF1 expression level
LCP1 expression level
FLT3 expression level
EZR expression level
CHST1 expression level
SAV1 expression level
ABCC9 expression level
METRK expression level
Normal Tumor
C
LEF1 expression (2^-ΔΔCt)
LCP1 expression (2^-ΔΔCt)
FLT3 expression (2^-ΔΔCt)
EZR expression (2^-ΔΔCt)
CHST1 expression (2^-ΔΔCt)
SAV1 expression (2^-ΔΔCt)
ABCC9 expression (2^-ΔΔCt)
MERTK expression (2^-ΔΔCt)
MCF10A MCF7
D
LCP1 (208885_at)
HR = 0.57 (0.37 − 0.87)
logrank P = 0.0086
LEF1 (210948_s_at)
HR = 0.61 (0.38 − 0.97)
logrank P = 0.035
MERTK (206028_s_at)
HR = 1.58 (1.02 − 2.44)
logrank P = 0.037
FLK2 (206674_at)
HR = 0.44 (0.28 − 0.67)
logrank P = 0.00011
ABCC9 (208561_at)
HR = 1.8 (1.02 − 3.17)
logrank P = 0.039
CHST1 (205567_at)
HR = 1.73 (1.08 − 2.79)
logrank P = 0.022
EZR (208623_s_at)
HR = 2.14 (1.35 − 3.4)
logrank P = 0.00097
SAV1 (218276_s_at)
HR = 0.38 (0.21 − 0.67)
logrank P = 0.00064
Probability
Expression
low
high
Time (months)

Figure 7. Verification of the expression and prognosis of the 8 hub tumor driven genes. A. HPA database verifies MERTK, ABCC9, EZR, SAV1 protein expression. B. Oncomine verifies the 8 hub tumor driver gene mRNA expression. C. The expression of eight hub tumor driver genes in MCF10A and MCF 7 was verified by rt-PCR. D. Kaplan-Meier plotter verifies the prognosis of the 8 hub genes.

riched in EMT, Angiogenesis, MTORC1 and Glycolysis. EMT is a process in which epithelial cells separate from their neighboring cells and acquire the characteristics of interstitial cell migration, which is crucial for initiating the metastatic cascade that allows cancer cells to leave the primary tumor, which causes tumor cells to spread to distant organs [30]. Pastushenko et al. found that FAT1, the most frequently mutated driver gene in multiple tumors, has a tumor suppressor effect, and that the loss of FAT1 function will promote heterozygous EMT, metastasis and drug resistance [31]. Importantly, the growth of tumor tissues must rely on angiogenesis to provide sufficient oxygen and nutrients to maintain growth. The most common driver gene, vascular endothelial growth factor (VEGF), not only directly promotes angiogenesiss, also indirectly stimulates angiogenesis by recruiting tumor-associated macrophages that support angiogenesis and secrete VEGF into the tumor microenvironment [32]. In addition, the energy metabolism of tumor cells has special characteristics. In normal cells, glycolysis is a highly regulated and conserved metabolic process in the cytoplasm, and oxidative phosphorylation is the main energy production process [33]. Although glycolysis is a metabolic method that produces less energy, Warburg confirmed that the conversion rate of glucose to lactic acid in rat liver cancer tissue increased by about 10 times in the presence of oxygen [34]. Chen et al. found that the new ovarian cancer driver genes TBC1D8 and TBC1D8 are amplified in ovarian cancer tissues, which combined with the key rate-limiting enzyme of sugar metabolism in tumor cells, PKM2, and inhibit the tetramerization of PKM2 and the activity of pyruvate metabolizing enzymes, then mediating the metabolic reprogramming of ovarian cancer cells, and ultimately driving the occurrence, development, and invasion of ovarian cancer [35]. The treatment of specific energy metabolism of tumor cells will be an effective anti-cancer strategy [35]. For mTORC1, it regulates mutations in a lot of oncogenic pathways, such as the Ras/Raf/Mek/Erk (MAPK) pathway and the PI3K/Akt pathway, and controls tumor cell proliferation and migration [36]. The above information suggested that BRCA samples in the high-risk group of clinical cases and molecular mechanisms have a higher malignancy and a worse prognosis.

We identified 8 hub-driven genes with independent prognosis, MERTK, ABCC9, CHST1, EZR with a HR>1, which were considered as dangerous genes; and SAV1, FTL3, LCP1, LEF1 with a HR<1, which were considered as protective genes. We have conducted multiple types of verification on the expression and prognosis of these genes, confirming our analysis. MERTK is a TAM tyrosine kinase that participates in multiple biological processes, including cell proliferation, survival, migration and immune regulation, apoptotic cell clearance, platelet aggregation, which leads to the activation of several classic carcinogenic signal pathways [37]. Huang et al. showed that MerTK inhibition in tumor leukocytes reduced the growth and metastasis of breast cancer [38]. MERTK also promotes breast cancer progression by combining oncogenic signals and host anti-tumor immunity evasion [39]. Studies in renal cancer may shed light on the tumor-promoting mechanism of MERTK, and Xu et al. reported that MERTK-mediated phosphorylation of Akt drives tumorigenesis and therapy resistance [40]. ABCC9 is a member of the ABC transporter family, which utilizes the energy of ATP to transport specific substrates and is closely related to the drug resistance of tumors [41]. EZR is a member of the ERM protein family, which acts as an intermediate between the plasma membrane and the actin cytoskeleton [42-44]. This protein plays a key role in the adhesion, migration, and organization of cell surface structures, and it is associated with various human cancers. Zhang et al. showed that the high expression of EZR in breast cancer is associated with poor prognosis [42]. It not only promotes the proliferation of cancer cells, but also promotes drug resistance by anchoring drug-resistant proteins on the cell membrane [40, 43, 44]. Xu et al believe that the mechanism by which EZR promotes tumor progression is through the activation of Akt signaling [45]. For risk genes,

the specific role of CHST1 in tumors has not been reported, and the research evidence of the same protective gene FTL3 is also insufficient. For SAV1, it's a member of the Hippo pathway, and studies proved that it inhibits the proliferation and metastasis of tumor cells and plays a tumor suppressor role [46]. However, LCP1 and LEF1 have shown their cancer-promoting effects in multiple reports. As shown by Nir Pillar and others, inhibiting LCP1 limits the progression of breast cancer [47]. The expression of LEF1 can combine the expression of Homeobox 2 with Slug and zinc finger E-box and MMP7 to promote tumor proliferation and invasion [48, 49]. This may be due to the heterogeneity of tumor cells resulting in changes in the role of genes.

In general, this is the first study to explore the expression characteristics of tumor-drive genes in BRCA, and analyze their potential molecular mechanisms. The potential ability of driver genes to predict breast cancer prognosis was also explored by bioinformatics methods. A prognostic model and nomogram were constructed through multiple validation methods, which confirmed its accurate predictive performance.

Acknowledgements

Natural Science Foundation of Jiangxi Province, Grant/Award Numbers: 20192BAB205079. Special clinical research project of the Second Affiliated Hospital of Nanchang University, Grant/Award Numbers: 2019YNLZ12002.

Disclosure of conflict of interest

None.

Address correspondence to: Chuan Liu and Zhiqing Chen, Key Laboratory of Molecular Medicine of Jiangxi Province, The Second Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi, China. E-mail: lc61116@126.com (CL); czq033-021@163.com (ZQC)

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A and Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021; 71: 209-249.

[2] Pernas S, Barroso-Sousa R and Tolaney SM. Optimal treatment of early stage HER2-positive breast cancer. Cancer 2018; 124: 4455-4466.

[3] Koshiba M. Molecular targeted therapy and laboratory tests. Rinsho Byori 2016; 64: 709-716.

[4] Beane J, Campbell JD, Lel J, Vick J and Spira A. Genomic approaches to accelerate cancer interception. Lancet Oncol 2017; 18: e494-e502.

[5] Weber BL. Cancer genomics. Cancer Cell 2002; 1: 37-47.

[6] Cheng FX, Zhao JF and Zhao ZM. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. Brief Bioinform 2016; 17: 642-656.

[7] Wang SW, Han CC, Liu TY, Ma ZF, Qiu MT, Wang J, You QJ, Zheng XF, Xu WZ, Xia WJ, Xu YT, Hu JW, Xu L and Yin R. FAM83H-AS1 is a noncoding oncogenic driver and therapeutic target of lung adenocarcinoma. Clin Transl Med 2021; 11: e316.

[8] Qian J, Hassanein M, Hoeksema MD, Harris BK, Zou Y, Chen HD, Lu PC, Eisenberg R, Wang J, Espinosa A, Ji XM, Harris FT, Rahman SM and Massion PP. The RNA binding protein FXR1 is a new driver in the 3q26-29 amplicon and predicts poor prognosis in human cancers. Proc Natl Acad Sci U S A 2015; 112: 3469-3474.

[9] Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H, Gonzalez-Perez A and Lopez-Bigas N. A compendium of mutational cancer driver genes. Nat Rev Cancer 2020; 20: 555-572.

[10] Wang GL, Zhang XF, Cheng W, Mo YX, Chen J, Cao ZM, Chen XG, Cui HQ, Liu SS, Huang L, Liu M, Ma L and Ma NF. CHD1L prevents lipopolysaccharide-induced hepatocellular carcinomar cell death by activating hnRNP A2/B1-nmMYLK axis. Cell Death Dis 2021; 12: 891.

[11] Zhang J, Qin X, Sun Q, Guo H, Wu X, Xie F, Xu Q, Yan M, Liu J, Han Z and Chen W. Transcriptional control of PAX4-regulated miR-144/451 modulates metastasis by suppressing ADAMs expression. Oncogene 2015; 34: 3283-3295.

[12] Wang JC, Zhao HK, Zheng L, Zhou Y, Wu L, Xu YQ, Zhang X, Yan GF, Sheng HL, Xin R, Jiang L, Lei J, Zhang JG, Chen Y, Peng J, Chen Q, Yang S, Yu K, Li DS, Xie QC and Li YS. FGF19/SOCE/NFATc2 signaling circuit facilitates the self-renewal of liver cancer stem cells. Theranostics 2021; 11: 5045-5060.

[13] GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet 2013; 45: 580-585.

[14] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C and Stuart JM. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013; 45: 1113-1120.

[15] Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, Yakovleva A, Palmieri T and Ciccarelli FD. The network of cancer genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol 2019; 20: 1.

[16] Ritchie ME, Phipson B, Wu D, Hu YF, Law CW, Shi W and Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.

[17] Ito K and Murphy D. Application of ggplot2 to pharmacometric graphics. CPT Pharmacometrics Syst Pharmacol 2013; 2: e79.

[18] Wu TZ, Hu EQ, Xu SB, Chen MJ, Guo PF, Dai ZH, Feng TZ, Zhou L, Tang WL, Zhan L, Fu XC, Liu SS, Bo XC and Yu GC. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation (Camb) 2021; 2: 100141.

[19] Walter W, Sanchez-Cabo F and Ricote M. GOplot: an R package for visually combining expression data with functional analysis. Bioinformatics 2015; 31: 2912-2914.

[20] Engebretsen S and Bohlin J. Statistical predictions with glmnet. Clin Epigenetics 2019; 11: 123.

[21] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005; 102: 15545-15550.

[22] Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, Edfors F, Sanli K, von Feilitzen K, Oksvold P, Lundberg E, Hober S, Nilsson P, Mattsson J, Schwenk JM, Brunnström H, Glimelius B, Sjöblom T, Edqvist PH, Djureinovic D, Micke P, Lindskog C, Mardinoglu A and Ponten F. A pathology atlas of the human cancer transcriptome. Science 2017; 357: eaan2507.

[23] Rhodes DR, Yu JJ, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A and Chinnaiyan AM. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 2004; 6: 1-6.

[24] Gyorffy B. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. Comput Struct Biotechnol J 2021; 19: 4101-4109.

[25] Moore AR, Rosenberg SC, McCormick F and Malek S. RAS-targeted therapies: is the undruggable drugged? Nat Rev Drug Discov 2020; 19: 533-552.

[26] Shah S, Brock EJ, Ji K and Mattingly RR. Ras and Rap1: a tale of two GTPases. Semin Cancer Biol 2019; 54: 29-39.

[27] Drosten M and Barbacid M. Targeting the MAPK pathway in KRAS-driven tumors. Cancer Cell 2020; 37: 543-550.

[28] Fang JY and Richardson BC. The MAPK signalling pathways and colorectal cancer. Lancet Oncol 2005; 6: 322-327.

[29] Fruman DA, Chiu H, Hopkins BD, Bagrodia S, Cantley LC and Abraham RT. The PI3K pathway in human disease. Cell 2017; 170: 605-635.

[30] Dongre A and Weinberg RA. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. Nat Rev Mol Cell Biol 2019; 20: 69-84.

[31] Liao CC, Wang Q, An JX, Long Q, Wang H, Xiang ML, Xiang ML, Zhao YJ, Liu YL, Liu JG and Guan XY. Partial EMT in squamous cell carcinoma: a snapshot. Int J Biol Sci 2021; 17: 3036-3047.

[32] Fukumura D, Kloepper J, Amoozgar Z, Duda DG and Jain RK. Enhancing cancer immunotherapy using antiangiogenics: opportunities and challenges. Nat Rev Clin Oncol 2018; 15: 325-340.

[33] Wu M, Neilson A, Swift AL, Moran R, Tamagnine J, Parslow D, Armistead S, Lemire K, Orrell J, Teich J, Chomicz S and Ferrick DA. Multiparameter metabolic analysis reveals a close link between attenuated mitochondrial bioenergetic function and enhanced glycolysis dependency in human tumor cells. Am J Physiol Cell Physiol 2007; 292: C125-136.

[34] Warburg O. über den Stoffwechsel der Carcinomzelle. Klin Wochenschr 1925; 4: 534-536.

[35] Cheong JH, Park ES, Liang J, Dennison JB, Tsavachidou D, Nguyen-Charles C, Wa Cheng K, Hall H, Zhang D, Lu Y, Ravoori M, Kundra V, Ajani J, Lee JS, Ki Hong W and Mills GB. Dual inhibition of tumor energy pathway by 2-deoxyglucose and metformin is effective against a broad spectrum of preclinical cancer models. Mol Cancer Ther 2011; 10: 2350-2362.

[36] Saxton RA and Sabatini DM. mTOR signaling in growth, metabolism, and disease. Cell 2017; 168: 960-976.

[37] Cummings CT, Deryckere D, Earp HS and Graham DK. Molecular pathways: MERTK signaling in cancer. Clin Cancer Res 2013; 19: 5275-5280.

[38] Cook RS, Jacobsen KM, Wofford AM, DeRyckere D, Stanford J, Prieto AL, Redente E, Sandahl M, Hunter DM, Strunk KE, Graham DK and Earp HS 3rd. MerTK inhibition in tumor

leukocytes decreases tumor growth and metastasis. J Clin Invest 2013; 123: 3231-3242.
[39] Davra V, Kumar S, Geng K, Calianese D, Mehta D, Gadiyar V, Kasikara C, Lahey KC, Chang YJ, Wichroski M, Gao C, De Lorenzo MS, Kotenko SV, Bergsbaken T, Mishra PK, Gause WC, Quigley M, Spires TE and Birge RB. Axl and mertk receptors cooperate to promote breast cancer progression by combined oncogenic signaling and evasion of host antitumor immunity. Cancer Res 2021; 81: 698-712.
[40] Jiang Y, Zhang YQ, Leung JY, Fan C, Popov KI, Su SY, Qian JY, Wang XD, Holtzhausen A, Ubil E, Xiang Y, Davis I, Dokholyan NV, Wu G, Perou CM, Kim WY, Earp HS and Liu PD. MERTK mediated novel site Akt phosphorylation alleviates SAV1 suppression. Nat Commun 2019; 10: 1515.
[41] Zhang RN, Li SW, Liu LJ, Yang J, Huang GF and Sang Y. TRIM11 facilitates chemoresistance in nasopharyngeal carcinoma by activating the β-catenin/ABCC9 axis via p62-selective autophagic degradation of Daple. Oncogenesis 2020; 9: 45.
[42] Zhang RJ, Zhang SH, Xing RG and Zhang Q. High expression of EZR (ezrin) gene is correlated with the poor overall survival of breast cancer patients. Thorac Cancer 2019; 10: 1953-1961.
[43] Yano K, Okabe C, Fujii K, Kato Y and Ogihara T. Regulation of breast cancer resistance protein and P-glycoprotein by ezrin, radixin and moesin in lung, intestinal and renal cancer cell lines. J Pharm Pharmacol 2020; 72: 575-582.
[44] Konstantinovsky S, Davidson B and Reich R. Ezrin and BCAR1/p130Cas mediate breast cancer growth as 3-D spheroids. Clin Exp Metastasis 2012; 29: 527-540.
[45] Xu J and Zhang W. EZR promotes pancreatic cancer proliferation and metastasis by activating FAK/AKT signaling pathway. Cancer Cell Int 2021; 21: 521.
[46] Lin ZJ, Xie RL, Guan KL and Zhang MJ. A WW tandem-mediated dimerization mode of SAV1 essential for hippo signaling. Cell Rep 2020; 32: 108118.
[47] Pillar N, Polsky AL and Shomron N. Dual inhibition of ABCE1 and LCP1 by microRNA-96 results in an additive effect in breast cancer mouse model. Oncotarget 2019; 10: 2086-2094.
[48] Bucan V, Mandel K, Bertram C, Lazaridis A, Reimers K, Park-Simon TW, Vogt PM and Hass R. LEF-1 regulates proliferation and MMP-7 transcription in breast cancer cells. Genes Cells 2012; 17: 559-567.
[49] Huang FI, Chen YL, Chang CN, Yuan RH and Jeng YM. Hepatocyte growth factor activates Wnt pathway by transcriptional activation of LEF1 to facilitate tumor invasion. Carcinogenesis 2012; 33: 1142-1148.