

Key Genes and Prognostic Analysis in HER2+ Breast Cancer

Yujie Weng, MS^{1,*}, Wei Liang, MS^{1,*}, Yucheng Ji, MS¹, Zhongxian Li, MS¹, Rong Jia, MS¹, Ying Liang, MS¹, Pengfei Ning, MS^{1,*,#}, and Yingqi Xu, PhD^{2,*,#} 

Technology in Cancer Research & Treatment
Volume 20: 1-14
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1533033820983298
journals.sagepub.com/home/tct



Abstract

Human epidermal growth factor 2 (HER2)+ breast cancer is considered the most dangerous type of breast cancers. Herein, we used bioinformatics methods to identify potential key genes in HER2+ breast cancer to enable its diagnosis, treatment, and prognosis prediction. Datasets of HER2+ breast cancer and normal tissue samples retrieved from Gene Expression Omnibus and The Cancer Genome Atlas databases were subjected to analysis for differentially expressed genes using R software. The identified differentially expressed genes were subjected to gene ontology and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analyses followed by construction of protein-protein interaction networks using the STRING database to identify key genes. The genes were further validated via survival and differential gene expression analyses. We identified 97 upregulated and 106 downregulated genes that were primarily associated with processes such as mitosis, protein kinase activity, cell cycle, and the p53 signaling pathway. Visualization of the protein-protein interaction network identified 10 key genes (*CCNA2*, *CDK1*, *CDC20*, *CCNB1*, *DLGAP5*, *AURKA*, *BUB1B*, *RRM2*, *TPX2*, and *MAD2L1*), all of which were upregulated. Survival analysis using PROGgeneV2 showed that *CDC20*, *CCNA2*, *DLGAP5*, *RRM2*, and *TPX2* are prognosis-related key genes in HER2+ breast cancer. A nomogram showed that high expression of *RRM2*, *DLGAP5*, and *TPX2* was positively associated with the risk of death. *TPX2*, which has not previously been reported in HER2+ breast cancer, was associated with breast cancer development, progression, and prognosis and is therefore a potential key gene. It is hoped that this study can provide a new method for the diagnosis and treatment of HER2 + breast cancer.

Keywords

bioinformatics, HER2+ breast cancer, key gene, prognosis-related gene, *TPX2*, therapeutic target

Abbreviations

HER2, human epidermal growth factor 2; GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEG, differentially expressed gene; PPI, protein-protein interaction; ER, estrogen receptor; PR, progesterone receptor; GEPIA, The Gene Expression Profiling Interactive Analysis; *CCNA2*, cyclin-A2; *CDC20*, cell-division cycle protein 20; *RRM2*, ribonucleoside-diphosphate reductase subunit M2; *TPX2*, targeting protein for Xklp2; OS, overall survival; SMAR1, scaffold/matrix attachment region 1; *DLGAP5*, DLG-associated protein 5

Received: August 24, 2020; Revised: November 1, 2020; Accepted: November 18, 2020.

Introduction

Breast cancer has become the second most common malignant tumor after lung cancer and poses a threat to life worldwide. A woman's lifetime risk of breast cancer is 12%, and the mortality rate of breast cancer is increasing.^{1,2} Early diagnosis and improvements in the awareness of cancer prevention have reduced mortality, but the incidence rate of cancer has remained high.^{3,4} HER2+ accounts for 25–30% of all breast cancer cases.⁵ In HER2+ breast cancer, HER2 protein is highly expressed; p53 is mutated in 75% of these cancers, which

¹ College of Computer and Information, Inner Mongolia Medical University, Hohhot, Inner Mongolia Autonomous Region, China

² College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang Province, China

* co-contributed for this work

co-corresponding author

Corresponding Authors:

Pengfei Ning, College of Computer and Information, Inner Mongolia Medical University, Hohhot, Inner Mongolia Autonomous Region, China.
Email: ningpengfei@immu.edu.cn

Yingqi Xu, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, Heilongjiang Province, China.
Email: xuyingqi1987@163.com



increases the degree of malignancy and leads to earlier recurrence and metastasis and poor prognosis.⁶⁻⁸

In recent years, with the development of high-throughput sequencing, DNA microarray and sequencing have become indispensable and effective methods in cancer research. In this study, bioinformatics methods were used to analyze DNA microarray data obtained from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases⁹ to provide basis for potential therapeutic targets and prognosis evaluation of HER2+ breast cancer. TCGA database is the largest cancer gene information database at present. It uses high-throughput genomic analysis technology to improve the understanding of cancer and the ability to prevent, diagnose, and treat cancer.

HER2+ breast cancer shows high degrees of malignancy and drug resistance. Many people have carried out in-depth research on HER2+ breast cancer. Al-Juboori et al showed that high expression of *PYK2* was related to a low survival rate of patients with HER2+ breast cancer, and revealed the role of *PYK2* in drug resistance during the treatment of patients.¹⁰ Lin et al found that overexpression of *RAC1* and *RRM2* is associated with poor prognosis of patients with HER2+ breast cancer, providing new targets for the diagnosis and treatment of these patients.¹¹ Lu et al showed that *PHLPP1*, *UBC*, *TGFBI*, and *AURKA* are associated with brain metastasis in patients with HER2+ breast cancer.¹² Bouchal et al found that *CDK1* was associated with the status of estrogen receptor (ER), tumor grading and HER2, which is helpful for the classification of breast tumors.¹³

Many cancers, including breast cancer, are caused by gene mutation. Therefore, the regulation of key genes can improve the survival rate of patients. Identifying and verifying key genes related to HER2+ breast cancer by using bioinformatics methods can provide new targets for diagnosis and treatment.¹⁴ In this study, HER2+ breast cancer data and normal tissue data were downloaded from TCGA database and GEO database to analyze differentially expressed genes. Additionally, Gene Ontology (GO) function analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, and protein interaction network analysis were carried out. Survival analysis was performed to screen key genes related to survival. Finally, effective key genes were verified by various methods and their roles were analyzed, including by examining key gene expression and survival and performing receiver operating characteristic analysis and multivariate logistic regression analysis.¹⁵

Material and Methods

Sources of DNA Microarray Data

The gene expression data for breast cancer were retrieved from GEO and TCGA databases. The dataset GSE45827 from the GEO public gene expression profile database was used to screen for differentially expressed genes (DEGs) in HER2+ breast cancer and normal breast tissues. The microarray dataset, which was constructed on the GPL570 platform,

comprised 155 samples, including 30 HER2+ breast cancer samples and 11 normal tissue samples. The gene expression data downloaded from TCGA database comprised 134 samples, including 47 HER2+ breast cancer samples and 87 normal tissue samples.

Identification of DEGs

The GEO dataset GSE45827 was screened for DEGs using the GEO2R web tool, which was used with gene expression matrix files using the limma package in R language. The p-values were adjusted to false discovery rates by the Benjamini-Hochberg method.

Functional Enrichment Analysis of DEGs

To analyze the biological processes underlying the pathogenesis of HER2+ breast cancer, the DEGs were subjected to GO enrichment analysis to evaluate biological processes, cellular components, and molecular functions. Additionally, the DEGs were subjected to KEGG pathway enrichment analysis using the DAVID online analysis tool and ClueGO plug-in in Cytoscape software. $P < 0.05$ indicated statistical significance. We also constructed a bar chart of P-values for the functions and pathways of upregulated and downregulated genes (DEGs) to further improve the interpretation of their biological significance.

Construction of PPI Networks for DEGs and Screening for Key Genes

The DEGs were uploaded to the STRING database (<https://stringdb.org/>), and the resulting protein-protein interaction (PPI) networks were imported into Cytoscape software for visualization and to remove isolated nodes showing no interaction with other proteins. Further, the PPI networks were subjected to modular analysis using the MCODE plug-in in Cytoscape software. Finally, key genes identified using the CytoHubba plug-in with maximal clique centrality and degree algorithms were intersected to generate the final list of key genes.

Survival Analysis of Key Genes

The PROGgeneV2 online tool (<http://genomics.jefferson.edu/proggene/>) was used to validate the correlation between these 10 key genes and HER2+ breast cancer as well as the effects of the key genes on the survival of patients with HER2+ breast cancer, and to construct the survival curve based on data from patients with HER2+ breast cancer in the GEO database.

Analysis of Expression Levels of Key Genes

The Gene Expression Profiling Interactive Analysis (GEPIA) database (<http://gepia.cancer-pku.cn/>) can be used to analyze DEGs between different types of tumors and normal tissues.

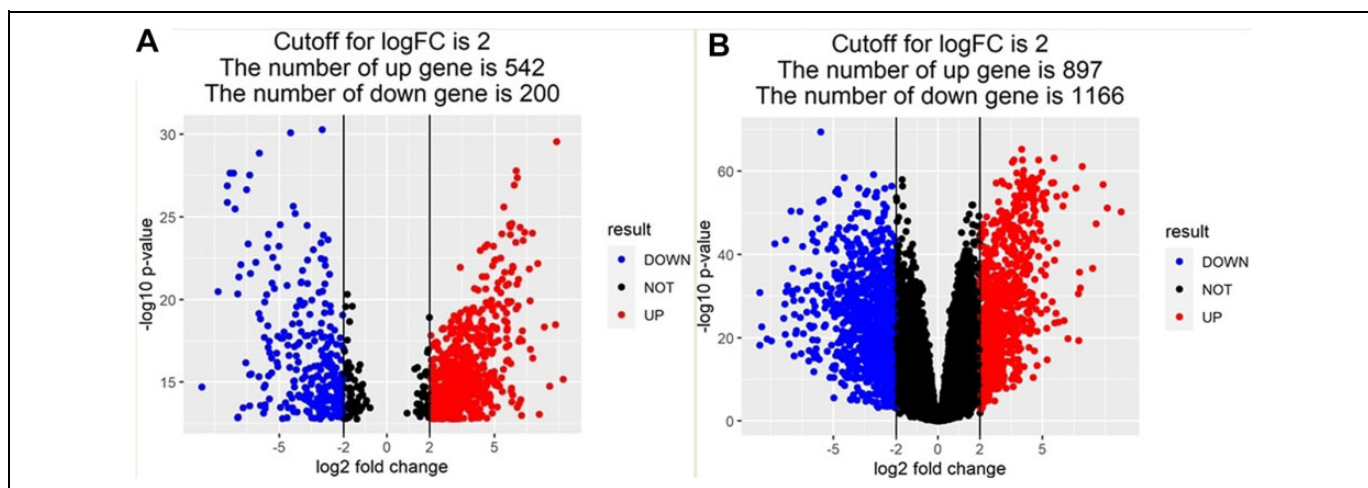


Figure 1. Volcano plots of DEGs (A. Dataset GSE45827; B. TCGA datasets). DEG: differentially expressed gene; FC, fold-change.

The expression of these key genes in HER2+breast cancer and normal tissues was further verified using the GEPIA database.

Validation of Protein Expressions

The Human Protein Atlas (<https://www.proteinatlas.org/>) database was used to view the expression of key gene in tissues, cells, and tumors.

Multi-Factor Logistic Regression Analysis

The clinical data obtained from HER2+ patients were downloaded from TCGA and screened for variables including age, tumor stage, and survival status which were included in multi-factorial logistic regression analysis along with key genes screened using survival analysis. A nomogram was developed and used to evaluate the discrimination, calibration, and clinical validity of the results using calibration curves.

Oncomine Analysis

Oncomine (<https://www.Oncomine.org>), a publicly accessible online cancer gene expression profile database, was used for genome-wide expression analyses. We used this tool to analyze the differential expression of key genes in different tumors and their association with survival status.

cBioPortal for Cancer Genomics Analysis

The cBioPortal (<https://www.cbioportal.org>), a website that integrates data from 126 tumor genome studies including TCGA and ICGC, was used to identify mutations in key genes in breast cancer.

Table 1. Number of DEGs in Datasets Retrieved From the 2 Databases.

Database	Upregulated DEGs	Downregulated DEGs	Total
GSE45827	542	200	742
TCGA datasets	897	1166	2063

Results

Identification of DEGs

A total of 1,017 DEGs (comprising 542 significantly upregulated and 200 significantly downregulated genes) were obtained from the GSE45827 dataset by GEO2R using screening thresholds of $P < 0.05$ and $|\log \text{fold-change}| \geq 2.0$ (Figure 1A). Gene expression data of 134 samples from TCGA database were normalized and subjected to differential gene expression analysis using the limma package in RStudio (an integrated development environment for R language) with screening thresholds of $P < 0.05$ and $|\log \text{fold-change}| \geq 2.0$, resulting in 897 significantly upregulated genes and 1,166 significantly downregulated genes (Figure 1B).

Common DEGs

The results of differential gene expression analysis performed on the GSE45827 and TCGA datasets are shown in Table 1. The subsequent intersection using the VennDiagram package yielded 203 common DEGs, comprising 97 upregulated (Figure 2A) and 106 downregulated genes (Figure 2B).

Functional Enrichment Analysis of DEGs

To further evaluate the functions of the 97 upregulated and 106 downregulated genes, the DEGs were subjected to GO functional enrichment and KEGG pathway enrichment

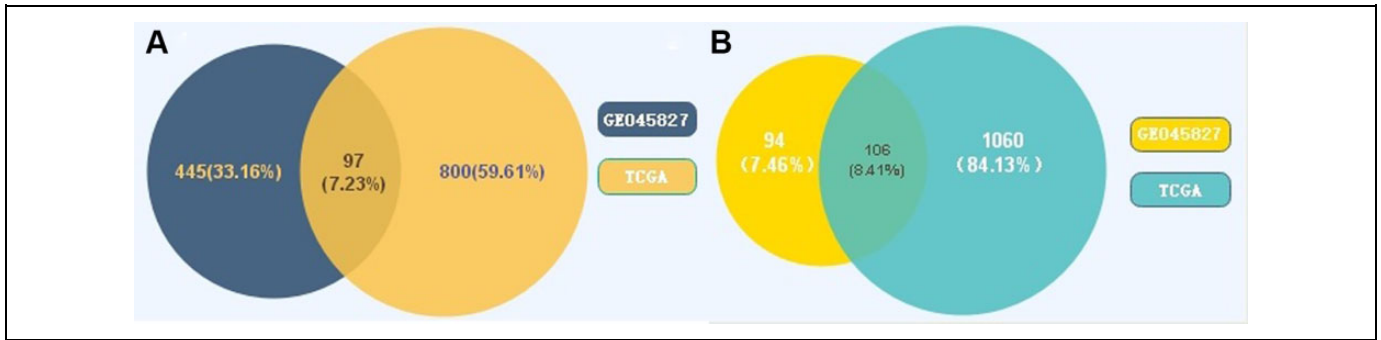


Figure 2. Venn diagrams of DEGs (A. Upregulated genes; B. Downregulated genes). DEG, differentially expressed gene.

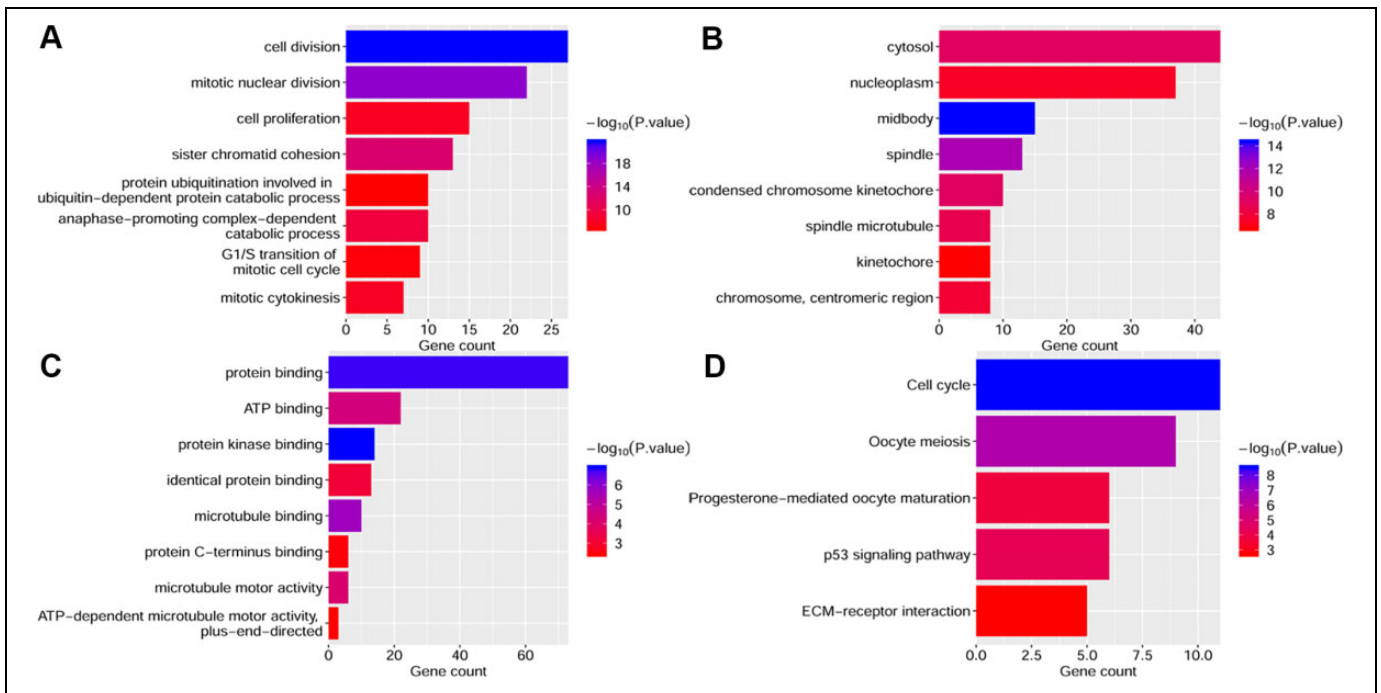


Figure 3. Functional enrichment analysis of upregulated genes (A. Biological processes; B. Cellular components; C. Molecular functions; D. KEGG pathway enrichment). DEG, differentially expressed gene; KEGG, Kyoto Encyclopedia of Genes and Genomes.

analyses using the DAVID online analysis tool (<https://david.ncifcrf.gov/>).

The enriched GO terms for upregulated genes primarily included cell division, mitosis, and cell proliferation, among others, for biological processes; spindle and spindle microtubule, among others, for cellular components; and protein kinase binding and microtubule binding, among others, for molecular functions. In contrast, the enriched GO terms for downregulated genes primarily included cellular response to cAMP and regulation of blood pressure, among others, for biological processes; extracellular space and proteinaceous extracellular matrix, among others, for cellular components; and glycosaminoglycan binding and peptide hormone receptor binding, among others, for molecular functions (Figure 3-4).

KEGG pathway enrichment analysis revealed that the upregulated genes were primarily involved in cell cycle and oocyte

meiosis, among others, whereas the downregulated genes were mainly involved in axon guidance and cAMP signaling pathway, among others (Figures 3 and 4).

Construction of PPI Networks for DEGs and Screening for Key Genes

PPI networks. The 97 upregulated and 106 downregulated genes were uploaded to the STRING database to construct PPI networks, and the results were exported to tab-separated values files, which were subsequently imported into Cytoscape software for visualization and labeling of upregulated and downregulated genes with different colors (Figure 5). The PPI networks were subjected to modular analysis using the MCODE plug-in in Cytoscape software with the following parameters: Node Score Cutoff ≥ 0.2 , K-core ≥ 2 , and

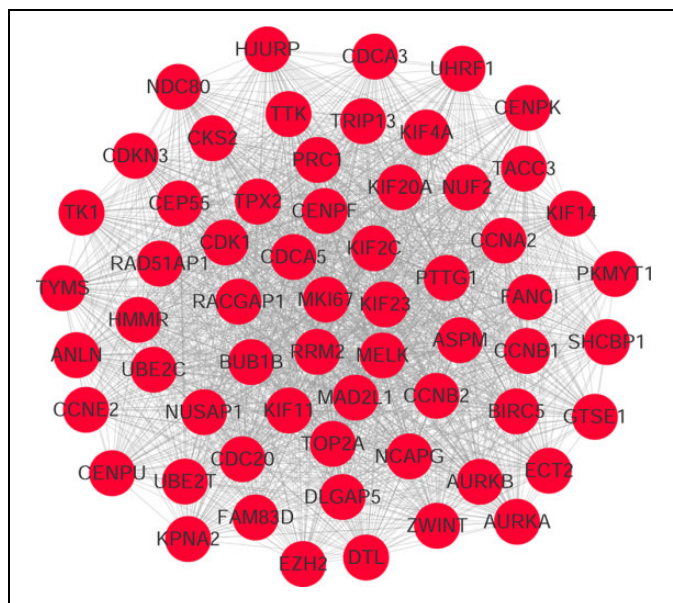


Figure 6. Module 1, comprising 61 nodes (all of which are upregulated genes) and 1,715 edges.

Maximum Depth = 100. This yielded a total of 6 modules, of which Module 1 with a score ≥ 20 was selected for further analysis (Figure 6). Module 1 comprised 61 nodes (all of which were upregulated genes) and 1,715 edges.

Module 1 was further analyzed using the ClueGO plug-in (Table 2). The enriched GO terms in Module 1 primarily included mitotic cell cycle process, nuclear division, and regulation of chromosome segregation for biological processes; spindle, condensed chromosome, spindle microtubule, cyclin-dependent protein kinase, and centralspindlin complex for cellular components; and microtubule motor activity and histone kinase activity for molecular functions. The results of KEGG pathway enrichment analysis suggested that Module 1 is primarily involved in cell cycle, p53 signaling pathway, and pyrimidine metabolism.

Screening for Key Genes

The key genes were screened using the CytoHubba plug-in, and the top 15 key genes were obtained using the multiscale curvature classification algorithm. To ensure that the key genes

Table 2. GO and KEGG Analyses for Module 1.

Category	GOID	GO term	Count	Term P value	Associated genes found
GO_BP	GO:1903047	mitotic cell cycle process	44	5.4E-47	[ANLN, AURKA, AURKB, BIRC5, BUB1B, CCNA2, CCNB1, CCNB2, CCNE2, CDC20, CDCA5, CDK1, CDKN3, CENPF, CENPK, CEP55, CKS2, DLGAP5, ECT2, GTSE1, KIF11, KIF14, KIF20A, KIF23, KIF2C, KIF4A, MAD2L1, MELK, NCAPG, NDC80, NUF2, NUSAP1, PKMYT1, PRC1, RACGAP1, RRM2, TACC3, TOP2A, TPX2, TRIP13, TTK, TYMS, UBE2C, ZWINT]
GO_BP	GO:0000280	nuclear division	31	1.7E-37	[ASPM, AURKA, AURKB, BIRC5, BUB1B, CCNB1, CCNE2, CDC20, CDCA5, CENPK, CKS2, DLGAP5, KIF11, KIF14, KIF23, KIF2C, KIF4A, MAD2L1, NCAPG, NDC80, NUF2, NUSAP1, PRC1, PTTG1, RACGAP1, TOP2A, TPX2, TRIP13, TTK, UBE2C, ZWINT]
GO_BP	GO:0051983	regulation of chromosome segregation	18	2.5E-24	[AURKB, BUB1B, CCNB1, CDC20, CDCA5, CENPF, DLGAP5, ECT2, KIF2C, MAD2L1, MKI67, NDC80, PTTG1, RACGAP1, TACC3, TRIP13, TTK, UBE2C]
GO_BP	GO:0007088	regulation of mitotic nuclear division	20	1.8E-23	[ANLN, AURKA, AURKB, BUB1B, CCNB1, CDC20, CDCA5, CENPF, DLGAP5, KIF11, MAD2L1, MKI67, NDC80, NUSAP1, PKMYT1, PTTG1, TACC3, TRIP13, TTK, UBE2C]
GO_CC	GO:0005819	spindle	27	6.49E-28	[ASPM, AURKA, AURKB, BIRC5, BUB1B, CCNB1, CDC20, CDK1, CENPF, DLGAP5, ECT2, FAM83D, KIF11, KIF14, KIF20A, KIF23, KIF2C, KIF4A, MAD2L1, NCAPG, NUSAP1, PRC1, RACGAP1, SHCBP1, TACC3, TPX2, TTK]
GO_CC	GO:0000779	condensed chromosome, centromeric region	15	5.24E-19	[AURKA, AURKB, BIRC5, BUB1B, CCNB1, CENPF, CENPK, CENPU, HJURP, KIF2C, MAD2L1, NCAPG, NDC80, NUF2, ZWINT]
GO_CC	GO:0005876	spindle microtubule	8	1.59E-10	[AURKA, AURKB, BIRC5, CDK1, KIF11, KIF4A, NUSAP1, PRC1]
GO_CC	GO:0000307	cyclin-dependent protein kinase holoenzyme complex	6	2.88E-08	[CCNA2, CCNB1, CCNB2, CCNE2, CDK1, CKS2]
GO_CC	GO:0097149	centralspindlin complex	3	3.07E-08	[ECT2, KIF23, RACGAP1]
GO_MF	GO:0003777	microtubule motor activity	6	9.01E-07	[KIF11, KIF14, KIF20A, KIF23, KIF2C, KIF4A]
GO_MF	GO:0035173	histone kinase activity	4	3.47E-06	[AURKA, AURKB, CCNB1, CDK1]
GO_MF	GO:0016538	cyclin-dependent protein serine/threonine kinase regulator activity	5	3.81E-06	[CCNA2, CCNB1, CCNB2, CCNE2, CKS2]
GO_MF	GO:0097472	cyclin-dependent protein kinase activity	3	3.84E-04	[CCNA2, CDK1, CDKN3]
KEGG	KEGG:04110	Cell cycle	11	8.36E-14	[BUB1B, CCNA2, CCNB1, CCNB2, CCNE2, CDC20, CDK1, MAD2L1, PKMYT1, PTTG1, TTK]
KEGG	KEGG:04115	p53 signaling pathway	6	1.10E-07	[CCNB1, CCNB2, CCNE2, CDK1, GTSE1, RRM2]
KEGG	KEGG:00240	Pyrimidine metabolism	3	8.88E-04	[RRM2, TK1, TYMS]

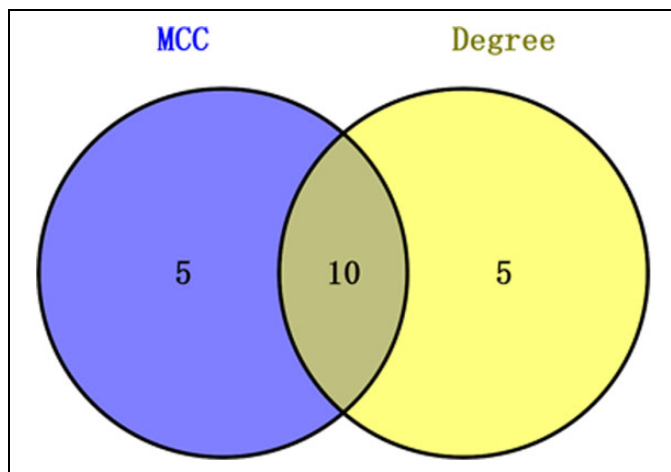


Figure 7. Venn diagram of key genes.

Table 3. Ten Key Genes Identified in HER2+ Breast Cancer.

Rank	Gene symbol	Degree
1	CCNA2	69
1	CDK1	69
3	CDC20	68
3	CCNB1	68
5	DLGAP5	66
5	AURKA	66
5	BUB1B	66
8	RRM2	65
8	TPX2	65
8	MAD2L1	65

were accurately identified, the top 15 key genes were intersected with another set of top 15 key genes ranked based on the degree score in the PPI network (Figure 7). The intersection yielded 10 key genes: Cyclin-A2 (*CCNA2*), *CDK1*, cell-division cycle protein 20 (*CDC20*), *CCNB1*, *DLGAP5*, *AURKA*, *BUB1B*, ribonucleoside-diphosphate reductase subunit M2 (*RRM2*), targeting protein for *Xklp2* (*TPX2*), and *MAD2L1*, all of which were upregulated. See Table 3 for the degree score of each node.

Survival Analysis of Key Genes

The PROGgeneV2 online tool was employed to further validate the correlation between these key genes and HER2+ breast cancer, in which the dataset GSE19783 from the GEO database was used. We constructed the overall survival (OS) curve with the samples grouped based on their median expression values. The results revealed that 5 genes were associated with the prognosis of HER2+ breast cancer: *CDC20*, *CCNA2*, *DLGAP5*, *RRM2*, and *TPX2*, among which *CDC20* displayed the highest correlation [hazard ratio:1.42 (1.13–1.79), $P:0.0026$] (Figure 8).

Survival analysis and ROC curve prediction of the survival time were performed for the 5 key genes. The validation data

was the GSE21653 data from GEO database, which included 4 subtypes of breast cancer. The clinical data and expression data of HER2+ were selected in this study.

The results showed that high expression of *TPX2*, *RRM2*, *DLGAP5*, *CDC20*, and *CCNA2* were all related to the low survival of HER2+ patients. The P values of *TPX2*, *DLGAP5*, and *CDC20* survival analysis were all less than 0.05, which was significant. The P value of *CDC20* was lowest at 0.0015, which is consistent with the survival analysis results (Figure 9).

The survivalROC package of R language was used to predict the 5-year survival rate of patients with HER2+, and the ROC curve was drawn. The expression of 5 genes was used as a marker, and the algorithm was the nearest neighbor estimation. The results showed that the area under the ROC curve of *TPX2* was highest at 0.863, and was the best predictor of the 5-year survival rate (Figure 10).

Expression Levels of Key Genes

We uploaded the key genes to the GEPIA database for online analysis to verify our results. Expression analysis of each key gene showed that *CCNA2*, *CDC20*, *DLGAP5*, *RRM2*, and *TPX2* were significantly upregulated in cancer tissues, which agrees with the DNA microarray analysis results (Figure 11).

Validation of Protein Expression

To further validate the gene results, the Human Protein Atlas database was used to evaluate expression in normal breast and breast cancer tissues (Figure 12). The *RRM2* and *DLGAP5* proteins were not expressed in normal breast tissues, whereas they showed low and high protein expression in breast cancer tissues, respectively. *CDC20* protein showed medium expression in normal tissues and high expression in breast cancer tissue.

Multi-Factor Logistic Regression Analysis

Variables in the logistic regression analysis included age, tumor stage, survival status, and key genes (*CCNA2*, *CDC20*, *RRM2*, *DLGAP5*, and *TPX2*). The key genes in the nomogram were labeled as 0 for low expression, 1 for high expression, and 1–4 for T1–T4 staging, respectively. The nomogram showed that high expression of *RRM2*, *DLGAP5*, and *TPX2* was positively correlated with the death score line, whereas *CCNA2* and *CDC20* were negatively correlated and the risk of death increased with age and tumor stage. In addition, the distribution curve of the nomogram's apparent predicted value fit well with the best curve of the calibration map with an ROC of 0.754, showing that the model had good discrimination, calibration, and clinical effectiveness (Figure 13).

Oncomine Analysis

Oncomine analysis showed that the *TPX2* and *DLGAP5* mRNA levels were significantly elevated in breast cancer, as well as in colorectal and lung cancer, suggesting that their high expression was associated with various cancers. Immunohistochemistry

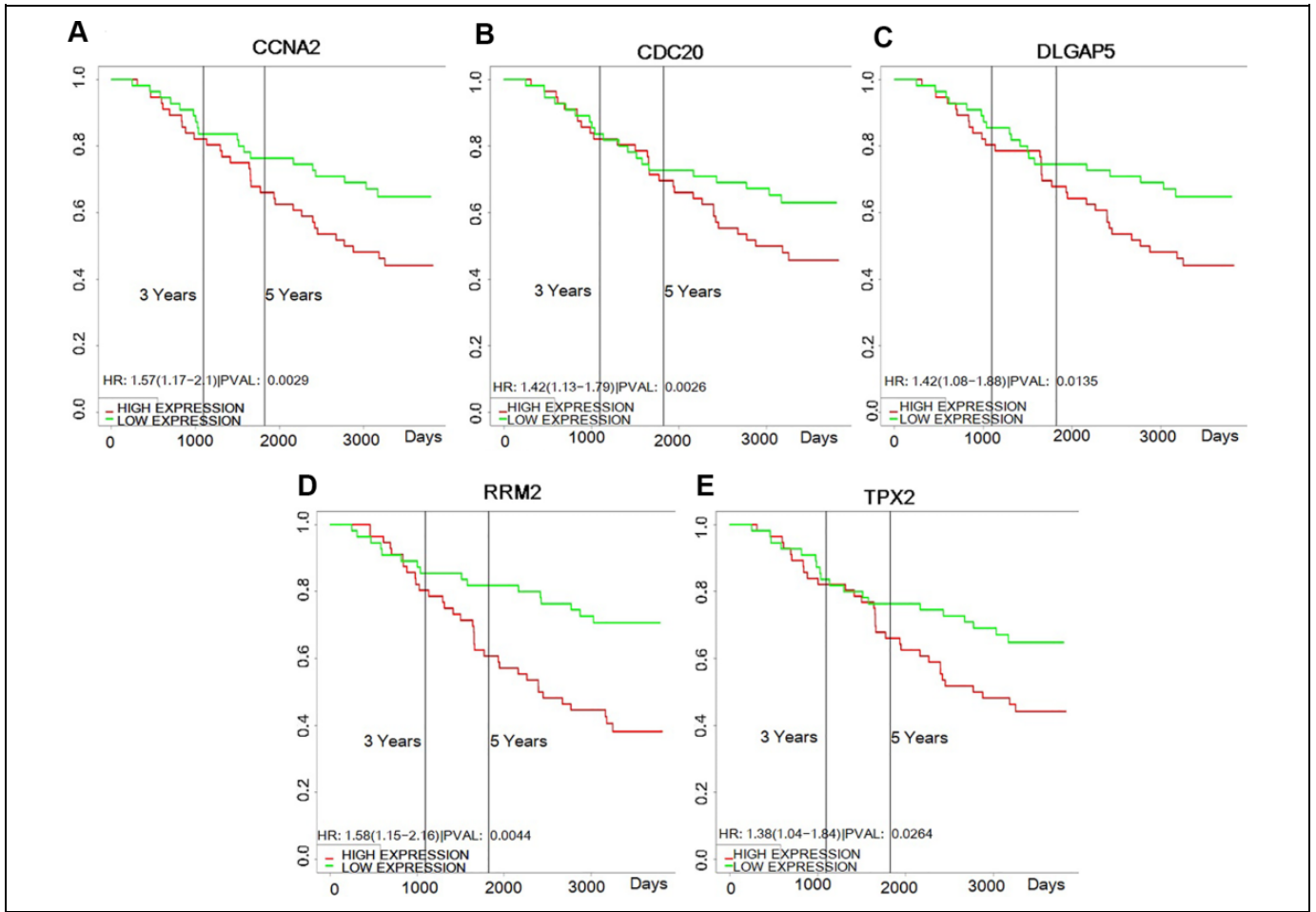


Figure 8. Survival analysis of 5 key genes associated with the prognosis of HER2+ breast cancer with GSE19783 dataset.

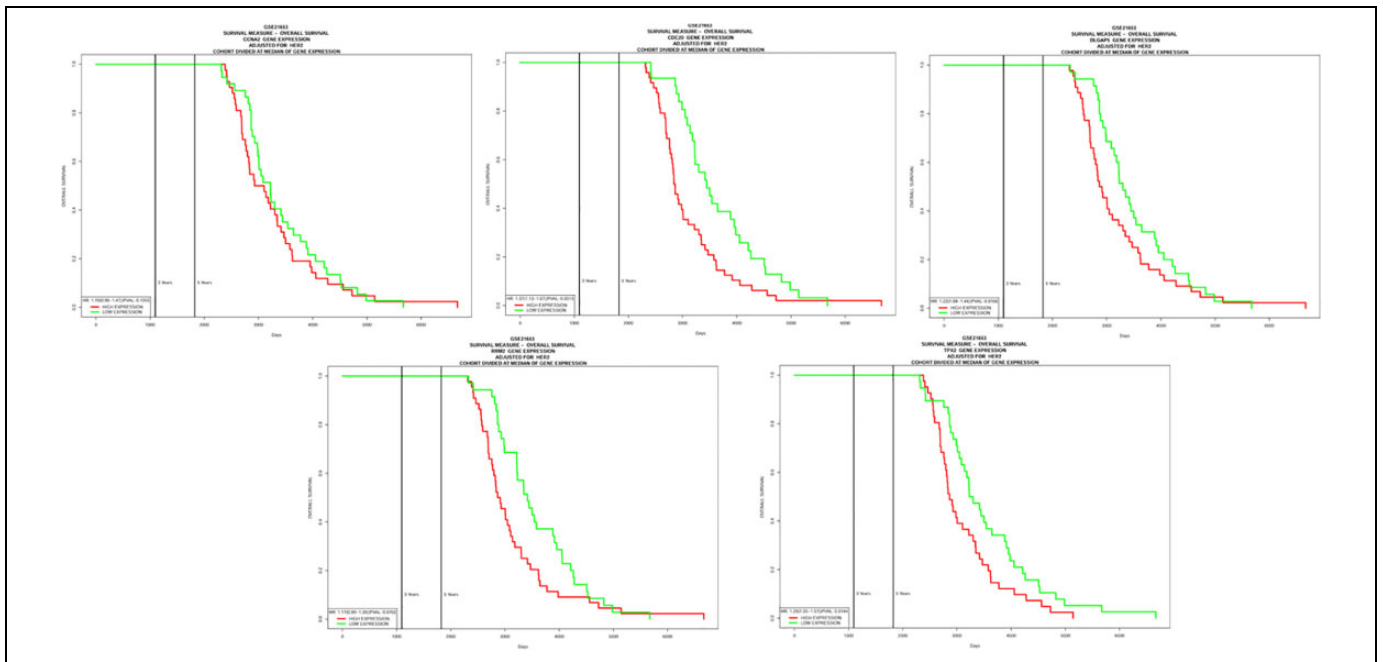


Figure 9. Survival analysis of 5 key genes associated with the prognosis of HER2+ breast cancer with GSE21653 dataset.

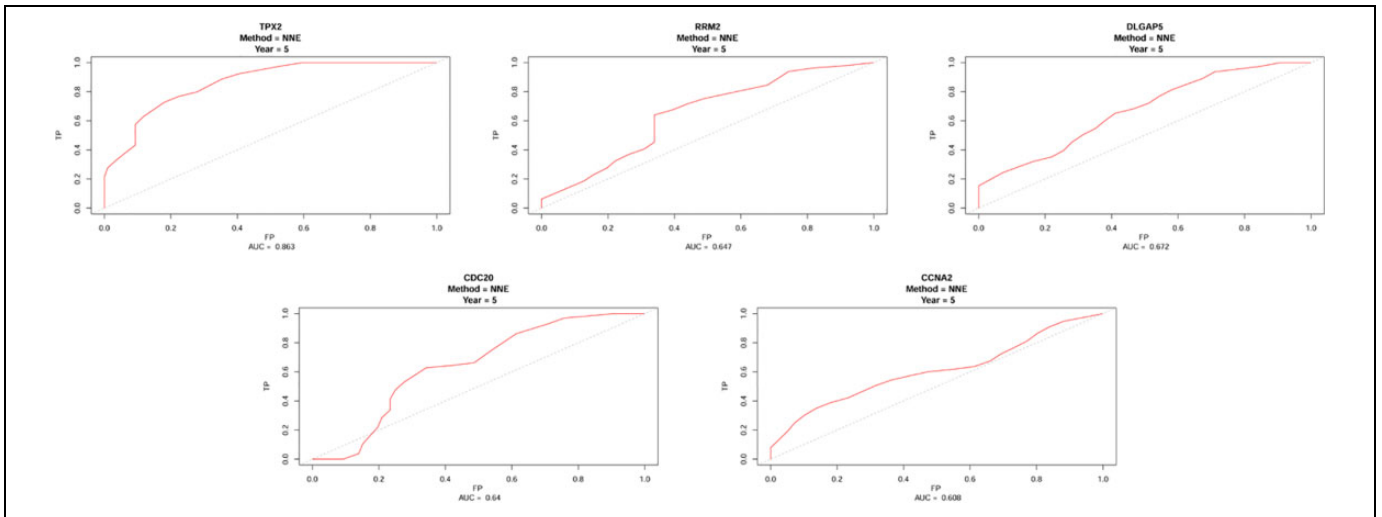


Figure 10. ROC curve of 5 key genes in predicting 5-year survival rate.

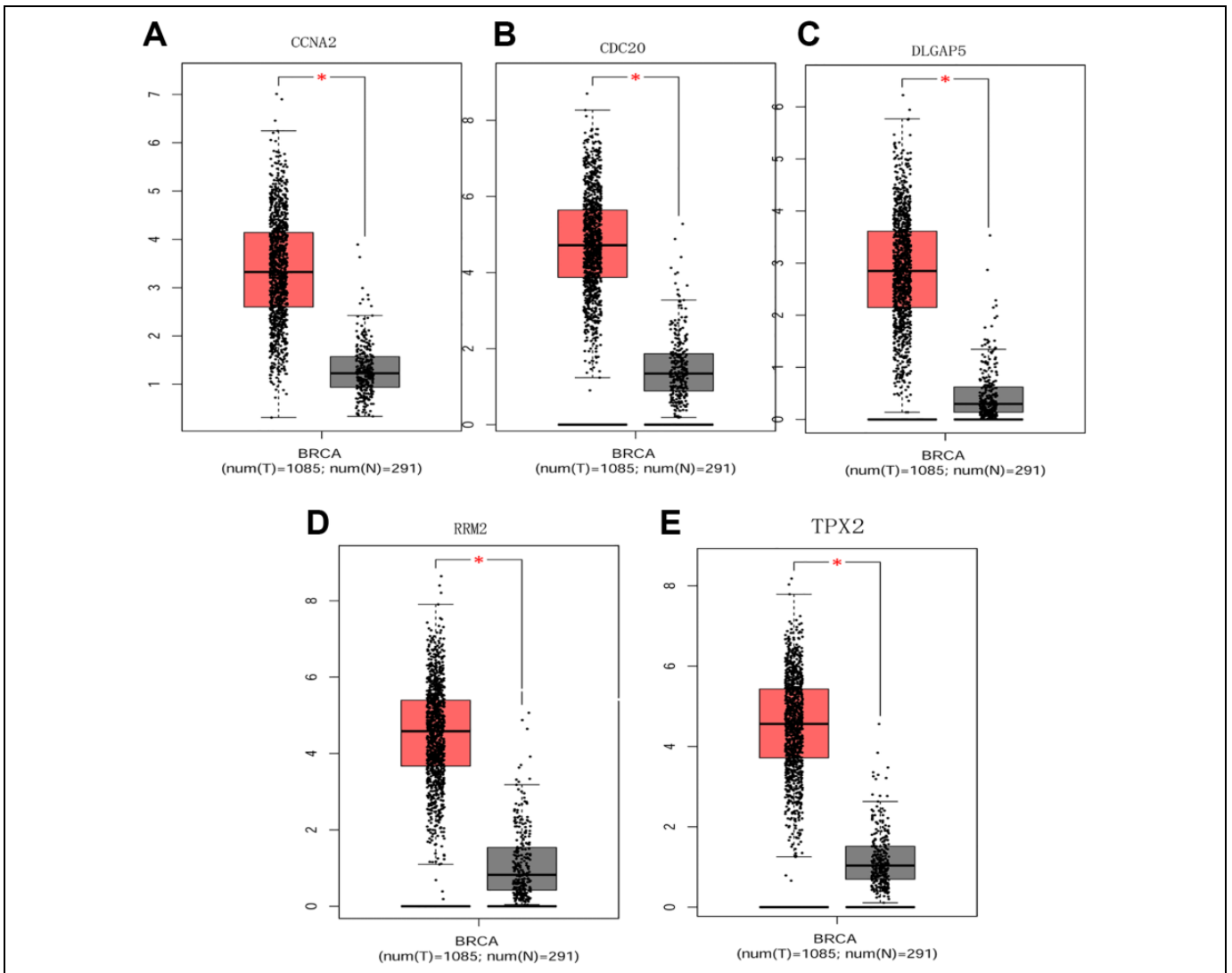


Figure 11. Expression of 5 key genes associated with the prognosis of HER2+ breast cancer (gene expression in breast cancer tissues and normal tissues is marked in red and black, respectively).

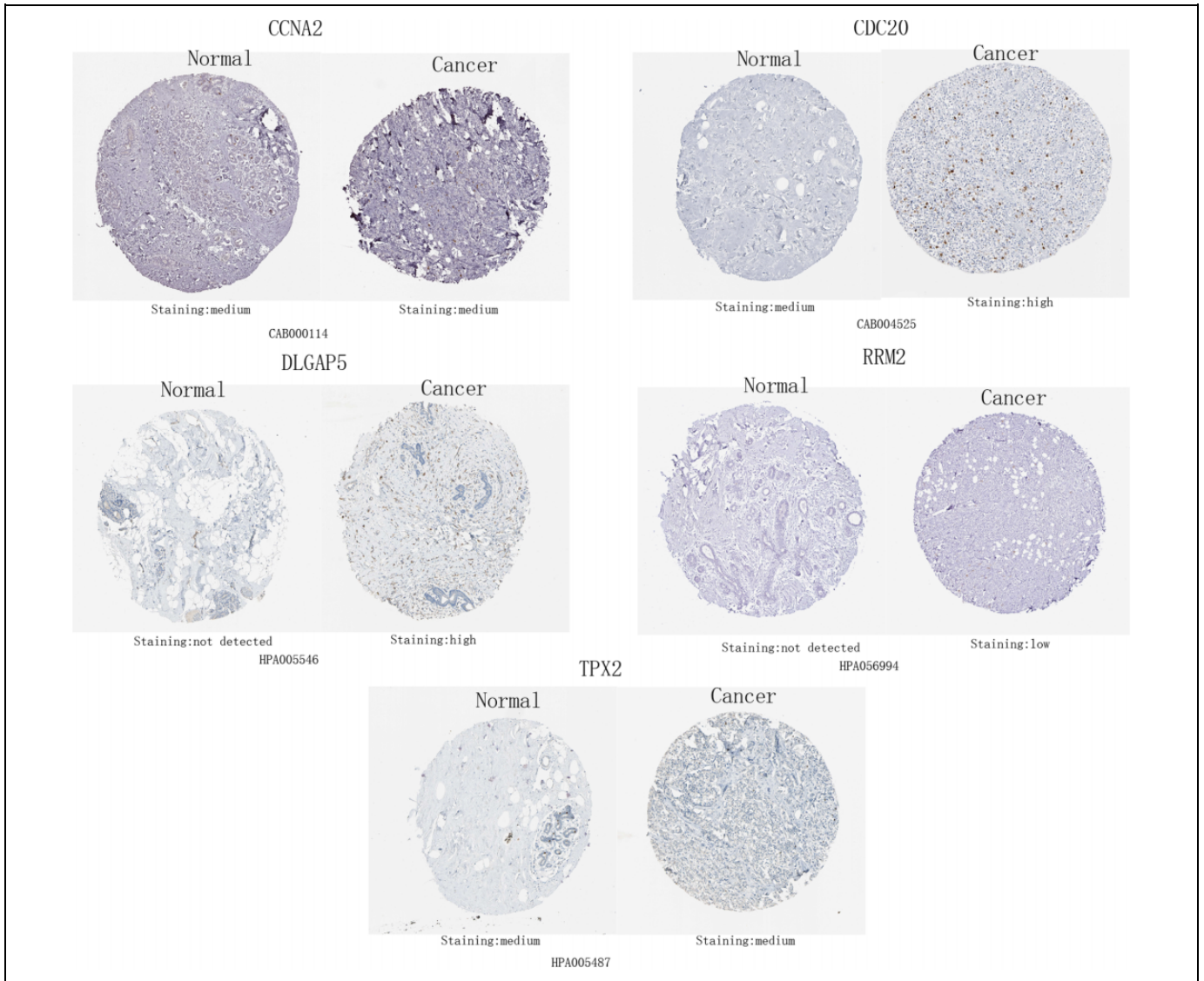


Figure 12. Immunohistochemistry of genes in breast cancer and normal tissues.

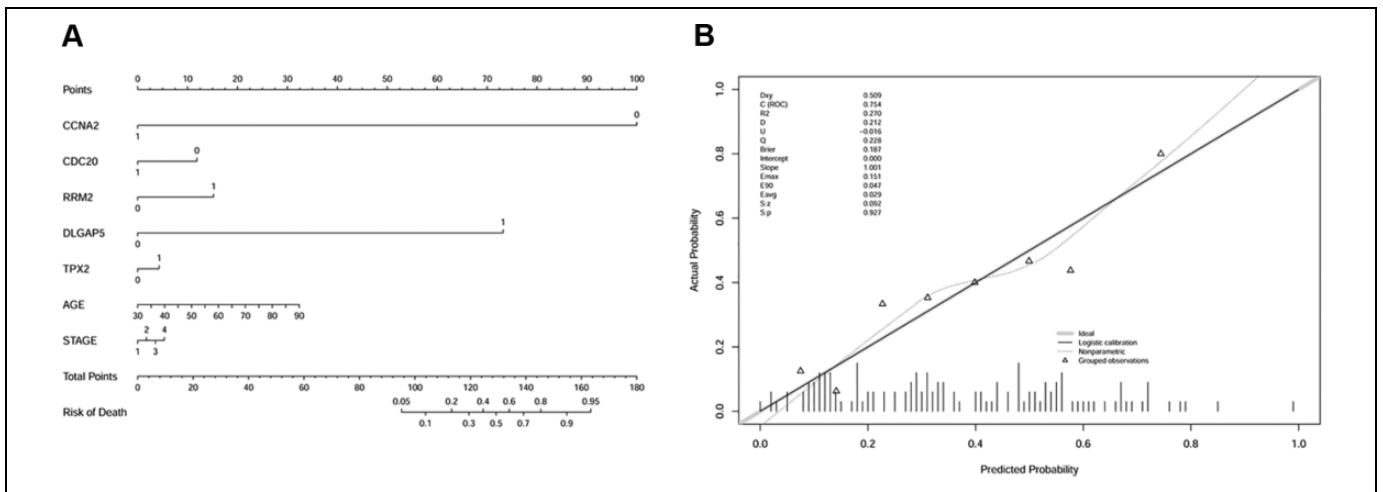


Figure 13. Logistic regression analysis (A. Nomogram;B. Calibration graph).

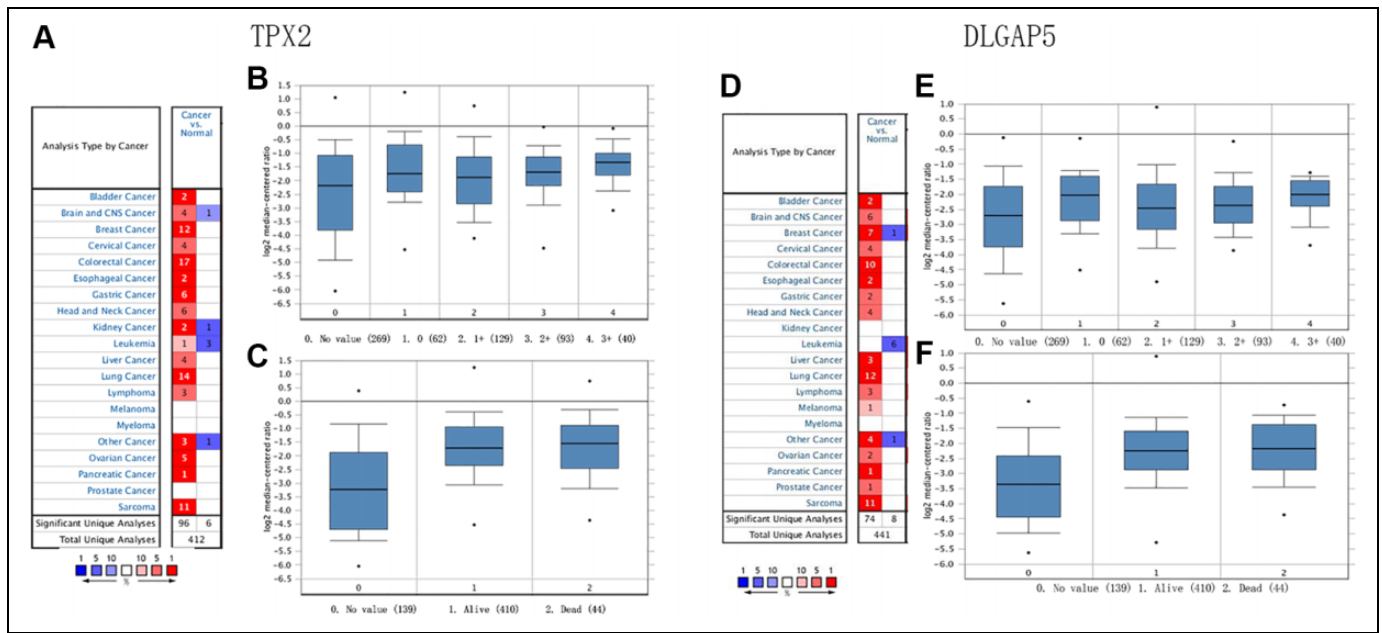


Figure 14. OncoPrint analysis of *TPX2* and *DLGAP5* mRNA expression levels (A. *TPX2* expression values in different cancer types; B. Expression of *TPX2* in different HER2 immunohistochemical levels; C. Expression of *TPX2* in patients with breast cancer at different status levels. D. *DLGAP5* expression values in different cancer types; E. Expression of *DLGAP5* at different HER2 immunohistochemical levels; F. Expression of *DLGAP5* in patients with breast cancer with different status levels).

analysis of HER2 showed that expression of *TPX2* and *DLGAP5* was higher in tissues from HER2¹⁺, 2⁺, and 3⁺ cases than in normal tissues, showing an increasing trend. According to the overall survival status of patients with breast cancer, *TPX2* and *DLGAP5* levels were higher in deceased patients than in surviving patients, suggesting that these genes were associated with survival (Figure 14).

cBioPortal for Cancer Genomics Analysis

TPX2 mutation types include mRNA-high and amplification in breast cancer. Among the different types of breast cancer, such as breast invasive ductal carcinoma and metaplastic breast cancer, high mRNA levels were observed in a large number of patients, which is consistent with our findings. The main mutation of *DLGAP5* resulted in amplification. In addition, we investigated the molecules downstream of *TPX2* and calculated the correlation coefficient between a specific gene in relation to other genes based on expression levels. A larger correlation coefficient indicated a closer association between 2 genes (positive numbers represent positive correlation, whereas negative numbers represent negative correlation), suggesting a greater association between the functions of upstream and downstream genes. *BUB1* is associated with *TPX2* and *DLGAP5*, which should be further evaluated in future studies. Scatter plots of *TPX2* showed increasing copy number of this gene with a progressively higher expression value, indicating that *TPX2* amplification is positively correlated with high expression (Figure 15).

Discussion

In this study, microarray datasets of HER2⁺ breast cancer retrieved from TCGA and GEO databases were subjected to differential gene expression analysis using bioinformatics methods. The intersection of DEGs from the 2 databases yielded 203 DEGs, including 97 upregulated and 106 downregulated genes. GO and KEGG pathway enrichment analysis demonstrated that these significantly upregulated DEGs are closely associated with mitosis, protein kinase activity, cell cycle, and p53 signaling pathway. The PPI network visualization identified 10 key genes: *CCNA2*, *CDK1*, *CDC20*, *CCNB1*, *DLGAP5*, *AURKA*, *BUB1B*, *RRM2*, *TPX2*, and *MAD2L1*, all of which were upregulated. The validation results of their expression levels using the GEPIA database are consistent with those of microarray data analysis. Survival analysis using the PROG-geneV2 online tool showed that the results for *CDC20*, *CCNA2*, *DLGAP5*, *RRM2*, and *TPX2* were significant ($P < 0.05$). The nomograph showed that high expression of *RRM2*, *DLGAP5*, and *TPX2* was positively associated with the risk of death. A literature search revealed that both *RRM2* and *DLGAP5* have previously been reported, whereas *TPX2* has not; thus, *TPX2* was further analyzed in this study.

In breast cancer, *CDC20* binds to and promotes the protein degradation of scaffold/matrix attachment region 1 (*SMAR1*). The significant correlation between the 2 proteins in breast cancer cells and patient samples suggests that *CDC20* has an important negative regulatory effect on *SMAR1* in advanced cancer by enhancing the migratory and invasive abilities of cancer cells.¹⁶⁻²⁰ It has been shown that *CDC20* is positively

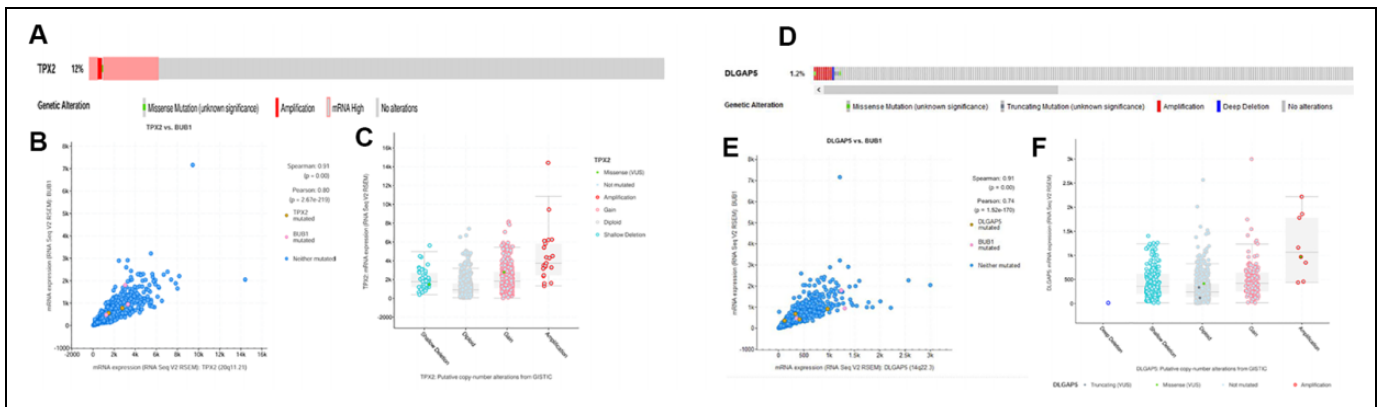


Figure 15. *TPX2* variation in breast cancer (A. *TPX2* mutation types in breast cancer; B. actions of molecules downstream of *TPX2*; C. scatter plots of *TPX2* gene and copy number changes. D. *DLGAP5* mutation types in breast cancer; E. actions of molecules downstream to *DLGAP5*; F. scatter plots of *DLGAP5* gene and copy number changes).

correlated with the grade of breast cancer and HER2 status in patients, with low expression of *CDC20* predicting a better prognosis,²¹ indicating that *CDC20* plays a very important role in the diagnosis and treatment of breast cancer. These results are consistent with those of our study.

Previous studies showed that *CCNA2* has an important role in relapse-free survival and the OS of patients with ER+ breast cancer.²²⁻²⁴ *CCNA2* is also highly expressed in colorectal cancer, and knockout of *CCNA2* may affect cell cycle progression and induce apoptosis.²⁵ In addition, ectopic expression of *CCNA2* promotes *in vitro* migration and invasion of non-small cell lung carcinoma cells.²⁶

RRM2 is an important marker for breast cancer metastasis and is a regulator associated with the cell cycle, cell migration, cell adhesion, and other cancer-related cellular functions. Hence, it is considered as a target for new complex pharmacotherapy and intervention regimens against the metastasis of breast cancer.²⁷⁻²⁸ *RRM2* is an upregulated gene in breast cancer regardless of the subtype; thus, its overexpression may serve as a biomarker in patients with breast cancer with relatively low OS and relapse-free survival.^{11,29,30}

DLGAP5 (DLG-associated protein 5) is a protein-coding gene involved in spindle assembly and separation of sister chromatids. Tumor growth depends, to some extent, on increased mitotic activity. Its key steps include spindle assembly and separation of sister chromatids.³¹⁻³³ In patients with lung cancer, upregulation of *DLGAP5* is associated with lower OS and relapse-free survival and therefore may be important for early diagnosis and treatment.^{34,35} Our study showed that the Aurora A signaling pathway is related to *DLGAP5*. GO analysis related to this gene revealed an association with phosphoprotein phosphate activity. Tumor analysis showed that the level of *DLGAP5* mRNA in breast cancer tissue was increased, and increased with an increasing HER2 status. The cBioPortal results showed that the *DLGAP5* mutation type of breast cancer was mainly amplification, and *BUB1* is a key downstream gene.

TPX2, a spindle assembly factor, is required for normal mitotic spindle assembly and normal apoptotic microtubule

assembly. It was previously reported that *TPX2* is a direct target protein of microRNA-491 in breast cancer cells. Restoration of microRNA-491 expression inhibited the invasion and migration of breast cancer cells and knockout of *TPX2* significantly reversed this inhibition.³⁶ *TPX2* expression gradually increases during the different stages of breast cancer.^{37,38} In addition, high expression of *TPX2* in gastric cancer tissues is associated with the aggravation of tumors and low survival rate.³⁹ Our study showed that *TPX2*-related pathways include the regulation of *PLK1* activity and gene expression during G2/M phase, whereas *TPX2*-associated GO terms include GTP binding and protein kinase binding. Additionally, high expression of *TPX2* can also reduce the survival time of patients with HER2+ breast cancer. Oncomine analysis showed that the mRNA level of *TPX2* was significantly elevated in breast cancer, and was higher than normal in HER21+, 2+ and 3+ cancer, showing an increasing trend. The cBioPortal results indicate that the major *TPX2* mutation types in breast cancer include mRNA-high and amplification and that *BUB1* is a key downstream gene, which warrants further investigation.

Although follow-up studies are needed to validate our results, the identified genes are potential targets for the treatment and prognostic assessment of HER2+ breast cancer.

Conclusion

In this study, key genes associated with patients with HER2+ breast cancer were screened using bioinformatics methods and analyzed to predict risk models for these patients. Further analysis of *TPX2* revealed that its expression level was positively correlated with breast cancer development and prognosis. Therefore, our study provides new targets for investigating HER2+ breast cancer.

Authors' Note

Yujie Weng, Wei Liang, Pengfei Ning, and Yingqi Xu are co-first contributors for this work. Our study did not require ethical board approval because it did not involve human or animal trials.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Yingqi Xu  <https://orcid.org/0000-0002-5426-8147>

References

- Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA Cancer J Clin.* 2016;66(1):7-30.
- Torre LA, Sauer AM, Chen MS, Kagawa-Singer M, Jemal A, Siegel RL. Cancer statistics for Asian Americans, Native Hawaiians, and Pacific Islanders, 2016: converging incidence in males and females. *CA Cancer J Clin.* 2016;66(3):182-202.
- Bianchini G, Balko JM, Mayer IA, Sanders ME, Gianni L. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nat Rev Clin Oncol.* 2016;13(11):674-690.
- Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. *CA Cancer J Clin.* 2016;66(2):115-132.
- Akram M, Iqbal M, Daniyal M, Khan AU. Awareness and current knowledge of breast cancer. *Biol Res.* 2017;50(1):33.
- Figuerola-Magalhães MC, Jelovac D, Connolly R, Wolff AC. Treatment of HER2+ breast cancer. *Breast.* 2014;23(2):128-136.
- Pavlenko IA, Zavalishina LE, Povilaitite PE. HER2/neu gene amplification as a mechanism of clonal heterogeneity in breast cancer [in Russian]. *Arkh Patol.* 2019;81(6):49-55.
- Ma CX, Ellis MJ. The cancer genome atlas: clinical applications for breast cancer. *Oncology (Williston Park).* 2013;27(12):1263-1269, 1274-1279.
- Chang TH, Wang F, Chapin W, Huang RS. Identification of microRNAs as breast cancer prognosis markers through the cancer genome atlas. *PLoS One.* 2016;11(12):e0168284.
- Al-Juboori SIK, Vadakekolathu J, Idri S, et al. PYK2 promotes HER2-positive breast cancer invasion. *J Exp Clin Cancer Res.* 2019;38(1):210.
- Lin Y, Fu F, Lv J, et al. Identification of potential key genes for HER-2 positive breast cancer based on bioinformatics analysis. *Medicine (Baltimore).* 2020;99(1):e18445.
- Lu X, Gao C, Liu C, et al. Identification of the key pathways and genes involved in HER2-positive breast cancer with brain metastasis. *Pathol Res Pract.* 2019;215(8):152475.
- Bouchal P, Schubert OT, Faktor J, et al. Breast cancer classification based on proteotypes obtained by SWATH mass spectrometry. *Cell Rep.* 2019; 16; 28(3):832-843.e7.
- Cui C, Li L, Zhen J. Bioinformatic analysis reveals the key pathways and genes in early-onset breast cancer. *Med Oncol.* 2018; 35(5):67.
- Hou L, Chen M, Wang M, et al. Systematic analyses of key genes and pathways in the development of invasive breast cancer. *Gene.* 2016;593(1):1-12.
- Paul D, Ghorai S, Dinesh US, Shetty P, Chattopadhyay S, Santra MK. Cdc20 directs proteasome-mediated degradation of the tumor suppressor SMAR1 in higher grades of cancer through the anaphase promoting complex. *Cell Death Dis.* 2017;8(6):e2882.
- Sabbaghi M, Gil-Gómez G, Guardia C, et al. Defective cyclin B1 induction in trastuzumab-emptansine (T-DM1) acquired resistance in HER2-positive breast cancer. *Clin Cancer Res.* 2017;23(22):7006-7019.
- Mukherjee A, Joseph C, Craze M, Chrysanthou E, Ellis IO. The role of BUB and CDC proteins in low-grade breast cancers. *Lancet.* 2015;385(1):S72.
- Tang J, Lu M, Cui Q, et al. Overexpression of ASPM, CDC20, and TTK confer a poorer prognosis in breast cancer identified by gene co-expression network analysis. *Front Oncol.* 2019;9:310.
- Karra H, Repo H, Ahonen I, et al. Cdc20 and securin overexpression predict short-term breast cancer survival. *Br J Cancer.* 2014; 110(12):2905-2913.
- Cheng L, Huang YZ, Chen WX, Shi L, Li Z, Zhang X. Cell division cycle proteininising prognostic biomarker of breast cancer. *Biosci Rep.* 2020;40(5):BSR20191227.
- Gao T, Han Y, Yu L, Ao S, Li Z, Ji J. CCNA2 is a prognostic biomarker for ER+ breast cancer and tamoxifen resistance. *PLoS One.* 2014;9(3):e91771.
- Wang Y, Kojetin D, Burris TP. Anti-proliferative actions of a synthetic REV-ERB α/β agonist in breast cancer cells. *Biochem Pharmacol.* 2015;96(4):315-322.
- Deng JL, Xu YH, Wang G. Identification of potential crucial genes and key pathways in breast cancer using bioinformatic analysis. *Front Genet.* 2019;10:695.
- Gan Y, Li Y, Li T, Shu G, Yin G. CCNA2 acts as a novel biomarker in regulating the growth and apoptosis of colorectal cancer. *Cancer Manag Res.* 2018;10:5113-5124.
- Ruan JS, Zhou H, Yang L, Wang L, Jiang ZS, Wang SM. CCNA2 facilitates epithelial-to-mesenchymal transition via the integrin $\alpha\beta3$ signaling in NSCLC. *Int J ClinExpPathol.* 2017;10(8):8324-8333.
- Bell R, Barraclough R, Vasieva O. Gene expression meta-analysis of potential metastatic breast cancer markers. *Curr Mol Med.* 2017;17(3):200-210.
- Gong MT, Ye SD, Lv W, He K, Li WX. Comprehensive integrated analysis of gene expression datasets identifies key anti cancer targets in different stages of breast cancer. *Exp Ther Med.* 2018;16(2):802-810.
- Chen WX, Yang LG, Xu LY, et al. Bioinformatics analysis revealing prognostic significance of RRM2 gene in breast cancer. *Biosci Rep.* 2019;39(4):BSR20182062.
- Qi L, Zhou B, Chen J, et al. Significant prognostic values of differentially expressed-aberrantly methylated hub genes in breast cancer. *J Cancer.* 2019;10(26):6618-6634.
- Liu R, Guo CX, Zhou HH. Network-based approach to identify prognostic biomarkers for estrogen receptor-positive breast cancer treatment with tamoxifen. *Cancer Biol Ther.* 2015;16(2):317-324.
- Zhang X, Pan Y, Fu H, Zhang J. Nucleolar and spindle associated protein 1 (NUSAP1) inhibits cell proliferation and enhances susceptibility to epirubicin in invasive breast cancer cells by

- regulating cyclin D kinase (CDK1) and DLGAP5 expression. *Med Sci Monit.* 2018;24:8553-8564.
33. Chen X, Thiaville MM, Chen L, et al. Defining NOTCH3 target genes in ovarian cancer. *Cancer Res.* 2012;72(9):2294-2303.
34. Shi YX, Yin JY, Shen Y, Zhang W, Zhou HH, Liu ZQ. Genome-scale analysis identifies NEK2, DLGAP5 and ECT2 as promising diagnostic and prognostic biomarkers in human lung cancer. *Sci Rep.* 2017;7(1):8072.
35. Schneider MA, Christopoulos P, Muley T, et al. AURKA, DLGAP5, TPX2, KIF11 and CKAP5: Five specific mitosis-associated genes correlate with poor prognosis for non-small cell lung cancer patients. *Int J Oncol.* 2017;50(2):365-372.
36. Tan GZ, Li M, Tan X, Shi ML, Mou K. MiR-491 suppresses migration and invasion via directly targeting TPX2 in breast cancer. *Eur Rev Med Pharmacol Sci.* 2019;23(22):9996-10004.
37. Huang C, Han Z, Wu D. Effects of TPX2 gene on radiotherapy sensitization in breast cancer stem cells. *Oncol Lett.* 2017;14(2):1531-1535.
38. Yang Y, Li DP, Shen N, et al. TPX2 promotes migration and invasion of human breast cancer cells. *Asian Pac J Trop Med.* 2015;8(12):1064-1070.
39. Tomii C, Inokuchi M, Takagi Y, et al. TPX2 expression is associated with poor survival in gastric cancer. *World J Surg Oncol.* 2017;15(1):14.