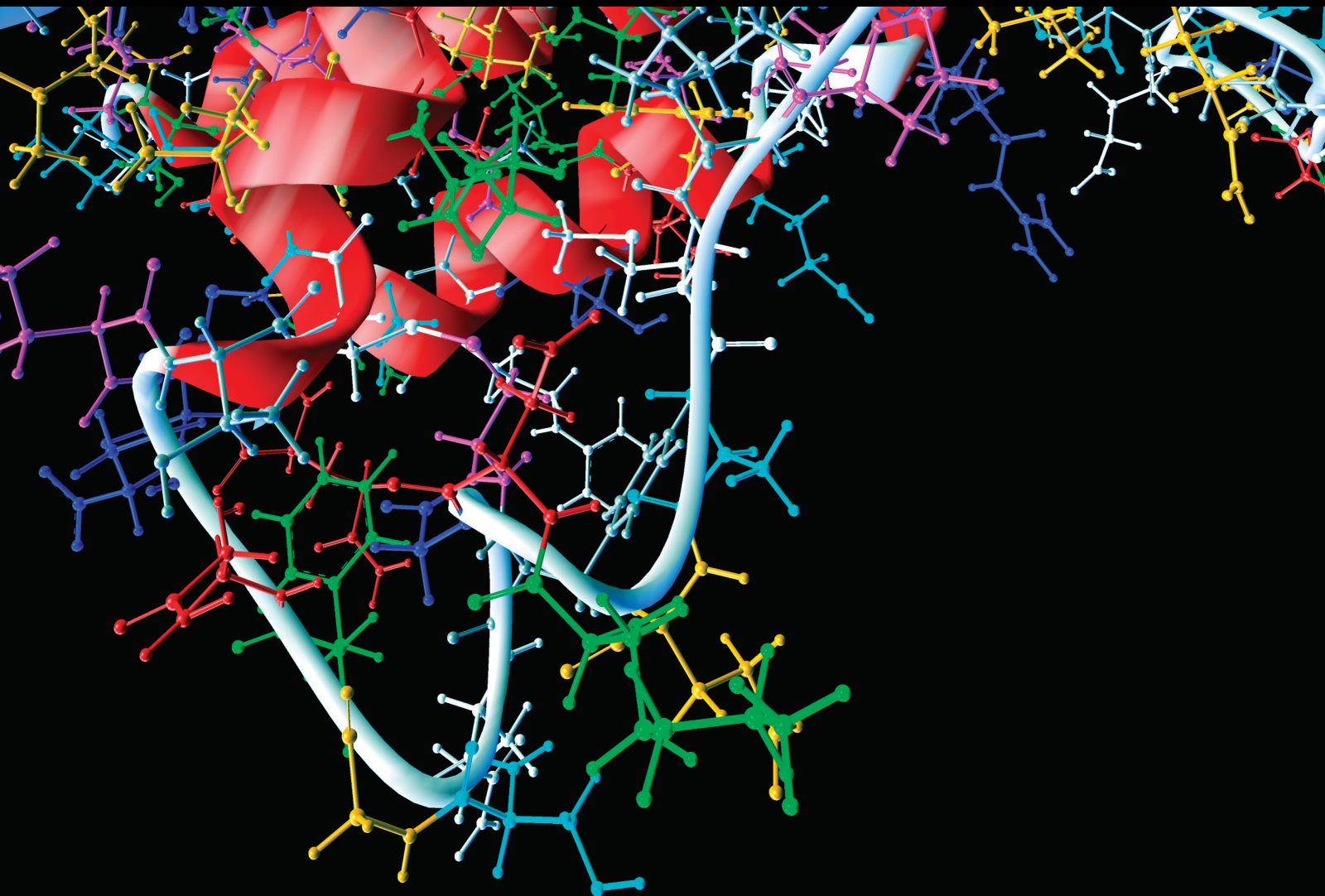


Developing and Applying Machine Learning-Based Methods in Special Function Protein Identification 2021

Lead Guest Editor: Hui Ding

Guest Editors: Balachandran Manavalan and Watshara Shoombuatong





**Developing and Applying Machine Learning-
Based Methods in Special Function Protein
Identification 2021**

Computational and Mathematical Methods in Medicine

**Developing and Applying Machine
Learning-Based Methods in Special
Function Protein Identification 2021**

Lead Guest Editor: Hui Ding




Guest Editors: Balachandran Manavalan and
Watshara Shoombuatong



Copyright © 2023 Hindawi Limited. All rights reserved.

This is a special issue published in “Computational and Mathematical Methods in Medicine.” All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Associate Editors

Ahmed Albahri, Iraq
Konstantin Blyuss , United Kingdom
Chuangyin Dang, Hong Kong
Farai Nyabadza , South Africa
Kathiravan Srinivasan , India

Academic Editors

Laith Abualigah , Jordan
Yaser Ahangari Nanehkaran , China
Mubashir Ahmad, Pakistan
Sultan Ahmad , Saudi Arabia
Akif Akgul , Turkey
Karthick Alagar, India
Shadab Alam, Saudi Arabia
Raul Alcaraz , Spain
Emil Alexov, USA
Enrique Baca-Garcia , Spain
Sweta Bhattacharya , India
Junguo Bian, USA
Elia Biganzoli , Italy
Antonio Boccaccio, Italy
Hans A. Braun , Germany
Zhicheng Cao, China
Guy Carrault, France
Sadaruddin Chachar , Pakistan
Prem Chapagain , USA
Huiling Chen , China
Mengxin Chen , China
Haruna Chiroma, Saudi Arabia
Watcharaporn Cholamjiak , Thailand
Maria N. D.S. Cordeiro , Portugal
Cristiana Corsi , Italy
Qi Dai , China
Nagarajan Deivanayagam Pillai, India
Didier Delignières , France
Thomas Desaive , Belgium
David Diller , USA
Qamar Din, Pakistan
Irina Doytchinova, Bulgaria
Sheng Du , China
D. Easwaramoorthy , India

Esmaeil Ebrahimie , Australia
Issam El Naqa , USA
Ilias Elmouki , Morocco
Angelo Facchiano , Italy
Luca Faes , Italy
Maria E. Fantacci , Italy
Giancarlo Ferrigno , Italy
Marc Thilo Figge , Germany
Giulia Fiscon , Italy
Bapan Ghosh , India
Igor I. Goryanin, Japan
Marko Gosak , Slovenia
Damien Hall, Australia
Abdulsattar Hamad, Iraq
Khalid Hattaf , Morocco
Tingjun Hou , China
Seiya Imoto , Japan
Martti Juhola , Finland
Rajesh Kaluri , India
Karthick Kanagarathinam, India
Rafik Karaman , Palestinian Authority
Chandan Karmakar , Australia
Kwang Gi Kim , Republic of Korea
Andrzej Kloczkowski, USA
Andrei Korobeinikov , China
Sakthidasan Sankaran Krishnan, India
Rajesh Kumar, India
Kuruva Lakshmana , India
Peng Li , USA
Chung-Min Liao , Taiwan
Pinyi Lu , USA
Reinoud Maex, United Kingdom
Valeri Makarov , Spain
Juan Pablo Martínez , Spain
Richard J. Maude, Thailand
Zahid Mehmood , Pakistan
John Mitchell , United Kingdom
Fazal Ijaz Muhammad , Republic of Korea
Vishal Nayak , USA
Tongguang Ni, China
Michele Nichelatti, Italy
Kazuhisa Nishizawa , Japan
Bing Niu , China

Hyuntae Park , Japan
Jovana Paunovic , Serbia
Manuel F. G. Penedo , Spain
Riccardo Pernice , Italy
Kemal Polat , Turkey
Alberto Policriti, Italy
Giuseppe Pontrelli , Italy
Jesús Poza , Spain
Maciej Przybyłek , Poland
Bhanwar Lal Puniya , USA
Mihai V. Putz , Romania
Suresh Rasappan, Oman
Jose Joaquin Rieta , Spain
Fathalla Rihan , United Arab Emirates
Sidheswar Routray, India
Sudipta Roy , India
Jan Rychtar , USA
Mario Sansone , Italy
Murat Sari , Turkey
Shahzad Sarwar, Saudi Arabia
Kamal Shah, Saudi Arabia
Bhisham Sharma , India
Simon A. Sherman, USA
Mingsong Shi, China
Mohammed Shuaib , Malaysia
Prabhishek Singh , India
Neelakandan Subramani, India
Junwei Sun, China
Yung-Shin Sun , Taiwan
Min Tang , China
Hongxun Tao, China
Alireza Tavakkoli , USA
João M. Tavares , Portugal
Jlenia Toppi , Italy
Anna Tsantili-Kakoulidou , Greece
Markos G. Tsipouras, North Macedonia
Po-Hsiang Tsui , Taiwan
Sathishkumar V E , Republic of Korea
Durai Raj Vincent P M , India
Gajendra Kumar Vishwakarma, India
Liangjiang Wang, USA
Ruisheng Wang , USA
Zhouchao Wei, China
Gabriel Wittum, Germany
Xiang Wu, China

KI Yanover , Israel
Xiaojun Yao , China
Kaan Yetilmezsoy, Turkey
Hiro Yoshida, USA
Yuhai Zhao , China



Contents

Retracted: Simulation and Prediction of Fungal Community Evolution Based on RBF Neural Network

Computational and Mathematical Methods in Medicine






Retraction (1 page), Article ID 9869064, Volume 2023 (2023)

Characterization of a Pyroptosis-Related Signature for Prognosis Prediction and Immune Microenvironment Infiltration in Prostate Cancer

Guian Zhang , Yong Luo, Weimin Dong, and Weide Zhong 




Research Article (51 pages), Article ID 8233840, Volume 2022 (2022)

Identification of Nine mRNA Signatures for Sepsis Using Random Forest

Jing Zhou , Siqing Dong , Ping Wang , Xi Su , and Liang Cheng 

Research Article (7 pages), Article ID 5650024, Volume 2022 (2022)

Prediction of New Risk Genes and Potential Drugs for Rheumatoid Arthritis from Multiomics Data

Anteneh M. Birga, Liping Ren, Huaichao Luo , Yang Zhang , and Jian Huang 



Research Article (11 pages), Article ID 6783659, Volume 2022 (2022)

Identification of *Helicobacter pylori* Membrane Proteins Using Sequence-Based Features

Mujixin Liu , Hui Chen , Dong Gao , Cai-Yi Ma , and Zhao-Yue Zhang 



Research Article (7 pages), Article ID 7493834, Volume 2022 (2022)

Dysregulation of Circadian Clock Genes as Significant Clinic Factor in the Tumorigenesis of Hepatocellular Carcinoma

Youfang Liang, Shaoxiang Wang, Xin Huang, Ruihuan Chai, Qian Tang, Rong Yang, Xiaoqing Huang, Xiao Wang , and Kai Zheng 




Research Article (14 pages), Article ID 8238833, Volume 2021 (2021)

Prediction of Metal Ion Binding Sites of Transmembrane Proteins

Jing Qu, Sheng S. Yin , and Han Wang 






Research Article (11 pages), Article ID 2327832, Volume 2021 (2021)

iMPT-FDNPL: Identification of Membrane Protein Types with Functional Domains and a Natural Language Processing Approach

Wei Chen , Lei Chen , and Qi Dai 

Research Article (10 pages), Article ID 7681497, Volume 2021 (2021)

[Retracted] Simulation and Prediction of Fungal Community Evolution Based on RBF Neural Network

Xiao-Wei Cai , Ya-Qian Bao , Ming-Feng Hu , Jia-Bao Liu , and Jia-Ming Zhu 

Research Article (13 pages), Article ID 7918192, Volume 2021 (2021)

Retraction

Retracted: Simulation and Prediction of Fungal Community Evolution Based on RBF Neural Network

Computational and Mathematical Methods in Medicine

Received 20 June 2023; Accepted 20 June 2023; Published 21 June 2023

Copyright © 2023 Computational and Mathematical Methods in Medicine. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] X. Cai, Y. Bao, M. Hu, J. Liu, and J. Zhu, "Simulation and Prediction of Fungal Community Evolution Based on RBF Neural Network," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 7918192, 13 pages, 2021.

Research Article

Characterization of a Pyroptosis-Related Signature for Prognosis Prediction and Immune Microenvironment Infiltration in Prostate Cancer

Guian Zhang ^{1,2}, Yong Luo,³ Weimin Dong,⁴ and Weide Zhong ^{1,2}

¹School of Medicine, South China University of Technology, Guangzhou, China

²Department of Urology, Guangdong Key Laboratory of Clinical Molecular Medicine and Diagnostics, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China

³Department of Urology, The Second People's Hospital of Foshan, Affiliated Foshan Hospital of Southern Medical University, Foshan 528000, China

⁴Department of Urology, The Fifth Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

Correspondence should be addressed to Weide Zhong; eyweidezhong@scut.edu.cn

Received 28 December 2021; Accepted 28 March 2022; Published 27 April 2022

Academic Editor: Hui Ding

Copyright © 2022 Guian Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This study was aimed at constructing a pyroptosis-related signature for prostate cancer (PCa) and elucidating the prognosis and immune landscape and the sensitivity of immune checkpoint blockade (ICB) therapy in signature-define subgroups of PCa. We identified 22 differentially expressed pyroptosis-related genes in PCa from The Cancer Genome Atlas (TCGA) database. The pyroptosis-related genes could divide PCa patients into two clusters with differences in survival. Seven genes were determined to construct a signature that was confirmed by qRT-PCR to be closely associated with the biological characteristics of malignant PCa. The signature could effectively and independently predict the biochemical recurrence (BCR) of PCa, which was validated in the GSE116918 and GSE21034. We found that patients in the high-risk group were more prone to BCR and closely associated with high-grade and advanced-stage disease progression. Outperforming clinical characteristics and nine published articles, our signature demonstrated excellent predictive performance. The patients in the low-risk group were strongly related to the high infiltration of various immune cells including CD8+ T cells and plasma B cells. Furthermore, the high-risk group with higher TMB levels and expression of immune checkpoints was more likely to benefit from immune checkpoint therapy such as PD-1 and CTLA-4 inhibitors. The sensitivity to chemotherapy, endocrine, and targeted therapy showed significant differences in the two risk groups. Our signature was a novel therapeutic strategy to distinguish the prognosis and guide treatment strategies.

1. Introduction

Prostate cancer (PCa) is the second most widespread male cancer with high lethality, causing more than 370000 deaths worldwide in 2020 [1]. Meanwhile, more than one-third of patients eventually experience biochemical recurrence (BCR) after definitive treatment [2]. Patients with BCR were more likely to develop clinical recurrence, metastases, and cancer-specific mortality [3]. Therefore, early detection of BCR was essential for the management and treatment of PCa patients. The existing clinical indicators cannot effec-

tively predict BCR and guide treatment, necessitating representative and robust clinical models to promote preclinical translational and mechanistic studies of treatment in PCa.

Pyroptosis is considered to be a form of programmed cell necrosis triggered by proinflammatory signals and associated with inflammation [4]. Pyroptotic cells undergo cytoplasmic swelling and membrane pore formation, leading to loss of plasma membrane integrity and ultimately to leakage of cytoplasmic contents. The occurrence of pyroptosis requires the activation of caspase-1, which is responsible for the maturation of proinflammatory cytokines through

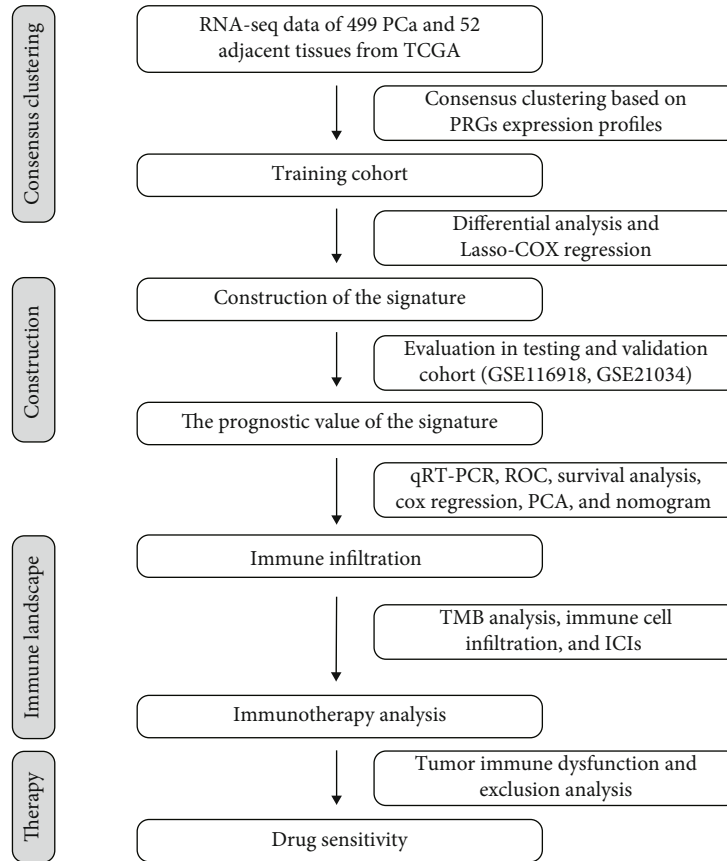


FIGURE 1: The workflow of this study.

inflammasome-dependent pathways, such as interleukin 1 β (IL-1 β) and IL-18 [5]. Meanwhile, gasdermin D (GSDMD) cleaved by activated caspase-1 locks into the plasma membrane to form pores [6]. More and more studies on the relationship between pyroptosis and tumors had shown that pyroptosis played an important role in the proliferation, invasion, and metastasis of tumor cells and affected the prognosis and therapeutic effects of tumors. GSDME-mediated pyroptosis promoted the development of colitis-related colorectal cancer, inducing tumor cell proliferation and proliferating cell nuclear antigen expression [7]. Gasdermin E-dependent pyroptosis might be indispensable in mediating the immunotherapy response of BRAF mutant melanoma [8].

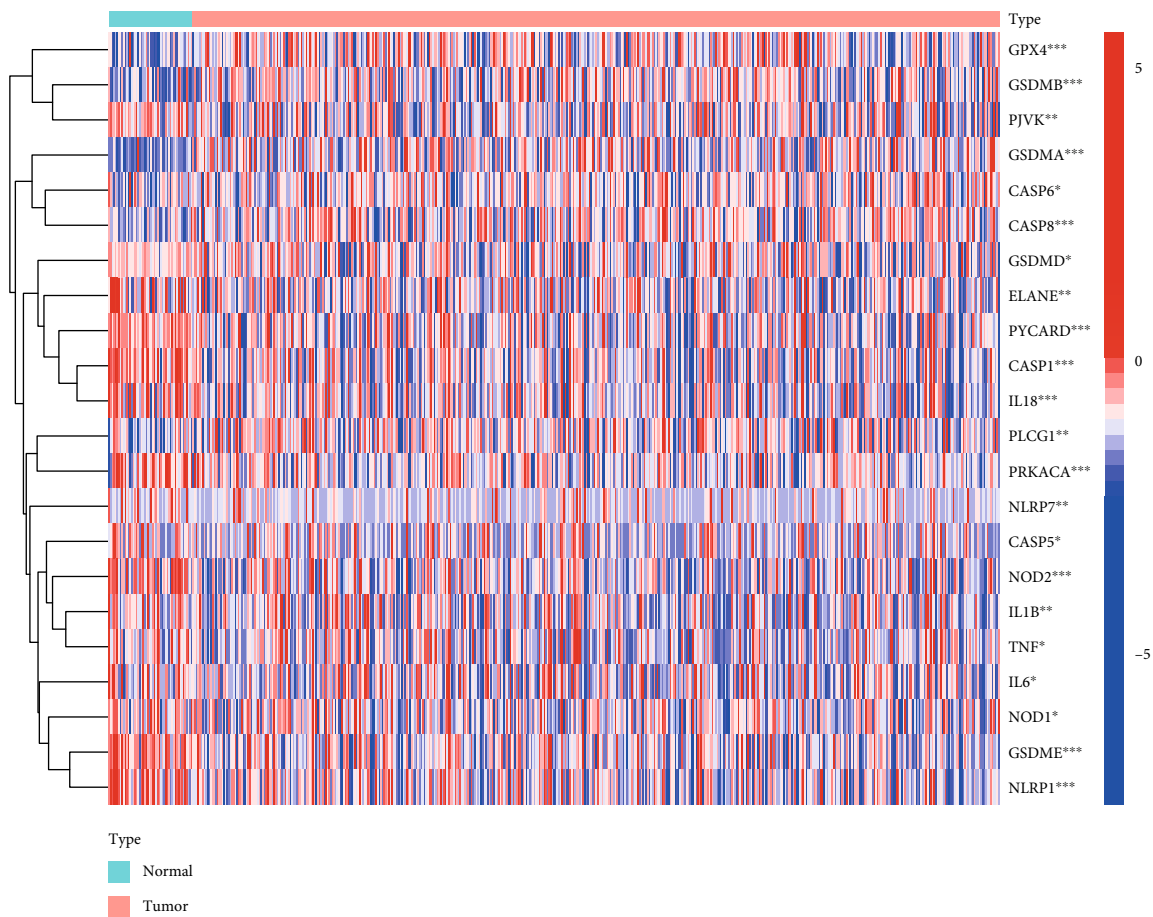
The tumor microenvironment (TME) has been confirmed to play a central role in tumorigenesis, immune escape, progression, and metastasis [9]. Tumor cells actively secrete inflammatory factors and growth factors to recruit stromal cells, inflammatory, and immune cells. The interaction between tumor cells and nontumor cells shapes TME, which in turn affects tumor progression and evades immune surveillance [10]. Characterized as inflammatory, pyroptosis recruited and activated immune cells through the inflammatory factors released during cell death to bridge innate immunity and adaptive immunity to regulate the TME and induce immune responses [11]. Meanwhile, neoantigens produced during the process of pyroptosis further induced new immune responses and hindered the development of

tumors [12]. The study by Z. Zhang et al. showed that the infiltration of CD8+ T cells and natural killing cells in the pyroptosis-activated TME could promote pyroptosis and form a positive feedback loop [13]. The important role of pyroptosis in the efficacy of cancer immunotherapy, such as immune checkpoint blockade (ICB), and the new approaches of pyroptosis to aid immunotherapy were receiving increasing attention [14]. Therefore, there was a need to identify the different risk stratification of PCa patients for immunotherapy through a comprehensive and deep insight into TME by pyroptosis.

In this study, we sought to develop a prognostic signature for PCa, which can effectively stratify patients and predict the prognosis and treatment efficacy of patients with different risk levels. The results revealed that the predictive ability of our signature was superior to traditional clinical features. On this basis, we systematically explored the role of the signature in the TME. Our signature was a promising prognostic biomarker to guide and determine the subgroup of PCa patients more suitable for endocrine therapy, chemotherapy, and immunotherapy.

2. Materials and Methods

2.1. Data Source and Preprocessing. Transcriptome RNA sequencing data and corresponding clinical information of PCa samples, which was the training cohort, were downloaded from the TCGA program (<https://tcga-data.nci.nih>



(a)

FIGURE 2: Continued.

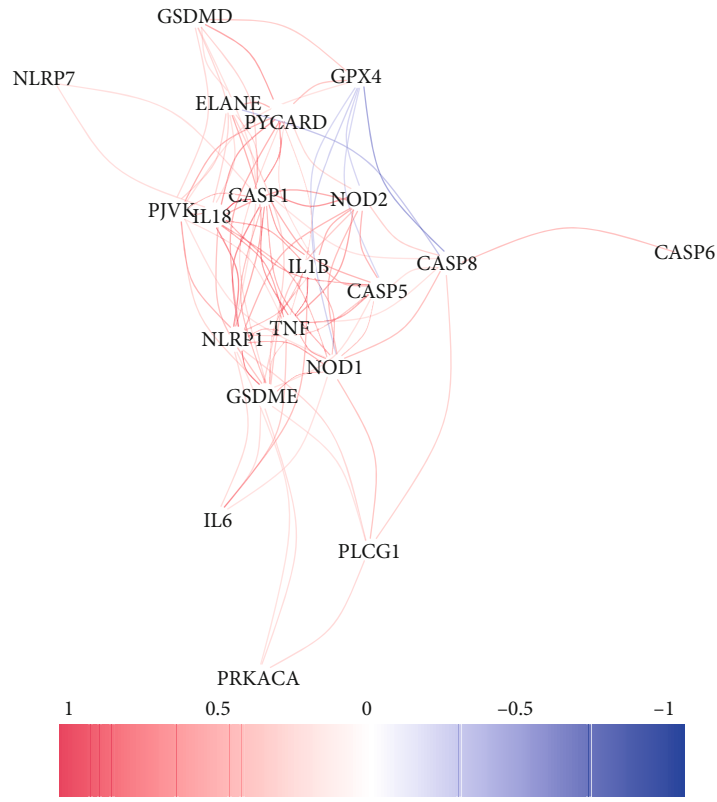
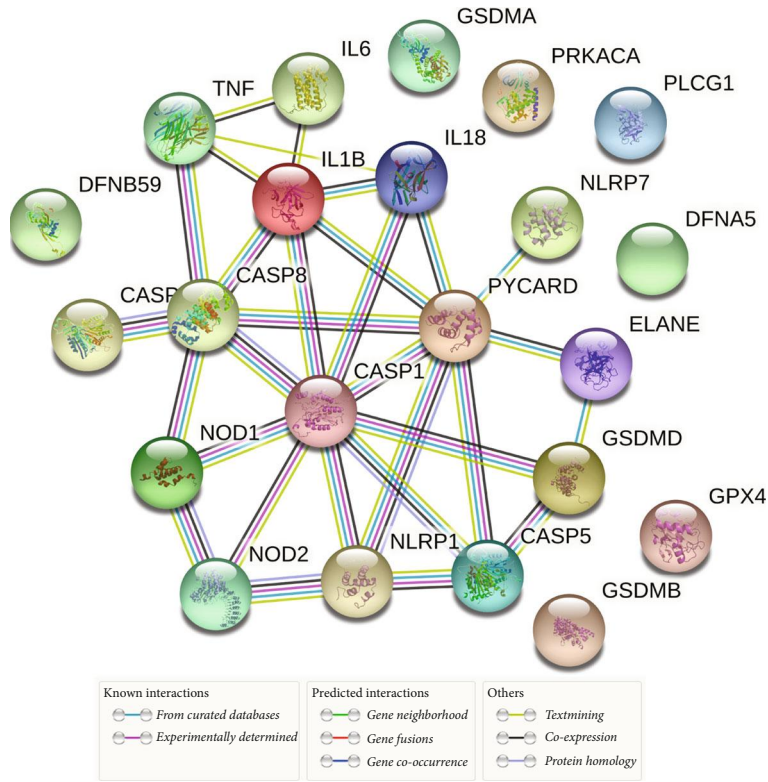


FIGURE 2: Continued.

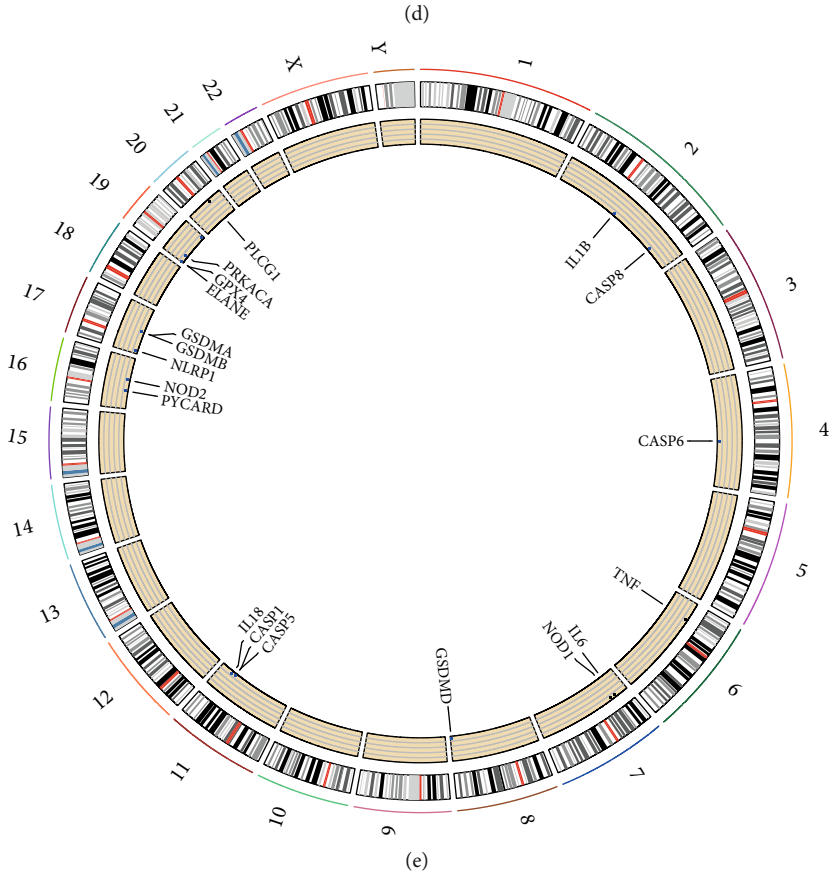
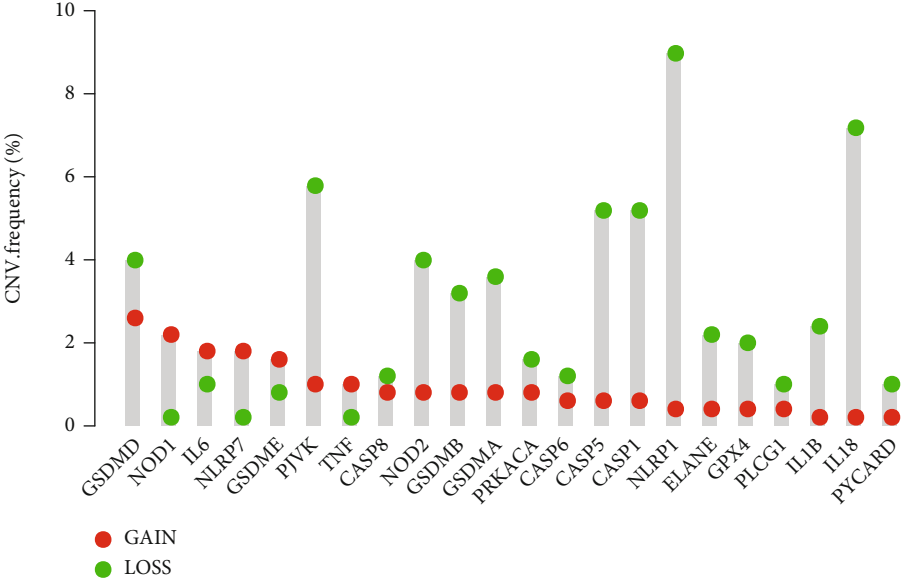
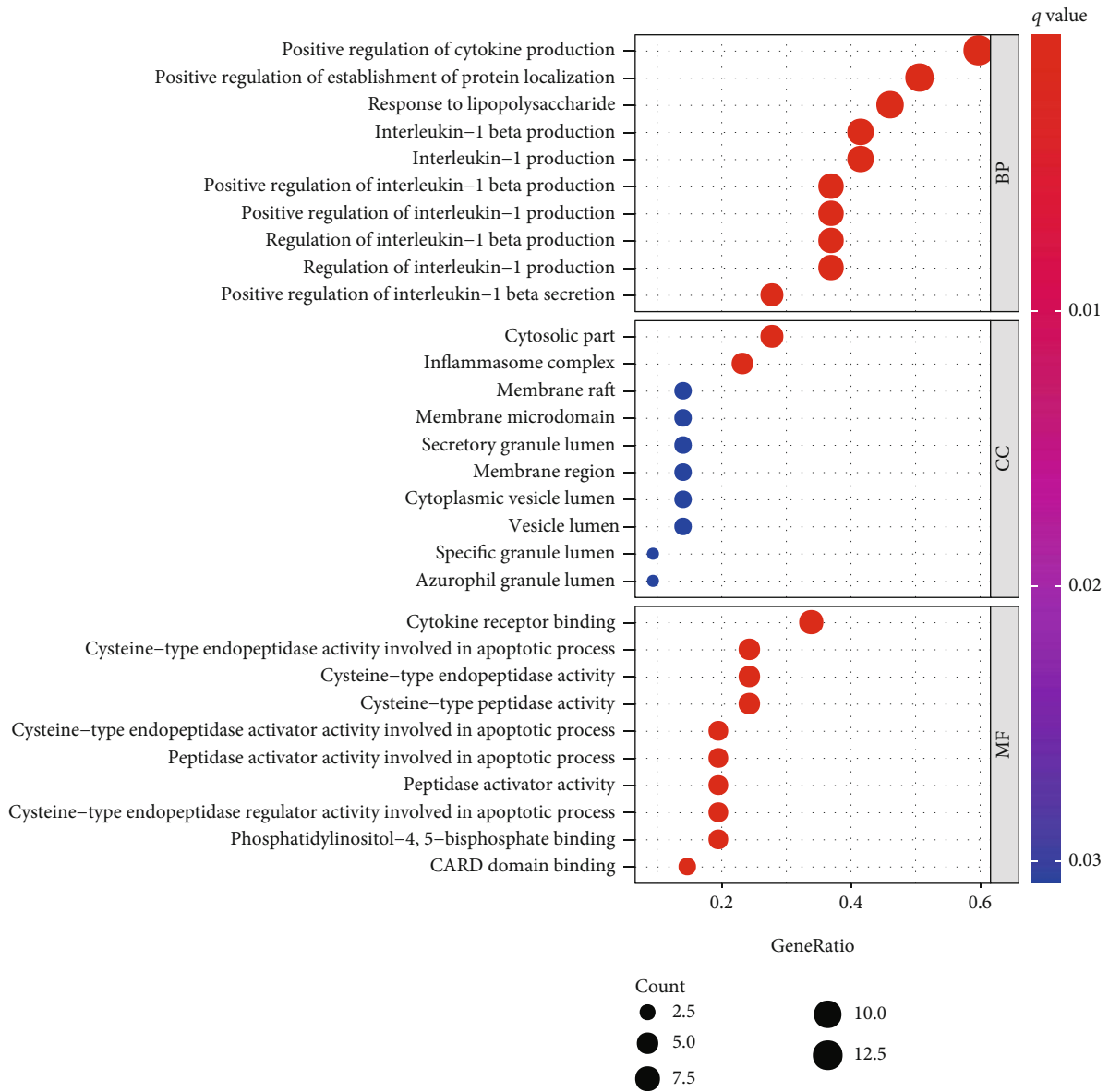


FIGURE 2: Continued.



(f)

FIGURE 2: Continued.

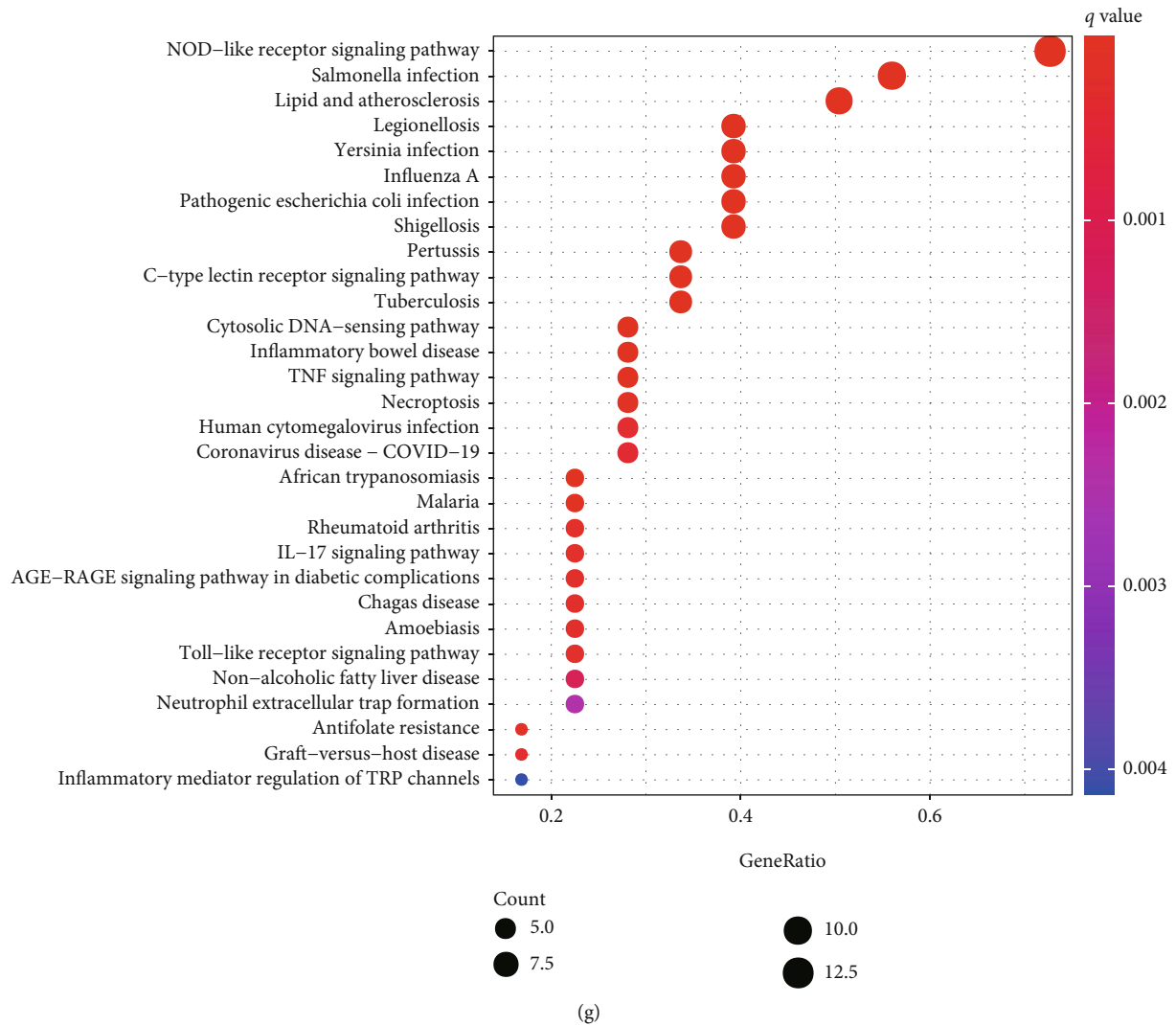


FIGURE 2: Expression and interactions of the pyroptosis-related genes in PCa. (a) Heatmap of differentially expressed pyroptosis-related genes in tumor and normal tissues. (b) Protein-protein interaction network of 22 DEGs. (c) The correlation network of DEGs. (d) The CNV variation frequency of DEGs. (e) The location of CNV alteration of DEGs on chromosomes. (f) Bubble graph for GO enrichment and (g) KEGG pathways.

.gov/tcga/). The GSE116918 dataset as testing cohort and GSE21034 dataset as validation cohort were extracted from the Gene Expression Omnibus (GEO) dataset (<https://www.ncbi.nlm.nih.gov/geo/>). The ComBat algorithm of SVA package was applied to correct the batch impact of nonbiotechnical bias. The training cohort was appointed to build signature, and the testing and validation cohorts were used to validate it. The R package *maftools* was used to visualize the mutation landscape, and the CNV feature in human chromosomes was investigated by the *Rcircos* package. The *rms* package was used to build a predictive nomogram for predicting the 1-, 2-, and 3-year overall survival.

2.2. Identification of Differentially Expressed Pyroptosis-Related Genes. A total of 33 pyroptosis-related genes were selected based on the previously published literature [15]. The difference in pyroptosis-related genes with a *P* value

< 0.05 was identified by *limma* package. We constructed a protein-protein interaction (PPI) network using the Search Tool for Retrieval of Interacting Genes (STRING).

2.3. Consensus Clustering. To identify different pyroptosis modifications, we applied consensus clustering to identify different pyroptosis patterns associated with the expression of pyroptosis-related genes. The *ConsensusClusterPlus* package was applied to determine the number of clusters and their stability, performing 1000 replications. The clusters were selected based on the relative change in the area under the cumulative distribution function (CDF) curve, the number of samples in the cluster, and the relevance of the cluster.

2.4. Construction of the Signature. The Cox regression analysis was conducted to assess the correlation between the expression level of each gene and its prognosis.

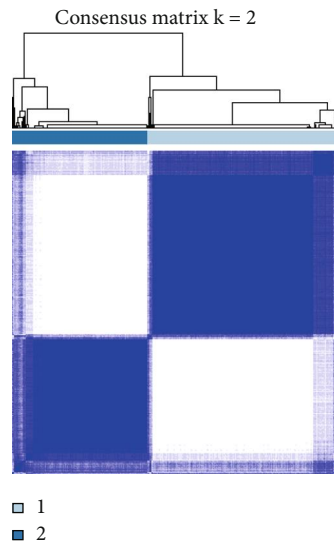
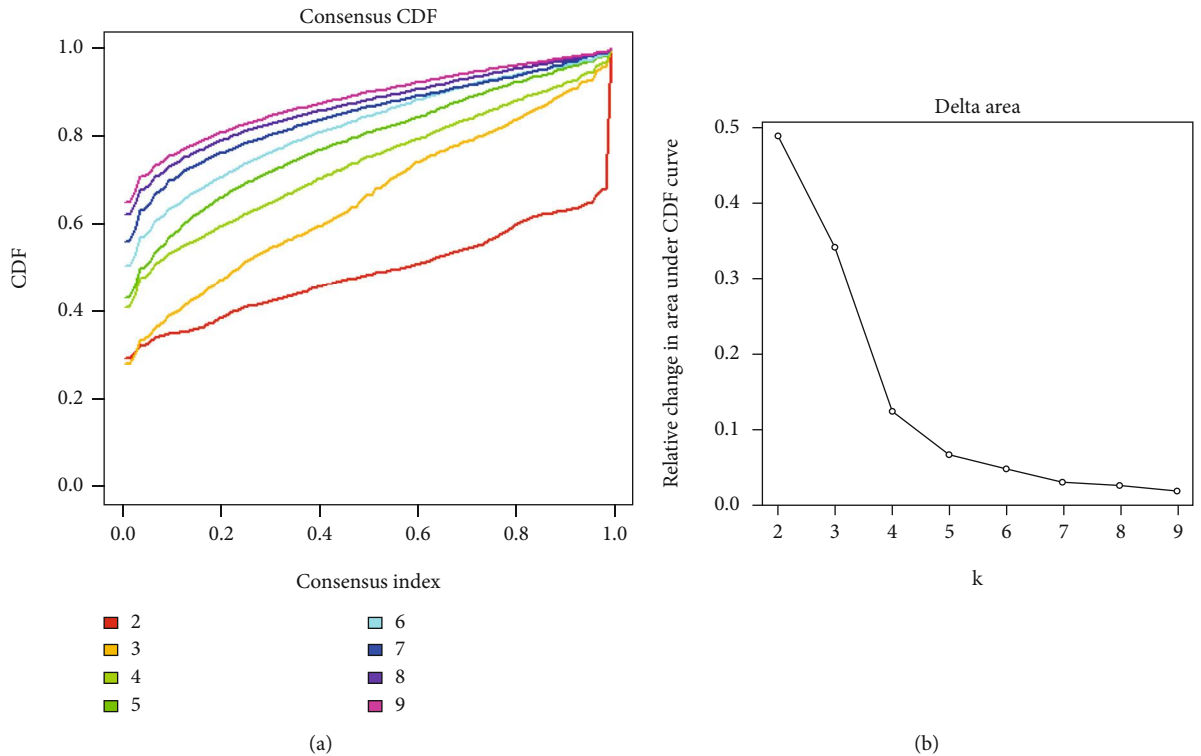
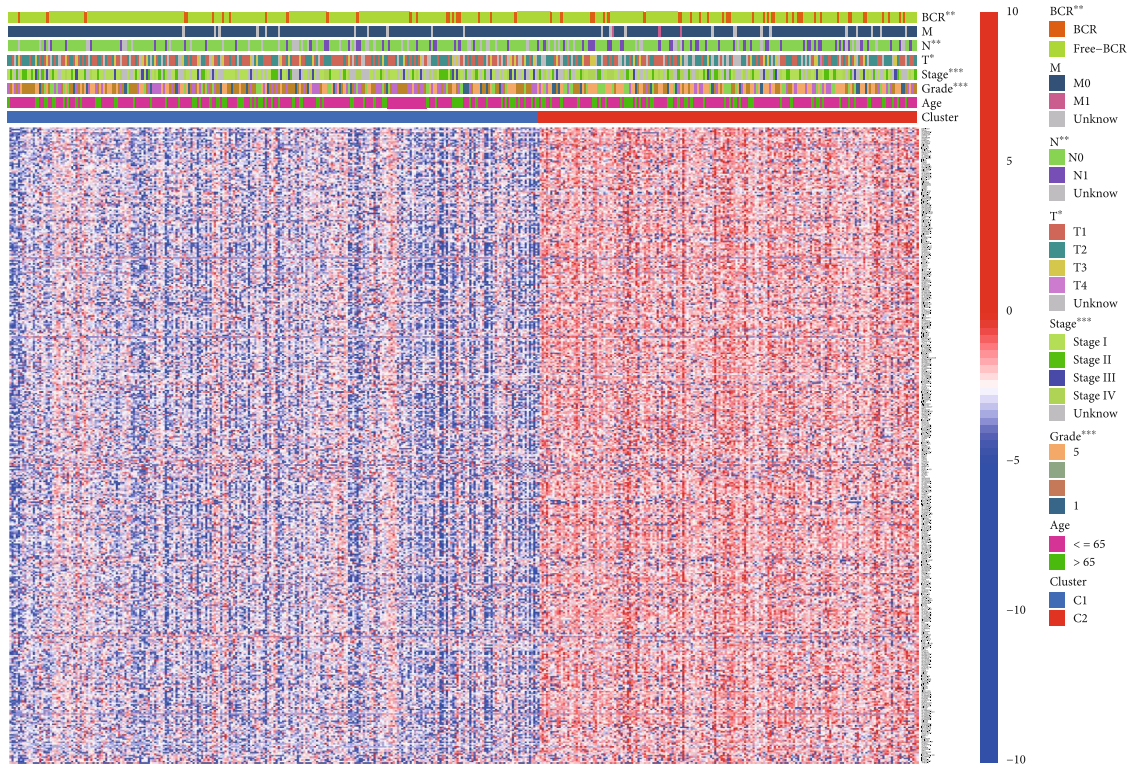
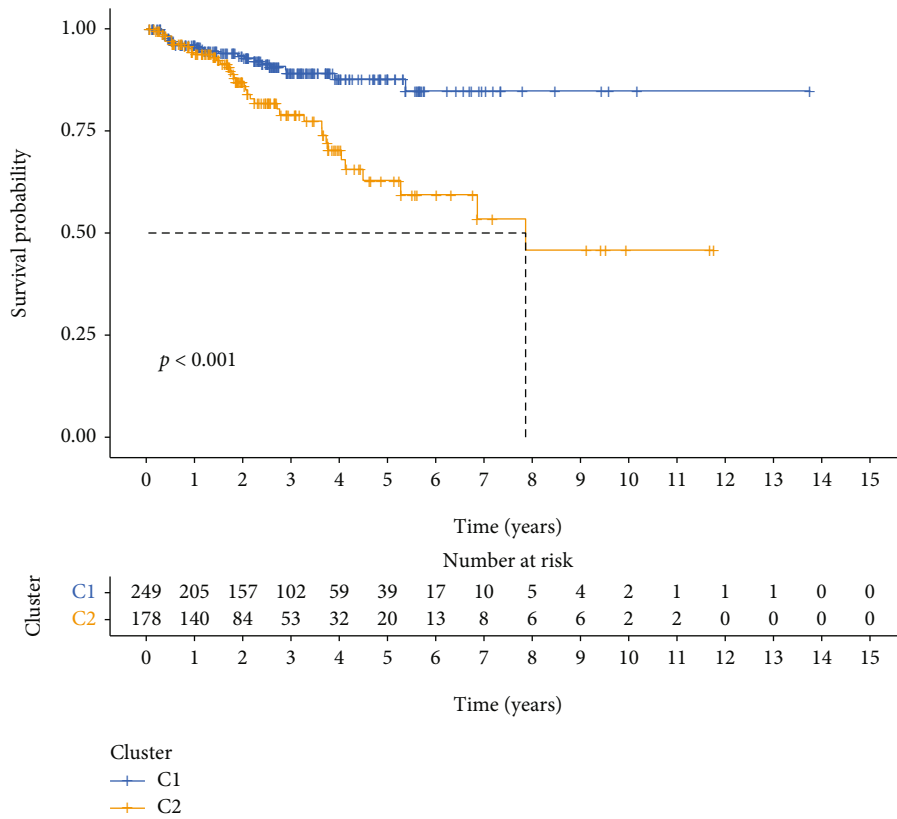


FIGURE 3: Continued.

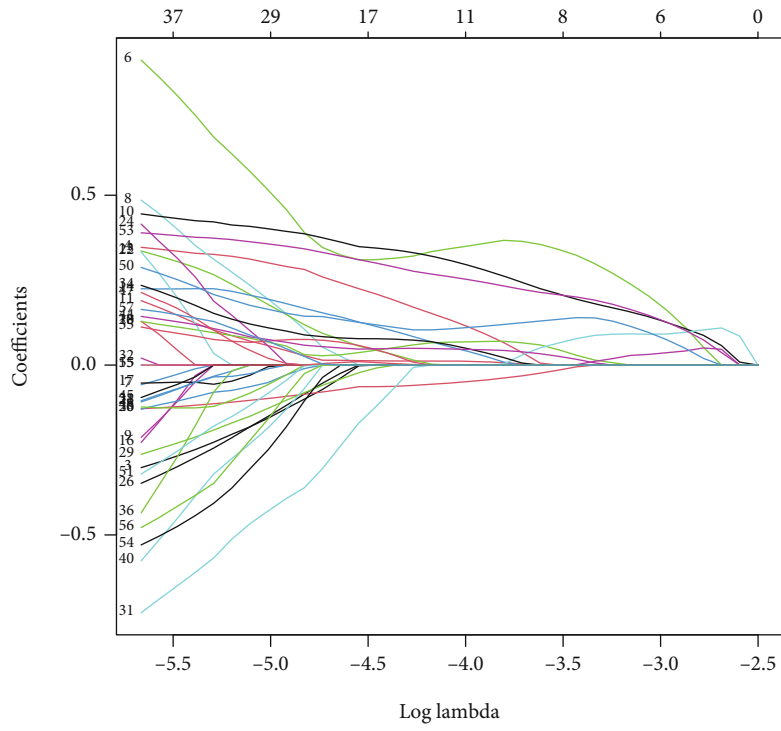


(d)

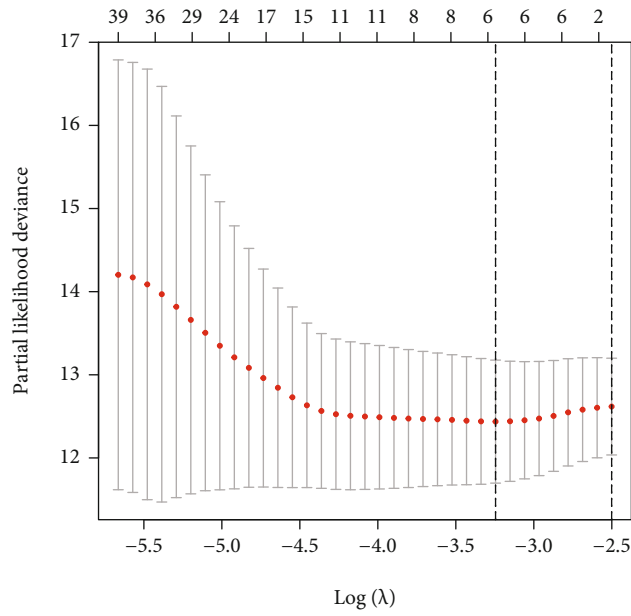


(e)

FIGURE 3: Clinical characteristics of PCa clusters. (a) CDF curves in clustering PCa patients. (b) Relative changes in the AUC of CDF curves. (c) PCa patients were divided into two clusters based on consensus clustering matrix. (d) The clinical characteristics of the two clusters in the heatmap. (e) Survival analysis in the two clusters.

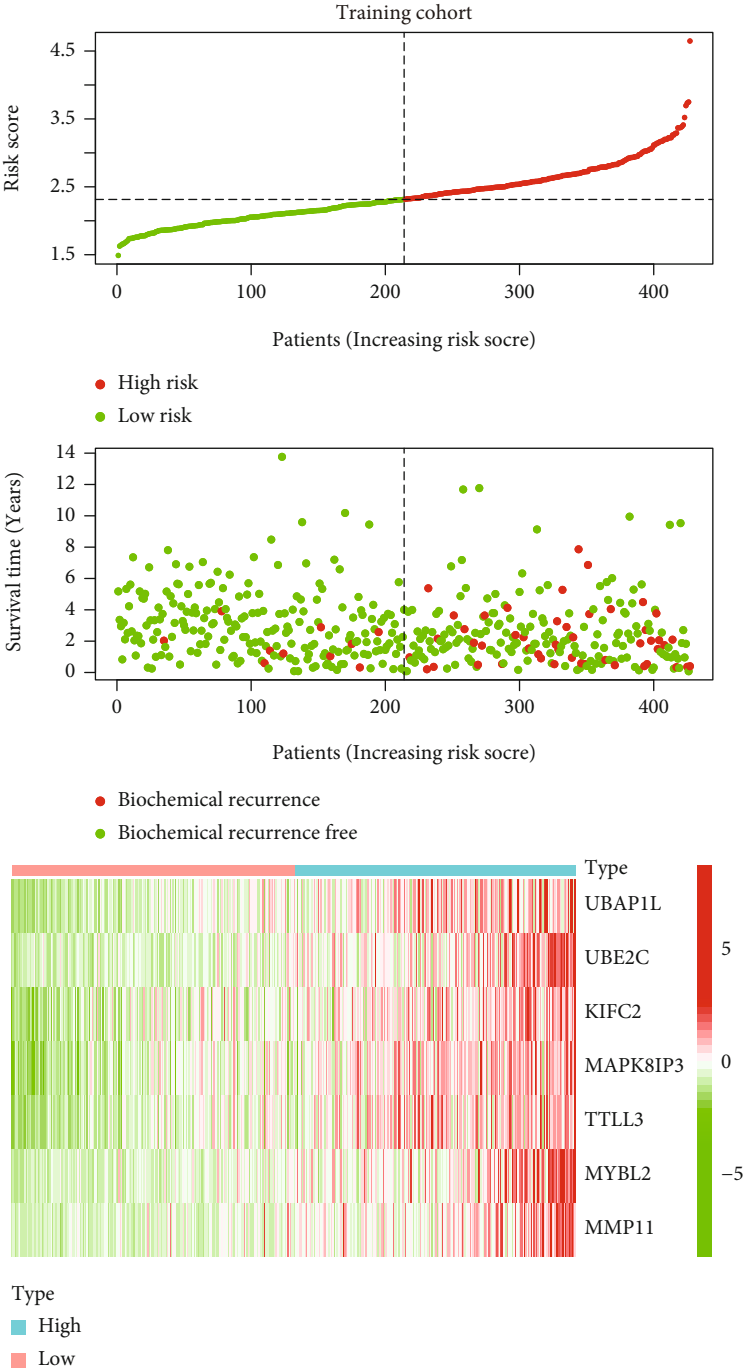


(a)



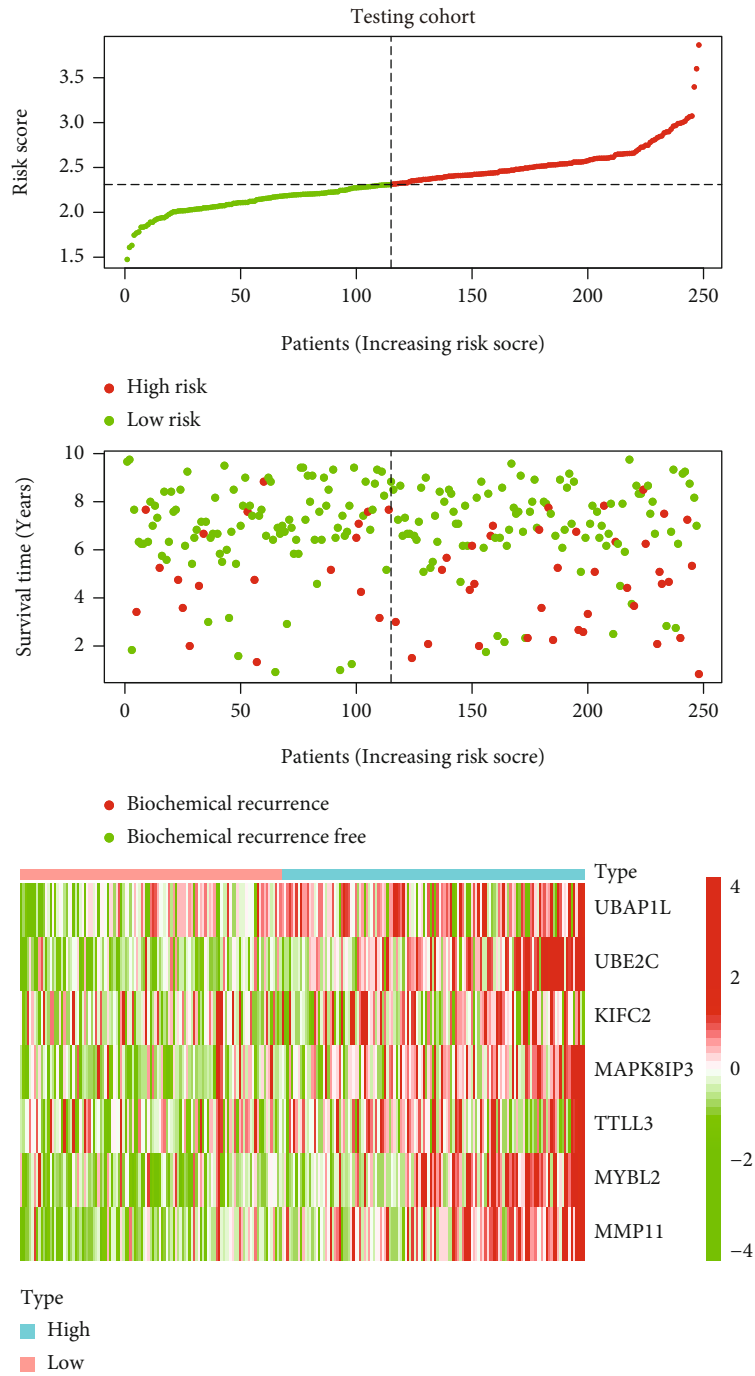
(b)

FIGURE 4: Continued.



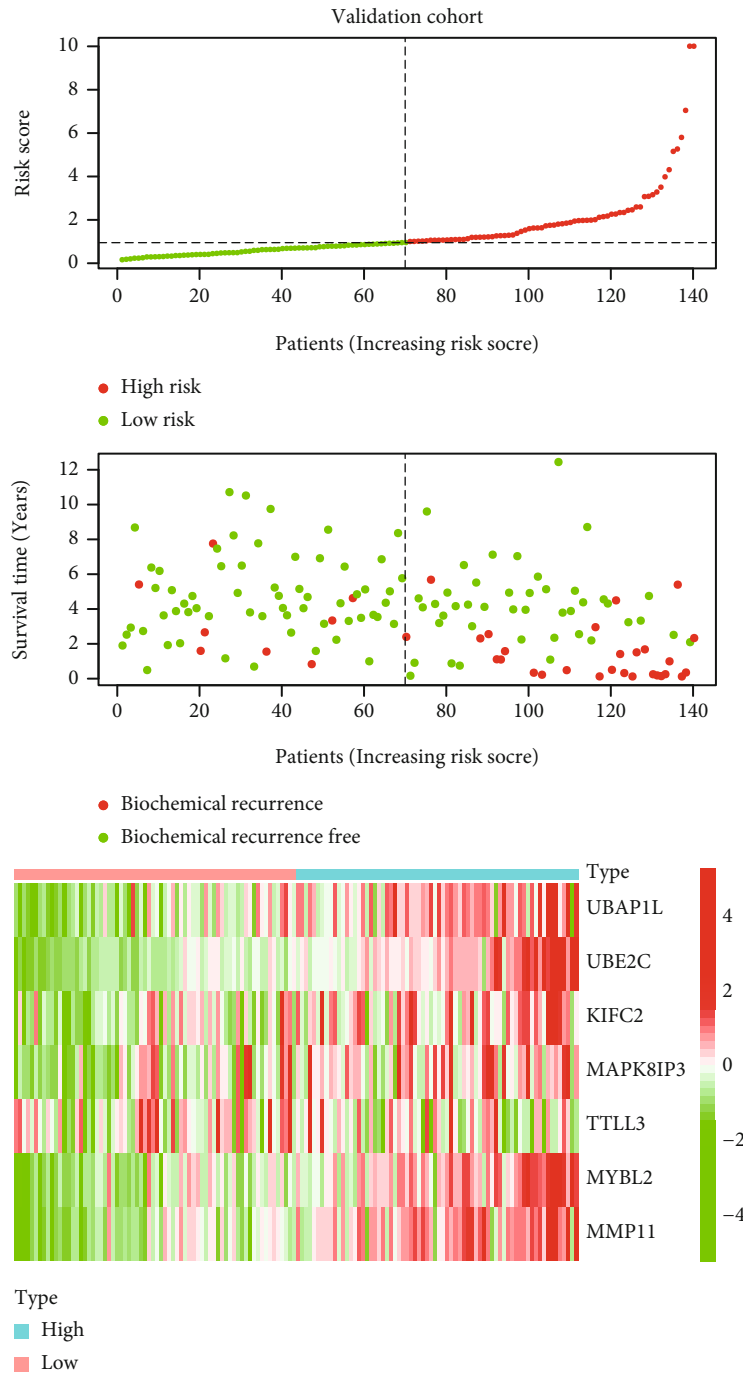
(c)

FIGURE 4: Continued.



(d)

FIGURE 4: Continued.



(e)

FIGURE 4: Identification of a signature to predict the BCR of PCa. (a and b) Process of variable selection in Lasso Cox regression and the optimal values of the penalty parameter were determined by 10-fold cross-validation in the training cohort. The risk score, survival status, and heatmap of the signature in the (c) training cohort, (d) testing cohort, and (e) validation cohort.

Furthermore, we obtained candidate gene through the least absolute shrinkage and selection operator (Lasso) with 10-fold cross-validation. In the end, we kept the 7 genes and the coefficients, and the penalty parameter (λ) was determined by the minimum criterion. The formula to calculate the risk score was as follows: Risk Score = $\sum_i^\lambda \beta_i S_i$, where β is the coefficients and S is the gene expression level.

2.5. Evaluation of the Signature. The area under curve (AUC) value of ROC curves was used to assess the sensitivity and specificity. A risk score was assigned to each patient according to the signature. Furthermore, we divided the PCa patients into high- and low-risk groups by the median value of risk score. Survival curves were plotted by the Kaplan-Meier analysis to assess the overall survival of

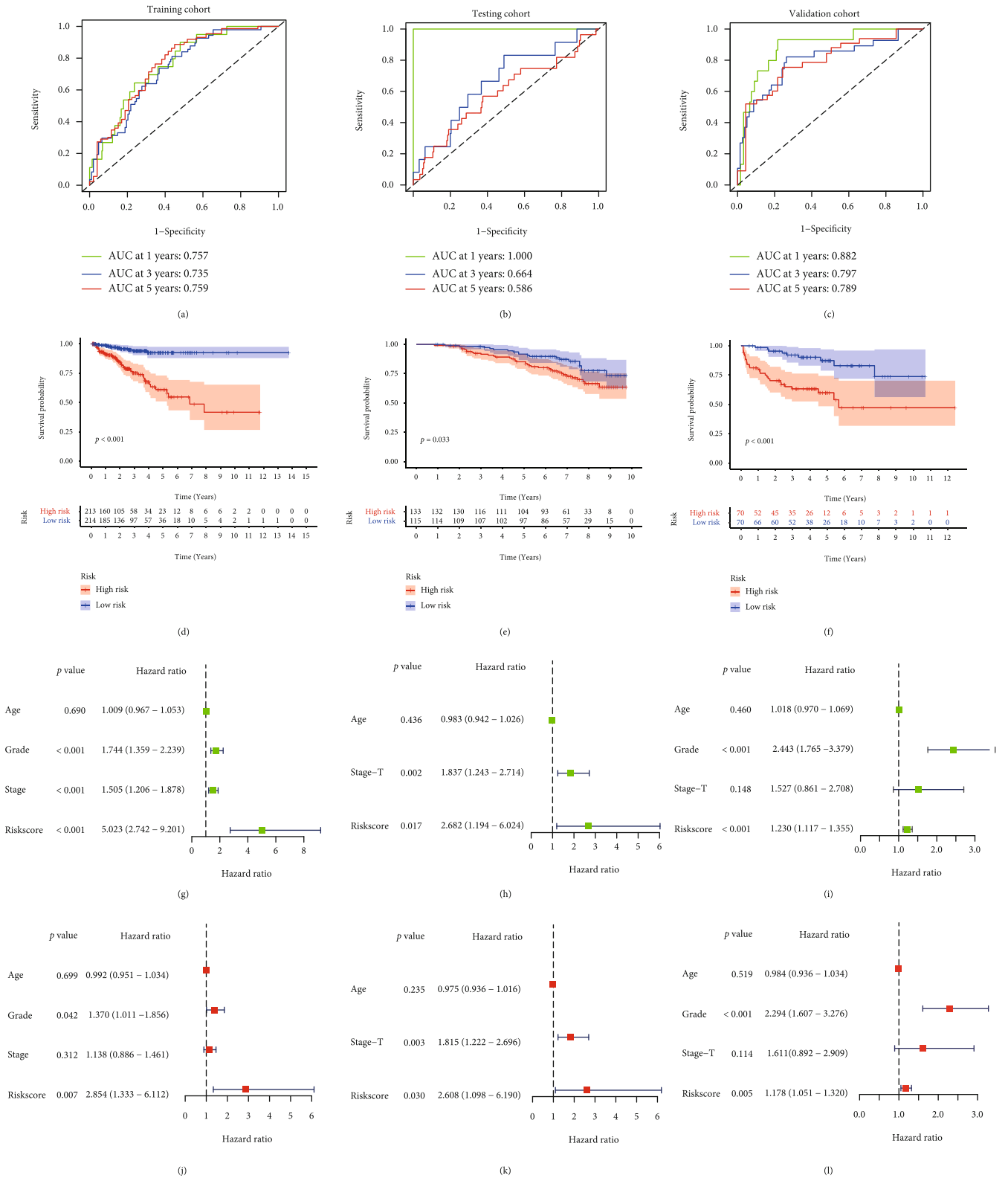
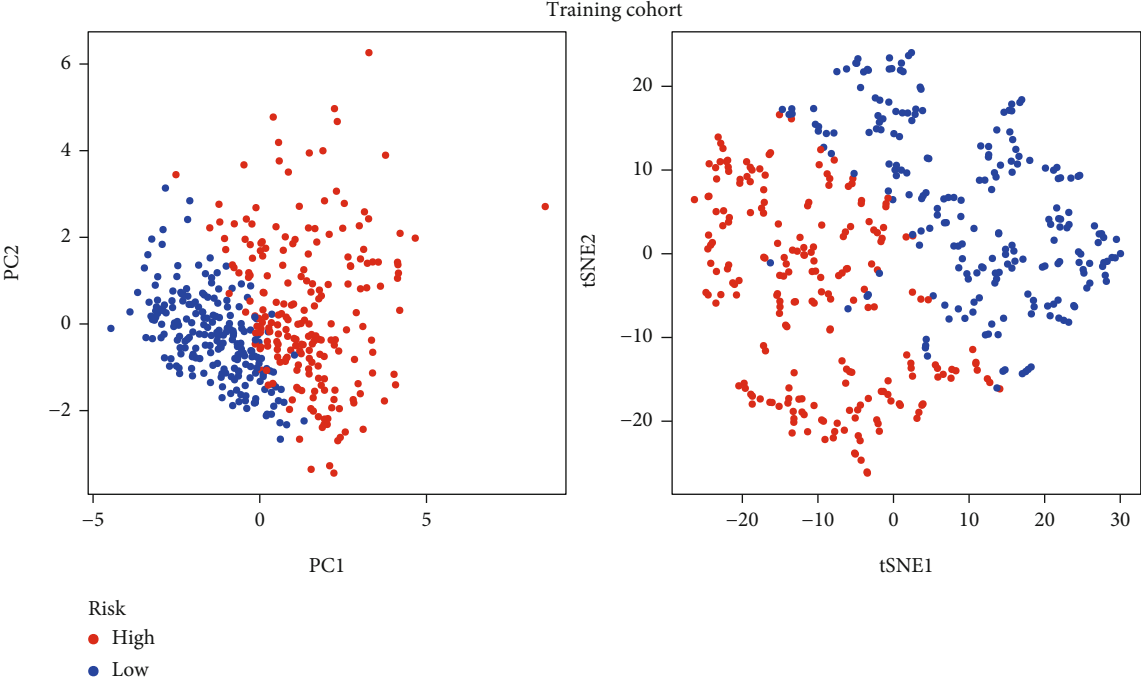
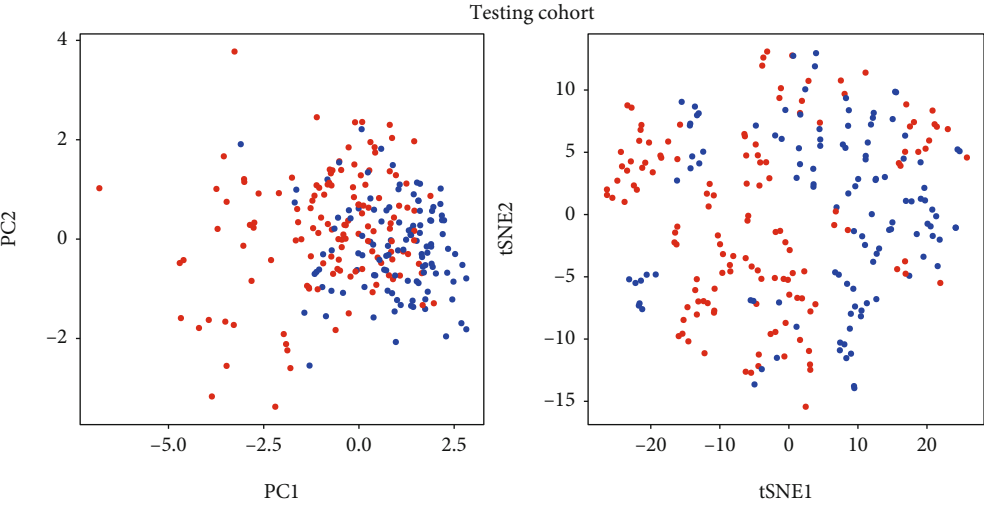


FIGURE 5: Validation of the signature in multiple cohorts. Time-dependent ROC curves analysis in the (a) training cohort, (b) testing cohort, and (c) validation cohort. The Kaplan-Meier survival curves based on the signature in the (d) training cohort, (e) testing cohort, and (f) validation cohort. Univariate analysis in the (g) training cohort, (h) testing cohort, and (i) validation cohort. Multivariate Cox regression in the (j) training cohort, (k) testing cohort, and (l) validation cohort.



(a)



(b)

FIGURE 6: Continued.

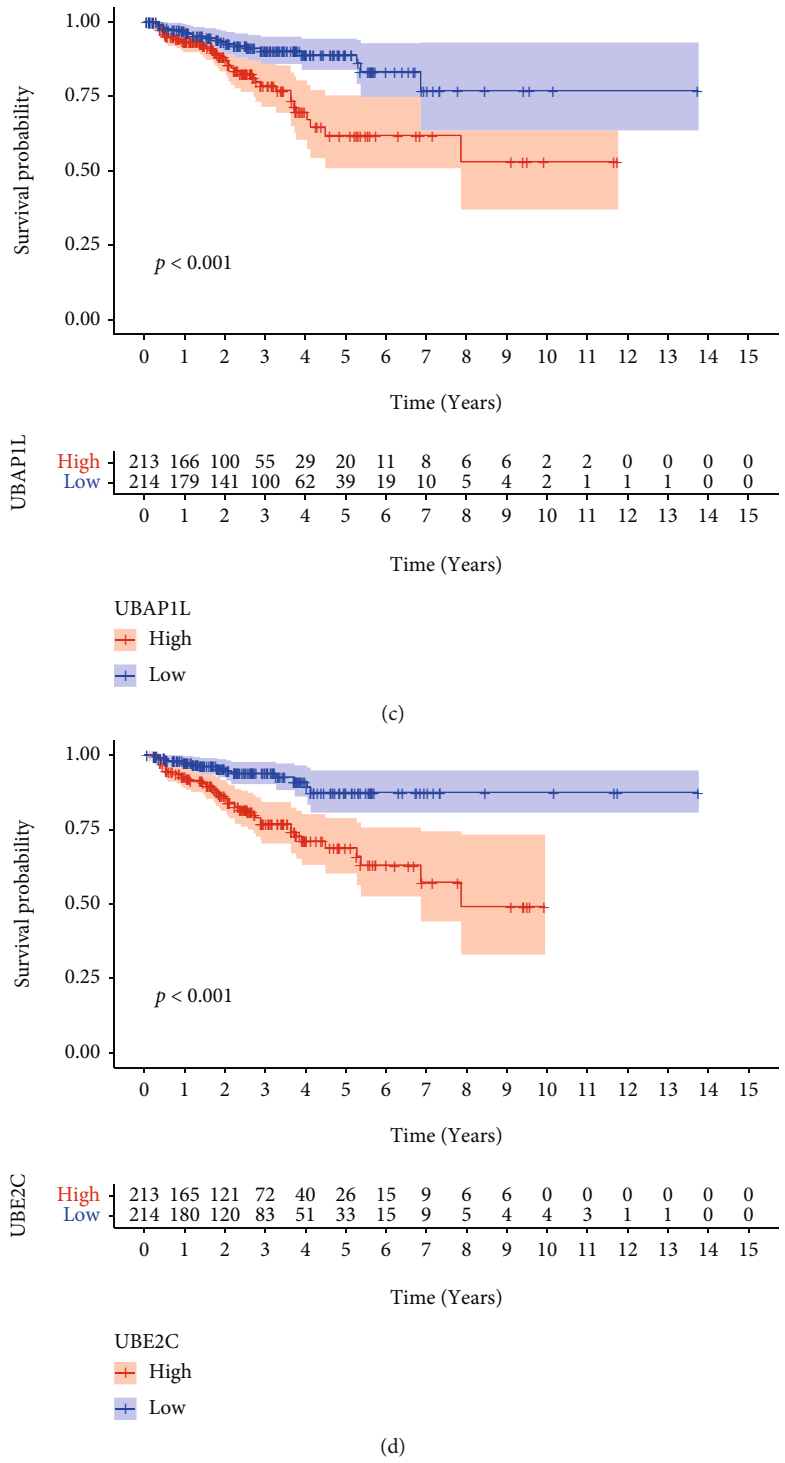


FIGURE 6: Continued.

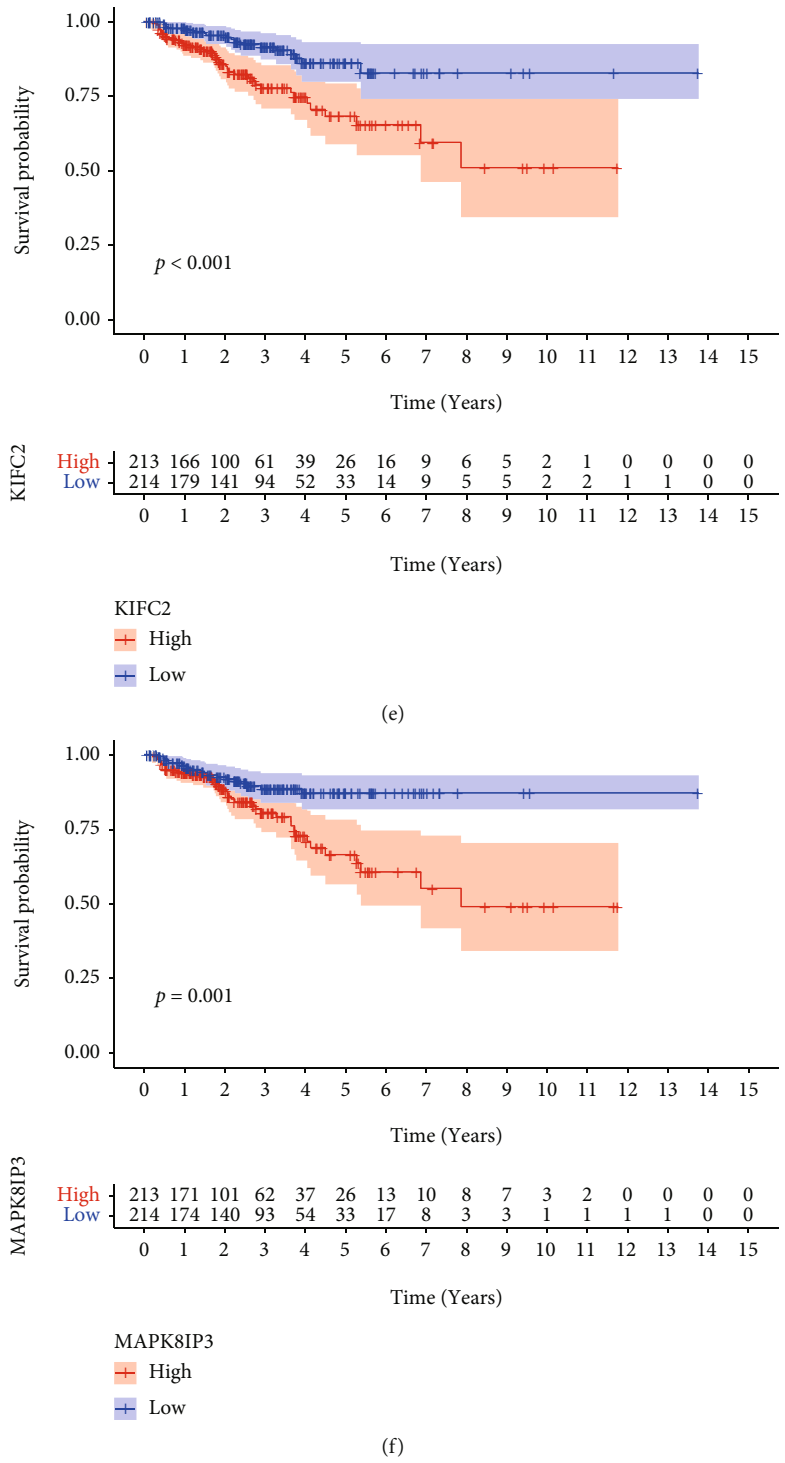
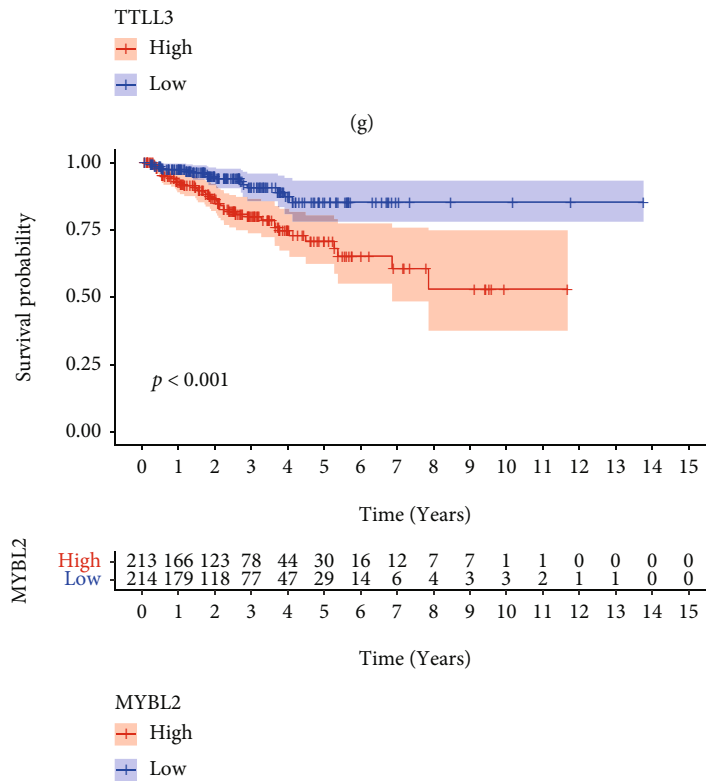
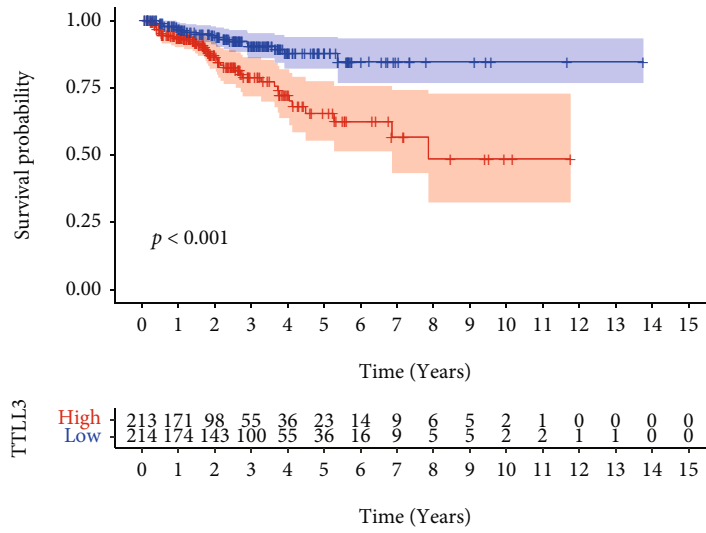
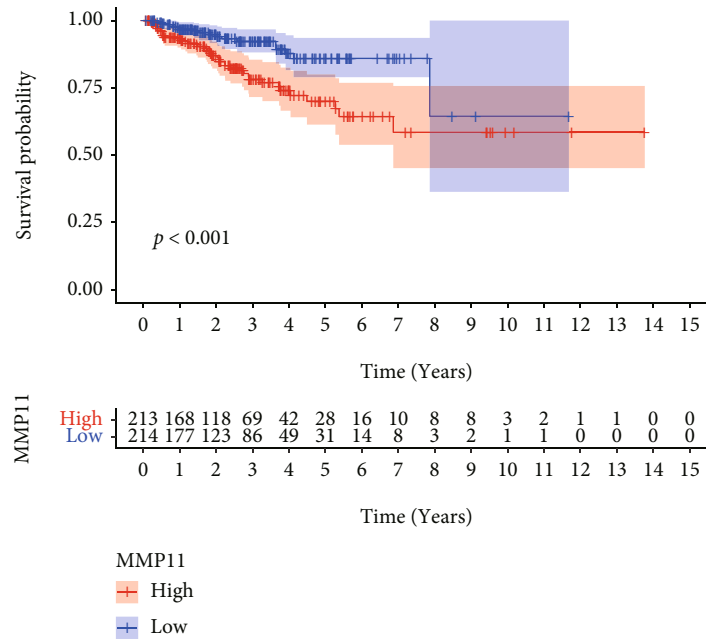


FIGURE 6: Continued.



(h)

FIGURE 6: Continued.



(i)

FIGURE 6: Distribution pattern and Kaplan-Meier survival analysis. (a) 2D PCA plot and t-SNE analysis between the high- and low-risk groups in the training cohort. (b) 2D PCA plot and t-SNE analysis between the two groups in the testing cohort. (c-i) The Kaplan-Meier survival curve of 7 genes between the two groups.

patients in the high- and low-risk groups. The univariate and multivariate Cox regression analyses were implemented to evaluate the independent prognostic value. These R software packages include *timeROC*, *survival*, and *survminer*.

2.6. Functional Enrichment Analysis. Gene Ontology (GO) including biological process (BP), cellular component (CC), and molecular function (MF) categories and the Kyoto Encyclopedia of Genes and Genomes (KEGG) were analyzed by the *clusterProfiler* R package.

2.7. Immune Landscape and TIDE Analysis. In order to explore the difference of the abundance of immune infiltrates in the high- and low-risk groups, we used the algorithms including EPIC, XCELL, MCPCOUNTER, QUANTISEQ, CIBERSORT-ABS, CIBERSORT, and TIMER to score the infiltration of each immune cell subtype. The significance threshold was set to a *P* value less than 0.05. The Wilcoxon sign-rank test was used to analyze the difference in the abundance of immune infiltrating cells between the high- and low-risk groups. The tumor immune dysfunction and exclusion (TIDE) of the PCa patients was calculated from the website (<http://tide.dfc.harvard.edu/>). The tumor inflammation signature (TIS) score was computed as the mean of log2-scale normalized expression of 18 signature genes [16].

2.8. Association between the Signature and the Treatments. To investigate the potential role of the signature in immunotherapy, we analyzed the relationship between the signature and immune checkpoints expression. Here, we adopted the

ggpubr package. In addition, we explored the function of signature in endocrine therapy and chemotherapy by analyzing the half-maximal inhibitory concentration (IC50) of the drugs. The difference in targeted therapy between the high- and low-risk groups was found by the Wilcoxon signed-rank test. The R packages used here were *pRRophetic* and *ggplot2*. NCI-60 database of 60 different tumor cell lines from 9 different tumor types was provided by CellMiner (<https://discover.nci.nih.gov/cellminer>). Pearson’s correlation analysis was carried to analyze the drug sensitivity between the expression of genes and 263 drugs approved by the FDA or in clinical trials.

2.9. Cell Line Culture and qRT-PCR. All human cell lines were purchased from the American Type Culture Collection (ATCC, USA), including DU145, PC3, and BPH-1. All cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium (Gibco, USA; catalog number: C11875500BT) supplemented with 10% fetal bovine serum (FBS; Gibco, USA; Cat.10270–106), 0.1 mg/mL streptomycin, and 100 U/mL penicillin (Gibco, USA; catalog number: 15,140–122) and were maintained in a humidified incubator at 37°C containing 5% CO₂. Total RNA was obtained with the RNeasy mini kit (QIAGEN, Germany, Cat. No. 74,104) and reverse transcribed with the RT kit (TaKaRa, Japan, Cat. No. NR037A). The cDNA products were then subjected to real-time PCR using Fast SYBR® Green Master Mix (Life technology, USA; Cat. No: 4,385,610). The sequences of all primers used for PCR were documented in the supplementary materials.

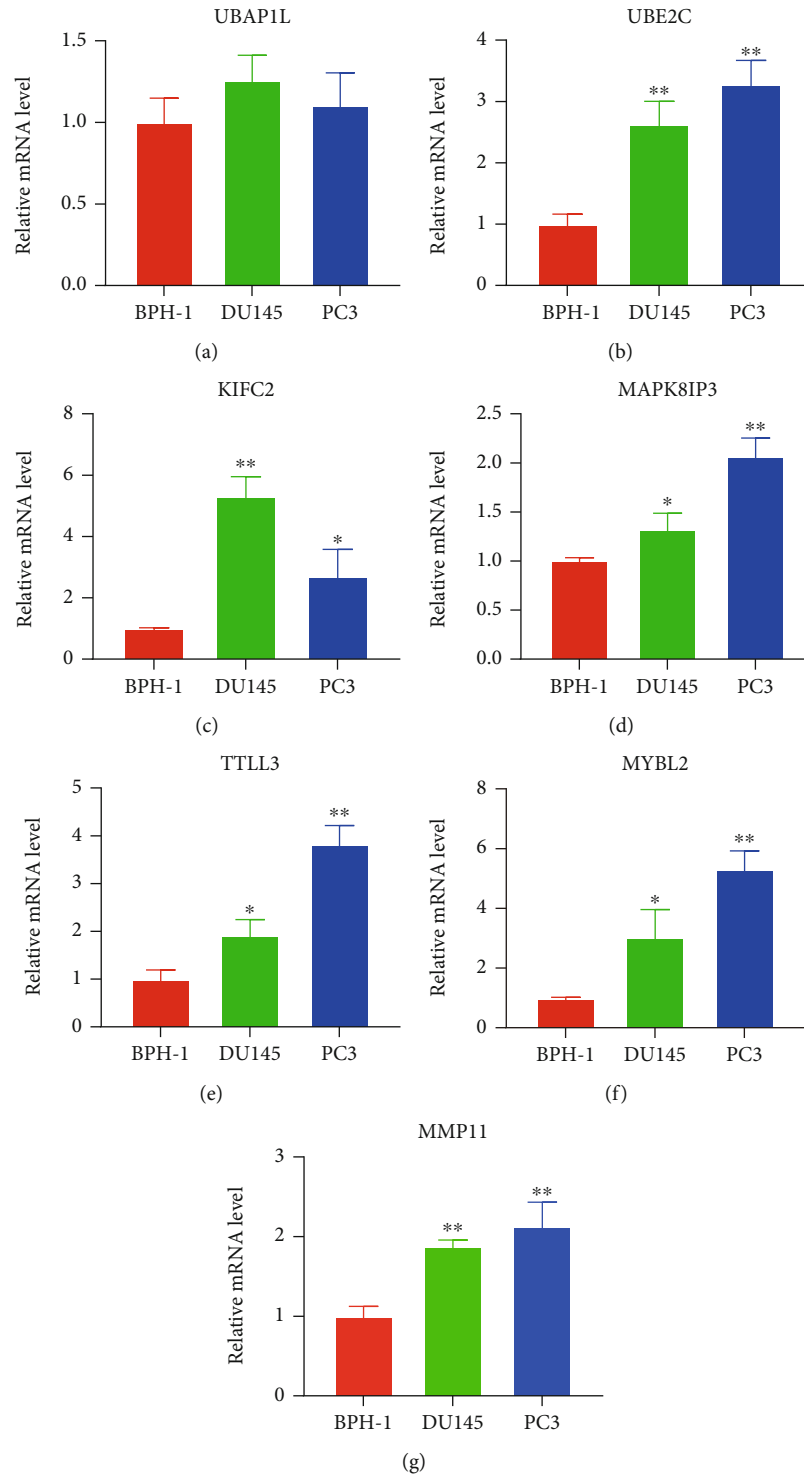
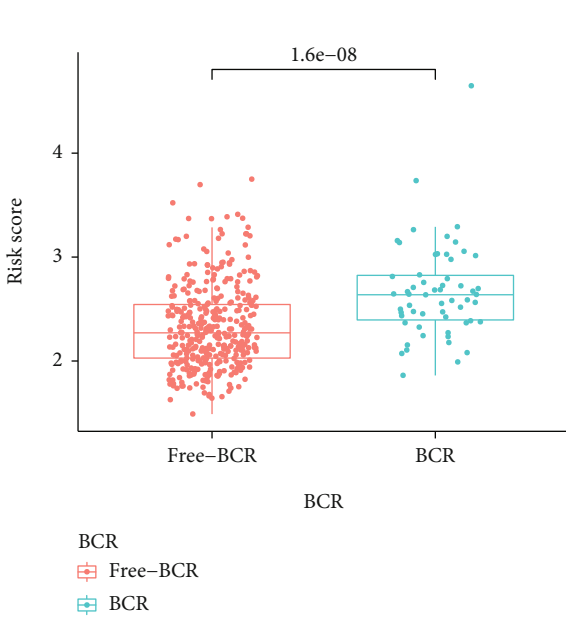
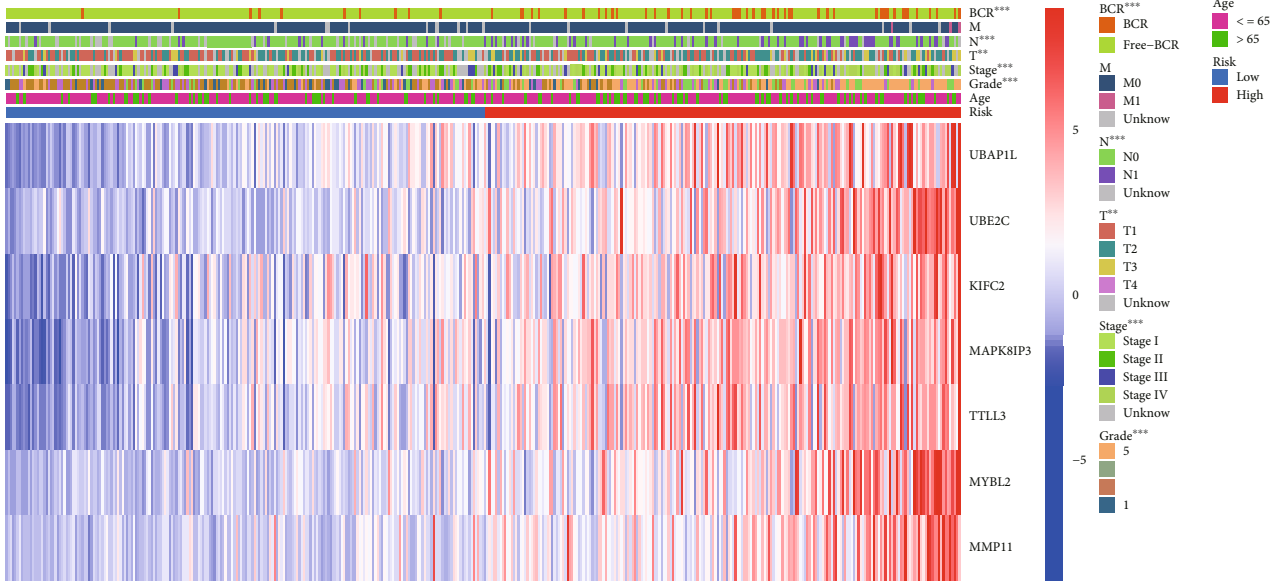


FIGURE 7: The expression of seven genes in PCa cell lines. (a–g) The relative mRNA levels of UBAP1L, UBE2C, KIFC2, MAPK8IP3, TTLL3, MYBL2, and MMP11 in DU145, PC3, and BPH-1.

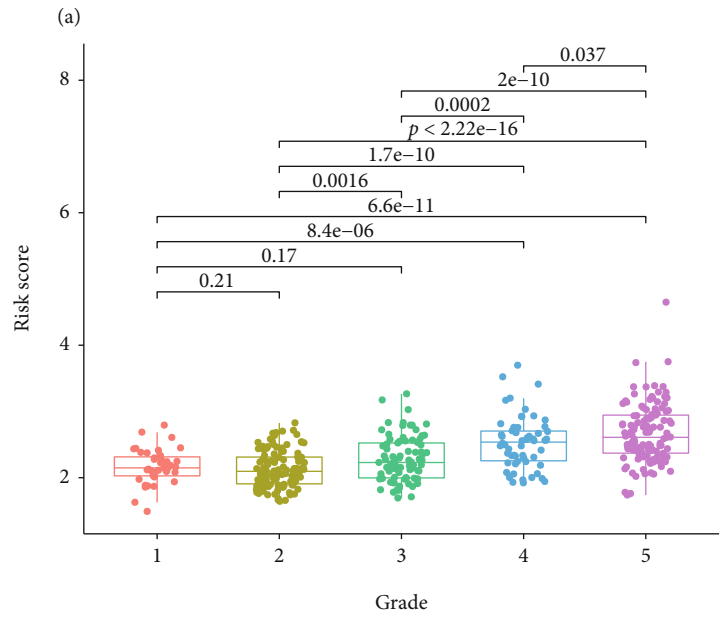
2.10. Statistical Analysis. All statistical analyses were applied by R version 4.1.1 (Institute for Statistics and Mathematics, Vienna, Austria; <https://www.r-project.org>), and some related packages were applied to all statistical analyses. $P < 0.05$ was considered the significantly statistical difference.

3. Result

3.1. Screening Differentially Expressed Pyroptosis-Related Genes. The brief process of this research was depicted in Figure 1. Initially, we compared the expression of 33



(b)



(c)

FIGURE 8: Continued.

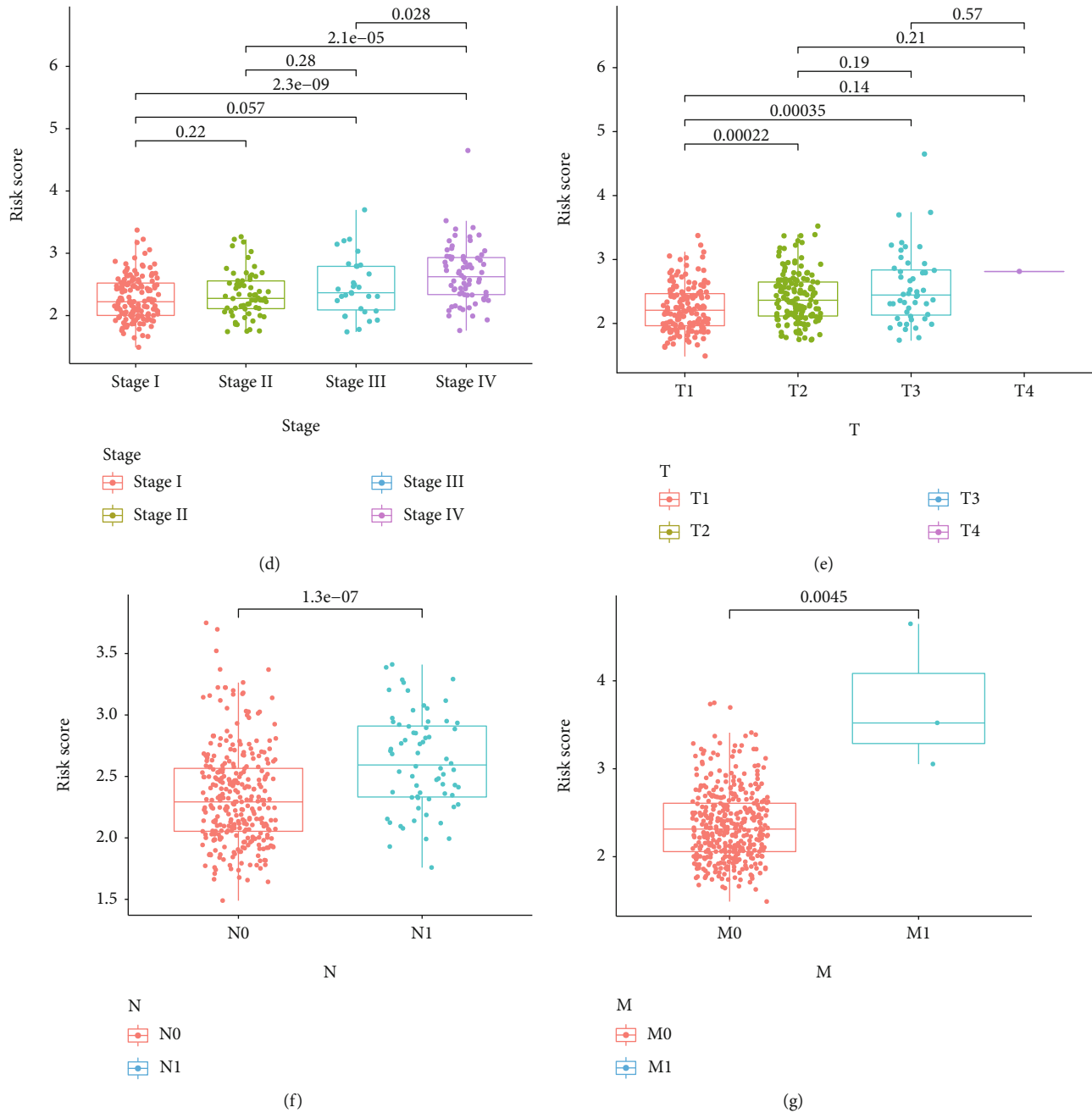
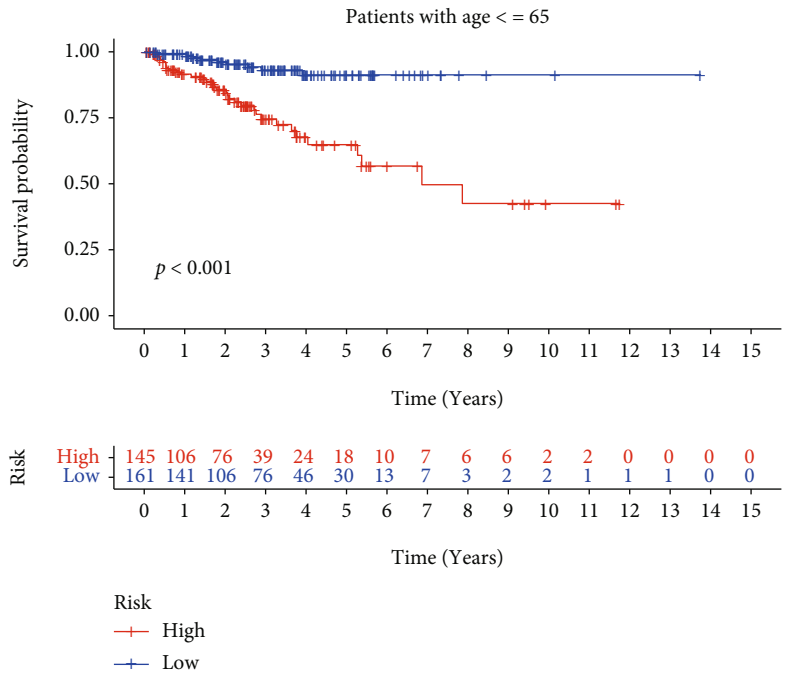


FIGURE 8: Evaluating the relationship between the signature and clinical characteristics of PCa. (a) The distribution of clinicopathological factors between the high- and low-risk groups. Risk scores were significantly associated with BCR (b), tumor grade (c), tumor stage (d), T stage (e), N stage (f), and M stage (g).

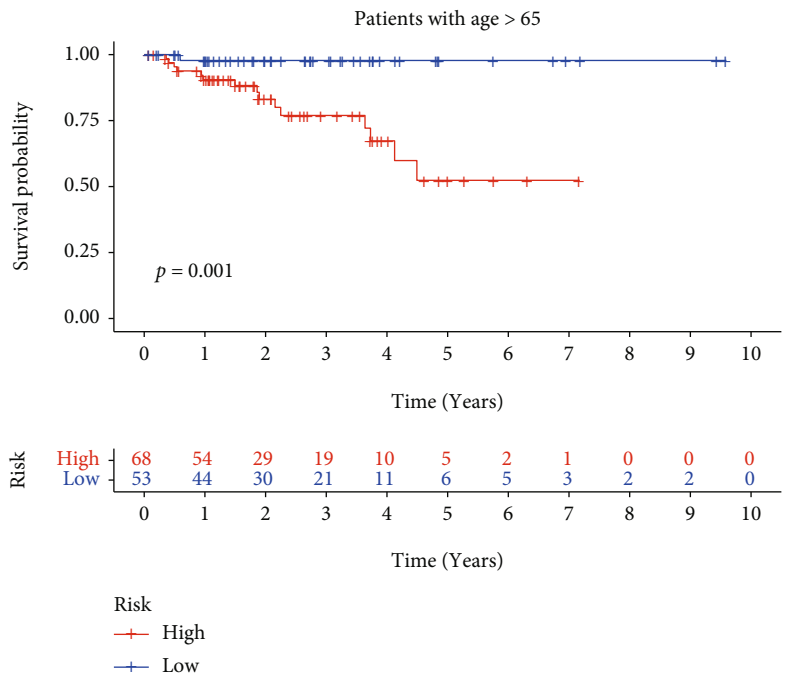
pyroptosis-related genes in 52 normal tissues and 499 PCa samples from the TCGA database and identified 22 differentially expressed genes (DEGs), which were depicted in the heatmap (all $P < 0.001$) (Figure 2(a)). Protein-protein interaction (PPI) analysis with the minimum required interaction score of 0.9 was employed to investigate the interactions of these DEGs. CASP1, CASP8, IL1B, and PYCARD were identified as hub genes (Figure 2(b)). Furthermore, the correlation network of the DEGs was illustrated in Figure 2(c). The analysis of CNV alteration frequency exhibited that most DEGs were focused on copy number reduction

(Figure 2(d)). We further annotated the sites of CNV alterations of DEGs on the chromosome (Figure 2(e)). In order to further explore the biological processes and potential molecular mechanisms that the DEGs involved, we conducted GO analysis and KEGG pathway, revealing the participation of many biological processes and signaling pathways (Figures 2(f) and 2(g)).

3.2. Classification of PCa Patients Based on Pyroptosis-Related Genes. The empirical CDF was depicted to identify the optimum k values for the distribution of samples with



(a)



(b)

FIGURE 9: Continued.

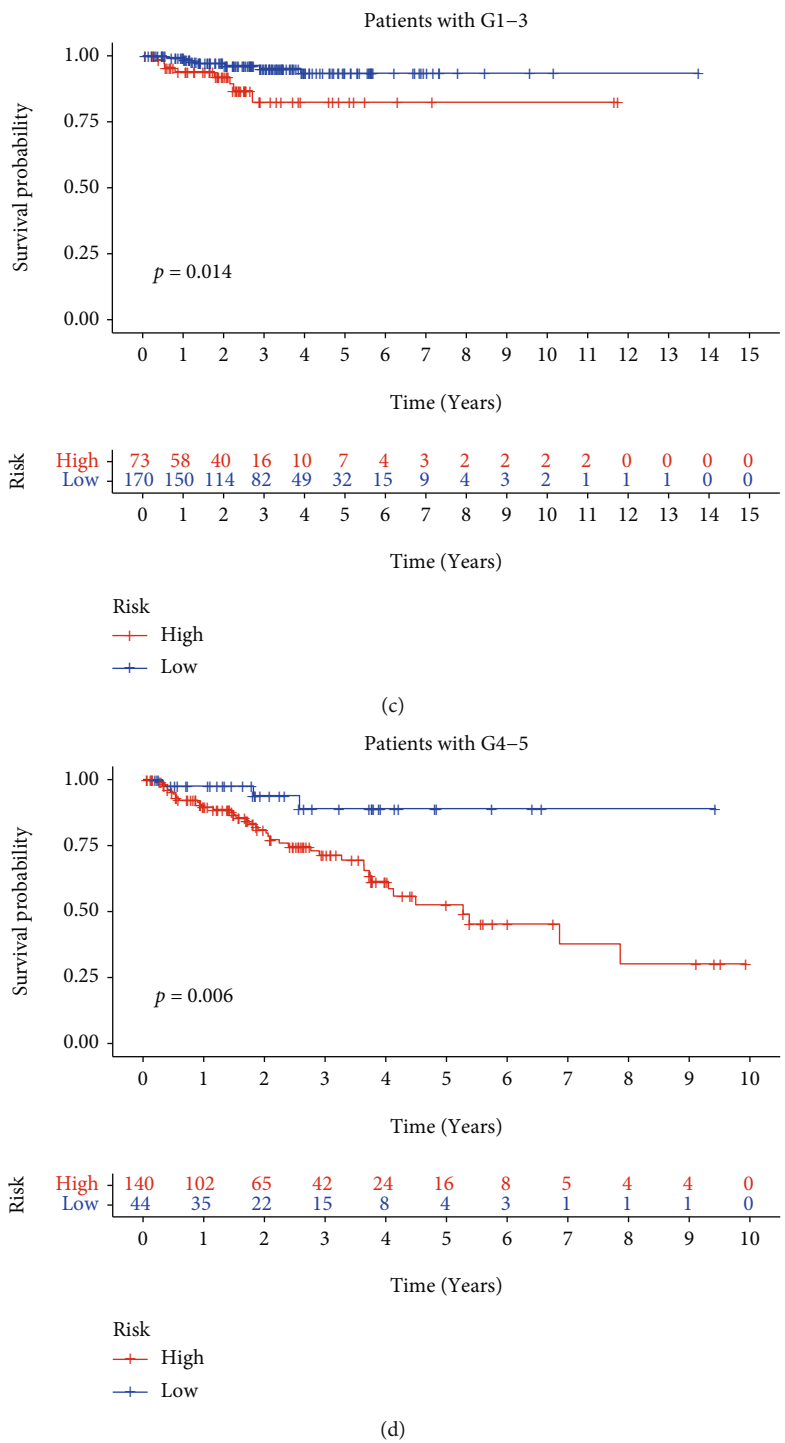
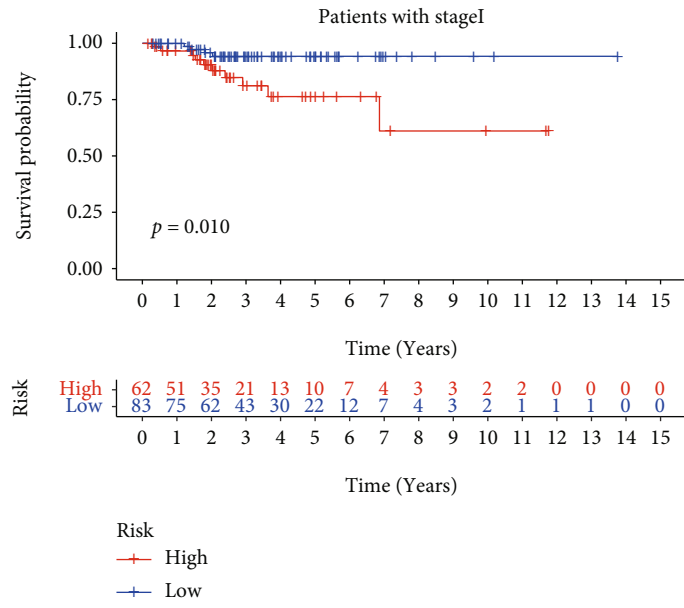
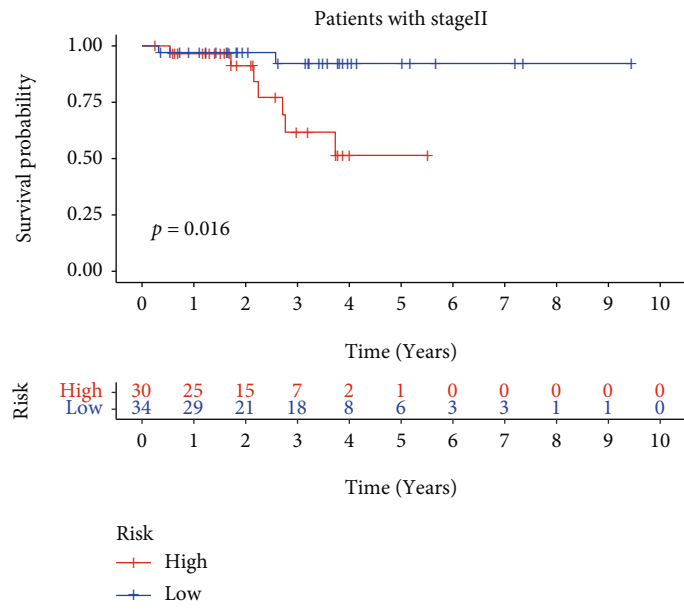


FIGURE 9: Continued.



(e)



(f)

FIGURE 9: Continued.

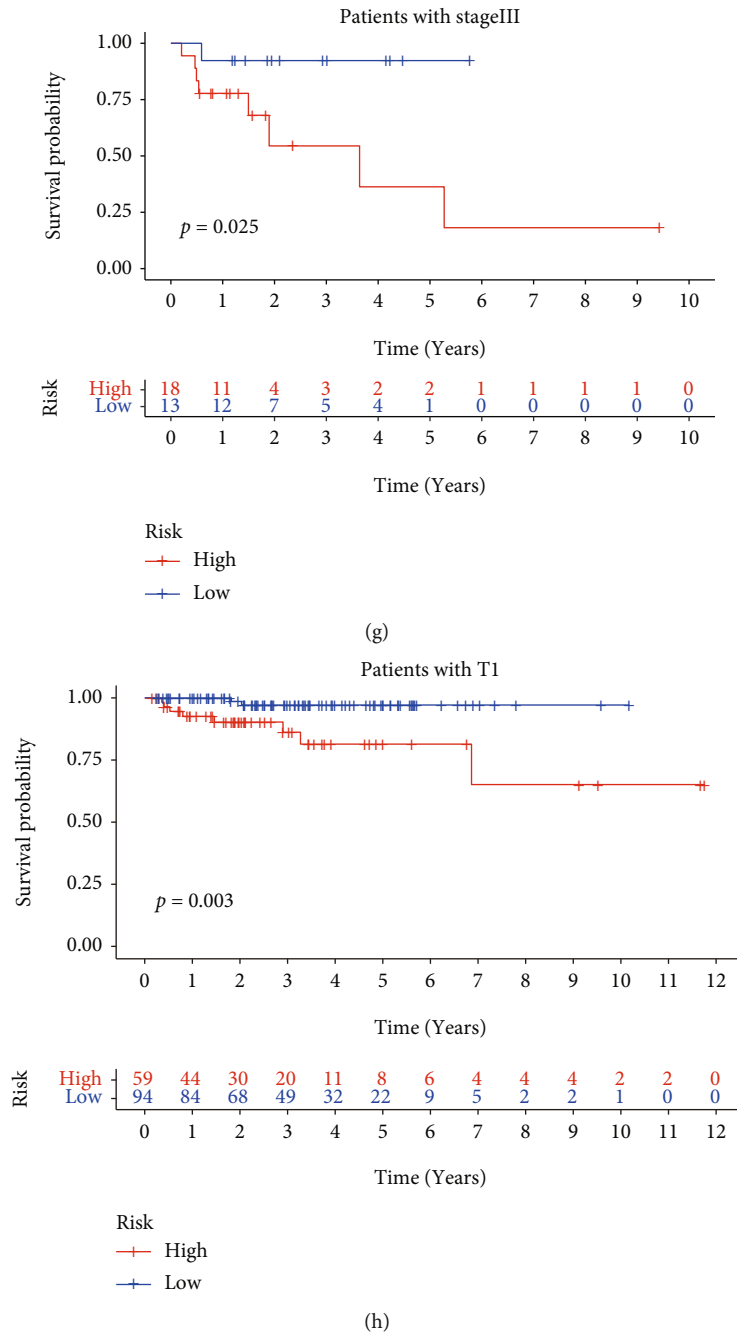


FIGURE 9: Continued.

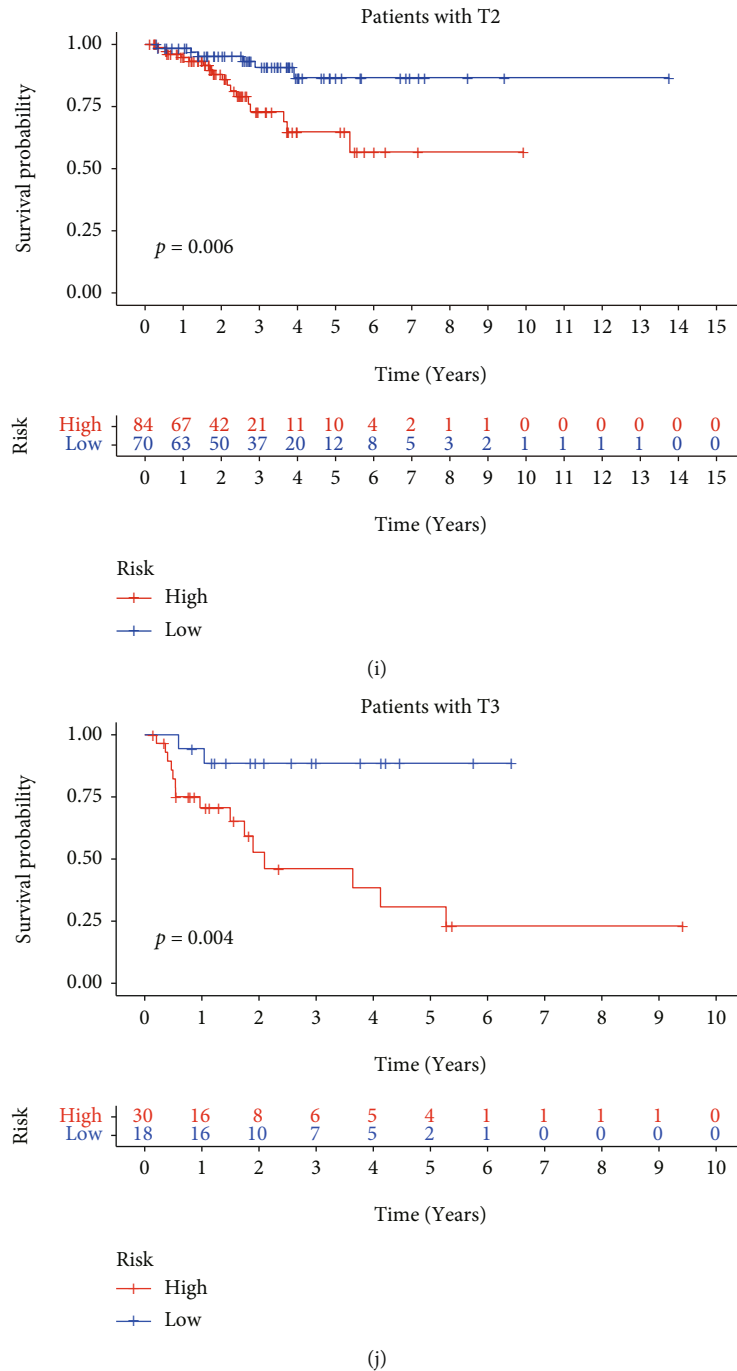
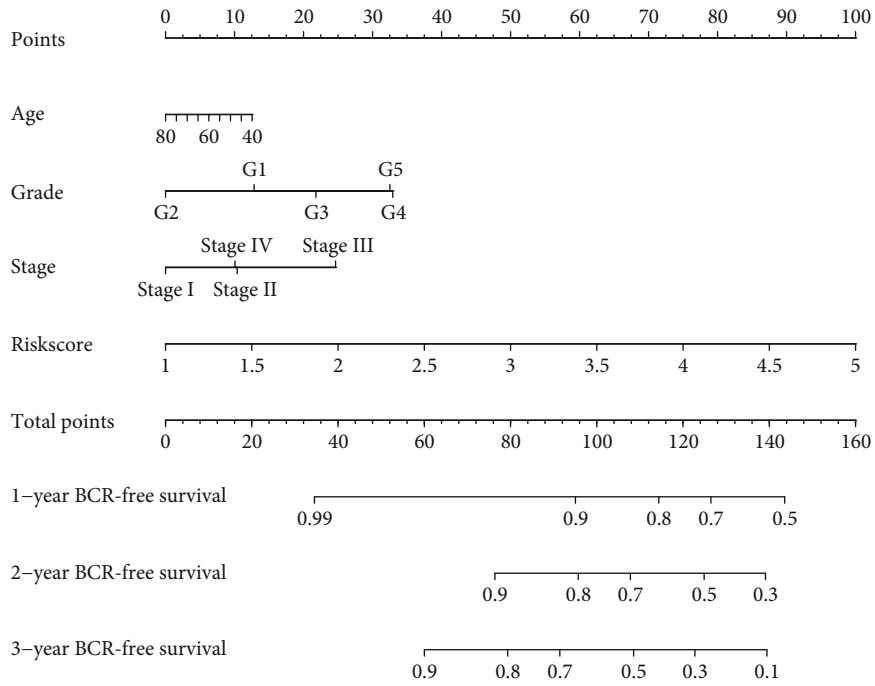


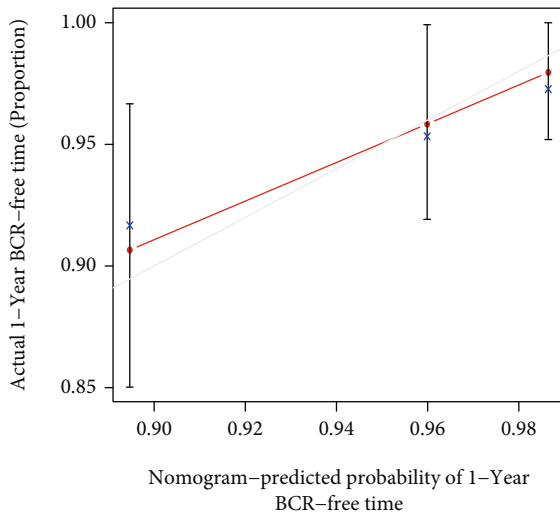
FIGURE 9: Stratification survival analyses. (a-j) The Kaplan-Meier curve analyses of overall survival in subgroups stratified by different clinical features.

maximal stability (Figures 3(a) and 3(b)). The result of consensus matrices suggested that PCa patients can be divided into two completely different clusters when clustering variable (k) = 2 (Figure 3(c)). We found significant differences in the clinical characteristics including BCR, M stage, N stage, T stage, tumor stage, and tumor grade between these two different clusters (Figure 3(d)). In addition, the Kaplan-Meier survival analysis confirmed that patients in cluster 2 had a shorter BCR-free time than those in cluster 1 ($P < 0.001$) (Figure 3(e)).

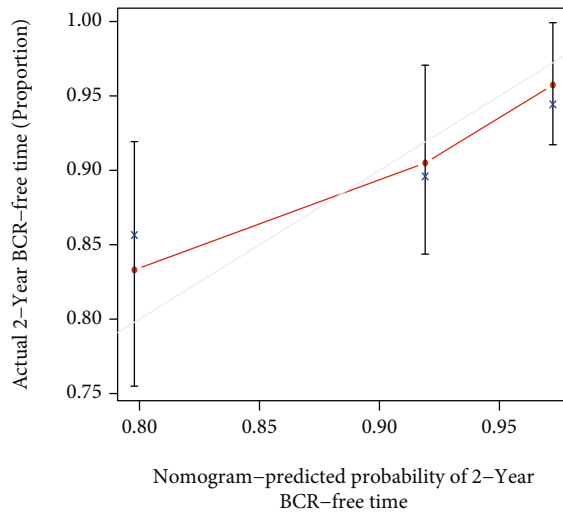
3.3. Construction and Evaluation of Prognostic Signature for PCa. To identify a specific prognostic signature for disease diagnosis and treatment, we explored differentially expressed genes between the above two clusters. Then, we performed univariate Cox regression and Lasso regression analysis, in which the best values of the penalty parameter were determined by 10-fold cross-validation (Figures 4(a) and 4(b)). Finally, 7 effective genes for the construction of the risk signature were determined. The PCa patients were stratified into high-risk and low-risk groups according to



(a)



(b)



(c)

FIGURE 10: Continued.

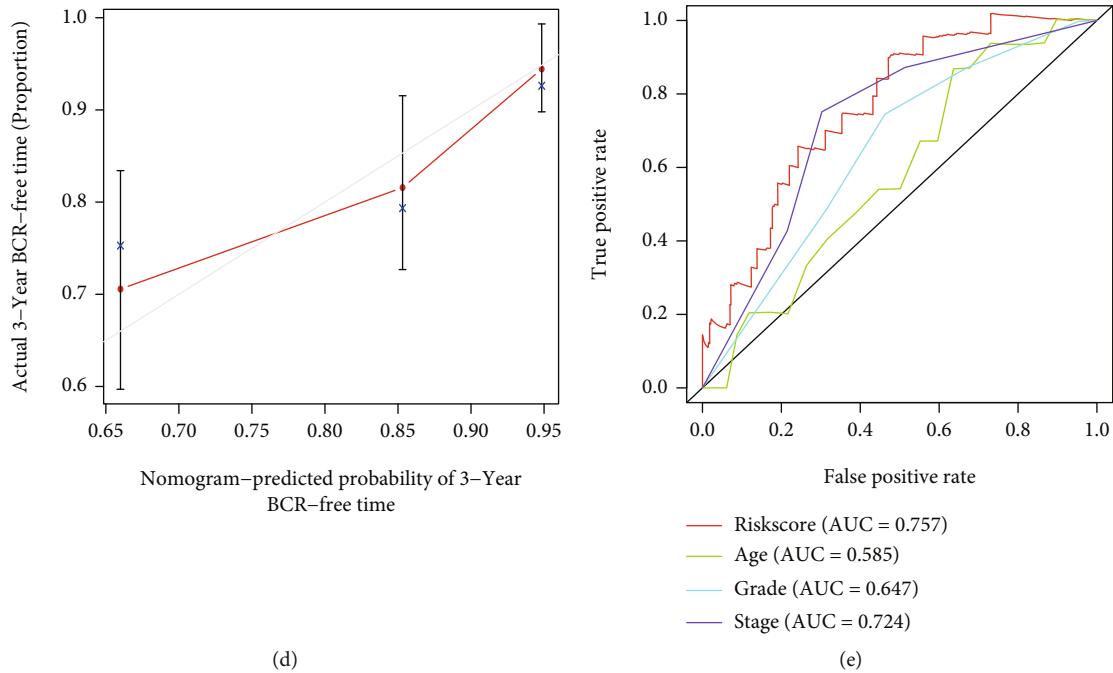


FIGURE 10: Construction and validation of nomogram. (a) The nomogram for predicting the probability of the 1-, 2-, and 3-year BCR-free survival. (b–d) Calibration curves for the validation of the nomogram. (e) Time-dependent ROC curves analysis of signature and the clinical factors.

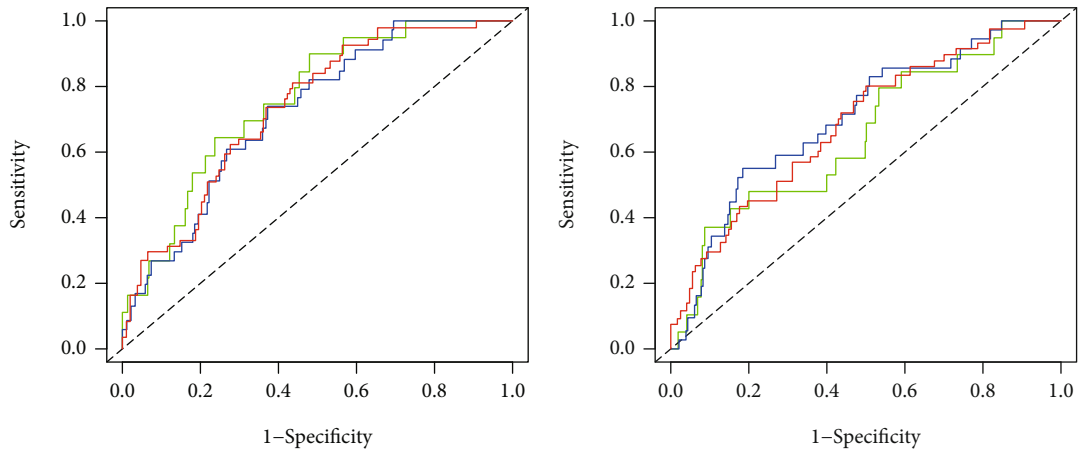
the median risk score as the cut-off point. The distribution of risk score showed a significant difference in BCR-free time among the training cohort, testing cohort, and independent external validation cohort, with a gradual increase in the probability of BCR as the risk score increased (Figures 4(c) – 4(e)). Furthermore, we performed time-dependent ROC analysis and calculated the AUC at 1, 3, and 5 years, showing good sensitivity and specificity of the signature for prognosis of PCa patients in three cohorts (Figures 5(a) – 5(c)). The result of the Kaplan-Meier survival curve indicated that the patients in the high-risk group suffered shorter BCR-free time, showing the same outcome in all three cohorts (Figures 5(d) – 5(f)). The univariate and multivariate Cox regression proved that the signature could serve as a robust and independent prognostic factor for PCa patients (Figures 5(g) – 5(l)).

3.4. Distribution Patterns of the High-Risk and Low-Risk Groups. PCA and t-SNE analyses were conducted to reduce dimensionality and showed a satisfactory separation between the high- and low-risk groups. The distribution of the high- and low-risk groups tended to be in different directions (Figures 6(a) and 6(b)). Furthermore, we explored the impact of the 7 genes used to construct the signature on BCR-free time. Surprisingly, patients had higher probability of BCR when each of these genes was highly expressed (Figures 6(c) – 6(d)). We further analyzed the mRNA expression of the 7 genes used to construct the signature in two PCa cell lines (DU145 and PC3) and benign prostatic hyperplasia cell (BPH-1) by qRT-PCR assays. These results indicated that the expression levels of UBE2C, KIFC2, MAPK8IP3, TTLL3, MYBL2, and MMP11 were significantly

upregulated in PCa cell lines, except for UBAP1L which did not show significant differences (Figures 7(a) – 7(g)).

3.5. Correlation between Clinicopathological Characteristics and the Signature. The distributed patterns between the signature and clinicopathological characteristics were illustrated on the heatmap (Figure 8(a)). The BCR, M stage, N stage, T stage, tumor stage, tumor grade, and age were diversely distributed in the high- and low-risk groups. To further investigate whether the signature was closely related to different clinicopathological conditions, we found that the clinical features including BCR, tumor grade, tumor stage, T stage, N stage, and M stage were significantly associated with the signature (Figures 8(b) – 8(g)). The high-grade and advanced-stage patients were more likely to be related to the high-risk group. In addition, the low-risk group was more inclined to low grade and early stage, which were equivalent to a better prognosis. We further divided PCa patients into different stratified groups according to age, gender, tumor grade, tumor stage, and T stage. There were significant differences between the high- and low-risk groups, suggesting that the low-risk group had longer BCR-free time in all stratification subgroups. (Figures 9(a) – 9(k)) Therefore, the signature might be significantly associated with the progression of PCa and had broad applicability and feasibility for prognosis prediction.

3.6. Construction and Evaluation of the Nomogram. We constructed a nomogram containing risk scores and clinical characteristics to predict the 1-, 2-, and 3-year BCR probability of PCa patients. A higher total score in the nomogram represented a worse prognosis (Figure 10(a)). The

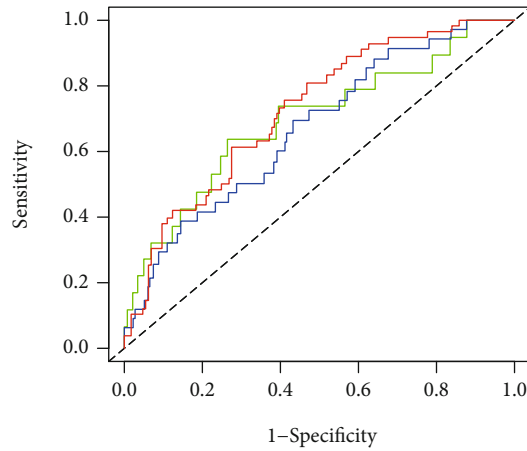


Pyroptosis signature
 — AUC at 1 years: 0.757
 — AUC at 2 years: 0.723
 — AUC at 3 years: 0.735

Luan1 signature
 — AUC at 1 years: 0.651
 — AUC at 2 years: 0.702
 — AUC at 3 years: 0.685

(a)

(b)



Shao signature
 — AUC at 1 years: 0.694
 — AUC at 2 years: 0.669
 — AUC at 3 years: 0.723

(c)

FIGURE 11: Continued.

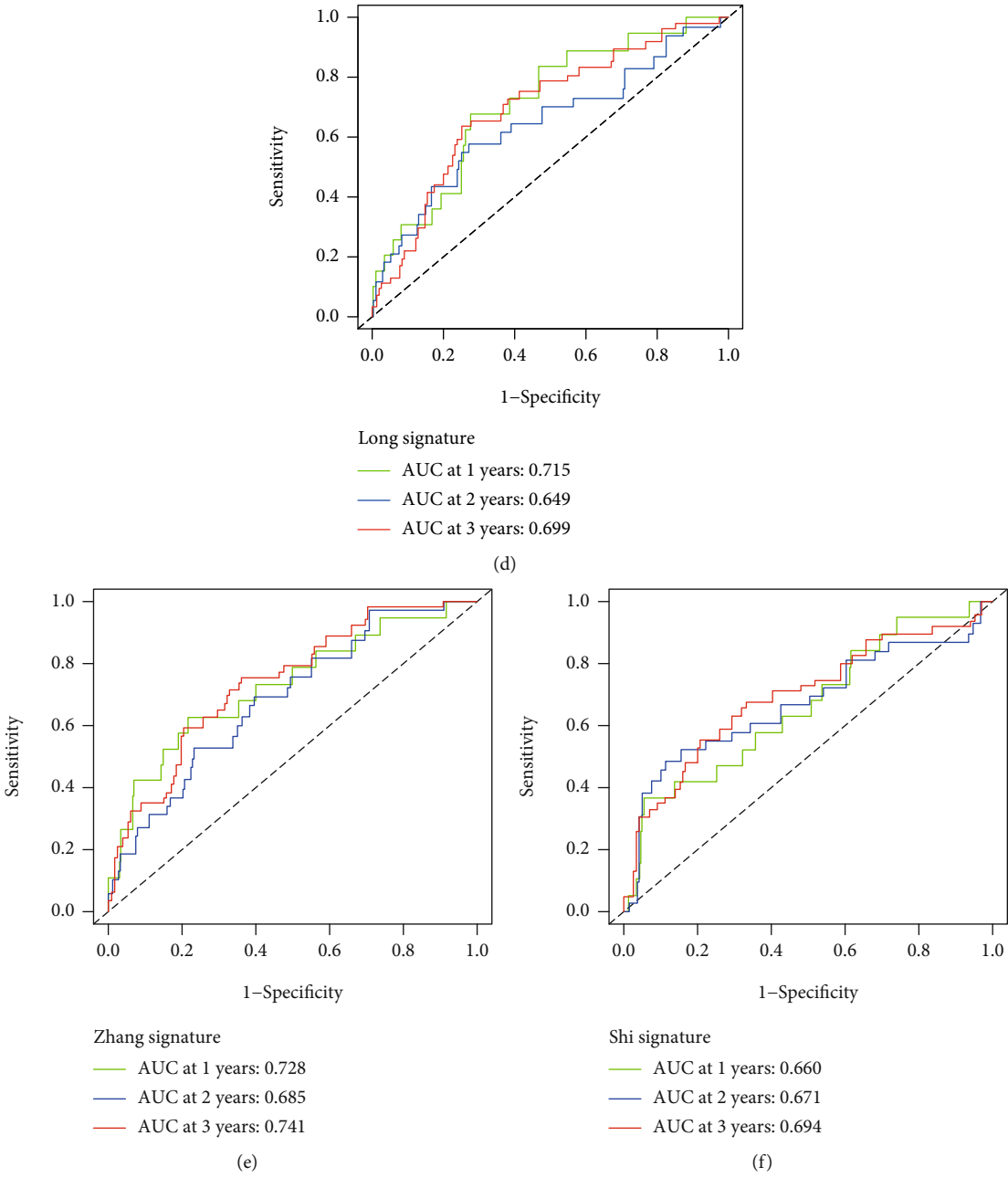
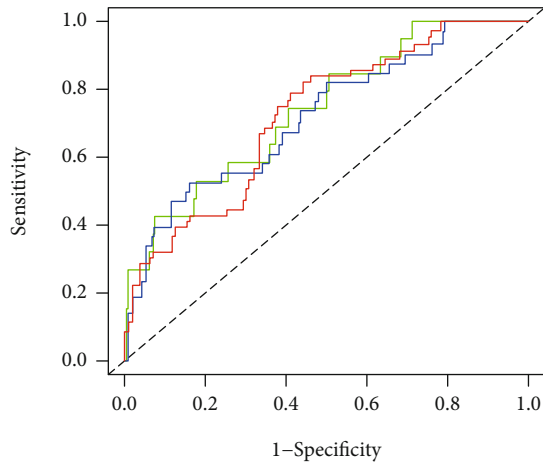
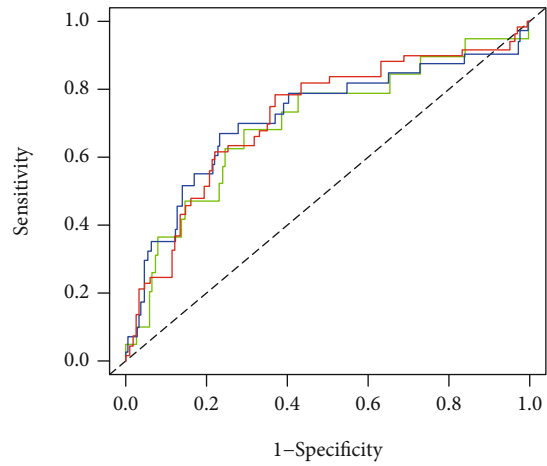


FIGURE 11: Continued.



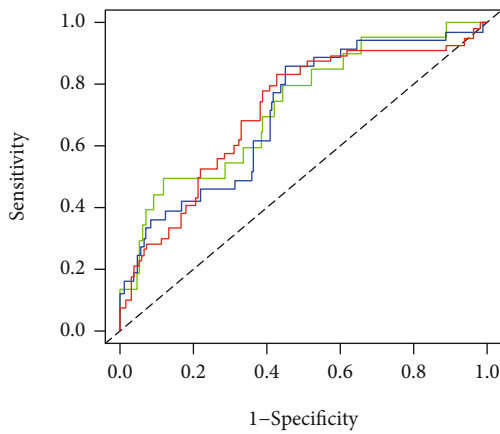
Gao signature
 — AUC at 1 years: 0.739
 — AUC at 2 years: 0.719
 — AUC at 3 years: 0.722

(g)



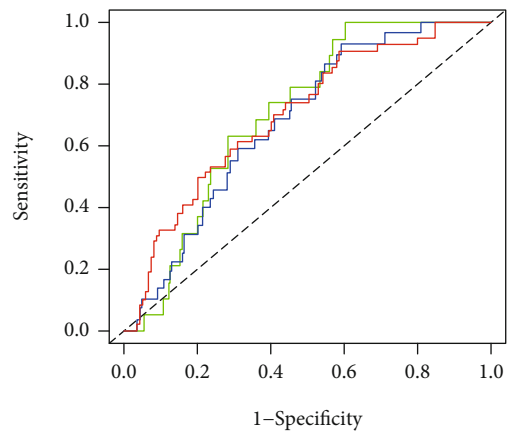
Yuan signature
 — AUC at 1 years: 0.699
 — AUC at 2 years: 0.717
 — AUC at 3 years: 0.718

(h)



Liu signature
 — AUC at 1 years: 0.725
 — AUC at 2 years: 0.708
 — AUC at 3 years: 0.711

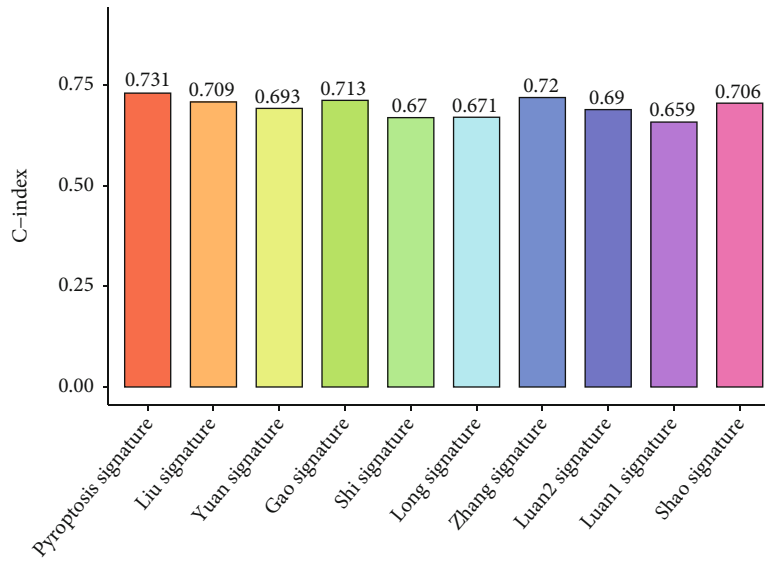
(i)



Luan2 signature
 — AUC at 1 years: 0.703
 — AUC at 2 years: 0.678
 — AUC at 3 years: 0.700

(j)

FIGURE 11: Continued.



(k)

FIGURE 11: Comparison with other 9 published gene signatures (a–j). (k) C-index of signatures.

calibration chart displayed excellent agreement between observed and predicted rates at 1, 2, and 3 years (Figures 10(b) – 10(d)). By comparing the AUC between the signature and clinical features, we found that our signature can predict BCR more accurately (Figure 10(e)). Thus, our nomogram based on the signature had good predictive ability in clinical practice.

3.7. Comparison with Other Gene Expression Signatures. To determine whether our signature was superior to other signatures, we compared the signatures constructed for PCa in 9 published articles [17–25]. We found that the accuracy and stability of our signature in 1, 2, and 3 years were better than those of the nine signatures in the ROC curves analysis (Figures 11(a) – 11(j)). Then, in order to further compare our signature with the predicted performance of these signatures, we calculated the concordance index (C-index). As the results depicted, the C-index of our signature was 0.731 (Figure 11(k)), which was better than other signatures.

3.8. Landscape of Somatic Mutations in PCa. We analyzed the TMB level of the high- and low-risk groups and found that the TMB level of the high-risk group was higher than the TMB level of the low-risk group and was proportional to the risk score (Figures 12(a) and 12(b)). PCa patients with high TMB levels were more likely to develop BCR (Figure 12(c)). After further dividing the patients into the high- and low-risk groups by TMB level, we noticed that the patients in the high-risk group with high TMB levels had the shortest BCR-free time (Figure 12(d)). We then compared the 20 genes with the highest mutation frequencies in the high- and low-risk groups, showing that these genes were mutated more frequently in the high-risk group, with more significant gene-to-gene coincidence and exclusivity relationships (Figures 12(e) – 12(j)).

3.9. Evaluation the Immune Landscape of PCa. We analyzed the correlation between the signature and the immune cell subtype infiltration, which showed that the signature was positively associated with multiple immune cells including CD8+ T cells, B plasma cells, B memory cells, and B naive cells (Figures 13(a) – 13(g)). Compared with the high-risk group, the abundance of infiltrating CD8+ T cells in the low-risk group was significantly higher. To figure out the relationship between the signature and the expression of immune checkpoint in PCa, we found that the high-risk group was positively correlated with high expression of TIGIT, LAG3, PD-1, and CTLA-4 (Figure 14(a)). The TIDE was applied to evaluate the potential response of ICIs for PCa patients (Figures 14(b) – 14(d)). TIDE value in the high-risk group was significantly lower than that in the low-risk group, demonstrating that the high-risk group deserved a better immunotherapy response and immunotherapy outcome. The time-dependent ROC analysis revealed that the prognostic performance of the signature was significantly higher than that of the newly discovered biomarkers including TIDE and TIS (Figure 14(e)).

3.10. Correlation Analysis between the Signature and Drug Treatments. Endocrine drugs and chemotherapeutic drugs are the conventional options for the nonsurgical treatment of PCa. Therefore, we analyzed the sensitivity of different risk groups to endocrine drugs, which suggested that bicalutamide had a lower IC50 in the low-risk group (Figure 15(a)). Chemotherapy combined with immunotherapy has been shown to have better efficacy than either therapy alone. Our results indicated that patients in the low-risk group were more sensitive to docetaxel. (Figure 15(b)) However, the high-risk group was more sensitive to chemotherapeutic agents such as cisplatin, paclitaxel, doxorubicin, etoposide, and mitomycin C than the low-risk group, implying that patients in the high-risk group were more likely to

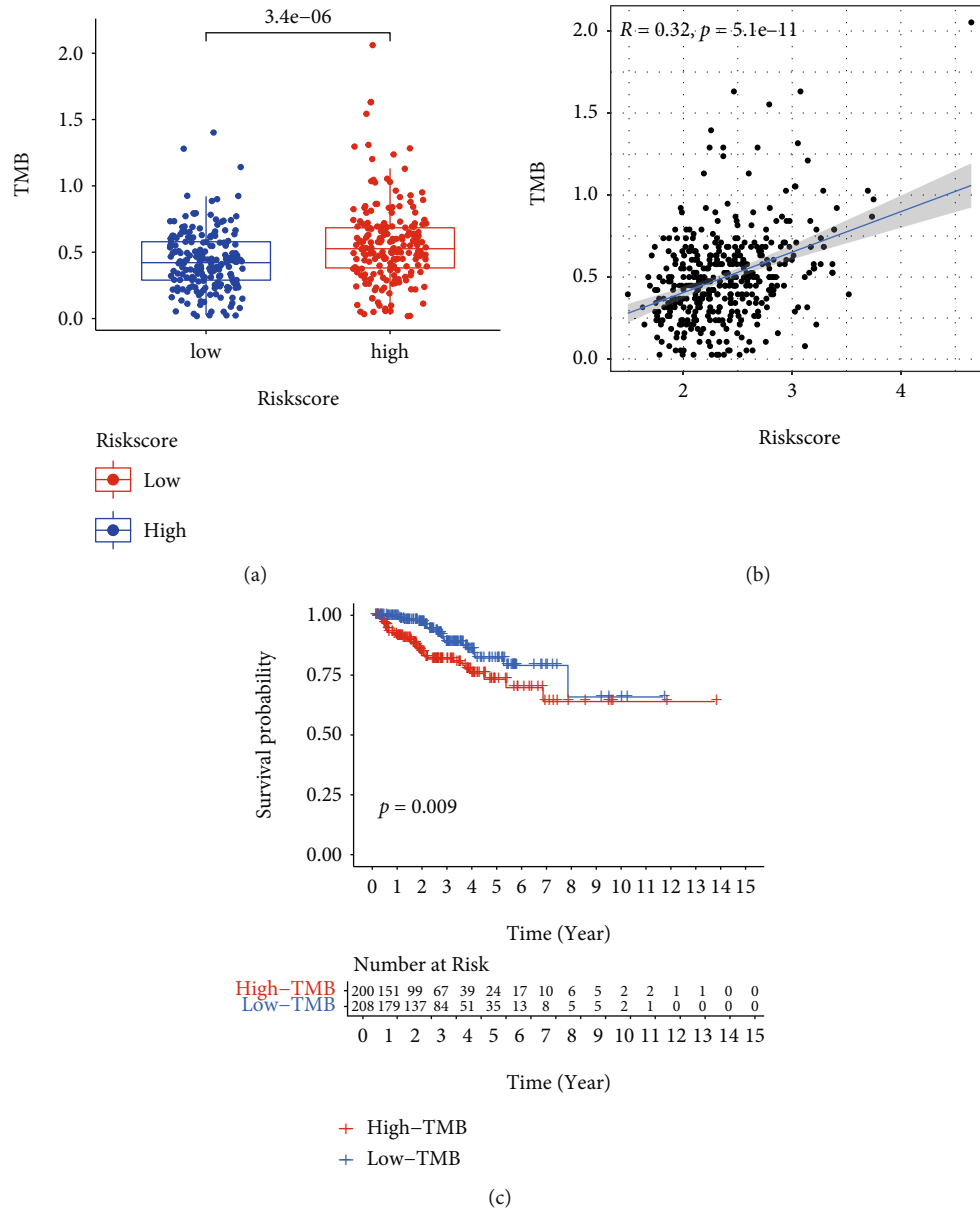
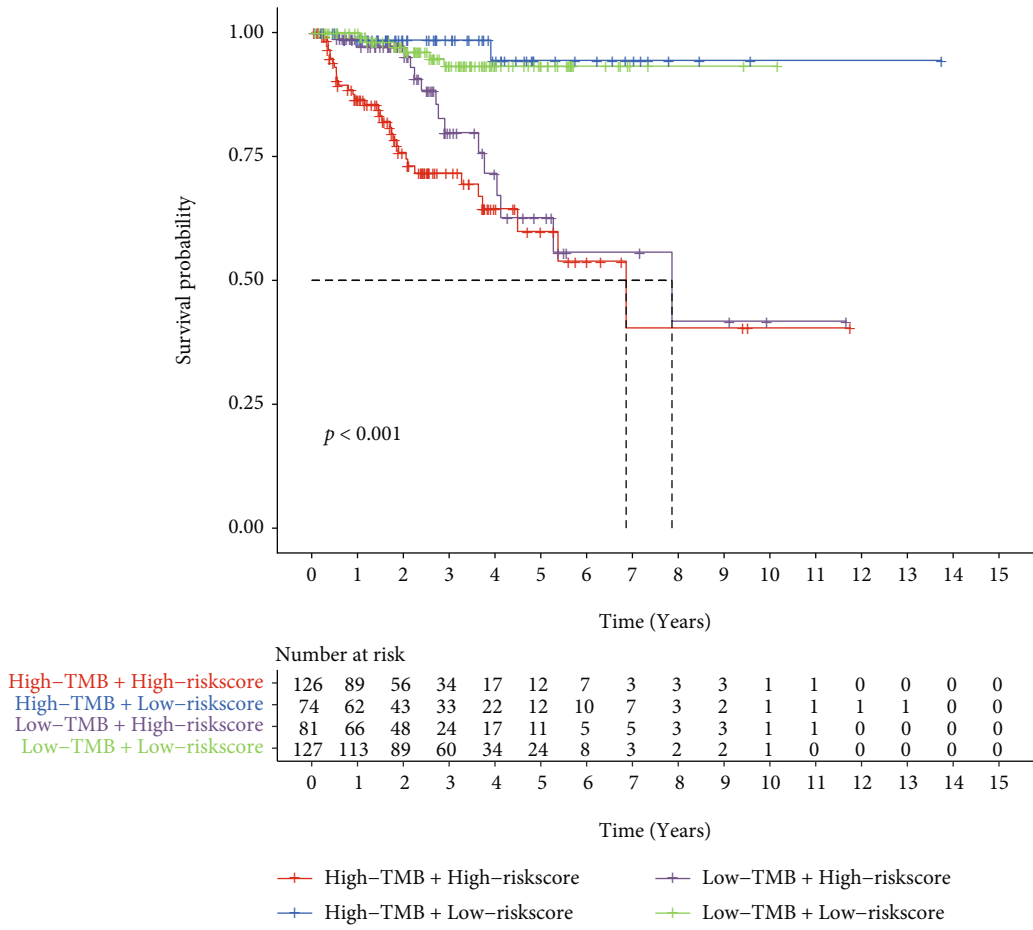
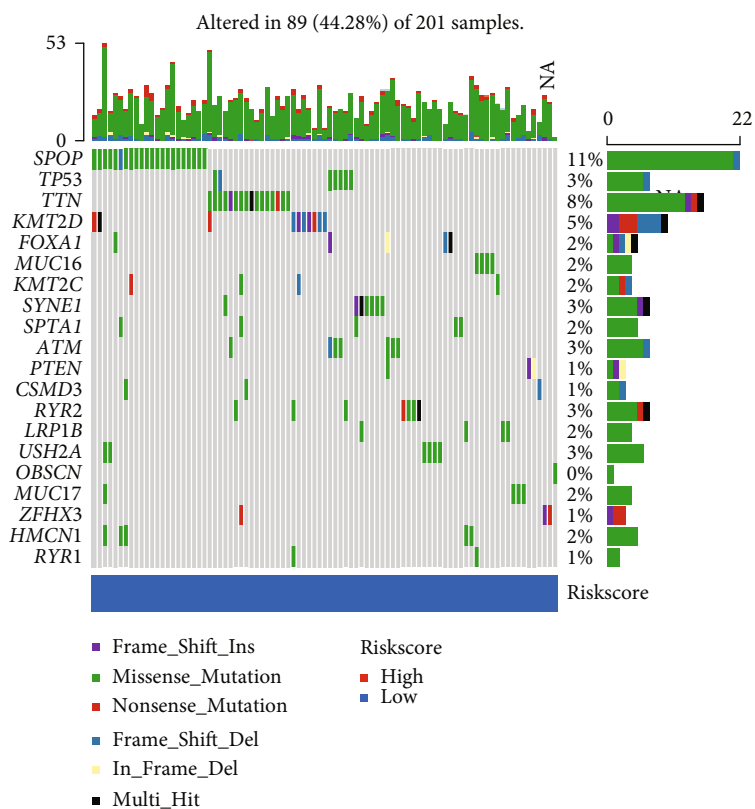


FIGURE 12: Continued.



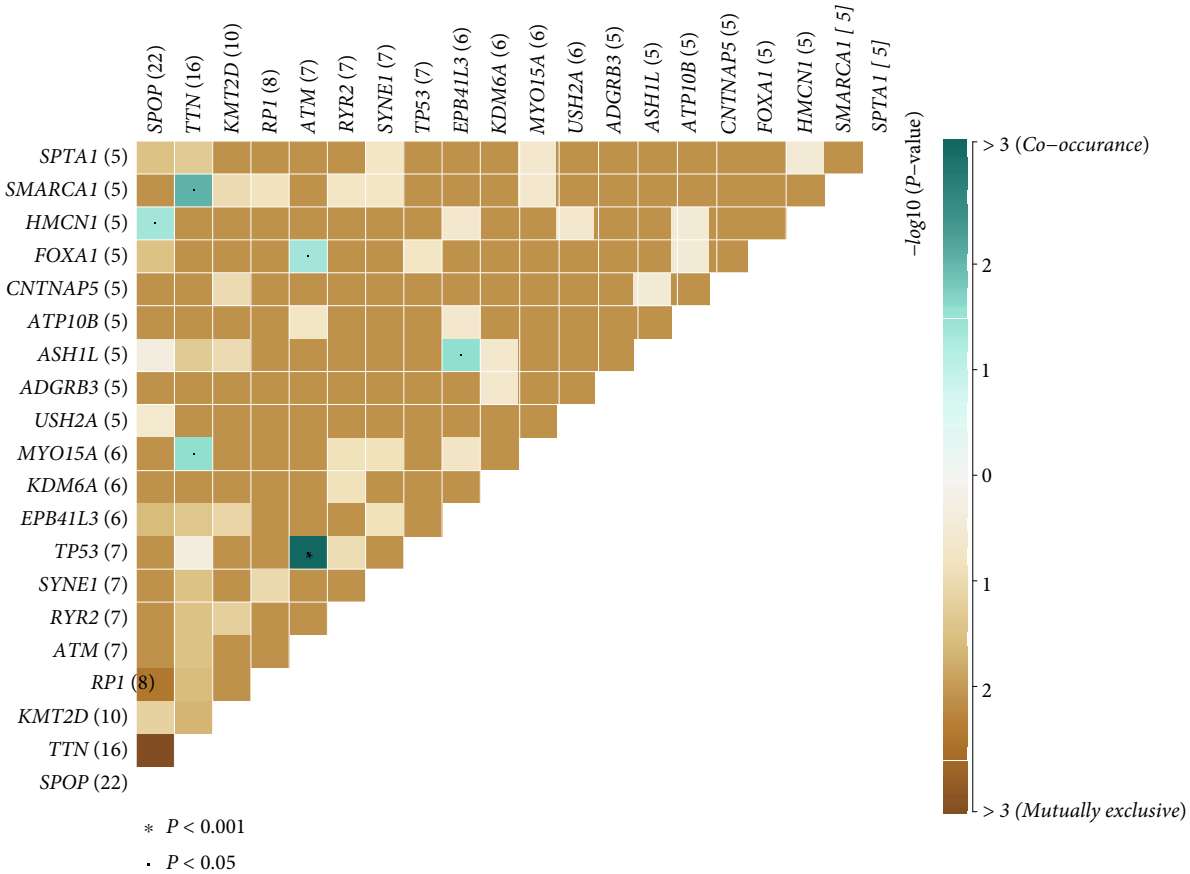
(d)

FIGURE 12: Continued.



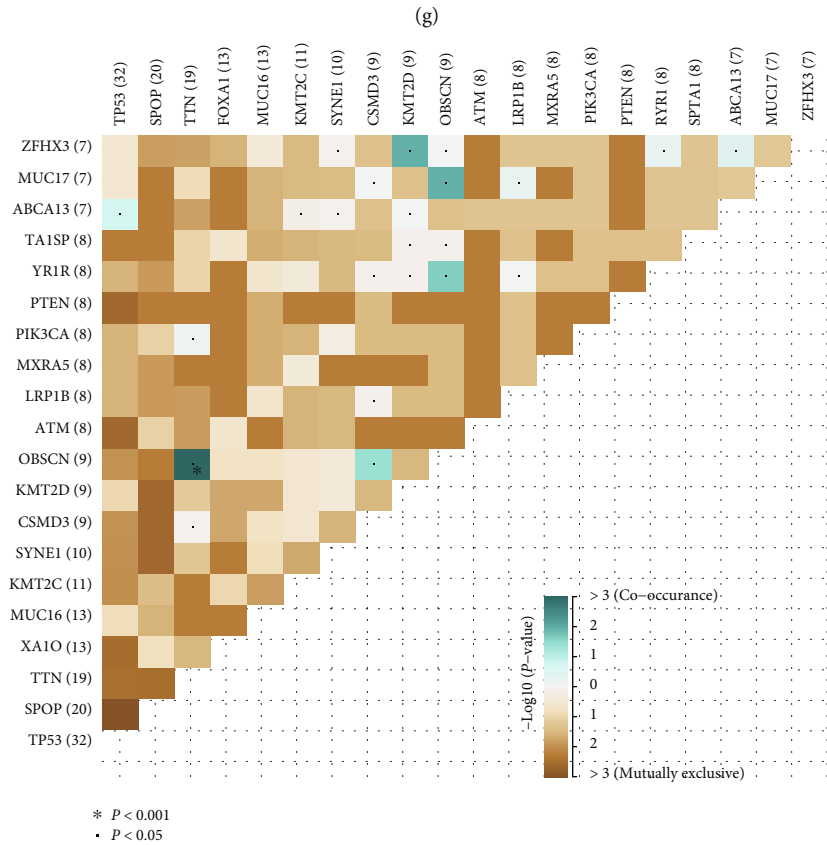
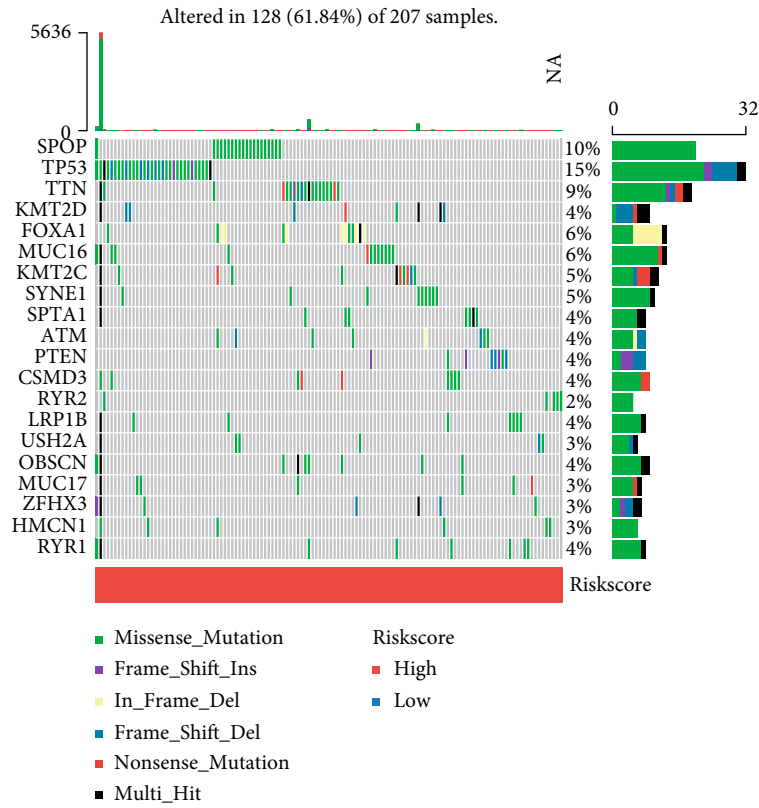
(e)

FIGURE 12: Continued.



(f)

FIGURE 12: Continued.



(h)

FIGURE 12: Continued.

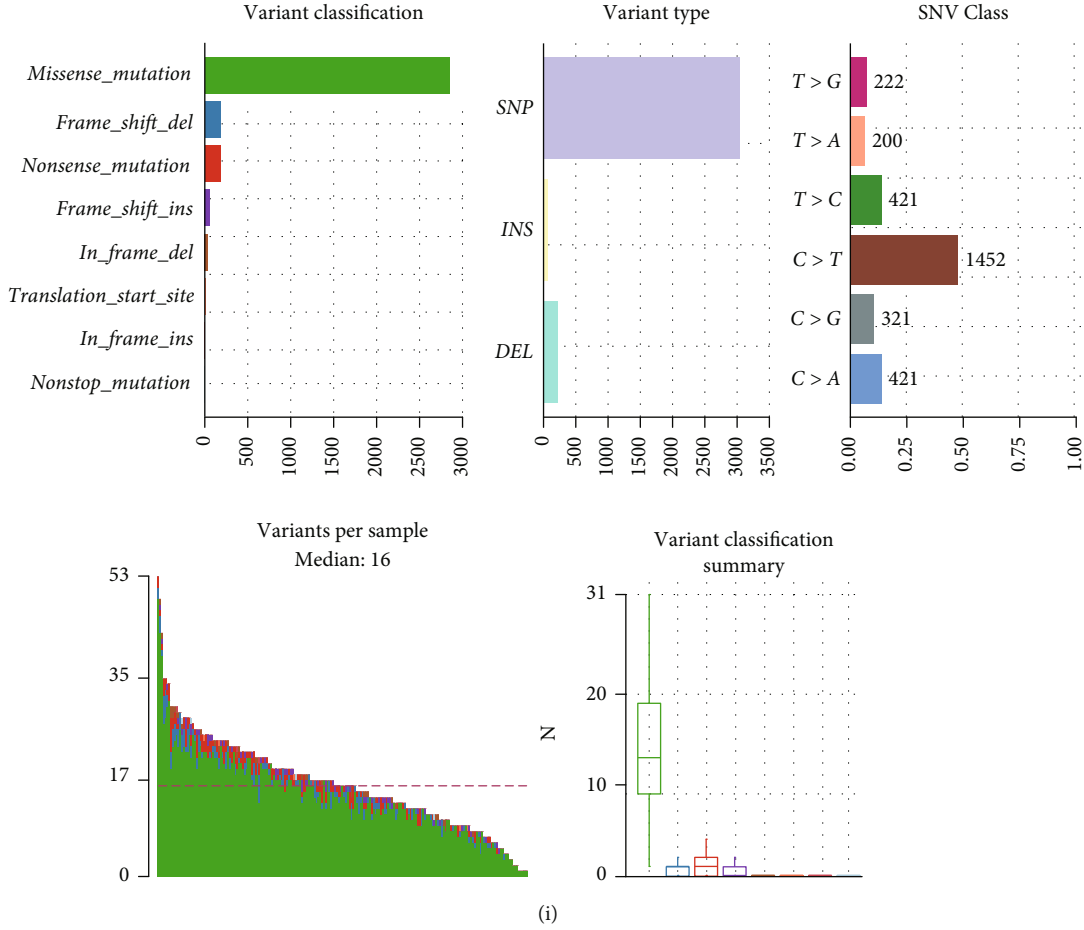


FIGURE 12: Continued.

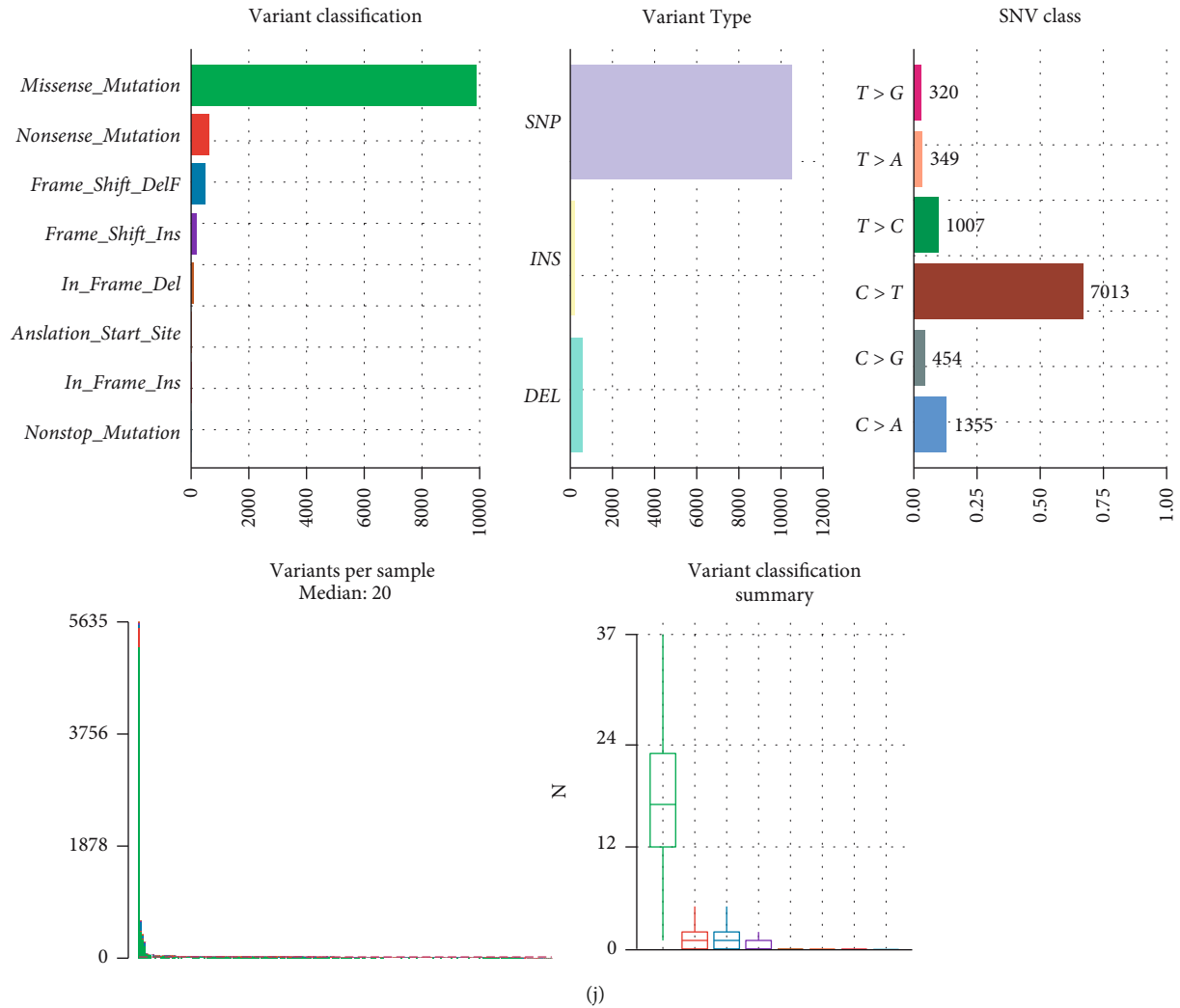


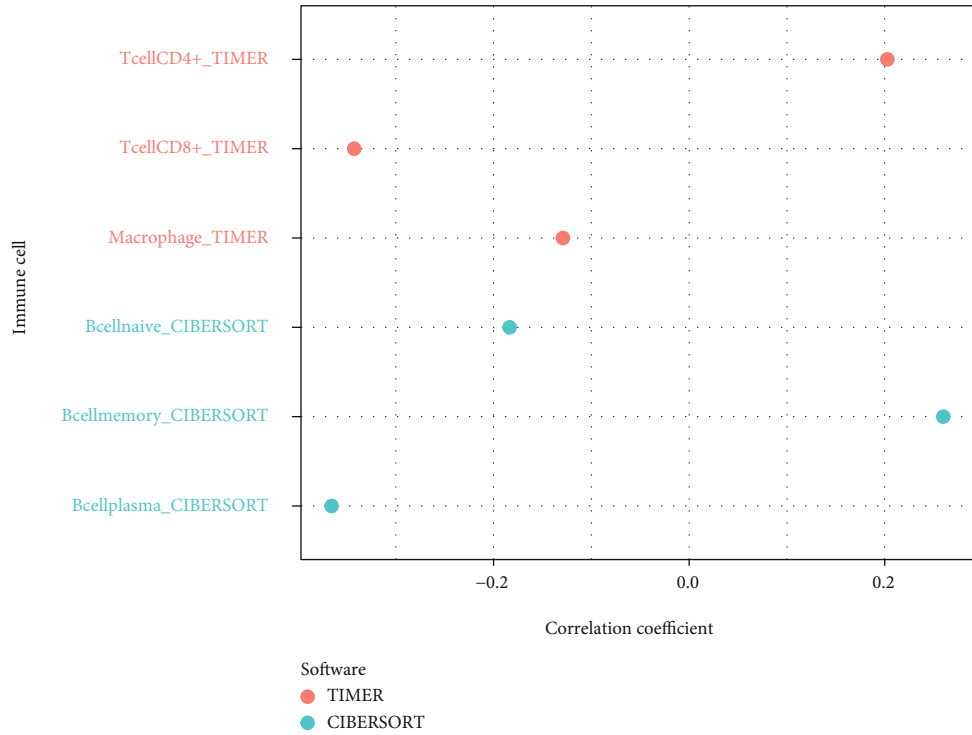
FIGURE 12: TMB analyses between the high- and low-risk groups. (a) TMB levels in the high- and low-risk groups. (b) The relationship between TMB levels and the risk score. (c) The Kaplan-Meier survival curves of patients with high and low TMB levels. (d) The Kaplan-Meier survival curves of patients with different TMB levels and risk groups. (e-j) Detailed mutation information in the two groups.

benefit from these agents (Figures 15(c) – 15(g)). Olaparib, a novel targeted drug, acted to inhibit poly ADP ribose polymerase protein [26]. The high-risk group was more sensitive to olaparib than the low-risk group (Figure 15(h)). Finally, we found that each of the seven genes was also closely related to multiple drugs (Figure 15(i)).

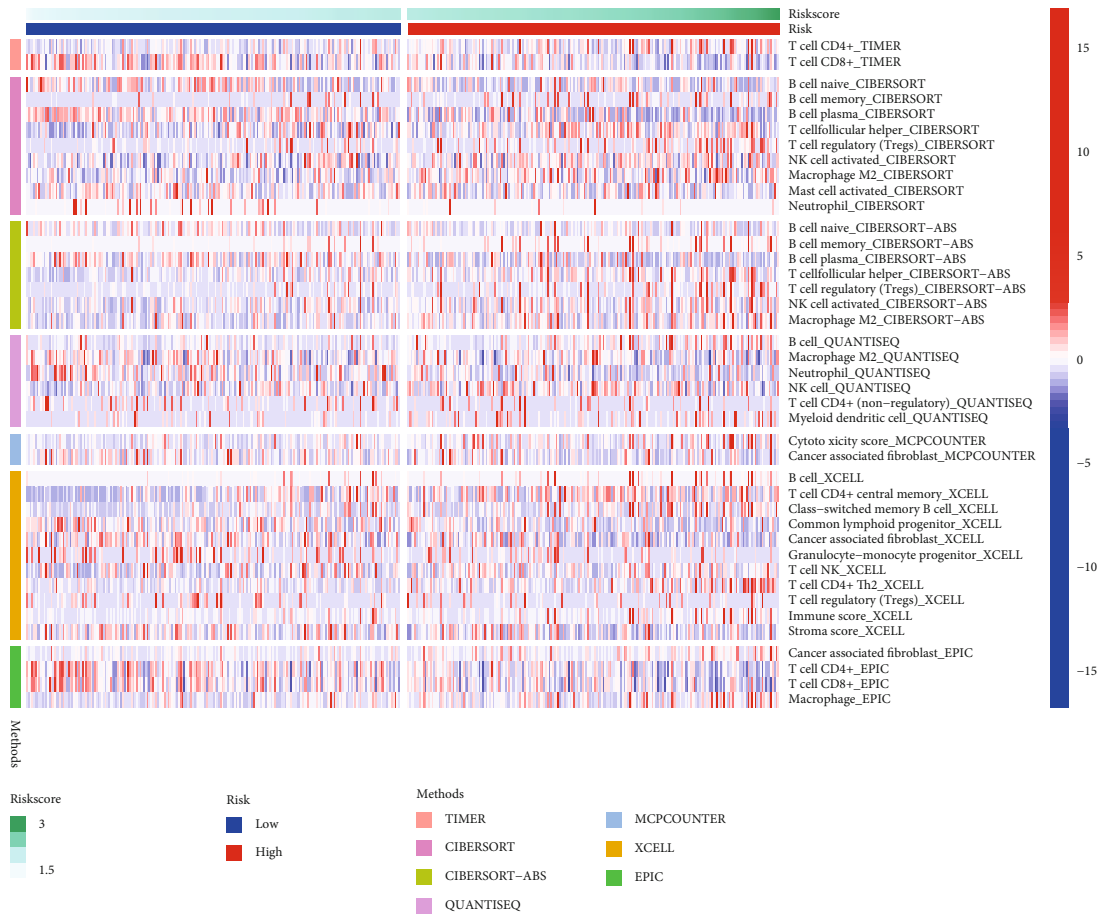
4. Discussion

Treatment strategies for PCa have evolved and progressed tremendously over the past decade yet remained unsatisfactory. More than half of patients with high-risk PCa experienced BCR postoperatively [27]. BCR was a significantly poor prognosis for PCa patients and was strongly associated with progression to metastatic castration-resistant prostate cancer (mCRPC) [28]. Accurately predicting the risk of BCR in PCa patients was essential for the clinical management of PCa and the prognosis of patients. Effective management of PCa could be achieved by precisely stratifying patients at low risk of BCR progression from those at high

risk of BCR progression. Watchful waiting (active surveillance) and curative therapies of patients at different risks of developing BCR could lead to a better prognosis for the patient population in greater need. However, there was currently no feasible way for risk stratification of PCa patients in clinical practice. Thus, this study focuses on a novel type of programmed cell death pyroptosis that played a complex and important role in tumor development and treatment. Normal cells might be transformed into cancer cells by the inflammatory factors released during the process of pyroptosis [29]. Meanwhile, the interaction between pyroptosis and immune cells in TME affected immune defense and antitumor immune function, which in turn had a significant impact on tumor growth, invasion, and metastasis [30]. Providing a novel and comprehensive insight into the relationship between pyroptosis and TME could lead to better identification of PCa and more precise treatments for the patients. As the first report of pyroptosis-related genes in PCa, this study accurately and effectively classified the risk of PCa patients by constructing a signature, which could



(a)



(b)

FIGURE 13: Continued.

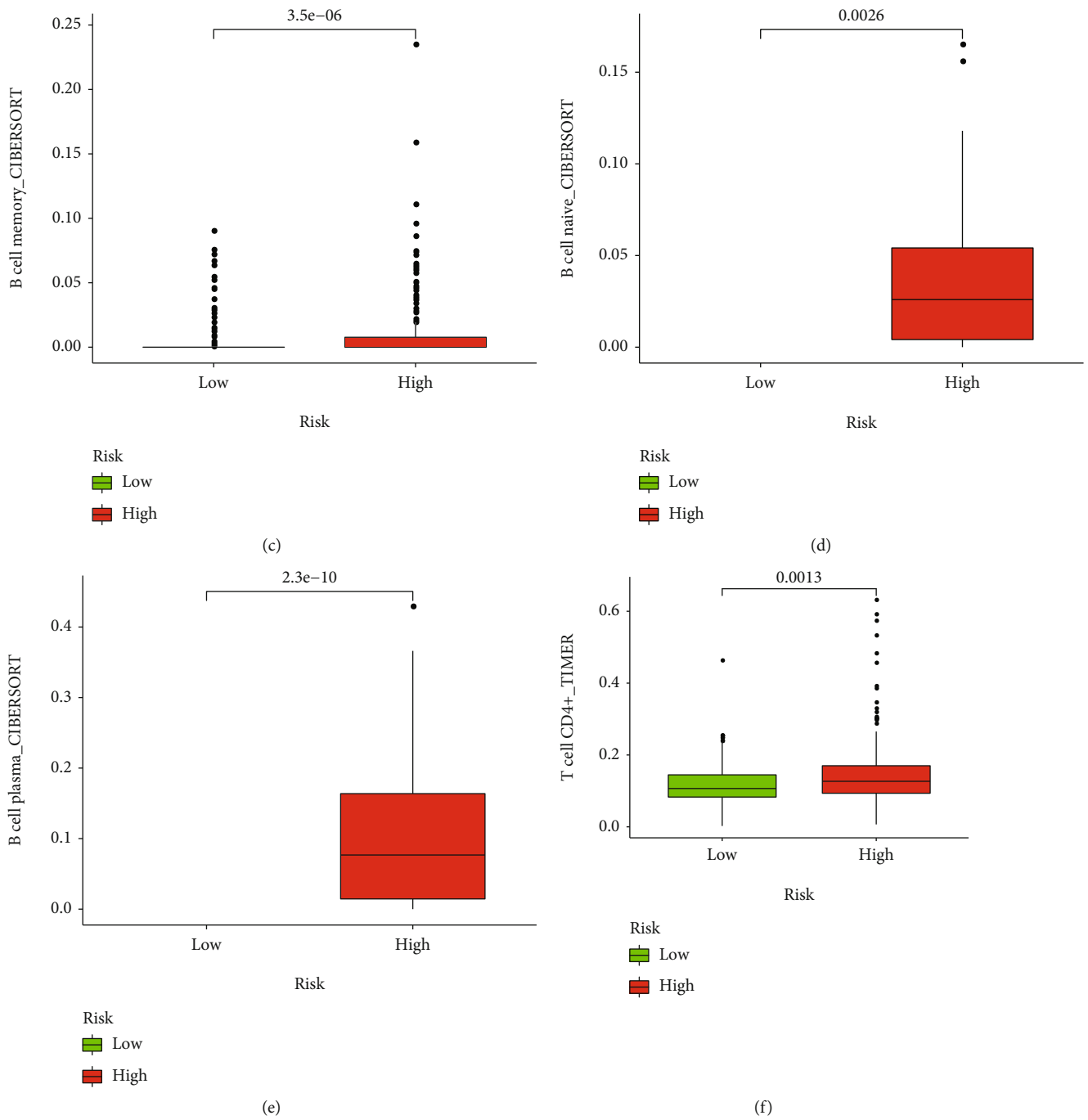


FIGURE 13: Continued.

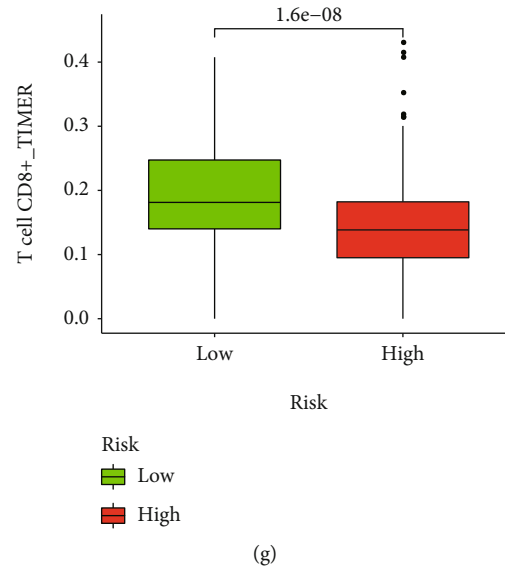


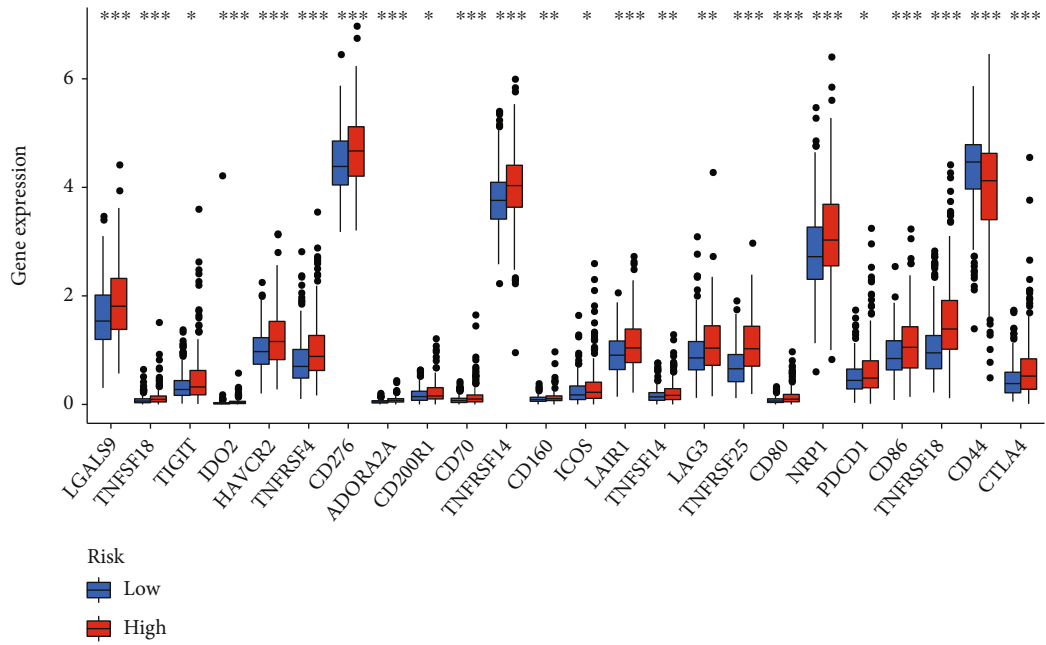
FIGURE 13: Correlation between the signature and the immune infiltration. (a) The difference between the signature and tumor-infiltrating immune cells in multiple algorithms. (b) The distribution of the immune cells in the high- and low-risk groups. The abundance of (c) B memory cells, (d) B naive cells, (e) plasma B cells, (f) CD4+ T cells, and (g) CD8+ T cells in the two groups.

predict the BCR and sensitivity to chemotherapy, endocrine therapy, and immunotherapy for PCa patients at different risk groups. Our signature could provide clinicians with new ideas for managing the risk of BCR in PCa patients and guiding clinical treatment strategies.

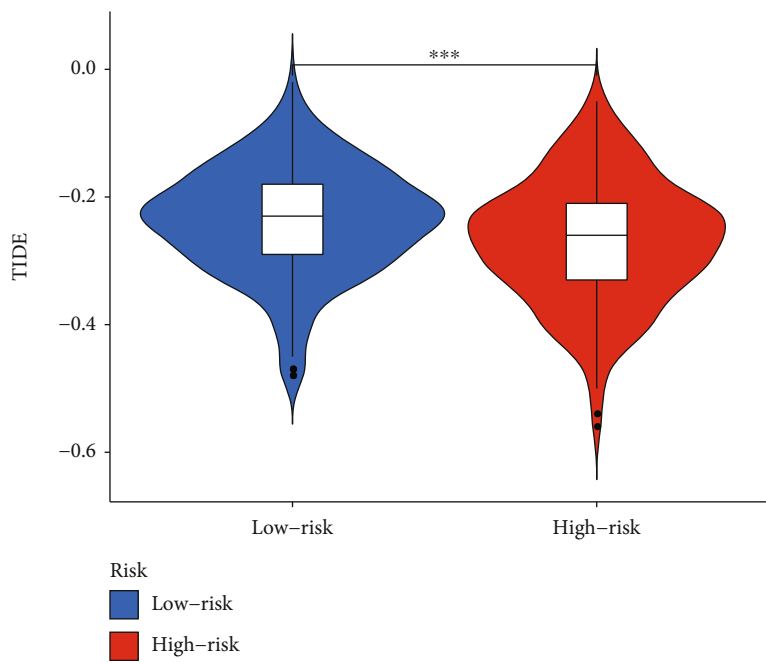
In this study, first, we determined the expression levels of 33 known pyroptosis-related genes in PCa and normal tissues and identified 22 differentially expressed pyroptosis-related genes related to prognosis. Second, sample classification based on predefined gene expression features was a proven method [31]. In order to verify the prognostic value of pyroptosis-related genes, we found that the expression of pyroptosis-related genes occurred differently in patients divided into two groups, resulting in a completely different prognosis. Patients in cluster 2 had higher expression levels of pyroptosis-related genes and a poorer prognosis. Third, a signature composed of 7 genes through Lasso regression analysis was constructed. The independent and powerful ability of the signature to predict the prognosis of PCa patients was verified in the two independent datasets GSE116918 and GSE21034. Fourth, our signature that was closely associated with various stages of PCa could effectively judge the prognosis of patients in different pathological conditions. There were significant differences between the two risk groups in N stage, T stage, and tumor stage and grade, suggesting that our signature was closely related to the existing clinical characteristics. A total of 5 grading groups from grade 1 to grade 5 were proposed based on the Gleason score [32]. Our results found that our signature was closely related to grade, and that grade increased with increasing risk score, indicating that our signature was strongly associated with the existing scoring systems such as Gleason score. Additionally, we then constructed a nomogram that combined our signature and clinical characteristics to predict the 1-, 2-, and 3-year BCR-free survival rates of PCa patients. Fifth, we compared our signature with nine published signatures

constructed for PCa and showed that our signature possesses excellent and accurate prognostic performance superior to the currently established PCa signatures. Overall, our signature had the unexpected predictive ability as well as excellent predictive accuracy to classify PCa patients according to the risk of BCR, which would facilitate clinicians to better treat patients with higher risk.

Chronic inflammation and the associated sustained immune response were thought to contribute to the development and progression of PCa [33]. Pyroptosis was an inflammatory programmed cell death caused by inflammatory caspases and was involved in the inflammatory response to enhance host protective immunity [34]. The tumor microenvironment played a key role in the pathogenesis and disease progression. As the interaction between cancer cells and the tumor microenvironment triggered complex physiological changes that lead to disease severity, cancer metastasis, and resistance to conventional therapies [35]. Q. Wang et al. found that less than 15% pyroptosis of tumor cells could induce the elimination of entire 4T1 tumor grafts in tumor-bearing mice by activating cytotoxic T cells and CD4+ T helper cells in the TME, which was not reproduced in immunodeficient mice [36]. The plasma B cells were considered to be the driving factor of the immune response of PCa, which could improve recurrence-free survival after surgery, and the way that plasma cells participated in the immune system for therapy might be a potential biomarker of the target for therapeutic response to immunotherapy for future prospective evaluation [37]. CD8+ T cells were active antitumor lymphocytes with strong prognostic relevance in many solid tumors [38]. Vicier et al. revealed that low density of CD8+ T cells was influential as an independent poor prognostic marker for BCR and risk of metastatic recurrence in a study of 109 patients with primary PCa [39]. Collectively, it could be seen that the poor prognosis and outcome of PCa were closely related to

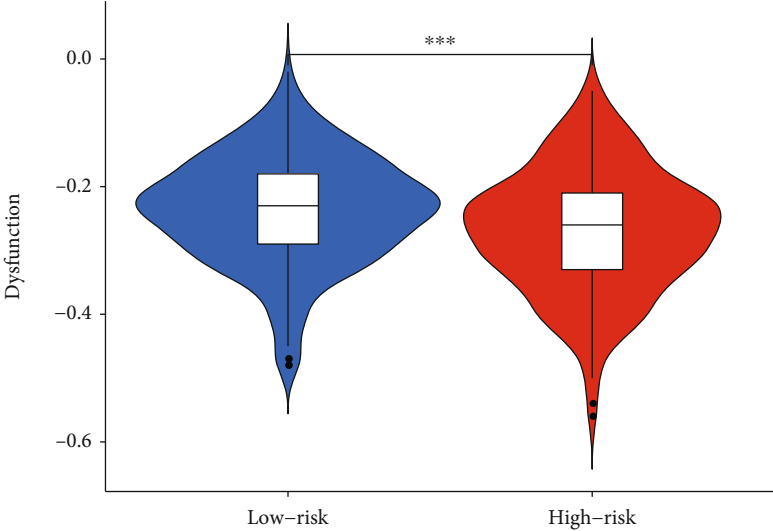


(a)



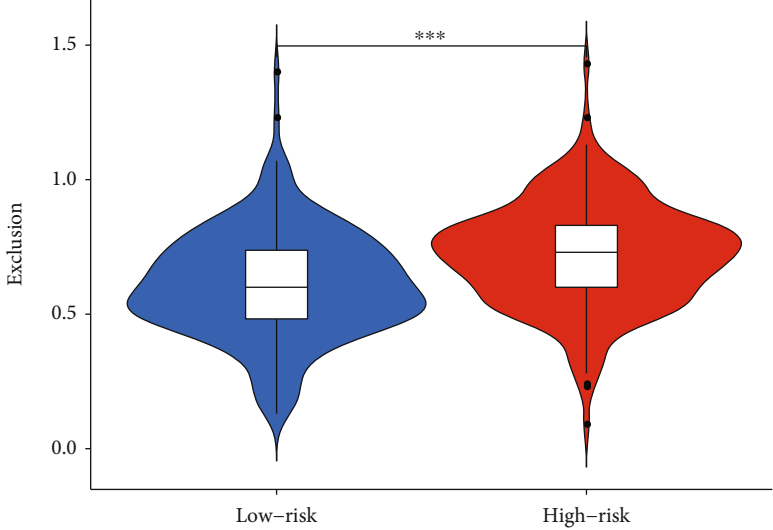
(b)

FIGURE 14: Continued.



Risk
■ Low-risk
■ High-risk

(c)



Risk
■ Low-risk
■ High-risk

(d)

FIGURE 14: Continued.

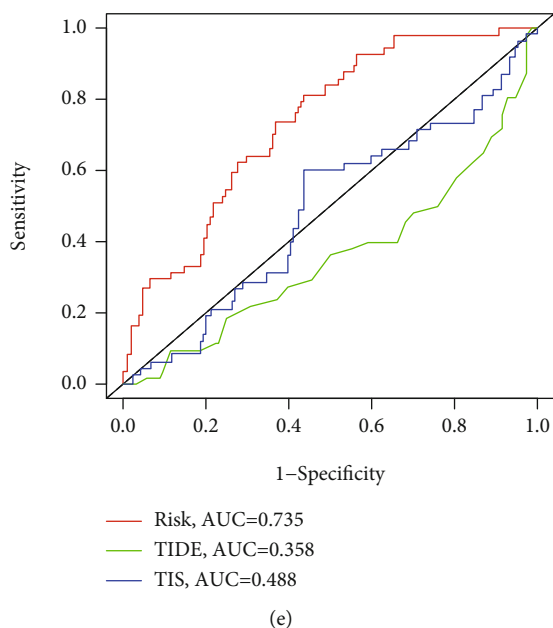


FIGURE 14: Immune function and TIDE analysis. (a) The expression of immune checkpoints in the high- and low-risk groups. (b) TIDE, (c) dysfunction score, (d) T cell exclusion in the high- and low-risk groups. (e) Time-dependent ROC curves analysis of signature, TIDE, and TIS.

immune cell infiltration, which was consistent with our results. As we have discovered, the patients in the high-risk group had a significantly shorter time to BCR, while the high-risk group was negatively associated with the immune cells such as CD8⁺ T cells and plasma B cells. The signature distinguished different groups and thus determined different degrees of immune cell infiltration, leading to different outcomes in PCa. Paying more attention to immune cell infiltration might become a future treatment strategy and further affect the clinical outcome of PCa patients.

One promising PCa treatment method currently under study was immunotherapy, which used the antitumor immune response of the innate immune system to destroy tumorigenesis. ICB therapy targeting CTLA-4, PD-1, and PD-L1 had shown significant therapeutic benefit and become an attractive treatment option for several malignant cancers, such as melanoma, bladder cancer, and lung cancer [40]. It was previously widely believed that PCa did not show a desirable therapeutic response to immunotherapy. However, a small percentage of PCa patients had shown impressive and durable responses to immunotherapy PD-1 inhibition according to the results of KEYNOTE-028 trial [41]. Meanwhile, the immunosuppressive microenvironment of PCa suppressed tumor-specific T cell responses and promoted tumor progression and invasion. A renewed focus on the tumor immune environment was needed to determine prognostic and predictive biomarkers and to guide novel immunotherapies for precise cancer treatment. KEYNOTE-199, the largest ongoing clinical study to date evaluating anti-PD-1 therapy in mCRPC, noted that patients with higher TMB after treatment with pembrolizumab were strongly associated with better prostate-specific antigen (PSA) response and time to PSA progression [42]. Moreover, in the subgroup of patients with mCRPC receiving

docetaxel and endocrine therapy, pembrolizumab demonstrated favorable antitumor activity and disease control, which was durable and encouraging [43]. As seen above, a key challenge in managing PCa was clinical heterogeneity, where patients with the same disease may have different outcomes depending on the tumor microenvironment and whether they were treated with a combination of chemotherapy and endocrine therapy, which was difficult to predict with the available biomarkers. In this study, we tried to provide novel insight to explore the immune landscape and immunotherapy in PCa by our signature. We compared the expression of immune checkpoints in the high- and low-risk groups and found that most immune checkpoints such as PD-1, CTLA-4, LAG3, and TIGIT were more expressed in the high-risk group than in the low-risk group. The previous studies reported that increased expression of PD-1 and PD-L1 was associated with more aggressive PCa [44, 45], which was in line with our findings that patients in the high-risk group were more likely to develop BCR and were associated with high-grade and advanced-stage PCa. Meanwhile, patients with high levels of immune checkpoint gene expression were prone to develop immunosuppressive microenvironment to promote tumor immune escape [46], suggesting that PCa patients in the high-risk group were more likely to benefit from immune checkpoint inhibitor therapy. TMB, TIS, and TIDE were newly identified predictors of immunotherapy [16, 47]. In particular, TIDE had been shown to have better performance than other biomarkers or indicators in predicting immunotherapeutic response [48]. We adopted TIDE to assess the potential clinical efficacy of immunotherapy in the high- and low-risk groups. Higher TIDE represents less likely to benefit from immunotherapy, such as PD-1 and CTLA-4 inhibition therapy. Based on our results, patients in the high-risk group

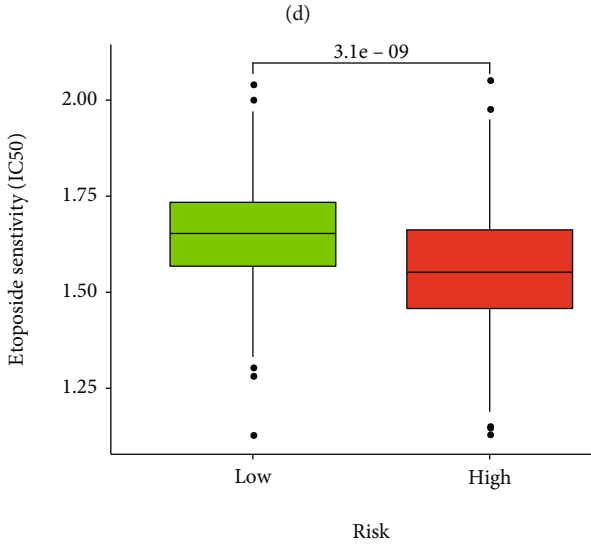
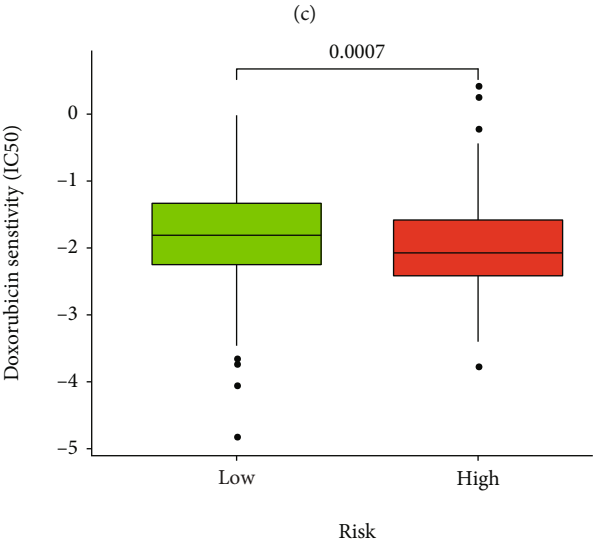
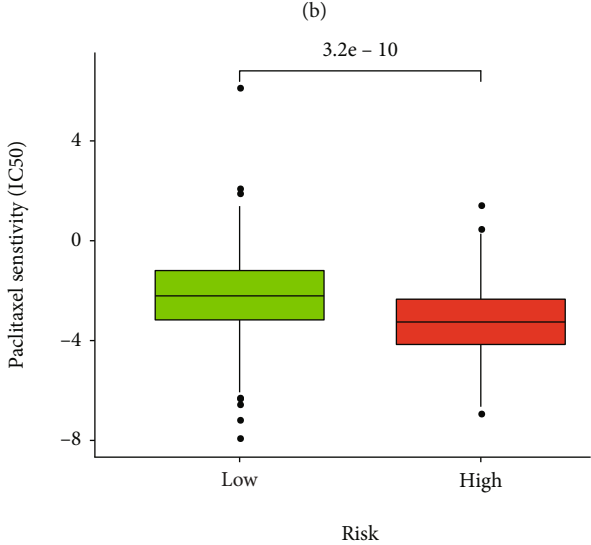
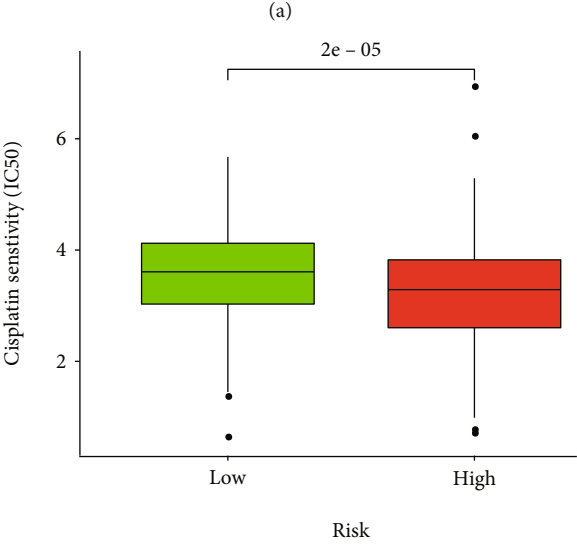
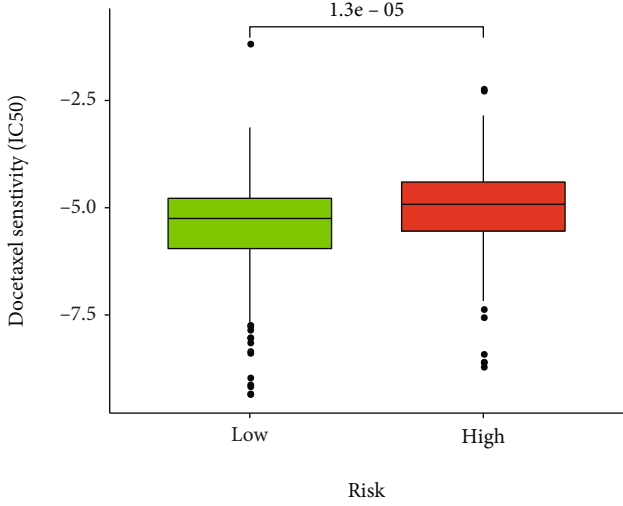
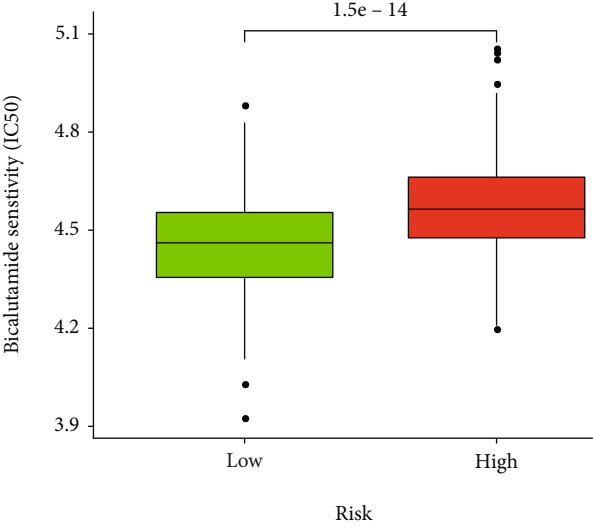
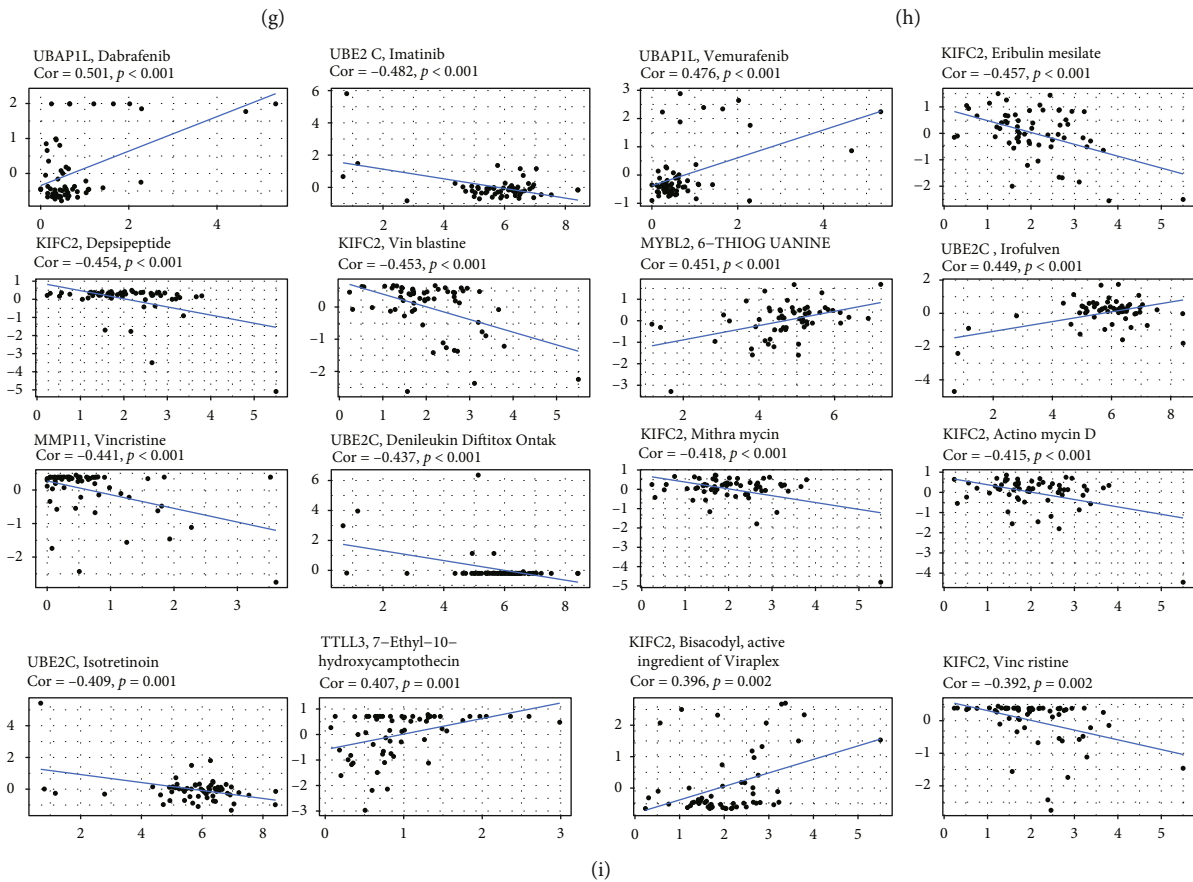
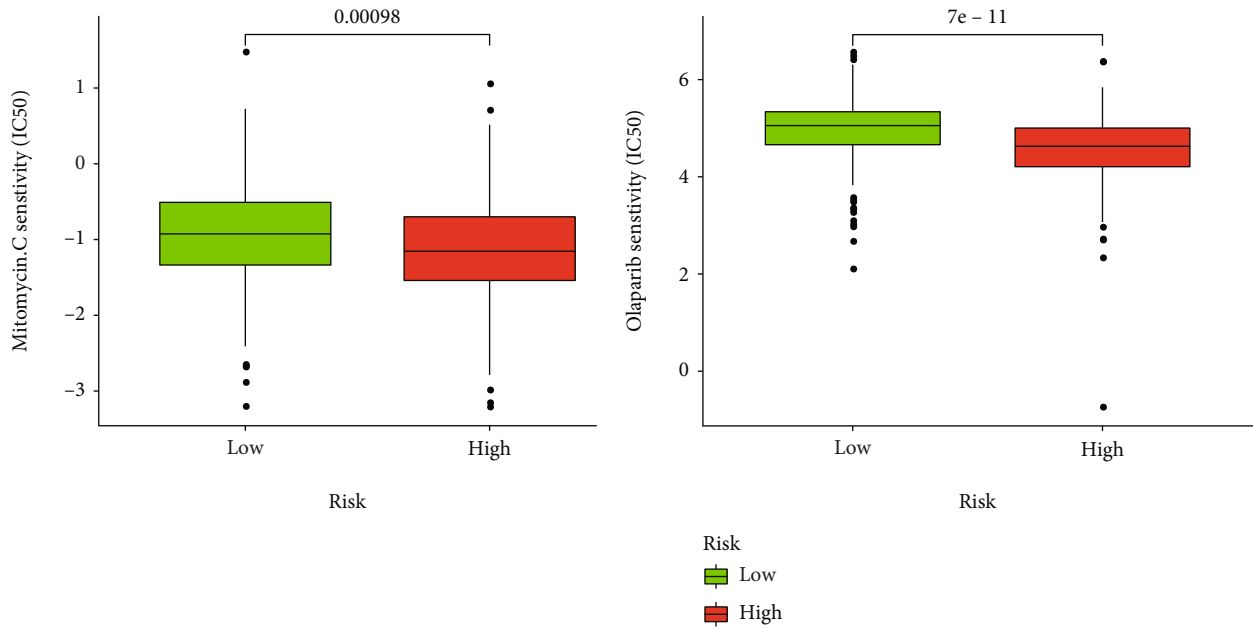


FIGURE 15: Continued.



(i)

FIGURE 15: Assessment of the drug sensitivity. The high- and low-risk groups had significant differences in IC50 of drugs such as (a) bicalutamide, (b) docetaxel, (c) cisplatin, (d) paclitaxel, (e) doxorubicin, (f) etoposide, (g) mitomycin C, and (h) olaparib. (i) The relation between multiple drugs and 7 genes.

with low TIDE were more suitable for immunotherapy. Our signature sheds new light on the effective identification of subgroups of PCa patients who can benefit from immunotherapy. In addition, by comparing the AUC values of our signature with other biomarkers in time-dependent ROC

analysis, we observed that our signature had better predictive performance and superiority. Therefore, it was suggested that our signature was not only effective as an efficacy predictor to discriminate PCa patients with greater benefit from immunotherapy but also had higher accuracy

and specificity to predict the prognosis than other existing biological indicators. We have proved that our signature could effectively stratify the risk of PCa patients into subgroups that were more suitable for immunotherapy and had the potential as an indicator of immunotherapy response in PCa.

Bicalutamide is a nonsteroidal androgen receptor inhibitor widely used in the endocrine therapy of PCa. A prospective randomized trial demonstrated that the use of bicalutamide significantly reduced the risk of objective disease progression in patients with locally advanced PCa [49]. The sensitivity analysis of bicalutamide in the high- and low-risk groups revealed that the low-risk group had a lower IC50, which meant that patients in the low-risk group had a higher sensitivity for bicalutamide. Chemotherapy is a common treatment for advanced PCa, among which docetaxel is the first choice for chemotherapy in most cases. Combined docetaxel and prednisone was the first-line treatment for mCRPC [50]. Chemotherapy drugs were designed to attack rapidly dividing cells, which include not only cancer cells but also normal cells in the body, and this is where the side effects of chemotherapy arise. The side effects of chemotherapy were determined by the type of drug and the dose and period of taking the drug. Common side effects included hair loss, diarrhea, and infections [51]. However, there was currently no biological indicator for the choice of chemotherapy drugs used in clinical practice. Our results showed that patients in the low-risk group were more sensitive to docetaxel and patients in the high-risk group could benefit more from cisplatin, doxorubicin, etoposide, mitomycin C, and paclitaxel. Subgroups of prostate patients stratified according to the signature had different sensitivities to chemotherapeutic agents. Targeted administration of chemotherapeutic agents based on their sensitivity will not only improve treatment outcomes but also reduce the adverse effects of chemotherapy. In addition, the available clinical trial results indicated that the targeted drug olaparib could bring unexpectedly better results to PCa patients [52]. Our results showed that the high-risk group was more likely to benefit from olaparib. Our signature was a promising and reliable predictor of chemotherapy, endocrine, and targeted therapy in PCa, providing a novel approach to get a better prognosis for patients.

5. Conclusion

In short, we have constructed a pyroptosis-related signature that could serve as an independent prognostic factor for PCa. The role of the signature in the immune landscape and treatments was fully elaborated. It was expected to become a robust and promising signature to guide the treatment of PCa.

Data Availability

All data generated or analyzed during this study are included in this article or are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no potential conflicts of interest.

Authors' Contributions

Weide Zhong and Guian Zhang conceived and designed the project. Guian Zhang, Yong Luo, and Weimin Dong acquired the data. Guian Zhang, Yong Luo, and Weimin Dong analyzed and interpreted the data. Guian Zhang wrote the paper. Guian Zhang and Yong Luo contributed equally to this work.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (82072813 and 81571427) and Guangzhou Municipal Science and Technology Project (201803040001).

Supplementary Materials

The sequences of all the primers. (*Supplementary Materials*)

References

- [1] H. Sung, J. Ferlay, R. L. Siegel et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.
- [2] L. Moris, M. G. Cumberbatch, T. Van den Broeck et al., "Benefits and risks of primary treatments for high-risk localized and locally advanced prostate cancer: an international multidisciplinary systematic review," *European Urology*, vol. 77, no. 5, pp. 614–627, 2020.
- [3] M. J. Roobol and S. V. Carlsson, "Risk stratification in prostate cancer screening," *Nature Reviews Urology*, vol. 10, no. 1, pp. 38–48, 2013.
- [4] D. Bertheloot, E. Latz, and B. S. Franklin, "Necroptosis, pyroptosis and apoptosis: an intricate game of cell death," *Cellular & Molecular Immunology*, vol. 18, no. 5, pp. 1106–1121, 2021.
- [5] Y. Tan, Q. Chen, X. Li et al., "Pyroptosis: a new paradigm of cell death for fighting against cancer," *Journal of Experimental & Clinical Cancer Research*, vol. 40, no. 1, p. 153, 2021.
- [6] X. Liu, Z. Zhang, J. Ruan et al., "Inflammasome-activated gasdermin D causes pyroptosis by forming membrane pores," *Nature*, vol. 535, no. 7610, pp. 153–158, 2016.
- [7] G. Tan, C. Huang, J. Chen, and F. Zhi, "HMGB1 released from GSDME-mediated pyroptotic epithelial cells participates in the tumorigenesis of colitis-associated colorectal cancer through the ERK1/2 pathway," *Journal of Hematology & Oncology*, vol. 13, no. 1, p. 149, 2020.
- [8] D. A. Erkes, W. Cai, I. M. Sanchez et al., "Mutant BRAF and MEK inhibitors regulate the tumor immune microenvironment via pyroptosis," *Cancer Discovery*, vol. 10, no. 2, pp. 254–269, 2020.
- [9] D. F. Quail and J. A. Joyce, "Microenvironmental regulation of tumor progression and metastasis," *Nature Medicine*, vol. 19, no. 11, pp. 1423–1437, 2013.

- [10] M. D. Wellenstein and K. E. de Visser, "Cancer-cell-intrinsic mechanisms shaping the tumor immune landscape," *Immunity*, vol. 48, no. 3, pp. 399–416, 2018.
- [11] S. K. Hsu, C. Y. Li, I. L. Lin et al., "Inflammation-related pyroptosis, a novel programmed cell death pathway, and its crosstalk with immune therapy in cancer treatment," *Theranostics*, vol. 11, no. 18, pp. 8813–8835, 2021.
- [12] R. Tang, J. Xu, B. Zhang et al., "Ferroptosis, necroptosis, and pyroptosis in anticancer immunity," *Journal of Hematology & Oncology*, vol. 13, no. 1, p. 110, 2020.
- [13] Z. Zhang, Y. Zhang, S. Xia et al., "Gasdermin E suppresses tumour growth by activating anti-tumour immunity," *Nature*, vol. 579, no. 7799, pp. 415–420, 2020.
- [14] L. Li, M. Jiang, L. Li et al., "Pyroptosis, a new bridge to tumor immunity," *Cancer Science*, vol. 112, no. 10, pp. 3979–3994, 2021.
- [15] Y. Ye, Q. Dai, and H. Qi, "A novel defined pyroptosis-related gene signature for predicting the prognosis of ovarian cancer," *Cell Death Discov*, vol. 7, no. 1, p. 71, 2021.
- [16] M. Ayers, J. Lunceford, M. Nebozhyn et al., "IFN- γ -related mRNA profile predicts clinical response to PD-1 blockade," *Journal of Clinical Investigation*, vol. 127, no. 8, pp. 2930–2940, 2017.
- [17] J. Luan, Q. Zhang, L. Song et al., "Identification and validation of a six immune-related gene signature for prediction of biochemical recurrence in localized prostate cancer following radical prostatectomy," *Transl Androl Urol*, vol. 10, no. 3, pp. 1018–1029, 2021.
- [18] N. Shao, H. Tang, Y. Mi, Y. Zhu, F. Wan, and D. Ye, "A novel gene signature to predict immune infiltration and outcome in patients with prostate cancer," *OncImmunity*, vol. 9, no. 1, p. 1762473, 2020.
- [19] G. Long, W. Ouyang, Y. Zhang et al., "Identification of a DNA repair gene signature and establishment of a prognostic nomogram predicting biochemical-recurrence-free survival of prostate cancer," *Frontiers in Molecular Biosciences*, vol. 8, article 608369, 2021.
- [20] Q. Zhang, K. Zhao, L. Song et al., "A novel apoptosis-related gene signature predicts biochemical recurrence of localized prostate cancer after radical prostatectomy," *Frontiers in Genetics*, vol. 11, article 586376, 2020.
- [21] R. Shi, X. Bao, J. Weischenfeldt et al., "A novel gene signature-based model predicts biochemical recurrence-free survival in prostate cancer patients after radical prostatectomy," *Cancers*, vol. 12, no. 1, 2019.
- [22] L. Gao, J. Meng, Y. Zhang et al., "Development and validation of a six-RNA binding proteins prognostic signature and candidate drugs for prostate cancer," *Genomics*, vol. 112, no. 6, pp. 4980–4992, 2020.
- [23] P. Yuan, L. Ling, Q. Fan et al., "A four-gene signature associated with clinical features can better predict prognosis in prostate cancer," *Cancer Medicine*, vol. 9, no. 21, pp. 8202–8215, 2020.
- [24] B. Liu, X. Li, J. Li, H. Jin, H. Jia, and X. Ge, "Construction and validation of a robust cancer stem cell-associated gene set-based signature to predict early biochemical recurrence in prostate cancer," *Disease Markers*, vol. 2020, Article ID 8860788, 8 pages, 2020.
- [25] J. C. Luan, Q. J. Zhang, K. Zhao et al., "A novel set of immune-associated gene signature predicts biochemical recurrence in localized prostate cancer patients after radical prostatectomy," *Journal of Cancer*, vol. 12, no. 12, pp. 3715–3725, 2021.
- [26] T. Helleday, "PARP inhibitor receives FDA breakthrough therapy designation in castration resistant prostate cancer: beyond germline BRCA mutations," *Annals of Oncology*, vol. 27, no. 5, pp. 755–757, 2016.
- [27] T. Wiegel, D. Bartkowiak, D. Bottke et al., "Adjuvant radiotherapy versus wait-and-see after radical prostatectomy: 10-year follow-up of the ARO 96-02/AUO AP 09/95 trial," *European Urology*, vol. 66, no. 2, pp. 243–250, 2014.
- [28] S. A. Boorjian, R. H. Thompson, M. K. Tollefson et al., "Long-term risk of clinical progression after biochemical recurrence following radical prostatectomy: the impact of time from surgery to recurrence," *European Urology*, vol. 59, no. 6, pp. 893–899, 2011.
- [29] R. Karki and T. D. Kanneganti, "Diverging inflammasome signals in tumorigenesis and potential targeting," *Nature Reviews Cancer*, vol. 19, no. 4, pp. 197–214, 2019.
- [30] R. Loveless, R. Bloomquist, and Y. Teng, "Pyroptosis at the forefront of anticancer immunity," *Journal of Experimental & Clinical Cancer Research*, vol. 40, no. 1, p. 264, 2021.
- [31] R. Cristescu, J. Lee, M. Nebozhyn et al., "Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes," *Nature Medicine*, vol. 21, no. 5, pp. 449–456, 2015.
- [32] J. I. Epstein, L. Egevad, M. B. Amin et al., "The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system," *American Journal of Surgical Pathology*, vol. 40, no. 2, pp. 244–252, 2016.
- [33] S. Jossan, Y. Matsuoka, L. W. Chung, H. E. Zhau, and R. Wang, "Tumor-stroma co-evolution in prostate cancer progression and metastasis," *Seminars in Cell & Developmental Biology*, vol. 21, no. 1, pp. 26–32, 2010.
- [34] S. M. Man, R. Karki, and T. D. Kanneganti, "Molecular mechanisms and functions of pyroptosis, inflammatory caspases and inflammasomes in infectious diseases," *Immunological Reviews*, vol. 277, no. 1, pp. 61–75, 2017.
- [35] S. L. Shiao, G. C. Chu, and L. W. Chung, "Regulation of prostate cancer progression by the tumor microenvironment," *Cancer Letters*, vol. 380, no. 1, pp. 340–348, 2016.
- [36] Q. Wang, Y. Wang, J. Ding et al., "A bioorthogonal system reveals antitumour immune function of pyroptosis," *Nature*, vol. 579, no. 7799, pp. 421–426, 2020.
- [37] A. B. Weiner, T. Vidotto, Y. Liu et al., "Plasma cells are enriched in localized prostate cancer in black men and are associated with improved outcomes," *Nature Communications*, vol. 12, no. 1, p. 935, 2021.
- [38] B. Farhood, M. Najafi, and K. Mortezaee, "CD8+ cytotoxic T lymphocytes in cancer immunotherapy: A review," *Journal of Cellular Physiology*, vol. 234, no. 6, pp. 8509–8521, 2019.
- [39] C. Vicier, L. Werner, Y. Huang et al., "Immune infiltrate with CD8 low or PDL1 high associated with metastatic prostate cancer after radical prostatectomy (RP)," *Journal of Clinical Oncology*, vol. 37, 7_suppl, p. 86, 2019.
- [40] A. D. Waldman, J. M. Fritz, and M. J. Lenardo, "A guide to cancer immunotherapy: from T cell basic science to clinical practice," *Nature Reviews Immunology*, vol. 20, no. 11, pp. 651–668, 2020.
- [41] A. R. Hansen, C. Massard, P. A. Ott et al., "Pembrolizumab for advanced prostate adenocarcinoma: findings of the KEYNOTE-028 study," *Annals of Oncology*, vol. 29, no. 8, pp. 1807–1813, 2018.

- [42] E. S. Antonarakis, J. M. M. Piulats Rodriguez, M. Gross-Goupil et al., “Biomarker analysis from the KEYNOTE-199 trial of pembrolizumab in patients (pts) with docetaxel-refractory metastatic castration-resistant prostate cancer (mCRPC),” *Journal of Clinical Oncology*, vol. 38, 15_suppl, p. 5526, 2020.
- [43] E. S. Antonarakis, J. M. Piulats, M. Gross-Goupil et al., “Pembrolizumab for treatment-refractory metastatic castration-resistant prostate cancer: multicohort, open-label phase II KEYNOTE-199 study,” *Journal of Clinical Oncology*, vol. 38, no. 5, pp. 395–405, 2020.
- [44] N. Ness, S. Andersen, M. R. Khanehkenari et al., “The prognostic role of immune checkpoint markers programmed cell death protein 1 (PD-1) and programmed death ligand 1 (PD-L1) in a large, multicenter prostate cancer cohort,” *Oncotarget*, vol. 8, no. 16, pp. 26789–26801, 2017.
- [45] H. Gevensleben, D. Dietrich, C. Golletz et al., “The immune checkpoint regulator PD-L1 is highly expressed in aggressive primary prostate cancer,” *Clinical Cancer Research*, vol. 22, no. 8, pp. 1969–1977, 2016.
- [46] G. P. Dunn, A. T. Bruce, H. Ikeda, L. J. Old, and R. D. Schreiber, “Cancer immunoediting: from immunosurveillance to tumor escape,” *Nature Immunology*, vol. 3, no. 11, pp. 991–998, 2002.
- [47] J. J. Havel, D. Chowell, and T. A. Chan, “The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy,” *Nature Reviews Cancer*, vol. 19, no. 3, pp. 133–150, 2019.
- [48] P. Jiang, S. Gu, D. Pan et al., “Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response,” *Nature Medicine*, vol. 24, no. 10, pp. 1550–1558, 2018.
- [49] P. F. Schellhammer, “An evaluation of bicalutamide in the treatment of prostate cancer,” *Expert Opinion on Pharmacotherapy*, vol. 3, no. 9, pp. 1313–1328, 2002.
- [50] P. Cornford, R. van den Bergh, E. Briers et al., “EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer. Part II-2020 Update: Treatment of Relapsing and Metastatic Prostate Cancer,” *European Urology*, vol. 79, no. 2, pp. 263–282, 2021.
- [51] K. Nurgali, R. T. Jagoe, and R. Abalo, “Editorial: adverse effects of cancer chemotherapy: anything new to improve tolerance and reduce sequelae?,” *Frontiers in Pharmacology*, vol. 9, p. 245, 2018.
- [52] J. de Bono, J. Mateo, K. Fizazi et al., “Olaparib for metastatic castration-resistant prostate cancer,” *The New England Journal of Medicine*, vol. 382, no. 22, pp. 2091–2102, 2020.

Research Article

Identification of Nine mRNA Signatures for Sepsis Using Random Forest

Jing Zhou ^{1,2}, Siqing Dong ³, Ping Wang ⁴, Xi Su ⁵, and Liang Cheng ⁴

¹Intensive Care Unit, The Second Affiliated Hospital of Harbin Medical University, Harbin 150081, China

²Genomics Research Center, Harbin Medical University, Harbin 150081, China

³Beidahuang Industry Group General Hospital, Harbin 150001, China

⁴College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

⁵Foshan Maternity & Child Healthcare Hospital, Southern Medical University, Foshan 528000, China

Correspondence should be addressed to Xi Su; xisu_fsfy@163.com and Liang Cheng; liangcheng@hrbmu.edu.cn

Received 9 February 2022; Accepted 28 February 2022; Published 19 March 2022

Academic Editor: Hui Ding

Copyright © 2022 Jing Zhou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sepsis has high fatality rates. Early diagnosis could increase its curating rates. There were no reliable molecular biomarkers to distinguish between infected and uninfected patients currently, which limit the treatment of sepsis. To this end, we analyzed gene expression datasets from the GEO database to identify its mRNA signature. First, two gene expression datasets (GSE154918 and GSE131761) were downloaded to identify the differentially expressed genes (DEGs) using Limma package. Totally 384 common DEGs were found in three contrast groups. We found that as the condition worsens, more genes were under disorder condition. Then, random forest model was performed with expression matrix of all genes as feature and disease state as label. After which 279 genes were left. We further analyzed the functions of 279 important DEGs, and their potential biological roles mainly focused on neutrophil thrashing, neutrophil activation involved in immune response, neutrophil-mediated immunity, RAGE receptor binding, long-chain fatty acid binding, specific granule, tertiary granule, and secretory granule lumen. Finally, the top nine mRNAs (MCEMP1, PSTPIP2, CD177, GCA, NDUFAF1, CLIC1, UFD1, SEPT9, and UBE2A) associated with sepsis were considered as signatures for distinguishing between sepsis and healthy controls. Based on 5-fold cross-validation and leave-one-out cross-validation, the nine mRNA signature showed very high AUC.

1. Introduction

As a clinical syndrome, sepsis has been accompanied by human society from ancient times to the present [1]. Sepsis and septic shock have high fatality rates and consume a large amount of medical resources. Since the launch of save sepsis in the early 2000s, the treatment outcomes of patients with sepsis have improved. But the case fatality rate for sepsis remains at 25 to 30 percent, and when shock occurs, it can be as high as 40 to 50 percent [2]. After decades of research, there is still no specific treatment for sepsis. The improvement in patient outcomes came primarily from nonspecific interventions, including fluid resuscitation, early application of antibiotics, and elimination of the source of infection ([3] #5; [4] #8478; [5] #8582; [6] #49). An important reason for

this disheartening situation is that the definitions of sepsis and septic shock cover a very heterogeneous population of patients. The causes are so varied that it is difficult to find a common treatment for these conditions.

How to classify patients with sepsis is one of the key areas of research on sepsis and other diseases [7–9], though biomarkers have been the subject of intensive research for decades ([10] #71; [11] #15; [12] #8853; [13] #431; [14] #673; [15] #50). For example, procalcitonin has been included in treatment guidelines [16], but there is currently no reliable biomarker to distinguish between infected and uninfected patients. Only 30–40% of patients with sepsis or septic shock have positive blood cultures. New technologies such as high-throughput technologies (genomics, transcriptomics, etc.) have been used to better identify subsets

of patients with sepsis, to identify patients at high risk of developing sepsis, and to provide the possibility for rapid and accurate diagnosis of infection [17, 18, 19].

This study analyzed microarray dataset from public gene expression database, to obtain differentially expressed genes (DEGs) between sepsis and healthy people, and then, a random forest model was performed on the DEGs to select more important biomarkers. Next, we performed gene functional enrichment analysis on the DEGs selected to analyze the function module of the DEGs and to uncover how the DEGs contribute to sepsis. Our study aims to detect neglected biomarkers of sepsis to better distinguish between sepsis patients and healthy controls.

2. Materials and Methods

2.1. Data Resource. To identify potential gene signatures associated with sepsis, we got two gene expression datasets (GSE154918 and GSE131761) [20, 21] from the GEO database (<https://www.ncbi.nlm.nih.gov/geo/>), GSE154918 dataset as primary research data and GSE131761 as supplementary data.

Totally, 109 samples from GSE154918 dataset were collected from 19 septic shock patients, 20 sepsis patients, 12 uncomplicated infection patients, and 40 healthy volunteers. Supplementary validation dataset GSE131761 was collected from 81 septic shock patients and 15 healthy volunteers. All samples were collected from peripheral blood. The diagnosis of septic shock was according to the Sepsis 3.0 criteria [3].

2.2. Identification of DEGs. The workflow is shown in Figure 1. DEGs were calculated between sepsis samples (uncomplicated infection, sepsis, and septic shock) and healthy control using Limma package [22] with p value < 0.05 and $|\log_{2}FC| > 1$ as threshold.

2.3. Random Forest Model. Random forest model was performed using Python machine learning library Scikit-learn [23], with expression matrix of all genes as feature and disease state as label as other researches [1]. We set 1000 random forest trees and operated 5-fold cross-validation and leave-one-out cross-validation to evaluate the performance of the model. Feature importance was collected from the random forest model after training and assessment, and then, we sorted the features by feature importance and chose top n features to reconstruct random forest, accessing the best combination of gene signatures [24].

2.4. Functional and Pathway Enrichment Analyses. Gene enrichment analysis of DEGs was based on Gene Ontology (GO) database from molecular function, cellular component, and biological process using R package ClusterProfiler [25]; pathways with adjusted p value < 0.05 were selected as significant enriched pathways [26].

3. Results

3.1. Identification of DEGs. Gene expression difference was calculated between three sepsis groups and healthy control,

respectively. 530 differentially expressed genes (DEGs) were found in uncomplicated infection patients compared with healthy control, 727 DEGs were found in sepsis samples, 1414 DEGs were found in sepsis shock samples, and 384 common DEGs were found in the above three contrast groups (Figure 2). We found that as the condition worsens, more genes were under disorder condition.

3.2. Features Selected by Random Forest. Next, we performed random forest [15, 27] to select important genes of the DEGs of the above three contrast groups, with gene expression matrix of DEGs as feature and health state as label; we selected the genes with feature *importance* > 0 as the most important genes. We, respectively, found 440, 657, and 1018 DEGs in uncomplicated infection vs. healthy control, sepsis vs. healthy control, and sepsis shock vs. healthy control, and 279 genes were common among the three (Figure 3).

3.3. Enrichment Analysis of Intersection Important DEGs. We further performed functional analyses for 279 important DEGs to explore the underlying biological roles. Multiple GO-BP terms were associated with neutrophil degranulation, neutrophil activation involved in immune response, and neutrophil-mediated immunity. The DEGs played essential roles in GO-MF terms containing 2 more enriched terms: RAGE receptor binding and long-chain fatty acid binding. The GO-CC revealed that these DEGs were mainly enriched in specific granule, tertiary granule, and secretory granule lumen (Figure 4).

3.4. Biomarkers Distinguishing Disease States. To detect biomarkers to distinguish three disease-state patients and healthy patients, we further performed random forest on three disease states and healthy samples together, sorted the feature importance, and selected the top n (1-50) gene features to reconstruct random forest model to access the best biomarker combination. We found that when 6 features were selected, the accuracy of the model reached 0.895, and the accuracy of model began to decline since more than 9 features were selected. Therefore, the first 9 characteristics were selected as potential biomarkers to predict different sepsis states (Figure 5).

3.5. Supplement Validation. In order to assess the availability of the 9 biomarkers, we used a supplementary validation dataset to build random forest model using these 9 biomarkers as feature. Since the biomarker MCEMP1 was not sequenced because of lacking probe, only 8 biomarkers were sequenced in supplement dataset, so we only evaluated the 8 biomarkers. We found that the model performed good to distinguish sepsis from healthy control in supplement dataset. Then, we further calculated the gene expression difference between sepsis and healthy control samples in supplement dataset, and we found that 6 out of 8 biomarkers were DEGs in supplement dataset (Table 1).

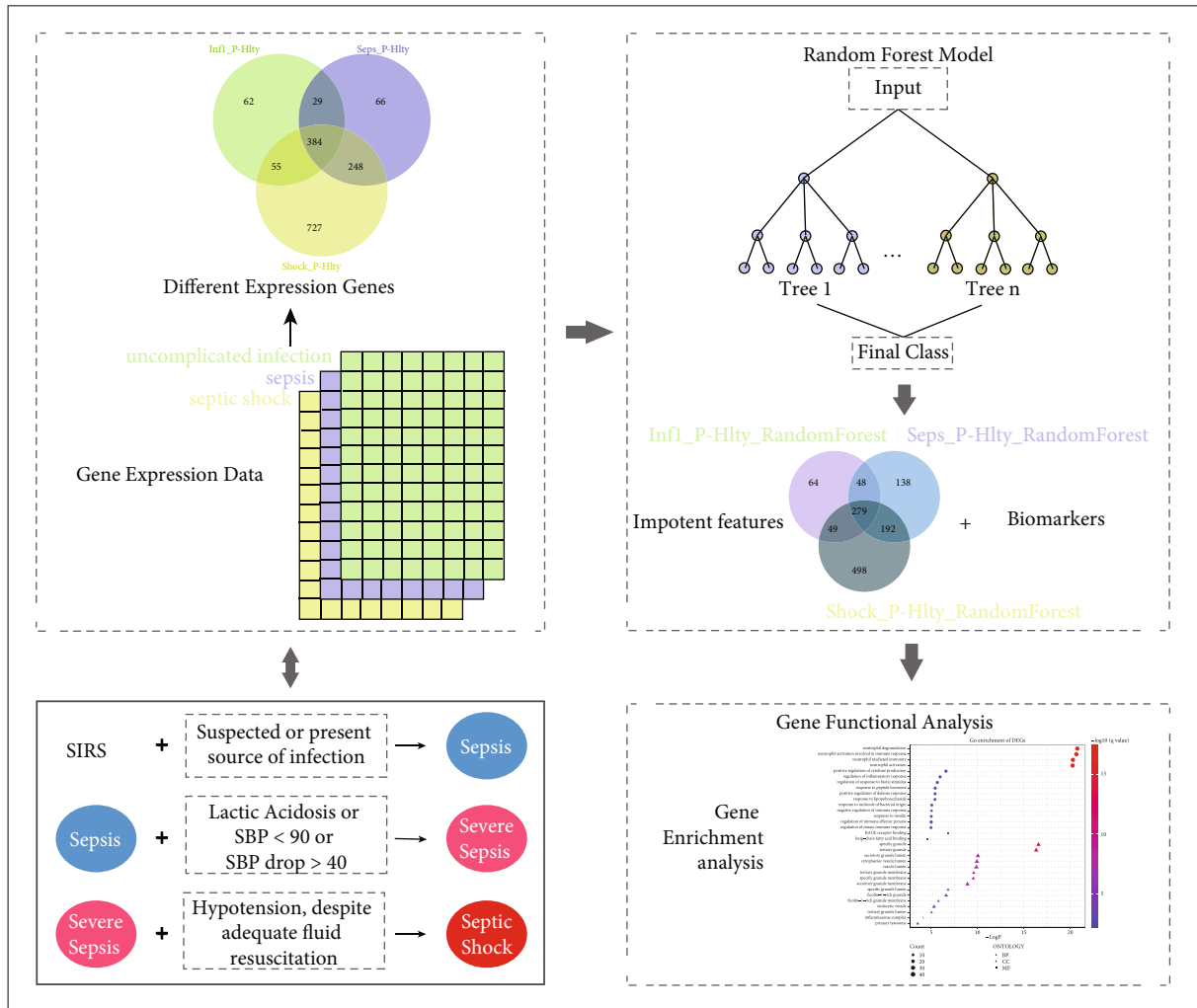


FIGURE 1: The workflow of identifying molecular signature of sepsis. Gene different expression analysis was firstly performed on gene expression data, and then, random forest model was performed with expression of DEGs as feature. Next, the function of selected important DEGs from random forest was analyzed.

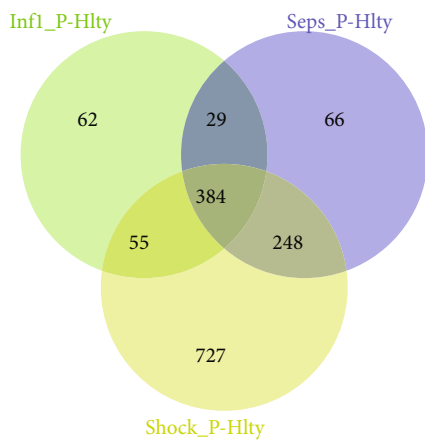


FIGURE 2: DEG numbers of the three contrast groups. The Venn diagram displays the DEG numbers of three contrast groups Infl_P vs. healthy, Sepsis_P vs. healthy, and Shock_P vs. healthy.

4. Discussion

According to the Sepsis 3.0 definition, sepsis is a life-threatening organ dysfunction resulting from an infection-induced host response disorder. Neutrophils are the main immune-cell barrier against pathogens, but they can be a double-edged sword in sepsis because they play a role in both proinflammatory response and anti-inflammatory response. We hypothesize that the immune signature of sepsis can be determined early by the phenotype of neutrophils and distinguish sepsis from noninfectious inflammatory syndromes. It is important to screen for features that are considered important in the biology of sepsis but alone are not distinguishable to clearly distinguish sepsis. Sepsis is thought to be an immune imbalance in which pathogens evade the host’s defense mechanisms and continue to stimulate and destroy host cells. Many of the protective immune mechanisms activated early in the disease become harmful and are associated with excessive inflammation and immunosuppression. The host response of sepsis involves the coexistence of inflammatory and anti-inflammatory

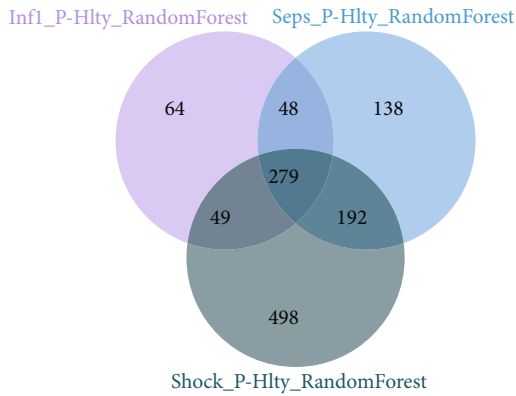


FIGURE 3: Gene feature numbers of the three contrast groups after performing random forest. The Venn diagram displays the gene feature numbers of the three contrast groups Infl_P vs. healthy, Sepsis_P vs. healthy, and Shock_P vs. healthy.

responses, involving different organs, systems, and cell types [28].

We conducted a differential analysis of data from a group of three sepsis severity levels and healthy controls and found that the number of differential genes in sepsis patients increased according to the severity of the disease, suggesting that more genes became dysregulated with the severity of the disease. There were 279 differentially expressed genes in all three kinds of severe infection, which may play roles in the onset and progression of sepsis. Functional enrichment analysis showed that biological processes were most significantly enriched in neutrophil activation, immune activation, inflammatory response, and bacterial response.

In GO-MF analysis, DEGs were significantly enriched in RAGE receptor binding and long-chain fatty acid binding. A meta-analysis showed that RAGE inhibition had a significant advantage in multiple microbial infections. For G^+ bacterial infection, RAGE suppression reduced bacterial growth and transmission, inflammatory cell flow, plasma cytokine levels, and lung damage. This paper concluded that RAGE inhibition had beneficial effects on the outcomes of animal models of sepsis with different causes [29]. There are few studies on long-chain fatty acid binding and sepsis. This article is one of them. Extraenteral pathogenic *E. coli* can cause diseases such as urinary tract infections and sepsis. Mucus is the main nutrient source of *Escherichia coli* in the intestinal tract, and genes directly or indirectly related to the fatty acid oxidation pathway contribute to the adaptation and migration of ExPEC [30].

In our study, the remarkable GO-CC terms are associated with neutrophil degranulation, such as tertiary granule, specific granule, and secretory granule lumen. Neutrophils are one of the most important cells in the host's natural defense. The following are the granules in neutrophil cytoplasm: azurophilic granule, specific granules, gelatinase granules, and secretory vesicles. They all play very important roles. Neutrophil dysregulation is present in sepsis. Many evidence suggest that neutrophil threshing molecules are of

value in the diagnosis and prognosis of sepsis. Monitoring neutrophil function may help identify early sepsis [31].

We used the random forest to select 9 characteristic genes as potential biomarkers for predicting sepsis: MCEMP1, PSTPIP2, CD177, GCA, NDUFAF1, CLIC1, UFD1, SEPT9, and UBE2A. Some of these genes have been confirmed in experiments or have also been widely concerned in bioinformatics studies.

MCEMP1 is involved in the regulation of mast cell differentiation or innate immune response. In our study, MCEMP1 gene expression was increased in sepsis. Chen et al. [32]. established a cecal ligation and puncture-induced sepsis mouse model to determine the expression of mast cell expression membrane protein 1 (MCEMP1). They observed that MCEMP1 was highly expressed in septic mice. Loss of MCEMP1 can promote T lymphocyte and NK cell activity, increase immunoglobulin expression, inhibit the release of inflammatory factors, and reduce T lymphocyte apoptosis. They also found that downregulation of lncRNA NEAT1 could inhibit MCEMP1, thereby promoting the immunosuppression effect of Mir-125 on sepsis mice. This may be a potential therapeutic target for sepsis. Xie et al. found that MALAT1 upregulates MCEMP1 by binding to Mir-23a, thereby promoting inflammatory response in sepsis mice [33].

Proline-serine-threonine-phosphatase-interacting protein 2 (PSTPIP2) belongs to the F-BAR family of proteins and is mainly expressed in macrophages. In recent years, PSTPIP2 has been found to play an important role in congenital immune diseases and acquired immune diseases (AIDS) [34]. Chen et al. [35] studied biomarkers of *Escherichia coli*-induced sepsis. They analyzed 4 microarray datasets from GEO database and identified 54 DEGs. Eight different genes were found between sepsis patients and controls. Furthermore, differential expression of the candidate gene was verified by human blood model in vitro. qPCR results suggested that PSTPIP2 may be closely related to *Escherichia coli*-induced sepsis.

Neutrophils play an important role in the pathophysiology of sepsis and are the primary defense against infection. A transcriptome study was performed on purified neutrophils from patients with septic shock to identify genes that were differentially expressed during the first week of illness compared with healthy controls. The results were confirmed at the protein level. They found that 364 differentially expressed genes were upregulated and 328 downregulated in patients with sepsis. CD177mRNA showed the most significant difference between patients and healthy controls. This is consistent with our findings, which also found that CD177 was significantly upregulated in sepsis patients [36].

Yang and Li [37] applied bioinformatics to study the molecular mechanism of sepsis. Transcriptome data (GSE12624) were downloaded from Gene Expression Omnibus database for protein-protein interaction network analysis. Twenty-four differentially expressed clusters were identified by ANCOVA global test, including 12 clusters in sepsis samples and 12 clusters in nonsepsis samples. 207 biomarker genes were extracted from the first 6 clusters by SVM method, and 10 genes including GCA were considered as potential biomarkers.

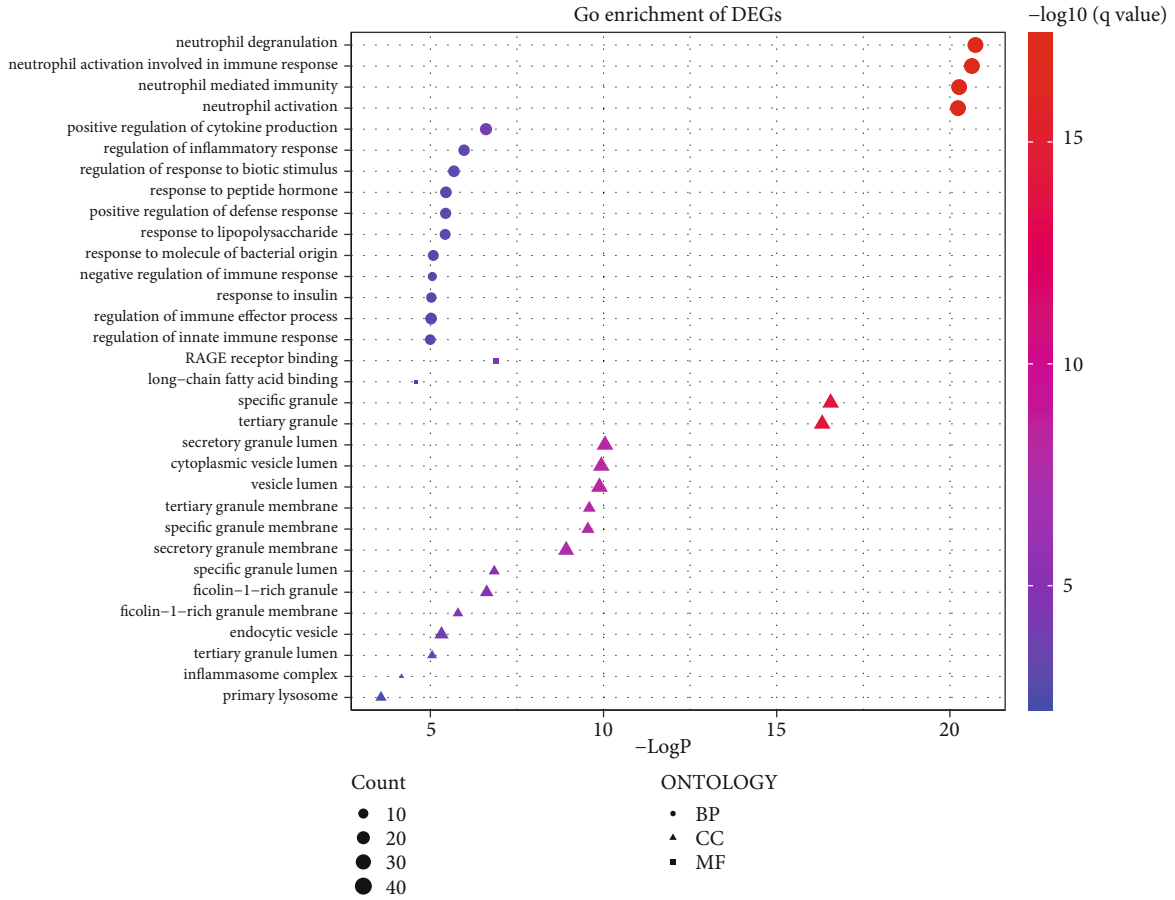


FIGURE 4: Gene enrichment result of 279 important features. Based on Gene Ontology database, from molecular function, cellular component, and biological process, respectively ($p < 0.05$).

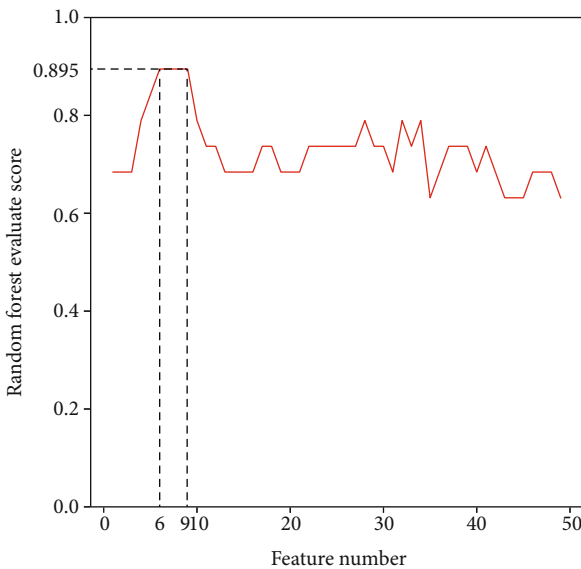


FIGURE 5: Relationship between the selected top feature number and model evaluation score. The line chart shows that the random forest evaluation score changes with the increase of feature number.

Tang et al. [38] explored the relationship between septic shock and AKI by analyzing codifferentially expressed genes (co-DEGs) in the hope of identifying possible genetic markers for septic shock-associated AKI. They downloaded two gene expression datasets (GSE30718 and GSE57065). DEGs related to septic shock and AKI were searched to clarify the molecular mechanism of DEGs through function analysis (GO), pathway enrichment analysis (KEGG), and protein interaction (PPI) network analysis. They also assessed co-DEGs and corresponding predictive miRNAs associated with septic shock and AKI. 16 genes, including NDUFAF1, were found to be involved in septic shock-associated AKI. Our study also found that NDUFAF1 expression was upregulated in patients with sepsis.

UBE2A, also called HHR6A or UBC2, can be expressed in a variety of tissues. Current studies mainly focus on cognitive impairment and skeletal muscle metabolism, but we have not found reports that UBE2A is directly related to sepsis. UBE2A may be associated with increased skeletal muscle protein catabolic activity in a number of diseases and malnutrition states, such as cancer, sepsis, and diabetes. Sepsis is often accompanied by septic encephalopathy, which is mainly manifested by changes in cognitive function and state of consciousness. It is necessary to further study whether UBE2A expression is abnormal in patients with septic encephalopathy [39, 40].

TABLE 1: DEG analysis in supplement dataset of 9 biomarkers. MCEMP1 was not sequenced in supplement dataset.

Biomarkers	logFC	Adj. p values
CLIC1	1.063383	2.56E-17
MCEMP1	—	—
PSTPIP2	1.257539	2.26E-12
UFD1	0.627806	3.35E-12
CD177	4.637112	9.32E-23
SEPT9	-1.2021	2.32E-22
NDUFAF1	1.206761	6.17E-16
GCA	1.582514	6.45E-16
UBE2A	0.595547	1.55E-12

In our study, there were 9 major differential genes involved in the development of sepsis. Five of these genes have been reported, indicating that the biomarkers selected by our random forest model have high diagnostic value. CLIC1, UFD1, SEPT9, and UBE2A are new biomarkers found by us through the random forest model, and there is no research report related to sepsis so far. These four genes may serve as relevant targets for the diagnosis and treatment of sepsis. Future in vitro and in vivo studies are needed to analyze the functions and pathways of these genes in the pathophysiology of sepsis. Further studies in more sepsis patients are needed to confirm the diagnostic value of the selected genes.

5. Conclusions

In this study, bioinformatics methods were used to analyze two septic shock-related datasets (GSE154918 and GES131761) and identify differentially expressed genes (DEGs) from GEO. We found that the number of differentially expressed genes increased with the increase of sepsis severity. It indicates that there are more genetic disorders from sepsis to septic shock. GO gene enrichment analysis showed that differential gene expression was significantly enriched in neutrophil activation and degranulation pathways. RAGE pathway has been found to be closely related to the occurrence of sepsis. Nine genes, including MCEMP1, PSTPIP2, CD177, GCA, NDUFAF1, CLIC1, UFD1, SEPT9, and UBE2A, were identified to be associated with sepsis. Further studies of the role of these pathways and genes in sepsis patients or experiments are needed.

Data Availability

The data used to support the findings of this study have been deposited in the GEO database repository (GSE154918 and GES131761).

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

Jing Zhou, Siqing Dong, and Ping Wang contribute equally to this work whether that is in the conception, study design, execution, or data analyzing. Jing Zhou, Siqing Dong, and Ping Wang contribute equally to this work.

Acknowledgments

This work was supported by grants from the Heilongjiang Postdoctoral Foundation (No. LBH-Z18219).

References

- [1] T. Zhao, Y. Hu, and L. Cheng, "Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches," *Briefings in Bioinformatics*, vol. 22, no. 4, 2021.
- [2] O. Huet and J. P. Chin-Dusting, "Septic shock: desperately seeking treatment," *Clinical Science (London, England)*, vol. 126, no. 1, pp. 31–39, 2014.
- [3] M. Singer, C. S. Deutschman, C. W. Seymour et al., "The third international consensus definitions for sepsis and septic shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.
- [4] H.-R. Yun, G. Lee, M. J. Jeon et al., "Erythropoiesis stimulating agent recommendation model using recurrent neural networks for patient with kidney failure with replacement therapy," *Computers in Biology and Medicine*, vol. 137, p. 104718, 2021.
- [5] C. Kok, V. Jahmunah, S. L. Oh et al., "Automated prediction of sepsis using temporal convolutional network," *Computers in Biology and Medicine*, vol. 127, p. 103957, 2020.
- [6] Z. Liu, A. Khojandi, A. Mohammed et al., "HeMA: a hierarchically enriched machine learning approach for managing false alarms in real time: a sepsis prediction case study," *Computers in Biology and Medicine*, vol. 131, p. 104255, 2021.
- [7] L. Cheng, C. Qi, H. Zhuang, T. Fu, and X. Zhang, "gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions," *Nucleic Acids Research*, vol. 48, no. D1, pp. D554–D560, 2020.
- [8] A. Rafiei, A. Rezaee, F. Hajati, S. Gheisari, and M. Golzan, "SSP: early prediction of sepsis using fully connected LSTM-CNN model," *Computers in Biology and Medicine*, vol. 128, p. 104110, 2021.
- [9] P. Shah, V. Chavda, S. Patel, S. Bhadada, and G. M. Ashraf, "Promising anti-stroke signature of Voglibose: investigation through in-silico molecular docking and virtual screening in in-vivo animal studies," *Current Gene Therapy*, vol. 20, no. 3, pp. 223–235, 2020.
- [10] C. Qi, C. Wang, L. Zhao et al., "SCovid: single-cell atlases for exposing molecular characteristics of COVID-19 across 10 human tissues," *Nucleic Acids Research*, vol. 50, no. D1, pp. D867–D874, 2022.
- [11] F. Mo, Y. Luo, D. A. Fan et al., "Integrated analysis of mRNA-seq and miRNA-seq to identify c-MYC, YAP1 and miR-3960 as major players in the anticancer effects of caffeic acid phenethyl ester in human small cell lung cancer cell line," *Current Gene Therapy*, vol. 20, no. 1, pp. 15–24, 2020.
- [12] T. Zhu, Q. Dai, and P.-A. He, "Identification of potential immune-related biomarkers in gastrointestinal cancers," *Current Bioinformatics*, vol. 16, no. 9, pp. 1203–1213, 2021.

- [13] W. Tang, S. Wan, Z. Yang, A. E. Teschendorff, and Q. Zou, "Tumor origin detection with tissue-specific miRNA and DNA methylation markers," *Bioinformatics*, vol. 34, no. 3, pp. 398–406, 2018.
- [14] J. Long, H. Yang, Z. Yang et al., "Integrated biomarker profiling of the metabolome associated with impaired fasting glucose and type 2 diabetes mellitus in large-scale Chinese patients," *Clinical and Translational Medicine*, vol. 11, no. 6, article e432, 2021.
- [15] L. Zhang, Y. Yang, L. Chai et al., "A deep learning model to identify gene expression level using cobinding transcription factor signals," *Briefings in Bioinformatics*, vol. 23, no. 1, 2022.
- [16] J. Y. Lu, D. J. Dai, and J. Zhou, "Research progress and future prospective of time in range (TIR)," *Zhonghua Yi Xue Za Zhi*, vol. 100, no. 38, pp. 2961–2965, 2020.
- [17] R. Biswas, D. Ghosh, B. Dutta, U. Halder, P. Goswami, and R. Bandopadhyay, "Potential non-coding RNAs from microorganisms and their therapeutic use in the treatment of different human cancers," *Current Gene Therapy*, vol. 21, no. 3, pp. 207–215, 2021.
- [18] S. Bourcier, P. Hindlet, B. Guidet, and A. Dechartres, "Reporting of organ support outcomes in septic shock randomized controlled trials: a methodologic review—the sepsis organ support study," *Critical Care Medicine*, vol. 47, no. 7, pp. 984–992, 2019.
- [19] H. Yang, Y. Luo, X. Ren et al., "Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators," *Information Fusion*, vol. 75, pp. 140–149, 2021.
- [20] V. Herwanto, B. Tang, Y. Wang et al., "Blood transcriptome analysis of patients with uncomplicated bacterial infection and sepsis," *BMC Research Notes*, vol. 14, no. 1, p. 76, 2021.
- [21] P. Martinez-Paz, M. Aragon-Camino, E. Gomez-Sanchez, M. Lorenzo-Lopez, E. Gomez-Pesquera, and A. Fadrigue-Fuentes, "Distinguishing septic shock from non-septic shock in postsurgical patients using gene expression," *The Journal of Infection*, vol. 83, no. 2, pp. 147–155, 2021.
- [22] M. E. Ritchie, B. Phipson, D. Wu et al., "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, p. e47, 2015.
- [23] A. Abraham, F. Pedregosa, M. Eickenberg et al., "Machine learning for neuroimaging with scikit-learn," *Frontiers in Neuroinformatics*, vol. 8, p. 14, 2014.
- [24] S. He, F. Guo, Q. Zou, and H. Ding, "MRMD2.0: a python tool for machine learning with feature ranking and reduction," *Current Bioinformatics*, vol. 15, no. 10, pp. 1213–1221, 2021.
- [25] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, "clusterProfiler: an R package for comparing biological themes among gene clusters," *OMICS*, vol. 16, no. 5, pp. 284–287, 2012.
- [26] L. Cheng, C. Qi, H. Yang et al., "gutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites," *Nucleic Acids Research*, vol. 50, no. D1, pp. D795–D800, 2022.
- [27] L. Liu, L. R. Zhang, F. Y. Dao, Y. C. Yang, and H. Lin, "A computational framework for identifying the transcription factors involved in enhancer-promoter loop formation," *Mol Ther Nucleic Acids*, vol. 23, pp. 347–354, 2021.
- [28] T. van der Poll, M. Shankar-Hari, and W. J. Wiersinga, "The immunology of sepsis," *Immunity*, vol. 54, no. 11, pp. 2450–2464, 2021.
- [29] X. Zhao, Y. N. Liao, and Q. Huang, "The impact of RAGE inhibition in animal models of bacterial sepsis: a systematic review and meta-analysis," *The Journal of International Medical Research*, vol. 46, no. 1, pp. 11–21, 2018.
- [30] C. W. Russell, A. C. Richards, A. S. Chang, and M. A. Mulvey, "The rhomboid protease GlpG promotes the persistence of extraintestinal pathogenic *Escherichia coli* within the gut," *Infection and Immunity*, vol. 85, no. 6, 2017.
- [31] M. Martin-Fernandez, A. Tamayo-Velasco, R. Aller, H. Gonzalo-Benito, P. Martinez-Paz, and E. Tamayo, "Endothelial dysfunction and neutrophil degranulation as central events in sepsis physiopathology," *International Journal of Molecular Sciences*, vol. 22, no. 12, p. 6272, 2021.
- [32] J. X. Chen, X. Xu, and S. Zhang, "Silence of long noncoding RNA NEAT1 exerts suppressive effects on immunity during sepsis by promoting microRNA-125-dependent MCEMP1 downregulation," *IUBMB Life*, vol. 71, no. 7, pp. 956–968, 2019.
- [33] W. Xie, L. Chen, L. Chen, and Q. Kou, "Silencing of long non-coding RNA MALAT1 suppresses inflammation in septic mice: role of microRNA-23a in the down-regulation of MCEMP1 expression," *Inflammation Research*, vol. 69, no. 2, pp. 179–190, 2020.
- [34] J. J. Xu, H. D. Li, X. S. Du, J. J. Li, X. M. Meng, and C. Huang, "Role of the F-BAR family member PSTPIP2 in autoinflammatory diseases," *Frontiers in Immunology*, vol. 12, 2021.
- [35] H. Chen, Y. Li, T. Li et al., "Identification of potential transcriptional biomarkers differently expressed in both *S. aureus*- and *E. coli*-induced sepsis via integrated analysis," *BioMed Research International*, vol. 2019, Article ID 2487921, 2019.
- [36] J. Demaret, F. Venet, J. Plassais et al., "Identification of CD177 as the most dysregulated parameter in a microarray study of purified neutrophils from septic shock patients," *Immunology Letters*, vol. 178, pp. 122–130, 2016.
- [37] Y. X. Yang and L. Li, "Identification of potential biomarkers of sepsis using bioinformatics analysis," *Experimental and Therapeutic Medicine*, vol. 13, no. 5, pp. 1689–1696, 2017.
- [38] Y. Tang, X. Yang, H. Shu et al., "Bioinformatic analysis identifies potential biomarkers and therapeutic targets of septic-shock-associated acute kidney injury," *Hereditas*, vol. 158, no. 1, p. 13, 2021.
- [39] O. A. Adegoke, N. Bedard, H. P. Roest, and S. S. Wing, "Ubiquitin-conjugating enzyme E214k/HR6B is dispensable for increased protein catabolism in muscle of fasted mice," *American Journal of Physiology. Endocrinology and Metabolism*, vol. 283, no. 3, pp. E482–E489, 2002.
- [40] C. Polge, R. Leulmi, M. Jarzaguet, A. Claustre, L. Combaret, and D. Bechet, "UBE2B is implicated in myofibrillar protein loss in catabolic C2C12 myotubes," *Journal of Cachexia, Sarcopenia and Muscle*, vol. 7, no. 3, pp. 377–387, 2016.

Research Article

Prediction of New Risk Genes and Potential Drugs for Rheumatoid Arthritis from Multiomics Data

Anteneh M. Birga,¹ Liping Ren,² Huaichao Luo ,^{1,3} Yang Zhang ,⁴ and Jian Huang ¹

¹School of Life Science and Technology, University of Electronic Science and Technology of China (UESTC), Chengdu, China

²School of Health Care Technology, Chengdu Neusoft University, Chengdu, China

³Department of Clinical Laboratory, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, University of Electronic Science and Technology of China (UESTC), Chengdu, China

⁴Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, China

Correspondence should be addressed to Huaichao Luo; luo1987cc@163.com, Yang Zhang; zhy1001@alu.uestc.edu.cn, and Jian Huang; hj@uestc.edu.cn

Received 17 October 2021; Revised 8 December 2021; Accepted 12 January 2022; Published 31 January 2022

Academic Editor: Chung-Min Liao

Copyright © 2022 Anteneh M. Birga et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Rheumatoid arthritis (RA) is an autoimmune and inflammatory disease for which there is a lack of therapeutic options. Genome-wide association studies (GWASs) have identified over 100 genetic loci associated with RA susceptibility; however, the most causal risk genes (RGs) associated with, and molecular mechanism underlying, RA remain unknown. In this study, we collected 95 RA-associated loci from multiple GWASs and detected 87 candidate high-confidence risk genes (HRGs) from these loci via integrated multiomics data (the genome-scale chromosome conformation capture data, enhancer-promoter linkage data, and gene expression data) using the Bayesian integrative risk gene selector (iRIGS). Analysis of these HRGs indicates that these genes were indeed, markedly associated with different aspects of RA. Among these, 36 and 46 HRGs have been reported to be related to RA and autoimmunity, respectively. Meanwhile, most novel HRGs were also involved in the significantly enriched RA-related biological functions and pathways. Furthermore, drug repositioning prediction of the HRGs revealed three potential targets (ERBB2, IL6ST, and MAPK1) and nine possible drugs for RA treatment, of which two IL-6 receptor antagonists (tocilizumab and sarilumab) have been approved for RA treatment and four drugs (trastuzumab, lapatinib, masoprocol, and arsenic trioxide) have been reported to have a high potential to ameliorate RA. In summary, we believe that this study provides new clues for understanding the pathogenesis of RA and is important for research regarding the mechanisms underlying RA and the development of therapeutics for this condition.

1. Introduction

Rheumatoid arthritis (RA) is an autoimmune and inflammatory disease in which the immune system mistakenly attacks healthy joint tissues, thereby causing inflammation that primarily affects the joints [1]. It is a multifactorial disease involving complex traits affected by many genetic and environmental factors, as well as the potential interactions among these factors [2]. Although the etiology underlying RA development is not fully understood, investigators have determined that abnormal immune system responses are

the core cause of RA-associated inflammation and joint destruction [3].

Currently, there is no cure for RA. Disease-modifying antirheumatic drugs (DMARDs) still represent the main treatment strategy for RA. These drugs mainly act on the immune system and slow the progression of RA; they can efficiently attenuate disease symptoms and substantially decrease and/or delay joint deformity [4]. DMARDs can be classified as follows: conventional DMARDs and biologic DMARDs [5]. Commonly used conventional DMARDs include methotrexate, leflunomide, hydroxychloroquine,

and sulfasalazine. Recently, many biological DMARDs, including TNF inhibitors (adalimumab, infliximab, and etanercept), anti-CD20 antibodies (rituximab), IL-6 receptor antibodies (sarilumab), RANKL antibodies (denosumab), and Janus kinase inhibitors (baricitinib), have been developed [6, 7]. Despite the increasing numbers of new drugs and treatment regimens, agents that completely cure RA or long-acting agents for RA are still far from being developed; thus, novel therapeutics and/or targets for this condition are required.

Hereditary factors show a clear causal relationship with RA [8]. And elucidating the pathogenesis of RA from the genomics and genetics standpoints is an important means for clinical therapeutics and drug discovery [9]. At present, genome-wide association studies (GWASs) have identified over 100 genetic loci associated with RA susceptibility [10, 11]. Although genetic information indicates an association between genetic factors and RA, the most causal risk genes (RGs) associated with RA and the molecular mechanisms underlying this disease remain unknown [12]. Mo et al. [13] predicted the RA-associated susceptibility genes by the summary data-based Mendelian randomization (SMR) analysis and identified 140 genes that showed causal association with RA. Moreover, thus far, only a few effective drug targets have been identified through GWASs [14].

In this study, to identify RA-associated RGs and predict candidate drug targets for RA, we collected 95 RA-associated loci from different GWASs and detected the candidate RGs from these loci via integrated multiomics data (the genome-scale chromosome conformation capture data, enhancer-promoter linkage data, and gene expression data) using the Bayesian integrative risk gene selector (iRIGS) [15]. Then, we evaluated the relevance between the candidate RGs and RA progression in the context of multiple aspects, such as biological functions, gene expression, and gene regulatory patterns. Finally, we predicted the candidate targets and drugs of these RA-associated RGs using the drug repositioning prediction approach (Figure 1(a)).

2. Methods

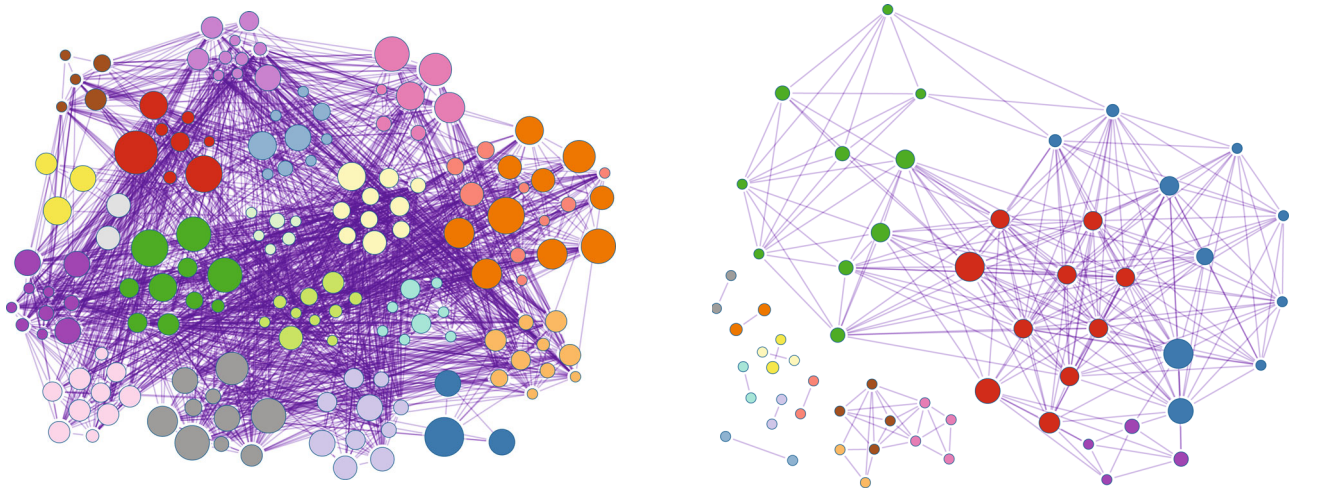
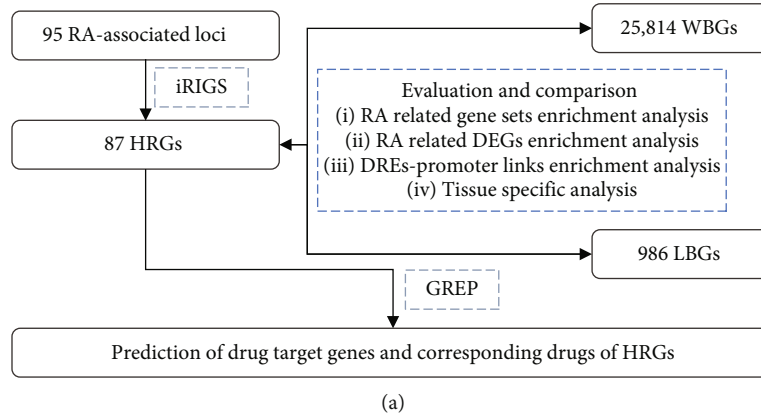
2.1. RA-Associated Loci. We collected over 100 RA-associated loci from multiple GWASs, including 101 loci collected from a meta-analysis GWAS containing over 100,000 subjects of European and Asian ancestries (29,880 RA cases vs. 73,758 controls) [16], two loci collected from a GWAS containing over 1,600 subjects (397 RA cases vs. 1,211 controls) [17], and four loci collected from a case-control GWAS of a cohort of Arab subjects (511 RA cases vs. 352 controls) [18]. Finally, a total of 104 RA-associated loci were collected (there are 3 duplicated SNPs). After excluding 12 loci for which SNP IDs were unavailable, 95 RA-associated loci were included in this study.

2.2. Identifying RGs for the RA-Associated Loci. The high-confidence risk genes (HRGs) of RA were inferred by iRIGS (GRCh38/hg38) [15], which is a powerful tool for RG identification that integrates multiomics data and gene networks. Here, the omics data include two RA-associated gene expres-

sion datasets, i.e., GSE55235 [19] and GSE77298 [20], two distal regulatory element- (DRE-) promoter linkage datasets, 1,618,000 DRE promoter linkages obtained from genome-scale chromosome conformation capture (Hi-C) [21], and 66,899 enhancer-promoter linkages obtained from the FANTOM5 project [22]. All these omics data have been processed and deposited in iRIGS. Furthermore, the GO network data containing gene-gene relationships obtained by the iRIGS method were also integrated. A total of 1,972 candidate genes located within a 2Mb region centered at the index SNP were collected as the candidate genes for iRIGS analysis. The posterior probability (PP) value was calculated by a Bayesian framework embedded in iRIGS [15], which is the index of possibility for genes to serve as an RG for RA. For each GWAS locus, one or more RGs can be selected according to the PP value. In this study, we only selected one risk gene with the highest PP for each locus. For evaluation of HRGs, we constructed two background gene lists for comparison with the HRGs: (1) the local background genes (LBGs), which is defined as the genes with P values less than the median PP of all candidate genes (1,972 genes located within a 2Mb region of the RA-associated loci). Ultimately, a total of 986 LBGs were obtained; (2) the whole-genome background genes (WBGs), which are defined as the genes that included all the human genes (obtained from the R package of iRIGS) except the HRGs. Ultimately, a total of 25,814 WBGs were obtained.

2.3. Data Collection. Five RA-associated keyword gene sets (keywords: “Arthritis,” “Rheumatic,” “Autoimmune,” “Joint,” and “Connective Tissue”) were constructed from the GeneCards database (<http://www.genecards.org>). At first, the five keywords were used to research the related genes in the GeneCards database; then, the genes with a relevance score greater than 10 were considered as the keyword-related genes. Finally, it was found that the “Connective Tissue” gene set contained 507 genes, the “Joint” gene set contained 1,063 genes, the “Autoimmune” gene set contained 457 genes, the “Arthritis” gene set contained 422 genes, and the “Rheumatic” gene set contained 65 genes. Furthermore, an immune system-related gene set containing 1,534 genes was collected from the ImmPort database (<https://www.immport.org>) [23]. The tissue-specific gene expression profiles (FPKM, reads per kilobase of transcript per million mapped reads) were collected from GTEx release V8 data source [24].

2.4. Drug Repositioning Prediction of the HRGs. To predict the drug-specific target genes and corresponding drugs specific to the HRGs, a command-line Python software, Genome for REPositioning drugs (GREP), was used [25]. The GREP software quantifies the enrichment of drug targets by using DrugBank and the Therapeutic Target Database. Approximately 22,300 drugs and 2,029 genes were categorized based on the Anatomical Therapeutic Chemical (ATC) and World Health Organization (WHO) classification system; the P values and odds ratios for this categorization were calculated using Fisher’s exact test.



- | | |
|--|--|
| <ul style="list-style-type: none"> ■ GO:0019900: kinase binding ■ GO:0008134: transcription factor binding ■ GO:0071396: cellular response to lipid ■ GO:0051403: stress-activated MAPK cascade ■ GO:0007159: leukocyte cell-cell adhesion ■ GO:0044389: ubiquitin-like protein ligase binding ■ GO:0019902: phosphatase binding ■ GO:0007169: transmembrane receptor protein tyrosine kinase signaling pathway ■ GO:0001568: blood vessel development ■ GO:0001959: regulation of cytokine-mediated signaling pathway ■ GO:0010035: response to inorganic substance ■ GO:0009896: positive regulation of catabolic process ■ GO:0009615: response to virus ■ GO:0071407: cellular response to organic cyclic compound ■ GO:0060485: mesenchyme development ■ GO:0032663: regulation of interleukin-2 production ■ GO:0006469: negative regulation of protein kinase activity ■ GO:0005925: focal adhesion ■ GO:1901652: response to peptide ■ GO:1904019: epithelial cell apoptotic process | <ul style="list-style-type: none"> ■ hsa05161: Hepatitis B ■ hsa05169: Epstein-Barr virus infection ■ ko04659: Th17 cell differentiation ■ hsa05166: HTLV-I infection ■ ko05203: Viral carcinogenesis ■ hsa05202: Transcriptional misregulation in cancer ■ hsa04114: Oocyte meiosis ■ hsa04810: Regulation of actin cytoskeleton ■ hsa05130: Pathogenic Escherichia coli infection ■ ko04141: Protein processing in endoplasmic reticulum ■ hsa04530: Tight junction ■ hsa05010: Alzheimer's disease ■ M00177: Ribosome, eukaryotes ■ ko04152: AMPK signaling pathway ■ hsa04110: Cell cycle |
|--|--|

(b)

FIGURE 1: A flowchart depicting the steps in our study and the function enrichment analysis of the HRGs. (a) A flowchart detailing the steps followed in this study. (b) The GO and KEGG pathway analyses of the HRGs.

TABLE 1: Information of some RA or autoimmunity-related HRGs.

HRG	SNP	PMID	RA related	Autoimmunity related
IL6ST	rs7731626	16646038	Yes	Yes
SUMO1	rs6715284	30562482; 17360386	Yes	
XPO1	rs13385025, rs34695944	24965445	Yes	
FOXO1	rs9603616	24812285	Yes	Yes
HIF1A	rs3783782	27445820	Yes	Yes
DUSP22	rs9378815	29287311	Yes	
GATA3	rs12413578, rs3824660	19248112; 29097726	Yes	Yes
AKT1	rs2582532	28559961	Yes	
CD40	rs4239702	28455435	Yes	Yes
EGR2	rs6479800, rs71508903	24058814		Yes

TABLE 2: Information of some HRGs without direct evidence linking to RA.

HRGs	SNP	PP value	Description
PTPRC	rs17668708	0.429	Associated with response to TNF α therapy
ANXA11	rs726288	0.427	Antigen associated with systemic autoimmune diseases
SPRED1	rs8032939	0.369	Suppressor of the Ras-ERK pathway
PRDM1	rs9372120	0.366	PRDM1 is belonging to the B cell development pathway
BUB1	rs6732565	0.351	Differentially expressed in RA chondrocytes
LCLAT1	rs10175798	0.327	Related to triacylglycerol biosynthesis and fatty acyl-CoA biosynthesis
AZI2	rs3806624	0.292	Activator of NFKB
GDI2	rs947474	0.284	Is a candidate biomarker in synovial fluid of RA
CNOT6L	rs10028001	0.2766	Differentially expressed in RA
RFTN1	rs4452313	0.271	Involved in T-cell antigen receptor-mediated signaling

2.5. Statistical Analysis. The differentially expressed genes (DEGs) were identified using the Limma package in the R software (adjusted. $P < 0.05$) [26]. The GO and pathway enrichment analyses were performed using Metascape [27]. One-sided Fisher’s exact test and one-sided Wilcoxon rank-sum test were performed using the R software. The Jensen–Shannon divergence (JSD) score was calculated using the R package “philtropy.” The P values were adjusted using the Bonferroni correction method.

3. Results

3.1. Predicting HRGs for RA. A total of 87 HRGs related to the 95 RA-associated loci were inferred using iRIGS; most of these genes have been implicated in RA and/or autoimmunity (see Table 1 and Supplementary Table 1). Some of the well-known drug targets for RA treatment, such as IRAK1, HIF1A, and IL6ST, have been identified as HRGs for RA [28]. Further, 36 and 46 genes have been reported to be related to RA and autoimmunity, respectively. For instance, IL6/IL6ST signaling plays a key role in the progression of RA, and some IL6 receptor antagonists have been proved to be effective in altering leukocyte trafficking and reducing the severity of RA [29]. GATA-3 has been shown to protect against severe joint inflammation and reduce the differentiation of Th17 cells in mice with RA [30]. EGR2 acts as a key regulator for systemic autoimmunity by regulating cytokine production and cell

proliferation [31]. Meanwhile, we also investigated the rest HRGs which have no direct evidence linking to RA and found that these HRGs might also be close to RA or autoimmunity diseases (Table 2). For example, PTPRC is associated with response to antitumor necrosis factor- α therapy, which is a mainstay of treatment in rheumatoid arthritis [32]. ANXA11 is an antigen associated with multiple systemic autoimmune diseases [33]. GDI2 is a candidate biomarker in synovial fluid of RA [34]. And there are seven genes (TNFAIP3, XPO1, GDI2, GATA3, EGR2, DDB1, and ABI2) supported by more than one SNP. Most of which are related to the RA. TNFAIP3 showed differential expression between RA and osteoarthritis synoviocytes [35]. XPO1 has been indicated to serve as new candidate therapeutic targets for RA [36]. Moreover, the GO and KEGG pathway enrichment analyses of the HRGs showed that these genes were enriched mainly in intercellular communication and immune-related functions and pathways, such as leukocyte cell-cell adhesion, focal adhesion, regulation of cytokine-mediated signaling pathways, tight junction formation, Th17 cell differentiation, and regulation of interleukin-2 production (Figure 1(b)). These functions and pathways have been reported to be critical for RA progression [37, 38].

3.2. Evaluation of the HRGs. To assess the reliability of the HRGs, we constructed two background gene lists for comparison with the HRGs: the local background genes (LBGs)

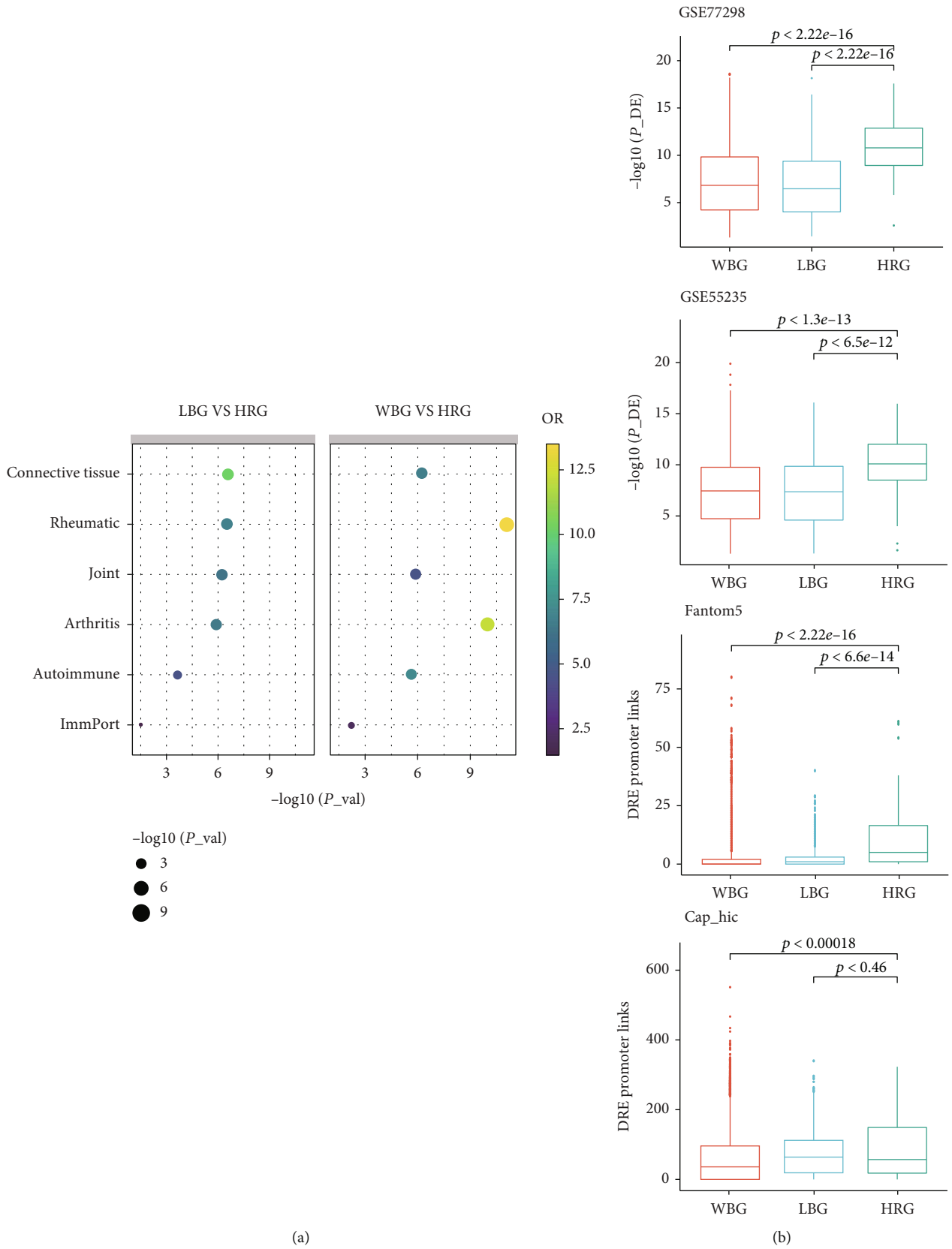


FIGURE 2: Continued.

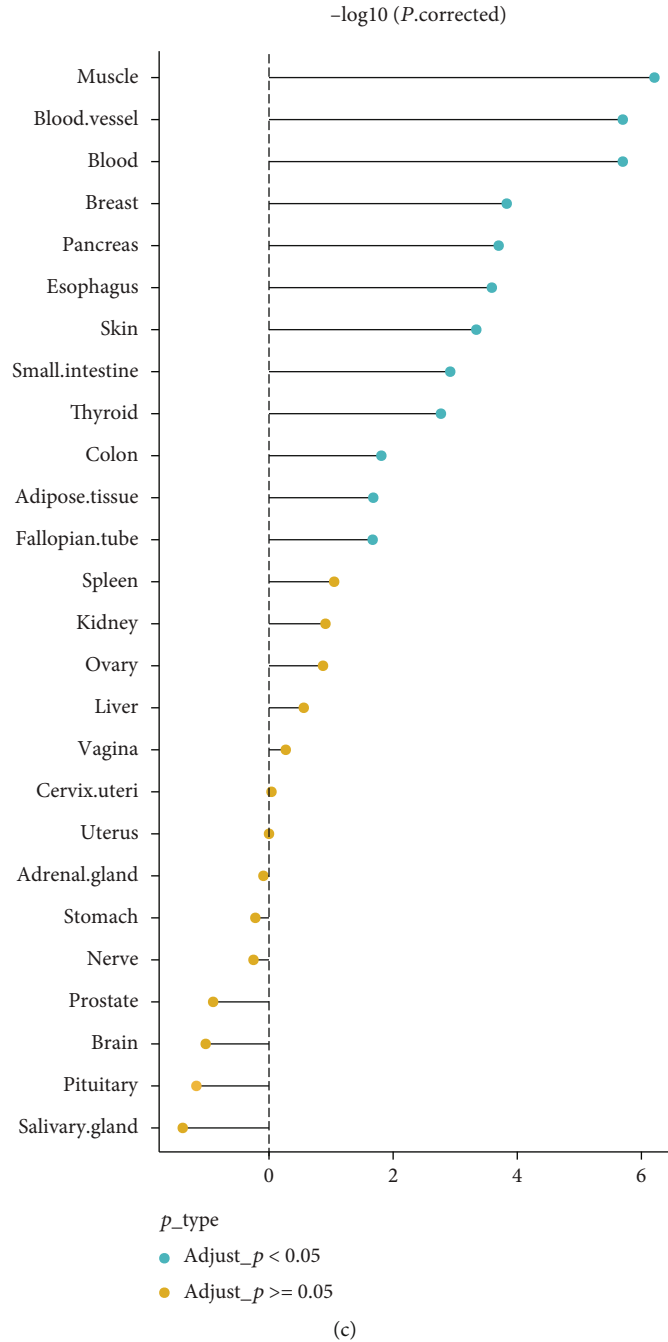


FIGURE 2: Comparison of the HRGs with the local background genes (LBGs) and whole-genome background genes (WBGs). (a) Comparison of the HRGs with the LBGs and WBGs using the six RA-related gene sets: the “Arthritis,” “Rheumatic,” “Autoimmune,” “Joint,” “Connective Tissue,” and “ImmPort” gene sets. (b) Comparison of the HRGs with the LBGs and WBGs using the two gene expression datasets GSE77298 and GSE55235 and the two DRE-promoter linkage datasets obtained using the Hi-C and FANTOM5. (c) Tissue-specificity analysis of the HRGs (one-sided Wilcoxon rank-sum test).

included 986 genes with PP values less than the median PP of all candidate genes, and the whole-genome background genes (WBGs) included all the human genes except the HRGs (25,814 genes). At first, concerning biological function, we compared the HRGs with the LBGs and WBGs using the six RA-related gene sets, i.e., the “Arthritis,” “Rheumatic,” “Autoimmune,” “Joint,” “Connective Tissue,” and “ImmPort” gene sets (see Methods for details). As

shown in Figure 2(a), HRGs were significantly enriched in all the six RA-related gene sets (one-sided Fisher’s exact test: P value < 0.05). Next, about gene expression, we compared the HRGs with the LBGs and WBGs using the two gene expression datasets GSE77298 and GSE55235; as shown in Figure 2(b), the HRGs were more likely to serve as the DEGs in these two RA gene expression profiles (one-sided Wilcoxon rank-sum test: P value < 0.05). Then, with regard to

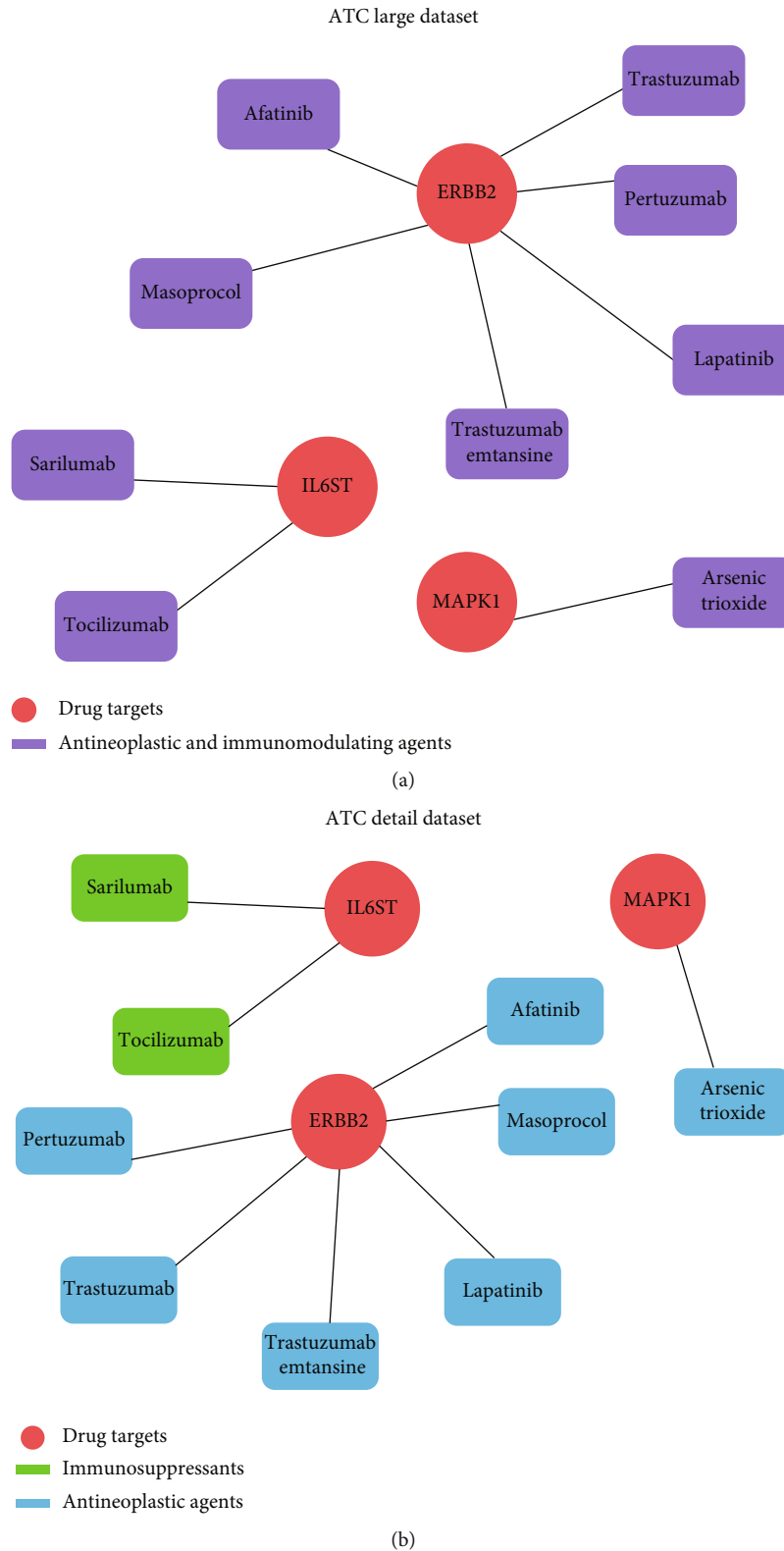


FIGURE 3: Drug repositioning prediction of the HRGs based on (a) the ATC large dataset and (b) the detailed ATC dataset.

gene regulation, we compared the HRGs with the LBGs and WBGs using the two DRE-promoter linkage datasets obtained using the Hi-C and FANTOM5 methods. These results also showed that the HRGs were significantly associ-

ated with a large number of DREs (Figure 2(b); one-sided Wilcoxon rank-sum test: P value < 0.05). To investigate the tissue specificity of the HRGs, we converted the RPKM GTEx data to JSD scores to represent the tissue specificity

of each gene for each tissue. Moreover, compared to the LBGs, the HRGs showed a significantly high expression in the muscles, blood vessels, blood, etc. (see Figure 2(c), one-sided Wilcoxon rank-sum test: adjusted P value < 0.05). These tissues have been proved involved in RA progression. For example, muscle deterioration (myositis and weakness) and inflammation of blood vessels (vasculitis and ulcers) are common complications of RA [39].

3.3. Predicting the Targets and Corresponding Drugs for the HRGs. To investigate whether some HRGs could serve as targets of existing repositioned drugs for RA therapy, we used GREP to perform enrichment analysis to ascertain the targets of the existing and approved drugs (see Methods for details). As shown in Figure 3 and Supplementary Table 2, three HRGs, ERBB2, IL6ST, and MAPK1, were identified to be related to the targets of immunosuppressants and antineoplastic agents. A total of six potential drugs (trastuzumab, pertuzumab, trastuzumab emtansine, lapatinib, afatinib, and masoprocol) were predicted to target ERBB2. Of these, trastuzumab, pertuzumab, and trastuzumab emtansine are HER2/ErbB2 receptor monoclonal antibodies approved for the treatment of metastatic HER2-positive breast cancer, and trastuzumab has been reported to inhibit RA synovial cell growth [40]. Lapatinib has been reported to ameliorate experimental arthritis in rats by targeting epidermal growth factor receptors (EGFRs) [41]. Li et al. [42] found that masoprocol significantly reduces the severity of bone destruction and osteoclast recruitment in the ankle joint of rats with adjuvant-induced arthritis and indicated the potential utility of masoprocol as a therapeutic agent for RA. Pertuzumab and afatinib have also been approved as antineoplastic agents. Two potential drugs (tocilizumab and sarilumab) were predicted to target IL6ST. Tocilizumab, which functions by targeting IL-6 receptors, was the first DMARD to be approved for RA treatment [43]. Sarilumab was the second IL-6 receptor antagonist to be approved for the treatment of RA [44]. Arsenic trioxide, which has been reported as a potential therapeutic agent for RA, was predicted to target MAPK1; it has also been approved to treat leukemia and reported to regulate the Treg and Th17 cell balance by modulating STAT3 expression in treatment-naïve RA patients [45].

4. Discussion

To date, the exact cause of the immune system's faulty response in RA remains unclear [46]. Though some genes have been identified to be responsible for the increased risk of developing RA, such as HLA complex, STAT4, TRAF1, and PTPN22 [47], most RA-related RGs and their causal variants remain unknown [48]. Recently, GWASs have been utilized to identify RA-associated genetic variants on a genome-wide scale, and over 100 RA-associated loci were obtained [10, 11]. However, the presence of most GWAS variants (90%) in noncoding regions hinders the identification of disease-related RGs [49], which also obscures the interpretation of their mode of action and the correct iden-

tification of the target gene via which the causal variant may affect the phenotype [50]. Herein, to fill this gap, we identified 87 HRGs from 95 RA-associated loci collected from different GWASs based on multiomics data. The assessment of the HRGs indicated that they were markedly correlated with RA progression. In addition, using drug repositioning prediction, we also identified several targets of these genes and the drugs associated with their function. Some of these identified drugs have already been approved for RA treatment.

The inspection of previously published literature revealed that 36 and 46 HRGs have been implicated in RA progression and autoimmunity, respectively. Besides the well-known drug targets for RA treatment, such as IRAK1, HIF1A, and IL6ST, some HRGs, including XPO1, GATA3, MYC, and CD40, have also been indicated to serve as new candidate therapeutic targets for RA [36, 51, 52]. The function enrichment analysis of the HRGs showed that they were enriched mainly in the immune system- and intercellular communication-related functions and pathways. It is known that RA is a classic autoimmune and inflammatory disease that strongly involves multiple innate and adaptive immune-related processes [53]. Additionally, the dysfunction of several intercellular signaling pathways, including the JAK/STAT, SAPK/MAPK, and PI-3K/AKT/mTOR signaling pathways, plays a critical role in RA [37]. Cell-cell crosstalk mediates various biological processes in the tissue microenvironment in RA. Therefore, many studies have focused on the development of new therapeutics for RA by considering the intercellular communications in RA [54–56]. These results indicate that the HRGs identified herein are markedly involved in RA progression and are of importance for research regarding the mechanism underlying RA and therapeutic strategies for this condition. Moreover, some of the rest HRGs without direct evidence linking to RA are also involved in autoimmunity disease-related functions or pathways. This part of HRGs is probably more worth exploring than the well-known RA-related HRGs.

The comparison of the HRGs with the LBGs and HRGs showed that the HRGs are markedly associated with RA-related functions and RA-related DEGs and indicated that the expression levels of the HRGs tend to be regulated by DREs. Interestingly, the HRGs showed a markedly high expression in the muscle tissues, blood vessels, and blood. Muscle deterioration (myositis and weakness) and inflammation of blood vessels (vasculitis and ulcers) are common complications of RA [39]. Therefore, the high expression of HRGs in these tissues may implicate them in the progression of RA and may highlight them as potential therapeutic targets for RA. Further, the expression of HRGs in the blood may mainly influence RA-related immune processes [57, 58]; this may also implicate these HRGs as factors governing, and ultimately, as candidate biomarkers for, the progression of RA.

Drug repositioning prediction of the HRGs yielded three targets and nine drugs. Two IL-6 receptor antagonist drugs, tocilizumab and sarilumab, have been approved for RA treatment. Meanwhile, trastuzumab, lapatinib, masoprocol,

and arsenic trioxide have been reported to ameliorate the symptoms of RA in patients or model animals and may serve as candidate DMARDs for RA treatment. The other drugs, pertuzumab, trastuzumab emtansine, and afatinib, have also been approved as immunosuppressants and/or antineoplastic agents. These results not only indicate that these HRGs are markedly involved in RA progression but also provide a trajectory for screening effective drugs for RA treatment.

5. Conclusion

In this study, we collected 95 RA-associated loci from different GWASs of RA and obtained 87 HRGs from these loci using a multiomics-based method. The analysis and evaluation of these HRGs indicated that these genes were indeed, highly involved in RA. Moreover, the drug repositioning prediction of the HRGs suggested several potential targets and drugs for RA treatment. In summary, this study predicted new RGs, drug targets, and drugs for RA using the GWAS and multiomics data. We believe that our study provides more clues for understanding the pathogenesis of RA and will be important for research regarding the mechanisms underlying RA and the possible therapeutic strategies for this condition.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We thank Dr. Quan Wang for his professional advice on using iRIGS and Mr. Hamza B. Abagna for kindly copyediting the manuscript. This work was supported by the National Natural Science Foundation of China (Grant No. 62071099) and the Basic and Applied Basic Research Fund of Guangdong Province (Grant No. 2019A1515110701).

Supplementary Materials

Supplementary Table 1: list of the 87 HRGs. Supplementary Table 2: predicting the targets and corresponding drugs for the HRGs by GREP. (*Supplementary Materials*)

References

- [1] M. Prete, V. Racanelli, L. Digiglio, A. Vacca, F. Dammacco, and F. Perosa, "Extra-articular manifestations of rheumatoid arthritis: an update," *Autoimmunity Reviews*, vol. 11, no. 2, pp. 123–131, 2011.
- [2] J. S. Smolen, D. Aletaha, A. Barton et al., "Rheumatoid arthritis," *Nature Reviews Disease Primers*, vol. 4, no. 1, p. 18001, 2018.
- [3] E. Marcucci, E. Bartoloni, A. Alunno et al., "Extra-articular rheumatoid arthritis," *Reumatismo*, vol. 70, no. 4, pp. 212–224, 2018.
- [4] Y. J. Lin, M. Anzaghe, and S. Schülke, "Update on the pathomechanism, diagnosis, and treatment options for rheumatoid arthritis," *Cells*, vol. 9, no. 4, 2020.
- [5] A. Rubbert-Roth, M. Z. Szabó, M. Kedves, G. Nagy, F. Atzeni, and P. Sarzi-Puttini, "Failure of anti-TNF treatment in patients with rheumatoid arthritis: the pros and cons of the early use of alternative biological agents," *Autoimmunity Reviews*, vol. 18, no. 12, p. 102398, 2019.
- [6] Q. Guo, Y. Wang, D. Xu, J. Nossent, N. J. Pavlos, and J. Xu, "Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies," *Bone research*, vol. 6, no. 1, p. 15, 2018.
- [7] Y. Huang, J. Wang, Y. Zhao et al., "cncRNAdb: a manually curated resource of experimentally supported RNAs with both protein-coding and noncoding function," *Nucleic Acids Research*, vol. 49, no. D1, pp. D65–D70, 2021.
- [8] J. Kurkó, T. Besenyi, J. Laki, T. T. Glant, K. Mikecz, and Z. Szekanecz, "Genetics of rheumatoid arthritis - a comprehensive review," *Clinical Reviews in Allergy and Immunology*, vol. 45, no. 2, pp. 170–179, 2013.
- [9] L. E. Dedmon, "The genetics of rheumatoid arthritis," *Rheumatology (Oxford)*, vol. 59, no. 10, pp. 2661–2670, 2020.
- [10] E. Ha, S.-C. Bae, and K. Kim, "Large-scale meta-analysis across East Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci," *Annals of the Rheumatic Diseases*, vol. 80, no. 5, pp. 558–565, 2021.
- [11] S. Onuora, "New insights into RA genetics from GWAS meta-analysis," *Nature Reviews Rheumatology*, vol. 17, no. 3, pp. 128–128, 2021.
- [12] Y. Okada, S. Eyre, A. Suzuki, Y. Kochi, and K. Yamamoto, "Genetics of rheumatoid arthritis: 2018 status," *Annals of the Rheumatic Diseases*, vol. 78, no. 4, pp. 446–453, 2019.
- [13] X.-B. Mo, Y.-H. Sun, Y.-H. Zhang, and S. F. Lei, "Integrative analysis highlighted susceptibility genes for rheumatoid arthritis," *International Immunopharmacology*, vol. 86, p. 106716, 2020.
- [14] H. Fang, L. Chen, and J. C. Knight, "From genome-wide association studies to rational drug target prioritisation in inflammatory arthritis," *The Lancet Rheumatology*, vol. 2, no. 1, pp. e50–e62, 2020.
- [15] Q. Wang, R. Chen, F. Cheng et al., "A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data," *Nature Neuroscience*, vol. 22, no. 5, pp. 691–699, 2019.
- [16] Y. Okada, D. Wu, G. Trynka et al., "Genetics of rheumatoid arthritis contributes to biology and drug discovery," *Nature*, vol. 506, no. 7488, pp. 376–381, 2014.
- [17] R. M. Plenge, C. Cotsapas, L. Davies et al., "Two independent alleles at 6q23 associated with risk of rheumatoid arthritis," *Nature Genetics*, vol. 39, no. 12, pp. 1477–1482, 2007.
- [18] R. Saxena, R. M. Plenge, A. C. Bjornnes et al., "A multinational Arab genome-wide association study identifies new genetic associations for rheumatoid arthritis," *Arthritis & Rheumatology*, vol. 69, no. 5, pp. 976–985, 2017.
- [19] D. Woetzel, R. Huber, P. Kupfer et al., "Identification of rheumatoid arthritis and osteoarthritis patients by transcriptome-

- based rule set generation,” *Arthritis Research & Therapy*, vol. 16, no. 2, p. R84, 2014.
- [20] M. G. Broeren, M. de Vries, M. B. Bennink et al., “Disease-regulated gene therapy with anti-inflammatory interleukin-10 under the control of the CXCL10 promoter for the treatment of rheumatoid arthritis,” *Human Gene Therapy*, vol. 27, no. 3, pp. 244–254, 2016.
- [21] B. Mifsud, F. Tavares-Cadete, A. N. Young et al., “Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C,” *Nature Genetics*, vol. 47, no. 6, pp. 598–606, 2015.
- [22] A. R. R. Forrest, H. Kawaji, M. Rehli et al., “A promoter-level mammalian expression atlas,” *Nature*, vol. 507, no. 7493, pp. 462–470, 2014.
- [23] S. Bhattacharya, S. Andorf, L. Gomes et al., “ImmPort: disseminating data to the public for the future of immunology,” *Immunologic Research*, vol. 58, no. 2-3, pp. 234–239, 2014.
- [24] S. Jiang, S.-J. Cheng, L.-C. Ren et al., “An expanded landscape of human long noncoding RNA,” *Nucleic Acids Research*, vol. 47, no. 15, pp. 7842–7856, 2019.
- [25] S. Sakaue and Y. Okada, “GREP: genome for REPositioning drugs,” *Bioinformatics*, vol. 35, no. 19, pp. 3821–3823, 2019.
- [26] I. Diboun, L. Wernisch, C. A. Orengo, and M. Koltzenburg, “Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma,” *BMC Genomics*, vol. 7, no. 1, p. 252, 2006.
- [27] Y. Zhou, B. Zhou, L. Pache et al., “Metascape provides a biologist-oriented resource for the analysis of systems-level datasets,” *Nature Communications*, vol. 10, no. 1, p. 1523, 2019.
- [28] X. Feng and Y. Chen, “Drug delivery targets and systems for targeted treatment of rheumatoid arthritis,” *Journal of Drug Targeting*, vol. 26, no. 10, pp. 845–857, 2018.
- [29] P. J. Richards, M. A. Nowell, S. Horiuchi et al., “Functional characterization of a soluble gp130 isoform and its therapeutic capacity in an experimental model of inflammatory arthritis,” *Arthritis and Rheumatism*, vol. 54, no. 5, pp. 1662–1672, 2006.
- [30] J. P. van Hamburg, A. M. Mus, M. J. de Bruijn et al., “GATA-3 protects against severe joint inflammation and bone erosion and reduces differentiation of Th17 cells during experimental arthritis,” *Arthritis and Rheumatism*, vol. 60, no. 3, pp. 750–759, 2009.
- [31] S. Sumitomo, K. Fujio, T. Okamura, and K. Yamamoto, “Egr2 and Egr3 are the unique regulators for systemic autoimmunity,” *Jakstat*, vol. 2, no. 2, p. e23952, 2013.
- [32] J. Cui, S. Saevarsdottir, B. Thomson et al., “Rheumatoid arthritis risk allele PTPRC is also associated with response to anti-tumor necrosis factor α therapy,” *Arthritis and Rheumatism*, vol. 62, no. 7, pp. 1849–1861, 2010.
- [33] C. S. Jürgensen, G. Levantino, G. Houen et al., “Determination of autoantibodies to annexin XI in systemic autoimmune diseases,” *Lupus*, vol. 9, no. 7, pp. 515–520, 2000.
- [34] S. M. Mahendran, E. C. Keystone, R. J. Krawetz, K. Liang, E. P. Diamandis, and V. Chandran, “Elucidating the endogenous synovial fluid proteome and peptidome of inflammatory arthritis using label-free mass spectrometry,” *Clinical Proteomics*, vol. 16, no. 1, pp. 23–23, 2019.
- [35] L. M. Elsby, G. Orozco, J. Denton, J. Worthington, D. W. Ray, and R. P. Donn, “Functional evaluation of TNFAIP3 (A20) in rheumatoid arthritis,” *Clinical and Experimental Rheumatology*, vol. 28, no. 5, pp. 708–714, 2010.
- [36] O. Perwitasari, S. Johnson, X. Yan et al., “Verdinexor, a novel selective inhibitor of nuclear export, reduces influenza A virus replication in vitro and in vivo,” *Journal of Virology*, vol. 88, no. 17, pp. 10228–10243, 2014.
- [37] C. J. Malemud, “Intracellular signaling pathways in rheumatoid arthritis,” *Journal of clinical & cellular immunology*, vol. 4, no. 4, p. 160, 2013.
- [38] I. B. McInnes, C. D. Buckley, and J. D. Isaacs, “Cytokines in rheumatoid arthritis – shaping the immunological landscape,” *Nature Reviews Rheumatology*, vol. 12, no. 1, pp. 63–68, 2016.
- [39] I. B. McInnes and G. Schett, “The pathogenesis of rheumatoid arthritis,” *The New England Journal of Medicine*, vol. 365, no. 23, pp. 2205–2219, 2011.
- [40] L. L. Gompels, N. M. Malik, L. Madden et al., “Human epidermal growth factor receptor bispecific ligand trap RB200: abrogation of collagen-induced arthritis in combination with tumour necrosis factor blockade,” *Arthritis Research & Therapy*, vol. 13, no. 5, p. R161, 2011.
- [41] M. Ozgen, S. S. Koca, A. Karatas et al., “Lapatinib ameliorates experimental arthritis in rats,” *Inflammation*, vol. 38, no. 1, pp. 252–259, 2015.
- [42] Y. J. Li, A. Kukita, T. Watanabe et al., “Nordihydroguaiaretic acid inhibition of NFATc1 suppresses osteoclastogenesis and arthritis bone destruction in rats,” *Laboratory Investigation*, vol. 92, no. 12, pp. 1777–1787, 2012.
- [43] A. Kaneko, “Tocilizumab in rheumatoid arthritis: efficacy, safety and its place in therapy,” *Therapeutic advances in chronic disease*, vol. 4, no. 1, pp. 15–21, 2013.
- [44] E. G. Boyce, E. L. Rogan, D. Vyas, N. Prasad, and Y. Mai, “Sarilumab: review of a second IL-6 receptor antagonist indicated for the treatment of rheumatoid arthritis,” *The Annals of Pharmacotherapy*, vol. 52, no. 8, pp. 780–791, 2018.
- [45] C. Li, J. Zhang, W. Wang, H. Wang, Y. Zhang, and Z. Zhang, “Arsenic trioxide improves Treg and Th17 balance by modulating STAT3 in treatment-naive rheumatoid arthritis patients,” *International Immunopharmacology*, vol. 73, pp. 539–551, 2019.
- [46] M. Pajares, A. I Rojo, G. Manda, L. Boscá, and A. Cuadrado, “Inflammation in Parkinson’s disease: mechanisms and therapeutic implications,” *Cells*, vol. 9, no. 7, 2020.
- [47] A. W. Morgan, J. I. Robinson, P. G. Conaghan et al., “Evaluation of the rheumatoid arthritis susceptibility loci HLA-DRB1, PTPN22, OLIG3/TNFAIP3, STAT4 and TRAF1/C5 in an inception cohort,” *Arthritis Research & Therapy*, vol. 12, no. 2, p. R57, 2010.
- [48] K. D. Deane, M. K. Demoruelle, L. B. Kelmenson, K. A. Kuhn, J. M. Norris, and V. M. Holers, “Genetic and environmental risk factors for rheumatoid arthritis,” *Best Practice and Research Clinical rheumatology*, vol. 31, no. 1, pp. 3–18, 2017.
- [49] S. L. Edwards, J. Beesley, J. D. French, and A. M. Dunning, “Beyond GWASs: illuminating the dark road from association to function,” *American Journal of Human Genetics*, vol. 93, no. 5, pp. 779–797, 2013.
- [50] M. T. Maurano, R. Humbert, E. Rynes et al., “Systematic localization of common disease-associated variation in regulatory DNA,” *Science*, vol. 337, no. 6099, pp. 1190–1195, 2012.
- [51] T. Pap, M. Nawrath, J. Heinrich et al., “Cooperation of Ras- and c-Myc-dependent pathways in regulating the growth and invasiveness of synovial fibroblasts in rheumatoid arthritis,” *Arthritis and Rheumatism*, vol. 50, no. 9, pp. 2794–2802, 2004.

- [52] Y. Guo, A. M. Walsh, U. Fearon et al., “CD40L-dependent pathway is active at various stages of rheumatoid arthritis disease progression,” *Journal of Immunology*, vol. 198, no. 11, pp. 4490–4501, 2017.
- [53] A. Gierut, H. Perlman, and R. M. Pope, “Innate immunity and rheumatoid arthritis,” *Rheumatic Diseases Clinics of North America*, vol. 36, no. 2, pp. 271–296, 2010.
- [54] P. Wehr, H. Purvis, S. C. Law, and R. Thomas, “Dendritic cells, T cells and their interaction in rheumatoid arthritis,” *Clinical and Experimental Immunology*, vol. 196, no. 1, pp. 12–27, 2019.
- [55] Y. Zhang, T. Liu, X. Hu et al., “CellCall: integrating paired ligand–receptor and transcription factor activities for cell–cell communication,” *Nucleic Acids Research*, vol. 49, no. 15, pp. 8520–8534, 2021.
- [56] Y. Zhang, T. Liu, J. Wang et al., “Cellinker: a platform of ligand–receptor interactions for intercellular communication analysis,” *Bioinformatics*, vol. 37, no. 14, pp. 2025–2032, 2021.
- [57] O. O. Olumuyiwa-Akeredolu and E. Pretorius, “Platelet and red blood cell interactions and their role in rheumatoid arthritis,” *Rheumatology International*, vol. 35, no. 12, pp. 1955–1964, 2015.
- [58] L. J. O’Neil and M. J. Kaplan, “Neutrophils in rheumatoid arthritis: breaking immune tolerance and fueling disease,” *Trends in Molecular Medicine*, vol. 25, no. 3, pp. 215–227, 2019.

Research Article

Identification of *Helicobacter pylori* Membrane Proteins Using Sequence-Based Features

Mujiexin Liu ¹, Hui Chen ², Dong Gao ³, Cai-Yi Ma ³, and Zhao-Yue Zhang ^{2,3}

¹Ineye Hospital of Chengdu University of TCM, Chengdu University of TCM, Chengdu 610084, China

²School of Healthcare Technology, Chengdu Neusoft University, 611844 Chengdu, China

³School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China

Correspondence should be addressed to Zhao-Yue Zhang; zyzhang@uestc.edu.cn

Received 7 November 2021; Accepted 16 December 2021; Published 12 January 2022

Academic Editor: Balachandran Manavalan

Copyright © 2022 Mujiexin Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Helicobacter pylori (*H. pylori*) is the most common risk factor for gastric cancer worldwide. The membrane proteins of the *H. pylori* are involved in bacterial adherence and play a vital role in the field of drug discovery. Thus, an accurate and cost-effective computational model is needed to predict the uncharacterized membrane proteins of *H. pylori*. In this study, a reliable benchmark dataset consisted of 114 membrane and 219 nonmembrane proteins was constructed based on UniProt. A support vector machine- (SVM-) based model was developed for discriminating *H. pylori* membrane proteins from nonmembrane proteins by using sequence information. Cross-validation showed that our method achieved good performance with an accuracy of 91.29%. It is anticipated that the proposed model will be useful for the annotation of *H. pylori* membrane proteins and the development of new anti-*H. pylori* agents.

1. Introduction

Helicobacter pylori (*H. pylori*) is a Gram-negative spiral-shaped bacterium that infects half of the human population worldwide. *H. pylori* causes gastric mucosa damage, chronic inflammation, and dysregulation of the gut community, increasing the risk of gastric cancer [1–3]. Attachment to the gastric mucosa is the first step in establishing bacterial colonization [4]. *H. pylori* membrane proteins such as antigen-binding adhesin (BabA), sialic acid-binding adhesin (SabA), outer inflammatory protein (OipA), and outer membrane protein Q (HopQ) can act as putative virulence factors that mediate the host-pathogen interactions, induce the release of inflammatory cytokines, and enhance the virulence property of the bacterium [4–6]. Thus, the identification of *H. pylori* membrane protein receptors contributes to the design of therapeutic drugs and vaccine development [7, 8].

Although *H. pylori* membrane proteins play a key role in attachment to and entry into host cells, only few have been described so far. There are some efforts in the prediction of membrane proteins [9, 10] for other germs like *Mycobacte-*

rial [11] and *Chlamydiae* [12]. However, there are no machine learning-based approaches for the prediction of the *H. pylori* membrane proteins. In this study, we developed a comprehensive in silico approach for discriminating novel *H. pylori* membrane proteins using amino acid sequence-based criteria. First, the benchmark dataset was constructed based on a reliable source. Second, sequence-based feature encoding methods were used to represent protein sequences. Next, the incremental feature selection (IFS) technique with multiple feature ranking methods was applied to obtain the optimal feature set. Finally, a membrane protein prediction model was established based on the optimal feature set. The workflow can be seen in Figure 1.

2. Materials and Methods

2.1. Benchmark Dataset. An objective and strict benchmark dataset is fundamental for a robust prediction model construction [13–18]. The Universal Protein Resource (UniProt) [19] is a comprehensive resource for proteins and can be freely accessed at <https://www.uniprot.org/>. The

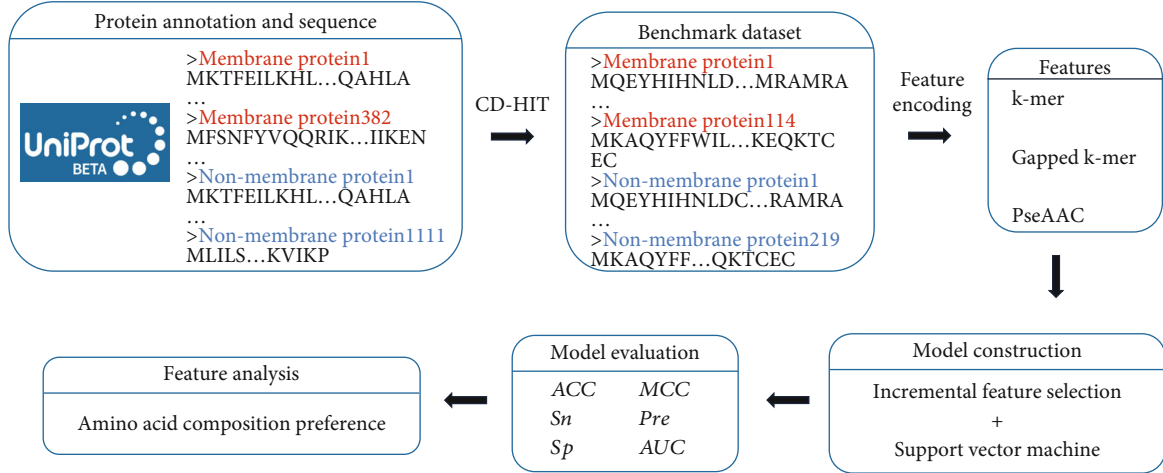


FIGURE 1: The workflow diagram of developing the *H. pylori* membrane protein prediction model.

382 *H. pylori* membrane protein sequences and 1111 nonmembrane protein sequences were obtained from the UniProt. If a sequence contains nonstandard letters, the sequence was removed from the dataset. To avoid the influence of sequence similarity [20], CD-HIT [21] with 0.3 sequence identity was used to exclude highly similar membrane proteins. Finally, 114 (29.8% of the original) membrane proteins and 219 (19.7% of the original) non-membrane proteins remained in the benchmark dataset.

2.2. Feature Encoding. Generally, feature encoding plays a crucial role for machine learning in model construction [22–28]. The feature encoding method determines the degree of sequence information mining. In this work, k -mer amino acid composition [29–31], gapped k -mer method [32], and pseudo-amino acid composition (PseAAC) [33–39] were used to formulate sequences.

Let the protein \mathbf{S} be expressed as follows:

$$\mathbf{S} = R_1 R_2 R_3 R_4 R_5 \cdots R_i R_{i+1} \cdots R_L, \quad (1)$$

where L denotes the length of the protein sequence and R_i is the i -th amino acid.

By using k -mer amino acid composition, a primary protein sequence \mathbf{S} can be transferred into a vector \mathbf{V}_k with 20^k elements according to the following formula:

$$\mathbf{V}_k = \left[f_1^{k\text{-mer}} f_2^{k\text{-mer}} \cdots f_i^{k\text{-mer}} \cdots f_{20^k}^{k\text{-mer}} \right]^T, \quad (2)$$

where the symbol \mathbf{T} means the transposition of a vector and $f_i^{k\text{-mer}}$ is the normalized frequency of the i -th k -mer amino acid component occurring in \mathbf{S} and can be calculated by

$$f_i^{k\text{-mer}} = \frac{n_i}{\sum_{i=1}^{20^k} n_i} = \frac{n_i}{L - k + 1}, \quad (3)$$

where n_i means the number of occurrences of the i -th k -mer amino acid component in the sequence \mathbf{S} .

With the increase of k , one protein sequence may have many k -mers absent, and its feature vector will contain a

large number of zero values. To overcome this sparse problem, gapped k -mer (k -mer with g gap) was used. For example, “GG” with 3 gaps constitute the patterns “GNNNG,” where N represent any kind of amino acid. By using the gapped k -mer method, a primary protein sequence \mathbf{S} can be transferred into a vector \mathbf{V}_g with 20^{k-g} elements according to the following formula:

$$\mathbf{V}_g = \left[f_1^{gk\text{-mer}} f_2^{gk\text{-mer}} \cdots f_i^{gk\text{-mer}} \cdots f_{20^{k-g}}^{gk\text{-mer}} \right]^T, \quad (4)$$

where the $f_i^{gk\text{-mer}}$ is the normalized frequency of the i -th k -mer with g gap amino acid component occurring in \mathbf{S} .

PseAAC can represent a protein sequence in a discrete model without completely losing its sequence-order information. A primary protein sequence \mathbf{S} can be transferred into a vector \mathbf{V}_p with PseAAC according to the following formula:

$$\mathbf{V}_p = [x_1 \cdots x_{20} x_{20+1} \cdots x_{20+\lambda}]^T, \quad (5)$$

$$x_i = \begin{cases} \frac{f_i}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \Theta_j}, & 1 \leq i \leq 20, \\ \frac{\omega \Theta_i - 20}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \Theta_j}, & 20 + 1 \leq i \leq 20 + \lambda, \end{cases} \quad (6)$$

where f_i is the normalized frequency of i -th amino acid, and Θ_j is the j -th sequence correlation factor that can be calculated by the product of the six physicochemical property numerical values between amino acids at different positions. ω is the weight factor for short range and long range.

2.3. Feature Selection and Modeling. To exclude noise and improve computational efficiency, feature selection is an indispensable step [23, 40–45]. Binomial distribution is one of the wonderful feature selection techniques that have been successfully applied in many works [46–48]. The high binomial distribution score indicates that the presence of the k

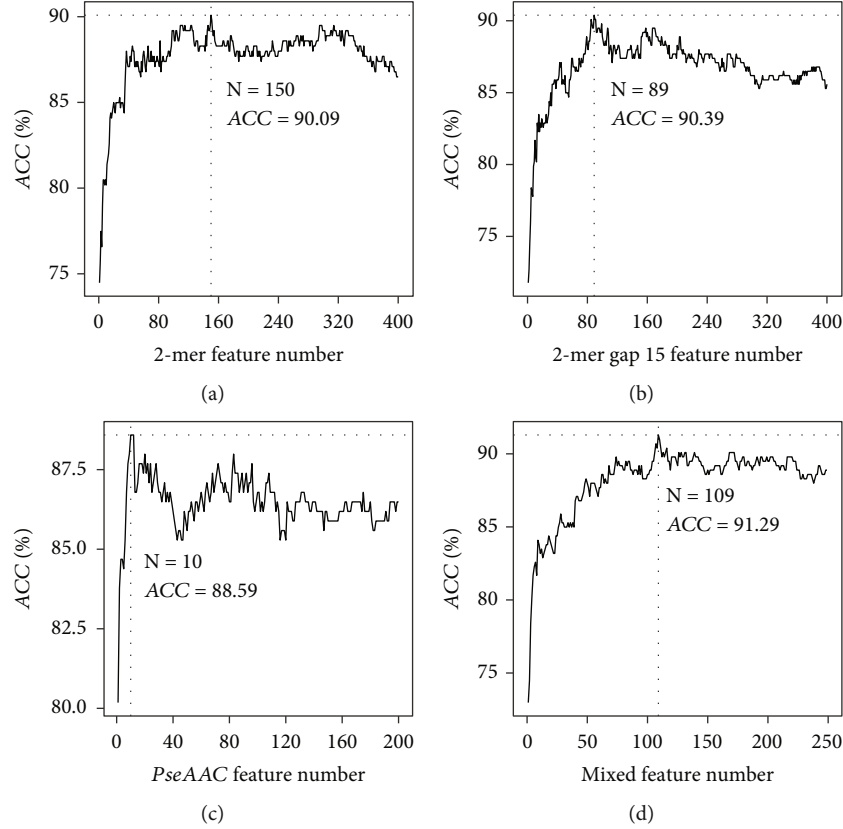


FIGURE 2: The IFS curves for (a) 2-mer features, (b) gapped 2-mer features, (c) PseAAC features, and (d) merged features.

-mer amino acid in a membrane protein sequence is not accidental. Analysis of variance (ANOVA) tests the ratio of the variance between groups and the variance within the groups to analyse the differences among group means [30]. The high ANOVA score means there is a big feature difference between the membrane protein group and the non-membrane protein group. In this study, binomial distribution was used on k -mer features, and ANOVA was used on gapped k -mer and PseAAC features to winnow out the irrelevant features. Then, ANOVA was used to re prune all the redundant features.

After ranking the features according to their statistical scores, the IFS strategy with support vector machine (SVM) was adopted to determine the optimal feature set [49–53]. SVM is a classification algorithm that finds the optimal classification hyperplane in the high-dimensional feature space. The IFS strategy added features one by one to the feature set from a higher-ranked to a lower-ranked score. Once a new feature set was composed, LIBSVM [54] with 5-fold cross-validation was performed to train and test prediction models. The optimal feature set is defined based on the principle that the prediction model based on such features could achieve maximum accuracy. Finally, an SVM model was constructed based on the optimal feature subset for the membrane protein prediction.

2.4. Performance Evaluation Metrics. In order to assess the capability of the binary prediction method, six indexes, namely, accuracy (ACC), sensitivity (Sn), specificity (Sp),

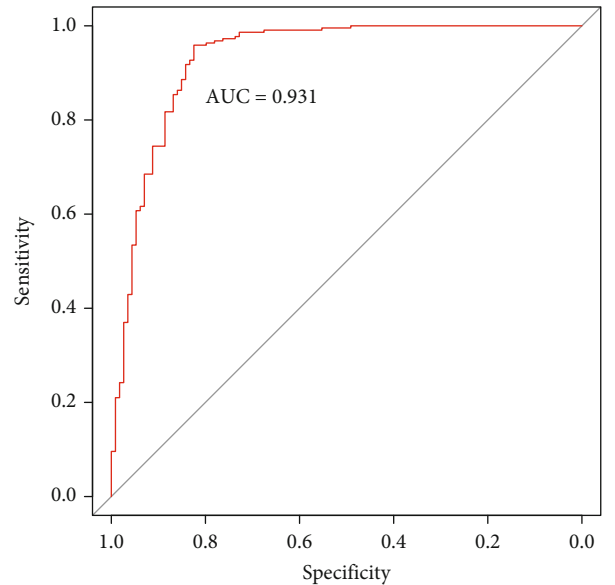


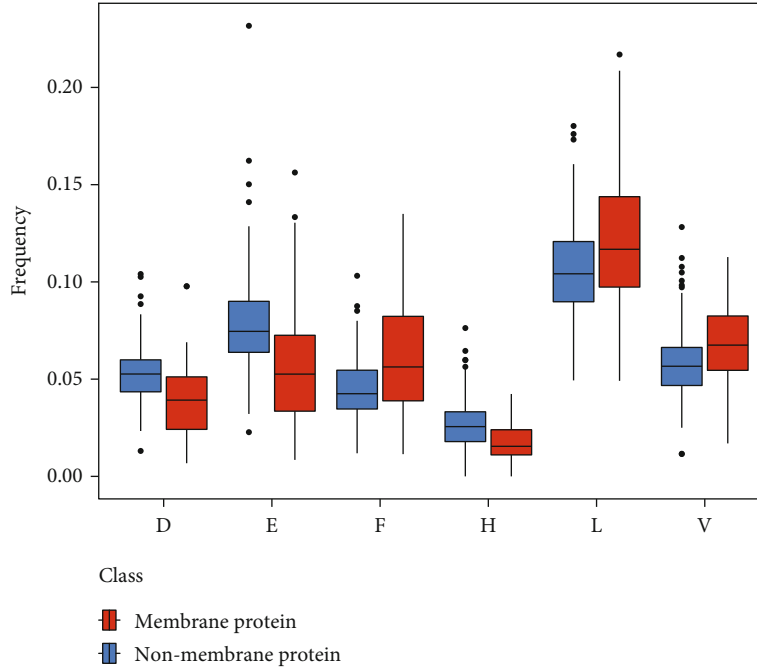
FIGURE 3: The ROC curves of the 5-fold cross-validation test.

precision (Pre), Matthew's correlation coefficient (MCC), and the area under the receiver operating characteristic curve (AUC) [55–60], were used and formulated as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad (7)$$



(a)



(b)

FIGURE 4: (a) The heat map of AAC of the model features. (b) The frequency of the six amino acids in the two classes.

$$Sn = \frac{TP}{TP + FN}, \quad (8)$$

$$Sp = \frac{TN}{TN + FP}, \quad (9)$$

$$Pre = \frac{TP}{TP + FP}, \quad (10)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}, \quad (11)$$

where TP (true positive) and TN (true negative) present the numbers of correctly identified membrane proteins and nonmembrane proteins, respectively. FP (false positive) and FN (false negative) denote the number of nonmembrane proteins incorrectly classified as membrane proteins and the number of membrane proteins incorrectly classified as nonmembrane proteins, respectively. Receiver operating characteristics (ROC) analysis was used to measure

the performance of the model with the varying decision thresholds [61–63]. Due to the small sample size, the result of the 5-fold cross-validation was used to evaluate the model performance.

3. Results and Discussion

3.1. Feature Optimization. As shown in equations (3), (4), and (5), the description of the protein sequences depends on parameters k , g , ω , and λ . For k -mer feature encoding, $k = 2, 3, 4$ was tried in this study. The model achieved the best accuracy of 90.09% with the top 150 binomial distribution-ranked 2-mer features (Figure 2(a)). For gapped k -mer feature encoding, we set $k = 2$ and traverse g from 1 to 20, when $g = 15$, and the model achieved the best accuracy of 90.39% with the top 89 ANOVA-ranked features (Figure 2(b)). For PseAAC, we set the weight factor $\omega = 0.5$ and parameter λ from 1 to 70 with step size 5, and the best performance achieved was 88.59% when the λ is 20 and feature number is 10 (Figure 2(c)). To represent the sequence

information comprehensively, all best feature subsets were merged and ranked by ANOVA. IFS was performed again to filter out the redundant features. As we can see in Figure 2(d), the model achieved the best accuracy of 91.29% when the top 109 ANOVA-ranked features were used to train the model.

3.2. Model Construction and Evaluation. Finally, 109 features were used to construct the SVM-based model for the prediction of membrane proteins. And the soft margin SVM penalty coefficient c and Gaussian kernel function width parameter γ are 0.5.

To show the prediction capability of the final model, six evaluation metrics were calculated based on the result of the 5-fold cross-validation. The model achieved the ACC of 91.29%, Sn of 82.46%, Sp of 95.9%, Pre of 91.26%, and MCC of 0.804. We also drew the ROC curve in Figure 3. It shows that the AUC reaches the value of 0.931, suggesting that the proposed model has an excellent prediction capability on membrane protein classification.

3.3. Amino Acid Composition (AAC) of Optimal Features. The AAC of the model features was used to analyse the preference of membrane proteins for specific amino acids. Among the optimal feature set, there are 83 2-mer features, 16 gapped 2-mer features, and 10 PseAAC features. Focusing on the 2-mer and gapped 2-mer features, we found that the occurrence of leucine (L), glutamic acid (E), aspartic acid (D), phenylalanine (F), valine (V), and histidine (H) exceeds 50% of the total (Figure 4(a)). And the frequencies of F, L, and V in membrane protein sequences are significantly higher than those in nonmembrane protein sequences ($p < 0.001$). In contrast, the frequencies of D, E, and H in nonmembrane protein sequences are significantly higher than those in membrane proteins ($p < 0.001$) (Figure 4(b)).

4. Conclusions

H. pylori membrane proteins are an important class of molecules that play key roles in host-pathogen interactions. However, it is a new area in the prediction of *H. pylori* membrane proteins with machine learning methods. Hence, we developed an *H. pylori* membrane proteins predictor on the basis of sequence-based information. The model will powerfully support the discovery of *H. pylori* membrane proteins and the research of *H. pylori* infection. It has the potential to be significant in novel vaccine candidate antigens and drug development [64, 65]. In the future, we will stay focused on the *H. pylori* membrane protein prediction issues and screen the possible vaccine candidates and drug targets. Moreover, we will collect more data to train a deep learning model [66–71] to improve prediction performance.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62102067).

References

- [1] Z. Li, T. Zhang, H. Lei et al., “Research on gastric cancer’s drug-resistant gene regulatory network model,” *Current Bioinformatics*, vol. 15, no. 3, pp. 225–234, 2020.
- [2] L. Cheng, C. Qi, H. Zhuang, T. Fu, and X. Zhang, “gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions,” *Nucleic Acids Research*, vol. 48, no. D1, pp. D554–D560, 2020.
- [3] L. Cheng, C. Qi, H. Yang et al., “gutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites,” *Nucleic Acids Research*, 2021.
- [4] Y. Matsuo, Y. Kido, and Y. Yamaoka, “Helicobacter pylori outer membrane protein-related pathogenesis,” *Toxins (Basel)*, vol. 9, no. 3, p. 101, 2017.
- [5] S. Ansari, E. T. Kabamba, P. K. Shrestha et al., “Helicobacter pylori Bab characterization in clinical isolates from Bhutan, Myanmar, Nepal and Bangladesh,” *PLoS One*, vol. 12, no. 11, article e0187225, 2017.
- [6] M. Sukanuma, M. Kurusu, S. Okabe et al., “Helicobacter pylori membrane protein 1: a new carcinogenic factor of Helicobacter pylori,” *Cancer Research*, vol. 61, no. 17, pp. 6356–6359, 2001.
- [7] Y. Yamaoka, O. Ojo, S. Fujimoto et al., “Helicobacter pylori outer membrane proteins and gastroduodenal disease,” *Gut*, vol. 55, no. 6, pp. 775–781, 2006.
- [8] L. Yu, M. Xia, and Q. An, “A network embedding framework based on integrating multiplex network for drug combination prediction,” *Briefings in Bioinformatics*, 2021.
- [9] M. Kabir, M. Arif, F. Ali, S. Ahmad, Z. N. K. Swati, and D. J. Yu, “Prediction of membrane protein types by exploring local discriminative information from evolutionary profiles,” *Analytical Biochemistry*, vol. 564–565, pp. 123–132, 2019.
- [10] Y. C. Zuo, W. X. Su, S. H. Zhang et al., “Discrimination of membrane transporter protein types using K-nearest neighbor method derived from the similarity distance of total diversity measure,” *Molecular BioSystems*, vol. 11, no. 3, pp. 950–957, 2015.
- [11] C. Ding, L. F. Yuan, S. H. Guo, H. Lin, and W. Chen, “Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions,” *Journal of Proteomics*, vol. 77, pp. 321–328, 2012.
- [12] E. Heinz, P. Tischler, T. Rattei, G. Myers, M. Wagner, and M. Horn, “Comprehensive in silico prediction and analysis of chlamydial outer membrane proteins reflects evolution and life style of the Chlamydiae,” *BMC Genomics*, vol. 10, p. 634, 2009.
- [13] D. Zhang, H.-D. Chen, H. Zulfiqar et al., “iBLP: an XGBoost-based predictor for identifying bioluminescent proteins,” *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021.

- [14] W. Su, M. L. Liu, Y. H. Yang et al., “PPD: a manually curated database for experimentally verified prokaryotic promoters,” *Journal of Molecular Biology*, vol. 433, no. 11, article ???, 2021.
- [15] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, “DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function,” *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, 2018.
- [16] L. Wei, W. He, A. Malik, R. Su, L. Cui, and B. Manavalan, “Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework,” *Briefings in Bioinformatics*, vol. 22, no. 4, 2021.
- [17] M. M. Hasan, M. A. Alam, W. Shoombuatong, H. W. Deng, B. Manavalan, and H. Kurata, “NeuroPred-FRL: an interpretable prediction model for identifying neuropeptide using feature representation learning,” *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.
- [18] P. Charoenkwan, C. Nantasenamat, M. M. Hasan, B. Manavalan, and W. Shoombuatong, “Bert4bitter: a bidirectional encoder representations from transformers (Bert)-based model for improving the prediction of bitter peptides,” *Bioinformatics*, vol. 37, no. 17, pp. 2556–2562, 2021.
- [19] C. UniProt, “UniProt: the universal protein knowledgebase in 2021,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D480–D489, 2021.
- [20] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, “Sequence clustering in bioinformatics: an empirical study,” *Briefings in Bioinformatics*, vol. 21, no. 1, pp. 1–10, 2020.
- [21] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “Cd-Hit: accelerated for clustering the next-generation sequencing data,” *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.
- [22] H. Zulfiqar, Z. J. Sun, Q. L. Huang et al., “Deep-4mCW2V: a sequence-based predictor to identify N4-methylcytosine sites in *Escherichia coli*,” *Methods*, 2021.
- [23] D. Zhang, Z. C. Xu, W. Su et al., “PROBselect: accurate prediction of protein-binding residues from proteins sequences via dynamic predictor selection,” *Bioinformatics*, vol. 36, Supplement_2, pp. i735–i744, 2020.
- [24] H. Yang, Y. Luo, X. Ren et al., “Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators,” *Information Fusion*, vol. 75, pp. 140–149, 2021.
- [25] J. Long, H. Yang, Z. Yang et al., “Integrated biomarker profiling of the metabolome associated with impaired fasting glucose and type 2 diabetes mellitus in large-scale Chinese patients,” *Clinical and Translational Medicine*, vol. 11, no. 6, article e432, 2021.
- [26] H. Lv, F. Y. Dao, Z. X. Guan, H. Yang, Y. W. Li, and H. Lin, “Landscape of cancer diagnostic biomarkers from specifically expressed genes,” *Briefings in Bioinformatics*, vol. 21, no. 6, pp. 2175–2184, 2020.
- [27] L. Yu, M. Wang, Y. Yang et al., “Predicting therapeutic drugs for hepatocellular carcinoma based on tissue-specific pathways,” *PLoS Computational Biology*, vol. 17, no. 2, article e1008696, 2021.
- [28] X. G. Chen, W. W. Shi, and L. Deng, “Prediction of disease comorbidity using HeteSim scores based on multiple heterogeneous networks,” *Current Gene Therapy*, vol. 19, no. 4, pp. 232–241, 2019.
- [29] M. L. Liu, W. Su, J. S. Wang, Y. H. Yang, H. Yang, and H. Lin, “Predicting preference of transcription factors for methylated DNA using sequence information,” *Mol Ther Nucleic Acids*, vol. 22, pp. 1043–1050, 2020.
- [30] H. Tang, Y. W. Zhao, P. Zou et al., “HBPred: a tool to identify growth hormone-binding proteins,” *International Journal of Biological Sciences*, vol. 14, no. 8, pp. 957–964, 2018.
- [31] Y. Zuo, Y. Li, Y. Chen, G. Li, Z. Yan, and L. Yang, “PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition,” *Bioinformatics*, vol. 33, no. 1, pp. 122–124, 2017.
- [32] J. X. Tan, S. H. Li, Z. M. Zhang et al., “Identification of hormone binding proteins based on machine learning methods,” *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 2466–2480, 2019.
- [33] L. Zheng, D. Liu, W. Yang, L. Yang, and Y. Zuo, “Location deviations of DNA functional elements affected SNP mapping in the published databases and references,” *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1293–1301, 2020.
- [34] L. Zheng, S. Huang, N. Mu et al., “RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou’s five-step rule,” *Database: The Journal of Biological Databases and Curation*, vol. 2019, 2019.
- [35] Y. Y. Cao, C. L. Yu, S. H. Huang, S. Y. Wang, Y. C. Zuo, and L. Yang, “Characterization and prediction of presynaptic and postsynaptic neurotoxins based on reduced amino acids and biological properties,” *Current Bioinformatics*, vol. 16, no. 3, pp. 364–370, 2021.
- [36] H. B. Shen and K. C. Chou, “PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition,” *Analytical Biochemistry*, vol. 373, no. 2, pp. 386–388, 2008.
- [37] S. Naseer, W. Hussain, Y. D. Khan, and N. Rasool, “Sequence-based identification of arginine amidation sites in proteins using deep representations of proteins and PseAAC,” *Current Bioinformatics*, vol. 15, no. 8, pp. 937–948, 2021.
- [38] M. A. M. Hasan, M. K. Ben Islam, J. Rahman, and S. Ahmad, “Citruination site prediction by incorporating sequence coupled effects into PseAAC and resolving data imbalance issue,” *Current Bioinformatics*, vol. 15, no. 3, pp. 235–245, 2020.
- [39] S. Amanat, A. Ashraf, W. Hussain, N. Rasool, and Y. D. Khan, “Identification of lysine carboxylation sites in proteins by integrating statistical moments and position relative features via general PseAAC,” *Current Bioinformatics*, vol. 15, no. 5, pp. 396–407, 2020.
- [40] X. Han, Q. Kong, C. Liu, L. Cheng, and J. Han, “Subtypedrug: a software package for prioritization of candidate cancer subtype-specific drugs,” *Bioinformatics*, vol. 37, no. 16, pp. 2491–2493, 2021.
- [41] Y. Sheng, Y. Jiang, Y. Yang et al., “Selecting gene features for unsupervised analysis of single-cell gene expression data,” *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.
- [42] W. Yang, X. J. Zhu, J. Huang, H. Ding, and H. Lin, “A brief survey of machine learning methods in protein sub-Golgi localization,” *Current Bioinformatics*, vol. 14, pp. 234–240, 2019.
- [43] S. He, F. Guo, Q. Zou, and H. Ding, “MRMD2.0: a Python tool for machine learning with feature ranking and reduction,” *Current Bioinformatics*, vol. 15, no. 10, pp. 1213–1221, 2021.
- [44] X. Wu and L. Yu, *EPSOL: sequence-based protein solubility prediction using multidimensional embedding*, Bioinformatics, Oxford, England, 2021.

- [45] J. W. Li, X. Y. Wang, N. Li et al., “Feasibility of mesenchymal stem cell therapy for Covid-19: a mini review,” *Current Gene Therapy*, vol. 20, no. 4, pp. 285–288, 2020.
- [46] Z. Y. Zhang, Y. H. Yang, H. Ding, D. Wang, W. Chen, and H. Lin, “Design powerful predictor for mRNA subcellular location prediction in Homo sapiens,” *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 526–535, 2021.
- [47] C. Q. Feng, Z. Y. Zhang, X. J. Zhu et al., “iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators,” *Bioinformatics*, vol. 35, no. 9, pp. 1469–1477, 2019.
- [48] H. Wang, P. Liang, L. Zheng, C. Long, H. Li, and Y. Zuo, “Correction to: ncDLRES: a novel method for non-coding RNAs family prediction based on dynamic LSTM and ResNet,” *Bioinformatics*, vol. 22, no. 1, 2021.
- [49] F. Y. Dao, H. Lv, F. Wang et al., “Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique,” *Bioinformatics*, vol. 35, no. 12, pp. 2075–2083, 2019.
- [50] C. Ao, L. Yu, and Q. Zou, “Prediction of bio-sequence modifications and the associations with diseases,” *Briefings in Functional Genomics*, vol. 20, no. 1, pp. 1–18, 2021.
- [51] S. Basith, G. Lee, and B. Manavalan, “Stallion: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction,” *Briefings in Bioinformatics*, 2021.
- [52] S. Basith, M. M. Hasan, G. Lee, L. Wei, and B. Manavalan, “Integrative machine learning framework for the identification of cell-specific enhancers from the human genome,” *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.
- [53] M. M. Hasan, N. Schaduagr, S. Basith, G. Lee, W. Shoombuatong, and B. Manavalan, “HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation,” *Bioinformatics*, vol. 36, no. 11, pp. 3350–3356, 2020.
- [54] C. C. Chang and C. J. Lin, “LIBSVM,” *ACM transactions on intelligent systems and technology*, vol. 2, no. 3, pp. 1–27, 2011.
- [55] B. Manavalan, T. H. Shin, and G. Lee, “PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine,” *Frontiers in Microbiology*, vol. 9, p. 476, 2018.
- [56] H. Tang, R. Z. Cao, W. Wang, T. S. Liu, L. M. Wang, and C. M. He, “A two-step discriminated method to identify thermophilic proteins,” *International Journal of Biomathematics*, vol. 10, no. 4, p. 1750050, 2017.
- [57] L. Cheng, H. Shi, Z. Wang et al., “IntNetLncSim: an integrative network analysis method to infer human lncRNA functional similarity,” *Oncotarget*, vol. 7, no. 30, pp. 47864–47874, 2016.
- [58] F. Mo, Y. Luo, D. A. Fan et al., “Integrated analysis of mRNA-seq and miRNA-seq to identify c-MYC, YAP1 and miR-3960 as major players in the anticancer effects of caffeic acid phenethyl ester in human small cell lung cancer cell line,” *Current Gene Therapy*, vol. 20, no. 1, pp. 15–24, 2020.
- [59] R. G. Govindaraj, S. Subramaniyam, and B. Manavalan, “Extremely-randomized-tree-based prediction of N(6)-methyladenosine sites in *saccharomyces cerevisiae*,” *Current Genomics*, vol. 21, no. 1, pp. 26–33, 2020.
- [60] S. Basith, B. Manavalan, T. Hwan Shin, and G. Lee, “Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening,” *Medicinal Research Reviews*, vol. 40, no. 4, pp. 1276–1314, 2020.
- [61] C. E. Metz, “Basic principles of ROC analysis,” *Seminars in Nuclear Medicine*, vol. 8, no. 4, pp. 283–298, 1978.
- [62] H. Lv, F. Y. Dao, H. Zulfiqar, and H. Lin, “DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of Sars-Cov-2 infection using a deep learning-based approach,” *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.
- [63] Q. An and L. Yu, “A heterogeneous network embedding framework for predicting similarity-based drug-target interactions,” *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.
- [64] D. Liu, G. Li, and Y. Zuo, “Function determinants of Tet proteins: the arrangements of sequence motifs with specific codes,” *Briefings in Bioinformatics*, vol. 20, no. 5, pp. 1826–1835, 2019.
- [65] B. F. Xu, D. Y. Liu, Z. R. Wang, R. X. Tian, and Y. C. Zuo, “Multi-substrate selectivity based on key loops and non-homologous domains: new insight into ALKBH family,” *Cellular and Molecular Life Sciences*, vol. 78, no. 1, pp. 129–141, 2021.
- [66] D. Wang, Z. Zhang, Y. Jiang et al., “DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism,” *Nucleic Acids Research*, vol. 49, no. 8, article e46, 2021.
- [67] F. Y. Dao, H. Lv, W. Su, Z. J. Sun, Q. L. Huang, and H. Lin, “iDHS-Deep: an integrated tool for predicting DNase I hypersensitive sites by deep neural network,” *Briefings in Bioinformatics*, vol. 22, no. 5, 2021.
- [68] Y. Zhang, J. Yan, S. Chen et al., “Review of the applications of deep learning in bioinformatics,” *Current Bioinformatics*, vol. 15, no. 8, pp. 898–911, 2020.
- [69] F. Cui, Z. Zhang, and Q. Zou, “Sequence representation approaches for sequence-based protein prediction tasks that use deep learning,” *Briefings in Functional Genomics*, vol. 20, no. 1, pp. 61–73, 2021.
- [70] X. Peng, L. Chen, and J.-P. Zhou, “Identification of carcinogenic chemicals with network embedding and deep learning methods,” *Current Bioinformatics*, vol. 15, no. 9, pp. 1017–1026, 2021.
- [71] Z. B. Lv, C. Y. Ao, and Q. Zou, “Protein function prediction: from traditional classifier to deep learning,” *Proteomics*, vol. 19, no. 14, p. 2, 2019.

Research Article

Dysregulation of Circadian Clock Genes as Significant Clinic Factor in the Tumorigenesis of Hepatocellular Carcinoma

Youfang Liang,¹ Shaoxiang Wang,¹ Xin Huang,² Ruihuan Chai,¹ Qian Tang,³ Rong Yang,¹ Xiaoqing Huang,¹ Xiao Wang^{ID,3} and Kai Zheng^{ID,1}

¹School of Pharmaceutical Sciences, Health Science Center, Shenzhen University, Shenzhen 518060, China

²Shenzhen Key Laboratory for Systemic Aging and Intervention, National Engineering Research Center for Biotechnology (Shenzhen), Medical Research Center, Shenzhen University Health Science Center, Shenzhen 518055, China

³Department of Pharmacy, The Second Clinical Medical College (Shenzhen People's Hospital), Jinan University, Shenzhen, China

Correspondence should be addressed to Xiao Wang; wangxiao0719@163.com and Kai Zheng; zhengk@szu.edu.cn

Received 16 September 2021; Accepted 9 October 2021; Published 29 October 2021

Academic Editor: Hui Ding

Copyright © 2021 Youfang Liang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Hepatocellular carcinoma (HCC) is the leading cause of cancer-related mortality worldwide due to its asymptomatic onset and poor survival rate. This highlights the urgent need for developing novel diagnostic markers for early HCC detection. The circadian clock is important for maintaining cellular homeostasis and is tightly associated with key tumorigenesis-associated molecular events, suggesting the so-called chronotherapy. An analysis of these core circadian genes may lead to the discovery of biological markers signaling the onset of the disease. In this study, the possible functions of 13 core circadian clock genes (CCGs) in HCC were systematically analyzed with the aim of identifying ideal biomarkers and therapeutic targets. Profiles of HCC patients with clinical and gene expression data were downloaded from The Cancer Genome Atlas and International Cancer Genome Consortium. Various bioinformatics methods were used to investigate the roles of circadian clock genes in HCC tumorigenesis. We found that patients with high *TIMELESS* expression or low *CRY2*, *PER1*, and *RORA* expressions have poor survival. Besides, a prediction model consisting of these four CCGs, the tumor-node-metastasis (TNM) stage, and sex was constructed, demonstrating higher predictive accuracy than the traditional TNM-based model. In addition, pathway analysis showed that these four CCGs are involved in the cell cycle, PI3K/AKT pathway, and fatty acid metabolism. Furthermore, the network of these four CCGs-related coexpressed genes and immune infiltration was analyzed, which revealed the close association with B cells and nTreg cells. Notably, *TIMELESS* exhibited contrasting effects against *CRY2*, *PER1*, and *RORA* in most situations. In sum, our works revealed that these circadian clock genes *TIMELESS*, *CRY2*, *PER1*, and *RORA* can serve as potential diagnostic and prognostic biomarkers, as well as therapeutic targets, for HCC patients, which may promote HCC chronotherapy by rhythmically regulating drug sensitivity and key cellular signaling pathways.

1. Introduction

Liver cancer is the sixth most common type of cancer and the fourth highest cause of cancer-associated death globally [1]. Hepatocellular carcinoma (HCC) accounts for 85–90% of all primary liver cancers with increased incidence and mortality [2]. Although there are several therapeutic treatments of HCC, including surgery, radiotherapy, and chemotherapy, the five-year survival of HCC patients remains low primarily due to the delayed diagnoses [3]. Alpha-fetoprotein (AFP) is a tumor marker

commonly used for diagnosing patients with HCC. However, the lack of specificity and accuracy limits its application for early-stage HCC detection. Therefore, it is urgent to search for novel biomarkers to facilitate early detection of HCC and improve the clinical survival rate of HCC patients.

Previous research has demonstrated the link between the circadian clock and key tumorigenesis-associated molecular events [4], suggesting the so-called chronotherapy [5]. The circadian clock is an internal timing system that adjusts behaviors and rhythm according to

geophysical time. Similarly, the mammalian circadian clock describes an internal timekeeping mechanism regulating physiology and behavior [6]. A set of core “clock genes” that form a feedback loop of gene transcription and translation has been identified to generate circadian rhythms in cells. The key “positive” transcriptional regulators CLOCK and BMAL1 bind to E-box regulatory elements and transactivate the transcription of the “negative” elements PERs and CRYs, as well as multiple other rhythmically expressed genes.

Conversely, PER and CRY act as repressors to inhibit the CLOCK : BMAL1 complex. Notably, by rhythmically transcriptionally regulating the gene expression and gene activity throughout the genome, circadian clock genes play critical roles in biological processes such as apoptosis, cellular senescence, DNA damage repair, and metastasis [7]. Accumulating evidence has shown the importance of circadian clock genes in the diagnosis, therapy, and prognosis of different kinds of cancers. For instance, the expression alterations of most circadian clock genes were associated with overall survival, tumor-node-metastasis stage, and cellular sensitivity to anticancer drugs [8]. Besides, PER1 and CLOCK were reported as potential biomarkers for head and neck squamous cell carcinoma [9], whereas PER2 was reported to be associated with vital tumor-related genes in oral cancer [10]. Until now, little is known about the roles of circadian clock genes in HCC.

Herein, we systematically characterized the expression pattern of core circadian clock genes, including *ARNTL*, *CLOCK*, *CRY1*, *CRY2*, *DBP*, *NPAS2*, *NR1D1*, *NR1D2*, *PER1*, *PER2*, *PER3*, *RORA*, and *TIMELESS*, and their clinical significances in HCC. The expression and clinical information profiles were extracted from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) databases. Various bioinformatics methods were applied to analyze the data to screen vital hits possibly involved in the development of HCC. We also established a prediction model with high performance to predict the overall survival of HCC patients. Moreover, we comprehensively analyzed the mutation, drug sensitivity, immune infiltration, key cellular signaling pathway, and coexpression network of circadian clock genes in the HCC tumor microenvironment.

2. Materials and Methods

2.1. Patient Data. The gene expression profiles and clinical information of HCC patients were downloaded from TCGA (<https://portal.gdc.cancer.gov/>) and ICGC (<https://dcc.icgc.org/>) databases, containing 50 normal and 374 tumor samples (TCGA) and 202 normal and 240 tumor samples (ICGC), respectively. Univariate and multivariate Cox regression analyses were performed to investigate the correlation between clinicopathological characteristics and overall survival (OS) by R software (4.0.2).

2.2. Analysis of Differential Expressed Gene. To investigate the expression difference of circadian clock genes between the tumor and normal samples, 374/424 of tumor samples

from TCGA and 240/442 of tumor samples from ICGC were analyzed using the ‘edgeR’ package and ‘limma’ package, respectively. Log2 fold change (logFC), *P* value, and false discovery rate (FDR) were calculated. Genes with $P < 0.05$ and $FDR < 0.05$ were regarded as differentially expressed genes (DEGs). The expression difference of each gene was shown by boxplots. Besides, a Venn diagram was drawn to show the overlapping genes which represent similar expression tendency in all HCC cases.

2.3. Validation of DEGs between HCC and Normal Liver Tissues. Methylation and copy number variation (CNV) analysis were performed to validate the differentially expressed genes between normal liver tissues and tumor tissues. Student’s *t*-test was used to analyze the methylation difference between the normal and tumor samples. The correlation between gene CNV and mRNA expression in HCC was also built. A Venn diagram was drawn to present circadian clock genes regulated by both methylation and CNV. The Human Protein Atlas (HPA) (<https://www.proteinatlas.org/>) database was used to validate the protein expression of DEGs between normal liver tissues and HCC tissues.

2.4. Survival Analysis. After dividing patients into the high- and low-expression groups, survival curves were drawn according to the Kaplan-Meier method by ‘survival’ package in R software, with significance set at $P < 0.05$. Besides, the receiver operating characteristic (ROC) curves were generated to determine the survival parameters, while the area under the curve (AUC) value determined the prognostic performance of the survival model. In addition, to further verify the result of survival analysis, the hazard ratio (HR) and *P* value of circadian clock genes were calculated through the univariate Cox regression based on the gene expression and overall survival.

2.5. Prognosis Prediction Models. Prediction models were used to predict the prognosis of HCC patients based on survival analysis. Through a stepwise multivariate Cox hazard regression analysis, a four-gene model was established. The risk score of each HCC patient was calculated by the following formula:

$$\text{Risk score} = \sum_{i=1}^n \text{Coef}_i \times \text{Exp}_i, \quad (1)$$

where n , Coef, and Exp represent the number of included circadian clock genes, the coefficient of each gene, and the gene expression level, respectively. The ROC curve was then constructed for the cohorts from TCGA and ICGC. The AUC representing the predictability of 3-year survival was also calculated by the ‘survival ROC’ package. When the AUC value was >0.6 , the prediction method was considered reliable. Furthermore, the HCC patients were grouped into the high-risk and low-risk groups according to the median risk score, and the survival curve was then obtained.

2.6. Weighted Gene Coexpression Network Analysis (WGCNA). WGCNA was performed to construct a gene coexpression network, aimed at finding genes coexpressing with circadian clock genes in HCC tissues. The coexpression network was drawn using Cytoscape software (version 3.8.0).

2.7. Immune Infiltrate Analysis. The connection between the gene expression and immune cell infiltration in each sample was evaluated by Immune Cell Abundance Identifier (ImmuCellAI). ImmuCellAI is a database-derived web tool to estimate the abundance of 24 immune cells from gene expression datasets, including RNA-Seq and microarray data, which provides infiltration scores of pancancer.

2.8. Pathway Analysis. The potential mechanism of circadian clock genes was explored by Gene Set Cancer Analysis (<http://bioinfo.life.hust.edu.cn/web/GSCALite/>), which is an online research tool for genomics analysis. A pie chart describes several critical cancer pathways in which the circadian clock genes play different roles. To further determine the underlying mechanism of circadian clock genes, the expression profiles of tumor samples downloaded from TCGA were used to conduct Gene Set Enrichment Analysis (GSEA). Hallmark gene sets (*h*) and Kyoto Encyclopedia of Genes and Genomes gene sets (*c2*) were used as references. A significant enrichment pathway was used to screen which circadian clock genes were upregulated in the high-risk group, with $P < 0.05$ set as the threshold. Furthermore, drug sensitivity analysis was carried out to investigate the correlation between clock genes and anticancer drugs.

3. Results

3.1. Circadian Rhythm of Core Circadian Clock Genes in the Liver. Herein, we investigated the possible roles of 13 core circadian clock genes in HCC, including *ARNTL*, *CLOCK*, *CRY1*, *CRY2*, *DBP*, *NR1D1*, *NR1D2*, *NPAS2*, *PER1*, *PER2*, *PER3*, *RORA*, and *TIMELESS*. The expression profiles of core circadian genes in liver tissue were explored by RNA sequencing at different intervals [11]. The corresponding expression fluctuations of these genes are shown in Figure 1. Apparently, all these genes showed significant circadian rhythms in liver tissue except *TIMELESS*. Besides, *ARNTL* and *CLOCK*, two central circadian clock regulators controlling the circadian rhythm of *PERs*, *CRYs*, *NR1Ds*, *RORA*, *DBP*, and *TIMISS* [6], exhibited the most regular rhythms.

3.2. Clinicopathological Characteristics of the HCC Patients. To investigate the functions of circadian clock genes in HCC, 424 samples from TCGA and 442 samples from ICGC were analyzed by univariate and multivariate Cox regression analyses, respectively. In univariate analysis, the poor overall survival of patients was related to tumor-node-metastasis (TNM) stage and T stage in TCGA. It was significantly associated with TNM stage and sex in ICGC (Tables 1 and 2). Clinicopathological characteristics observed with $P < 0.3$ in the univariate analysis were further screened and used for

multivariate analysis, revealing that sex and TNM stage might be independent prognostic factors for patients with HCC (Table 2).

3.3. Identification of Differentially Expressed Circadian Clock Genes. The differential expression of the circadian clock genes between the tumor and normal samples was described using a boxplot (Figures 2(a) and 2(b)). Besides, the overlapping genes that exhibited similar expression levels in tumor samples from both the TCGA and ICGC databases were shown in a Venn diagram, including *DBP*, *NPAS2*, *PER1*, *RORA*, and *TIMELESS* (Figure 2(c)). Next, we analyzed the copy number variation (CNV) and methylation, two important factors influencing the mRNA expression, of these circadian clock genes. As shown in Figure 2(d), the methylation levels of *CRY2*, *DBP*, and *RORA* were statistically higher in HCC tissues than in normal liver tissues. Besides, most of the circadian clock genes were regulated by methylation except for *ARNTL* and *PER1* (Figure 2(e)). The result of the CNV analysis indicated that the mRNA expressions of all circadian clock genes, except for *DBP* and *NPAS2*, were regulated by copy number variation (Figure 2(f)). Moreover, a Venn diagram was drawn to demonstrate that these genes were regulated by both methylation and CNV (Figure 2(g)).

Furthermore, the protein expression levels of *TIMELESS* and *CRY2* were validated using the HPA database. The protein expression level of *TIMELESS* was increased, and that of *CRY2* was decreased in cancerous tissues compared to those in adjacent noncancerous tissues in HCC patients (Fig. S1), which was in agreement with the bioinformatics analysis. Finally, to investigate the interrelationship between circadian clock genes, the Pearson correlation coefficient was applied to draw the correlation coefficient heatmap based on the gene expression profiles. As shown in Figure 2(h), three circadian clock genes *CRY2*, *PER1*, and *RORA*, were positively and closely related to each other, indicating their similar effects on HCC patients. Additionally, the correlation between each gene was investigated by R software (Fig. S2), which further verified the close relationship between *CRY2*, *PER1*, and *RORA*. On the contrast, *TIMELESS* showed a low relevance to the expression of *CRY2*, *PER1*, and *RORA*, which were slightly negatively associated. Indeed, *CRY2*, *PER1*, and *RORA* were downregulated, and *TIMELESS* was upregulated in tumor tissues, suggesting that *TIMELESS* may play a different role in HCC.

3.4. Circadian Clock Genes as Prognostic Biomarkers for HCC Patients. HCC patients were grouped into the high- and low-risk groups according to the expression of the targeted gene. The survival curves of circadian clock genes were plotted using the K-M method (Figures 3(a) and 3(b)). Among 13 circadian clock genes, *CRY2*, *PER1*, *RORA*, and *TIMELESS* were the only four genes associated with the overall survival of HCC patients (Fig. S3). Patients with higher *TIMELESS* expression had poorer overall survival rates ($P = 0.01$ in TCGA and $P = 0.003$ in ICGC). On the

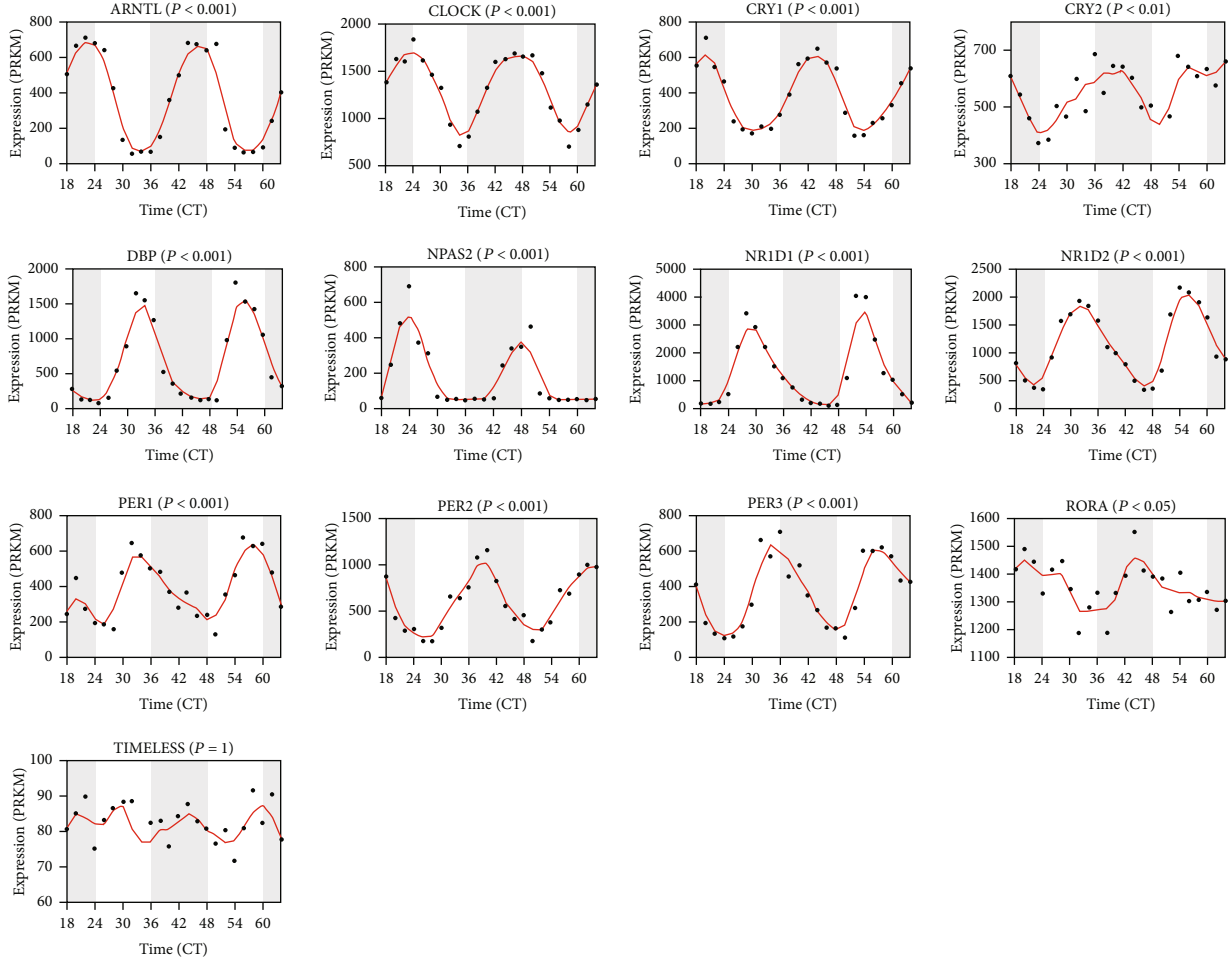


FIGURE 1: Core circadian clock genes in HCC. The circadian rhythm of core circadian genes in HCC, including *ARNTL*, *CRY1*, *CRY2*, *CLOCK*, *DBP*, *NR1D1*, *NR1D2*, *NPAS2*, *PER1*, *PER2*, *PER3*, *RORA*, and *TIMELESS*. RNA-seq data are from ref. [12].

contrary, patients with lower *CRY2*, *PER1*, and *RORA* expressions exhibited poor overall survival rates ($P = 0$, $P = 0.001$, and $P = 0.018$ in TCGA and $P = 0.003$, $P = 0.005$, and $P = 0.004$ in ICGC, respectively). Collectively, these results suggested that *CRY2*, *PER1*, *RORA*, and *TIMELESS* were closely associated with the prognosis of HCC.

3.5. Circadian Clock Gene-Based Prediction Models. Subsequently, a circadian clock gene-based prediction model was established to predict patient survival using the multivariate Cox regression analysis. As shown in Figures 4(a) and 4(b), ROC curves of the single-gene model (*CRY2*, *PER1*, *RORA*, and *TIMELESS*, respectively) showed unsatisfactory predictive effects, with the AUC value of 0.6 approximately (0.63, 0.673, 0.586, and 0.62 in TCGA and 0.641, 0.672, 0.62, 0.696 in ICGC, respectively). Furthermore, the traditional TNM stage-based prediction model was constructed, and it was observed that the AUC value was 0.642 in both TCGA and ICGC, which is nearly equal to the single-gene-based model (Figures 4(c) and 4(d)). In addition, the combinatory prediction models consisting of a single circadian clock gene and the TNM stage were constructed, which still exhibited unsatisfactory prediction (Fig. S4). A four-gene-based pre-

diction model combined with two clinicopathological risk factors, TNM stage and sex, was established to further improve predictive frequency (Figures 4(e) and 4(f)). Risk scores of the patients were calculated according to the following formulas:

$$\begin{aligned} \text{Risk Score (TCGA)} = & (-0.235 * \text{CRY2}_{\text{Exp}}) \\ & + (-0.031 * \text{RORA}_{\text{Exp}}) + (-0.267 * \text{PER1}_{\text{Exp}}) \\ & + (0.077 * \text{TIMELESS}_{\text{Exp}}) + (-0.130 * \text{SEX}) \\ & + (0.905 * \text{TNM}), \end{aligned} \quad (2)$$

$$\begin{aligned} \text{Risk Score (ICGC)} = & (-0.576 * \text{CRY2}_{\text{Exp}}) \\ & + (0.193 * \text{RORA}_{\text{Exp}}) + (-0.236 * \text{PER1}_{\text{Exp}}) \\ & + (0.913 * \text{TIMELESS}_{\text{Exp}}) + (-1.109 * \text{SEX}) \\ & + (1.135 * \text{TNM}). \end{aligned}$$

As a result, the AUC value reached 0.743 in the TCGA database and 0.806 in the ICGC database. Finally, patients were divided into the high-risk and low-risk groups according to the median point, and survival curves were plotted,

TABLE 1: Univariate and multivariate analyses of clinicopathological characteristics for overall survival in HCC patients from the TCGA dataset ($N = 318$).

Variables	n (%)	Univariate analysis		Multivariate analysis	
		HR (95% CI)	P	HR (95% CI)	P
Age					
<60	152 (47.8%)	1 (reference)			
>60	166 (52.2%)	1.173 (0.796-1.730)	0.421		
Sex					
Female	99 (31.1%)	1 (reference)		1 (reference)	
Male	219 (68.9%)	0.804 (0.539-1.198)	0.284	0.864 (0.579-1.287)	0.472
TNM stage					
I+II	237 (73.9%)	1 (reference)		1 (reference)	
III+IV	83 (26.1%)	2.815 (1.909-4.151)	<0.001	1.522 (0.206-11.219)	0.68
Tumor grade					
G1+G2	197 (61.9%)	1 (reference)			
G3+G4	121 (38.1%)	1.077 (0.724-1.603)	0.713		
T stage					
T1+T2	237 (74.5%)	1 (reference)		1 (reference)	
T3+T4	81 (25.5%)	2.839 (1.923-4.189)	<0.001	1.822 (0.247-13.464)	0.556

Note: characteristics with $P < 0.3$ in the univariate analysis were further screened in the multivariate analysis. HR: hazard ratio; CI: confidence interval; TNM stage: tumor-node-metastasis stage; T stage: stage of tumor invasion.

TABLE 2: Univariate and multivariate analyses of clinicopathological characteristics for overall survival in HCC patients from the ICGC dataset ($N = 231$).

Variables	n (%)	Univariate analysis		Multivariate analysis	
		HR (95% CI)	P	HR (95% CI)	P
Age					
<60	44 (19.0%)	1 (reference)			
>60	187 (81.0%)	0.890 (0.426-1.862)	0.758		
Sex					
Female	61 (26.4%)	1 (reference)		1 (reference)	
Male	170 (73.6%)	0.502 (0.268-0.940)	0.031	0.389 (0.203-0.744)	0.004
TNM stage					
I+II	141 (61.0%)	1 (reference)		1 (reference)	
III+IV	90 (39.0%)	2.492 (1.351-4.599)	0.003	3.003 (1.598-5.645)	<0.001

Note: characteristics with $P < 0.3$ in the univariate analysis were further screened in the multivariate analysis. HR: hazard ratio; CI: confidence interval; TNM stage: tumor-node-metastasis stage.

demonstrating a similar tendency. Collectively, the results showed that the prognostic model proposed in this study effectively predicted the survival of HCC patients.

3.6. Nomogram Analysis Indicates the Sampling Time of HCC Patients. Furthermore, nomogram analysis was performed based on genes showing significant circadian rhythms in liver tissue, which showed that CCGs, including *CRY2*, *PER1*, and *RORA*, have significant impacts on the predictive accuracy of the 4-CCG-based predictive model (Figure 5(a)). The nomogram results also revealed that lower expression levels of *CRY2*, *PER1*, and *RORA* were associated with higher predictive ability. More importantly, due to the rhythmic expression of CCGs in the liver, the time course of CCG's predictive accuracy was plotted based on their different expression levels (Figure 5(b)). Previous research indi-

cates that the expression peak phase of CCGs shifted by ~12 hours between the mouse and baboon [12]. Accordingly, we found that, when patients sampling at night (8:00 pm), *CRY2* and *PER1* reached their peak, resulting in higher risk scores and facilitating the early diagnosis of patients. Therefore, it is better to sample the HCC patients in the evening to obtain a more accurate predictive function.

3.7. Molecular Mechanisms of Circadian Clock Genes in HCC. To investigate the underlying mechanisms of circadian clock genes in the prognosis and diagnosis of HCC, firstly, WGCNA was performed to construct a coexpression gene network of the four core clock genes. As shown in Figure 5, these four clock genes are marked as large red nodes, whereas blue nodes represent the other coexpressed genes. Notably, gene *CRY2*, *PER1*, and *RORA* were closely

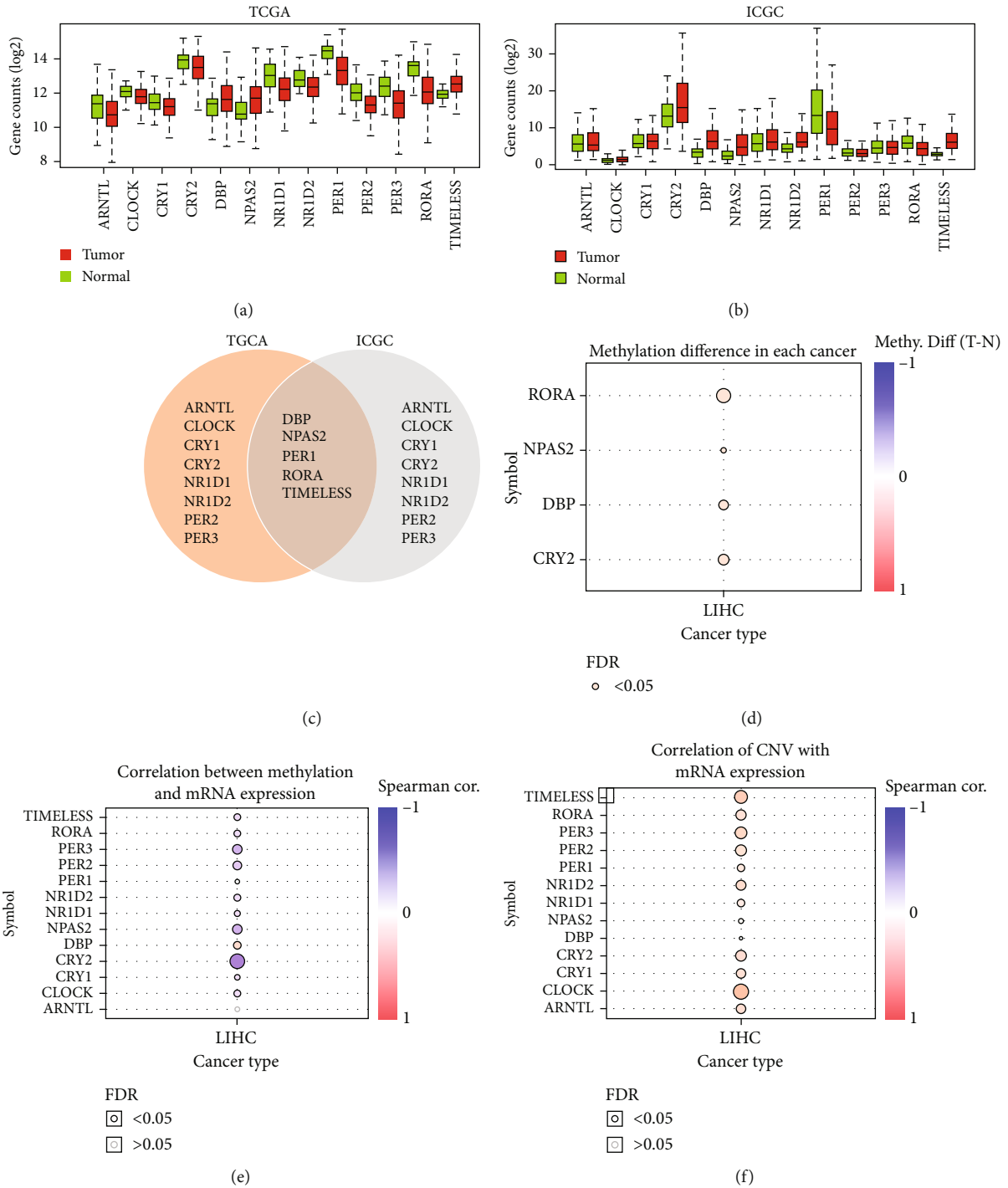


FIGURE 2: Continued.

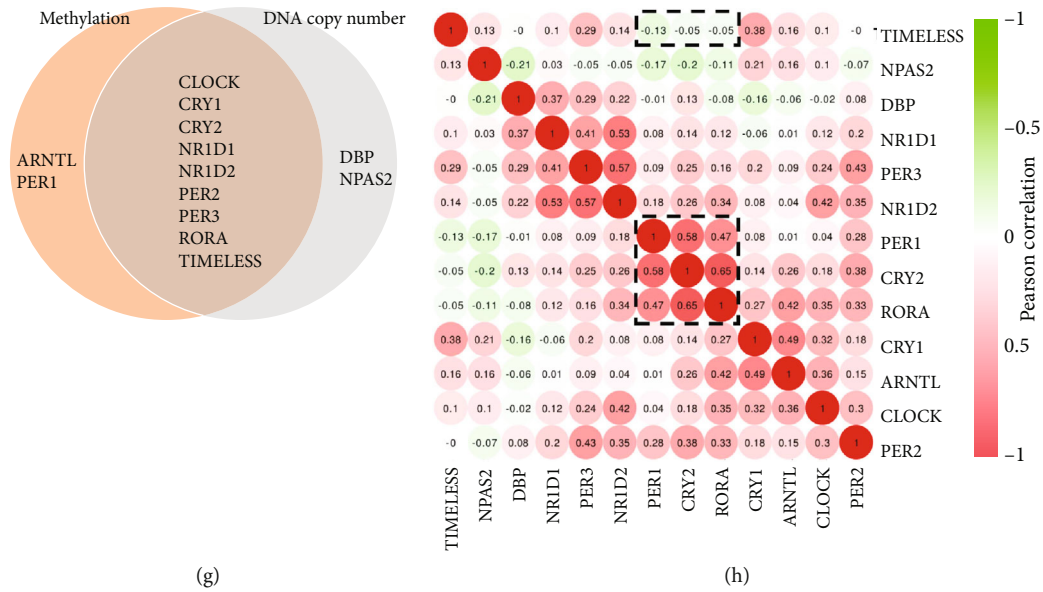


FIGURE 2: Differential expression analysis of circadian clock genes between HCC and normal tissues. (a, b) Box diagrams showing the expression levels of 13 circadian clock genes in tumor samples compared with normal samples in TCGA and ICGC. The P values of the differential expressed four CCGs (*CRY2*, *PER1*, *RORA*, and *TIMELESS*) were >0.05 . (c) The circadian clock genes showing a similar expression tendency in TCGA and ICGC. (d) Methylation difference between normal and tumor tissues. (e) Correlation between methylation and mRNA expression. (f) Correlation of copy number variation (CNV) with mRNA expression. (g) Venn diagram showing clock genes that were regulated by both methylation and CNV. (h) The interrelationship between circadian clock genes. TCGA: The Cancer Genome Atlas; ICGC: International Cancer Genome Consortium.

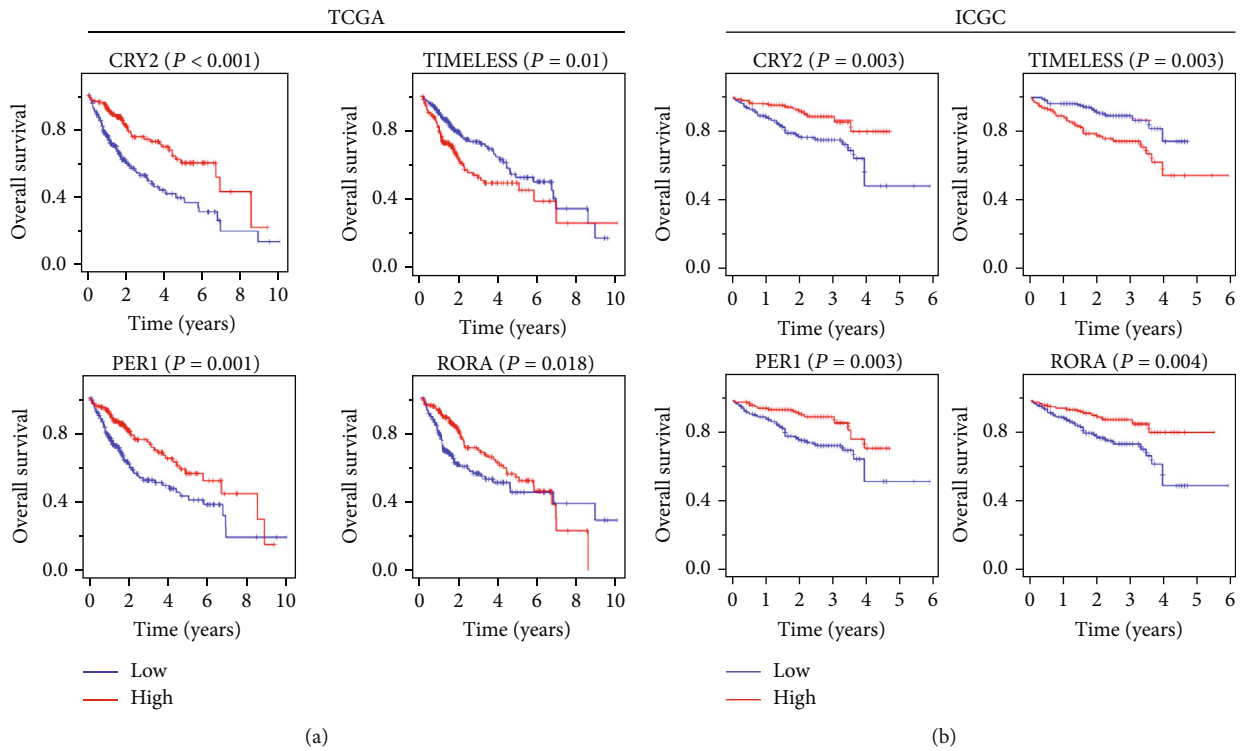


FIGURE 3: The prognostic value of circadian clock gene in HCC. The role of circadian clock genes in the overall survival of HCC patients based on the TCGA database (a) or ICGC database (b).

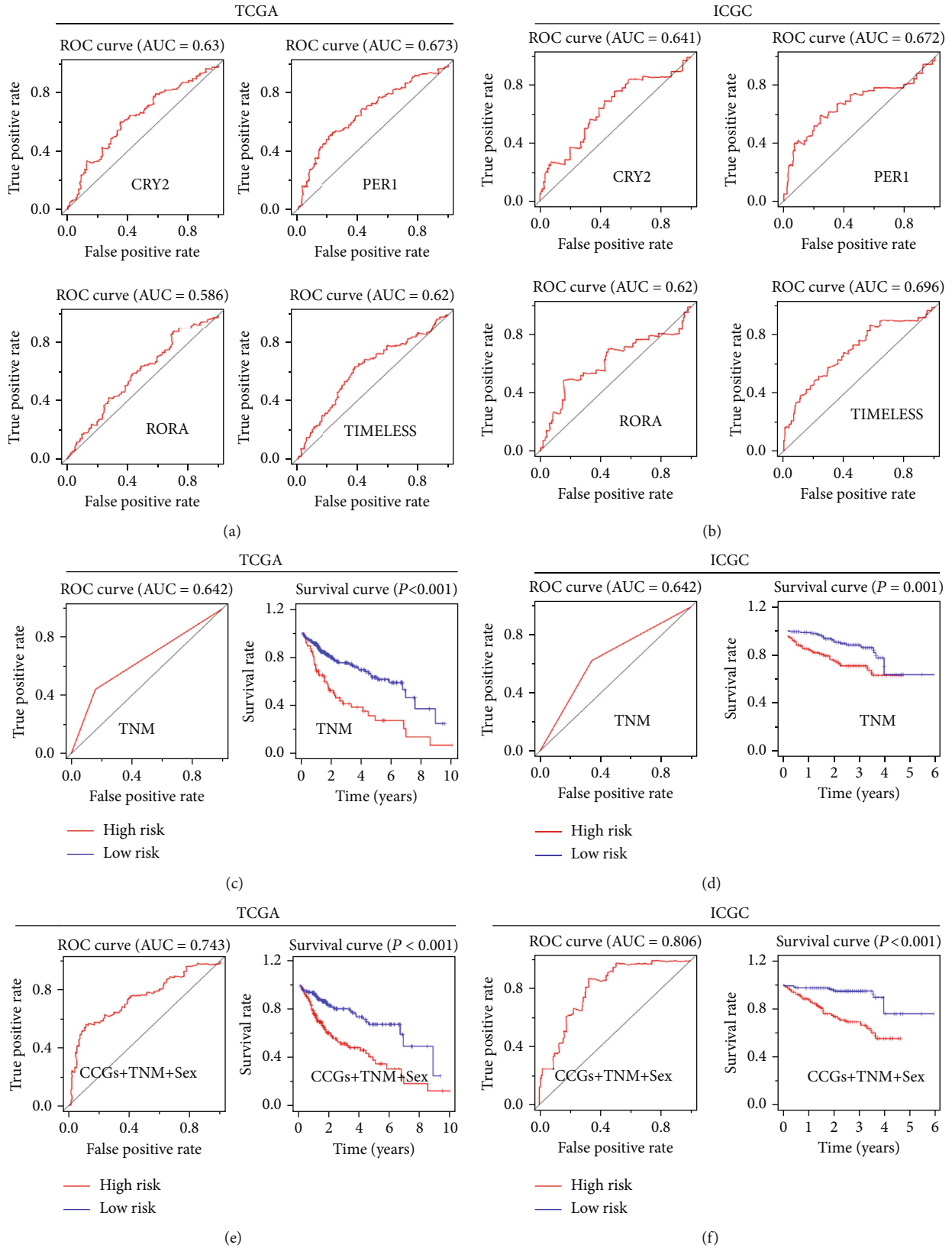


FIGURE 4: Prediction models to predict the survival of HCC patients. (a, b) ROC and survival curves of single-gene-based models in TCGA and ICGC, respectively. (c, d) ROC and survival curves of TNM stage-based model in TCGA and ICGC, respectively. (e, f) ROC and survival curves of the model consisting of survival-related four genes significantly associated with TNM stage and sex in TCGA and ICGC, respectively. CCGs: the four circadian clock genes.

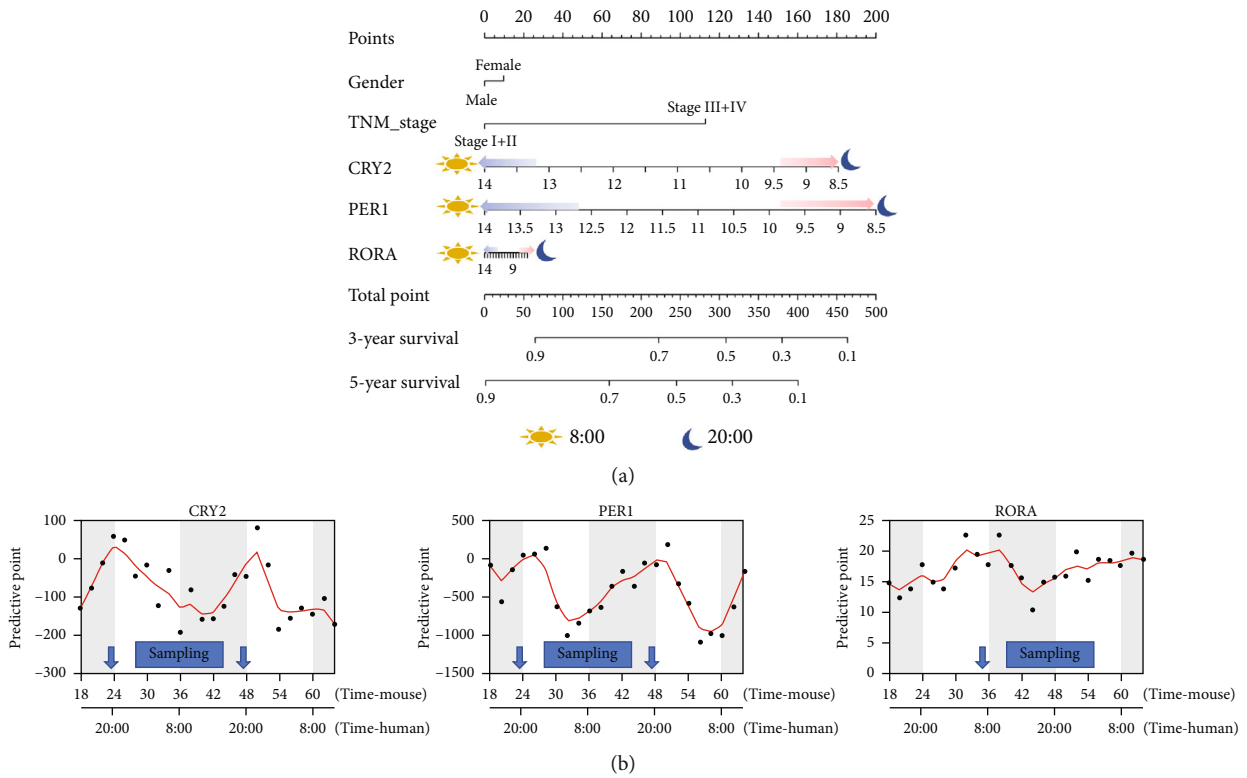


FIGURE 5: Nomogram analysis showed risk scores in HCC patients. (a) Nomogram based on genes that showed significant circadian rhythms in liver tissue. (b) Detailed display of the predictive point based on gene rhythmic expression.

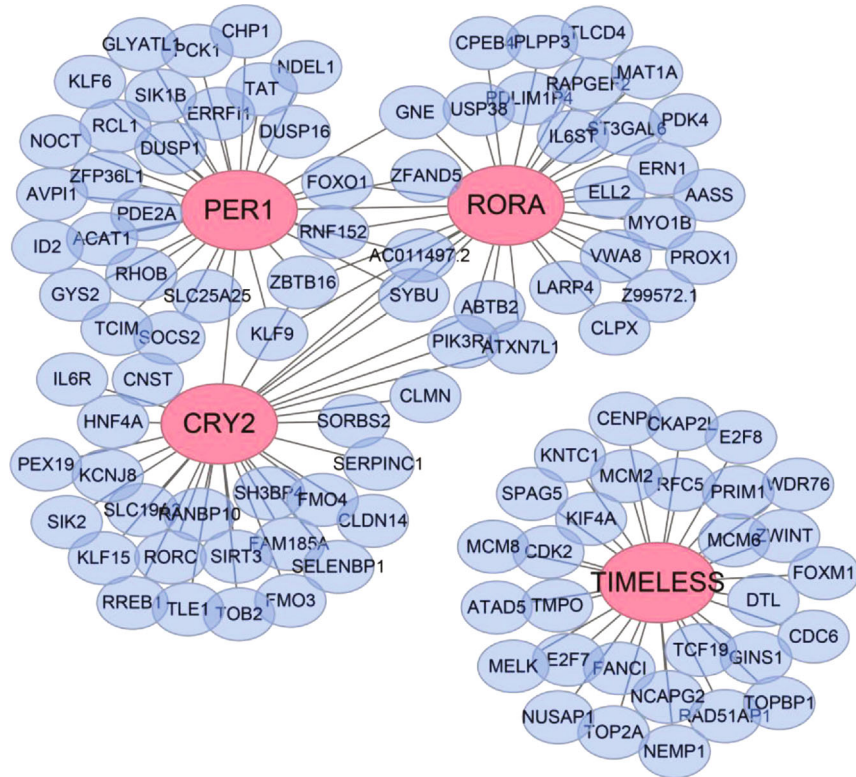


FIGURE 6: Coexpression network of circadian clock genes. The red nodes are circadian clock genes, while the blue nodes are the coexpressed genes.

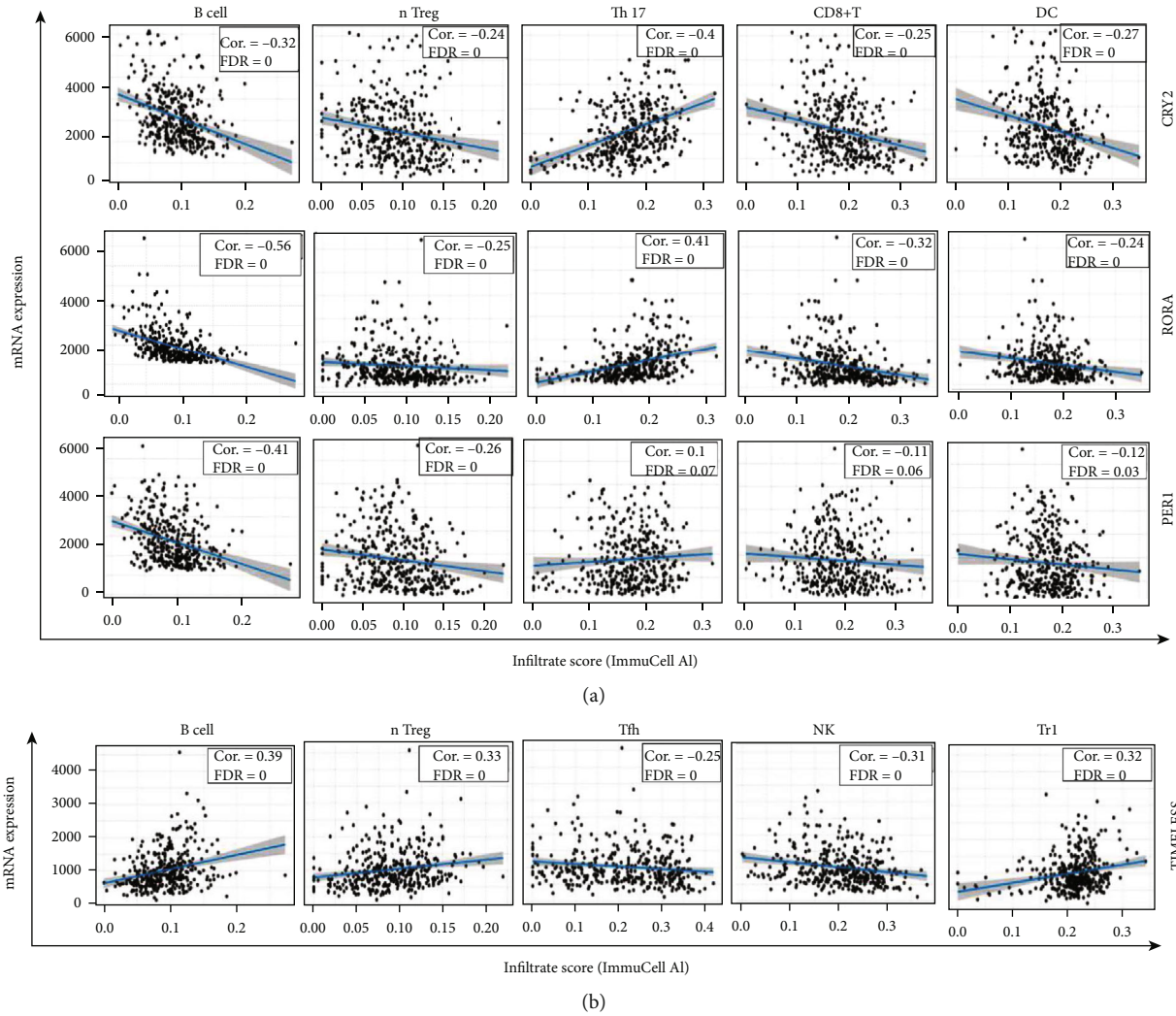


FIGURE 7: The correlation between circadian clock genes and the immune infiltration level in HCC. nTreg: natural regulator T cell; Th17: T helper 17 cells; DC: dendritic cell; Tfh: T follicular helper cell; NK, natural killer cell; Tr1: type 1 regulatory T cell.

associated, possessing several mutual cooperators (hereafter referred to as Cluster 1). However, *TIMELESS* was a relatively independent part of the coexpression gene network (Figure 6). This result was in accordance with the interrelationship between circadian clock genes (Figure 2(h)).

Previous studies have revealed the connection between the circadian rhythm and tumor microenvironment [13]. However, the role of the circadian clock in the tumor microenvironment remains unclear. Next, a correlation analysis was performed between the four core circadian clock genes and the infiltration levels of different immune cells (Figure 7). It was observed that Cluster 1 was significantly negatively associated with B cell, natural CD4⁺ regulatory T cell (nTreg), CD8⁺ T cell, and dendritic cell (DC) and positively related with the infiltration of T helper 17 (Th17) cell. On the contrary, *TIMELESS* was positively associated with B cell and nTreg cell. *TIMELESS* was also correlated with Tfh cell, NK cell, and Tr1 cell. These results indicated that Cluster 1 and *TIMELESS* might affect the survival of HCC patients by regulating immune infiltration levels, especially B cell and nTreg cell.

In addition, the role of circadian clock genes in cancer-related signaling pathways, including TSC/mTOR, RTK, RAS/MAPK, PI3K/AKT, hormone ER, hormone AR, EMT, DNA damage response, cell cycle, and apoptosis pathways, were examined (Figures 8(a) and 8(b)). As shown in pan-cancer analysis (Figure 8(a)) or liver cancer analysis (Figure 8(b)), Cluster 1 and *TIMELESS* exerted opposite effects on the same signaling pathway; that is, Cluster 1 activated, whereas *TIMELESS* inhibited the same pathway and vice versa. Besides, Cluster 1 mainly inhibited apoptosis, cell cycle, and DNA damage response, which play a critical role in maintaining uncontrolled proliferation and chemoresistance of cancer cells. For a better understanding of the molecular functions underlying the oncogenesis of early HCC, Gene Set Enrichment Analysis (GSEA) was performed, which showed that each clock gene of Cluster 1 was enriched in the same pathway, such as fatty acid metabolism, adipogenesis, bile acid metabolism, and peroxisome pathway based on the Hallmark Gene Sets. By contrast, *TIMELESS* was involved in pathways, including mitotic spindle, oxidative phosphorylation, and the E2F pathway.

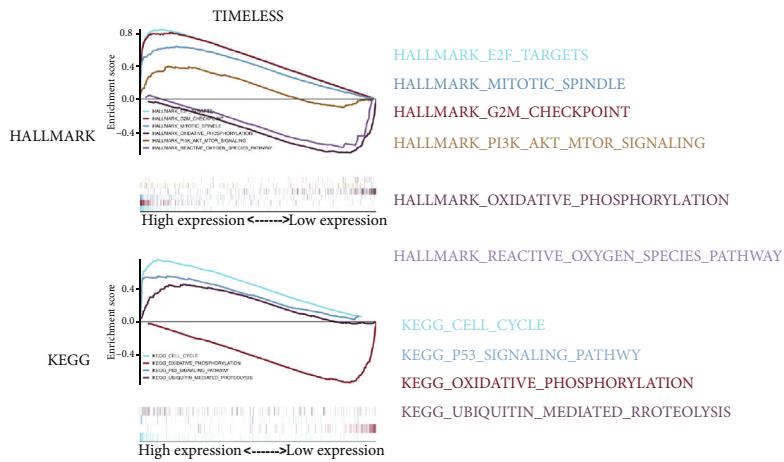
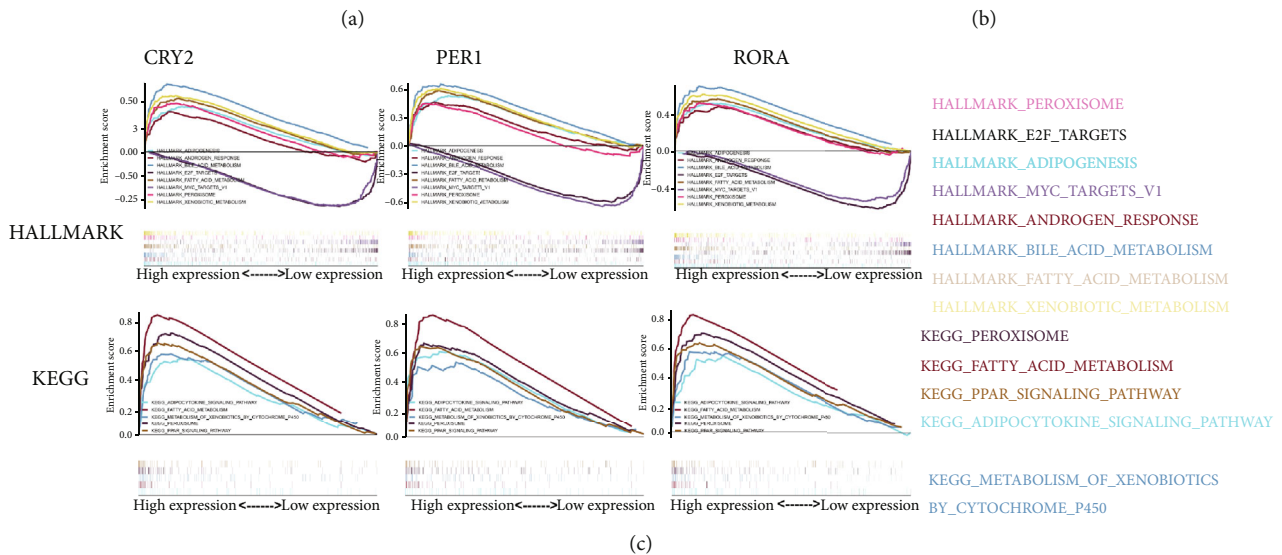
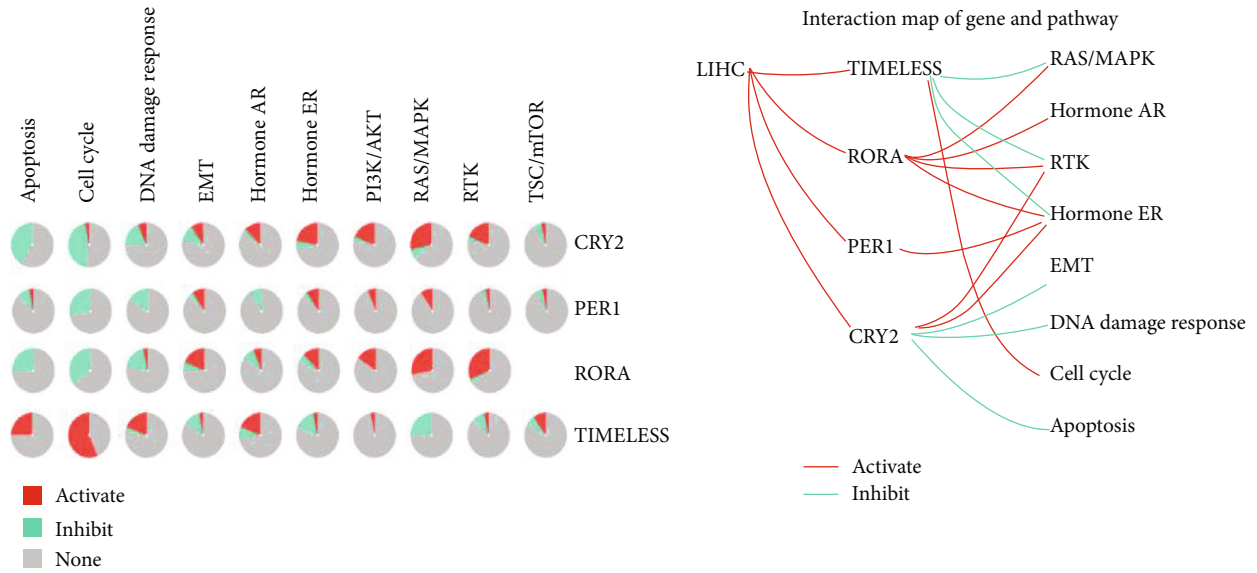


FIGURE 8: Continued.

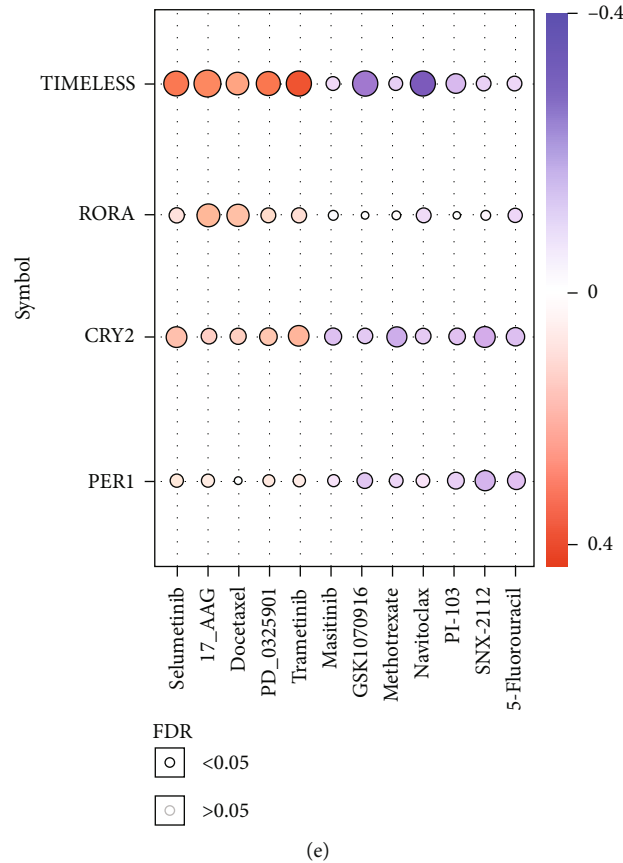


FIGURE 8: Pathway associated with survival-related clock genes. (a, b) Correlation between survival-related genes and cancer-related pathways. Pancancer (a) or liver cancer analysis (b) was performed to find the key cellular processes associated with the four CCGs. (c, d) GSEA-enriched terms. (e) Drug sensitivity of four survival-related genes.

KEGG gene sets were also applied as a reference cohort (Figures 8(c) and 8(d)). It was also observed that Cluster 1 was closely positively related to the metabolism of amino acids, whereas *TIMELESS* was related to DNA replication and DNA repair-associated signaling pathways (Fig. S5).

Cancer chronotherapy, a therapeutic treatment at a specific time following circadian rhythms, may improve the antitumor effects and reduce toxicity [14]. Accordingly, the correlation between clock gene expression and drug sensitivity was also investigated using datasets from Genomics of Drug Sensitivity in Cancer (GDSC), in which high expression means resistance to a particular anticancer drug. We found that higher expression of Cluster 1 exhibited a similar positive correlation with chemoreagents such as selumetinib, 17-AAG, docetaxel, PD-0325901, and trametinib. Conversely, *TIMELESS* showed a stronger negative correlation with masitinib, GSK1070916, methotrexate, navitoclax, PI-103, SNX-2112, and 5-fluorouracil (Figure 8(e)). These results suggested that inhibition of Cluster 1 or activation of *TIMELESS* might enhance the chemotherapeutic sensitivity toward special anticancer drugs.

4. Discussion

This study demonstrated that four circadian clock genes, including *CRY2*, *PER1*, *RORA*, and *TIMELESS*, could be

potential diagnostic and prognostic biomarkers for HCC patients. We also established a prediction model consisting of these four genes, TNM stage, and sex, demonstrating high predictive ability. In addition, it was shown that Cluster 1 (*CRY2*, *PER1*, and *RORA*) and *TIMELESS* exerted opposite impacts on interactive gene network, infiltration of immune cells, cancer-related signaling pathways, and cellular sensitivity to clinically used drugs.

Disruption of the circadian rhythm always leads to physiological disorders of homeostasis in mammals, which is closely associated with the development of cancer [4]. Gene expression, cell cycle, and DNA repair are regulated by the clock genes, providing the base to the hypothesis that disruption of biorhythms may predispose individuals to cancer [6]. Considering the possibility that circadian clock genes play a pivotal role in the physiological functions of mammals, rendering individuals towards the development of cancer [15], the differential expression of core circadian clock genes between HCC tissues and normal tissues was discussed. It was observed that *DBP*, *NPAS2*, *PER1*, *RORA*, and *TIMELESS* showed similar expression tendency in HCC tissues in the TCGA and ICGC databases. The mRNA expression was either affected by methylation [16] or by copy number variation (CNV), and the fluctuation of DNA copy number was found responsible for the alteration in coding RNA

expression level [17]. It was also observed that the expression of genes such as *CRY2*, *DBP*, *NPAS2*, and *RORA* was significantly affected by methylation (Figure 2(d)), and all circadian clock genes, except for *DBP* and *NPAS2*, exhibited a significant correlation with CNV. Collectively, the results mentioned above implied the involvement of methylation or CNV in the dysregulation of circadian clock genes.

In addition, we demonstrated that the dysregulation of circadian clock genes was associated with the prognosis of HCC patients. High expression of Cluster 1 (*CRY2*, *PER1*, and *RORA*), or low expression of *TIMELESS*, was correlated with prolonged overall survival (OS) of patients (Figure 3). The investigation of the molecular mechanisms revealed that Cluster 1 and *TIMELESS* counteractively regulated the infiltration of several immune cells such as B cells and nTreg cells. Inherently, B cells can inhibit tumor growth by producing antibodies and presenting tumor antigens, while nTreg cells control the inflammatory microenvironment to restrict tumor development [18–20]. High expression of Cluster 1, or low expression of *TIMELESS*, might inhibit both the infiltration of B cells and nTreg cells (Figure 7), suggesting that the dysregulation of circadian clock genes may manifest HCC by disrupting the tumor microenvironment.

Another important finding of this study was that dysregulation of the circadian clock genes was also found to be associated with several cancer-related pathways (Figure 8), such as DNA damage response, cell cycle, and apoptosis, which is in accordance with previous research that the circadian clock genes influenced cancer susceptibility through DNA damage and apoptosis [21]. Although the cell cycle and circadian clock genes are considered two different biological oscillators, their close relation and interaction have been reported [22]. The GSEA results showed that gene sets of E2F targets, fatty acid metabolism, AKT/mTOR, and p53 signal pathway were significantly enriched. Similarly, Cluster 1 and *TIMELESS* exerted effects on these signaling pathways conversely. Moreover, AKT/mTOR and p53 pathways played vital roles in regulating cell proliferation, and *TIMELESS* could promote the proliferation of HCC cells by inhibiting the p53-dependent signals [23], affirming the finding that high expression of *TIMELESS* is related to poor survival of HCC patients (Figure 3).

Furthermore, the interaction between the circadian clock genes and cellular sensitivity to an anticancer drug was analyzed. Several chemoreagents, such as 5-FU [24] and docetaxel [25], have demonstrated potent antiliver cancer activities. It was observed that higher expression of Cluster 1 might enhance the chemoresistance of these anticancer reagents, implying that inhibition of Cluster 1, or activation of *TIMELESS*, may render liver cancer cells more sensitive to chemotherapy.

5. Conclusion

This work demonstrated that four CCGs, including *CRY2*, *PER1*, *RORA*, and *TIMELESS*, could be potential diagnos-

tic and prognostic biomarkers for HCC patients. Besides, *CRY2*, *PER1*, and *RORA* exerted opposite impacts against *TIMELESS* on immune cell infiltration and cancer-related signaling pathways, affecting the overall survival of HCC patients. Selective regulation of circadian clock genes may further assist in precise chronotherapy of HCC patients.

Abbreviations

HCC:	Hepatocellular carcinoma
PLC:	Primary liver cancer
TCGA:	The Cancer Genome Atlas
ICGC:	International Cancer Genome Consortium
logFC:	Log ₂ foldchange
FDR:	False discovery rate
DEGs:	Differentially expressed genes
CNV:	Copy number variation
HR:	Hazard ratio
WGCNA:	Weighted gene coexpression network analysis
GSEA:	Gene Set Enrichment Analysis
TNM stage:	Tumor-node-metastasis stage.

Data Availability

The datasets presented in this study, including TCGA and ICGC, can be found online.

Conflicts of Interest

The authors declare no conflict of interest, financial or otherwise.

Authors' Contributions

XW and KZ conceived the project. YL and SW carried out the experiments. XH, RC, and QT analyzed data. YL, RY, and XH collected clinical data. XW and KZ acquired funding. YL and KZ wrote the original draft of the manuscript. KZ reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version. Youfang Liang and Shaoliang Wang contributed equally to this work.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (82073937), Natural Science Foundation of Guangdong Province (2018A030313122), Shenzhen Science and Technology Project (JCYJ20180305163658916, JCYJ20180228175059744), Shenzhen Key Medical Discipline Construction Fund (SZXK059), Shenzhen Healthcare Research Project (SZBC2018007), the Shenzhen University funding, and SZU Medical Young Scientists' program.

Supplementary Materials

Supplementary material containing four figures is available on the publisher's website along with the published article. (*Supplementary Materials*)

References

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] A. Ozakyol, "Global epidemiology of hepatocellular carcinoma (HCC epidemiology)," *Journal of Gastrointestinal Cancer*, vol. 48, no. 3, pp. 238–240, 2017.
- [3] T. E. Huang, Y. N. Deng, J. L. Hsu et al., "Evaluation of the anticancer activity of a bile acid-dihydroartemisinin hybrid ursodeoxycholic-dihydroartemisinin in hepatocellular carcinoma cells," *Frontiers in Pharmacology*, vol. 11, p. 599067, 2020.
- [4] S. Masri and P. Sassone-Corsi, "The emerging link between cancer, metabolism, and circadian rhythms," *Nature Medicine*, vol. 24, no. 12, pp. 1795–1803, 2018.
- [5] C. R. Cederroth, U. Albrecht, J. Bass et al., "Medicine in the fourth dimension," *Cell Metabolism*, vol. 30, no. 2, pp. 238–250, 2019.
- [6] A. Sancar and R. N. van Gelder, "Clocks, cancer, and chronotherapy," *Science*, vol. 371, no. 6524, article eabb0738, 2021.
- [7] D. Gu, S. Li, S. Ben et al., "Circadian clock pathway genes associated with colorectal cancer risk and prognosis," *Archives of Toxicology*, vol. 92, no. 8, pp. 2681–2689, 2018.
- [8] C. Cadenas, L. van de Sandt, K. Edlund et al., "Loss of circadian clock gene expression is associated with tumor progression in breast cancer," *Cell Cycle*, vol. 13, no. 20, pp. 3282–3291, 2014.
- [9] C. M. Hsu, P. M. Lin, C. C. Lai, H. C. Lin, S. F. Lin, and M. Y. Yang, "PER1 and CLOCK: potential circulating biomarkers for head and neck squamous cell carcinoma," *Head & Neck*, vol. 36, no. 7, pp. 1018–1026, 2014.
- [10] H. Xiong, Y. Yang, K. Yang, D. Zhao, H. Tang, and X. Ran, "Loss of the clock gene PER2 is associated with cancer development and altered expression of important tumor-related genes in oral cancer," *International Journal of Oncology*, vol. 52, no. 1, pp. 279–287, 2018.
- [11] R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch, "A circadian gene expression atlas in mammals: implications for biology and medicine," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 45, pp. 16219–16224, 2014.
- [12] L. S. Mure, H. D. le, G. Benegiamo et al., "Diurnal transcriptome atlas of a primate across major neural and peripheral tissues," *Science*, vol. 359, no. 6381, 2018.
- [13] Y. Yang, G. Yuan, H. Xie et al., "Circadian clock associates with tumor microenvironment in thoracic cancers," *Aging (Albany NY)*, vol. 11, no. 24, pp. 11814–11828, 2019.
- [14] Y. Ye, Y. Xiang, F. M. Ozguc et al., "The genomic landscape and pharmacogenomic interactions of clock genes in cancer chronotherapy," *Cell Systems*, vol. 6, no. 3, pp. 314–328.e2, 2018.
- [15] L. Fu and N. M. Kettner, "The circadian clock in cancer development and therapy," *Progress in Molecular Biology and Translational Science*, vol. 119, pp. 221–282, 2013.
- [16] R. Feil and M. F. Fraga, "Epigenetics and the environment: emerging patterns and implications," *Nature Reviews Genetics*, vol. 13, no. 2, pp. 97–109, 2012.
- [17] L. Liang, J. Y. Fang, and J. Xu, "Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy," *Oncogene*, vol. 35, no. 12, pp. 1475–1482, 2016.
- [18] R. D. Leone and J. D. Powell, "Metabolism of immune cells in cancer," *Nature Reviews Cancer*, vol. 20, no. 9, pp. 516–531, 2020.
- [19] Y. Zhang, T. Liu, X. Hu et al., "CellCall: integrating paired ligand-receptor and transcription factor activities for cell-cell communication," *Nucleic Acids Research*, vol. 49, no. 15, pp. 8520–8534, 2021.
- [20] Y. Zhang, T. Liu, J. Wang et al., "Cellinker: a platform of ligand-receptor interactions for intercellular communication analysis," *Bioinformatics*, vol. 37, no. 14, pp. 2025–2032, 2021.
- [21] A. Angelousi, E. Kassi, N. Ansari-Nasiri, H. Randeva, G. Kaltsas, and G. Chrousos, "Clock genes and cancer development in particular in endocrine tissues," *Endocrine-Related Cancer*, vol. 26, no. 6, pp. R305–R317, 2019.
- [22] E. Farshadi, G. T. J. van der Horst, and I. Chaves, "Molecular links between the circadian clock and the cell cycle," *Journal of Molecular Biology*, vol. 432, no. 12, pp. 3515–3524, 2020.
- [23] J. S. Zhang, P. Yuan, Z. Y. Yan et al., "Timeless promotes the proliferation of hepatocellular carcinoma cell by reprogramming of glucose metabolism," *Zhonghua Zhong Liu Za Zhi*, vol. 40, no. 7, pp. 499–505, 2018.
- [24] Z. Zhang, K. Hu, K. Miyake et al., "A novel patient-derived orthotopic xenograft (PDOX) mouse model of highly-aggressive liver metastasis for identification of candidate effective drug-combinations," *Scientific Reports*, vol. 10, no. 1, p. 20105, 2020.
- [25] H. L. Lin, T. Y. Liu, G. Y. Chau, W. Y. Lui, and C. W. Chi, "Comparison of 2-methoxyestradiol-induced, docetaxel-induced, and paclitaxel-induced apoptosis in hepatoma cells and its correlation with reactive oxygen species," *Cancer*, vol. 89, no. 5, pp. 983–994, 2000.

Research Article

Prediction of Metal Ion Binding Sites of Transmembrane Proteins

Jing Qu,^{1,2} Sheng S. Yin ¹ and Han Wang ²

¹Systems Engineering Research Institute, Beijing, China

²Institute of Computational Biology, School of Information Science and Technology, Northeast Normal University, Changchun, China

Correspondence should be addressed to Sheng S. Yin; 552180276@qq.com and Han Wang; wangh101@nenu.edu.cn

Received 17 September 2021; Accepted 1 October 2021; Published 22 October 2021

Academic Editor: Hui Ding

Copyright © 2021 Jing Qu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The metal ion binding of transmembrane proteins (TMPs) plays a fundamental role in biological processes, pharmaceuticals, and medicine, but it is hard to extract enough TMP structures in experimental techniques to discover their binding mechanism comprehensively. To predict the metal ion binding sites for TMPs on a large scale, we present a simple and effective two-stage prediction method TMP-MIBS, to identify the corresponding binding residues using TMP sequences. At present, there is no specific research on the metal ion binding prediction of TMPs. Thereby, we compared our model with the published tools which do not distinguish TMPs from water-soluble proteins. The results in the independent verification dataset show that TMP-MIBS has superior performance. This paper explores the interaction mechanism between TMPs and metal ions, which is helpful to understand the structure and function of TMPs and is of great significance to further construct transport mechanisms and identify potential drug targets.

1. Introduction

Metal ions are vital to live organisms involving in various biological processes. They can enter cells to regulate the expression and activation of multiple biomolecules, participate in cell signal transduction, and complete various functions. For example, Ca^{2+} signaling is essential for T cell activation, autoantigen tolerance, differentiation, and development [1]. Mg^{2+} regulates ion channels' activity in cardiac cells, affecting the myocardium's electrical properties [2]. Zn^{2+} is a multitasking tool necessary to stimulate various enzyme activities [3] that lack and excess can cause central nervous system diseases [4, 5]. Also, other metal ions [6–10] perform their respective biological functions. Their homeostasis disorders involve neurodegenerative diseases, cardiovascular diseases, bone diseases, asthma, cancer, and diabetes [11]. Therefore, maintaining the correct levels of metal ions in the cytoplasm is essential for life and health.

As we know, metal ions cannot directly penetrate the cell membrane unless the transporter's assistance is on the cell membrane. According to the transport mode and spatial structure, the transporter protein can be roughly divided

into channel and carrier proteins, all transmembrane proteins (TMPs). These particular proteins cross through the biomembranes by their transmembrane domains and exist therein whole life, constitute 15-30% of the genome [12]. TMPs, as the primary carrier of metal ions, participate in signal transduction, intracellular trafficking, and maintaining homeostasis [13, 14]. However, knowledge about the transport mechanism of metal ions that bind to proteins across membranes is still insufficient and varies for different metals. Exploring the TMPs' metal ion binding site (MIBs) provides a practical means to explore the ion selectivity and crucial abilities and even further construct the ion transport mechanism.

Experimental techniques such as AFM [15], MS [16], IMAC [17, 18], NMR [19], and X-ray crystallography [20] are comprehensive to identify the crystal structures of protein and characterize the binding sites in proteins. However, these techniques had not achieved large-scale application compared to the water-soluble proteins since the TMPs' folding, native structure, stability, and activity are reached only within the lipid bilayer [21]. With sequencing technology development, the time has come to study the work

related to TMPs. Previous extensive research on water-soluble proteins has provided ideas in computational methods for use as reference. These methods, computation-based, reduce the cost to discover the potential MIBs that also have near-accurate predictions.

Over the last decade, computational methods have been made significant advances in identifying MIBs. Yang et al. combined two methods based on substructure and sequence (COACH [22]) to identify protein-ligand binding sites and achieved Matthews correlation coefficient (MCC) of 0.54. Zhao et al. present a 3D template-based metal site prediction (TEMSP [23]) to predict zinc binding sites and completed sensitivity of 0.86. Lin et al. used a fragment transformation method (MIB [24]) to predict twelve metal ion-binding sites with overall accuracy from 0.92 to 0.95. Yu et al. report a ligand-specific template-free predictor (TargetS [25]) for identifying protein-ligand binding sites that contain five metal ions overall MCC from 0.14 to 0.69. Hu et al. proposed a ligand-specific and template-based components approach (IonCom [26]) to predict 13 ions and achieved MCC from 0.14 to 0.69. Cao et al. [27] used only sequence information for multiple metals and yielded an overall accuracy from 0.62 to 0.84. Kumar [28] used the amino acid sequence information and machine learning approach to predict six metal ion binding sites' accuracy 0.86 to 0.87. Qiao and Xie developed a sequence-based ligand-specific predictor (MIONSite [29]) to predict 12 metal ion binding sites and completed MCC from 0.17 to 0.68. Haberal and Ogul [30] present deep learning architectures to predict metal binding of histidines (HIS) and cysteines (CYS) amino acids and acquire the precision 0.79 and recall 0.82.

Although the prediction of the binding site of metal ions and proteins has been fruitful, it cannot be directly applied to TMPs [31–34]. First of all, the above methods can be roughly summarized into structure-based and sequence-based. The former has better prediction performance, but the latter is more common. In the past, one of the impediments to this effort related to ion channels is that TMPs' structures have been notoriously difficult to obtain. Therefore, the performance of structure-based is limited in TMPs. Then, sequence-based methods of MIBs did not distinguish between TMPs and water-soluble proteins, while TMPs have significant conformational differences with those water-soluble proteins. Structurally, metal ions pass through the body of TMPs while they had never done so to any water-soluble protein. Functionally, water-soluble proteins cannot take on the responsibility of transporting metal ions inside and outside the membrane. Finally, TMPs have selective specificity for metal ions, which allows only a suitable size of metal ions to pass through. Therefore, ignoring the natures as mentioned above is incompatible with biological significance.

In this study, we proposed a metal-specific method for predicting the binding sites of the transmembrane protein and metal ions (TMP-MIBS) from protein sequence information. We selected five kinds of TMPs' specific structural or biochemical features: evolutionary information, physicochemical properties, solvent-accessible surface area, topology structure, and Z-coordinate features. TMP-MIBS was

well trained against an up-to-date dataset collected from the PDBTM database. The sliding windows were introduced to build feature spaces, and random undersampling was utilized to tackle sample imbalance. The performance of the model is gradually improved through a two-stage learning process. In the first stage, metal ion binding sites of TMPs were identified. In the second stage, specific recognition models were constructed for the metal-specific. We have not yet found any tool specifically for binding site prediction about metal ion prediction and TMPs, so we compared ours with the published means for metal ions and general proteins. Our model achieves the best performance except for Ca^{2+} . The work has culminated in a relatively effective tool for predicting metal binding sites without 3D structures. It has guiding significance for understanding and ultimately controlling the binding ability of metal ions and their application in drug and disease treatment in the future.

2. Materials and Methods

2.1. Datasets. The PDBTM [35] database (available at <http://pdbtm.enzim.hu>), which aims to collect all the TMPs from the protein structure database (PDB) and keep up to date with PDB, is the source of the data in this work. We screen protein data containing metal ion binding sites and parse sequences from the PDB file by applying the following criteria.

- (1) Only keep chains with residues that participate in binding metal ions when the proteins have more than one polypeptide chain
- (2) The length of the polypeptide chain is required to exceed 50 residues
- (3) Removing the protein sequences with sequence similarities greater than or equal to 40% by Cd-Hit [36]

Finally, there are 427 protein chains left as the experimental dataset. To evaluate the effectiveness of our model, we divide the training dataset and the independent verification dataset as listed in Table 1.

2.2. Feature Extraction

2.2.1. Evolutionary Information. The sequence-based methods mainly rely on residue conservation analyses assuming that ligand binding residues are functionally important and should be conserved in the evolution [37–39]. By running the PSI-BLAST program on the server, iteratively searched the NR database three times and used 0.001 as the *E*-value cutoff of multiple sequence alignments to obtain evolution information of the protein sequence. We generated the position-specific scoring matrix (PSSM). The *L* residue's protein sequence generates an $L \times 20$ matrix.

2.2.2. Physicochemical Properties. Early studies in the prediction of transmembrane (TM) helices had widely used physicochemical properties (PCP) such as hydrophobicity analysis [40], the positive inside rule [41–43], and charge bias which are indeed valid. Besides, the residues binding

TABLE 1: The training set and independent test set.

Category	Training dataset		Independent verification dataset	
	NProta	Nrecb	NProta	Nrecb
K ⁺	63	202	6	17
Ca ²⁺	78	388	10	51
Na ⁺	54	223	7	46
Zn ²⁺	52	241	5	26
Mg ²⁺	64	209	10	27
Hg ²⁺	8	83	3	25
Cu ²⁺	14	54	2	9
W ⁶⁺	13	56	1	4
Cd ²⁺	8	31	2	10
Ni ²⁺	10	28	3	9
Fe ³⁺	7	19	0	0
Mn ²⁺	7	32	0	0
Cu ₂	2	12	0	0
Rb ⁺	4	18	0	0
Au ⁺	2	13	0	0
Cs ⁺	6	23	0	0
Pb ²⁺	1	10	0	0
Fe ²⁺	3	7	0	0
Pt ²⁺	1	2	0	0
Sr ²⁺	1	4	0	0
Li ⁺	1	4	0	0
Co ²⁺	3	4	0	0
Pr ³⁺	1	3	0	0
Mo ⁶⁺	1	1	0	0

NProta: number of protein entries; Nrecb: number of protein receptors bound with ions.

with metal ions have many distinctive properties, such as electron-acceptor ability, positive charge, ion size, specific ligand affinity, varying valence state, and low or high spin configuration [44]. We collected the 553 physicochemical properties that influence the microenvironment of proteins. They were obtained from AAindex [45]. The protein sequence of the L residue generates an $L \times 553$ matrix.

2.2.3. Solvent-Accessible Surface Area. The solvent-accessible residues could be responsible for acquiring metal and may act as a potential metallochaperone to deliver metal to the TM region [46]. We calculate the relative solvent accessibility surface area (rASA) by MemBrain [47] for each residue to provide the residues' relative positions, which characterize TMPs' structure. The protein sequence of the L residue generates an $L \times 1$ matrix.

2.2.4. Topology Structure. Knowledge of the TM helices' presence and the exact location is essential for functional annotation and direct functional analysis. The prediction of topology structure (TOPO) serves to quickly obtain fundamental structural knowledge of TM proteins [48]. We used TMHMM-2.0 [49], which predicts the sequence's most

probable location and orientation of transmembrane helices. The protein sequence of the L residue generates an $L \times 3$ matrix.

2.2.5. Z-Coordinate. The Z-coordinate (Zcoord) is defined as the residue's distance to the center of the membrane [50] and reflects the high correlation with the ligand binding and the protein-protein binding regions [51]. It implicitly contains information about TMPs' secondary structure, such as re-entrant helices, interfacial helices, a TM helix's tilt, and loop lengths. TOPCONS [52] was used to predict the Zcoord. The protein sequence of the L residue generates an $L \times 1$ matrix.

2.3. Methods

2.3.1. Outline. TMP-MIBS employs a two-stage learning process and an ensemble of models to improve prediction performance gradually. The obtained data were preprocessed and extracted the protein sequence and feature. All binding residues (the 24 kinds of metal ions) are predicted to identify the MIBs in the first stage. The second stage indicates the most probable location and binding probability of MIBs for seven classes which are K⁺, Ca²⁺, Na⁺, Zn²⁺, Mg²⁺, Hg²⁺, and others. We test two-stage models on the independent verification dataset to examine the performance of the model. More details on how our final model was built and trained are explained below.

2.3.2. The First Stage of the Learning Process. When constructing the feature space is generated as the input of the first stage of the model, the sliding window strategy is used to intercept the amino acid fragments, and the random undersampling is introduced to extract some negative samples. Random forest (RF) is used as the prediction model and vote for binary class and selects the classification having the most votes. For a given protein sequence, the classifier outputs the exact conclusion that each residue is or is not a MIBs. This stage only predicted whether the residue would be binding with one in the 24 metal ions.

2.3.3. The Second Stage of the Learning Process. The second stage learning process takes into account the ligand-specific. After the first model training stage, two prediction results, "1" and "0", are output, corresponding to MIBs and non-MIBs. To further predict the binding of amino acid residues to metal ions, it is necessary to model the samples with the prediction result of "1" and enter the second stage of model learning. The second stage models the seven types of metal ions with the most significant number of sites, respectively. The OVR strategy in the multiclassification problem is adopted. Each time, the examples in one class are regarded as positive classes, and all other classes are taken as counterexamples. Finally, the seven classifiers for seven class metal ions output the probabilities for each residue binding residue in the given protein sequence.

2.4. Random Undersampling. Undersampling is a common technique among the existing technologies to overcome the sample imbalance problems. All the binding sites (positive

samples) are kept, and the nonbinding sites (negative samples) as an original dataset S will generate a new set S' . The numbers are N times the positive samples (N takes an integer). We set the ratio parameter of positive and negative samples to 1:5, with the N design details explained in Section 3.4.

2.5. Sliding Windows. The structural state of a residue is determined not only by amino acid residue itself but also by neighboring residues. The interception of the neighbor residue length is critical to the description of the target residue. Underintroducing the information of neighbor residues is not conducive to distinguishing, but overintroducing may cause noise. The sliding window strategy is widely used to contemplate the influence of neighbor residues for the target residuals, located in the middle, and $(w - 1)/2$ adjacent residues are found on both sides (w size, being an odd number). Since the volume of metal ions is usually small, the optimal window length of metal ions should be smaller than that of the larger ligands, such as ATP and NAD ligands (17 in general) [53]. We computed and analyzed evaluation indicators for seven class metal ions to determine the optimal sliding window length.

2.6. Validation and Evaluation Metrics. Random 10-fold cross-validation was used to validate model and tuning parameters, which one set was used for testing, and the remaining sets were used for training. We randomly divided the dataset into ten sets. Repeat this process ten times, and the final score was obtained by averaging the performances. We used five evaluation measures to evaluate the generalization ability of the model, which are accuracy (ACC), specificity (SPE), sensitivity (SEN), Matthews correlation coefficient (MCC), and area under ROC curve (AUC), respectively [54–57]. The training dataset is used to fine-tune the proposed methods' parameters, and the independent test is used to test the methods.

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (1)$$

$$\text{SPE} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad (2)$$

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (3)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}, \quad (4)$$

$$\text{AUC} = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} + x_i) \cdot (y_i + y_{i+1}), \quad (5)$$

where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative.

3. Results and Discussion

3.1. Specific Binding of Metal Ions and Amino Acids. It is well known that ion channels are highly selective for controlling ions in and out, which can be reflected by combining different ions with amino acid residues [27]. We counted the frequency of amino acids and nonamino acids bound by metal ions to TMPs, as shown in Figure 1. Interestingly, (b) Ca^{2+} , (d) Zn^{2+} , (e) Mg^{2+} , and (f) Hg^{2+} have higher specificity when combined with residues than (a) K^+ and (c) Na^+ . For Zn^{2+} , it is more likely to connect with His (H), Cys (C), and Glu (E), which are polar amino acids. Mg^{2+} is more likely to bind ASP (D) and more minor to nonpolar amino acids. Hg^{2+} has the highest tendency to combine with amino acids containing the neutral R group, while Ca^{2+} has the most increased tendency to combine with acidic amino acids. Careful observation shows that metal ions are more likely to connect with hydrophilic amino acids than hydrophobic amino acids. The finding supports the hypothesis of solvent-accessible residues that act as a potential metallochaperone and participate in delivering metal to the TM region.

In addition, we can conclude that the difference in the binding frequency with amino acids reflects metal ions' physical and chemical properties. Metal ions under the main analogous group have similar chemical properties and also have similar selectivity. The selection of binding amino acid residues by metal ions of different main groups is also quite different.

3.2. Position Conservation of Amino Acids. We further studied the conservative position information of the above six MIBs by WebLogo [58], as shown in Figure 2. Sequences were intercepted in window length L of 21 as an example for each metal ion class to analyze. The relative size of letters (amino acids) indicates their occurrence frequency in the sequence. The larger the letter, the higher the frequency. According to the illustration, no matter the binding site or nonbinding site of K^+ and Na^+ is remarkable, reflecting the proximity of the two metal ion sites in sequence and structure. But the status of other metal ions (Ca^{2+} , Zn^{2+} , Mg^{2+} , and Hg^{2+}) makes the difference, which reflection of the binding site is remarkable, but the neighboring residues' contribution limits during the crucial process.

The degree of conservation demonstrates the importance of amino acids in evolution. A commonly cited approximation is that the more critical amino acids realize protein function, the less likely they will mutate. Thus, the conservation of amino acid residues is a good indicator of protein-metal ion binding. It was selected as the feature information to develop an effective identification model further.

3.3. Contribution of Features. The feature space contains five feature information, which we introduced in Section 2.2 for classifier learning. We compared the effects of adding different features on the results to verify the selected feature's validity. Table 2 shows that five elements were verified by successively adding them into the classifier in the first stage of the learning process.

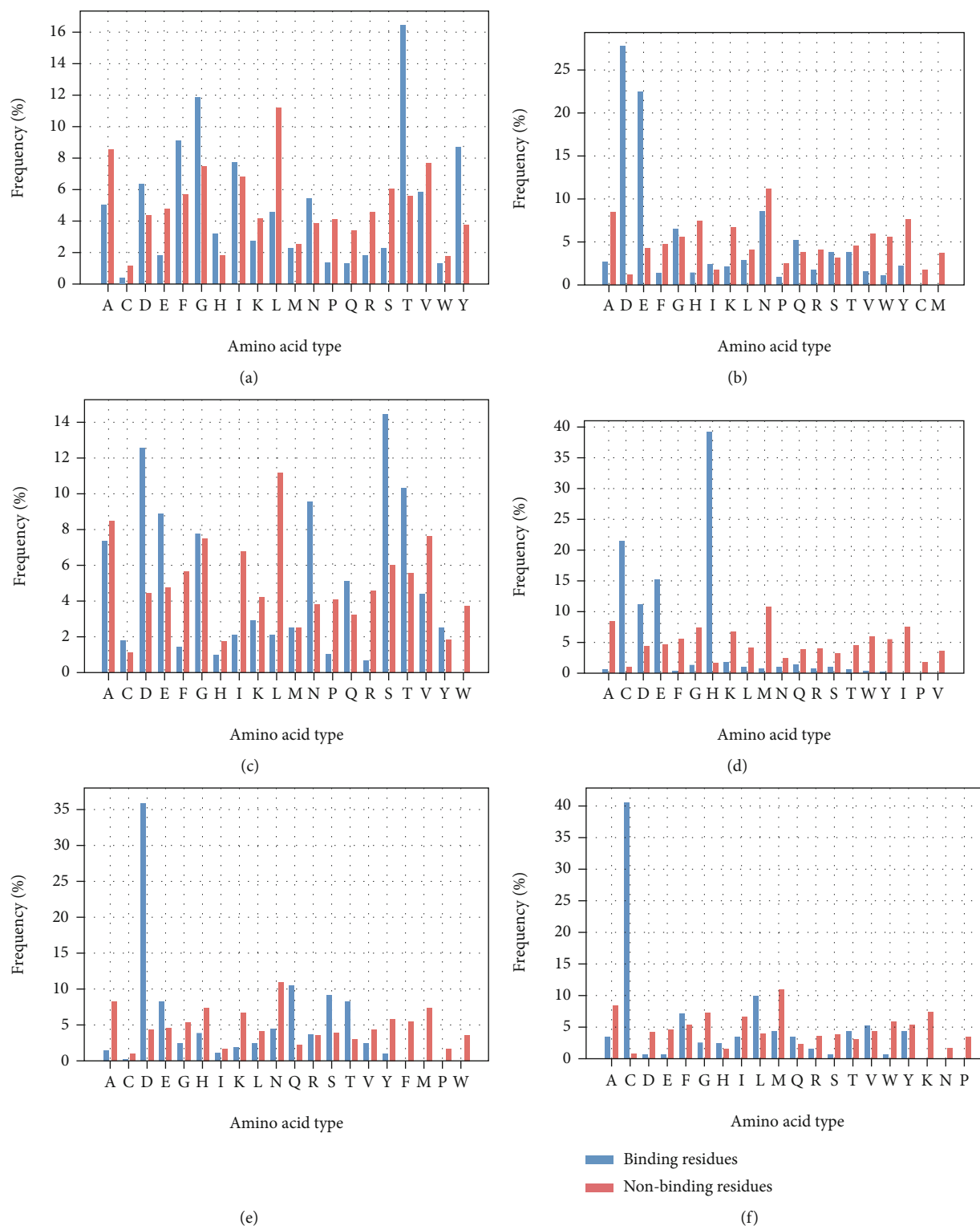


FIGURE 1: The amino acid binding frequency of six metal ions. The frequency of 20 kinds of amino acids on the binding site (blue) and nonbinding (red) was histogram. The abscissa represents the kinds of amino acids, and the ordinate represents the frequency (%); (a), (b), (c), (d), (e), and (f) represents K^+ , Ca^{2+} , Na^+ , Zn^{2+} , Mg^{2+} , and Hg^{2+} , respectively.

It can be seen from Table 2 that adding features in sequence from top to bottom plays a positive role for models in MCC indicators. On the one hand, the five characteristics selected in this experiment can better reflect the critical

information of the TMP sequence and help the model identify the MIBs. On the other hand, five features are relatively independent and can play a more significant role when combined.

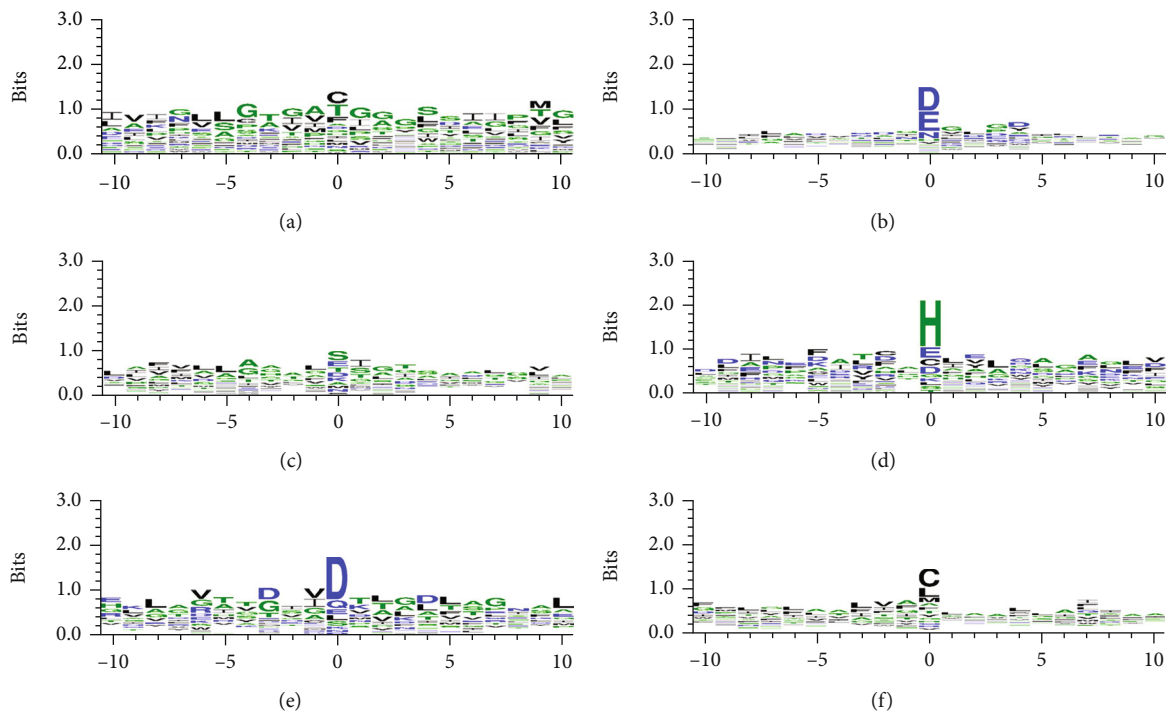


FIGURE 2: Position-specific conservation of amino acid residues. (a) K^+ , (b) Ca^{2+} , (c) Na^+ , (d) Zn^{2+} , (e) Mg^{2+} , and (f) Hg^{2+} .

TABLE 2: The performance of different combinations of features.

Feature combination	ACC	SPE	SEN	MCC	AUC
PSSM	0.695	0.765	0.624	0.395	0.695
PSSM, PCP	0.75	0.801	0.7	0.504	0.75
PSSM, PCP, rASA	0.754	0.832	0.676	0.515	0.754
PSSM, PCP, rASA, Zcoord	0.755	0.834	0.675	0.516	0.755
PSSM, PCP, rASA, Zcoord, TOPO	0.766	0.861	0.67	0.542	0.766

3.4. Random Undersampling Scheme. Exploring the ratio of binding and nonbinding residues is necessary to tackle the sample imbalance problem because too few negative samples will cause the loss of valuable information, and too many negative instances will increase the interference caused by redundant data. Figure 3 depicts the change of evaluation indexes with the shift in positive and negative sample proportion in the first stage. We can see that the MCC value in the 10-fold cross-validation sets shows a decreasing trend with the ratio increase. In contrast, the MCC value on the independent set is fluctuant, but it increased in general. The ratio of negative sample sampling is the key to influence the final results. We used the proportion of positive and negative samples which is 1:5 to improve the model's overall performance.

3.5. Comparison with Other Machine Learning Methods over Cross-Validation. TMP-MIBS is based on the RF algorithm. This section compares the random forest with other machine learning methods on the training dataset, such as support vector machine (SVM), naïve Bayes, and AdaBoost. These methods have shown excellent performance in common classification problems. To obtain fair and objective

experimental results, all models adopt the same dataset and preprocessing mechanism and finally get the test results shown in Table 3.

As shown in Table 3, the integration classes' performance is better than the others. Compare the two ensemble strategies AdaBoost and RF. The former adopts a boosting approach to ensemble base learners that adjust according to the previous one to generate prediction results serially, making the model susceptible to noise and outliers. Instead, the RF adopts a bagging strategy to make the base learner relatively independent and has no strong dependency. It can generate the prediction results in parallel, reduce outliers' influence on them, and have the natural advantage of solving the multidimensional unbalanced data. We further compared the prediction performance of different classifiers for each metal ion. The comparison similarly shows that the overall performance of the random forest classifier is optimal.

3.6. Comparison with Other Ligand-Specific Methods. To prove TMP-MIBS's robustness and effectiveness, we further tested the model on the independent testing dataset and compared it with two publicly available methods, including

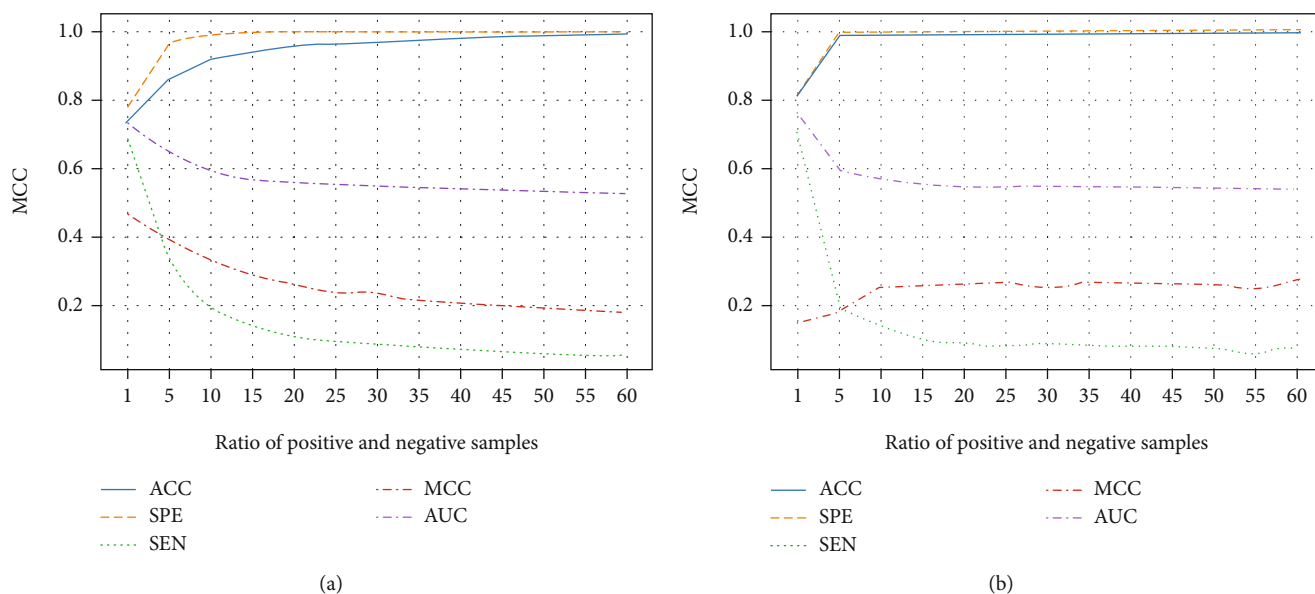


FIGURE 3: The ratio of nonbinding residues and binding residues. (a) 10-fold cross-validation test. (b) Independent validation test.

TABLE 3: Comparison of RF with other classifiers.

Classifier	ACC	SPE	SEN	MCC	AUC
SVM	0.658	0.648	0.703	0.267	0.676
Naïve Bayes	0.767	0.775	0.73	0.409	0.752
AdaBoost	0.808	0.804	0.649	0.428	0.745
RF	0.795	0.808	0.73	0.447	0.769

TABLE 4: Comparison with publicly available methods.

Ligand	Method	ACC	SPE	SEN	MCC	AUC
K^+	TMP-MIBS	0.981	1	0.118	0.34	0.559
Ca^{2+}	MIB	0.942	0.943	0.342	0.067	0.643
	TargetS	0.997	0.999	0.471	0.494	0.735
Na^+	TMP-MIBS	0.908	0.998	0.196	0.398	0.597
	TargetS	0.998	0.999	0.259	0.336	0.629
Zn^{2+}	TMP-MIBS	0.901	0.997	0.304	0.501	0.65
	MIB	0.945	0.946	0.538	0.101	0.742
Mg^{2+}	TargetS	0.996	0.997	0.231	0.151	0.614
	TMP-MIBS	0.979	1	0.154	0.388	0.577
Hg^{2+}	MIB	0.932	0.933	0.053	0	0.493
	TMP-MIBS	0.991	0.998	0.259	0.356	0.628
Others	MIB	0.948	0.949	0.56	0.104	0.754
	TMP-MIBS	0.973	0.998	0.259	0.502	0.63
Others	TMP-MIBS	0.976	0.987	0.056	0.041	0.521

TargetS [25] and MIB [24]. The prediction performance was calculated based on the same dataset (Tables 1 and 4). For the TargetS method, a ligand-specific template-free protein-ligand binding site predictor used classifier ensemble and spatial clustering. It has five metal ligands that overlap with this study. We submitted the protein sequence into

the webserver (<http://www.csbio.sjtu.edu.cn/TargetS/>) to obtain the predicted results and evaluate predictive performance. For the MIB method, which constructs metal ion binding templates for structural comparison between query proteins and templates and has four metal ligands identical to this study, we submitted and ran the MIB webserver (<http://bioinfo.cmu.edu.tw/MIB/>).

We observed that the performance of TMP-MIBS significantly outperforms the MIB on four metal ions. The average MCC value of Na^+ , Zn^{2+} , and Hg^{2+} is about 16–39% higher than the TargetS. The results show that our model is superior to the available metal ion predictors, whether template-based or non-template-based methods. It can be inferred that the results largely depend on our input data rather than the complicated method. Although the number of TMPs sequences is increasing, it is still quite limited compared with non-TMPs. MIB and TargetS training models do not distinguish TMPs, so the models mainly learn the information of non-TMPs. The differences between the TMPs and non-TMPs are reflected in the secondary structure through sequence information and determine their tertiary conformation and function. TMP-MIBS focuses on TMPs, and the final results also confirm our efforts.

3.7. Metal Ion Binding Motif Analysis. A motif is an approximate sequence pattern that repeatedly occurs in a group of related sequences. It was used to reflect the protein’s conservative information and discover novel information between different sequences. We tried to find out the motif within the metal ion binding domains to discover potential drug targets. The seven group metal ion binding domains were extracted for analysis. Figure 4 shows the sequence logos of motifs for six metal ions and the 3D visualizations of their examples. Note that we stipulate the MEME outputs with ten motifs for each metal ion class and select the highest E -value for reporting.

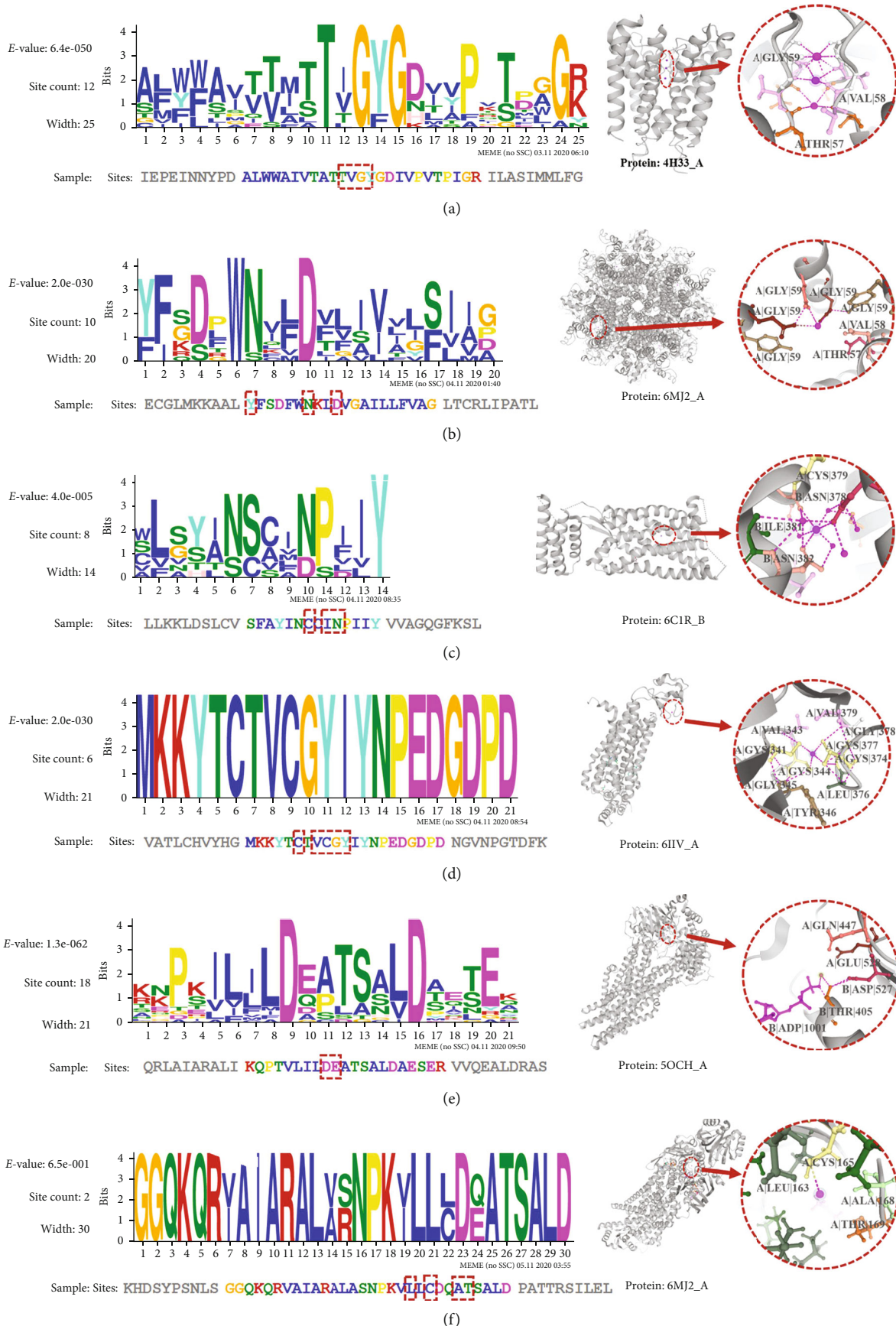


FIGURE 4: Continued.

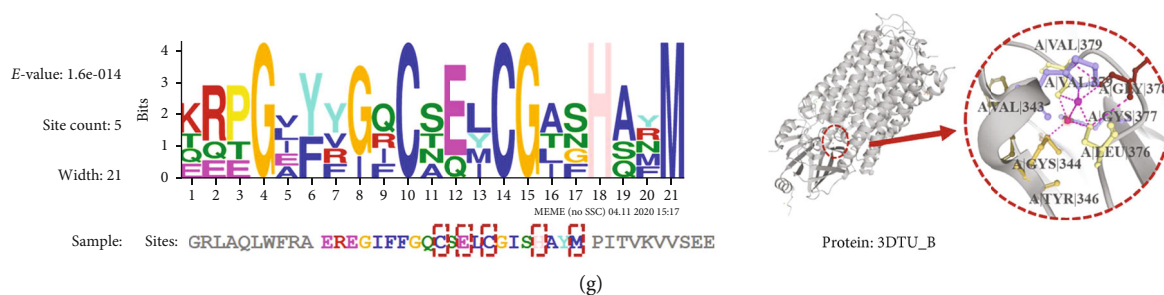


FIGURE 4: Sequence logos of motif within the seven metal ion binding domains. (a) K^+ , (b) Ca^{2+} , (c) Na^+ , (d) Zn^{2+} , (e) Mg^{2+} , (f) Hg^{2+} , and (g) others.

From Figures 4(a) to 4(g), describe the logo of K^+ , Ca^{2+} , Na^+ , Zn^{2+} , Mg^{2+} , Hg^{2+} , and (g) other metal ions respectively. “E-value” is an estimate of the expected number of motifs with the given log-likelihood ratio (or higher). “Site Count” represents the number of sites contributing to the construction of the motif. “Width” represents the width of the motif. Sequences where each position is independent and letters are chosen according to the background letter frequencies. The red dashed box indicates the TMP-MIBS prediction site. The 3D visualization on the right is an example of the corresponding motif. “Protein” represents the PDB ID_Chain (domain).

The relative size of letters indicates their frequency in the sequence. It can be seen from the figure that the higher the letter, the more likely it is to become a binding site. Based on the extraction of motif sequence, we can predict the potential binding sites, which is helpful to understand further the biological significance involved in various biological processes.

4. Conclusions

Metal ions regulate almost all organisms’ physiological cell functions, and their abnormal homeostasis usually leads to a variety of diseases and pathogenic states. They achieve homeostasis inside and outside the membrane and perform essential biological functions with TMPs’ assistance. This study proposed an effective method to predict the binding residues of seven class metal ions in TMPs. We used the combination of conservative structure, physical and chemical properties, topological structure, solution accessibility, and Z-coordinate to apply the random forest algorithm to identify metal ion binding residues. These characteristics positively affected the prediction in essence. Test results show that TMP-MIBS has excellent performance for metal ion binding residues. This indicates that the sequential approach alone can achieve pleasant performance and demonstrates the importance of input data. With more and more sequence information obtained in the future, our model will show more excellent performance.

In the current work, a significant problem of TMB-MIBS is that predicting fewer MIBs on the protein sequence is still challenging. However, it can accurately predict more sites than existing tools because the imbalance of positive and negative samples is the unavoidable normal state of such

problems. We will work to overcome this problem as the goal of the next phase.

Data Availability

TMP-MIBS’s code and dataset are available at https://github.com/QuJing785464/TMP_MIBS.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (Nos. 2412019FZ052 and 2412019FZ048).

References

- [1] K. M. Fracchia, C. Y. Pai, and C. M. Walsh, “Modulation of T cell metabolism and function through calcium signaling,” *Frontiers in Immunology*, vol. 4, p. 324, 2013.
- [2] K. Mubagwa, A. Gwanyanya, S. Zakharov, and R. Macianskiene, “Regulation of cation channels in cardiac and smooth muscle cells by intracellular magnesium,” *Archives of Biochemistry and Biophysics*, vol. 458, no. 1, pp. 73–89, 2007.
- [3] E. Mocchegiani, C. Bertoni-Freddari, F. Marcellini, and M. Malavolta, “Brain, aging and neurodegeneration: role of zinc ion availability,” *Progress in Neurobiology*, vol. 75, no. 6, pp. 367–390, 2005.
- [4] A. Takeda and H. Tamano, “Significance of the degree of synaptic Zn^{2+} signaling in cognition,” *Biomaterials*, vol. 29, no. 2, pp. 177–185, 2016.
- [5] C. J. Frederickson, M. A. Klitenick, W. I. Manton, and J. B. Kirkpatrick, “Cytoarchitectonic distribution of zinc in the hippocampus of man and the rat,” *Brain Research*, vol. 273, no. 2, pp. 335–339, 1983.
- [6] C. C. Bridges and R. K. Zalups, “The aging kidney and the nephrotoxic effects of mercury,” *Journal of Toxicology and Environmental Health. Part B, Critical Reviews*, vol. 20, no. 2, pp. 55–80, 2017.
- [7] K. Grzeszczak, S. Kwiatkowski, and D. Kosik-Bogacka, “The role of Fe, Zn, and Cu in pregnancy,” *Biomolecules*, vol. 10, no. 8, p. 1176, 2020.

- [8] M. R. Kapkaeva, O. V. Popova, R. V. Kondratenko et al., "Effects of copper on viability and functional properties of hippocampal neurons in vitro," *Experimental and Toxicologic Pathology*, vol. 69, no. 5, pp. 259–264, 2017.
- [9] K. Kubota, M. Dahbi, T. Hosaka, S. Kumakura, and S. Komaba, "Towards K-ion and Na-ion batteries as "beyond Li-ion", " *Chemical Record*, vol. 18, no. 4, pp. 459–479, 2018.
- [10] C. Wang, R. Zhang, X. Wei, M. Lv, and Z. Jiang, "Metalloimmunology: the metal ion-controlled immunity," *Advances in Immunology*, vol. 145, pp. 187–241, 2020.
- [11] J. H. de Baaij, J. G. Hoenderop, and R. J. Bindels, "Magnesium in man: implications for health and disease," *Physiological Reviews*, vol. 95, no. 1, pp. 1–46, 2015.
- [12] S. J. Fleishman, V. M. Unger, and N. Ben-Tal, "Transmembrane protein structures without X-rays," *Trends in Biochemical Sciences*, vol. 31, no. 2, pp. 106–113, 2006.
- [13] A. Guna and R. S. Hegde, "Transmembrane domain recognition during membrane protein biogenesis and quality control," *Current Biology*, vol. 28, no. 8, pp. R498–r511, 2018.
- [14] H. Y. Gee, J. Kim, and M. G. Lee, "Unconventional secretion of transmembrane proteins," *Seminars in Cell & Developmental Biology*, vol. 83, pp. 59–66, 2018.
- [15] C. Wallin, S. B. Sholts, N. Österlund et al., "Alzheimer's disease and cigarette smoke components: effects of nicotine, PAHs, and Cd(II), Cr(III), Pb(II), Pb(IV) ions on amyloid- β peptide aggregation," *Scientific Reports*, vol. 7, no. 1, p. 14423, 2017.
- [16] M. Butler and G. Cabrera, "A mass spectrometry-based method for differentiation of positional isomers of monosubstituted pyrazine N-oxides using metal ion complexes," *Journal of Mass Spectrometry*, vol. 50, no. 1, pp. 136–144, 2015.
- [17] S. Feng, C. Pan, X. Jiang et al., "Fe³⁺ immobilized metal affinity chromatography with silica monolithic capillary column for phosphoproteome analysis," *Proteomics*, vol. 7, no. 3, pp. 351–360, 2007.
- [18] G. Kaur-Atwal, D. J. Weston, P. S. Green, S. Crosland, P. L. R. Bonner, and C. S. Creaser, "On-line capillary column immobilised metal affinity chromatography/electrospray ionisation mass spectrometry for the selective analysis of histidine-containing peptides," *Journal of Chromatography B*, vol. 857, no. 2, pp. 240–245, 2007.
- [19] P. Rondeau, S. Sers, D. Jhurry, and F. Cadet, "Sugar interaction with metals in aqueous solution: indirect determination from infrared and direct determination from nuclear magnetic resonance spectroscopy," *Applied Spectroscopy*, vol. 57, no. 4, pp. 466–472, 2003.
- [20] K. B. Handing, E. Niedzialkowska, I. G. Shabalin, M. L. Kuhn, H. Zheng, and W. Minor, "Characterizing metal-binding sites in proteins with X-ray crystallography," *Nature Protocols*, vol. 13, no. 5, pp. 1062–1090, 2018.
- [21] M. N. Mbaye, Q. Hou, S. Basu, F. Teheux, F. Pucci, and M. Rooman, "A comprehensive computational study of amino acid interactions in membrane proteins," *Scientific Reports*, vol. 9, no. 1, pp. 12043–12043, 2019.
- [22] J. Yang, A. Roy, and Y. Zhang, "Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment," *Bioinformatics*, vol. 29, no. 20, pp. 2588–2595, 2013.
- [23] W. Zhao, M. Xu, Z. Liang et al., "Structure-based de novo prediction of zinc-binding sites in proteins of unknown function," *Bioinformatics*, vol. 27, no. 9, pp. 1262–1268, 2011.
- [24] Y.-F. Lin, C.-W. Cheng, C.-S. Shih, J.-K. Hwang, C.-S. Yu, and C.-H. Lu, "MIB: metal ion-binding site prediction and docking server," *Journal of Chemical Information and Modeling*, vol. 56, no. 12, pp. 2287–2291, 2016.
- [25] D. J. Yu, J. Hu, J. Yang, H. B. Shen, J. Tang, and J. Y. Yang, "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 4, pp. 994–1008, 2013.
- [26] X. Hu, Q. Dong, J. Yang, and Y. Zhang, "Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transfers," *Bioinformatics*, vol. 32, no. 21, pp. 3260–3269, 2016.
- [27] X. Cao, X. Hu, X. Zhang et al., "Identification of metal ion binding sites based on amino acid sequences," *PLoS One*, vol. 12, no. 8, article e0183756, 2017.
- [28] S. Kumar, "Prediction of metal ion binding sites in proteins from amino acid sequences by using simplified amino acid alphabets and random forest model," *Genomics & Informatics*, vol. 15, no. 4, pp. 162–169, 2017.
- [29] L. Qiao and D. Q. Xie, "MionSite: ligand-specific prediction of metal ion-binding sites via enhanced AdaBoost algorithm with protein sequence information," *Analytical Biochemistry*, vol. 566, pp. 75–88, 2019.
- [30] I. Haberal and H. Ogul, "Prediction of protein metal binding sites using deep neural networks," *Molecular Informatics*, vol. 38, no. 7, article 1800169, 2019.
- [31] J. Abbass and J.-C. Nebel, "Rosetta and the journey to predict proteins' structures, 20 years on," *Current Bioinformatics*, vol. 15, no. 6, pp. 611–628, 2020.
- [32] C. Peng, S. Tong, Z. Youzhi, and W. Bing, "A sequence-segment neighbor encoding schema for protein hotspot residue prediction," *Current Bioinformatics*, vol. 15, no. 5, pp. 445–454, 2020.
- [33] I. Murugan, R. N. Ahmed, and A. Subramanian, "A weighted association rule mining method for predicting HCV-human protein interactions," *Current Bioinformatics*, vol. 13, no. 1, pp. 73–84, 2018.
- [34] K. Neetu and V. Anshul, "Analysis of oncogene protein structure using small world network concept," *Current Bioinformatics*, vol. 15, no. 7, pp. 732–740, 2020.
- [35] D. Kozma, I. Simon, and G. E. Tusnády, "PDBTM: protein data bank of transmembrane proteins after 8 years," *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D524–D529, 2013.
- [36] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.
- [37] P. Chen, S. Hu, J. Zhang et al., "A sequence-based dynamic ensemble learning system for protein ligand-binding site prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 5, pp. 901–912, 2016.
- [38] A. Rausell, D. Juan, F. Pazos, and A. Valencia, "Protein interactions and ligand binding: from protein subfamilies to functional specificity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 5, pp. 1995–2000, 2010.
- [39] H. Guohua and L. Jincheng, "Feature extractions for computationally predicting protein post-translational modifications," *Current Bioinformatics*, vol. 13, no. 4, pp. 387–395, 2018.

- [40] P. Argos, J. K. Rao, and P. A. Hargrave, "Structural prediction of membrane-bound proteins," *European Journal of Biochemistry*, vol. 128, no. 2-3, pp. 565–575, 1982.
- [41] H. Gv, "Membrane proteins: from sequence to structure," *Annual Review of Biophysics and Biomolecular Structure*, vol. 23, no. 1, pp. 167–192, 1994.
- [42] G. Heijne, "The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology," *The EMBO Journal*, vol. 5, no. 11, pp. 3021–3027, 1986.
- [43] H. Shida, G. Fei, Z. Quan, and H. Ding, "MRMD2.0: a python tool for machine learning with feature ranking and reduction," *Current Bioinformatics*, vol. 15, no. 10, pp. 1213–1221, 2021.
- [44] T. Dudev, "Modeling metal binding sites in proteins by quantum chemical calculations," *Computational Chemistry*, vol. 2, no. 2, pp. 19–21, 2014.
- [45] K. Shuichi, P. Piotr, P. Maria, K. Andrzej, K. Toshiaki, and K. Minoru, "AAindex: amino acid index database, progress report 2008," *Nucleic Acids Research*, vol. 36, no. Database issue, pp. D202–D205, 2007.
- [46] A. T. Smith, D. Barupala, T. L. Stemmler, and A. C. Rosenzweig, "A new metal binding domain involved in cadmium, cobalt and zinc transport," *Nature Chemical Biology*, vol. 11, no. 9, pp. 678–684, 2015.
- [47] X. Yin, J. Yang, F. Xiao, Y. Yang, and H. B. Shen, "MemBrain: an easy-to-use online webserver for transmembrane protein structure prediction," *Nano-Micro Letters*, vol. 10, no. 1, pp. 1–8, 2018.
- [48] K. D. Tsirigos, C. Peters, N. Shu, L. Käll, and A. Elofsson, "The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides," *Nucleic Acids Research*, vol. 43, no. W1, pp. W401–W407, 2015.
- [49] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes¹," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, 2001.
- [50] E. Granseth, H. Viklund, and A. Elofsson, "ZPRED: predicting the distance to the membrane center for residues in α -helical membrane proteins," *Bioinformatics*, vol. 22, no. 14, pp. e191–e196, 2006.
- [51] C. Lu, Y. Gong, Z. Liu, Y. Guo, Z. Ma, and H. Wang, "TM-ZC: a deep learning-based predictor for the Z-coordinate of residues in α -helical transmembrane proteins," *IEEE Access*, vol. 8, pp. 40129–40137, 2020.
- [52] A. Bernsel, H. Viklund, A. Hennerdal, and A. Elofsson, "TOPCONS: consensus prediction of membrane protein topology," *Nucleic Acids Research*, vol. 37, no. Web Server, pp. W465–W468, 2009.
- [53] J. S. Chauhan, N. K. Mishra, and G. P. Raghava, "Identification of ATP binding residues of a protein from its primary sequence," *BMC Bioinformatics*, vol. 10, no. 1, p. 434, 2009.
- [54] Z. Lv, P. Wang, Q. Zou, and Q. Jiang, "Identification of sub-Golgi protein localization by use of deep representation learning features," *Bioinformatics (Oxford, England)*, vol. 36, no. 24, pp. 5600–5609, 2021.
- [55] Z. Lv, F. Cui, Q. Zou, L. Zhang, and L. Xu, "Anticancer peptides prediction with deep representation learning features," *Briefings in Bioinformatics*, vol. 22, no. 5, 2021.
- [56] Z. Lv, S. Jin, H. Ding, and Q. Zou, "A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 215, 2019.
- [57] Z. Lv, J. Zhang, H. Ding, and Q. Zou, "RF-PseU: a random forest predictor for RNA pseudouridine sites," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 134, 2020.
- [58] G. Crooks, G. Hon, J.-M. Chandonia, and S. Brenner, "WebLogo: a sequence logo generator," *Genome Research*, vol. 14, no. 6, pp. 1188–1190, 2004.

Research Article

iMPT-FDNPL: Identification of Membrane Protein Types with Functional Domains and a Natural Language Processing Approach

Wei Chen ¹, Lei Chen ¹, and Qi Dai ²

¹College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

²College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China

Correspondence should be addressed to Lei Chen; chen_lei1@163.com and Qi Dai; daailiu04@yahoo.com

Received 22 August 2021; Revised 15 September 2021; Accepted 27 September 2021; Published 11 October 2021

Academic Editor: Hui Ding

Copyright © 2021 Wei Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Membrane protein is an important kind of proteins. It plays essential roles in several cellular processes. Based on the intramolecular arrangements and positions in a cell, membrane proteins can be divided into several types. It is reported that the types of a membrane protein are highly related to its functions. Determination of membrane protein types is a hot topic in recent years. A plenty of computational methods have been proposed so far. Some of them used functional domain information to encode proteins. However, this procedure was still crude. In this study, we designed a novel feature extraction scheme to obtain informative features of proteins from their functional domain information. Such scheme termed domains as words and proteins, represented by its domains, as sentences. The natural language processing approach, word2vector, was applied to access the features of domains, which were further refined to protein features. Based on these features, Random k-labelsets with random forest as the base classifier was employed to build the multilabel classifier, namely, iMPT-FDNPL. The tenfold cross-validation results indicated the good performance of such classifier. Furthermore, such classifier was superior to other classifiers based on features derived from functional domains via one-hot scheme or derived from other properties of proteins, suggesting the effectiveness of protein features generated by the proposed scheme.

1. Introduction

Membrane protein refers to the protein that can bind to the cell membrane and is an important part of the cell membrane. It exposes a surface that is very suitable for merging to the membrane [1]. There are lots of membrane proteins in human. They perform various functions related to cell survival. About 30% of genes can encode membrane proteins [2], 60% of membrane proteins can be used as drug targets, and some membrane proteins can act as enzyme mediators in the immune system [3]. It is reported that the function of membrane protein is highly associated with its type. Identification of the types of membrane proteins is an important step to uncover their functions. Traditional experimental methods can provide solid results. However, they have some evident defects, such as low efficiency and high cost. The large-scale tests for identification of membrane

protein types via these methods are almost impossible. Thus, it is urgent to design quick and cheap methods.

In recent years, lots of new computational methods have proposed, providing strong technical support for designing classifiers for identification of membrane protein types. On the other hand, several online databases have been set up for collecting various information of proteins, giving strong data support. To date, several classifiers have been proposed to identify membrane protein types. Most classifiers are based on machine learning algorithms. These classifiers always learn patterns based on the information of membrane proteins, whose types have been determined. These patterns can be used to determine the types of given proteins. Several existing classifiers used features extracted from protein sequences [4–9]. Amino acid composition (AAC) and pseudo amino acid composition (PseAAC) are two classic schemes to access features from protein sequences. Functional domains are also

used to build classifiers for identification of membrane protein types [10–12]. The classifiers incorporating such information always provided good performance. However, the usage of functional domain information is still at a low level. One-hot scheme was used to encode proteins based their functional domain information. Through this scheme, each protein was encoded into a binary vector, where each component represented one domain. If the domain was annotated on a given protein, its corresponding component was set to one; otherwise, it was set to zero. However, such scheme had some evident defects. For example, the performance of the classifiers was quite sensitive to some domains. This study gave an investigation on the usage of functional domain information of proteins.

In this study, we set up a novel classifier to identify membrane protein types. This classifier adopted the novel features obtained from functional domain information of proteins via a natural language processing approach, word2-vector. These features were fed into a multilabel classification scheme, RANdom k-labELsets (RAKEL) [13], to set up the classifier. Classic classification algorithm, random forest (RF) [14], was selected as the base classifier in RAKEL. The proposed classifier was called iMPT-FDNPL. The ten-fold cross-validation indicated the good performance of such classifier. It was also superior to other classifiers that were constructed with other widely used feature extraction schemes, including the classifier using features derived from functional domain information via one-hot scheme.

2. Materials and Methods

2.1. Database. The data of human membrane proteins was sourced from Huang et al.’s study (dataset S1) [15]. 2883 membrane proteins, encoded by UniProt IDs, were obtained. In fact, these proteins were extracted from a larger dataset retrieved from the UniProt database (release 2012_09) [16] by using CD-HIT [17]. The sequence similarity of any two proteins was smaller than 0.7. These 2883 proteins were classified into six types: (1) GPI- (glycosyl phosphatidyl isohydrin-) anchored, (2) lipid-anchor, (3) multipass, (4) peripheral, (5) single-channel type I, and (6) single-pass II type [18]. Because we adopted functional domain information to encode proteins, those without such information were excluded. 2729 membrane proteins remained. These proteins were still classified into six abovementioned types. The distribution of 2729 membrane proteins on six types is shown in Table 1. The sum of protein numbers in all six types was 2810 (last row of Table 1), which was bigger than the number of different proteins. It was suggested that some proteins belonged to more than one types. As shown in Figure 1, 73 proteins belonged to two types, 4 proteins belonged to three types, whereas rest proteins belonged to one type. Thus, it is a multilabel classification problem to assign types to membrane proteins.

2.2. Feature Engineering. Feature engineering is an important step in designing efficient classifiers. In this study, we should extract features from each membrane protein, which can retain essential properties of proteins. Functional domain is

TABLE 1: Distribution of membrane proteins on six types.

Membrane protein type	Number of proteins
GPI-anchor	69
Lipid-anchor	211
Multipass	1306
Peripheral	530
Single-pass type I	539
Single-pass type II	155
Total	2810

widely used to investigate various protein-related problems, including membrane protein type prediction. The classic way to employ such information is one-hot scheme. Several classifiers have been built with such scheme, and they provided good performance [10–12]. As mentioned above, such scheme also had some defects. Here, we proposed a new scheme to adopt functional domain information, thereby encoding membrane proteins in a new way.

2.2.1. Domain Representation. The functional domain information of all human proteins was retrieved from the InterPro database (<http://ftp.ebi.ac.uk/pub/databases/interpro/>, accessed in February 2021) [19]. 17,410 IPR terms were annotated on 171,472 human proteins. In this study, we adopted a natural language processing approach to analyze this information. To this end, IPR terms were deemed as words and proteins, represented by one or more IPR terms, were termed as sentences. Accordingly, the well-known word2vector method was applied on them to learn a feature vector for each IPR term. This study used the word2vector program obtained from <https://github.com/RaRe-Technologies/gensim>. Default parameters were adopted.

2.2.2. Protein Representation. As mentioned above, the feature vector of each IPR term was learnt by word2vector. Based on them, we can further access the feature vectors of proteins. Here, a simple way was adopted. The feature vector of a given protein was defined as the average vector of feature vectors of IPR terms that was annotated on such protein. For example, for a certain protein A4D1S5, there are three IPR terms, say IPR001806, IPR005225, IPR027417, and the average vector of three vectors, representing above three IPR terms, respectively, was used to represent A4D1S5.

2.3. Multilabel Classifier. This study adopted a problem transformation method, RAKEL [13], to build the multilabel classifier, which has wide applications in dealing with several biological and medicine problems [20–27]. From the original multilabel classification problem, several single-label classification problems are derived as follows. Given a problem with l labels, denoted by L_1, L_2, \dots, L_l , it first randomly constructs m label subsets, each of which contains k labels, where $1 \leq k \leq l$. For each label subset, members in its power set are deemed as new labels. Samples are assigned new labels according to their original labels. For example, for the label subset $\{L_1, L_2, L_3\}$, the labels of each sample are

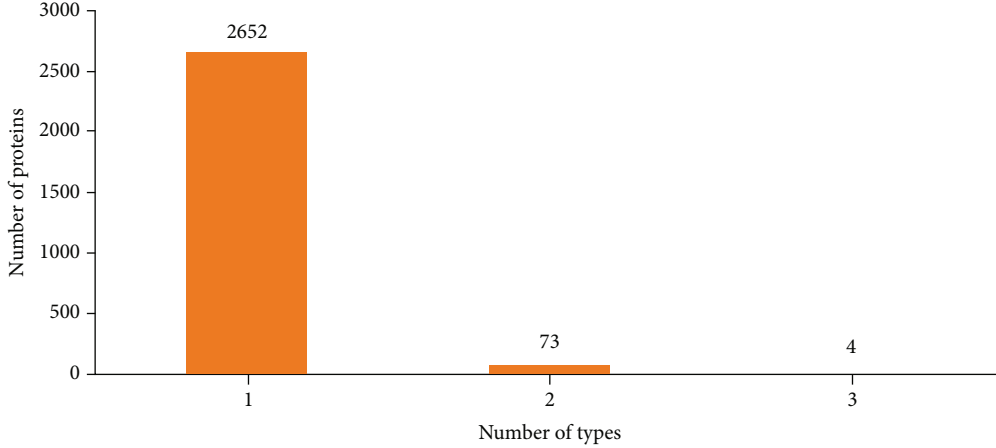


FIGURE 1: An illustration to show the distribution on the number of types a membrane belongs to. Four membrane proteins belong to three types, 73 proteins belong to two types, and rest 2652 proteins belong to one type.

first restricted to this subset, i.e., labels in this subset are picked up and the rest are discarded. Then, the remaining labels are put together as a new label. If the labels for one sample are L_1, L_2 and L_4 , L_1 and L_2 are first selected and $\{L_1, L_2\}$, a member of the power set of $\{L_1, L_2, L_3\}$, is assigned to such sample as its new label. Accordingly, each sample has exactly one new label. Then, a classifier can be built with a given base single-label classifier. The m label subsets induce m single-label classifiers. The final multilabel classifier integrates these single-label classifiers. In detail, given a query sample, each single-label classifier provides its prediction. Such prediction can be refined to the binary predictions for labels involved in this classifier. For each label, the binary predictions yielded by classifiers involving this label are selected and count the proportion of classifiers that predict this label. If this proportion is higher than a predefined threshold, which is always set to 0.5, the label is assigned to the query sample.

To quickly implement the RAKEL algorithm, we used the tool “RAKEL” in Meka [28], retrieved from <http://waikato.github.io/meke/>. Several values of m and k , the main parameters of RAKEL, were tried in this study. For convenience, the classifiers built by RAKEL were termed as RAKEL classifiers.

2.4. Base Classifier. The multilabel classifier built by RAKEL needs a base single-label classifier as mentioned above. One of the most classic algorithms, RF [14], was selected in this study. It is an ensemble classifier, consisting of several decision trees. Each decision tree is constructed by randomly selecting samples and features. Given a sample, each decision tree provides its prediction. RF counts these predictions and determines the final prediction using majority voting. Although decision tree is quite weak, RF is much more robust. Thus, it is always an important candidate to build classifiers for tackling different problems [29–39].

In this study, we adopted the tool “RandomForest” integrated in Meka [28], which implements RF.

2.5. Performance Measurement. All classifiers were assessed by tenfold cross-validation [40–44]. This method randomly and equally divides samples into ten subsets. Each subset is singled out to constitute the test set one by one, and rest subsets are put together to constitute the training set. Accordingly, each sample is predicted only once.

After obtaining the outcomes of tenfold cross-validation, we calculated three measurements to assess the quality of results, including exact matching, accuracy, and hamming loss [25–27], which can be computed by

$$\begin{cases} \text{Exact match} = \frac{1}{n} \sum_{i=1}^n \nabla(L_i, L_i'), \\ \text{Accuracy} = \frac{1}{n} \sum_{i=1}^n \left(\frac{\|L_i \cap L_i'\|}{\|L_i \cup L_i'\|} \right), \\ \text{Hamming loss} = \frac{1}{n} \sum_{i=1}^n \left(\frac{\|L_i \Delta L_i'\|}{m} \right), \end{cases} \quad (1)$$

where n denotes the overall number of samples, m stands for the number of labels ($m = 6$ in this study), L_i and L_i' represent the set of true labels and predicted labels of the i^{th} sample, respectively, Δ stands for the set symmetric difference operation, and ∇ is defined as follows:

$$\nabla(L_i, L_i') = \begin{cases} 1 & \text{If } L_i \text{ is identical to } L_i', \\ 0 & \text{Otherwise.} \end{cases} \quad (2)$$

Obviously, the higher exact matching and the accuracy, the better the performance of the classifier. For hamming loss, the lower the hamming loss, the better the performance. For easy comparisons, an integrated measurement, called integrated score, was defined as below

$$\text{Integrated score} = \text{exact match} * \text{accuracy} * (1 - \text{hamming loss}). \quad (3)$$

The higher the score, the better the classifier.

3. Results and Discussion

In this study, we set up a multilabel classifier, iMPT-FDNPL, for prediction of membrane protein types. Such classifier employed the features derived from functional domain information of proteins. The entire procedures are shown in Figure 2. In this section, we would give the evaluation results and comparisons with other classifiers.

3.1. Performance of iMPT-FDNPL. iMPT-FDNPL adopted the features derived from functional domain information via word2vector. Because the optimum dimension of features was unknown, several dimensions were tried, including dimensions from 50 to 500 with interval 50. Furthermore, the main parameter m in RAKEL was set to 10, and another main parameter k was set to all integers between 2 and 6. As for the parameter of RF, number of decision trees, it was set to integers from 100 to 500 with interval 100. RAKEL classifiers with all possible parameter settings were set up and assessed by tenfold cross-validation. The outcomes showed that when the dimension was set to 350, $k = 6$, $m = 10$, and the number of decision trees was 500, the RAKEL classifiers provided the highest integrated score of 0.6874. Thus, this classifier was the proposed multilabel classifier, iMPT-FDNPL. The exact match, accuracy, and hamming loss were 0.851, 0.853, and 0.053, respectively, which are listed in Table 2. The exact match and accuracy both exceed 0.850, suggesting the good performance of iMPT-FDNPL.

To fully assess the performance of iMPT-FDNPL under tenfold cross-validation, 20 additional tenfold cross-validations on this classifier were conducted. The obtained values of exact matching, accuracy, hamming loss, and integrated score are illustrated in Figure 3. We can see that exact match varied from 0.853 to 0.860, accuracy from 0.856 to 0.863, hamming loss from 0.049 to 0.052, and integrated score from 0.6921 to 0.7058. Above four measurements varied in a small interval, implying that the performance of iMPT-FDNPL was quite stable no matter how samples were divided.

3.2. Comparison of RAKEL Classifiers with Other Base Classifiers. The proposed classifier, iMPT-FDNPL, adopted RF as the base classifier. In fact, we also attempted another classic classification algorithm, support vector machine (SVM) [45]. Similar to RF, the tool “SMO” integrated in Meka was directly employed in this study, which implements one type of SVM, whose training procedures are optimized by the sequential minimal optimization algorithm [46, 47]. The kernel was polynomial kernel or RBF kernel. Various values of regularization parameter C were tried, including 1, 2, 3, and 4. The exponent of polynomial kernel was set to 1, 2, 3, and 4. As for parameter γ of RBF kernel, it was set to various values between 0.01 and 0.05. The feature dimensions and m, k in RAKEL were the same as those in Section 3.1. All RAKEL classifiers with possible parameter

settings were built and evaluated by tenfold cross-validation. The best performance (highest integrated score) of RAKEL classifiers with SVM using two different kernels is listed in Table 2. If the basic classifier was SVM (polynomial kernel), the integrated score was 0.6515, exact match was 0.831, accuracy was 0.834, and hamming loss was 0.060. If SVM (RBF kernel) was the base classifier, the integrated score was 0.6787, exact match was 0.846, accuracy was 0.848, and hamming loss was 0.054. The comparisons of those yielded by iMPT-FDNPL indicated that the proposed classifier was superior to these RAKEL classifiers. It was proper to select RF as the base classifier to construct the classifier.

3.3. Comparison of BR Classifiers. In this study, we adopted RAKEL to build the multilabel classifier. Here, another multilabel classifier construction method, Binary Relevance (BR) [48], was employed to build the classifiers. Similar to RAKEL, it also needs one base classifier. We still used three base classifiers mentioned above: RF, SVM with polynomial kernel, and SVM with RBF kernel. We tried the same parameter settings as those in above sections. With all possible parameter settings, several classifiers were set up and assessed by tenfold cross-validation. For convenience, these classifiers were called BR classifiers.

The best performance of BR classifiers with different base classifiers is listed in Table 2. The integrated scores of these BR classifiers were 0.5778, 0.6152, and 0.6544, respectively, which were all lower than that of the iMPT-FDNPL. Furthermore, the exact match and accuracy of iMPT-FDNPL were also higher than the corresponding measurements of three BR classifiers. As for hamming loss, iMPT-FDNPL provided lower performance than BR classifier with SVM (RBF kernel) as the base classifier. However, the hamming loss of iMPT-FDNPL was lower than those of other two BR classifiers. All these results indicated the superiority of the iMPT-FDNPL. In addition, given a base classifier, RAKEL classifiers always provided higher performance than BR classifiers, implying RAKEL was more powerful to construct multilabel classifiers for identifying membrane protein types than BR.

3.4. Comparison of Classifiers with Other Embedding Features. In this study, the multilabel classifier, iMPT-FDNPL, adopted features derived from functional domains via a natural language processing approach to encode membrane proteins. As mentioned above, one-hot scheme is a more widely used way to encode proteins. Here, each protein was encoded by such scheme. Then, the RAKEL and BR were employed to construct classifiers, and the base classifier was SVM or RF. With all possible parameter settings used above, several classifiers were built, each of which was assessed by tenfold cross-validation. The best performance for RAKEL and BR with one of the base classifiers is listed in Table 3, from which we can see that with such features, the RAKEL with SVM (polynomial kernel) provided the best performance. In detail, the integrated score was 0.6794, and three measurements (exact match, accuracy, and hamming loss) were 0.847, 0.848, and 0.054. Such performance was lower than that of the iMPT-FDNPL. Thus, features derived

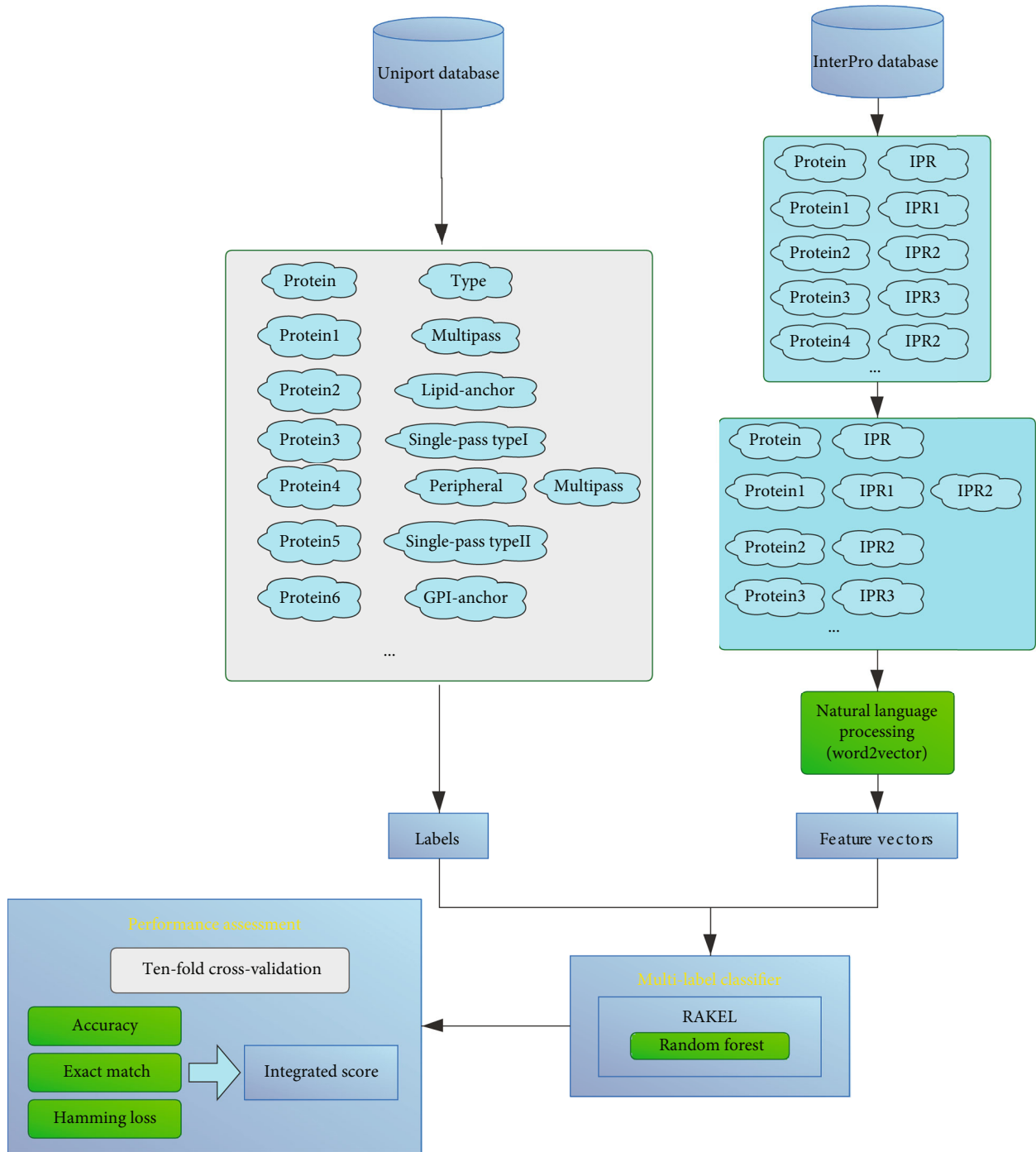


FIGURE 2: Entire procedures to construct and evaluate the multilabel classifier, iMPT-FDNPL. Membrane proteins and types are retrieved from the UniProt database. The types are termed as labels. Function domain information is obtained from the InterPro database. This information is processed by a natural language processing approach (word2vector), and the outcomes are used to encode proteins. Labels and vectors are fed into RAKEL with random forest as the base classifier to construct the multilabel classifier. This classifier is evaluated by tenfold cross-validation.

from functional domains via word2vector were more efficient than the features derived from functional domains via one-hot scheme for identifying membrane protein types.

Gene ontology (GO) [49] and KEGG pathway [50] information was also widely used to investigate protein- or gene-related problems. With the similar procedures that

were done for functional domains, GO terms and pathways were termed as words, whereas proteins, annotated by GO terms and pathways, were considered as sentences. We can obtain feature vectors of GO terms and pathways via word2-vector. Then, a membrane protein was represented by an average vector of vectors of GO terms and pathways that

TABLE 2: Performance of different multilabel classifiers with features derived from functional domain information via a natural language processing approach.

Scheme (base classifier)	Exact match	Accuracy	Hamming loss	Integrated score
RAKEL (RF) (iMPT-FDNPL)	0.851	0.853	0.053	0.6874
RAKEL (SVM-polynomial kernel)	0.831	0.834	0.060	0.6515
RAKEL (SVM-RBF kernel)	0.846	0.848	0.054	0.6787
BR (RF)	0.781	0.782	0.054	0.5778
BR (SVM-polynomial kernel)	0.804	0.815	0.061	0.6152
BR (SVM-RBF kernel)	0.829	0.831	0.050	0.6544

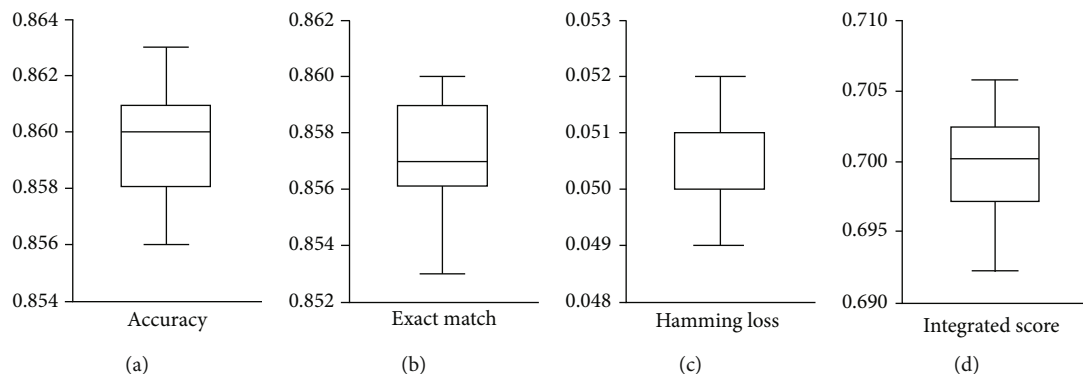


FIGURE 3: Boxplot to show the performance of iMPT-FDNPL using tenfold cross-validation for 20 times. (a) Accuracy; (b) exact match; (c) hamming loss; (d) integrated score. Each measurement varies in a same range.

TABLE 3: Performance of different multilabel classifiers with features derived from functional domain information via one-hot scheme.

Scheme (base classifier)	Exact match	Accuracy	Hamming loss	Integrated score
RAKEL (RF)	0.825	0.827	0.061	0.6406
RAKEL (SVM-polynomial kernel)	0.847	0.848	0.054	0.6794
RAKEL (SVM-RBF kernel)	0.846	0.847	0.054	0.6778
BR (RF)	0.785	0.788	0.049	0.5882
BR (SVM-polynomial kernel)	0.774	0.778	0.049	0.5726
BR (SVM-RBF kernel)	0.836	0.840	0.048	0.6685

TABLE 4: Performance of different multilabel classifiers with features derived from gene ontology and pathway information via a natural language processing approach.

Scheme (base classifier)	Exact match	Accuracy	Hamming loss	Integrated score
RAKEL (RF)	0.761	0.762	0.083	0.5324
RAKEL (SVM-polynomial kernel)	0.808	0.810	0.067	0.6106
RAKEL (SVM-RBF kernel)	0.808	0.810	0.068	0.6099
BR (RF)	0.584	0.584	0.087	0.3113
BR (SVM-polynomial kernel)	0.717	0.738	0.068	0.4931
BR (SVM-RBF kernel)	0.747	0.755	0.063	0.5284

were annotated on such protein. Likewise, several dimensions from 50 to 500 with interval 50 were generated. RAKEL or BR with SVM or RF as the base classifier was employed. Several classifiers were constructed with all possible parameter settings. All classifiers were evaluated by tenfold cross-validation. Similarly, the best performance

using RAKEL or BR with one base classifier is listed in Table 4. Evidently, in this case, RAKEL with SVM (polynomial kernel) generated the highest performance with integrated score of 0.6106. The exact match was 0.808, accuracy was 0.810, and hamming loss was 0.067. The exact match, accuracy, and integrated score were all lower than

TABLE 5: Performance of different multilabel classifiers with features derived from protein networks via a network embedding algorithm.

Scheme (base classifier)	Exact match	Accuracy	Hamming loss	Integrated score
RAKEL (RF)	0.758	0.759	0.085	0.5264
RAKEL (SVM-polynomial kernel)	0.805	0.807	0.068	0.6054
RAKEL (SVM-RBF kernel)	0.801	0.803	0.070	0.5981
BR (RF)	0.584	0.584	0.088	0.3110
BR (SVM-polynomial kernel)	0.712	0.730	0.068	0.4844
BR (SVM-RBF kernel)	0.746	0.756	0.063	0.5284

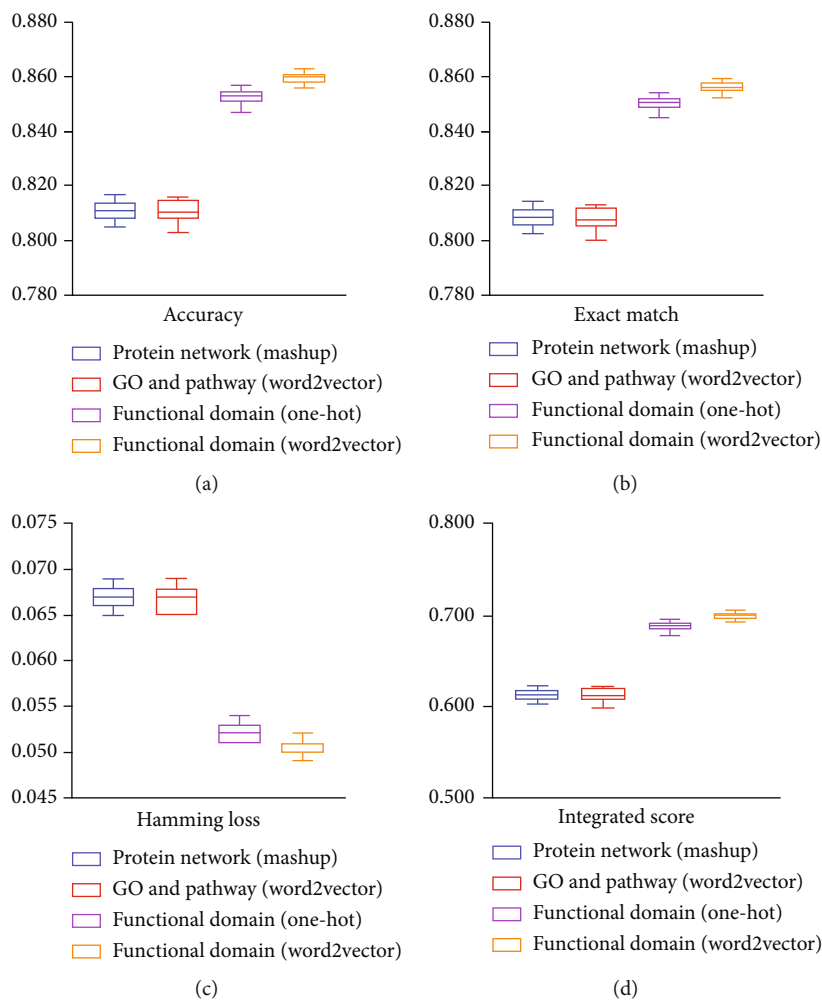


FIGURE 4: Boxplot to show the performance of classifiers with different feature types using tenfold cross-validation for 20 times. (a) Accuracy; (b) exact match; (c) hamming loss; (d) integrated score. Features derived from functional domain via word2vector are most efficient to identify membrane protein types.

those of iMPT-FDNPL, and the hamming loss was larger than that of iMPT-FDNPL. These results indicated that features derived from functional domains via word2vector were more powerful to identify membrane protein types than those derived from GO and pathways via the same natural language processing approach. It was also implied that functional domain information was more related to membrane protein types than GO and pathway information.

Network embedding algorithm is a type of recently proposed computational methods, which can abstract asso-

ciations of nodes in one or more networks and extract a feature vector for each node. It has also been applied to process some protein-related problems [25, 26, 34, 51–55]. Here, we used such method to extract protein features. To this end, eight protein networks were first built according to protein-protein interaction information reported in STRING (<https://www.string-db.org/>, version 10.0) [56]. The network embedding algorithm, Mashup [53], was applied on these networks to access the feature vectors of proteins. The dimensions included integers from 50 to 500

with interval 50. Obtained feature vectors of membrane proteins were fed into RAKEL or BR with SVM or RF as the base classifier to build the classifiers. All possible parameter settings used above were tried, and all constructed classifiers were assessed by tenfold cross-validation. Table 5 lists the best performance of RAKEL or BR classifiers with different base classifiers. Interestingly, the RAKEL with SVM (polynomial kernel) also provided the best performance. The integrated score of such classifier was 0.6054. Other three measurements were 0.805, 0.807, and 0.068, respectively. However, compared with the performance of iMPT-FDNPL (see Table 2), such performance was still lower. These results also suggested the effectiveness of features derived from functional domain via word2vector for prediction of membrane protein types.

With above arguments, we can conclude that features derived from functional domain via word2vector are quite effective to identify membrane protein types because classifiers based such features were more powerful than those based on other three types of features, which were derived from functional domain via one-hot scheme, from GO and pathway via word2vector, and from protein network via Mashup, respectively. To further confirm the superiority of features derived from functional domain via word2vector, the best classifiers using above three types of features were further evaluated by tenfold cross-validation for 20 times. Obtained values of exact match, accuracy, hamming loss, and integrated score are shown in Figure 4. For easy comparisons, those of the classifier (iMPT-FDNPL) using features derived from functional domain via word2vector are also shown in this figure. It is easy to observe that iMPT-FDNPL always generated highest exact match, accuracy, and integrated score and lowest hamming loss. All these further confirmed the superiority of the used features, which was the main reason why iMPT-FDNPL can provide such good performance.

4. Conclusions

This study sets up a multilabel classifier, iMPT-FDNPL, to identify membrane protein types. A novel feature extraction scheme was integrated in this classifier, which can extract efficient protein features by applying a natural language processing approach, word2vector, to functional domain information of proteins. The cross-validation results showed that such classifier was quite powerful and superior to classifiers using other types of protein features. Such results also indicated the superiority of features extracted by the proposed scheme. It is hopeful that such classifier can be a useful tool to identify membrane protein types, and the novel feature extraction scheme can be used to tackle other protein-related problems. All codes and data are available at <https://github.com/mufei111/iMPT-FDNPL>.

Data Availability

The original data used to support the findings of this study are available at the UniProt database.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (61772028), the key research and development plan of Zhejiang Province (2021C02039), and the Natural Science Foundation of Shanghai (17ZR1412500).

References

- [1] P. Yeagle, *The Membranes of Cells*, Academic Press, 2016.
- [2] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, 2001.
- [3] M. S. Almén, K. J. V. Nordström, R. Fredriksson, and H. B. Schiöth, "Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin," *BMC Biology*, vol. 7, no. 1, p. 50, 2009.
- [4] Y. D. Cai, P. W. Ricardo, C. H. Jen, and K. C. Chou, "Application of SVM to predict membrane protein types," *Journal of Theoretical Biology*, vol. 226, no. 4, pp. 373–376, 2004.
- [5] M. Wang, J. Yang, G. P. Liu, Z. J. Xu, and K. C. Chou, "Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition," *Protein Engineering, Design & Selection*, vol. 17, no. 6, pp. 509–516, 2004.
- [6] S. Q. Wang, J. Yang, and K. C. Chou, "Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 242, no. 4, pp. 941–946, 2006.
- [7] A. Mahdavi and S. Jahandideh, "Application of density similarities to predict membrane protein types based on pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 276, no. 1, pp. 132–137, 2011.
- [8] M. Hayat and A. Khan, "Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 271, no. 1, pp. 10–17, 2011.
- [9] E. S. Sankari and D. Manimegalai, "Predicting membrane protein types by incorporating a novel feature set into Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 455, pp. 319–328, 2018.
- [10] Y. D. Cai, G. P. Zhou, and K. C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," *Biophysical Journal*, vol. 84, no. 5, pp. 3257–3263, 2003.
- [11] P. Jia, Z. Qian, K. Feng, W. Lu, Y. Li, and Y. Cai, "Prediction of membrane protein types in a hybrid space," *Journal of Proteome Research*, vol. 7, no. 3, pp. 1131–1137, 2008.
- [12] Y. D. Cai and K. C. Chou, "Predicting membrane protein type by functional domain composition and pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 238, no. 2, pp. 395–400, 2006.

- [13] G. Tsoumakas and I. Vlahavas, *Random k-Labelsets: An Ensemble Method for Multilabel Classification*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.
- [14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [15] G. Huang, Y. Zhang, L. Chen, N. Zhang, T. Huang, and Y. D. Cai, "Prediction of multi-type membrane proteins in human by an integrated approach," *PLoS One*, vol. 9, no. 3, article e93553, 2014.
- [16] The UniProt Consortium, "The universal protein resource (UniProt) in 2010," *Nucleic Acids Research*, vol. 38, suppl_1, pp. D142–D148, 2010.
- [17] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.
- [18] K. C. Chou and Y. D. Cai, "Prediction of membrane protein types by incorporating amphipathic effects," *Journal of Chemical Information and Modeling*, vol. 45, no. 2, pp. 407–413, 2005.
- [19] R. Apweiler, T. K. Attwood, A. Bairoch et al., "The InterPro database, an integrated documentation resource for protein families, domains and functional sites," *Nucleic Acids Research*, vol. 29, no. 1, pp. 37–40, 2001.
- [20] J.-P. Zhou, L. Chen, and Z.-H. Guo, "iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs," *Bioinformatics*, vol. 36, no. 5, pp. 1391–1396, 2020.
- [21] J.-P. Zhou, L. Chen, T. Wang, and M. Liu, "iATC-FRAKEL: a simple multi-label web-server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only," *Bioinformatics*, vol. 36, no. 11, pp. 3568–3569, 2020.
- [22] H. Weng, Z. Liu, A. Maxwell et al., "Multi-label symptom analysis and modeling of TCM diagnosis of hypertension," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain, 2018.
- [23] A. Maxwell, R. Li, B. Yang et al., "Deep learning architectures for multi-label classification of intelligent health risk prediction," *BMC Bioinformatics*, vol. 18, Suppl 14, p. 523, 2017.
- [24] J. S. Saleema, B. Sairam, S. D. Naveen, K. Yuvaraj, and L. M. Patnaik, "Prominent label identification and multi-label classification for cancer prognosis prediction," in *TENCON 2012 IEEE Region 10 Conference*, Cebu, Philippines, 2012.
- [25] Y. Zhu, B. Hu, L. Chen, and Q. Dai, "iMPTCE-Hnetwork: a multilabel classifier for identifying metabolic pathway types of chemicals and enzymes with a heterogeneous network," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 6683051, 12 pages, 2021.
- [26] L. Chen, Z. Li, T. Zeng et al., "Predicting gene phenotype by multi-label multi-class model based on essential functional features," *Molecular Genetics and Genomics*, vol. 296, no. 4, pp. 905–918, 2021.
- [27] J. Che, L. Chen, Z. H. Guo, S. Wang, and Aorige, "Drug target group prediction with multiple drug networks," *Combinatorial Chemistry & High Throughput Screening*, vol. 23, no. 4, pp. 274–284, 2020.
- [28] J. Read, P. Reutemann, B. Pfahringer, and G. Holmes, "MEKA: a multi-label/multi-target extension to WEKA," *Journal of Machine Learning Research*, vol. 17, 2016.
- [29] Y. Yang and L. Chen, "Identification of drug–disease associations by using multiple drug and disease networks," *Current Bioinformatics*, vol. 16, 2021.
- [30] Y. Jia, R. Zhao, and L. Chen, "Similarity-based machine learning model for predicting the metabolic pathways of compounds," *IEEE Access*, vol. 8, pp. 130687–130696, 2020.
- [31] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Mathematical Biosciences*, vol. 306, pp. 136–144, 2018.
- [32] Y. H. Zhang, H. Li, T. Zeng et al., "Identifying transcriptomic signatures and rules for SARS-CoV-2 infection," *Frontiers in Cell and Development Biology*, vol. 8, p. 627302, 2021.
- [33] Y.-H. Zhang, Z. Li, T. Zeng et al., "Detecting the multiomics signatures of factor-specific inflammatory effects on airway smooth muscles," *Frontiers in Genetics*, vol. 11, p. 599970, 2021.
- [34] X. Pan, H. Li, T. Zeng et al., "Identification of protein subcellular localization with network and functional embeddings," *Frontiers in Genetics*, vol. 11, p. 626500, 2021.
- [35] K. K. Kandaswamy, K. C. Chou, T. Martinetz et al., "AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties," *Journal of Theoretical Biology*, vol. 270, no. 1, pp. 56–62, 2011.
- [36] Y. B. Marques, A. de Paiva Oliveira, A. T. Ribeiro Vasconcelos, and F. R. Cerqueira, "Miracle: machine learning with SMOTE and random forest for improving selectivity in pre-miRNA ab initio prediction," *BMC Bioinformatics*, vol. 17, no. S18, p. 474, 2016.
- [37] T.-T. Nguyen, J. Z. Huang, Q. Wu, T. T. Nguyen, and M. J. Li, "Genome-wide association data classification and SNPs selection using two-stage quality-based random forests," *BMC Genomics*, vol. 16, Supplement 2, p. S5, 2015.
- [38] F. Ahmad, A. Farooq, M. U. G. Khan, M. Z. Shabbir, M. Rabbani, and I. Hussain, "Identification of most relevant features for classification of *Francisella tularensis* using machine learning," *Current Bioinformatics*, vol. 15, no. 10, pp. 1197–1212, 2021.
- [39] E. Kwon, M. Cho, H. Kim, and H. S. Son, "A study on host tropism determinants of influenza virus using machine learning," *Current Bioinformatics*, vol. 15, no. 2, pp. 121–134, 2020.
- [40] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *In international joint conference on artificial intelligence*, Lawrence Erlbaum Associates Ltd, 1995.
- [41] Y.-H. Zhang, T. Zeng, L. Chen, T. Huang, and Y.-D. Cai, "Determining protein–protein functional associations by functional rules based on gene ontology and KEGG pathway," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1869, no. 6, p. 140621, 2021.
- [42] H. Liu, B. Hu, L. Chen, and L. Lu, "Identifying protein subcellular location with embedding features learned from networks," *Current Proteomics*, vol. 17, 2021.
- [43] X. G. Chen, W. W. Shi, and L. Deng, "Prediction of disease comorbidity using HeteSim scores based on multiple heterogeneous networks," *Current Gene Therapy*, vol. 19, no. 4, pp. 232–241, 2019.
- [44] L. Chen, S. Wang, Y. H. Zhang et al., "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.
- [45] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [46] J. Platt, *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*, B. Schölkopf, C. Burges, and A. Smola, Eds., *Advances in kernel methods: Support vector learning*, MIT press: Cambridge, MA, 1998.

- [47] J. Platt, *Sequential Minimal Optimizaton: A Fast Algorithm for Training Support Vector Machines*, Technical Report MSR-TR-98-14, 1998.
- [48] G. Tsoumakas and I. Katakis, “Multi-label classification: an overview,” *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.
- [49] M. Ashburner, C. A. Ball, J. A. Blake et al., “Gene ontology: tool for the unification of biology,” *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [50] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.
- [51] X. Zhang, L. Chen, Z. H. Guo, and H. Liang, “Identification of human membrane protein types by incorporating network embedding methods,” *IEEE Access*, vol. 7, pp. 140794–140805, 2019.
- [52] X. Pan, L. Chen, Liu, Z. Niu, T. Huang, and Y. D. Cai, “Identifying protein subcellular locations with embeddings-based node2loc,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1, 2021.
- [53] H. Cho, B. Berger, and J. Peng, “Compact integration of multi-network topology for functional analysis of genes,” *Cell Systems*, vol. 3, no. 6, pp. 540–548.e5, 2016.
- [54] X. Zhao, L. Chen, Z. H. Guo, and T. Liu, “Predicting drug side effects with compact integration of heterogeneous networks,” *Current Bioinformatics*, vol. 14, no. 8, pp. 709–720, 2019.
- [55] Y. Luo, X. Zhao, J. Zhou et al., “A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information,” *Nature Communications*, vol. 8, no. 1, p. 573, 2017.
- [56] C. von Mering, “STRING: known and predicted protein-protein associations, integrated and transferred across organisms,” *Nucleic Acids Research*, vol. 33, no. Database issue, pp. D433–D437, 2004.

Retraction

Retracted: Simulation and Prediction of Fungal Community Evolution Based on RBF Neural Network

Computational and Mathematical Methods in Medicine

Received 20 June 2023; Accepted 20 June 2023; Published 21 June 2023

Copyright © 2023 Computational and Mathematical Methods in Medicine. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] X. Cai, Y. Bao, M. Hu, J. Liu, and J. Zhu, "Simulation and Prediction of Fungal Community Evolution Based on RBF Neural Network," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 7918192, 13 pages, 2021.

Research Article

Simulation and Prediction of Fungal Community Evolution Based on RBF Neural Network

Xiao-Wei Cai ¹, Ya-Qian Bao ¹, Ming-Feng Hu ¹, Jia-Bao Liu ², and Jia-Ming Zhu ¹

¹School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233030, China

²School of Mathematics and Physics, Anhui Jianzhu University, Hefei 230601, China

Correspondence should be addressed to Jia-Ming Zhu; zhujm1973@163.com

Received 17 August 2021; Accepted 20 September 2021; Published 8 October 2021

Academic Editor: Hui Ding

Copyright © 2021 Xiao-Wei Cai et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Simulation and prediction of the scale change of fungal community. First, using the experimental data of a variety of fungal decomposition activities, a mathematical model of the decomposition rate and the relationship between the bacterial species was established, thereby revealing the internal mechanism of fungal decomposition activity in a complex environment. Second, based on the linear regression method and the principle of biodiversity, a model of fungal decomposition rate was constructed, and it was concluded that the interaction between mycelial elongation and moisture resistance could increase the fungal decomposition rate. Third, the differential equations are used to quantify the competitive relationship between different bacterial species, divide the boundaries of superior and inferior species, and simulate the long-term and short-term evolution trends of the community under the same initial environment. And an empirical analysis is made by taking the sudden change of the atmosphere affecting the evolution of the colony as an example. Finally, starting from summer, combining soil temperature, humidity, and fungal species data in five different environments such as arid and semiarid, a three-dimensional model and RBF neural network are introduced to predict community evolution. The study concluded that under given conditions, different strains are in short-term competition, and in the long-term, mutually beneficial symbiosis. Biodiversity is important for the biological regulation of nature.

1. Introduction

The carbon cycle is an important part of life on earth, where the decomposition of compounds allows carbon to be renewed and used in other forms [1]. The key component of this process is the decomposition of plant materials and wood fibers. Related studies have found that the decomposition rate of fungi, a key factor in the decomposition of plant materials and wood fibers, is influenced by temperature, humidity, time, growth rate, mycelial density, and moisture tolerance [2, 3]. And slow-growing fungi are more likely to survive and grow in the environment of humidity and temperature changes, while faster-growing fungi are less resistant to the same environmental changes [4]. At the same time, the decomposition rate of fungi determines the biomass and nutrient content of the forest surface and significantly affects the physical and chemical properties of the soil. By exploring the mycelial elongation rate of fungi and the moisture resis-

tance of fungus, it is possible to reveal the important role of fungi's decomposition mechanism of plant material and wood fiber, as well as the mutual adaptation of coupling modes between different species combinations in biodiversity [5].

He selected the suitable tree species of karst natural community habitat, Constructus and Yungui gooseberry as the research object to investigate how AM fungi regulate soil litter to achieve nutrient release and change soil properties under competitive conditions [6]. Zhang investigated the correlation between the culture products of corn stover in four treatments, basic properties, material content and biological enzyme activities, and the dynamic change pattern of the humus-like composition of culture products [7]. To investigate the effect of endophytic fungi on the decomposition of apoplasts, Chen investigated the effect of endophytic fungi on the decomposition of apoplasts by using different sampling methods and selecting endophytic fungi with

different dominance to participate in different community construction [8]. By combining indoor culture experiments and field experiments, Tan investigated the changes of soil microbial biomass, enzyme activities related to soil organic carbon decomposition, and the effects of ectomycorrhizal fungi community structure and diversity, and then analyzed the role of ectomycorrhizal fungi in forest soil organic carbon decomposition [9].

Under natural conditions, microorganisms do not degrade apoplankton independently, but the interactions between decomposer groups and their components cannot be ignored, but due to the complexity of the interactions between microorganisms, soil and apoplankton quality, this aspect has not been well studied. Most of the studies on fungal decomposition and community evolution are based on the biological level and are not well integrated with mathematical models. The data in this paper are from real experimental data, which are reliable and can be closer to the complex natural conditions.

In this paper, first, a mathematical model of the relationship between the decomposition rate of fungal species and fungal species was established based on the experimental data of decomposition activities of various fungi. Second, based on the principle of biodiversity, a linear regression method is used to construct a model of fungal decomposition. Then, using differential equations, a dynamic model of fungal competitiveness was established and empirically analyzed. Finally, the three-dimensional model and RBF neuron network were combined to predict the evolution of the community.

2. Basic Assumptions

The research question comes from Question A of the 2021 American College Students Mathematical Modeling Competition, and the data comes from National Center for Biotechnology Information. To explore the above issues, we make the following assumptions: (i) it is assumed that the substances produced by fungal decomposition have no significant impact on itself and the surrounding environment. (ii) It is assumed that only fungi participate in the decomposition process, and other microorganisms do not participate in the decomposition of the compound. (iii) The decomposition of plant material and wood fiber is independent of each other. (iv) It is assumed that the main factors affecting the shape of the fungus that affect the decomposition rate are the fungal hyphae elongation and moisture resistance, and the influence of other factors is negligible. (v) In the competition of fungi, the influence of fungal aerobic respiration, anaerobic respiration, or anaerobic respiration on the decomposition rate is not considered.

3. Construction of Fungal Decomposition Model Based on Multiple Linear Regression Method

3.1. Research Ideas. To describe the decomposition of organic matter by a variety of fungi, this paper selects seven common fungi as the research object, takes Chinese fir as an

example to be decomposed, and constructs a fungal decomposition model [10–12]. By consulting the relevant information, we found that temperature, humidity, colony abundance, time, soil sulfur and phosphorus content, mycelial elongation, and moisture resistance are the main factors affecting the decomposition rate of fungi, so we set them as independent variables and use K , H , N , T , P , G , and R to represent, respectively. Besides, we set the mass loss rate of each patch of plant per unit time as the dependent variable, denoted by V . The method of multiple linear regression is used to study the decomposition activity of fungi.

3.2. Analysis Steps

3.2.1. Construct a Multiple Linear Regression Model. In this paper, 7 strains of AM fungus, *Cladosporium*, *Trichoderma*, *Aspergillus flavus*, *Alternaria* spp, *Penicillium*, and *Chaetomium vulgare* were mixed [13], and they were simultaneously inoculated on a petri dish with Chinese fir as the decomposed substance. The experimental data was collected in one week and lasted for 12 weeks. Under the conditions of controlling the temperature and humidity of each group of experiments, the fungus' decomposition activity on the substrate was studied [14].

Standardize the data obtained in the experiment to eliminate the influence of dimensions on the model and use the data to construct a multiple linear regression model.

$$V = a + bK + cH + dT + eS + fG + gR. \quad (1)$$

Since this article is seven sets of parallel experiments under the control of temperature K , humidity H , and fungal richness N , the multiple linear regression models of each group can be obtained:

$$\begin{cases} V_1 = -1.062 - 0.056K + 1.093H - 0.212T + 0.218P - 0.0416G - 1.041R, N = 1, \\ V_2 = -1.173 - 1.041K - 1.043H - 0.031T + 0.237P + 2.153G - 1.092R, N = 2, \\ V_3 = -0.159 + 0.483K - 1.224H - 0.435T - 0.083P - 1.228G - 1.059R, N = 3, \\ V_4 = 0.338 + 2.087K - 1.006H - 0.052T - 0.163P + 2.221G + 0.162R, N = 4, \\ V_5 = 1.387 + 1.607K - 1.020H + 0.690T + 0.286P + 1.156G - 1.241R, N = 5, \\ V_6 = 1.710 + 1.609K - 0.183H - 0.039T + 0.189P + 0.376G - 0.197R, N = 6, \\ V_7 = 0.450 - 0.231K + 0.144H - 0.058T + 0.1429P - 0.107G - 0.317R, N = 7. \end{cases} \quad (2)$$

To obtain a fungal decomposition rate model closer to the natural environment, this paper uses the entropy method to weight the above equations to eliminate the influence of colony richness N on the difference in fungal decomposition rates between the experimental groups to meet the needs of biodiversity.

3.2.2. Entropy Method. By calculating the information entropy of each indicator in the indicator system and determining the weight of the indicator according to the relative change degree of the indicator and the contribution rate to the overall system, it is a method of combining static weighting and dynamic weighting. If the information entropy is smaller, the disorder degree of the information is lower,

the utility value of the information is larger, and the weight of the index is larger. Specific steps are as follows:

$$X = (x_{ij})_{mn}, \text{ namely, } X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}. \quad (3)$$

$$P(x_{ij}) = \frac{x'_{ij}}{\sum_{i=1}^m x'_{ij}}. \quad (4)$$

$$h_j = -k \sum_{i=1}^n P(x_{ij}) \ln P(x_{ij}), j = 1, 2, \dots, n. \quad (5)$$

Step 1. Define the initial evaluation matrix.

Step 2. Calculate the ratio of the index value $P(x_{ij})$.

Step 3. Calculate the entropy value of each index.

Among them, $P(x_{ij})$ represents the corresponding value, $K > 0$, generally take $K = 1/\ln k$, so $0 \leq h_j \leq 1$. The greater the degree of difference between the evaluation, indicators x_{ij} the greater the contribution rate of the indicator to the entire indicator system and the greater the weight assigned.

$$\omega_j = \frac{h_j}{\sum_{j=1}^n h_j}, j = 1, 2, \dots, n. \quad (6)$$

Step 4. Calculation of index weight.

The weight of the indicator reflects the degree of influence of the evaluation indicator on the overall performance.

To sum up, when $N = 1, 2, 3, 4, 5, 6, 7$ the weight assigned to each group should be (1.424, 1.729, 1.865, 1.870, 1.885, 1.918, 1.930).

That is, the relationship between the overall fungal decomposition rate and the fungal decomposition rate between groups is

$$1.424V_1 + 1.729V_2 + 1.865V_3 + 1.870V_4 + 1.885V_5 + 1.918V_6 + 1.930V_7. \quad (7)$$

Substituting the equations of the above experiment to obtain the decomposition rate model of fungi on fir in the presence of multiple fungi is

$$V = -1.703 + 2.081K - 1.334H - 0.046T + 0.356P + 2.128G - 1.093R. \quad (8)$$

3.3. Conclusion Analysis. It can be found from equation (12) that when the colony abundance $N = 7$, it has a greater impact on the overall fungal decomposition rate, and when

$N = 1$, it has a small impact on the overall fungal decomposition rate. To better adapt to the biodiversity of nature, expand the number of studies on colonies, and further explore the impact of biodiversity on fungal decomposition activities, this paper sets a dummy variable D_1 . The colony with $N > 2$ is defined as a multicolony group and $D_1 = 1$ is assigned; the colony with $N = 1, 2$ is defined as a single colony, and the value is $D_1 = 0$.

$$D_1 = \begin{cases} 0 & N = 1, 2, \\ 1 & N > 2. \end{cases} \quad (9)$$

The average value of the fungal decomposition rate in each group of experiments is selected to indicate the size of the fungal decomposition rate under different colony richness when the temperature, humidity, and other variables are unchanged. Introduce dummy variables in the form of addition, and establish a unary linear model of fungal decomposition rate and colony richness:

$$V = a + bN + D_1. \quad (10)$$

Substituting the standardized decomposition rate data into equation (10), the equation of decomposition rate and colony richness is obtained as:

$$V = 4.9671 + 0.3452N + 1.0387D_1. \quad (11)$$

The regression results show that the decomposition rate is positively correlated with the colony abundance, and the coefficient before the dummy variable D_1 is positive, indicating that the multicolony community is beneficial to the decomposition of fungi, that is, the decomposition rate will be significantly increased under the condition of multiple colonies.

From equation (13), it can be found that the fungal decomposition rate is related to temperature, humidity, decomposition time, soil antibiotic content, mycelial elongation, and moisture resistance.

In summary, in a variety of bacterial communities, the decomposition rate of fungi will be significantly increased. Therefore, under the conditions of coexistence of multiple fungi, higher temperature, higher soil phosphorus and sulfur content, and faster growth rate of fungi, the decomposition rate of fungi increases; in the presence of single fungi, higher humidity, longer decomposition time, and higher humidity resistance under the conditions, the decomposition rate of fungi slows down [15–18].

4. Construction of a Decomposition Rate Model Based on the Interaction between Fungi

4.1. Research Ideas. The rate of change of temperature and humidity has a certain relationship with the vitality of fungi [19]. According to Lustenhouwer et al., the relationship between mycelial elongation and wood decomposition rate is approximately positive and linear; under logarithmic transformation, the relationship between moisture tolerance

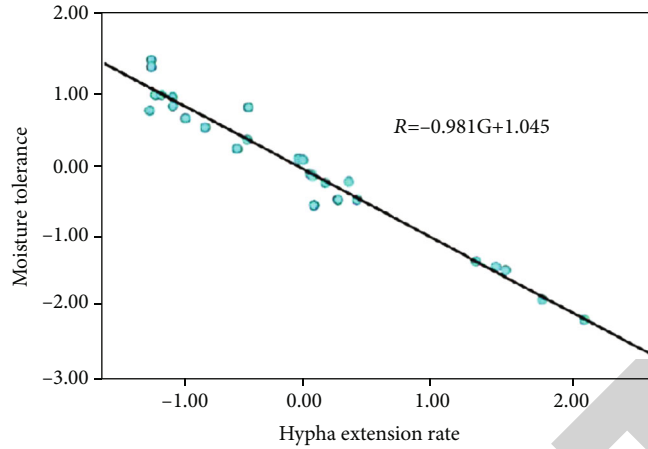


FIGURE 1: Mycelium elongation and moisture resistance.

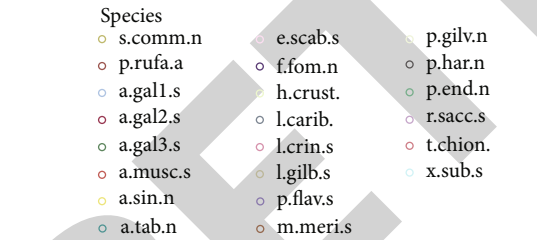
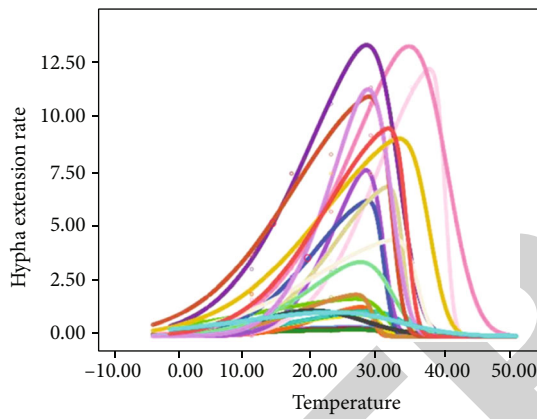


FIGURE 2: The relationship between temperature and mycelial elongation.

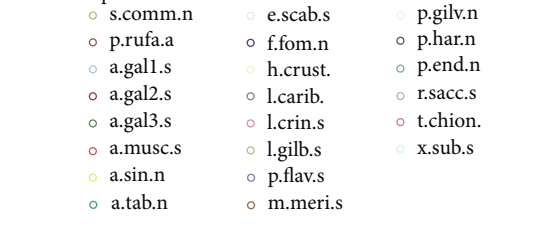
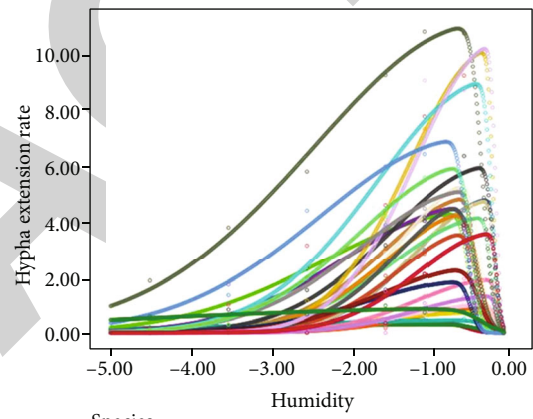


FIGURE 3: The relationship between humidity and mycelial elongation.

of different fungi and the final wood decomposition rate is also approximately positive and linear. It can be inferred that there may also be a linear relationship between the moisture resistance of different fungi and the elongation of their hyphae.

Therefore, this paper will use the characteristics of hyphae, elongation, and moisture resistance of different types of fungi to combine the types of fungi. First, we establish a univariate linear regression model to study the relationship between the moisture tolerance of different fungi and the hyphae elongation; second, the two characteristics of fungi are combined under the condition of keeping the fungal decomposition rate constant, that is, the hyphae extension rate represents moisture resistance; finally,

substituting the unary linear model into the above fungal decomposition rate model to obtain a modified model—a decomposition rate model based on the interaction between fungi.

4.2. *Analysis Steps.* Under the condition that the decomposition rate of fungi remains unchanged, the relevant data of mycelial elongation and moisture resistance are obtained through web crawling and standardized processing, and a unary linear regression model of mycelial elongation and moisture resistance is constructed:

$$R = a + bG. \tag{12}$$

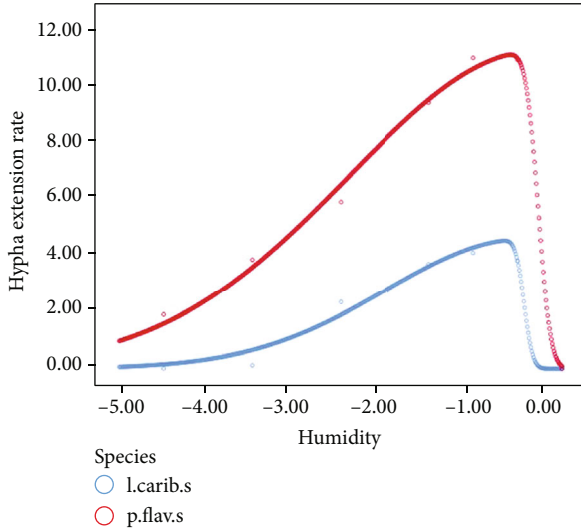


FIGURE 4: Comparison of different fungi under the same humidity.

Using SPSS25 software to fit the equations of mycelial elongation and moisture resistance, as shown in Figure 1.

$$R = -0.981G + 1.45. \tag{13}$$

Thanks for the suggestion, this sentence is revised to: Substitute the above results into the fungal decomposition rate model in section 3.3.2, and get an improved model based on the different interactions among hyphae, elongation, and moisture resistance:

$$V = -1.231 + 2.018K - 0.533H - 0.346T + 1.056P + 1.119G. \tag{14}$$

4.3. Conclusion Analysis. The change of environmental temperature and humidity is related to the vitality of fungi, that is, it will affect the growth rate of fungi. The decomposition rate of fungi is positively correlated with the growth rate and negatively correlated with the humidity resistance. Under the condition of controlling the decomposition rate unchanged, the fungus growth rate and humidity resistance are combined. It was found that there was a significant negative correlation between the growth rate and moisture tolerance of the fungus, and the interaction between the growth rate and moisture tolerance of the fungus had a positive effect on its decomposition rate.

5. Construction of Competitive Dynamic Model Based on Differential Equations

5.1. Research Ideas. Different fungi have different moisture resistance, so their mycelial elongation rate or growth rate is different in the same environment [20]. When the size of each colony is different, the limited survival resources will not get a reasonable and even distribution, and the competition relationship between the populations will occur, that is, the interaction [21–25]. The dynamic model is a model that

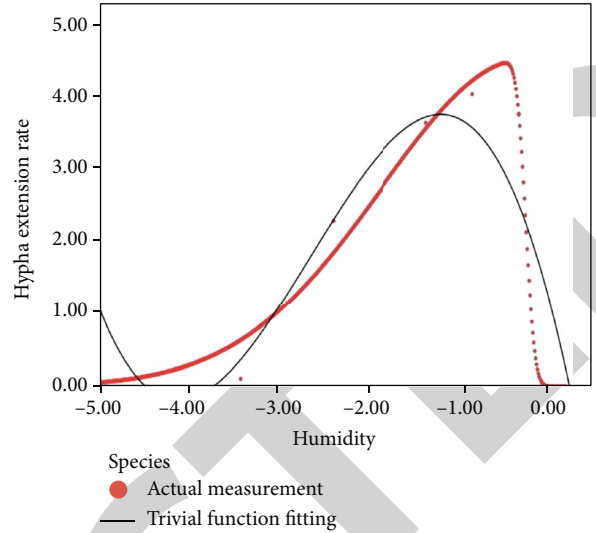


FIGURE 5: Changes in humidity of a fungus.

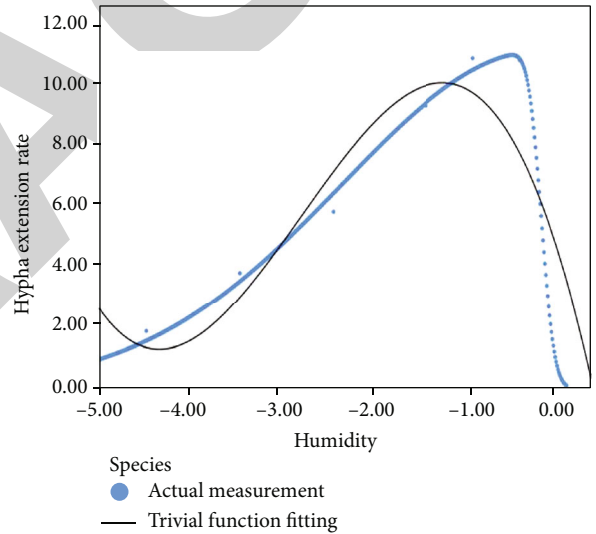


FIGURE 6: Effect of fungus on humidity.

describes the characteristics of the system related to time changes and the environment of the event. It can explore the overall behavior of the system, reduce the complexity of the system with the help of state diagrams or sequence diagrams, and can be completed while monitoring whether the conceptual system has defects. And show the internal operating mechanism of the system in detail [26].

Based on the above analysis, this paper firstly quantifies the internal interaction, namely, establishes a competitive dynamic model to simulate the long-term and short-term evolution trends of the community under the same initial environment under the premise of dividing the boundary between the advantages and disadvantages of the bacteria [27]. Second, consider the sudden external atmospheric changes. The overall impact is to explore the susceptibility of colony evolution to rapid environmental fluctuations.

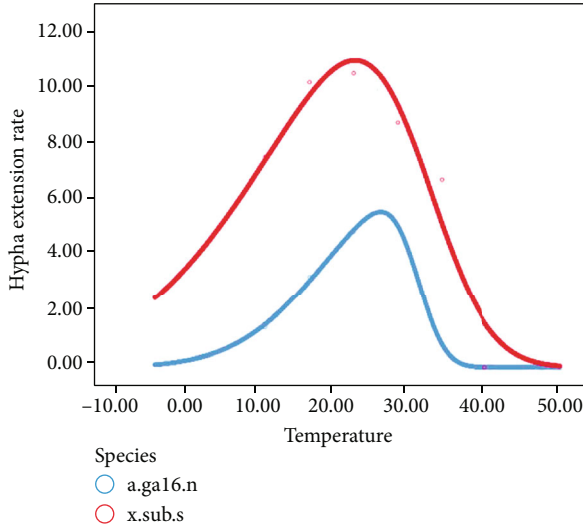


FIGURE 7: Comparison of different fungi at the same temperature.

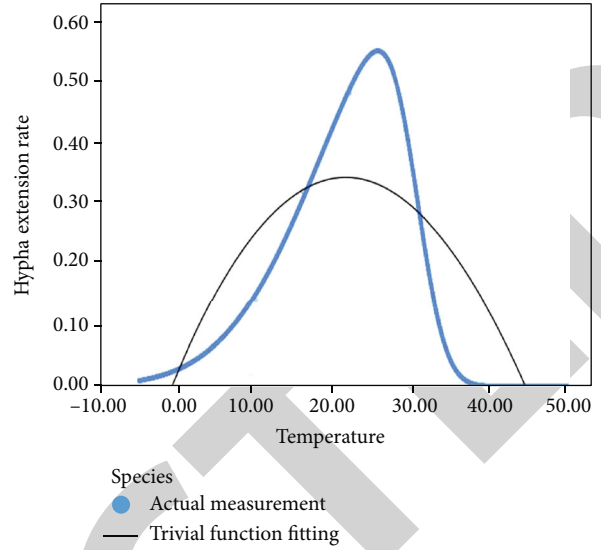


FIGURE 9: Temperature change of another fungus.

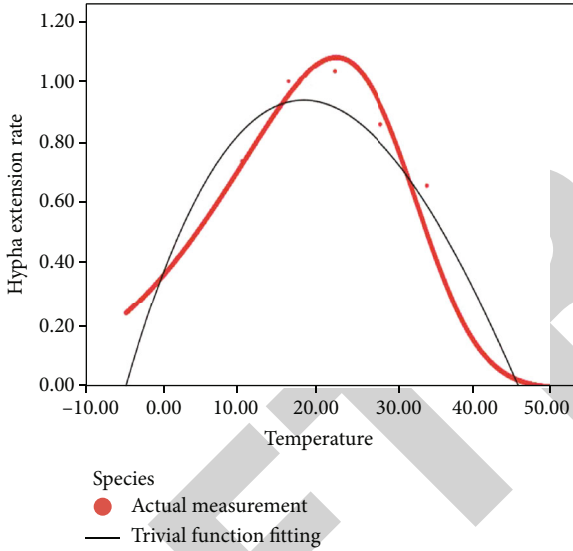


FIGURE 8: Temperature response of a fungus.

5.2. Analysis Steps

5.2.1. Comparative Analysis. According to the data of changes in the mycelial elongation of each fungus under different temperature and humidity conditions, a scatter plot was drawn with the mycelial elongation as the dependent variable, and temperature and humidity independent.

Variables are shown in Figures 2 and 3.

It can be seen from the above figure that the influence trends of temperature and humidity on different strains are roughly similar, but there are still some differences. To observe the degree of difference more intuitively, this paper selected two fungi from the aspects of temperature and humidity for comparison, to amplify the similarities and differences of the impact of environmental changes on different fungi, as shown in Figures 4–9.

It can be seen from the comparative analysis graph that the overall trend of the sensitivity of different fungi to

changes in the external environment is relatively consistent, and all of them can fit the cubic linear equation with a higher coefficient of determination. The impact of the environment on different strains is more significant, that is, under the same external environment, different fungi have different hyphae elongation rates, which in turn creates a competitive relationship between colonies.

5.2.2. Model Construction under General Conditions. The density of the hypha is equal to the ratio of the length of the hypha to the quality of the soil, which describes the density of fungal growth horizontally and can represent the growth scale of the colony. This article is represented by W .

According to the above temperature and humidity curve, the evaluation model is introduced to evaluate and score the competitiveness of different fungi, and the competitiveness ranking of each strain is obtained, as shown in Table 1.

According to the size of the competitive ranking, the bacteria are divided into two categories, A and B, based on the value equal to 0.5.

Type A fungi: the competitive ranking is less than 0.5, and it is greatly affected by temperature and humidity. It has greater hyphae elongation and hyphae density at the optimum temperature. This article defines it as an inferior strain.

Type B fungi: competitive ranking is greater than or equal to 0.5 and is less affected by temperature and humidity. It has low hyphae elongation and hypha density at the optimum temperature. This article defines it as a dominant species.

To explore the dynamic changes of the mycelial density D of these two types of bacteria under the long-term and short-term trends under this competitive situation, this paper uses $A(t)$ and $J(t)$ to indicate the density of hyphae of type A species and type B species, setting the initial temperature of the fungus to 22°C and 55% humidity under general conditions in week t .

TABLE 1: Ranking.

Fungus name	Competitive ranking	Mycelial elongation rate	Hypha density
Armillaria	0.281	0.398	0.420
Hyphodontia	0.630	3.590	0.080
Laetiporus	0.342	4.115	0.168
Lentinus	0.569	6.380	0.050
Mycoacia subconspersa	0.569	1.300	0.840
Merulius tremulosus	0.813	10.120	0.050
Phlebiopsis gigantea	0.634	6.105	0.480
Porodisculus pendulus	0.465	4.060	0.320
Phellinus robiniae	0.376	2.220	0.095
Phlebia acerina	0.973	8.510	0.270
Pycnoporus sanguineus	0.697	4.970	0.020
Schizophyllum commune	0.626	3.490	0.560
Tyromyces	0.805	3.880	0.060
Xylobolus subpileatus	0.493	0.770	1.740

(i) In the long-term process, from the time span of the 1st to 12th week, the following differential equations are obtained using experimental data:

$$\begin{cases} \frac{dA(t)}{dt} = -aJ(t), \\ \frac{dA(t)}{dt} = -bA(t), \\ A(0) = 0.2475, J(0) = 0.5570. \end{cases} \quad (15)$$

$$A(t) - A(0) = -a \sum_{n=1}^t J(n). \quad (16)$$

$$J(t) - J(0) = -b \sum_{n=1}^t A(n). \quad (17)$$

At week 12:

$$A(t) = 0.2063, J(t) = 0.7640. \quad (18)$$

$$\sum_{n=1}^t A(n) = 0.2269 \times 12 = 2.4958. \quad (19)$$

Substituting the above formula into equation (17), we can get

$$-2.4958b = 0.7640 - 0.5557, \text{ namely, } b = 0.0835. \quad (20)$$

Because $b < 0$, there is a coexistence relationship between A and B species in the long run.

(ii) In the short-term process, from the time span of the 1st to 5th week, the following differential equations are obtained using experimental data

$$\begin{cases} \frac{dA(t)}{dt} = -aJ(t), \\ \frac{dA(t)}{dt} = -bA(t), \\ A(0) = 0.2475, J(0) = 0.5570. \end{cases} \quad (21)$$

$$A(t) - A(0) = -a \sum_{n=1}^t J(n) = -0.2475. \quad (22)$$

$$J(t) - J(0) = -b \sum_{n=1}^t A(n) = -0.5570. \quad (23)$$

At week 5:

$$A(t) = 0.1633, J(t) = 0.4783. \quad (24)$$

$$\sum_{n=1}^t A(n) = 0.2054 \times 5 = 1.027. \quad (25)$$

Substituting the above formula into equation (23), we can get

$$-1.027b = 0.4783 - 0.5570, \text{ namely, } b = 0.0766. \quad (26)$$

Because $b > 0$, there is a competitive relationship between A and B strains in the short term.

In summary, under the conditions of a temperature of 22°C and a humidity of 55%, the A strain and the B strain have short-term competition and long-term coexistence.

5.2.3. Model Construction under Burst Conditions. In the absence of human interference, the changes in the external environment are usually relatively stable, but there will also be certain emergencies [28–30], in which atmospheric changes dominate. Therefore, the following will study the overall impact of atmospheric changes on the interaction between different species of fungi, select temperature and humidity as the main factors affecting the atmospheric level, and still use the mycelial density to represent the growth scale of the colony.

First of all, based on the example of B-type strains, which are dominant strains, a binary linear regression equation with hypha density as the dependent variable and temperature and humidity as independent variables is established using standardized related data.

It can be seen from Table 2 that the model determination coefficient R_2 is 0.087, indicating that the degree of fit is very small, and the P values of the coefficients are all greater than 0.05, indicating that the effect of independent variables in the model is not significant. Therefore, the binary linear

TABLE 2: Model fit test.

Variables	Coefficient	Standard deviation	T -statistic	Significance
C	0.372	0.441	0.843	0.413
K	-0.016	0.019	-0.847	0.411
H	0.212	0.201	1.056	0.309
Decidability factor			0.087	

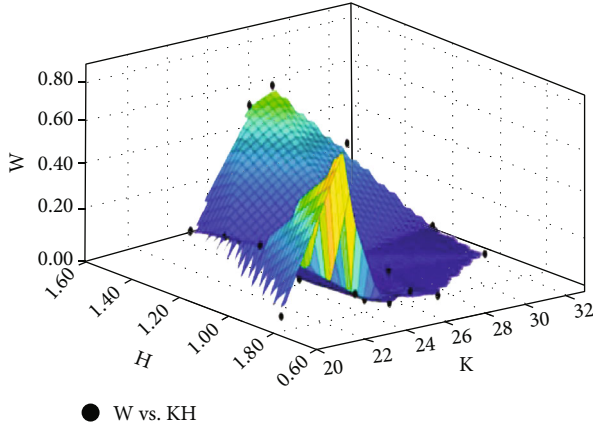


FIGURE 10: Fitting image of B-type fungus hypha density with temperature and humidity.

TABLE 3: Output results.

SSE	R -square	Adjusted R -square	RMSE
$2.046e-34$	1	NaN	NaN

regression equation is not suitable to explain this problem. This article will consider building a fitting model in a three-dimensional space.

Combine the temperature and humidity data to determine the plane corresponding to each value of the mycelium density, use Matlab software to build a three-dimensional fitting model, as shown in Figure 10.

From Table 3, it is clear that the SSE value of the dynamic model based on the dominant species is extremely small and negligible, and the correlation coefficient R -square value is 1, indicating that the model fits well and is able to explain and portray the overall effect of fungal atmospheric changes on the interaction between dominant fungi.

Similarly, the three-dimensional fitting model of atmospheric changes on the class A strain, i.e., the inferior strain, can be obtained, and the results are shown in Figure 11.

Thus, as shown in Table 4, the SSE numerical disadvantage dynamic model-based species can be ignored, the correlation coefficient R -square value of 1 indicates that the model fits well the effect, and can be interpreted to characterize changes in the atmosphere between disadvantage fungal fungi are mutually effect.

In summary, by constructing a dynamic model of the influence of temperature and humidity on mycelial density in a three-dimensional space, it is possible to reveal the nat-

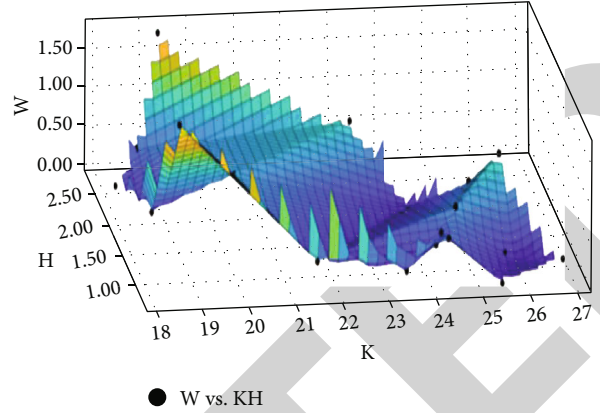


FIGURE 11: Three-dimensional fitting model of type A bacteria.

TABLE 4: Output results.

SSE	R -square	Adjusted R -square	RMSE
$7.754e-33$	1	NaN	NaN

ural law of interaction between atmospheric changes on different types of fungi and to explore the overall trend of atmospheric changes on the evolution of colonies.

5.3. Conclusion Analysis. Changes in the atmospheric environment mainly lead to changes in external temperature and humidity, which in turn affect the initial mycelial density D of the superior and inferior strains. Under the disturbance of rapid environmental fluctuations, it is possible that the original dominant strains will gradually weaken or even become inferior strains [31]. The conclusions on the sensitivity of colony evolution to rapid environmental fluctuations are specifically analyzed in this paper in conjunction with the models in Sections 5.2.2 and 5.2.3.

When the initial ambient temperature is 22.5°C and the humidity is 55%, the verification of the competitive dynamic model is constructed in Section 5.2.2. It is known that the relationship between A and B species is a state of short-term competition and long-term mutually beneficial symbiosis, and in the end, B species becomes the dominant species.

When the external environment changes, i.e., the temperature and humidity of the living environment of A and B bacteria change, according to the three-dimensional model constructed in Section 5.2.3, their respective mycelial density D will also change in the initial situation.

When the temperature changed to 16.6°C and the humidity changed to 32%, the initial mycelial densities corresponding to the A and B strains were 0.38 and 0.27. We applied the competitive dynamics model to simulate the evolution of the community and obtained that the mycelial densities corresponding to the moment of equilibrium point for A and B fungi were 0.47 and 0.26, respectively. In this case, the class A fungi became the dominant species, in contrast to the original environmental conditions in which they were the weaker species.

Based on the above conclusions, it can be concluded that changes in the environment have a greater impact on the

TABLE 5: Relevant conditions in five different environments.

Temperature band	x_1	x_2	x_3	x_4	x_5	x_6
Drought	34.5	10%	1848	8	31.4	15%
Semiarid	29.4	25%	5038	12	28.3	28%
Temperate	22.5	55%	10274	24	20.7	51%
Tree coniferous forest	8.6	28%	17893	24	10.2	29%
Tropical rainforest	31.5	68%	23137	48	30.2	65%

interrelationship between strains, and one manifestation of which is that they change the status of dominant and inferior strains. But in the overall situation, their interrelationship is all about short-term competition and long-term mutually beneficial symbiosis [32].

6. Prediction of Relative Advantages and Disadvantages among Species

6.1. Research Ideas. Due to the rich biodiversity in nature, the advantages and disadvantages of each species or combination of species are always relative [33–35]. From the abovementioned community evolution, with the change of time, the dominant species may become the inferior species, and predicting the relative advantages and disadvantages among species can determine the dominant species, which is conducive to exploring the inner mechanism of nature’s superiority and inferiority.

6.2. Prediction of Relative Strengths and Weaknesses. By reading various references [36–41], soil temperature, humidity, fungal species, and initial temperature and humidity of five different environmental types with a representative starting point in summer were collected in this paper [42], as shown in Table 5. Where x_1 , x_2 , x_3 , x_4 , x_5 , and x_6 denote temperature (°C), humidity, fungal species, initial temperature and humidity duration (weeks), equilibrium temperature (°C), and equilibrium humidity, respectively.

From the three-dimensional model of mycelial density and temperature and humidity in Section 5.2.3, it is known that if any combination of temperature and humidity is given, class A and class B strains will have corresponding initial mycelial densities, respectively, and our summary data is shown in Table 6. Where x_7 , x_8 , x_9 , and x_{10} denote initial mycelial density of class A fungi, initial mycelial density of class B fungi, mycelial density of class A fungi, in equilibrium and mycelial density of class B fungi in equilibrium, respectively.

From the data in the table, it can be seen that when the temperature is suitable and the humidity is lower, the mycelial density of the A strain is greater than that of the B strain, and the growth rate is faster. At this time, it is the dominant strain. When the temperature is suitable and the humidity is higher, the mycelial density of the B strain is higher than that of the A strain, and the growth rate is faster. At this time, it is the dominant strain.

On the whole, the B strains are less affected by the environment, the mycelial elongation rate is more stable, and the relative advantage is higher. However, due to the continuous changes of the environment over time, the B strain may

TABLE 6: Summary data of type A hyphae and type B hyphae.

Temperature band	x_7	x_8	x_9	x_{10}
Tree coniferous forest	0.017	0.009	0.023	0.016
Tropical rainforest	0.050	0.030	0.110	0.090
Tree coniferous forest	0.450	0.650	0.430	0.710
Tropical rainforest	0.210	0.270	0.280	0.220
Tree coniferous forest	0.760	0.940	0.740	0.980

evolve into a weak strain under special environmental conditions. On the contrary, the A inferior strain may also become a dominant strain.

If there are only a few species of bacteria in the soil, the growth and reproduction of fungi will be greatly affected by environmental changes, and the abundance of colonies will also be affected to a certain extent. If many strains of fungi are present in the soil, the environment, fungi, and fungus are predicted to establish a relationship between their strengths and weaknesses in the colony. In other words, when the environment changes, each species of fungus can complement each other in dynamics and ensure the biodiversity in the soil. After a period of time, the overall size of the colonies of various fungi and the abundance of the colony in the soil remained relatively stable, that is, in the evolution of the community, the various bacterial species compete in the short term, while the long-term different bacterial species present mutually beneficial symbiosis. This situation reflects the diversity and richness of biodiversity.

7. Biodiversity Prediction Based on RBF Neuron Networks

7.1. Research Ideas. It can be seen from the above that the rapid fluctuation of the external environment will affect the vitality of fungi. At the same time, the interaction between different bacterial species will also affect the colony size of each bacterial species.

7.2. Analysis Steps. From the above, it can be seen that the status of superior and inferior strains may be interchanged when the environment is changed, and at this time, the abundance of colonies will be affected to some extent. Diverse fungal communities are more “disturbance resistant” and “resilient” when considering the possibility of varying degrees of variability in the local environment, i.e., differences in initial conditions. That is, if the diversity of the fungal community is lost, the ecosystem can still be restored under certain natural laws.

Construction of neuronal network models. The multiple linear regression model has strong correlation between variables, i.e., high multicollinearity, so the accuracy of the model is still lacking. Considering this effect, we use RBF neural network algorithm to implement the prediction model with the help of Matlab software to achieve better prediction results [43].

RBF neural network has strong approximation ability, classification ability, and learning speed. Its working principle is to regard the network as an approximation to an

TABLE 7: Sample data on the variation of decomposition rate with each factor.

Growth rate	Mycelial density	Humidity resistance	Sulfur, phosphorus content	Temperature	Humidity	Decomposition rate
0.25	0.10	3.46	44.1	18.75	1.955	0.31
0.35	1.02	2.55	36.5	25.85	1.510	0.64
0.21	0.16	4.18	58.6	18.20	2.325	0.47
0.25	0.50	4.64	48.6	18.75	2.491	0.46
0.25	0.65	3.09	13.0	20.40	1.805	0.59
0.49	0.91	4.34	4.7	23.10	2.265	0.51
0.25	0.55	2.85	60.7	25.30	1.670	0.58
0.76	0.61	2.21	44.6	24.75	1.190	0.53
0.77	0.12	1.89	9.1	21.70	1.020	0.47
0.50	0.07	3.65	17.2	24.20	1.935	0.32
1.07	0.63	1.43	8.9	24.20	0.770	0.57
4.71	0.02	1.29	15.7	25.45	0.695	0.23
1.96	0.12	1.28	8.3	18.70	0.685	0.45
4.11	0.09	1.31	6.3	23.15	0.715	0.26
4.70	0.03	1.74	3.8	23.85	1.050	0.24
3.77	0.10	2.28	2.3	23.75	1.355	0.31
5.16	0.04	1.52	46.6	23.55	0.910	0.22
6.38	0.05	1.68	0.3	31.30	0.905	0.21
4.14	0.12	1.63	9.7	26.85	0.895	0.35
3.39	0.41	1.55	21.2	24.45	0.941	0.24
1.30	0.84	1.28	0.1	21.95	0.685	0.41
10.62	0.08	1.31	11.0	25.40	0.715	0.22
9.62	0.02	1.38	2.1	24.90	0.762	0.21
8.04	0.05	1.72	2.0	26.90	0.935	0.32
10.80	0.04	2.81	3.0	24.20	1.544	0.35
4.04	0.03	1.71	3.4	26.00	1.011	0.31
1.54	1.80	1.99	14.4	18.90	1.205	0.67
4.06	0.32	1.58	17.0	24.20	0.960	0.43
2.30	0.07	1.84	14.6	26.85	1.075	0.41
2.14	0.12	1.79	9.4	25.70	1.052	0.46
8.75	0.14	1.29	2.9	22.45	0.695	0.32
8.51	0.27	1.62	0.9	21.10	0.980	0.34
4.97	0.02	2.08	1.4	32.45	1.225	0.25
4.41	0.53	2.67	14.3	27.45	1.570	0.53
2.57	0.59	2.74	16.9	28.70	1.581	0.33
3.88	0.06	1.27	1.8	26.30	0.675	0.32
0.77	1.74	5.25	52.5	19.35	2.770	0.36

unknown function. Any function can be expressed as a weighted sum of a set of basis functions, that is, the transfer function of each hidden layer neuron is selected to form a set of basis functions to approximate the unknown. Build a neural network model and function. The RBF artificial neural network consists of an input layer, a hidden layer, and an output layer.

Build a general model: set the input layer as $X = [x_1, x_2, \dots, x_n]$, and the actual output layer as $Y = [y_1, y_2, \dots, y_p]$. The input layer realizes the nonlinear mapping from X to $R_i(X)$, the output layer realizes the linear mapping from

$R_i(X)$ to y_k , and the output of the k -th neuron network in the output layer is

$$\hat{y}_k = \sum_{i=1}^m w_{ik} R_i(X), k = 1, \dots, p. \quad (27)$$

In equation (27), n is the number of input nodes, and m is the number of hidden layer nodes; p is the number of output layer nodes; w_{ik} is the connection weight between the i -th neuron in the hidden layer and the k

-th neuron in the input layer; $R_i(X)$ is the hidden layer the action function of the i -th neuron in the layer, namely,

$$R_i(X) = \exp(-\|X - C_i\|^2/2\alpha_i^2), i = 1, \dots, m. \quad (28)$$

In equation (28), X is the n -dimensional input vector; C_i is the center of the i -th basis function, a vector with the same dimension as X ; α_i is the width of the i -th basis function; m is the number of perceptual units (the number of hidden layer nodes); the norm of the vector $\|X - C_i\|$, which usually represents the distance between X and C_i ; the unique maximum value of $R_i(X)$ at C_i . As $\|X - C_i\|$ increases, $R_i(X)$ decays rapidly to 0.

For a given input, only a small portion near the center of X is activated. Once the clustering centers C_i , weights w_{ik} , and α_i of the RBF network are all determined, the corresponding output values of the network can be given for a certain input.

In this paper, there are 6 independent variables and 1 dependent variable, the number of input neurons is taken as 6, the number of output neurons is taken as 1, and the number of neurons in the middle hidden layer is 0. The RBF network will be taken adaptively during the training process.

Use the data in Table 7 to fit the RBF neural network model to predict the decomposition effect of biologically diverse colonies when the local environment has different degrees of variability.

The data were imported into Matlab, and RBF neural networks were performed to fit and combine the predictions. The predictions were compared to single strain conditions to give an arbitrary combination of six initial factor values for colonies with species diversity.

For example, growth rate: 6.16; mycelial density: 0.04; humidity tolerance: 1.52; sulfur and phosphorus content: 46.6; temperature: 23.55; humidity: 0.91, enter code: $p_i = [6.16 \ 0.04 \ 1.52 \ 46.6 \ 23.55 \ 0.91]$.

The predicted value of decomposition rate t is 0.31, which is higher and better than the decomposition rate in the case of a single strain.

7.3. Conclusion Analysis. Based on the results of the above runs, we can find that the higher the species richness, the higher the decomposition rate. Because different species have different factors such as its mycelial growth rate and growth rate under different environmental conditions, i.e., different rates of decomposition of dead branches and leaves [44]. When the ambient temperature changes abruptly, the A class strains are affected more by weak strains, while the B class strains are affected less as the dominant strains, with a diversity of strains can complement each other between the AB class strains, the overall decomposition rate will not appear too big fluctuations, even if the community is affected more, it can gradually recover to the original level with time. Whereas a single strain is more affected by environmental changes, no complementary strain compensates for the decreased part of the decomposition rate [45]. In addition when it is affected more, it is less resilient than colonies with

material diversity. It is thus clear that biodiversity is important for the automatic balance of biological regulation in nature.

8. Conclusion

After our reasonable and rigorous model analysis, it was concluded that the decomposition rate of fungi was affected by temperature, humidity, colony abundance, time, soil sulfur and phosphorus content, mycelial elongation rate, and moisture tolerance. Under the given temperature and humidity conditions, different fungi have different mycelial densities, while they show competitive relationships in the short term and mutually beneficial symbiotic relationships in the long term. Different fungi are sensitive to environmental changes, so when environmental conditions change, the original dominant species may become the inferior species. For colonies with high species richness, there are dominant and inferior species in the colony, regardless of environmental changes. They have complementary strengths and weaknesses, and the rate of decomposition is dynamically balanced over time by the combined action of multiple fungi. In addition, plant communities are highly resistant to disturbance and recovery, so biodiversity provides stability to ecosystems.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was funded by the Teaching and Research Fund Project of the Education Department of Anhui Province (2019jyxm0186 and 2020jyxm0017), "First-Class Course" of Anhui University of Finance and Economics (acylk202008), and the Teaching and Research Fund Project of the Anhui University of Finance and Economics (ANJYYB2019053, acjyyb2019109, and acjyyb2020011).

References

- [1] G. H. Yu, S. G. Liu, and G. X. Cheng, "Protein: the store and transmitter of another type of information in biology—a reinterpretation of knowledge related to prion biology," *Advances in Biochemistry and Biophysics*, no. 8, pp. 698–706, 2005.
- [2] W. S. Zhao, Q. G. Guo, Z. H. Su et al., "Rhizosphere soil fungal community structure of healthy potato and Verticillium wilt plants and their utilization characteristics of carbon sources," *Chinese Agricultural Sciences*, vol. 54, no. 2, pp. 296–309, 2021.
- [3] S. Wang, S. Dou, Y. L. Liu, H. M. Li, and J. T. Cui, "Infrared spectroscopy study on the influence of microorganisms on the structural characteristics of humus after adding wheat

- straw to black soil," *Spectroscopy and Spectral Analysis*, vol. 32, no. 9, pp. 2409–2413, 2012.
- [4] Y. J. Liu, D. D. Fan, X. Z. Li, W. Q. Zhao, and Y. P. Kou, "Diversity of soil fungal communities in artificial and natural spruce forests and characteristics of bacterial community network relationships," *Journal of Applied Ecology*, no. 1, pp. 1–12, 2021.
 - [5] Z. H. Shen, Y. W. Li, K. Yang, and L. Chen, "The emerging cross-disciplinary studies of landscape ecology and biodiversity in China," *Journal of Geographical Sciences*, vol. 29, no. 7, pp. 1063–1080, 2019.
 - [6] M. H. He, *Effects of AM Fungi on Soil Litter Decomposition and Soil Nutrient Content under Different Competitive Treatments*, Guizhou University, 2019.
 - [7] Y. F. Zhang, *Effect of Fungi on Decomposition and Transformation of Corn Straw and Humus Formation*, Jilin Agricultural University, 2019.
 - [8] M. R. Chen, *Diversity of Endophytic Fungi in Fir Leaves and their Influence on the Decomposition Process of Apoplast*, Jishou University, 2020.
 - [9] F. J. Tan, *The Role of Biochar and Ectomycorrhizal Fungal Communities on Organic Carbon Decomposition in Forest Soils*, Jinan University, 2017.
 - [10] C. J. Lu, M. L. Lu, X. M. Liu, Y. H. Liu, C. H. Gao, and X. Y. Xu, "Diversity and antibacterial activity of gorgonian symbiotic fungi in Weizhou Island, Guangxi," *Journal of Tropical Oceanography*, no. 1, pp. 1–8, 2021.
 - [11] J. Hu, D. L. Meng, X. D. Liu, Y. L. Liang, H. Q. Yin, and H. W. Liu, "Response of soil fungal community to long-term chromium contamination," *Transactions of Nonferrous Metals Society of China*, vol. 28, no. 9, pp. 1838–1846, 2018.
 - [12] J. Lin, G. J. Zhao, L. X. Meng, and Z. P. Li, "Using X-ray diffraction technology and infrared spectroscopy to analyze wood eroded by fungi," *Spectroscopy and Spectral Analysis*, vol. 30, no. 6, pp. 1674–1677, 2010.
 - [13] S. S. Sun, X. M. Chen, and S. X. Guo, "Analysis of endophytic fungi in roots of *Santalum album* Linn. and its host plant *Kuhnia rosmarinifolia* Vent.," *Journal of Zhejiang University SCIENCE B*, vol. 15, no. 2, pp. 109–115, 2014.
 - [14] S. Y. Liu, Q. Z. Wang, H. X. Wu, S. S. Huang, and J. Feng, "The influence of common medium and optimized medium on fungal isolation and diversity in Fangchenggang waters," *Guangxi Science*, no. 1, pp. 1–9, 2021.
 - [15] F. J. Xiong, W. J. Liu, Z. Ding, J. H. Tang, S. L. Jiang, and W. Q. Guo, "Diversity and antibacterial activity of symbiotic fungi from peat moss in Jinggangshan National Nature Reserve," *Journal of Liaocheng University (Natural Science Edition)*, vol. 34, no. 3, pp. 101–110, 2021.
 - [16] M. Chen, Z. J. Chen, J. M. Liu et al., "Diversity analysis of soil microbes and endophytic fungi in the rhizosphere of *Phyllostachys edulis*," *Acta Ecologica Sinica*, no. 10, pp. 1–11, 2021.
 - [17] Y. C. Dai, Z. L. Yang, B. K. Cui et al., "Diversity and systematic study of important groups of macrofungi in China's forests," *Acta Physica Sinica*, pp. 1–36, 2021.
 - [18] G. Feng, X. C. Mi, H. Yan, F. Y. Li, J. C. Svenning, and K. Ma, "CForBio: a network monitoring Chinese forest biodiversity," *Science Bulletin*, vol. 61, no. 15, pp. 1163–1170, 2016.
 - [19] N. Lustenhouwer, D. S. Maynard, M. A. Bradford et al., "A trait-based understanding of wood decomposition by fungi," *Proceedings of the National Academy of Sciences*, vol. 117, no. 21, pp. 11551–11558, 2020.
 - [20] L. L. Liu and S. F. Chen, "Species diversity and distribution of *Lichella* fungi in soil," *Eucalyptus Science and Technology*, vol. 37, no. 4, pp. 48–59, 2020.
 - [21] Z. X. Guo, Y. I. Wang, B. W. Wu et al., "Study on the population genetic diversity and structure of ectomycorrhizal fungus *A. terrestris*," *Acta Mycology*, no. 1, pp. 1–16, 2021.
 - [22] J. Gong, H. Y. Ma, S. L. Zheng et al., "Effects of continuous cropping on potato phenolic acid autotoxins and rhizosphere fungal communities," *Northwest Agricultural Journal*, no. 3, pp. 1–8, 2021.
 - [23] W. W. Yan, F. Y. Zhao, and S. Y. Liu, "Diversity and community composition of fungi in the rhizosphere soil of *Morchella* in Liaoyuan area, Jilin Province," *Fungi Research*, no. 1, pp. 1–12, 2021.
 - [24] L. M. Bao, Y. F. Ding, Y. L. Wei, F. T. Zi, and Y. Tan, "Analysis on the composition and diversity of fungal communities in continuous cropping and fallow soils of *Panax notoginseng*," *Chinese Medicinal Materials*, no. 1, pp. 7–12, 2021.
 - [25] M. Li and X. H. Gao, "Community structure and driving factors of ectomycorrhizal fungi around the roots of *Betula platyphylla* in Daqingshan," *Journal of Ecology*, no. 1, pp. 1–10, 2021.
 - [26] J.-M. Zhu, L. Wang, and J.-B. Liu, "Eradication of Ebola based on dynamic programming," *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 1580917, 9 pages, 2016.
 - [27] S. Y. Liu, F. Ma, and J. Q. Zhang, "Study on algae succession and diversity index in the process of eutrophication simulation of landscape water body," *Acta Scientiae Circumstantiae*, no. 2, pp. 337–341, 2007.
 - [28] W. D. Chou and Q. Fang, "The effect of urbanization on soil microbial diversity," *Journal of Science & Technology Economics*, vol. 29, no. 5, pp. 124–125, 2021.
 - [29] L. Z. Chen, Y. L. Li, and Y. L. Yu, "Soil bacterial and fungal community successions under the stress of chlorpyrifos application and molecular characterization of chlorpyrifos-degrading isolates using ERIC-PCR," *Journal of Zhejiang University SCIENCE B*, vol. 15, no. 4, pp. 322–332, 2014.
 - [30] Y. Long, D. S. Shen, H. M. Lao, L. F. Hu, and Y. M. Zhu, "Distribution characteristics of microorganisms in different particle sizes of old garbage in domestic waste landfills," *Acta Scientiae Circumstantiae*, no. 9, pp. 1485–1490, 2007.
 - [31] X. L. Wang, M. X. Wang, X. G. Xie et al., "An amplification-selection model for quantified rhizosphere microbiota assembly," *Science Bulletin*, vol. 65, no. 12, pp. 983–986, 2020.
 - [32] H. M. Ge and R. X. Tan, "Symbiotic bacteria-an important source of new active natural products," *Progress in Chemistry*, vol. 21, no. 1, pp. 30–46, 2009.
 - [33] S. E. Lu, B. Xiao, F. M. Ren, W. Zhuo, and H. Y. Huang, "Analysis of rhizosphere soil fungal community structure and diversity of *Polygonatum* root rot based on Illumina Miseq," *World Science and Technology-Modernization of Traditional Chinese Medicine*, no. 1, pp. 1–7, 2021.
 - [34] Q. Ma, D. S. Xia, Z. H. Wu et al., "Diversity analysis of soil fungi in western Inner Mongolia," *Science Technology and Engineering*, vol. 20, no. 35, pp. 14447–14454, 2020.
 - [35] M. Wei, S. Wang, H. G. Xiao, B. D. Wu, K. Jiang, and C. Y. Wang, "Co-invasion of daisy fleabane and Canada goldenrod pose synergistic impacts on soil bacterial richness," *Journal of Central South University*, vol. 27, no. 6, pp. 1790–1801, 2020.