Research Article

Construction and Validation of a Newly Prognostic Signature for CRISPR-Cas9-Based Cancer Dependency Map Genes in Breast Cancer

Xin Yan [ID],[1] **Sai-Nan You** [ID],[2] **Yan Chen** [ID],[3] **and Ke Qian** [ID][1]

[1]*Liyang People's Hospital, Liyang 213399, China*
[2]*Suqian Hospital of Nanjing Drum Tower Hospital Group, Suqian 223800, China*
[3]*Jiangsu Province Geriatric Hospital, Nanjing 210029, China*

Correspondence should be addressed to Yan Chen; chenyanwws@sina.com and Ke Qian; 13813657139@163.com

Cancer Dependency Map (CDM) genes comprise an extensive series of genome-scale RNAi-based loss-of-function tests; hence, it served as a method based on the CRISPR-Cas9 technique that could assist scientists in investigating potential gene functions. These CDM genes have a role in tumor cell survival and proliferation, suggesting that they may be used as new therapeutic targets for some malignant tumors. So far, there have been less research on the involvement of CDM genes in breast cancer, and only a tiny percentage of CDM genes have been studied. In this study, information of patients with breast cancer was extracted from The Cancer Genome Atlas (TCGA), from which differentially expressed CDM genes in breast cancer were determined. A variety of bioinformatics techniques were used to assess the functions and prognostic relevance of these confirmed CDM genes. In all, 290 CDM genes were found differentially expressed. Six CDM genes (SRF, RAD51, PMF1, EXOSC3, EXOC1, and TSEN54) were found to be associated with the prognosis of breast cancer samples. Based on the expression of the identified CDM genes and their coefficients, a prognosis model was constructed for prediction, according to which patients with breast cancer were separated into two risk groups. Those with high risk had substantially poorer overall survival (OS) than patients in the other risk group in the TCGA training set, TCGA testing set, and an external cohort from Gene Expression Omnibus (GEO) database. The area under the receiver operating characteristic (ROC) curve for this prognostic signature was, respectively, 0.717 and 0.635 for TCGA training and testing sets, demonstrating its reliability in predicting the prognosis of patients with breast cancer. We next created a nomogram using the six CDM genes discovered to create a therapeutically useful model. The Human Protein Atlas database was used to acquire all immunohistochemistry staining images of the discovered CDM genes. The proportions of tumor-infiltrating immune cells, as well as the expression levels of checkpoint genes, varied substantially between the two risk groups, according to the analyses of immune response. In conclusion, the findings of this research may aid in the understanding of the prognostic value and biological roles of CDM genes in breast cancer.

1. Introduction

Breast cancer is one of the most prevalent malignancies in women [1]. Meanwhile, male breast cancer accounts for approximately 1% of all breast cancer cases globally, and clinical data suggest that it has been on the increase in recent years [2]. Breast cancer screening has been shown to decrease mortality from the disease [3]. Patients with breast cancer have a five-year survival rate of almost 98% if the illness is identified early [4]. Although the survival rate for breast cancer patients has improved, the survival rate for advanced-stage patients remains low, particularly for patients with locally advanced breast cancer [5], which affects up to 20% of patients with breast cancer [6]. Hence, understanding the molecular processes that cause breast cancer remains an important challenge for us to develop new treatments [7].

The Broad Institute and the Dana-Farber Cancer Institute started a project called "Defining a Cancer Dependency Map" to find genes that encouraged cancer cell

development [8]. This significant research demonstrated the significance of certain genes in the development and proliferation of cancer cells. These certain genes may be used as targets for the development of new targeted medicines. More than 500 human cancer cell lines that were often used to investigate the impact of shutting off specific genes on proliferation and growth for their ability of growing indefinitely were examined in this research. In detail, cancer cell lines were transformed with a lentiviral vector expressing Cas9 nuclease under cystatin selection (pXPR-311Cas9) [9]. Cas9 activity assays were performed on each Cas9-expressing cell line to characterize the effect of CRISPR/Cas9 on these cell lines. Cell lines that detected Cas9 activity below 45% were not eligible for further screening. Although the majority of these identified certain genes were cancer-specific, approximately 10% of them were shown to be involved in various malignant tumors, indicating that they had important biological roles. This research also found that studying gene activity patterns rather than concentrating on whether a particular gene was faulty was the best method to anticipate this dependency [10]. More relevant genes were identified and uploaded to the website (https://depmap.org/portal/) as the study progressed and real-time updates were made [9]. A total of 1246 CDM genes were included in the study in the most recent update of the dataset, and they were shown to be frequent and important in the incidence and development of different malignancies.

In this research, breast cancer data from TCGA was downloaded, and differentially expressed CDM genes were identified, as well as their potential roles and processes. A CDM gene-based prognosis signature was then created and further validated, based on which, some CDM genes may be utilized as possible prognostic biomarkers.

2. Materials and Methods

2.1. Data Collection and Processing. RNA-sequencing and clinical feature information of breast cancer were all downloaded online from a database named The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/). A total of 1,246 CDM genes were acquired from a project named Cancer Dependency Map online (https://depmap.org/portal/) [9]. CDM genes with FDR adjusted P values <0.05 along with the log2|fold change| values >0.5 were identified as candidate ones since they were significantly differentially expressed between breast cancer patients and normal samples in TCGA training set using packages of R software (version 4.0.5). Volcano plots and heatmaps were, respectively, visualized with the "ggplot2" [11] and "pheatmap" [12] packages in R software. The TCGA testing set was used to validate the results from training set. An external validation dataset of breast cancer, which included transcription profiles and clinical information, were downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53031).

2.2. Enrichment and Pathway Analyses. To explore the function of identified prognosis-related CDM genes, function and pathway enrichment analyses were carried out.

Pearson correlation test was also used to assess those candidate CDM genes. Genes were further classified via the package named "clusterProfiler" [13] in R software according to the projection of GO [14] (http://www.geneontology.org/) or KEGG [15] (http://www.genome.jp/kegg/pathway.html) pathways at a specific level. Of course, GO terms or KEGG pathway with a corrected-P < 0.05 were considered as significant enrichment.

2.3. Construction and Validation of a CDM Gene-Based Risk Model. The samples in the breast cancer cohort from TCGA were randomly separated into two sets: a training set containing 524 samples and a testing set containing 522 samples. Univariate Cox regression analysis was carried out on those identified prognosis-related CDM genes using R software. The log-rank test was performed to analyze the statistical significance. Multivariate Cox regression analysis was further carried out on the genes chosen by univariate Cox regression analysis to identify ones with independently prognostic value. The six CDM genes that were finally identified by multivariate Cox regression analysis were then used to construct a risk model using their expression levels and coefficient values. The risk score (RS) was calculated for each sample in TCGA training set, the median one of which was chosen as cut-off. According to the cut-off value, patients with breast cancer were divided into the high-risk group or low-risk group. Overall survival (OS) and receiver operating characteristic (ROC) curve analyses were carried out by packages in R software to evaluate the predictive value of the risk model in breast cancer, which was further confirmed by the TCGA testing set. Besides, we established a nomogram based on the constructed risk model and independent clinical factors, where significance was taken at P < 0.05.

2.4. Human Protein Atlas Database Analysis. To investigate the protein expression levels of the hub CDM genes involved in the signature in human normal and breast cancer tissues, we downloaded the representative images of the immunohistochemical assay from the Human Protein Atlas (http://www.proteinatlas.org) [16].

2.5. Immunity-Related Analysis. We further used the Tumor Immune Estimation Resource (TIMER) [17], CIBERSORT [18], QUANTISEQ [19], XCELL [20], MCP counter [21], and Estimating the Proportion of Immune and Cancer cells (EPIC) [22] to comparatively assess cell immune responses or cellular components between two different risk groups. Moreover, single-sample GSEA (ssGSEA) was utilized to compare and quantify the subgroups of the tumor-infiltrating immune cells, as well as their immunological function, in two groups with different risks. Previous research yielded a list of possible immunological checkpoints.

2.6. Statistical Analysis. Chi^2 tests were used to examine relationships between categorical and continuous variables in testing and training sets from TCGA. Univariate Cox proportional hazards regression (CPHR) analysis and the

Kaplan–Meier technique were used to confirm candidate CDM genes that were substantially linked to the OS of patients with breast cancer. Stepwise multivariate CPHR studies were used to identify both important clinical factors and the risk score calculation. Survival curves were produced using the Kaplan–Meier technique, which were then evaluated using the log-rank test method. A time-dependent ROC curve was used to evaluate the risk model's specificity and sensitivity in predicting patients' survival. The t-test was carried out to compare the differences between two risk groups in training and testing sets. R software with packages was used to conduct the statistical analyses in this research. Statistical significance was determined when the P value was less than 0.05.

3. Results

3.1. Differentially Expressed CDM Genes and Related Pathway Enrichment Analyses in Breast Cancer.

We first analyzed the expression levels of 1246 CDM genes between the patients and controls to identify differentially expressed ones. As shown in Figure 1(a), a total of 290 CDM genes were chosen, which composed of 116 downregulated genes and 174 upregulated ones. The heatmap in Figure 1(b) revealed the 290 differentially expressed genes between patients and controls with FDR<0.05 and log2|fold change|>0.5. To investigate these identified CDM genes' function and related mechanisms, we carried out functional analyses of them, which mainly concluded KEGG and GO pathway enrichment analyses. As shown in Figures 1(c) and 1(d), downregulated CDM genes were enriched in RNA splicing, translational initiation, and so on, while upregulated CDM genes were enriched in DNA replication, chromosome segregation, nuclear division, and so on. GO enrichment analysis revealed that those downregulated genes mainly enriched in RNA transport and citrate cycle, and upregulated genes enriched in cell cycle, DNA replication, and homologous recombination (Figures 1(e) and 1(f)).

3.2. Identification of Prognostic-Related CDM Genes and Construction of the Prognostic Model.

To explore the effects of CDM genes in predicting the prognosis of patients with breast cancer, we studied the relationship between the 290 differentially expressed CDM genes and the OS of patients with breast cancer by univariate Cox regression analysis. As shown in Figure 2(a), 22 CDM genes were selected as candidate genes for their significantly association with the OS of breast cancer patients. Then, we further carried out the multivariate Cox analysis and identified six hub genes with an independent prognostic value (Figure 2(b)). Based on the six candidate hub genes, we established a model to predict the prognosis for those patients with breast cancer: risk score = coef (i) * exp (i), where coef (i) meant the multivariate Cox regression coefficient and exp (i) meant the expression level. We assessed the risk score of each sample from the TCGA training set, and the median value was identified as the cut-off, based on which patients with breast cancer were divided into the low-risk or high-risk group.

3.3. Validation of the Risk Model in the TCGA and GEO Databases.

To evaluate the prognostic value of the constructed risk model, we performed Kaplan–Meier survival analysis and found that patients with high risk in the training set had poorer prognosis than those with low risk (Figure 3(a)). The time-dependent ROC analysis showed the area under the ROC curve (AUC) was 0.717 at 1 year (Figure 3(b)), which indicated the good prognostic value of the identified model in the TCGA training set. The heatmap of the hub 6 CDM genes and patients' survival status in the training set are shown in Figures 3(c) and 3(d). To further explore the predictive value of the identified risk model composed of 6 CDM genes in the TCGA testing set, we carried out survival analysis and ROC analysis. As shown in Figure 4(a), patients with lower risk scores in the testing set had significantly longer survival time than those with higher risk scores. Also, the AUC was 0.635 at 1 year in the testing set (Figure 4(b)). The survival status and risk scores of patients in the testing set, along with the expression heatmap of the 6 hub genes in testing set are shown in Figures 4(c) and 4(d).

Further validation of the risk model on the patients with breast cancer by GSE53031 cohort from the GEO database were carried out, and the results (Figure S1) showed that the six identified CDM genes were closely linked to the prognosis of patients not only from TCGA but also from an external dataset from GEO. To further validate the efficiency of the signature based on CDM genes in predicting the survival, we compared the constructed signature with another two signatures reported recently in the TCGA-breast cancer cohort. The nine-gene signature derived from Wang's research (WangSig) was built based on the analyses of the ferroptosis-related information from breast cancer [23]. However, the six-gene signature from Mo's study (MoSig) was constructed based on integrated multiomics data analysis [24]. As shown in Figure S2, CDMSig had a higher AUC value (AUC = 0.684) than that of the WangSig (AUC = 0.587) and the MoSig (AUC = 0.478), suggesting that the CDM gene-based signature we built had a good predictive performance.

3.4. Establishment of a Nomogram Combining with Clinical Characteristics.

Considering that the constructed risk model was significantly associated with poor prognosis, we combined the OS of breast cancer patients with their clinical features and performed univariate and multivariate Cox analysis to explore whether the signature we constructed could serve as an independent factor to predict the prognosis. As shown in Figure 5, all the P values were less than 0.001, indicating that the signature had an independent prognostic value. The heatmap in Figure 6(a) showed the association between the six identified CDM genes in the prognostic risk model and clinical manifestations of all breast cancer samples from TCGA. Then, by combing all these clinical features and the constructed signature, we established a nomogram to expand the application and availability of the constructed risk model in clinic (Figure 6(b)).
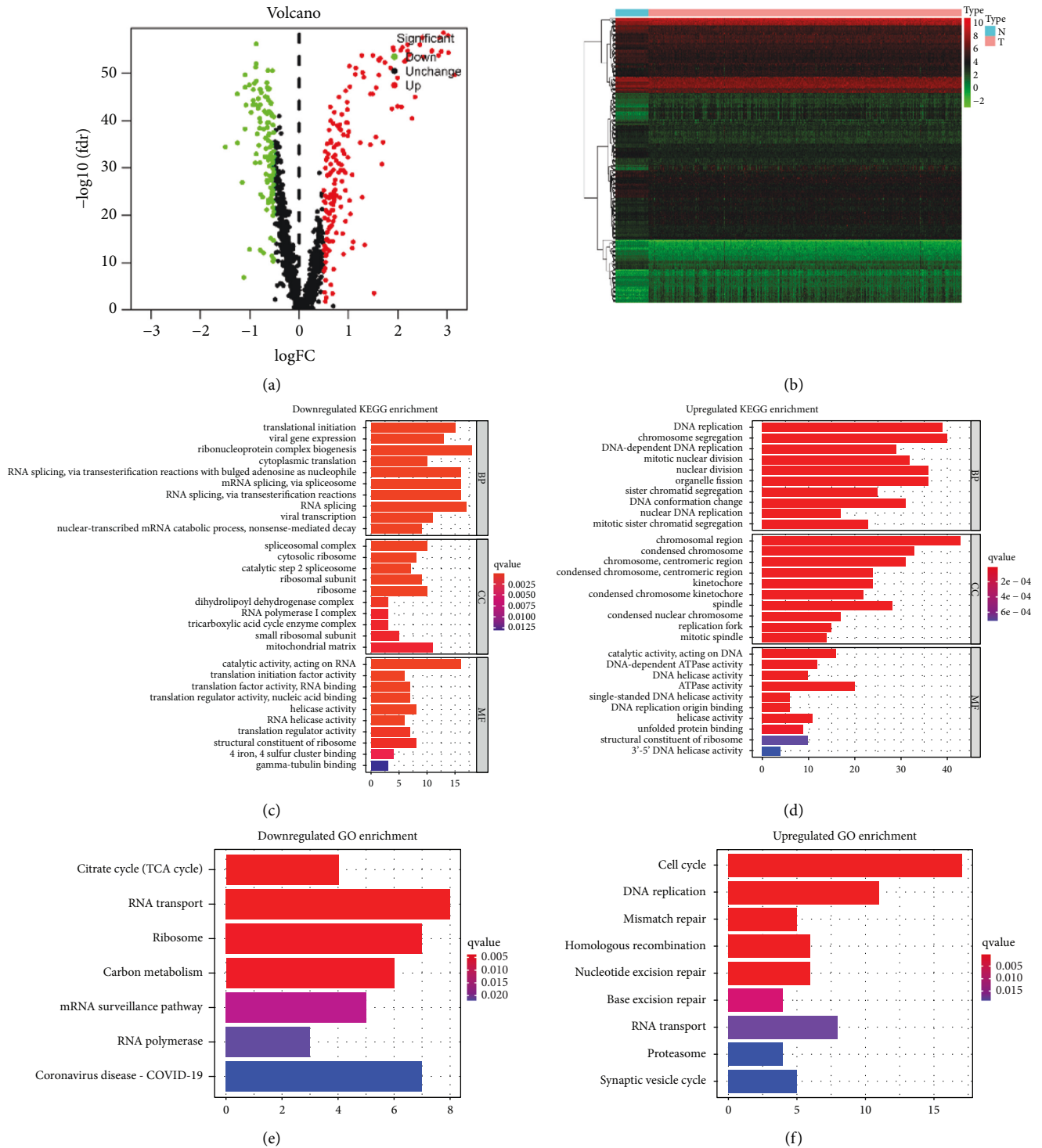
(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 1: Identification of differentially expressed Cancer Dependency Map (CDM) genes in breast cancer and pathway enrichment analyses. (a) Volcano plot of CDM genes in patients with breast cancer from the TCGA database. (b) Heatmap of CDM genes. KEGG pathway enrichment of those downregulated CDM genes (c) and those upregulated ones (d). GO pathway enrichment of those downregulated CDM genes (e) and those upregulated ones (f).
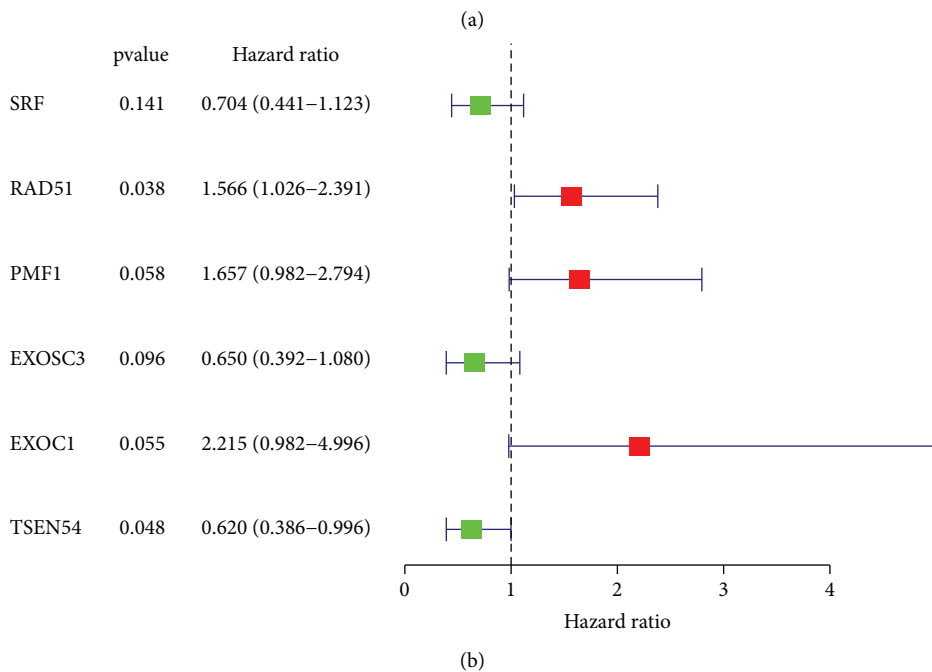
	pvalue	Hazard ratio
TXN	0.026	1.398 (1.042−1.877)
SRF	0.011	0.663 (0.484−0.909)
RPL31	0.006	0.652 (0.479−0.887)
ANAPC11	0.012	0.706 (0.538−0.928)
PLK1	0.025	1.247 (1.028−1.513)
PSMD14	0.013	1.760 (1.126−2.751)
ANLN	0.029	1.210 (1.020−1.437)
RACGAP1	0.019	1.347 (1.051−1.726)
TBCB	0.017	0.684 (0.501−0.935)
RPS15A	0.049	0.718 (0.516−0.999)
NSF	0.039	1.394 (1.017−1.910)
RAD51	0.025	1.377 (1.040−1.822)
PAFAH1B1	0.037	1.484 (1.024−2.151)
PMF1	0.026	0.724 (0.546−0.961)
MASTL	0.026	1.380 (1.039−1.831)
EIF3D	0.027	0.659 (0.455−0.954)
POLR1E	0.045	0.721 (0.524−0.992)
RPS29	0.007	0.666 (0.495−0.896)
RPL13	0.003	0.680 (0.529−0.875)
EXOSC3	0.027	0.673 (0.475−0.955)
EXOC1	0.013	1.866 (1.139−3.058)
TSEN54	0.028	0.694 (0.501−0.962)

(a)

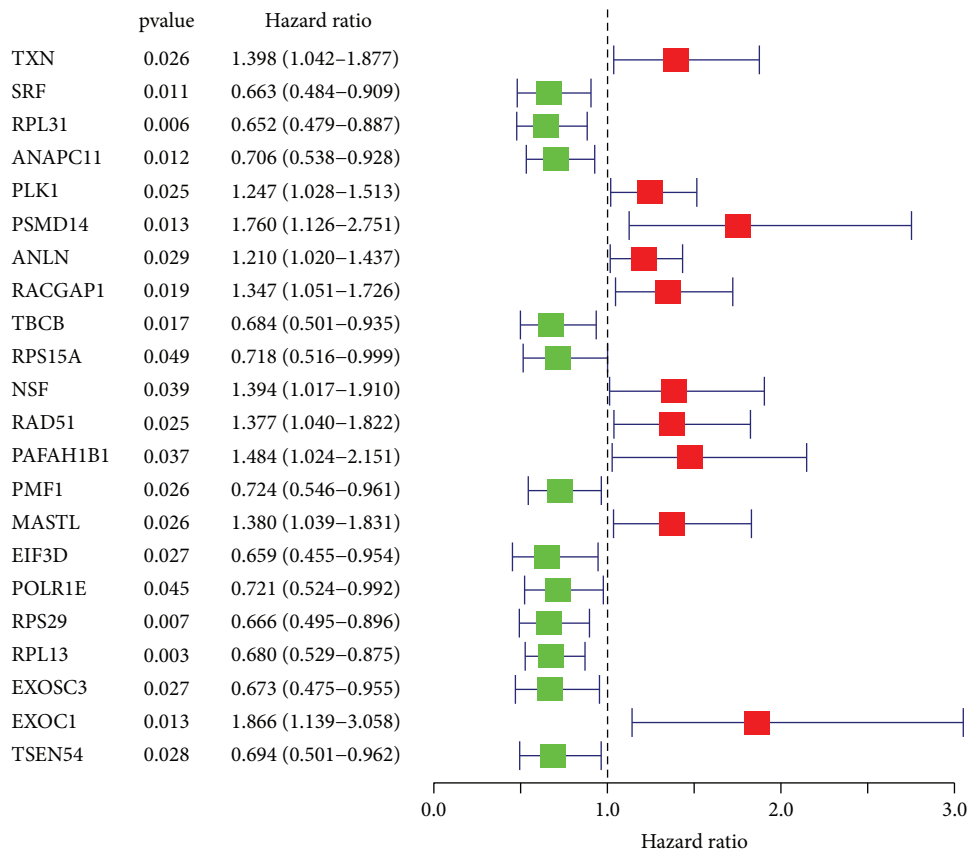|	pvalue	Hazard ratio
SRF	0.141	0.704 (0.441−1.123)
RAD51	0.038	1.566 (1.026−2.391)
PMF1	0.058	1.657 (0.982−2.794)
EXOSC3	0.096	0.650 (0.392−1.080)
EXOC1	0.055	2.215 (0.982−4.996)
TSEN54	0.048	0.620 (0.386−0.996)

(b)

FIGURE 2: Identification of CDM genes with the predictive value and construction of risk model for breast cancer. Univariate (a) and multivariate (b) Cox regression analyses on the differentially expressed CDM genes.

(a)

AUC at 1 years: 0.717
AUC at 3 years: 0.682
AUC at 5 years: 0.674
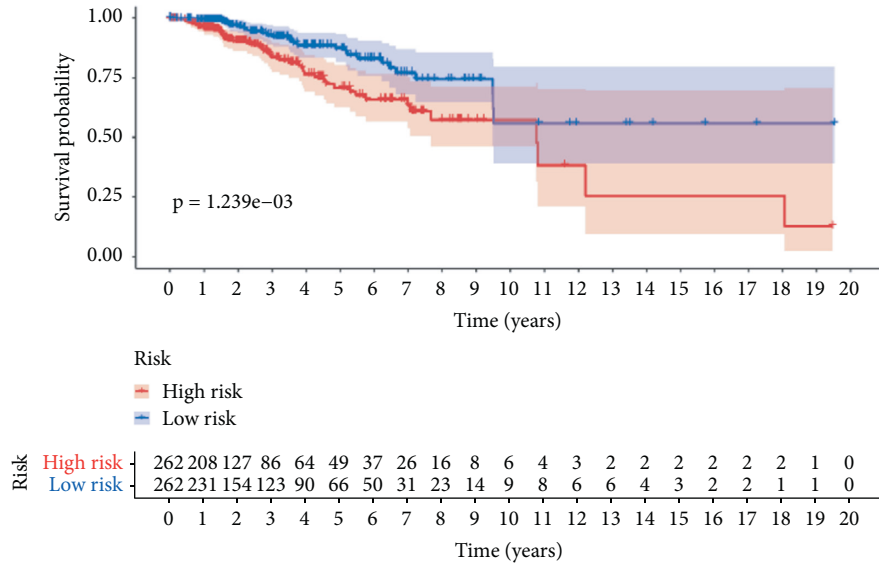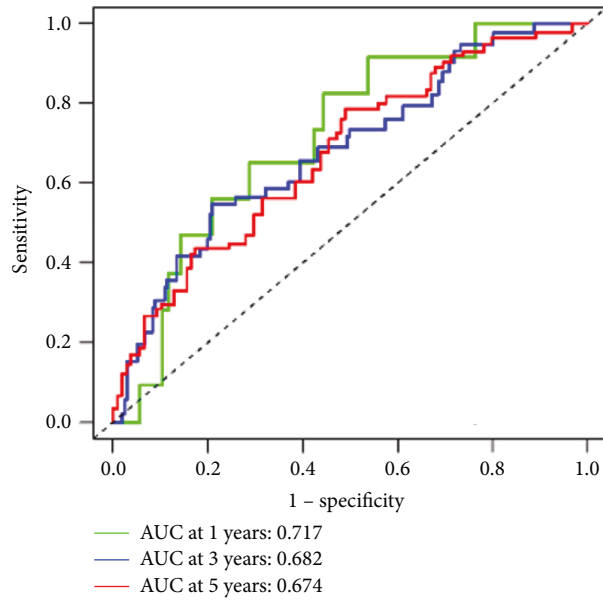
(b)
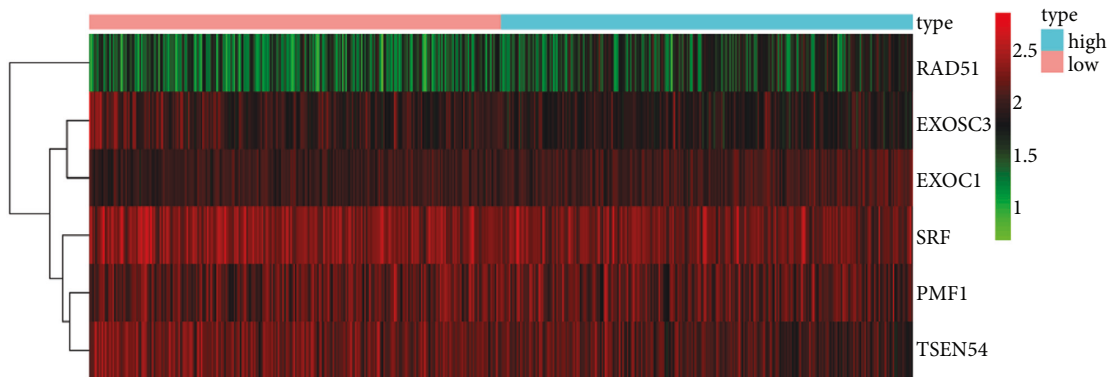
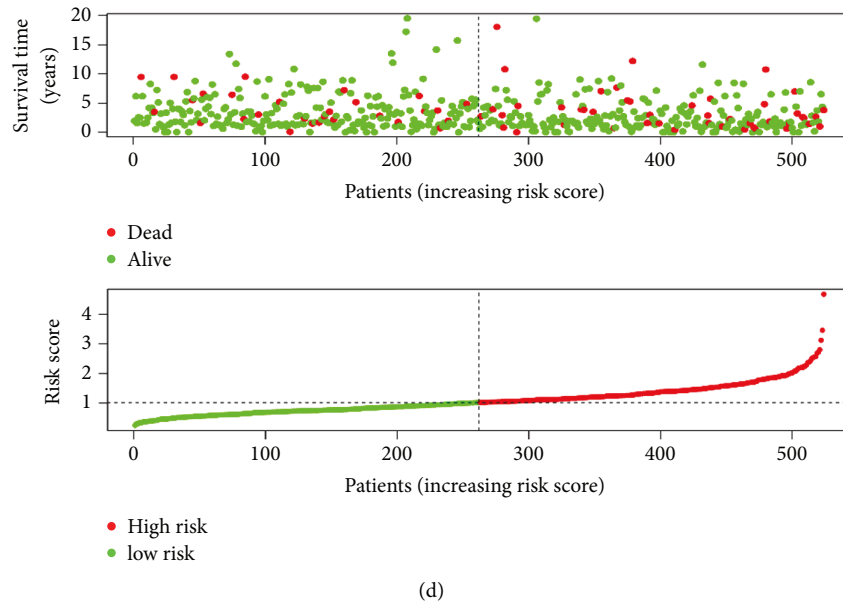(c)

Figure 3: Continued.

(d)

FIGURE 3: Validation of the six CDM gene-based signature in the TCGA training set. (a) The Kaplan–Meier curve for the overall survival (OS) of breast cancer samples based on the constructed signature in the TCGA training set. (b) Time-dependent receiver operating characteristic (ROC) curve analysis of the signature for predicting OS of patients with breast cancer. (c) Heatmap of the signature in the TCGA training set. (d) Risk score distribution and OS status in the TCGA training set.

3.5. Protein Expression Levels of Identified Hub CDM Genes. We used the Human Protein Atlas online to evaluate the protein expression levels of the 6 CDM genes in breast cancer sample and normal human mammary tissue. As shown in Figure 7, SRF and EXOSC3 expression levels were lower in tumor tissues than in control tissues, while RAD51 and EXOC1 were significantly higher in tumor tissues. These results indicated that RAD51 and EXOC1 might be pathogenic factors, while SRF and EXOSC3 might be protective factors in breast cancer, which were consistent with their coefficients in the risk model.
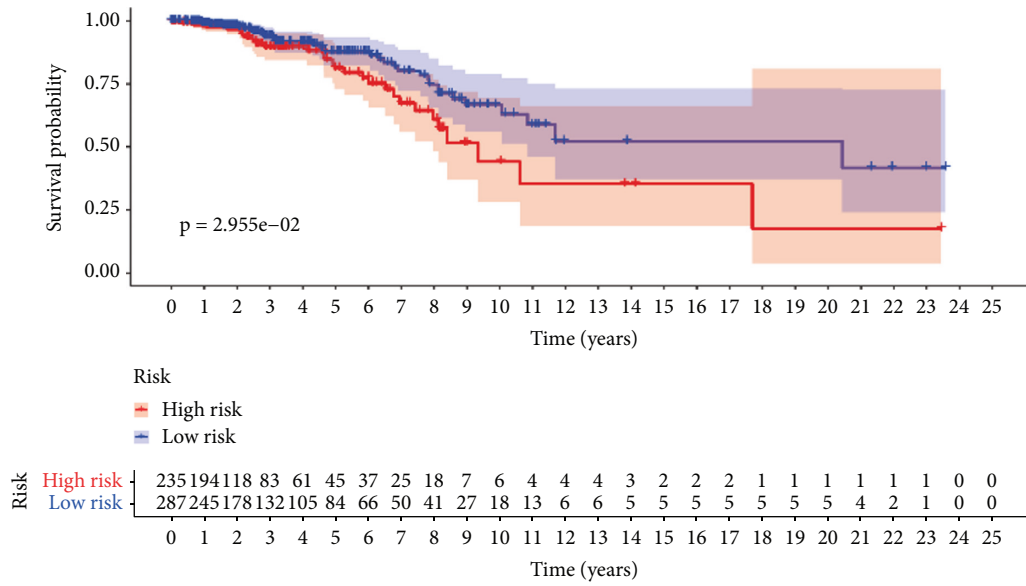
3.6. Immune-Related Analysis of the Established Risk Model. As the immune status has been reported significantly associated with breast cancer [25], we further explored the immune responses between two risk groups via TIMER algorithms, CIBERSORT, QUANTISEQ, XCELL, MCP counter, and EPIC (Figure 8). The enrichment fractions of immune cells were analyzed to explore the association between the immune status and risk scores in patients with breast cancer from the TCGA database. As shown in Figure 9, the antigen presentation process contents, including HLA, B cells, NK cells, and mast cells, were significantly declined in the high-risk group. Also, comparing with the patients in the low-risk group, the scores of Tfh cells, T cell co-stimulation, and CD8+ T cells were obviously lower in the patients with higher risk scores, which indicated the difference in the regulation of T cells between the two risk groups. In addition, the fractions of checkpoints, CCR, and the activity of type II IFN response were lower in patients in the high-risk group, while the macrophages scores showed the opposite. Furthermore, we explored the expression levels

of immune checkpoints and found that LAG3, CD27, PDCD1, CD40, and TNFRSF14 were significantly different between two groups with different risk scores (Figure 10).
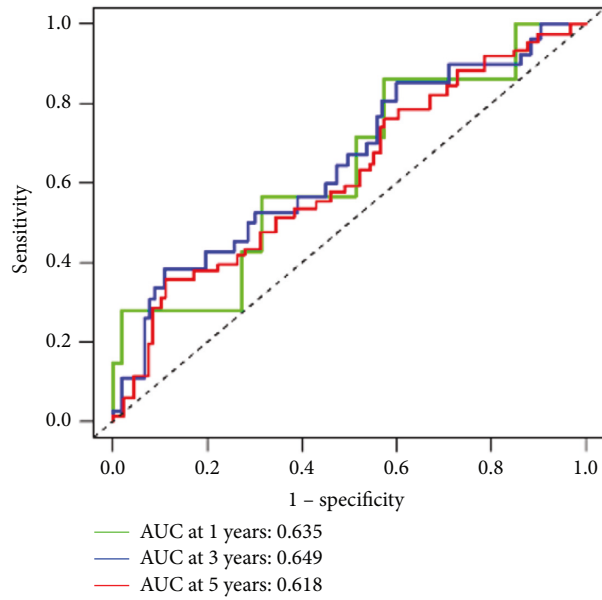
4. Discussion

It still remains difficult for us to identify genes whether they are essential for the survival of tumor, for that most tumors especially malignant ones are closely associated with genetic mutations. These genetic mutations are often related with the cell growth and specific vulnerabilities to specific damages [26]. It has been reported that some of these mutations have potential as therapeutic targets [27]. Hence, the challenge nowadays for us is how to identify specific targetable vulnerabilities for every cancer by using current tools [28]. Scientists have profiled about hundreds of cell models to elucidate genomic information. Large-scale databases have also been searched for hoping that the identification of genetic targets would help a lot in developing novel therapies, identifying the certain treatments for patients and allowing physicians to better understand the weaknesses of cancers. Actually, a project named Cancer Dependency Map, has committed to remain freely available to the public. The related information could be downloaded online freely (https://depmap.org/portal/depmap/). CDM has been used as a novel method to explore the priority targets, as well as the drug sensitivity to different cancers. Hence, it will help scientists discover new targeted therapies as soon as possible, thus facilitating progress in precision therapy [29].
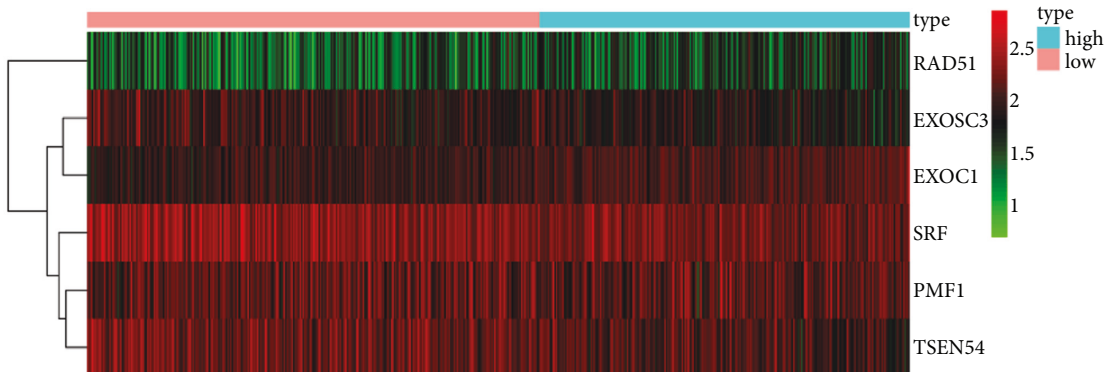
To date, studies on the roles of CDM genes are scarce [30, 31]. In this study, 1246 CDM genes were analyzed and a total of 290 differentially expressed CDM genes had been

p = 2.955e−02

Risk

High risk
Low risk

Risk	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
High risk	235	194	118	83	61	45	37	25	18	7	6	4	4	4	3	2	2	2	1	1	1	1	1	1	0	0
Low risk	287	245	178	132	105	84	66	50	41	27	18	13	6	6	5	5	5	5	5	5	5	4	2	1	0	0

Time (years)

(a)

— AUC at 1 years: 0.635
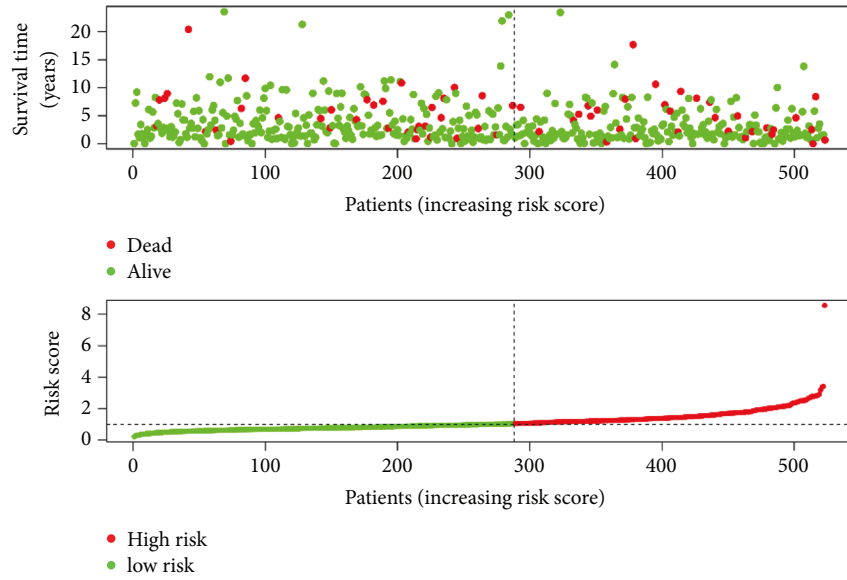— AUC at 3 years: 0.649
— AUC at 5 years: 0.618

(b)

(c)

FIGURE 4: Continued.

(d)

FIGURE 4: Validation of the six CDM gene-based signature in the TCGA testing set. (a) Kaplan–Meier curves for OS of patients with breast cancer in the TCGA testing set. (b) Time-dependent ROC curve analysis for predicting OS of breast cancer samples. (c) Heatmap of the signature in the TCGA testing set. (d) Risk score distribution and OS status in the TCGA testing set.
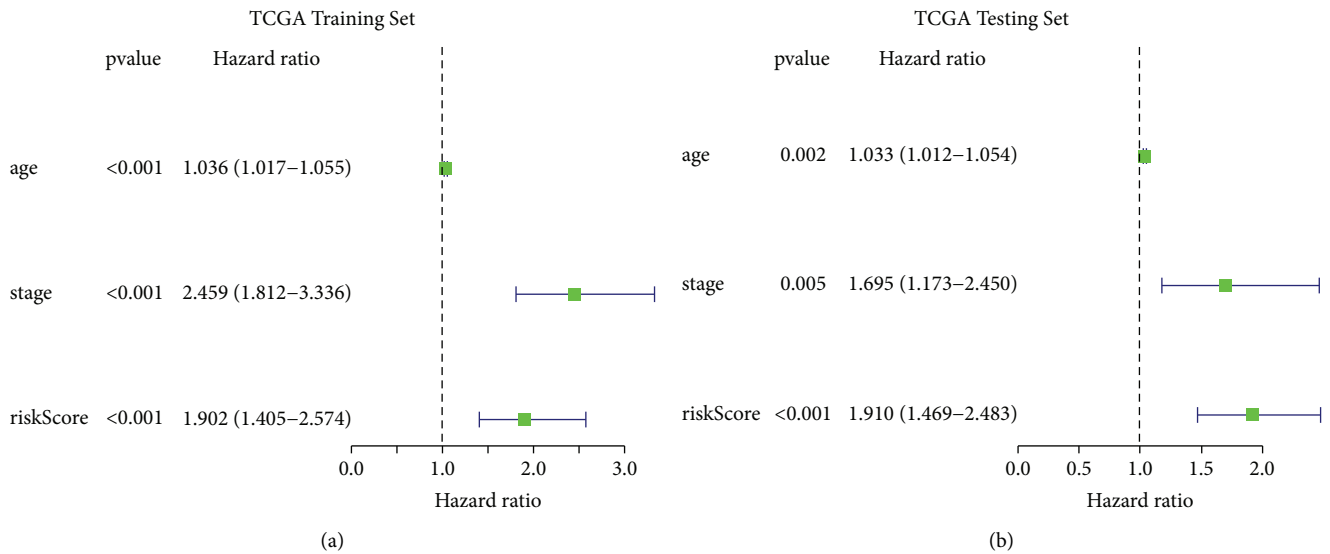

FIGURE 5: Continued.

TCGA Training Set

	pvalue	Hazard ratio	
age	<0.001	1.039 (1.020–1.059)	
stage	<0.001	2.490 (1.868–3.319)	
riskScore	<0.001	2.019 (1.493–2.730)	

Hazard ratio

(c)

TCGA Testing Set

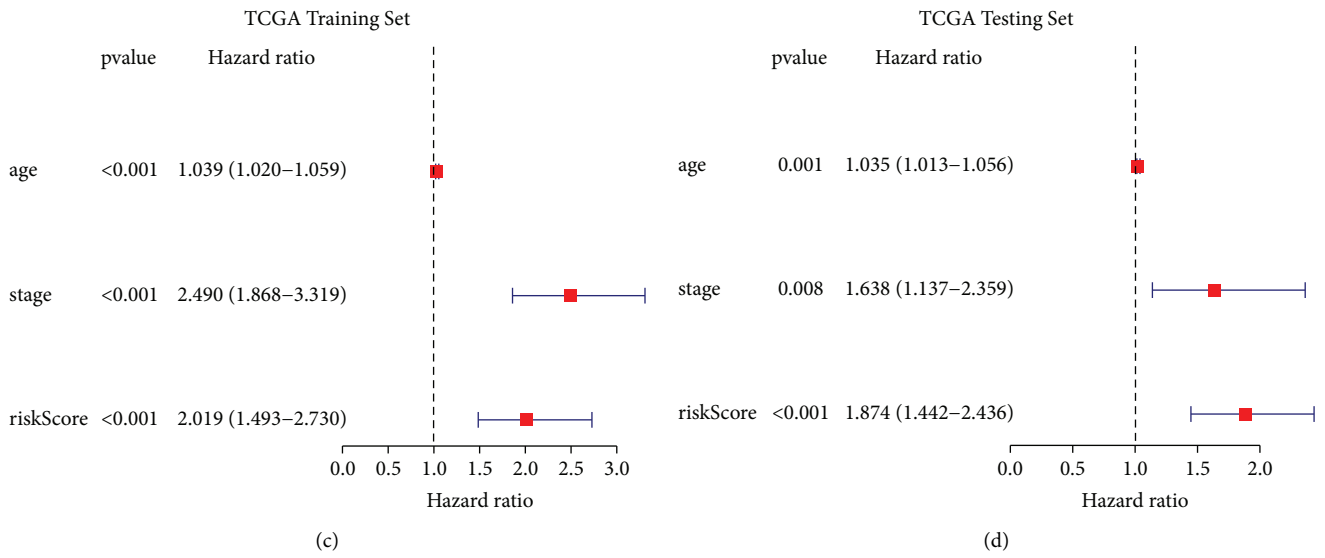|	pvalue	Hazard ratio	
age	0.001	1.035 (1.013–1.056)	
stage	0.008	1.638 (1.137–2.359)	
riskScore	<0.001	1.874 (1.442–2.436)	

Hazard ratio

(d)

FIGURE 5: Validation for the prognostic value of the constructed risk model along with different clinical characteristics. Univariate (a) and multivariate (c) COX regression analysis of the signature and different clinical features in the TCGA training set. Univariate (b) and multivariate (d) COX regression analysis of the signature and different clinical features in the TCGA testing set.
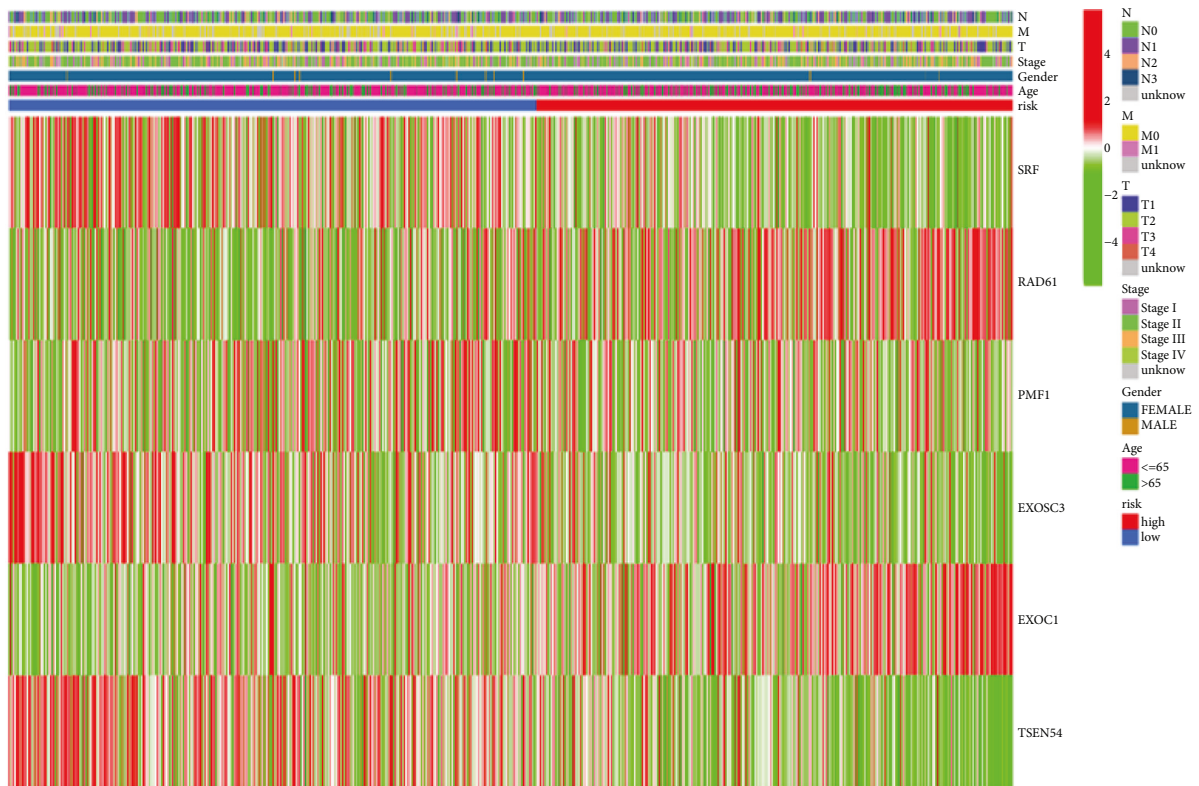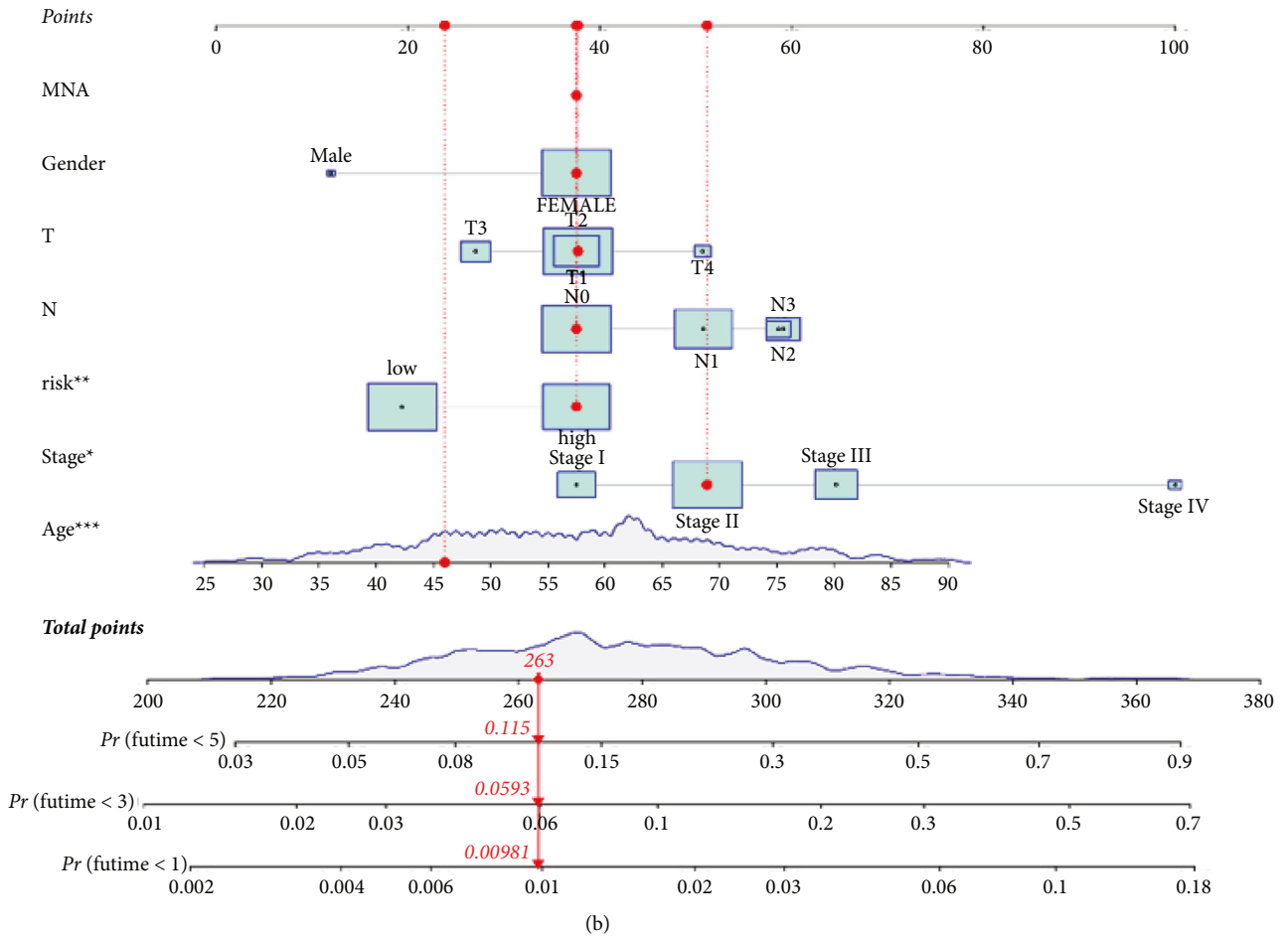

(a)

FIGURE 6: Continued.

FIGURE 6: Construction of a nomogram for prediction. (a) Heatmap for the CDM gene-based signature with clinicopathological manifestations in TCGA-breast cancer set. (b) Nomogram for predicting 1-, 3-, and 5-year OS of patients with breast cancer.

identified. The univariate and multivariate Cox regression analysis revealed that six CDM genes (SRF, RAD51, PMF1, EXOSC3, EXOC1 and TSEN54) were associated with the survival of patients with breast cancer. Hence, a CDM gene-related risk model was established in the training set based on the six genes and was further validated in TCGA testing set. It was further proven by ROC analysis to be reliable in predicting the prognosis of breast cancer patients. In addition, we constructed a nomogram to improve clinical practicality of the constructed risk model for patients with breast cancer. These results might make it easier for us to understand the mechanisms of breast cancer and help us explore new biomarkers for patients with breast cancer. Functional and enrichment pathway analyses revealed that upregulated CDM genes mainly took part in DNA replication, nuclear division, chromosome segregation, cell cycle, DNA replication, and homologous recombination, while downregulated CDM genes were mainly involved in RNA splicing, translational initiation, RNA transport, and citrate cycle. During the past few years, it has been reported that abnormal metabolism [32–34] and processing [35–37] of DNA play critical roles in many neoplastic diseases.

The results of the multivariate Cox regression analysis of the six identified CDM genes revealed that all of them had

prognostic value for breast cancer. Consistent with our findings, He et al. [38] demonstrated that SRF could regulate the promoter activity of HOTAIR, a negative prognostic factor for breast cancer, thus affecting the invasiveness and proliferation of breast cancer cells. The overexpression of RAD51 has been found in various cancers, including breast cancer, and it is revealed closely associated with poor survival [39]. Also, by combining with Nrf-2, PMF1 could regulate the expression of SSAT [40], which affects the apoptotic cell death in breast cancer cells [41]. Although there has been few researches on the roles of EXOSC3 and EXOC1 in breast cancer, the results from the Human Protein Atlas (Figure 8) showed that the expression of EXOSC3 and EXOC1 were significantly different between tumor tissues and control tissues, and their expression levels showed the same trend with the results of multivariate Cox regression analysis.

As breast cancer is characterized by tumor heterogeneity [42, 43], which is also the main reason that leads to the complexity of treatment [44], a more reliable model with prediction value is urgently needed. Although there have been several signatures constructed for predicting the prognosis of patients with breast cancer [45–47], the relationship between these indicators and breast cancer remains
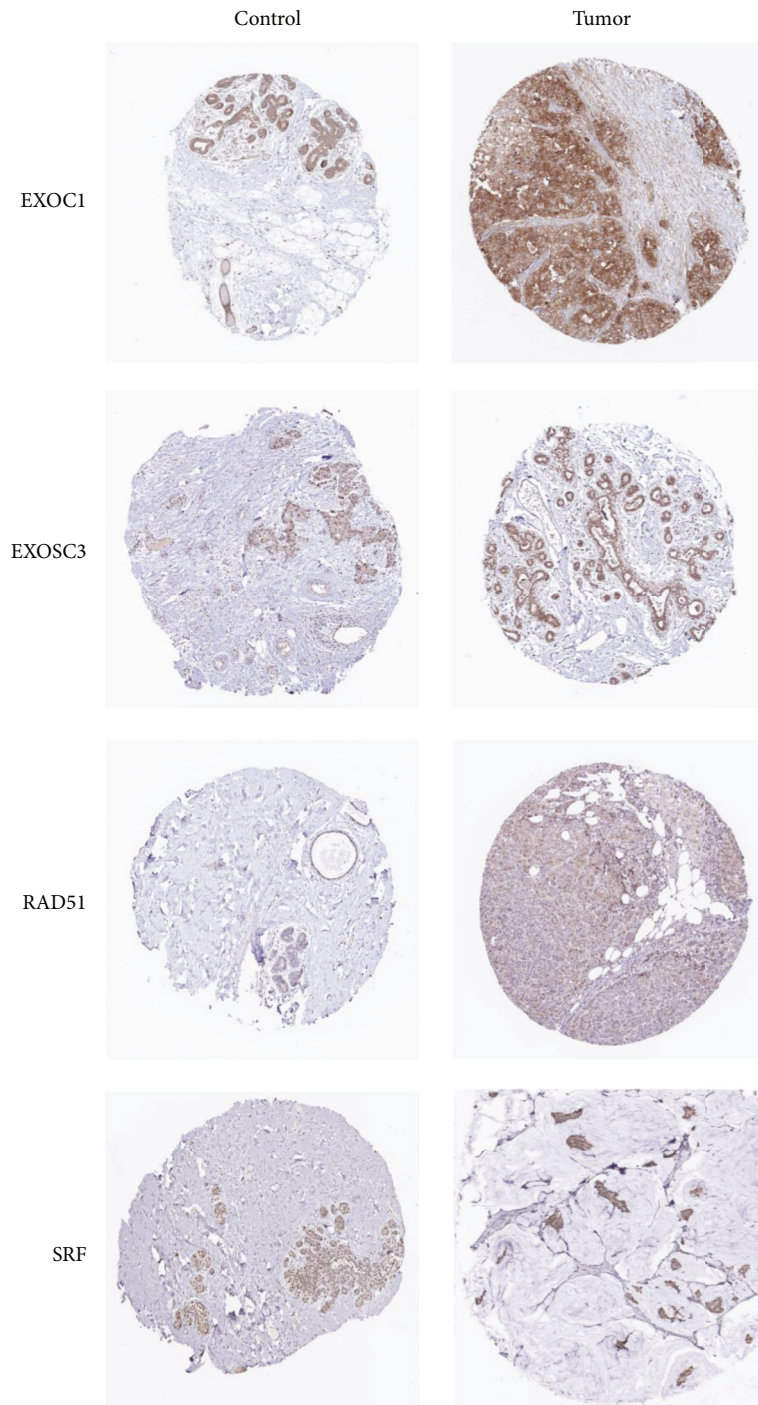
FIGURE 7: The representative images of the immunohistochemical images of identified CDM genes from the Human Protein Atlas database.

unclear. Then, using the six identified CDM genes, we established a risk model with prognostic value for patients with breast cancer by combining the data from the TCGA training set. The ROC analysis demonstrated that the constructed signature could select out patients with poor prognosis by calculating their values. The ROC analysis based on data from the TCGA testing set showed the same results. Further confirmation of the reliability of the constructed signature was done by Kaplan–Meier analysis.

What is more, the multiple Cox regression analyses in the TCGA training set and testing set both revealed that the CDM gene-based signature could serve as an independent indicator for predicting the prognoses of patients with breast cancer. All of the findings aforementioned pointed out the CDM gene-based prediction risk model had clinical practicality. A nomogram was further constructed to provide an easier method to enable the physicians predict the survival of patients with breast cancer.
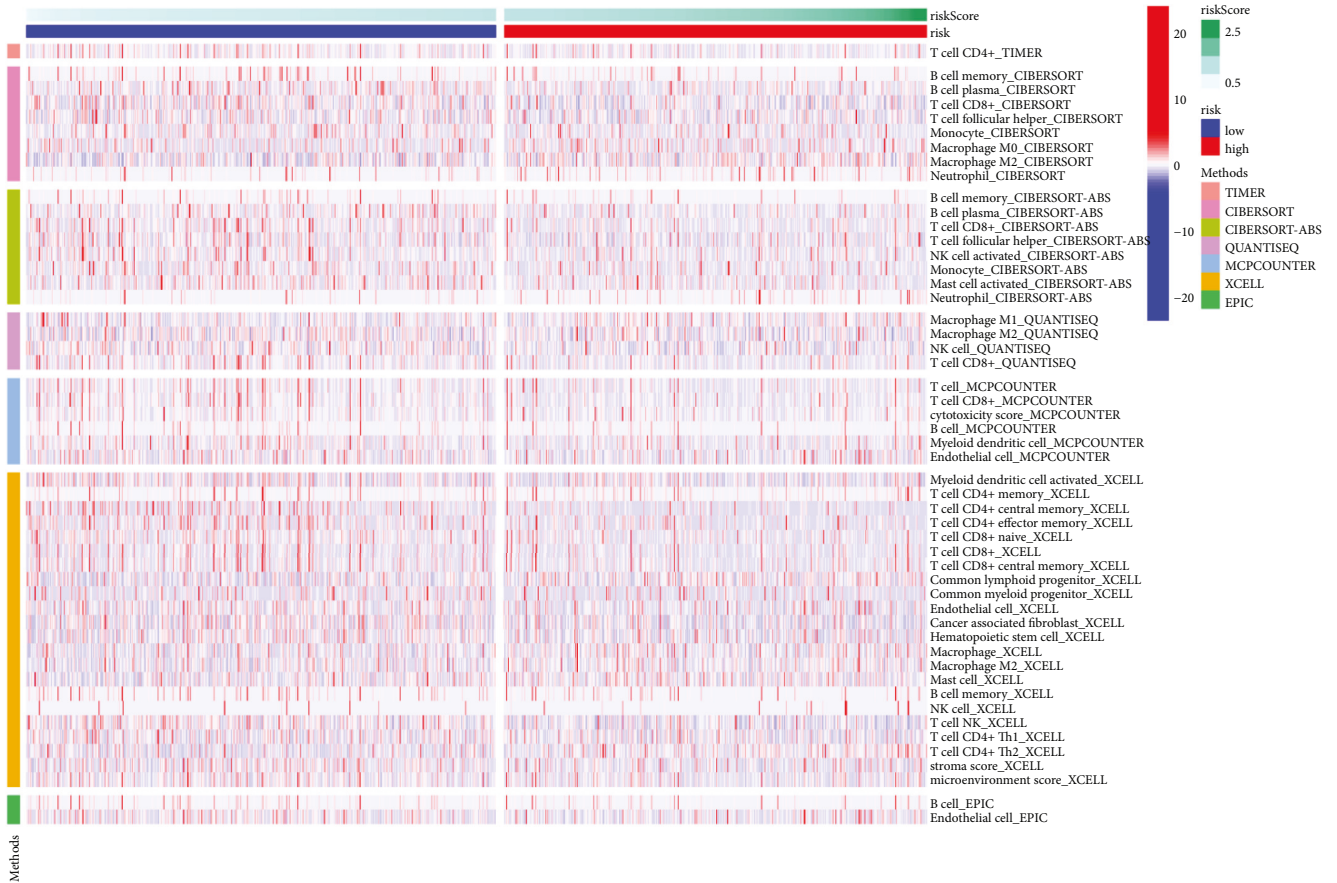
FIGURE 8: A heatmap of immune responses according to TIMER algorithms, CIBERSORT, QUANTISEQ, XCELL, MCP counter, and EPIC between the high- and low-risk groups of breast cancer.
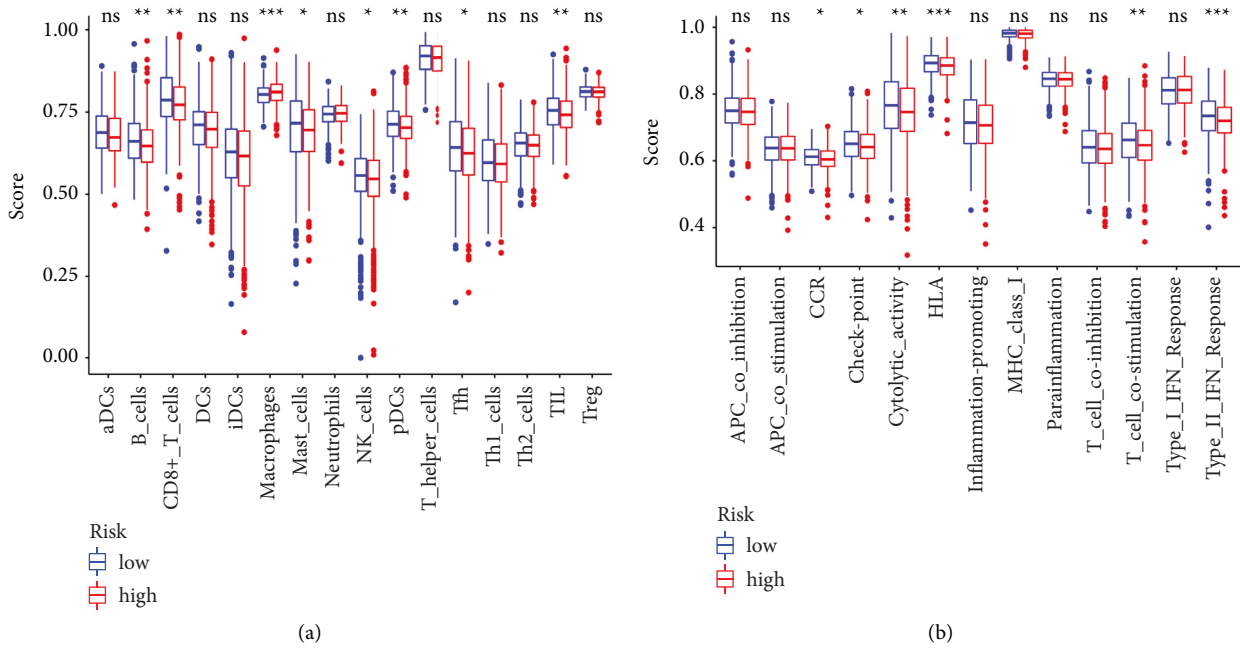

(a)

(b)

FIGURE 9: Immune-related analyses for the patients with breast cancer from TCGA. (a) Correlation analysis for immune cell subpopulations between the two breast cancer risk groups. (b) ssGSEA for the association between immune cell subpopulations and related functions between two risk groups.
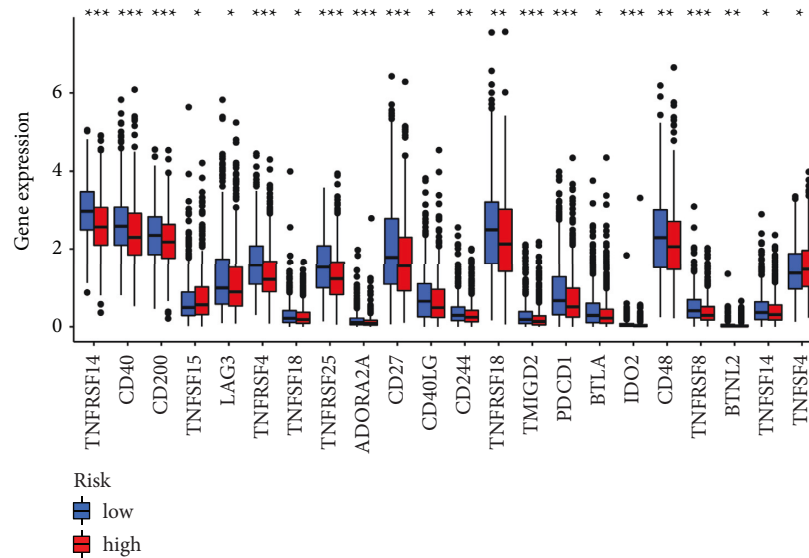
FIGURE 10: The expression levels of immune checkpoint genes between the two risk groups.

It is well known that in the pathogenesis of breast cancer, tumor-infiltrating immune cells play important roles [48] since over 70% of patients with breast cancer contain tumor-infiltrating lymphocytes [49]. To investigate the relationship of the constructed risk model and the immune status of patients with breast cancer, we divided all samples from TCGA into two risk groups according to their scores. Of note, consistent with previous researches, immune checkpoint genes, CD40, and TNFRSF14 were negatively related with the CDM gene-based signature. The activation of CD40 is important in the generation of T cell immunity [50]. Besides, the absence of TNFRSF14 would change the microenvironment of tumors [51, 52].

However, this study also has limitations. First, the information from TCGA database did not contain information of the preoperative and postoperative parameters, which made it impossible to carry out more comprehensive analyses with these potential and important factors. Second, since the data of this study were downloaded from TCGA, it was more in line with the characteristics of a retrospective study. Hence, to further validate the results, we need to carry out prospective clinical studies.

In summary, we explored the potential functions of the differentially expressed CDM genes in patients with breast cancer. A CDM gene-based risk model was constructed, which could predict the prognoses of patients with breast cancer. Besides, a clinically practical nomogram that can predict prognoses of breast cancer patients was constructed to help individual management in clinic.

Data Availability

The datasets generated and analyzed during the current study are available in TCGA (http://cancergenome.nih.gov/abouttcga) and GEO (https://www.ncbi.nlm.nih.gov/geo). The authors acknowledge TCGA and GEO database for providing the platform for uploading the meaningful datasets.

Ethical Approval

Not applicable.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Xin Yan and Sai Nan You contributed equally to this work. All listed authors took part actively in the research and read and approved the manuscript submitted.

Supplementary Materials

Supplementary File Figure S1: validation of the six CDM gene-based signature in an external dataset from GEO database. (A) Kaplan–Meier curves for OS of patients with breast cancer in GEO cohort. (B) Time-heatmap of the signature in GEO dataset, along with the risk score distribution and OS status. Supplementary File Figure S2: time-dependent ROC curves study of 3-year OS for the CDMSig, WangSig, and MoSig. (*Supplementary Materials*)

References

[1] Z. Niu, X. Li, S. Feng et al., "The deubiquitinating enzyme USP1 modulates ERα and modulates breast cancer progression," *Journal of Cancer*, vol. 11, no. 23, pp. 6992–7000, 2020.

[2] R. M. T. Staruch, M. J. Rouhani, and M. Ellabban, "The surgical management of male breast cancer: time for an easy access national reporting database?" *Annals of Medicine and Surgery*, vol. 9, pp. 41–49, 2016.

[3] L. F. Jensen, A. F. Pedersen, B. Andersen, and P. Vedsted, "Identifying specific non-attending groups in breast cancer screening—population-based registry study of participation and socio-demography," *BMC Cancer*, vol. 12, no. 1, p. 518, 2012.

[4] V. S. Shirure, N. M. Reynolds, and M. M. Burdick, "Mac-2 binding protein is a novel E-selectin ligand expressed by breast cancer cells," *PLoS One*, vol. 7, no. 9, Article ID e44529, 2012.

[5] S. K. Jeh, S. H. Kim, and B. J. Kang, "Comparison of the diagnostic performance of response evaluation criteria in solid tumor 1.0 with response evaluation criteria in solid tumor 1.1 on MRI in advanced breast cancer response evaluation to neoadjuvant chemotherapy," *Korean Journal of Radiology*, vol. 14, no. 1, pp. 13–20, 2013.

[6] H. Moghadas-Dastjerdi, H. R. Sha-E-Tallat, L. Sannachi, A. Sadeghi-Naini, and G. J. Czarnota, "A priori prediction of tumour response to neoadjuvant chemotherapy in breast cancer patients using quantitative CT and machine learning," *Scientific Reports*, vol. 10, no. 1, p. 10936, 2020.

[7] V. E. Strong, "Progress in gastric cancer," *Updates in Surgery*, vol. 70, no. 2, pp. 157–159, 2018.

[8] A. Tsherniak, F. Vazquez, P. G. Montgomery et al., "Defining a cancer dependency map," *Cell*, vol. 170, no. 3, pp. 564–576, 2017.

[9] R. M. Meyers, J. G. Bryan, J. M. McFarland et al., "Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells," *Nature Genetics*, vol. 49, no. 12, pp. 1779–1784, 2017.

[10] F. M. Behan, F. Iorio, G. Picco et al., "Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens," *Nature*, vol. 568, no. 7753, pp. 511–516, 2019.

[11] J. L. V. Maag, "Gganatogram: an R package for modular visualisation of anatograms and tissues based on ggplot2," *F1000Res*, vol. 7, p. 1576, 2018.

[12] T. Galili, A. O'Callaghan, J. Sidi, and C. Sievert, "Heatmaply: an R package for creating interactive cluster heatmaps for online publishing," *Bioinformatics*, vol. 34, no. 9, pp. 1600–1602, 2018.

[13] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "ClusterProfiler: an R package for comparing biological themes among gene clusters," *OMICS: A Journal of Integrative Biology*, vol. 16, no. 5, pp. 284–287, 2012.

[14] M. A. Harris, J. Clark, A. Ireland et al., "The gene ontology (GO) database and informatics resource," *Nucleic Acids Research*, vol. 32, pp. D258–D261, 2004.

[15] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, "KEGG: integrating viruses and cellular organisms," *Nucleic Acids Research*, vol. 49, no. D1, pp. D545–D551, 2021.

[16] M. Uhlén, L. Fagerberg, B. M. Hallström et al., "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, Article ID 1260419, 2015.

[17] T. Li, J. Fu, Z. Zeng et al., "TIMER2.0 for analysis of tumor-infiltrating immune cells," *Nucleic Acids Research*, vol. 48, no. W1, pp. W509–W514, 2020.

[18] A. M. Newman, C. L. Liu, M. R. Green et al., "Robust enumeration of cell subsets from tissue expression profiles," *Nature Methods*, vol. 12, no. 5, pp. 453–457, 2015.

[19] F. Finotello, C. Mayer, C. Plattner et al., "Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data," *Genome Medicine*, vol. 11, no. 1, p. 34, 2019.

[20] D. Aran, Z. Hu, and A. J. Butte, "xCell: digitally portraying the tissue cellular heterogeneity landscape," *Genome Biology*, vol. 18, no. 1, p. 220, 2017.

[21] E. Becht, N. A. Giraldo, L. Lacroix et al., "Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression," *Genome Biology*, vol. 17, no. 1, p. 218, 2016.

[22] J. Racle, K. de Jonge, P. Baumgaertner, D. E. Speiser, and D. Gfeller, "Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data," *Elife*, vol. 6, 2017.

[23] D. Wang, G. Wei, J. Ma et al., "Identification of the prognostic value of ferroptosis-related gene signature in breast cancer patients," *BMC Cancer*, vol. 21, no. 1, p. 645, 2021.

[24] W. Mo, Y. Ding, S. Zhao, D. Zou, and X. Ding, "Identification of a 6-gene signature for the survival prediction of breast cancer patients based on integrated multi-omics data analysis," *PLoS One*, vol. 15, Article ID e0241924, 2020.

[25] Y. H. Park, S. Lal, J. E. Lee et al., "Chemotherapy induces dynamic immune responses in breast cancers that impact treatment outcome," *Nature Communications*, vol. 11, no. 1, p. 6175, 2020.

[26] Q. Chang, M. Chandrashekhar, T. Ketela, Y. Fedyshyn, J. Moffat, and D. Hedley, "Cytokinetic effects of Wee1 disruption in pancreatic cancer," *Cell Cycle*, vol. 15, no. 4, pp. 593–604, 2016.

[27] C. Yu, A. M. Mannan, G. M. Yvone et al., "High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines," *Nature Biotechnology*, vol. 34, no. 4, pp. 419–423, 2016.

[28] J. Barretina, G. Caponigro, N. Stransky et al., "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, pp. 603–607, 2012.

[29] S. M. Corsello, J. A. Bittker, Z. Liu et al., "The drug repurposing hub: a next-generation drug library and information resource," *Nature Medicine*, vol. 23, no. 4, pp. 405–408, 2017.

[30] W. Wang, A. Malyutina, A. Pessia, J. Saarela, C. A. Heckman, and J. Tang, "Combined gene essentiality scoring improves the prediction of cancer dependency maps," *Ebiomedicine*, vol. 50, pp. 67–80, 2019.

[31] J. S. Boehm and T. R. Golub, "An ecosystem of cancer cell line factories to support a cancer dependency map," *Nature Reviews Genetics*, vol. 16, no. 7, pp. 373-374, 2015.

[32] Y. Wang, Y. Xia, and Z. Lu, "Metabolic features of cancer cells," *Cancer Communications*, vol. 38, no. 1, p. 65, 2018.

[33] G. Konopka, "Functional genomics of the brain: uncovering networks in the CNS using a systems approach," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 3, no. 6, pp. 628–648, 2011.

[34] E. V. Efimova, S. Takahashi, N. A. Shamsi et al., "Linking cancer metabolism to DNA repair and accelerated senescence," *Molecular Cancer Research*, vol. 14, no. 2, pp. 173–184, 2016.

[35] N. S. Gavande, P. S. VanderVere-Carozza, H. D. Hinshaw et al., "DNA repair targeted therapy: the past or future of cancer treatment?" *Pharmacology & Therapeutics*, vol. 160, pp. 65–83, 2016.

[36] N. Sohrabi, A. Valizadeh, S. M. Farkhani, and A. Akbarzadeh, "Basics of DNA biosensors and cancer diagnosis," *Artificial Cells, Nanomedicine, and Biotechnology*, vol. 44, no. 2, pp. 654–663, 2016.

[37] A. J. Medford, R. N. Gillani, and B. H. Park, "Detection of cancer DNA in early stage and metastatic breast cancer patients," *Methods in Molecular Biology*, vol. 1768, pp. 209–227, 2018.

[38] H. He, Z. Wei, F. Du et al., "Transcription of HOTAIR is regulated by RhoC-MRTF-A-SRF signaling pathway in human breast cancer cells," *Cellular Signalling*, vol. 31, pp. 87–95, 2017.

[39] Y. Deng, W. Guo, N. Xu, F. Li, and J. Li, "CtBP1 transactivates RAD51 and confers cisplatin resistance to breast cancer cells," *Molecular Carcinogenesis*, vol. 59, no. 5, pp. 512–519, 2020.

[40] B. Husbeck, D. E. Stringer, E. W. Gerner, and G. Powis, "Increased thioredoxin-1 inhibits SSAT expression in MCF-7 human breast cancer cells," *Biochemical and Biophysical Research Communications*, vol. 306, no. 2, pp. 469–475, 2003.

[41] A. Coker-Gurkan, M. Celik, M. Ugur et al., "Curcumin inhibits autocrine growth hormone-mediated invasion and metastasis by targeting NF-κB signaling and polyamine metabolism in breast cancer cells," *Amino Acids*, vol. 50, no. 8, pp. 1045–1069, 2018.

[42] D. Georgopoulou, M. Callari, M. Callari et al., "Landscapes of cellular phenotypic diversity in breast cancer xenografts and their impact on drug response," *Nature Communications*, vol. 12, no. 1, p. 1998, 2021.

[43] S. Kumar, A. Nandi, S. Singh et al., "Dll1+ quiescent tumor stem cells drive chemoresistance in breast cancer through NF-κB survival pathway," *Nature Communications*, vol. 12, no. 1, p. 432, 2021.

[44] W. Park, S. Park, G. Song, and W. Lim, "Inhibitory effects of osthole on human breast cancer cell progression via induction of cell cycle arrest, mitochondrial dysfunction, and ER stress," *Nutrients*, vol. 11, 2019.

[45] A. M. Braden, R. V. Stankowski, J. M. Engel, and A. A. Onitilo, "Breast cancer biomarkers: risk assessment, diagnosis, prognosis, prediction of treatment efficacy and toxicity, and recurrence," *Current Pharmaceutical Design*, vol. 20, pp. 4879–4898, 2014.

[46] G. F. Stark, G. R. Hart, B. J. Nartowt, and J. Deng, "Predicting breast cancer risk using personal health data and machine learning models," *PLoS One*, vol. 14, Article ID e0226765, 2019.

[47] C. Boeri, C. Chiappa, F. Galli et al., "Machine learning techniques in breast cancer prognosis prediction: a primary evaluation," *Cancer Medicine*, vol. 9, no. 9, pp. 3234–3243, 2020.

[48] R. Vishnubalaji, V. Sasidharan Nair, K. Ouararhni, E. Elkord, and N. M. Alajez, "Integrated transcriptome and pathway analyses revealed multiple activated pathways in breast cancer," *Frontiers in Oncology*, vol. 9, p. 910, 2019.

[49] H. Chen, W. Chong, X. Yang et al., "Age-related mutational signature negatively associated with immune activity and survival outcome in triple-negative breast cancer," *OncoImmunology*, vol. 9, no. 1, Article ID 1788252, 2020.

[50] R. H. Vonderheide, "The immune revolution: a case for priming, not checkpoint," *Cancer Cell*, vol. 33, no. 4, pp. 563–569, 2018.

[51] L. von Hoff, E. Kärgel, V. Franke et al., "Autocrine LTA signaling drives NF-κB and JAK-STAT activity and myeloid gene expression in Hodgkin lymphoma," *Blood*, vol. 133, no. 13, pp. 1489–1494, 2019.

[52] R. Kennedy and U. Klein, "A T cell-B cell tumor-suppressive axis in the germinal center," *Immunity*, vol. 51, no. 2, pp. 204–206, 2019.