Am J Transl Res 2022;14(9):6521-6535
www.ajtr.org /ISSN:1943-8141/AJTR0143677

Original Article

Identification and exploration of the pyroptosis-related molecular subtypes of breast cancer by bioinformatics and machine learning

Li Zhang, Xiu-Feng Chu, Jing-Wei Xu, Xue-Yuan Yao, Hong-Qiao Zhang, Yan-Wei Guo

Department of Oncology, The Fifth Affiliated Hospital of Zhengzhou University, Zhengzhou, China

Received April 21, 2022; Accepted August 4, 2022; Epub September 15, 2022; Published September 30, 2022

Abstract: Objectives: To classify breast cancer (BRCA) according to the expression of pyroptosis-related genes and explore their molecular characteristics. Methods: Nonnegative matrix factorization (NMF) was used for subtype classification based on 21 pyroptosis-related genes in the TCGA database. Survival analysis and t-distributed stochastic neighbor embedding (t-SNE) analysis were conducted to assess the NMF results' performance. XGBoost, CatBoost, logistic regression, neural network, random forest, and support vector machine were utilized to perform supervised machine learning and construct prediction models. Genetic mutations, tumor mutational burden, immune infiltration, methylation, and drug sensitivity were analyzed to explore the molecular signatures of different subtypes. Lasso, RF, and Cox regression were operated to construct a prognostic model based on differentially expressed genes. Results: BRCA patients were divided into two subtypes (named Cluster1 and Cluster2). Survival analysis (P = 0.02) and t-SNE analysis demonstrated that Cluster1 and Cluster2 were well classified. The XGBoost model achieved reliable predictions on both training and validation sets. Regarding molecular characteristics, Cluster1 had higher TMB, immune cell infiltration, and m^6A methylation-related gene expression than Cluster2. There was also a statistically significant difference between the two subtypes concerning drug susceptibility. Finally, a 5-gene prognostic model was constructed using Lasso, RF, and Cox regression and validated in the GEO database. Conclusion: Our study may provide new insights from bioinformatics and machine learning for exploring pyroptosis-related subtypes and their respective molecular signatures in BRCA. In addition, our models may be helpful for the treatment and prognosis of BRCA.

Keywords: Breast cancer, pyroptosis, subtype, bioinformatics, machine learning

Introduction

Breast cancer (BRCA) is now the most common cancer worldwide [1], which accounts for approximately 31% of new cancer cases in women [2]. Clinical outcomes of BRCA patients have been greatly improved due to the development of molecular characterization and its related targeted treatments, including *ER*, *PR*, *HER2* (*ERBB2*), and *Ki-67* (*MKI67*). Based on the molecular subtyping, the administration of chemotherapy, endocrine therapy, *ERBB2*-targeted interventions, or their combination could be given reasonably [3]. In addition, the efficacy of novel drugs such as CDK4 and CDK6 inhibitors, PI3K inhibitors, PARP inhibitors, and immune checkpoint inhibitors also depends on the molecular characteristics of BRCA [4]. So, it is of great significance to find new subtypes to explore their respective molecular features for the treatment of BRCA.

Pyroptosis, a type of programmed cell death characterized by the pore formation of gasdermins in the plasma membrane, can activate potent proinflammatory responses [5]. Many studies have shown that pyroptosis plays an essential role in many tumors and is closely related to tumor therapy [6-8]. Further, some studies have shown that pyroptosis is associated with chemotherapy-induced cell damage and may increase the efficacy of chemotherapy drugs [6, 9, 10]. Pyroptosis has also been involved in immunotherapy. A previous study showed that PD-L1 mediates an inflammatory form of cell death in tumor cells by activating the expression of *GSDMC*, which ultimately leads to tumor necrosis [11].

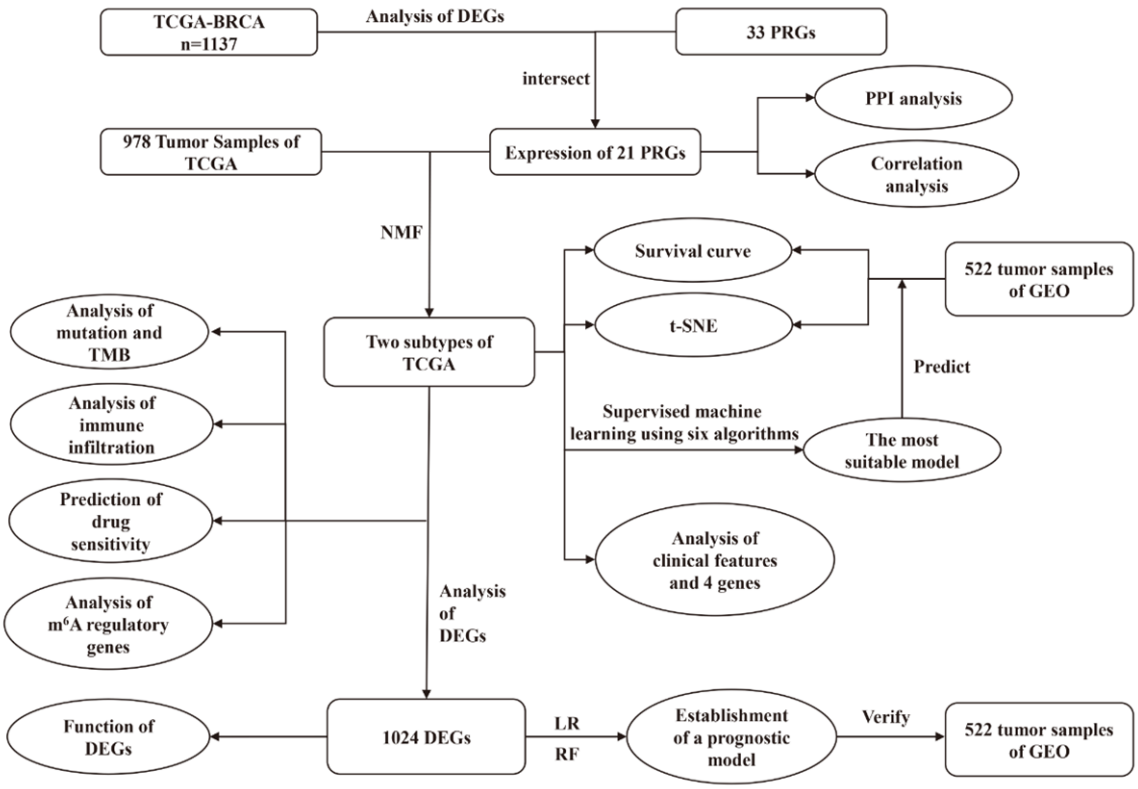

Figure 1. The flow chart of this study. DEGs, differentially expressed genes; PRGs, pyroptosis-related genes; PPI, protein-protein interaction; TMB, tumor mutation burden: t-SNE, t-distributed stochastic neighbor embedding; LR, logistic regression; RF, random forest.

As gene sequencing technology advances, researchers are starting to classify cancers based on the expression of specific genes, which can help identify new subtypes and tailor treatments to their characteristics.

In this study, gene expression matrices and the clinical information of samples were obtained from the public databases TCGA and GEO. Based on the expression of 21 pyroptosis-related genes (PRGs) obtained in previous studies and the differentially expressed genes (DEGs), BRCA in patients in the TCGA database was divided into two subtypes using nonnegative matrix factorization (NMF). Next, a prediction model was constructed using machine learning algorithms, which showed promising results in the internal validation set. We used this model to predict grouping in the GEO database. We then explored somatic variants, immune infiltration, drug sensitivity, and methylation between these two subtypes. Subsequently, we performed DEGs analysis between the two subtypes and constructed a prognostic model based on DEGs. The flow chart of this study is shown in **Figure 1**.

Materials and methods

Datasets

First, we obtained 1137 BRCA samples with gene expression and clinical information from the TCGA database, including 97 normal tissue samples and 1040 tumor tissue samples. Then, as the clinical data of tumor samples were processed, 978 tumor tissue samples were obtained by selecting samples with a survival time of more than 30 days. Then, we downloaded GSE20685 [12], GSE20711 [13], and GSE58812 [14] from the GEO database as external validation datasets and obtained 522 tumor tissue samples after the same data processing.

Identification of the subtypes

Previous studies [5, 15-17] identified 33 PRGs, as shown in Supplementary Table 1. We per-

formed the analysis of DEGs (adjusted *P* value < 0.05) using the DESeq2 R package [18] on normal tissues and tumor tissues in the TCGA database and selected 21 PRGs according to the intersection between DEGs and 33 PRGs. NMF, an unsupervised learning algorithm based on decomposition by parts, has successfully explained biologically meaningful categories. This method has recently been widely used to identify tumor subtypes [19]. We used this method to divide 978 BRCA samples from the TCGA database into two subtypes based on the expression of these 21 PRGs. To determine the accuracy of the NMF results, we performed survival analysis and t-distributed stochastic neighbor embedding (t-SNE) analysis on the groupings obtained by NMF.

Building the predictive model

According to the results of NMF in the TCGA database, we used six machine learning methods for supervised machine learning, including XGBoost, CatBoost, logistic regression (LR), neural network (NNET), random forest (RF), and support vector machine (SVM). These methods were implemented with the xgboost R package, the catboost R package, and the mlr3 R package. Receiver operating characteristic (ROC) curves were used to compare the performance of these six models. We next tested the classification performance of the model on the GEO database.

Analysis of the somatic variants of the two subtypes

We downloaded the masked mutation files of BRCA samples from the TCGA database and used the MAftools R package [20] to analyze the somatic mutations of the two subtypes. Next, we examined whether there was a significant difference between the two subtypes in tumor mutation burden (TMB), a valuable biomarker to assess immunotherapy efficiency [21].

Analysis of immune infiltration in the two subtypes

TIMER2.0 is a database that contains TIMER, CIBERSORT, quanTIseq, xCell, MCP-counter, and EPIC algorithms [22]. We used this database to perform a statistical analysis of the immune infiltration of the two subtypes. The immune infiltration is often used to evaluate the efficacy of immunotherapy [23].

Prediction of drug sensitivity for the two subtypes

We used the pRRophetic R package to predict the drug sensitivity of some commonly used drugs in BRCA [24]. This R package makes predictions based on the 50% inhibitory concentration (IC_{50}), and the smaller value has a higher sensitivity.

Expression analysis of N^6-methyladenosine regulatory genes in the two subtypes

N^6-methyladenosine (m^6A) is adenosine that is methylated at the N^6 position. This change is involved in the occurrence and development of tumors by regulating the expression of the tumor-related genes *BRD4*, *MYC*, *SOCS2*, and *EGFR* [25]. Fifteen m^6A regulatory genes were selected from previous studies [25] and the GeneCards database (Supplementary Table 1), and their expression was compared between the two subtypes.

DEGs functional enrichment analysis and the establishment of the prognostic model

We analyzed the DEGs between the two subtypes using the DESeq2 R package. We set the selection conditions for an adjusted *P* value < 0.05, |log2FC| > 1. The clusterProfiler R package [26] was used for Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis based on these DEGs. Next, we used Lasso regression [27], random forest, and Cox regression to identify related prognostic genes and build a prognostic model based on these genes. The risk score is equal to the result of matrix multiplication between the exponential operation of each mRNA's Cox partial regression coefficient and the matrix of mRNA expression. GSE20685, GSE20711, and GSE58812 were utilized to verify this scoring model.

Statistical analysis

All statistical analyses were performed using R version 4.0.5. The sva R package was used to remove batch effects. Nonnormally distributed data were tested using the Wilcoxon test. The chi-square test was used to analyze the distri-

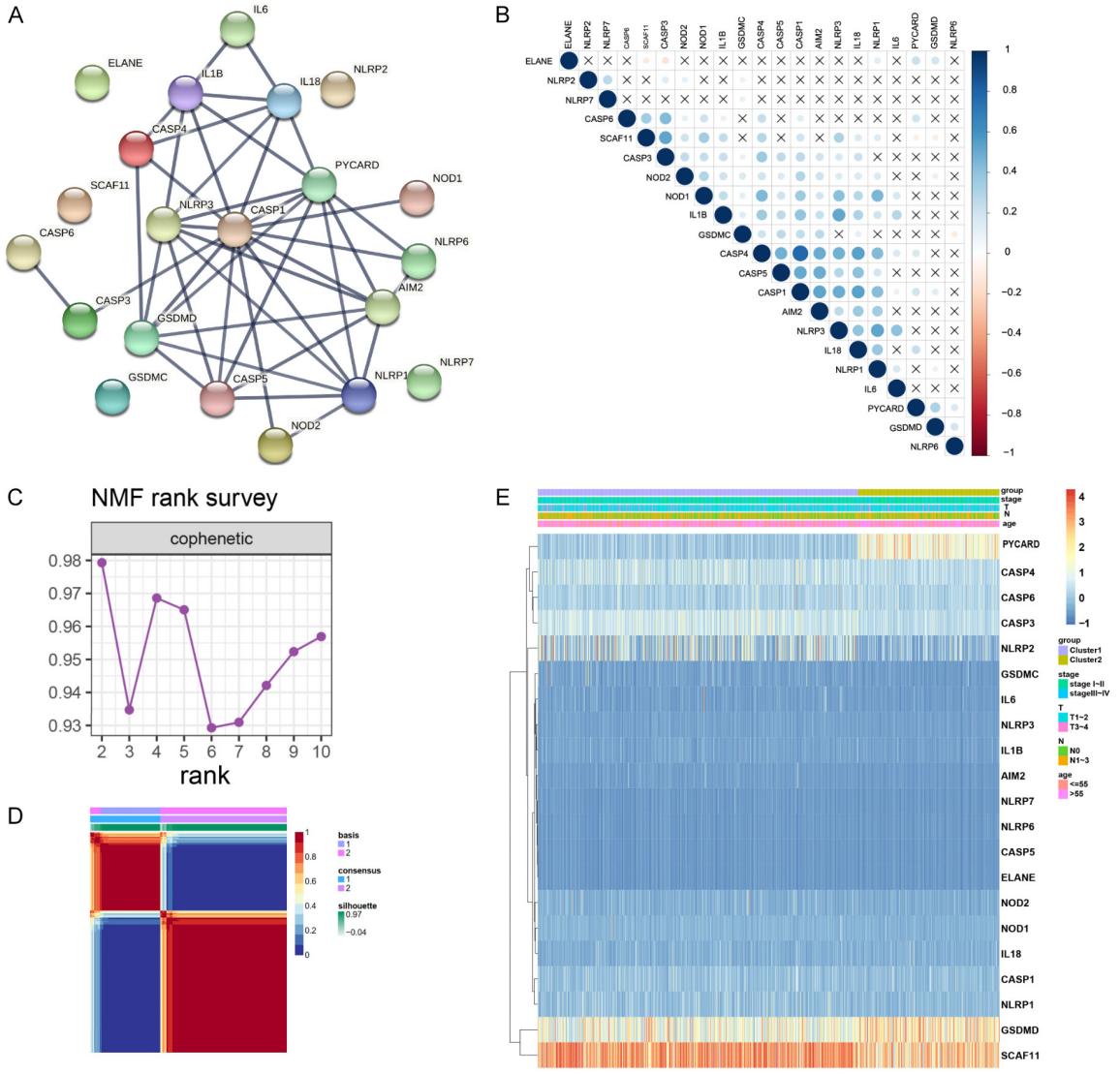

Figure 2. Analysis of the 21 PRGs and results of NMF. A: The network plot of PPI based on the 21 PRGs. B: The heatmap of correlation of the 21 PRGs. The "×" indicates that *P*-value is less than 0.01. C: The cophenetic value of each rank in NMF. D: The classification of 978 BRCA samples using NMF, with rank = 2. E: The heatmap of the 21 PRGs expressions.

bution proportions. Survival curves were compared by the log-rank test. Unless otherwise specified, *P* values < 0.05 were statistically significant.

Results

Identification and validation of BRCA subtypes

To determine which of the 33 PRGs was differentially expressed between normal and tumor samples of BRCA in the TCGA database, we performed differential gene expression analysis using the DESeq2 R package (adjusted *P* value < 0.05). This analysis obtained 21 PRGs, including *IL6*, *ELANE*, *NLRP1*, *IL1B*, *NOD1*, *NLRP3*, *CASP*, *CASP4*, *SCAF11*, *CASP3*, *CASP6*, *CASP5*, *GSDMD*, *NLRP2*, *IL18*, *NLRP6*, *NOD2*, *NLRP7*, *AIM2*, *PYCARD*, and *GSDMC*. We then performed an analysis of protein-protein interactions (PPIs) (**Figure 2A**) and the correlation of expression (**Figure 2B**) to explore the relationship between these 21 genes, and the minimum required interaction score for the PPI analysis was set at 0.9. We found that *CASP1*, *NLRP3*, *PYCARD*, *GSDMD*, and *NLRP1* interacted strongly among these 21 genes. Next, we

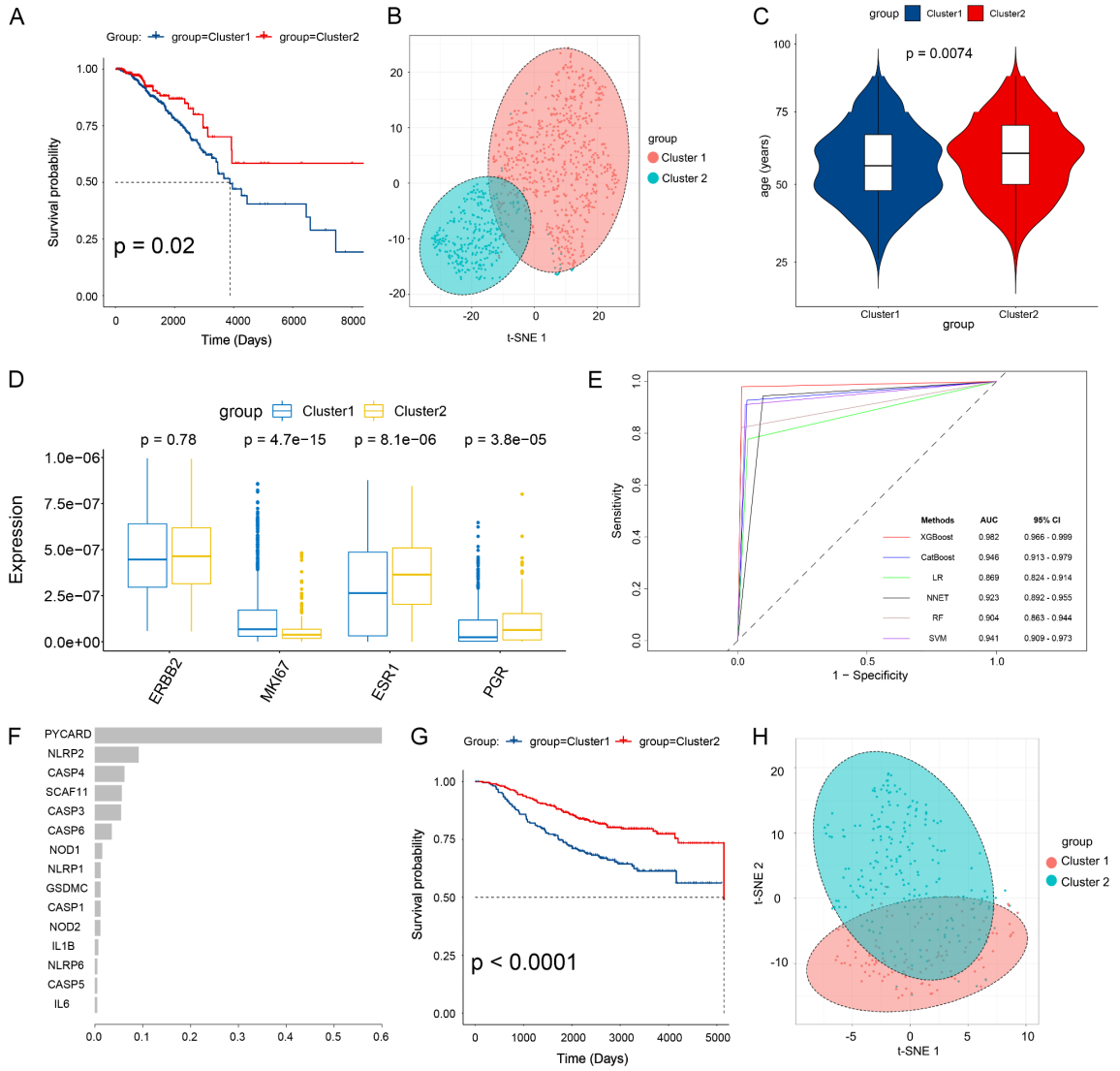

Figure 3. Analysis of the difference in characteristics of Cluster1 and Cluster2. A: Survival curves of the two subtypes in the TCGA database. B: The distribution analyzed by t-SNE of the two subtypes in the TCGA database. C: Differences in the age distribution of the two subtypes in the TCGA database. D: Four gene expression differences between the two subtypes in the TCGA database. E: Receiver operating characteristic (ROC) curves of the six machine learning algorithms based on the validation set. F: The importance of the 21 PRGs affecting the XGBoost model. G: Survival curves of the two subtypes predicted by the XGBoost model in the GEO database. H: The distribution analyzed by t-SNE of the two subtypes in the GEO database.

obtained rank (*rank* = 2-10) subtypes using NMF based on the expression of these 21 PRGs in 978 BRCA samples. As shown in **Figure 2C**, the polyline changed the most when rank = 2 and rank = 3, so rank = 2 was the front point where the polyline changed the most. As shown in **Figure 2D**, it is appropriate that the samples were divided into two subtypes. Therefore, we divided these 978 BRCA samples into two subtypes and named them Cluster1 (n = 683) and Cluster2 (n = 295). We drew a heatmap to explore the expression of the 21 PRGs of Cluster1 and Cluster2 (**Figure 2E**). The heatmap showed that the expression of *PYCARD* in Cluster2 was significantly higher than that in Cluster1.

Next, we plotted the survival curves of the two subtypes using the survival R package by the Kaplan-Meier method. In addition, we performed t-SNE analysis of the two subtypes using the Rtsen R package. **Figure 3A** shows

Table 1. Clinical information for Cluster1 and Cluster2

Characteristic	Cluster1, N = 683[1]	Cluster2, n = 295[1]	*p*-value[2]
Age			0.007
Mean (SD)	57.0 (12.9)	59.4 (12.8)	
T			0.215
T1	187 (27%)	73 (25%)	
T2	386 (57%)	177 (60%)	
T3	80 (12%)	40 (14%)	
T4	27 (4.0%)	5 (1.7%)	
TX	3 (0.4%)	0 (0%)	
N			0.084
N0	314 (46%)	132 (45%)	
N1	224 (33%)	112 (38%)	
N2	86 (13%)	23 (7.8%)	
N3	45 (6.6%)	25 (8.5%)	
NX	14 (2.0%)	3 (1.0%)	
M			0.062
cM0 (i+)	4 (0.6%)	2 (0.7%)	
M0	574 (84%)	232 (79%)	
M1	16 (2.3%)	4 (1.4%)	
MX	89 (13%)	57 (19%)	
Stage			0.221
stage I	121 (18%)	45 (15%)	
stage II	374 (55%)	180 (61%)	
stage III	155 (23%)	63 (21%)	
stage IV	14 (2.0%)	4 (1.4%)	
stage X	19 (2.8%)	3 (1.0%)	

[1]n (%); [2]Wilcoxon rank sum test; Fisher's exact test; Pearson's Chi-squared test.

that Cluster2 had better survival outcomes than Cluster1 (P = 0.02), and t-SNE analysis showed that the two subtypes were well differentiated (**Figure 3B**). **Figure 3A** and **3B** demonstrate that the results obtained by NMF were meaningful in the TCGA database. To obtain more information on the differences between the two subtypes, we analyzed the expression of some critical genes in BRCA and the clinical characteristics of the two subtypes (**Table 1**). The results of the analysis of the clinical characteristics showed that the age of Cluster1 was lower than that of Cluster2 (**Figure 3C**). The rest of the clinical characteristics, including stage (I-II or III-IV), T stage (1-2 or 3-4), and N stage (0 or 1-3), were not statistically significant. We also found that the expression of ERBB2 in the two subtypes was not statistically significant, the expression of MKI67 in Cluster1 was significantly higher than that in Cluster2, and the expression of PGR and ESR1 in Cluster1 was considerably lower than that in Cluster2 (**Figure 3D**).

According to the results of NMF, we used six machine learning methods, including XGBoost, CatBoost, LR, NNET, RF, and SVM, for supervised machine learning. We split the data into 70% and 30%, with the 70% part as the training set and the remaining 30% as the validation set. The area under curve (AUC) of XGBoost, CatBoost, LR, NNET, RF, and SVM was 0.982, 0.946, 0.869, 0.923, 0.904, and 0.941, respectively (**Figure 3E**). The 95% confidence intervals for these six algorithms are also shown in **Figure 3E**. The model obtained by XGBoost was the most suitable. We then assessed the importance of the genes affecting the model and found that *PYCARD* significantly impacted the model (**Figure 3F**). To further verify the results of NMF, we used the XGBoost model to make a prediction on 522 GEO database samples of BRCA. Based on the predicted results, we drew survival curves (**Figure 3G**) and performed t-SNE analysis (**Figure 3H**). Results of these two analyses were similar to those from the TCGA database, justifying the classification from the external validation.

Analysis of the somatic variants of Cluster1 and Cluster2

In this part, we used the maftools R package to analyze the differences in somatic mutations of the two subtypes in the TCGA database. **Figure 4A** and **4B** show the summary of the mutations of Cluster1 and Cluster2, respectively. Except for the apparent difference in the mutated genes, there was little difference in other somatic mutations, such as variant classification, variant types, and the classification of single nucleotide variants. Therefore, we drew detailed pictures of genetic mutations of the two subtypes (only the top 10 are shown). **Figure 4C** demonstrates that the primary genetic mutations in Cluster1 and Cluster2 were *TP53* (41%) and *PIK3A* (45%). **Figure 4D**, the forest plot, shows a statistical comparison of the gene mutations in Cluster1 and Cluster2 (only genes with P < 0.01 are shown). **Figure 4E** and **4F** show the TMB distribution of Cluster1 and Cluster2, respectively. We used a boxplot (**Figure 4G**) to determine that the differ-

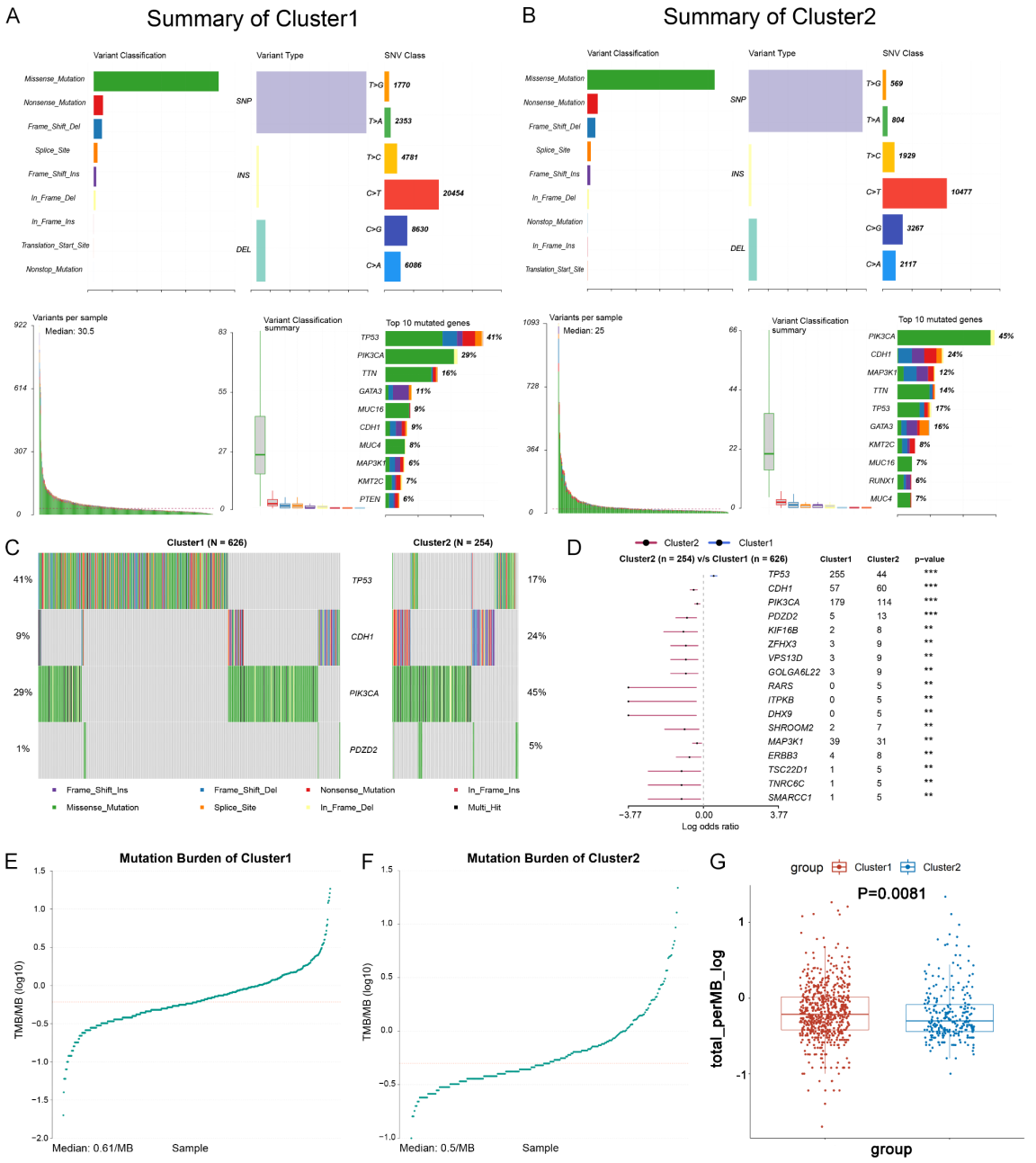

Figure 4. Somatic mutations and TMB analysis of Cluster1 and Cluster2. A, B: The Summary of mutations for Cluster1 and Cluster2, respectively. C: Comparison of gene mutations between Cluster1 and Cluster2 (only genes with P < 0.001 were shown). D: The forest plot comparing the gene mutations of Cluster1 and Cluster2 (only genes with P < 0.01 were shown). E: The TMB of each sample in Cluster1. F: TMB of each sample in Cluster2. G: The boxplot of TMB for Cluster1 and Cluster2. $^{**}P < 0.01$; $^{***}P < 0.001$.

ence in TMB between the two was statistically significant ($P = 0.0081$), i.e., Cluster1 had higher TMB than Cluster2.

Analysis of immune infiltration in Cluster1 and Cluster2

The immune infiltration is commonly used to evaluate the efficacy of immunotherapy [23], so we used TIMER, CIBERSORT, quanTIseq, xCell, MCP-counter, and EPIC to score the immune infiltration of the two subtypes. We screened out the items with statistical significance ($P < 0.05$) for plotting (xCell had too many significant items, and we only drew pictures related to the immune score). The results of the TIMER algorithm showed that the two subtypes had no

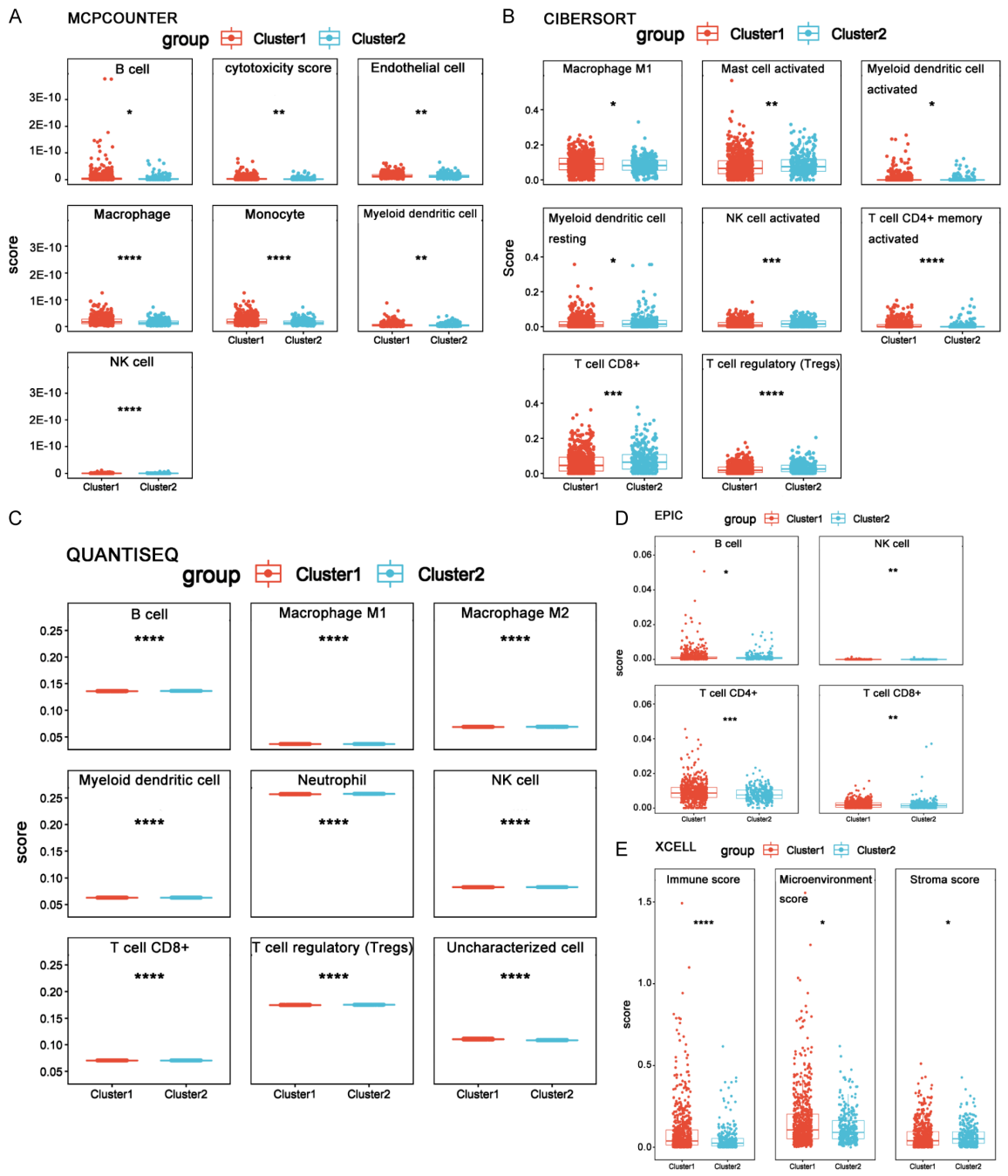

Figure 5. Statistical analysis of Cluster1 and Cluster2 in immune infiltration derived from several algorithms. A: Statistically significant items in MCP-counter. B: Statistically significant items in CIBERSORT. C: Statistically significant items in quanTIseq. D: Statistically significant items in EPIC. E: Statistically significant items in xCell (only showed items related to the immune score). *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$.

items with statistical significance. The results of CIBERSORT (**Figure 5B**) indicated that the degree of immune cell infiltration of the two subtypes could not be recapitulated in general. The results of quanTIseq (**Figure 5C**) showed that the degree of immune infiltration of Cluster2 was higher than that of Cluster1. The scores of MCP-counter (**Figure 5A**), EPIC (**Figure 5D**), and xCell (**Figure 5E**) all showed that the degree of immune infiltration of Cluster1 was higher than that of Cluster2. Combining the results of the six algorithms, the degree of immune infiltration of Cluster1 was higher than that of Cluster2.

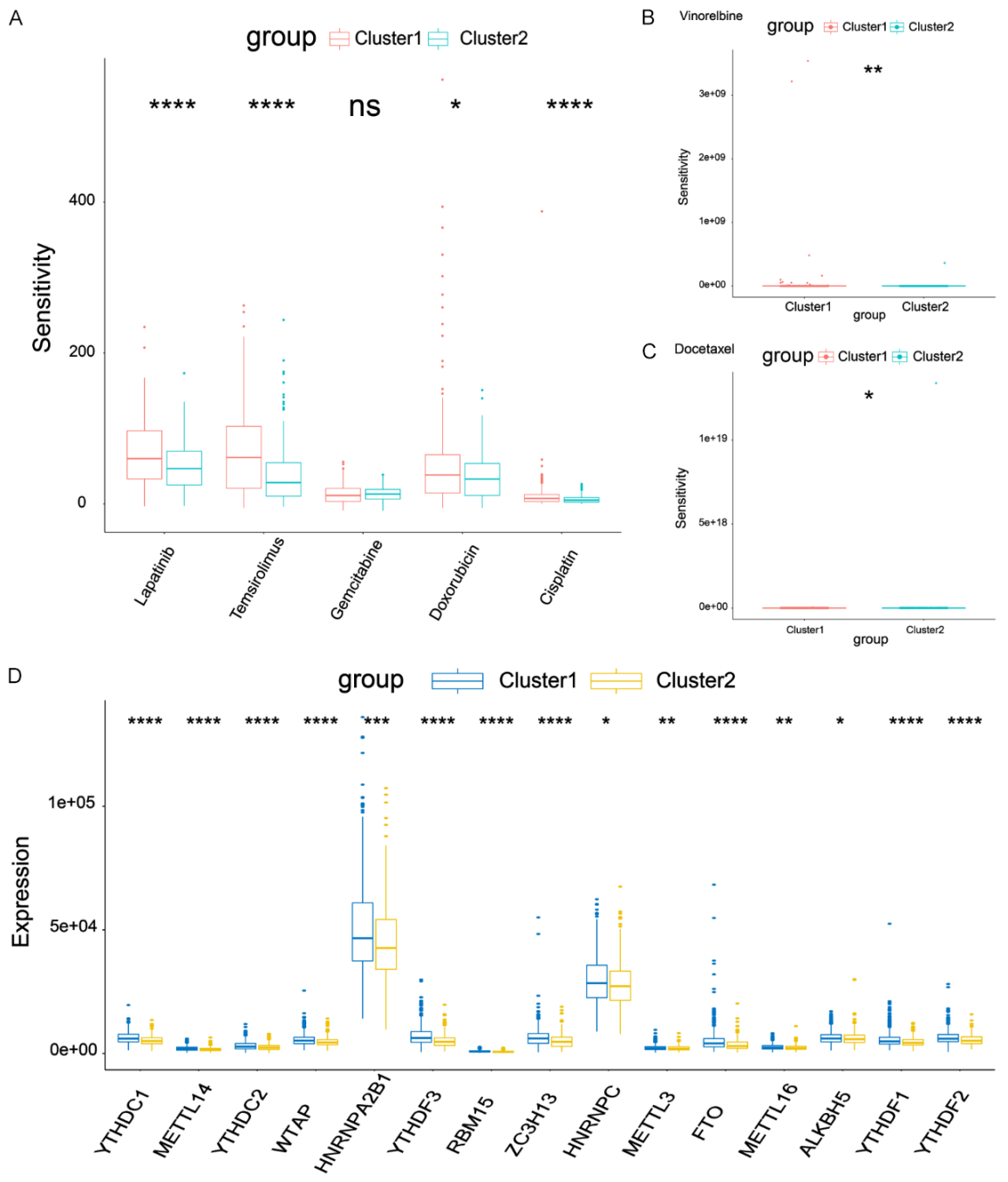

Figure 6. Drug sensitivity analysis and m^{6}A methylation-related gene expression analysis of Cluste1 and Cluster2. A: The boxplot of sensitivity of Cisplatin, Doxorubicin, Gemcitabine, Temsirolimus, and Lapatinib in the two subtypes. B: The boxplot of sensitivity of Docetaxel in the two subtypes. C: The boxplot of sensitivity of Vinorelbine in the two subtypes. D: Expression analysis of 15 m^{6}A methylation-related genes between the two subtypes. *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.001; ns, not significant.

Prediction of drug sensitivity for Cluster1 and Cluster2

To explore the difference in drug sensitivity of the two subtypes, we used the pRRophetic R package to perform a sensitivity analysis of some commonly used BRCA drugs for the two subtypes. These drugs are cisplatin, docetaxel, doxorubicin, gemcitabine, vinorelbine, temsirolimus, and lapatinib. The results showed no significant difference in the sensitivity to gemcitabine (**Figure 6A**). Cluster1 was more sensi-

tive to docetaxel than Cluster2 (**Figure 6C**), and Cluster2 was more sensitive to the other five drugs (cisplatin, doxorubicin, vinorelbine, temsirolimus, and lapatinib) than Cluster1 (**Figure 6A**, **6B**).

Expression analysis of N^6-methyladenosine regulatory genes in Cluster1 and Cluster2

In this section, we analyzed 15 genes related to N^6 methylation regulation. We found that these genes were expressed at higher levels in Cluster1 than in Cluster2 (**Figure 6D**). Furthermore, except for *ALKBH5*, *HNRNPC* ($0.01 < P < 0.05$), *METTL16*, *METTL3* ($0.001 < P < 0.01$) and *HNRNPA2B1* ($0.0001 < P < 0.001$), the differential expression of the remaining 10 genes was significant ($P < 0.0001$).

Functional enrichment analysis and the establishment of the prognostic model based DEGs of Cluster1 and Cluster2

First, we used the DESeq2 R package to identify the DEGs of the two subtypes and obtained 1024 DEGs (|log2FoldChange| > 1 and adjusted *P* value < 0.05). Based on these 1024 DEGs, we performed KEGG pathway analysis and GO enrichment analysis. The results of the KEGG pathway analysis showed that the neuroactive ligand-receptor interaction signaling pathway occupied a dominant position in both the number of genes and the statistical significance (**Figure 7A**). The results of the GO enrichment analysis of these DEGs showed that DEGs were mainly concentrated in various channels and transporters in molecular function (MF) (**Figure 7B**). There were many kinds of biological process (BP) results, and the humoral immune response and various keratinization were the main terms (**Figure 7C**). The cellular components (CC) were mainly synaptic membrane and transporter (**Figure 7D**), which were closely related to the results of MF.

Next, we used RF to obtain 27 prognosis-related genes and Lasso regression to obtain 20 prognosis-related genes. We took the intersection of the results obtained by these two methods and obtained five prognosis-related genes (**Figure 7E**). Cox regression was performed based on these five prognosis-related genes. According to the results of multivariate Cox regression, *APOH*, *IYD*, *LRRTM3*, *PAX7*, and *SMR3A* were identified as key prognostic genes. The risk score was equal to the result of matrix multiplication between the exponential operation of the Cox partial regression coefficient of each mRNA and the matrix of mRNA expression (risk score = 1.140281 × *APOH* + 1.096954 × *IYD* + 1.233628 × *LRRTM3* + 1.081889 × *PAX7* + 1.221493 × *SMR3A*). According to the median risk score, the BRCA samples of TCGA were divided into high-score and low-score groups, and survival analysis was performed. **Figure 7F** shows that the low group lived longer than the high group ($P < 0.0001$). We next made a dynamic nomogram using the DynNom R package (**Figure 7G**), which is convenient for clinical use, and a nomogram using the rms R package to predict the 3-, 5- and 10-year survival rates (**Figure 8A**). The next step was to evaluate the models. We evaluated the survival model of the high-score and low-score groups using the time-dependent ROC curve. The AUC values for three years, five years, and ten years were 0.66, 0.65, and 0.64, respectively (**Figure 8B**). In addition, we used calibration curves to evaluate the nomogram for predicting 3-, 5-, and 10-year survival (**Figure 8C-E**). The calibration curve showed that the nomogram predicted the nomogram well. We tried to predict GSE20685, GSE20711, and GSE58812 using the model obtained from these five genes, but we found that the GPL570 platform could not recognize *LRRTM3*, so we had to use *APOH*, *IYD*, *PAX7*, and *SMR3A* for a tentative prediction. **Figure 8F** displays that the difference in the survival curves between the high-score and low-score groups was statistically significant in the GEO database ($P = 0.028$).

Discussion

With the continuous development of gene sequencing and biomedical technology, the types of BRCA are also increasing, and the treatment methods are constantly being enriched. In this study, we analyzed the expression matrix of the 21 PRGs derived from the analysis of DEGs in the TCGA database by NMF and finally obtained two subtypes (Cluster1 and Cluster2). We performed supervised machine learning by six algorithms based on the results of NMF to build the XGBoost predictive model with an AUC value of 0.982. We used this model to predict the two subtypes in the GEO database to demonstrate that the classification was meaningful in external databases. Next,

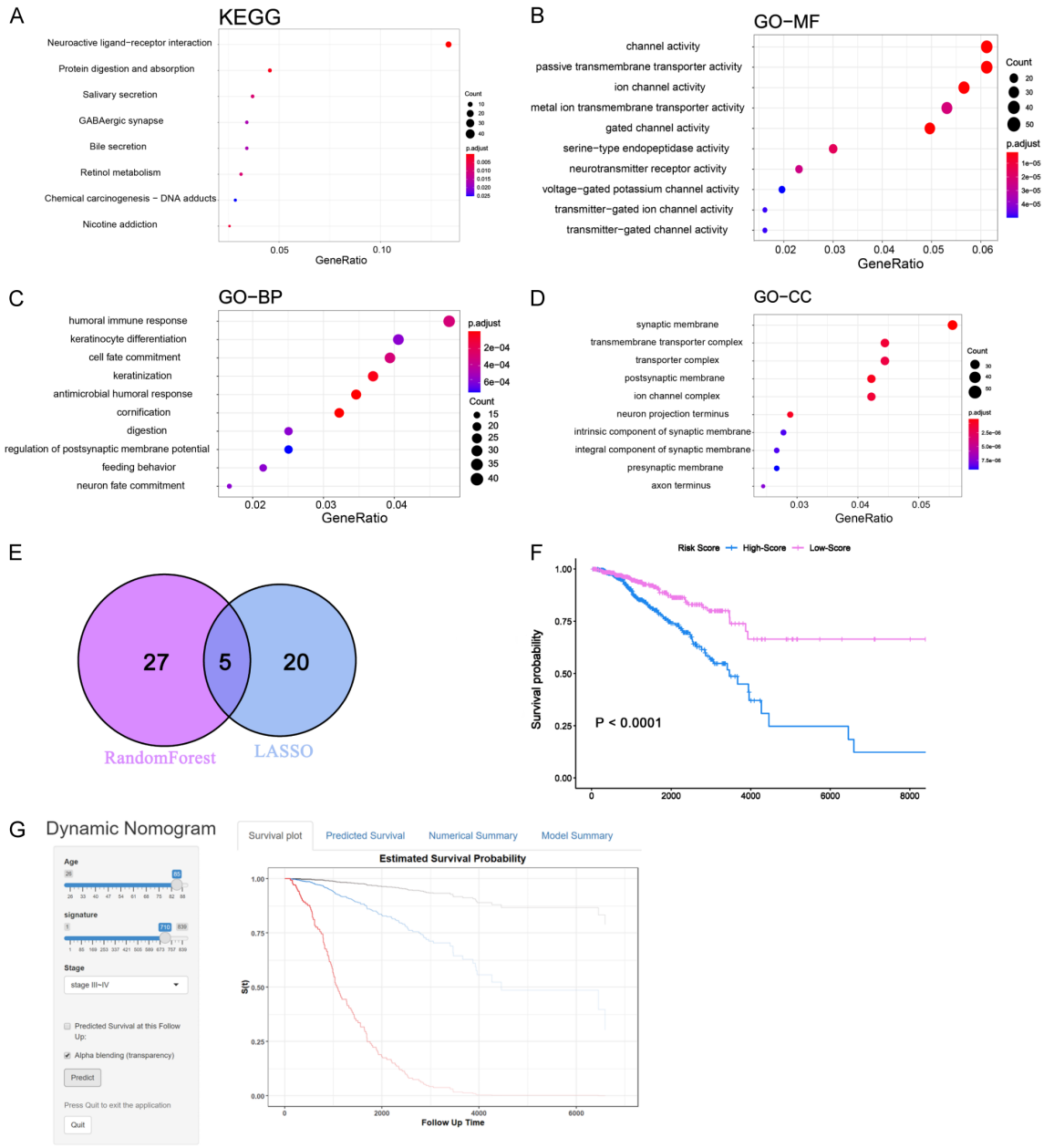

Figure 7. Functional analysis and prognostic correlation analysis of Cluste1 and Cluster2. A: The result of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis. B: The molecular function (MF) of Gene Ontology (GO) enrichment analysis. C: The biological process (BP) of GO enrichment analysis. D: The cellular components (CC) section of GO enrichment analysis. E: The Venn diagram of LASSO and random forest results. F: Survival curves for high-score and low-score groups using the Kaplan-Meier method. G: Part of function display of the dynamic nomogram.

we explored and demonstrated the differences between the two subtypes in somatic variants, immune infiltration, and N^6-methyladenosine. We then examined the sensitivity of the two subtypes to some drugs. We further performed GO enrichment analysis and KEGG pathway analysis based on the DEGs of the two subtypes, and the results showed that DEGs were mainly concentrated in synaptic membranes, channels, and transporters. Finally, we developed a five-gene prognostic model that predicted survival at 3, 5, and 10 years in the high- and low-risk groups and tested the model in the GEO database. Together, these results demonstrated that this classification was meaningful and provided new insights into the molecular

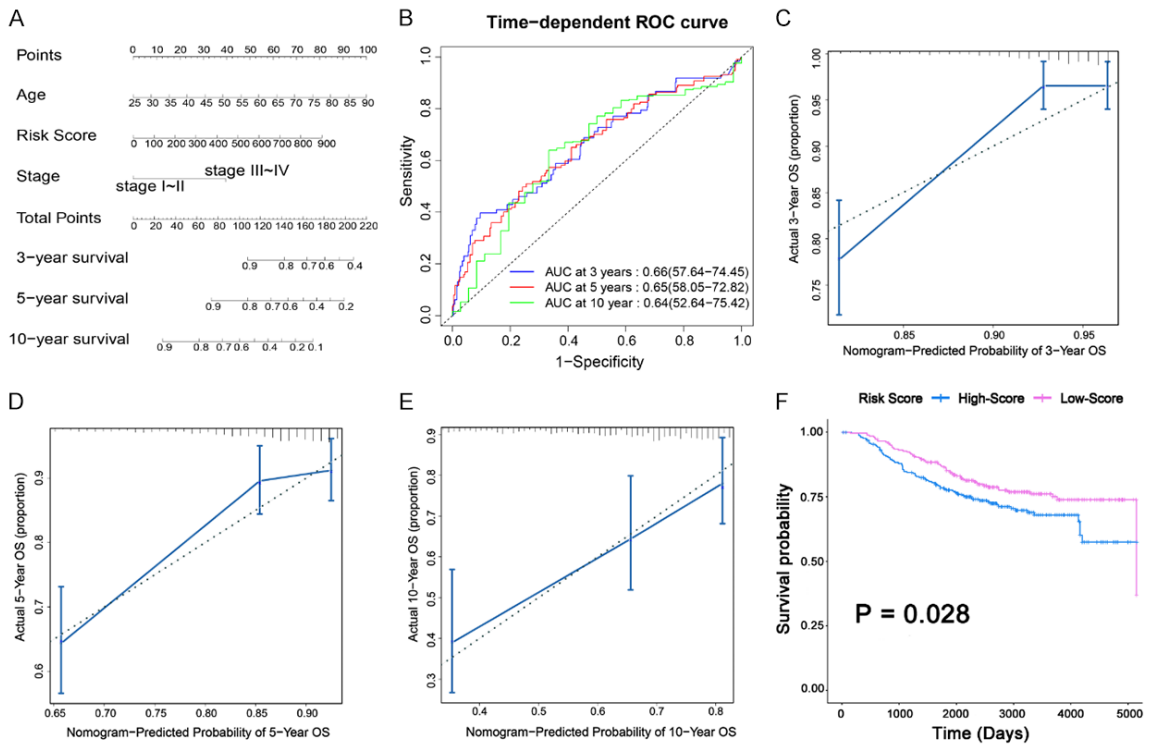

Figure 8. Analysis and validation of the prognostic model. A: The traditional nomogram. B: Time-dependent ROC curves for 3, 5, and 10 years. C-E: Calibration curves of the nomogram for predicting 3-year, 5-year, and 10-year survival rates. F: Validation of the prognostic model based on four genes in the GEO database.

characterization and clinical treatment of BRCA.

With the gene importance ranking that affects the classification provided by XGBoost, we discovered that *PYCARD*, *NLRP2*, *CASP4*, *SCAF11*, *CASP3*, and *CASP6* have essential effects on the model, especially *PYCARD*. *PYCARD*, a protein-coding gene, encodes an adaptor protein composed of two protein-protein interaction domains: *PYD* and *CARD*. This protein is involved in macrophage pyroptosis and is the major constituent of the ASC pyroptosome, which forms upon potassium depletion and rapidly recruits and activates *caspase-1* [28]. The expression of *PYCARD* was increased in Cluster2 compared to Cluster1. The GO annotations associated with this gene included transmembrane transporter binding, consistent with the subsequent GO enrichment analysis of DEGs between Cluster1 and Cluster2. *PYCARD* has not been definitively studied in BRCA, but the protein encoded by *PYCARD* is a crucial component of the inflammasome, which plays a vital role in BRCA progression and metastasis [29]. A previous study demonstrated that the *NLRP3* inflammasome in fibroblasts links tissue damage to inflammation in BRCA progression and metastasis [30].

From the perspective of somatic variants, we found that *TP53* mutations dominate Cluster1, and *PIK3CA* mutations dominate Cluster2. Regarding *TP53*, a recent study found that *TP53*-mutated patients had significantly higher antitumor immune signatures and TMB than *TP53* wild-type patients in BRCA [31]. The results of this study were consistent with the results obtained in our study. In addition, a study demonstrated shorter median overall survival in patients with *TP53* mutations than in patients with wild-type *TP53* (all patients, regardless of treatment) [32]. This conclusion may be one of the reasons why Cluster1 had a worse prognosis than Cluster2. Regarding *PIK3CA*, one study showed a significant improvement in progression-free survival in patients with *PIK3CA* mutant-specific ctDNA treated with alpelisib, a kind of PI3K inhibitor, suggesting that the efficacy of PI3Kα inhibitors is dependent on *PIK3CA* mutant tumors [33]. Another study showed that patients with *TP53*

mutations do not benefit from alpelisib [34]. Based on these two studies, we can speculate that PI3K inhibitors are more effective in Cluster2. Furthermore, *PIK3CA* mutations are known to activate the PI3K/AKT/mTOR pathway. Cluster2 was more sensitive to temsirolimus, an mTOR inhibitor, indicating that the PI3K/AKT/mTOR pathway is more activated in Cluster2 than in Cluster1. A recent study showed that *mTORC1* could activate pyroptosis, so we speculate that *PIK3CA* activates the PI3K/AKT/mTOR pathway more in Cluster2 and synthesizes more *mTORC1*, and *mTORC1* activates pyroptosis to cause *PYCARD* to be highly expressed in Cluster2 [35].

In the following analysis, we found that the TMB of Cluster1 was higher than that of Cluster2, and the results of the six immune algorithms showed that the degree of immune infiltration of Cluster1 was higher than that of Cluster2. These two markers are considered to assess the outcome of immunotherapy, so it is reasonable to assume that immune checkpoint inhibitors would be more effective in Cluster1 than in Cluster2.

In the methylation analysis, the expression of the 15 N^6-methylation-regulated genes we selected was higher in Cluster1 than in Cluster2. This result may be one of the reasons why Cluster1 had a worse prognosis than Cluster2 because several studies have demonstrated that *YTHDF1*, *YTHDF3*, *METTL14*, and *FTO* promote progression and are poor prognostic factors in BRCA [36-39]. At present, the relationship between m^6A and pyroptosis is still inconclusive. Some studies have concluded that *METTL14* and *METTL3* inhibit pyroptosis [40-43]. In our results, Cluster1 had a higher expression of 15 m^6A methylation-regulated genes, including *METTL14* and *METTL3*. This consequence may cause the expression of *PYCARD* in Cluster1 to be lower than that in Cluster2.

In summary, the Cluster1 and Cluster2 subtypes of BRCA were identified in the TCGA database based on 21 PRGs, and a predictive model was built using a supervised machine learning approach. We then confirmed the differences in gene mutation, immune infiltration, methylation, and drug sensitivity between the two subtypes. These analyses shed new light on understanding the underlying molecular features of BRCA and may offer different perspectives on personalizing treatment for patients. In addition, our prediction of drug efficacy for these two pyroptosis-based subtypes may provide new ideas for clinical research on some drugs. Of course, this study also has certain limitations. Our predictions were based only on the TCGA and GEO databases, and more data are needed to support our conclusions. The XGBoost model needs further optimization to facilitate clinical use. In addition, the speculation that *PIK3CA* mutations lead to high expression of *PYCARD* requires experimental validation, while the prediction of partial drug efficacy requires real-world studies.

Acknowledgements

We thank the TCGA and GEO databases for providing the curated data and the contributors who uploaded the precious data.

Disclosure of conflict of interest

None.

Abbreviations

BRCA, breast cancer; DEGs, differentially expressed genes; PRGs, pyroptosis-related genes; NMF, nonnegative matrix factorization; PPI, protein-protein interaction; TMB, tumor mutation burden; IC50, 50% inhibiting concentration; t-SNE, t-distributed stochastic neighbor embedding; LR, logistic regression; RF, random forest; SVM, support vector machine; NNET, neural network; 95% CI, 95% confidence interval; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; ROC, Receiver Operating Characteristic.

Address correspondence to: Yan-Wei Guo, Department of Oncology, The Fifth Affiliated Hospital of Zhengzhou University, Zhengzhou, China. E-mail: YanweiGuo1276@163.com

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018; 68: 394-424.

[2] Siegel RL, Miller KD, Fuchs HE and Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022; 72: 7-33.

[3] Waks AG and Winer EP. Breast cancer treatment: a review. JAMA 2019; 321: 288-300.
[4] Loibl S, Poortmans P, Morrow M, Denkert C and Curigliano G. Breast cancer. Lancet 2021; 397: 1750-1769.
[5] Shao F and Fitzgerald KA. Molecular mechanisms and functions of pyroptosis. J Mol Biol 2022; 434: 167461.
[6] Wu M, Wang Y, Yang D, Gong Y, Rao F, Liu R, Danna Y, Li J, Fan J, Chen J, Zhang W and Zhan Q. A PLK1 kinase inhibitor enhances the chemosensitivity of cisplatin by inducing pyroptosis in oesophageal squamous cell carcinoma. EBioMedicine 2019; 41: 244-255.
[7] Kayagaki N, Stowe IB, Lee BL, O'Rourke K, Anderson K, Warming S, Cuellar T, Haley B, Roose-Girma M, Phung QT, Liu PS, Lill JR, Li H, Wu J, Kummerfeld S, Zhang J, Lee WP, Snipas SJ, Salvesen GS, Morris LX, Fitzgerald L, Zhang Y, Bertram EM, Goodnow CC and Dixit VM. Caspase-11 cleaves gasdermin D for non-canonical inflammasome signalling. Nature 2015; 526: 666-671.
[8] Pizato N, Luzete BC, Kiffer L, Correa LH, de Oliveira Santos I, Assumpcao JAF, Ito MK and Magalhaes KG. Omega-3 docosahexaenoic acid induces pyroptosis cell death in triple-negative breast cancer cells. Sci Rep 2018; 8: 1952.
[9] Zhang CC, Li CG, Wang YF, Xu LH, He XH, Zeng QZ, Zeng CY, Mai FY, Hu B and Ouyang DY. Chemotherapeutic paclitaxel and cisplatin differentially induce pyroptosis in A549 lung cancer cells via caspase-3/GSDME activation. Apoptosis 2019; 24: 312-325.
[10] Wang Y, Gao W, Shi X, Ding J, Liu W, He H, Wang K and Shao F. Chemotherapy drugs induce pyroptosis through caspase-3 cleavage of a gasdermin. Nature 2017; 547: 99-103.
[11] Blasco MT and Gomis RR. PD-L1 controls cancer pyroptosis. Nat Cell Biol 2020; 22: 1157-1159.
[12] Kao KJ, Chang KM, Hsu HC and Huang AT. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer 2011; 11: 143.
[13] Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance M, Michiels S, Volkmar M, Deplus R, Luciani J, Lallemand F, Larsimont D, Toussaint J, Haussy S, Rothe F, Rouas G, Metzger O, Majjaj S, Saini K, Putmans P, Hames G, van Baren N, Coulie PG, Piccart M, Sotiriou C and Fuks F. DNA methylation profiling reveals a predominant immune component in breast cancers. EMBO Mol Med 2011; 3: 726-741.
[14] Jezequel P, Loussouarn D, Guerin-Charbonnel C, Campion L, Vanier A, Gouraud W, Lasla H, Guette C, Valo I, Verriele V and Campone M. Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. Breast Cancer Res 2015; 17: 43.
[15] Song W, Ren J, Xiang R, Kong C and Fu T. Identification of pyroptosis-related subtypes, the development of a prognosis model, and characterization of tumor microenvironment infiltration in colorectal cancer. Oncoimmunology 2021; 10: 1987636.
[16] Zhang Z, Zhang Y, Xia S, Kong Q, Li S, Liu X, Junqueira C, Meza-Sosa KF, Mok TMY, Ansara J, Sengupta S, Yao Y, Wu H and Lieberman J. Gasdermin E suppresses tumour growth by activating anti-tumour immunity. Nature 2020; 579: 415-420.
[17] Rogers C, Fernandes-Alnemri T, Mayes L, Alnemri D, Cingolani G and Alnemri ES. Cleavage of DFNA5 by caspase-3 during apoptosis mediates progression to secondary necrotic/pyroptotic cell death. Nat Commun 2017; 8: 14128.
[18] Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014; 15: 550.
[19] Devarajan K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput Biol 2008; 4: e1000029.
[20] Mayakonda A, Lin DC, Assenov Y, Plass C and Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res 2018; 28: 1747-1756.
[21] Chan TA, Yarchoan M, Jaffee E, Swanton C, Quezada SA, Stenzinger A and Peters S. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. Ann Oncol 2019; 30: 44-56.
[22] Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, Li B and Liu XS. TIMER2.0 for analysis of tumor-infiltrating immune cells. Nucleic Acids Res 2020; 48: W509-W514.
[23] Lei X, Lei Y, Li JK, Du WX, Li RG, Yang J, Li J, Li F and Tan HB. Immune cells within the tumor microenvironment: biological functions and roles in cancer immunotherapy. Cancer Lett 2020; 470: 126-133.
[24] Geeleher P, Cox N and Huang RS. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. PLoS One 2014; 9: e107468.
[25] He L, Li H, Wu A, Peng Y, Shu G and Yin G. Functions of N6-methyladenosine and its role in cancer. Mol Cancer 2019; 18: 176.
[26] Yu G, Wang LG, Han Y and He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012; 16: 284-287.

[27] Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010; 33: 1-22.
[28] Wang Y, Ning X, Gao P, Wu S, Sha M, Lv M, Zhou X, Gao J, Fang R, Meng G, Su X and Jiang Z. Inflammasome activation triggers caspase-1-mediated cleavage of cGAS to regulate responses to DNA virus infection. Immunity 2017; 46: 393-404.
[29] Karki R, Man SM and Kanneganti TD. Inflammasomes and cancer. Cancer Immunol Res 2017; 5: 94-99.
[30] Ershaid N, Sharon Y, Doron H, Raz Y, Shani O, Cohen N, Monteran L, Leider-Trejo L, Ben-Shmuel A, Yassin M, Gerlic M, Ben-Baruch A, Pasmanik-Chor M, Apte R and Erez N. NLRP3 inflammasome in fibroblasts links tissue damage with inflammation in breast cancer progression and metastasis. Nat Commun 2019; 10: 4375.
[31] Li L, Li M and Wang X. Cancer type-dependent correlations between TP53 mutations and antitumor immunity. DNA Repair (Amst) 2020; 88: 102785.
[32] Shahbandi A, Nguyen HD and Jackson JG. TP53 mutations and outcomes in breast cancer: reading beyond the headlines. Trends Cancer 2020; 6: 98-110.
[33] Mishra R, Patel H, Alanazi S, Kilroy MK and Garrett JT. PI3K inhibitors in cancer: clinical implications and adverse effects. Int J Mol Sci 2021; 22: 3464.
[34] Mayer IA, Abramson VG, Formisano L, Balko JM, Estrada MV, Sanders ME, Juric D, Solit D, Berger MF, Won HH, Li Y, Cantley LC, Winer E and Arteaga CL. A phase Ib study of alpelisib (BYL719), a PI3Kalpha-specific inhibitor, with letrozole in ER+/HER2- metastatic breast cancer. Clin Cancer Res 2017; 23: 26-34.
[35] Evavold CL, Hafner-Bratkovic I, Devant P, D'Andrea JM, Ngwa EM, Borsic E, Doench JG, LaFleur MW, Sharpe AH, Thiagarajah JR and Kagan JC. Control of gasdermin D oligomerization and pyroptosis by the Ragulator-Rag-mTORC1 pathway. Cell 2021; 184: 4495-4511, e19.
[36] Chen H, Yu Y, Yang M, Huang H, Ma S, Hu J, Xi Z, Guo H, Yao G, Yang L, Huang X, Zhang F, Tan G, Wu H, Zheng W and Li L. YTHDF1 promotes breast cancer progression by facilitating FOXM1 translation in an m6A-dependent manner. Cell Biosci 2022; 12: 19.
[37] Anita R, Paramasivam A, Priyadharsini JV and Chitra S. The m6A readers YTHDF1 and YTHDF3 aberrations associated with metastasis and predict poor prognosis in breast cancer patients. Am J Cancer Res 2020; 10: 2546-2554.
[38] Yi D, Wang R, Shi X, Xu L, Yilihamu Y and Sang J. METTL14 promotes the migration and invasion of breast cancer cells by modulating N6-methyladenosine and hsa-miR-146a-5p expression. Oncol Rep 2020; 43: 1375-1386.
[39] Niu Y, Lin Z, Wan A, Chen H, Liang H, Sun L, Wang Y, Li X, Xiong XF, Wei B, Wu X and Wan G. RNA N6-methyladenosine demethylase FTO promotes breast tumor progression through inhibiting BNIP3. Mol Cancer 2019; 18: 46.
[40] Diao MY, Zhu Y, Yang J, Xi SS, Wen X, Gu Q and Hu W. Hypothermia protects neurons against ischemia/reperfusion-induced pyroptosis via m6A-mediated activation of PTEN and the PI3K/Akt/GSK-3beta signaling pathway. Brain Res Bull 2020; 159: 25-31.
[41] Wang X, Li Y, Li J, Li S and Wang F. Mechanism of METTL3-mediated m(6)A modification in cardiomyocyte pyroptosis and myocardial ischemia-reperfusion injury. Cardiovasc Drugs Ther 2022; [Epub ahead of print].
[42] Meng L, Lin H, Huang X, Weng J, Peng F and Wu S. METTL14 suppresses pyroptosis and diabetic cardiomyopathy by downregulating TINCR lncRNA. Cell Death Dis 2022; 13: 38.
[43] Liu BH, Tu Y, Ni GX, Yan J, Yue L, Li ZL, Wu JJ, Cao YT, Wan ZY, Sun W and Wan YG. Total flavones of abelmoschus manihot ameliorates podocyte pyroptosis and injury in high glucose conditions by targeting METTL3-dependent m(6)A modification-mediated NLRP3-inflammasome activation and PTEN/PI3K/Akt signaling. Front Pharmacol 2021; 12: 667644.