Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/35/20240909

Establishment and validation of a prognostic model for major histocompatibility complex (MHC)-related genes in breast cancer

Shilong Yu[1], Zengjian Tian[1] and Qilun Liu[2, 3]

[1]The First School of Clinical Medicine, Ningxia Medical University, Yinchuan, Ningxia, China
[2]General Hospital of Ningxia Medical University, Yinchuan, Ningxia, China

[3]liuql6311@hotmail.com

Abstract. The major histocompatibility complex (MHC) is a group of genes involved in the immune system. In order to investigate this phenomenon, relevant sample data from human breast cancer can be downloaded from databases such as TCGA and GEO. Differential analysis of MHC-related genes that are differentially expressed (MHCRDEGs) can then be performed using single-factor Cox analysis. The identified characteristic genes can be subjected to differential analysis and protein interaction network analysis using multiple datasets. This analysis can aid in the selection of prognostic genes and the establishment of a clinically relevant MHCRDEG model, which can then be validated using multiple datasets. Through machine learning methods, six characteristic genes (LIFR, UGP2, F2RL2, SLC7A5, TUBA1C, IL12B) can be screened, and a diagnostic risk model can be developed. Finally, by comparing the results obtained from multiple datasets, four characteristic genes (LIFR, SLC7A5, TUBA1C, UGP2) can be identified. A clinical prognostic risk model can be established based on these genes, and its validity and accuracy can be confirmed using multiple datasets. This comprehensive study provides valuable insights into the underlying mechanisms of MHC-related genes in cancer.

Keywords: Breast cancer, Major Histocompatibility Complex,prognostic model,genes

1. Introduction

Breast cancer is a prevailing malignant neoplasm, and its global occurrence is increasing steadily. The etiology of this disease is intricate, encompassing various elements, such as genetic factors, environmental influences, lifestyle choices, and more [1]. In recent years, it has been discovered by scientists that there is a connection between the occurrence of breast cancer and genes associated with the major histocompatibility complex (MHC) [2]. The MHC-related genes encompass the coding of MHC molecules, which hold significant significance in the immune system as they contribute to the processes involving antigen recognition and immune response [2]. Several studies have established a significant association between MHC-related genes and the overall functionality and stability of the immune system. The various MHC genotypes can influence an individual's capacity to identify and eliminate tumor antigens, thus impacting the progression of breast cancer. Moreover, research has also suggested a correlation between MHC-related genes and the clinical attributes and prognosis of

DOI: 10.54254/2753-8818/35/20240909

individuals with breast cancer. Certain MHC genotypes have been found to be closely associated with important clinical characteristics such as stage, degree of differentiation, and metastasis propensity in breast cancer cases [3-6]. Some MHC genotypes may affect the immune evasion ability of tumor cells, thereby affecting patient prognosis and treatment effects. This study uses machine learning to screen MHC breast cancer differential genes and establishes clinically relevant models based on this, including differential gene diagnostic models and prognostic models.The role of MHC-related genes in breast cancer is further explored through mutation analysis, immune infiltration analysis, and immune checkpoint analysis. At the same time, drug sensitivity analysis provides new directions for future drug research. By focusing on MHC-related genes, this study aims to elucidate their mechanisms in breast cancer patients, ultimately providing a theoretical foundation for precision treatment of this disease.

2. Materials and Methods

2.1. Data Download

We used The TCGA biolinks package from the cancer genome Project (TCGA, The cancer genome atlas, (https://portal.gdc.cancer.gov/) to download Breast cancer Breast invasive carcinoma, BRCA) data set (TCGA - BRCA) and serves as a test set were analyzed, and the sample data in 1186 cases of Breast cancer (BRCA), A total of 113 adjacent normal samples (group: Control) and 1073 breast cancer samples (group: BRCA) were included, and the corresponding clinical data were obtained from the UCSC Xena database(http://genome.ucsc.edu).Tumor mutation burden (TMB) data and Microsatellite Instability (Microsatellite Instability, MSI) data were downloaded (cBioPortal for Cancer Genomics) (https:// www.cbioportal.org/).

We from the TCGA website (https://portal.gdc.cancer.gov/) to choose “Masked Somatic Mutation” data as the patients with breast cancer (TCGA - BRCA) single nucleotide polymorphisms (SNPS, data Single Nucleotide Polymorphism (SNP) data were preprocessed by VarScan software and visualized by maftools R package .To analyze the Copy Number Variations (CNV) in TCGA-BRCA patients, we used R’s TCGAbiolinks package to download the patient’s “Copy Number Segment” data. Then we performed GISTIC 2.0 analysis on the downloaded and processed CNV segments, and all the default parameters were used in the process of GISTIC 2.0 analysis.

In addition, we also downloaded the expression profile datasets GSE42568, GSE36295 of Breast invasive carcinoma (BRCA)using the R package GEOquery. he datasets GSE42568 and GSE36295 are both from Homo sapiens. The GSE42568 dataset includes 104 Breast invasive carcinoma (group: BRCA) samples and 17 normal breast tissue samples (group: BRCA). Control), the data platform was GPL570[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array; The samples of GSE36295 dataset were 45 Breast invasive carcinoma (BRCA) and 5 normal breast tissue samples (group: BRCA). GPL6244[HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array [transcript (gene) version] was used as the data platform. All the samples in datasets GSE42568 and GSE36295 were included in the subsequent analysis, and the probe annotation was performed using platform files.

In addition we we from GeneCards MHC related genes, MHCRGs). GeneCards database provides comprehensive information on human genes. We used “MHC” as the search keyword and only retained protein coding, and a total of 4489 MHC-related genes (MHCRGs) were obtained.

We have to “the Human Leukocyte Antigen” as the search key words from GeneCards database (https://www.genecards.org/) to search for and retain only begin with “HLA” protein-coding genes, A total of 22 HLA family genes were obtained. In addition, we searched 47 Immune checkpoint genes from the published literature .

TIDE (Tumor Immune Dysfunction and Exclusion) is a computational approach that simulates the two main mechanisms of tumor immune escape -- T-cell dysfunction and T-cell rejection. It can be used to predict the response of cancer patients to Immune checkpoint inhibitors (ICIs). TIDE scores for the TCGA-BRCA data set were obtained from the TIDE website (http://tide.dfci.harva rd.edu) after uploading gene-expression data, as described on the website.

DOI: 10.54254/2753-8818/35/20240909

2.2. Screen for MHC-related differentially expressed genes

We first performed data set correction on the breast cancer (BRCA) datasets TCGA-BRCA, GSE42568 and GSE36295. We used the to normalize TCGA and the three GEO datasets. Boxplot plot of the expression matrix of the dataset before and after removing batch effect was drawn for comparison and display.

Then we use difference analysis was carried out on the TCGA limma package, with | logFC | > 1 and p.a DJ < 0.05 for each screen with standard data sets of differentially expressed genes (differentially expressed genes, DEGs), We will TCGA DEGs with MHCRGs intersection, obtain differentially expressed genes linked to the MHC (MHC related differentially expressed genes, MHCRDEGs). Finally, we also used univariate Cox analysis to screen the MHCRDEGs and retained only the MHCRDEGs with p<0.05.

2.3. Functional enrichment analysis (GO) and pathway enrichment analysis (KEGG)

Gene Ontology (GO)analysis is a common method for large-scale functional enrichment studies, including biological process (BP), molecular function (KEGG), molecular function (KEGG). MF) and cellular component (CC). Kyoto Encyclopedia of Genes and Genomes (]We used R package clusterProfiler to perform GO functional enrichment analysis and KEGG pathway enrichment analysis of MHCRDEGs. The entry screening criterion was pvalue< 0.05, and the P value correction method was Benjamini-Hochberg (BH).

2.4. Construct MHCRDEGs diagnostic model

In order to obtain the Logistic diagnostic model of TCGA-BRCA dataset, we first performed Logistic regression analysis on MHCRDEGs. When the dependent variable was a binary variable, that is, BRCA group and Control group, Logistic regression was used to analyze the association between independent variables and dependent variables. pvalue < 0.05 was used as the standard to screen MHCRDEGs and construct a Logistic diagnostic model. The molecular expression of MHCRDEGs included in the Logistic regression model was displayed by Forest Plot. The prognostic KM curve of MHCRDEGs obtained by Logistic regression analysis was also drawn, and only the MHCRDEGs with a prognostic KM curve p<0.05 were retained.

Then, based on the expression matrix and grouping information (BRCA/Control) of the TCGA-BRCA dataset, we used SVM(Support Vector Machine) algorithm to construct the SVM model, and screened MHCRDEGs based on the number of genes with the highest accuracy and the lowest error rate.

RandomForest (RF) is an algorithm that integrates multiple decision trees through the idea of ensemble learning. It belongs to the bagging (bootstrap aggregation) ensemble algorithm of ensemble algorithm, which is composed of multiple algorithms. Random forest is also a commonly used model building method. By constructing multiple decision trees, when a sample needs to be predicted, the prediction results of each tree in the forest for the sample are counted, and then the final result is selected from these prediction results by voting. The expression level of MHCRDEGs in the expression matrix of the TCGA-BRCA dataset was used to construct the model using the randomForest package, with the parameters of set.seed (234) and ntree = 1000.

$$I(X = xi) = -log_2 p(x_i)$$

We also performed LASSO (Least Absolute Shrinkage and Selection) for MHCRDEGs through the R package glmnet Operator regression analysis was used to obtain the Logistic-LASSO regression model (seed number 2022). LASSO regression analysis is based on linear regression analysis, by adding a penalty term (lambda × absolute value of slope) to reduce the overfitting of the model and improve the generalization ability of the model. The diagnostic model plot and variable trajectory plot were used to visualize the results of LASSO regression analysis.

$$\text{risk}Score = \sum_i Coefficient\ (hub\ gene_i) * mRNA\ Expression\ (hub\ gene_i)$$

DOI: 10.54254/2753-8818/35/20240909

Then, we interposed the MHCRDEGs included in the Logistic-LASSO regression model with the SVM model and the MHCRDEGs included in the random forest model to draw a Venn diagram. Common MHCRDEGs (Common MHC related differentially expressed genes, Common MHCRDEGs) were obtained. Then we combined the coefficients of Common MHCRDEGs in the Logistic-LASSO regression model and the expression levels of the TCGA-BRCA dataset to obtain the MHCRDEGs diagnostic model and the corresponding Riskscores.

A Nomogram is a graph that uses a cluster of disjoint line segments to represent the functional relationship between multiple independent variables in a rectangular plane coordinate system. The R package rms was used to perform Logistic regression analysis and draw a Nomogram to show Common based on the expression level of genes in the MHCRDEGs diagnostic model obtained by Logistic LASSO regression analysis in the TCGA-BRCA data set The analysis results of MHCRDEGs. Decision Curve Analysis (is a simple method to evaluate clinical prediction models, diagnostic tests and molecular markers were evaluated by drawing DCA diagram using R package ggDCA.

2.5. GSEA enrichment analysis and GSVA analysis

After retaining only cancer samples, we divided the TCGA-BRCA dataset samples into High risk group and Low risk group (High/Low Riskscore group) according to the median Riskscore of MHCRDEGs diagnostic model.

2.6. Immune infiltration analysis between high and low risk groups in TCGA-BRCA dataset

We used the single-sample gene-set enrichment analysis (ssGSEA) algorithm to analyze the tumor microenvironment. Different human immune cell phenotypes in TME were distinguished with high sensitivity and specificity. The algorithm obtained 28 gene sets for labeling different tumor infiltrating immune cell types from published tumor immune infiltration articles CD8+ T cells, dendritic cells, macrophages, regulatory T cells and so on. We calculated the enrichment scores by the ssGSEA algorithm in the GSVA package of R package and used the enrichment scores to represent the infiltration levels of different types of immune cells in each sample. boxplot plot was used to show the differences in the infiltration abundance of immune cells between the high and low risk groups of MHCRDEGs diagnostic model in the TCGA-BRCA dataset. The correlation between immune cells was calculated by spearman's statistical algorithm and visualized by the R package ggplot2. The correlation between immune cells and Common MHCRDEGs was calculated by spearman statistical algorithm and the R package ggplot2 was used to draw correlation dot plot for display.

CIBERSORT is an immune infiltration analysis algorithm that deconvolutes the transcriptome expression matrix based on the principle of linear support vector regression to estimate the composition and abundance of immune cells in mixed cells. We upload the matrix data of the TCGA-BRCA dataset to CIBERSORT, combine with the LM22 feature gene matrix to screen out the data with immune cell enrichment score greater than zero, and finally obtain and display the specific results of immune cell infiltration abundance matrix.

2.7. Immune score of TCGA-BRCA dataset

We evaluated the immune activity of tumors based on the expression profile data of the TCGA-BRCA dataset through the ESTIMATE package. We quantitatively analyzed the Immune activity (immune infiltration level) in BRCA tumor samples based on gene expression profiling, and obtained three scores of BRCA samples in the TCGA-BRCA dataset, namely Stromal Score, immune Score, and ESTIMATE Score. Combined with the Riskscores high and low grouping information of MHCRDEGs diagnostic model in TCGA-BRCA data set, we presented the Stromal Score, Immune Score and ESTIMATE Score results by group comparison plot. In addition, we also plotted the scatter plot of the correlation between Stromal Score, Immune Score, ESTIMATE Score and Riskscores of MHCRDEGs diagnostic model.

DOI: 10.54254/2753-8818/35/20240909

2.8. Drug sensitivity analysis

Alterations in the cancer genome strongly influence the clinical response to therapy and are in many cases effective biomarkers of response to drug therapy. CTRP database for Cancer Therapeutics Response Portal (https://portals.broadinstitute.org/ctrp/), the database will be in the department of Cancer genetics, lineage, and other characteristics linked with small molecule sensitivity, The goal is to accelerate the discovery of cancer therapeutic molecules (drugs) that are matched to patients. The data include 860 cancer cell lines of mutation, gene expression, copy number variation and other omics data, as well as 481 small molecule sensitivity data. The Genomics of Drug Sensitivity in Cancer (GDSC) database (www.cancerRxgene.org) is the largest public resource for information on molecular markers of drug sensitivity and drug response in cancer cells. The Cancer Drug Sensitivity Genomics database can be used to find cancer drug response data and genomic sensitive markers. We performed drug sensitivity analysis of key genes based on the expression levels of Common MHCRDEGs in the TCGA-BRCA dataset and the drug data in the CTRP and GDSC databases, and the results were presented.

2.9. Group comparison diagram of Common MHCRDEGs among cancer control groups in the three datasets

Next, we used Mann-Whitney U test (Wilcoxon rank sum test) to analyze the expression differences of Common MHCRDEGs among cancer control groups in TCGA-BRCA, GSE42568, and GSE36295 and presented them by group comparison plots. The expression trend of Common MHCRDEGs in the three datasets was discussed by group comparison plot.

2.10. Construct PPI network and mRNA-TF interaction network

Protein-protein interaction network is composed of individual proteins that interact with each other to participate in various aspects of life processes such as biological signal transduction, gene expression regulation, energy and substance metabolism, and cell cycle regulation. Systematic analysis of the interactions of a large number of proteins in biological systems is of great significance for understanding the working principle of proteins in biological systems, understanding the reaction mechanism of biological signals and energy and substance metabolism under special physiological conditions such as diseases, and understanding the functional relationship between proteins. The STRING database is a database for searching the interactions between known and predicted proteins. In this study, we used the STRING database to construct protein-protein interaction networks related to Common MHCRDEGs (minimum required interaction score: low confidence (0.150)) and visualized using Cytoscape (version 3.9.1).

CHIPBase database (version 3.0) (https://rna.sysu.edu.cn/chipbase/) from the DNA binding protein ChIP - seq data identified thousands of combining base sequence matrix and its binding site, And the transcriptional regulatory relationships between millions of Transcription factors (TFS) and genes are predicted. We searched CHIPBase database (version 3.0) and hTFtarget database for transcription factors (TFS) that bind to Common MHCRDEGs, Number of samples found (upstream)>0 and Number of samples found (downstream)>0 were used as screening criteria to screen the interaction, and Cytoscape was used to visualize the interaction.

2.11. Construct the prognosis model of MHCRDEGs

Then we performed univariate Cox regression analysis on the Common MHCRDEGs using the TCGA-BRCA data set, and the variables meeting the threshold were used for multivariate Cox regression analysis and multivariate Cox prognostic model was constructed using $p<0.1$ as the standard. We based on the results of multivariate Cox regression analysis, the forest map and nomogram . The nomogram was constructed by the R package rms, and the nomogram was a graph that represented the functional relationship between multiple independent variables with a cluster of disjoint line segments in the rectangular coordinate system. On the basis of multivariate regression analysis, a certain scale is set to characterize each variable in the multivariate regression model, and finally the total score is calculated to predict the probability of the occurrence of the event.

DOI: 10.54254/2753-8818/35/20240909

After keeping only the disease samples, we calculated the Riskscore for the GSE42568 dataset (GSE36295) based on the coefficients of the Common MHCRDEGs in the TCGA-BRCA multivariate Cox prognostic model. According to the median Riskscore, cancer samples from GSE42568 (GSE36295) dataset were divided into high and low risk groups. Group comparison plots were also drawn to show the expression differences of Common MHCRDEGs between high and low risk groups of the MHCRDEGs prognostic model in GSE42568 (GSE36295) data set.

2.12. Construct a clinically relevant prognostic model

In order to determine build MHCRDEGs clinically relevant prognostic model, we will Riskscore MHCRDEGs prognosis model combined with BRCA patients' clinical information (Pathologic_T_stage Pathologic_N_stage, Pathologic_M_stage) were used to perform univariate Cox regression analysis, and variables meeting $p < 0.1$ in the univariate analysis results were included in the multivariate Cox regression analysis to construct a clinically relevant prognostic model. We then calculated risk scores for clinically relevant prognostic models based on the expression levels of the relevant variables in the TCGA-BRCA data set.

$$\text{Risk } Scores = \sum_{i} Coefficient\ (hub\ gene_i) * mRNA\ Expression\ (hub\ gene_i)$$

We then used forest plots to present the results of univariate Cox regression analysis. Then we used the R package rms to construct a nomogram (nomogram) to show the results of multivariate Cox regression analysis. The nomogram is a graph that uses a cluster of disjoint line segments to represent the functional relationship between multiple independent variables in a rectangular coordinate system, based on the multivariate regression analysis.Based on the multi-factor regression analysis, a certain scale was set to characterize the various variables in the multi-factor regression model, and the total score was finally calculated to predict the probability of the occurrence of events.

Finally, the calibration curve was used to evaluate the accuracy and resolution of the nomogram. The Calibration plot is used to evaluate the prediction effect of the model on the actual results by drawing the fitting of the actual probability and the predicted probability of the model under different conditions in the figure. It is mainly used for the fitting analysis of the model established by Cox regression method and the actual situation.DCA (Decision curve analysis) is a simple method for evaluating clinical prediction models, diagnostic tests and molecular markers.Finally, DCA chart was used to evaluate the accuracy and discrimination of clinically relevant prognostic models. We used the R package ggDCA to draw DCA diagrams to evaluate the effect of clinically relevant prognostic models.

2.13. Statistical analysis

All data processing and analysis in this article were based on R software (Version 4.1.2). For the comparison of two groups of continuous variables, the statistical significance of normally distributed variables was estimated by independent Student t test. The Mann-Whitney U test (WilCoxon rank sum test) was used to analyze the differences between variables that were not normally distributed. The survival package of R was used for survival analysis, Kaplan-Meier survival curves were used to show survival differences, and the Log-rank test was used to assess the significance of differences in survival time between the two groups. If not specified, the results were calculated by Spearman correlation analysis to calculate the correlation coefficient between different molecules, and $P < 0.05$ was considered statistically significant.

DOI: 10.54254/2753-8818/35/20240909

3. Results

3.1. Technology Rote map

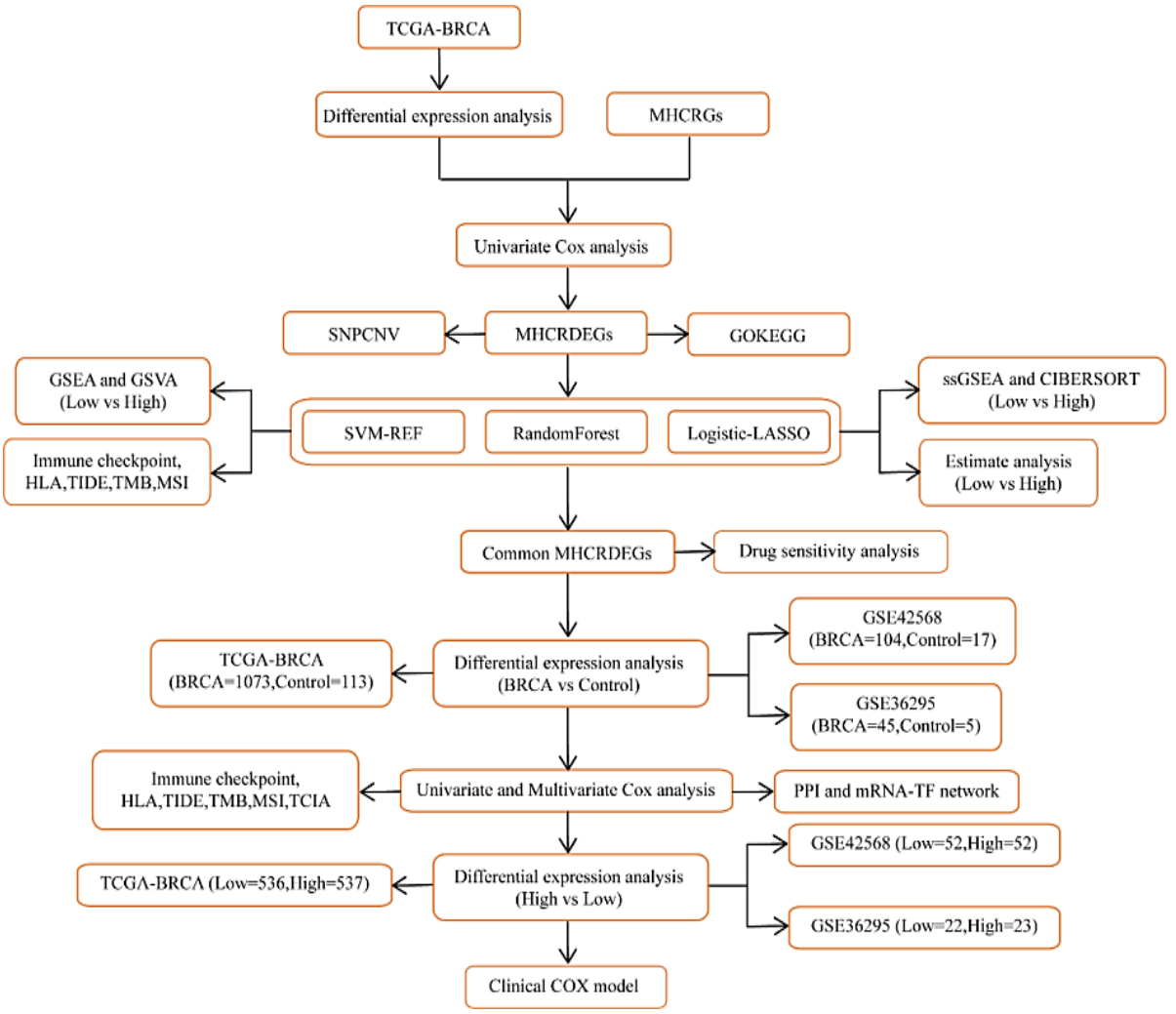

Figure 1. Technology Rote map

DOI: 10.54254/2753-8818/35/20240909

3.2. Data standardization

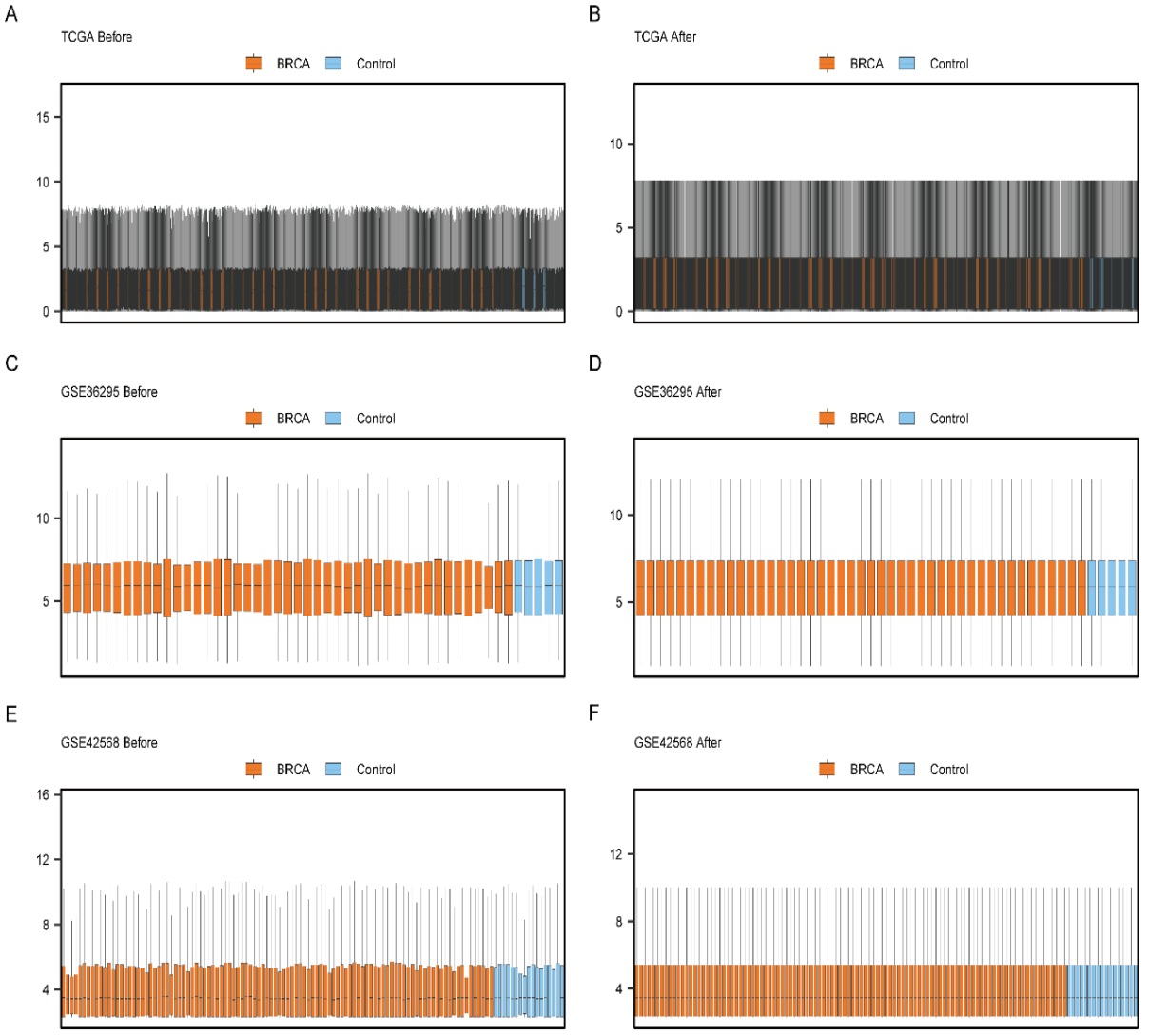

Figure 2. Normalization of data sets

Boxplot plots (A) and (B) display the TCGA-BRCA data before and after normalization. The boxplot plot (C) depicts the GSE42568 data before standardized treatment, and (D) represents the same data after standardization. Additionally, the boxplot plots (E) and (F) demonstrate the GSE36295 data before and after standardized treatment, respectively.

DOI: 10.54254/2753-8818/35/20240909

3.3. *Acquisition of differentially expressed MHC-related genes*

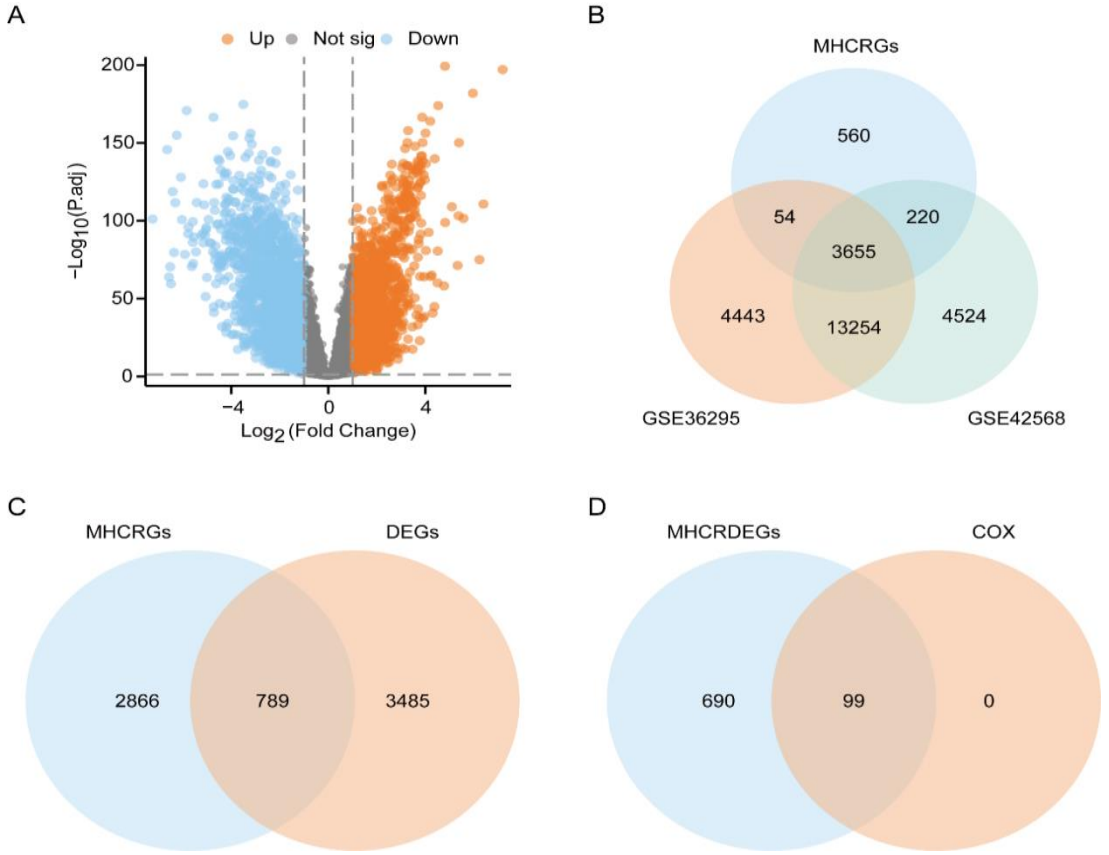

Figure 3. Analysis of Differential Phenotype Genes.

A.Volcano plot illustrating the results of differential analysis conducted between the cancer control groups in the TCGA-BRCA dataset. B. Venn diagram comparing the MHCRGs (MHC related genes) against all genes in the GSE42568 and GSE36295 datasets. C. Venn diagram is illustrated to display the intersection between differentially expressed genes and MHCRGs in the TCGA-BRCA dataset among the cancer controls. D. Results of univariate COX screening of MHCRDEGs (MHC related differentially expressed genes) along with the corresponding Venn diagram.

DOI: 10.54254/2753-8818/35/20240909

3.4. *Mutation analysis of MHCRDEGs in breast cancer patients*

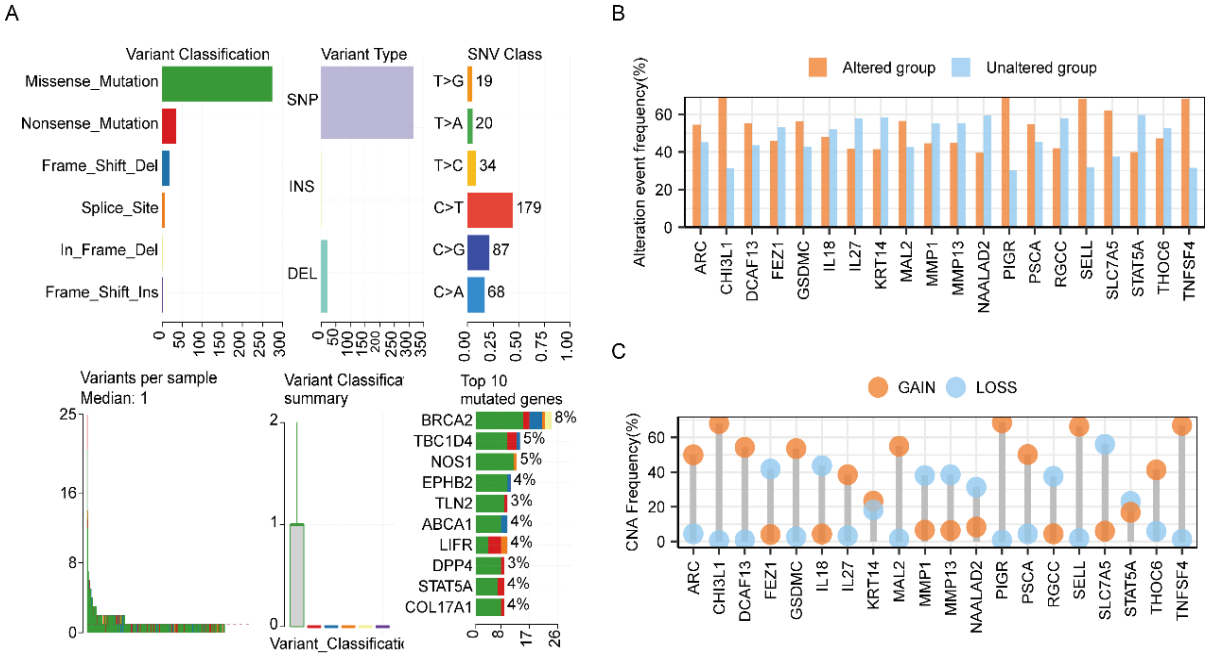

Figure 4. Analysis of Genetic Mutations in Breast Cancer Patients' MHCRDEGs

We present the single nucleotide polymorphisms (SNPs) found in the MHCRDEGs of the TCGA-BRCA dataset. B-C. Displaying the copy number variations (CNVs) of the top 20 MHCRDEGs with altered frequencies in the TCGA-BRCA dataset. MHCRDEGs, genes differentially expressed in MHC-related pathways;

DOI: 10.54254/2753-8818/35/20240909

3.5. *Functional enrichment analysis (GO) and pathway enrichment analysis (KEGG)*

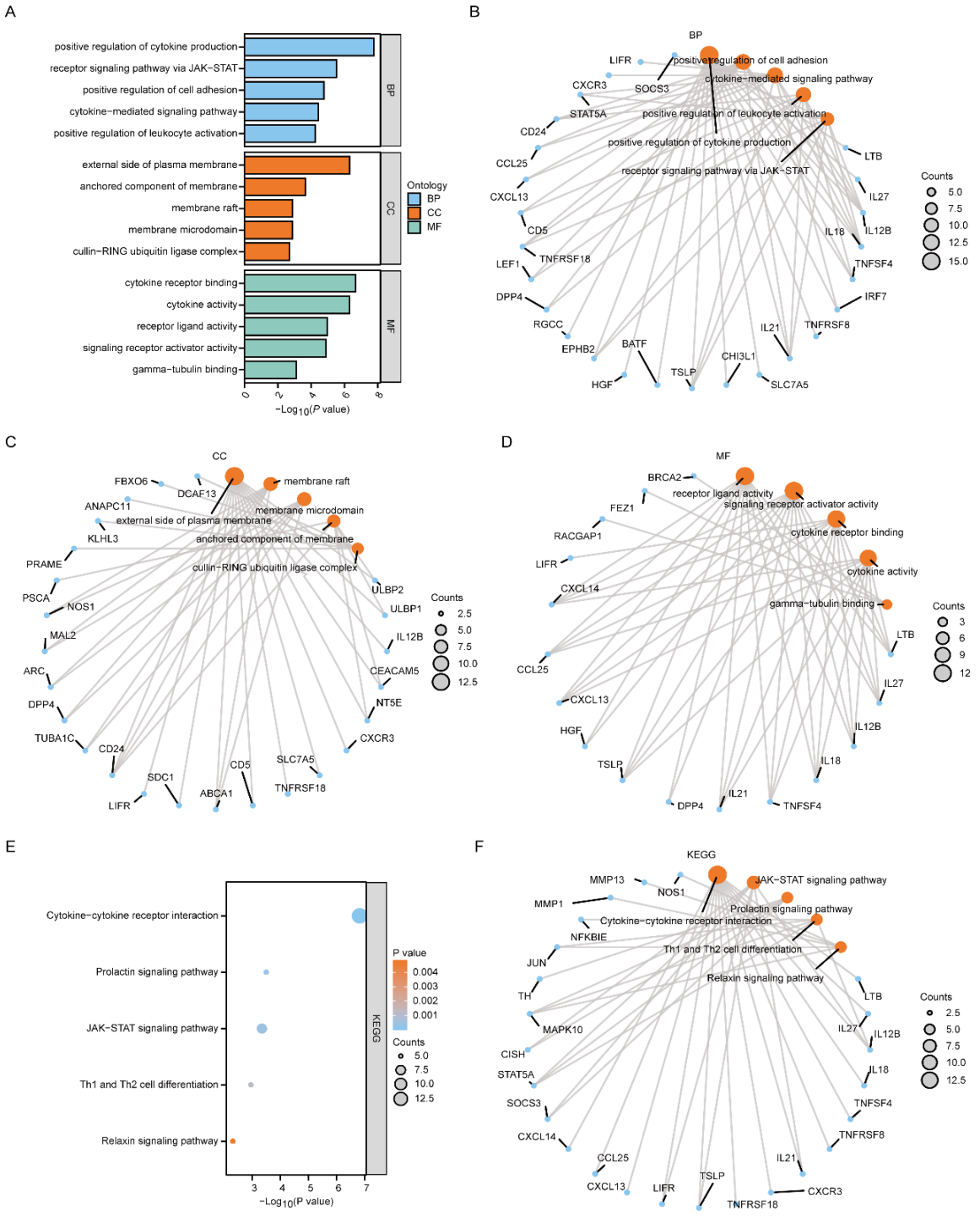

Figure 5. Displays the outcomes of functional enrichment analysis (GO) and pathway enrichment analysis (KEGG) for MHCRDEGs.

A. The presentation of the results obtained from conducting an enriching analysis for MHCRDEGs is done using a bar chart. B-D. The enriched analysis outcomes for the BP, MF and CC pathway. Bubble plot illustrating the findings of KEGG pathway enrichment analysis. F. Ring network diagram displaying the results of KEGG pathway enrichment analysis. In the bar chart (A), the y-axis is representative of

DOI: 10.54254/2753-8818/35/20240909

the GO terms, and the length of the bars represents the p-value associated with each GO term. In the network diagram (B, C, D, F), blue dots correspond to specific genes, while orange dots represent specific pathways. The criterion used for screening GO/KEGG enrichment items was a $p < 0.05$.

3.6. MHCRDEGs diagnostic model

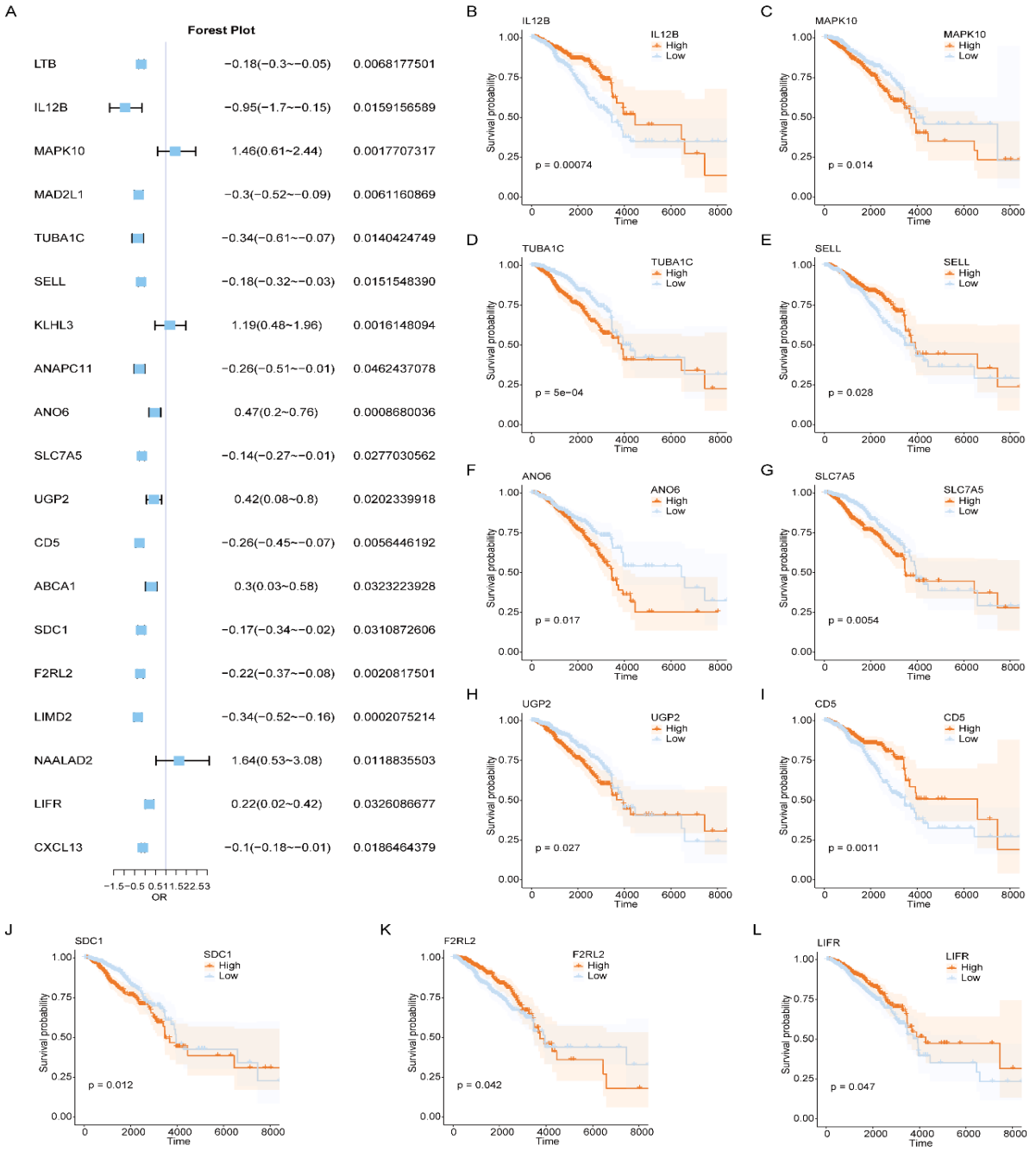

Figure 6. logistic regression and KM curves of MHCRDEGs

A. The results of Logistic regression screening are presented in a forest plot display. B-L. The KM curves for IL12B(B), MAPK10(C), TUBA1C(D), SELL(E), ANO6(F), SLC7A5(G), UGP2(H), CD5(I), SDC1(J), F2RL2(K), and LIFR(L) are shown. MHCRDEGs represent MHC-related differentially expressed genes.

DOI: 10.54254/2753-8818/35/20240909

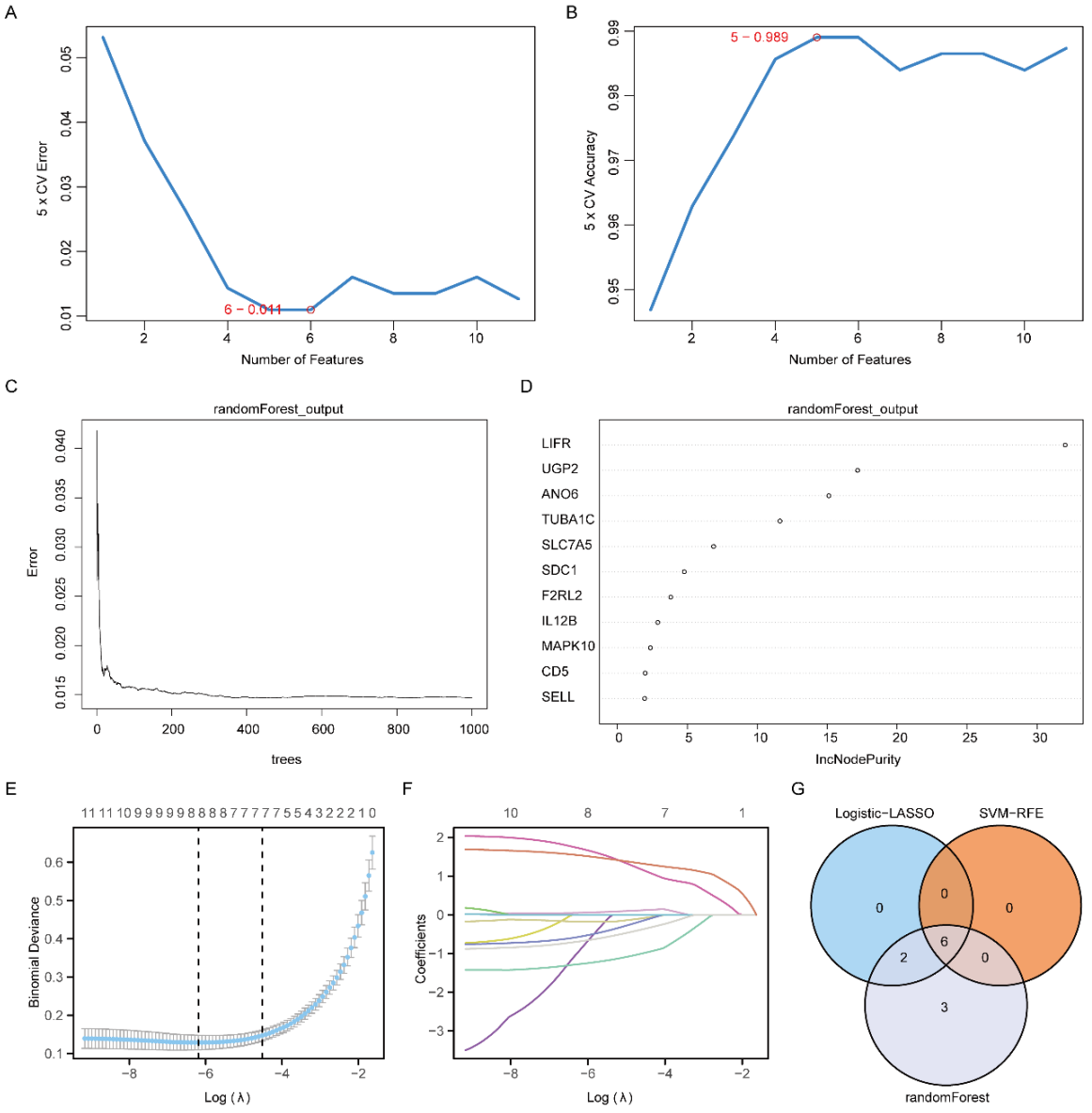

Figure 7. Construction of MHCRDEGs diagnostic model

A. The SVM algorithm yielded the number of genes with the lowest error rate. B. The number of genes exhibiting the highest accuracy was obtained using the SVM algorithm. C. The plot illustrating the training error of the random forest algorithm for model training. D. The random forest model presents MHCRDEGs (arranged in descending order of IncNodePurity). E. The Logistic-LASSO model plot displaying diagnostic outcomes. F. The variable trajectory plot for the Logistic-LASSO model. G. A Venn diagram illustrating the overlap of MHCRDEGs between the Logistic-LASSO model, SVM model, and random forest model.

DOI: 10.54254/2753-8818/35/20240909

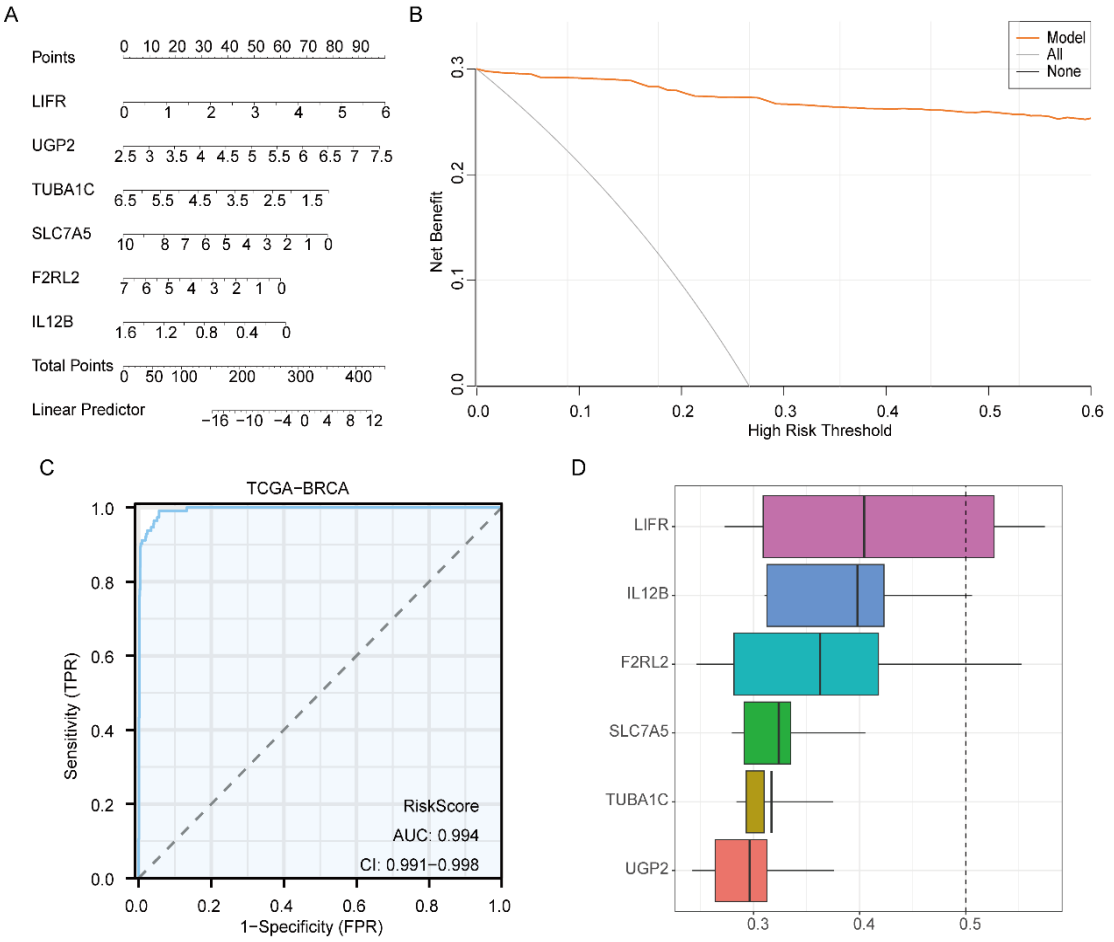

Figure 8. Validation of the MHCRDEGs diagnostic model

A.The nomogram demonstrates the Logistic regression model of MHCRDEGs, displaying the shared MHCRDEGs. B. To assess the effectiveness of the diagnostic model for MHCRDEGs, a decision curve analysis (DCA) is presented. C. The ROC curve of the MHCRDEGs diagnostic model is uncovered using the TCGA-BRCA dataset. D. The common MHCRDEGs' functional similarity analysis reveals the outcomes.

DOI: 10.54254/2753-8818/35/20240909

3.7. The TCGA-BRCA dataset was utilized to perform GSEA and GSVA analyses on the high-risk and low-risk groups of the MHCRDEGs diagnostic model.

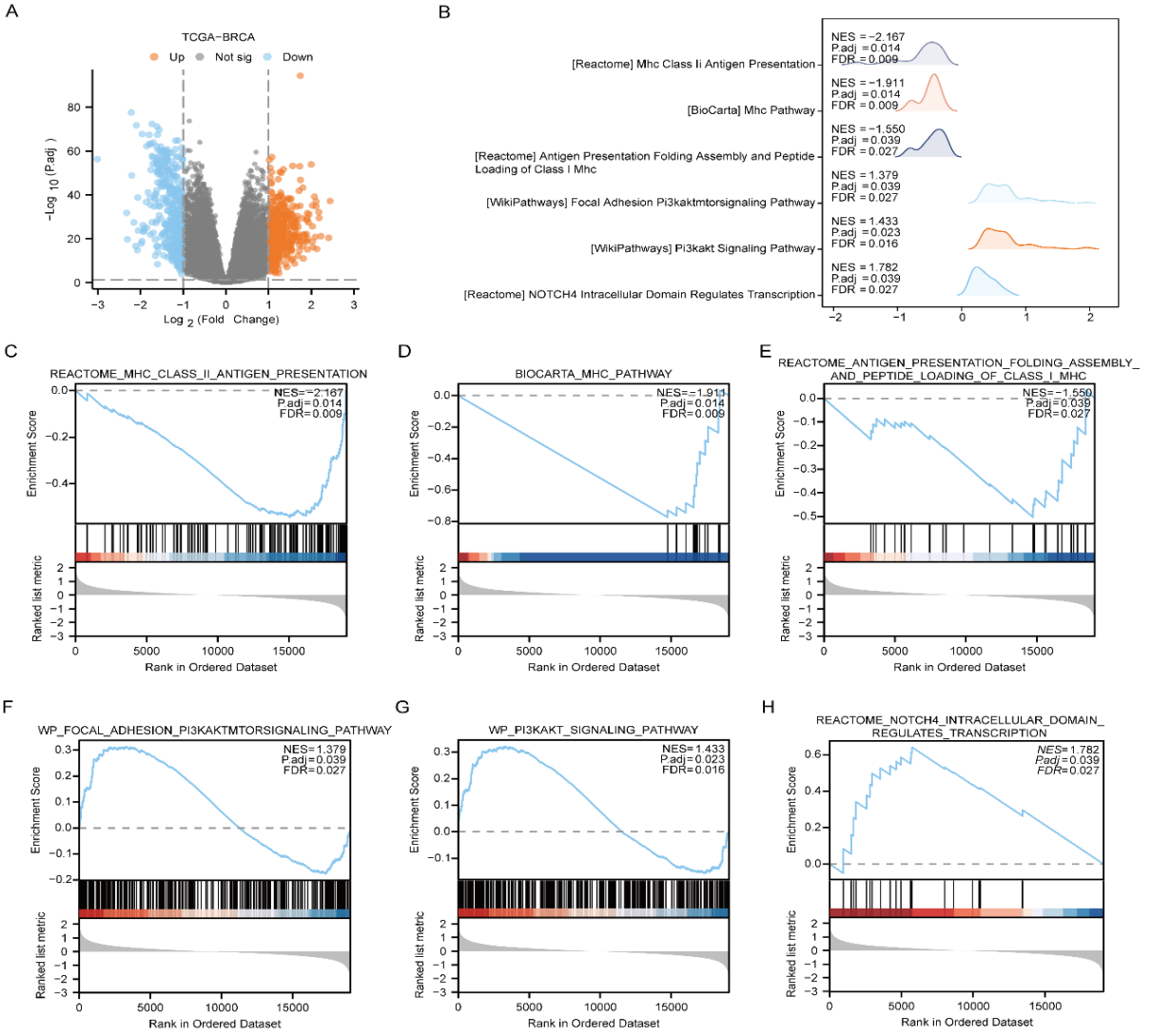

Figure 9. Displays the GSEA results of the MHCRDEGs diagnostic model for the high-risk and low-risk groups in the TCGA-BRCA dataset.

A.Through a volcano plot, the representation of the differential analysis results in the TCGA-BRCA dataset of the MHCRDEGs diagnostic model was observed. B. Utilizing GSEA enrichment analysis, six significant biological characteristics were identified .C-H. The genes found within the TCGA-BRCA dataset displayed noteworthy enrichment in diverse pathways, encompassing various biological processes. REACTOME_MHC_CLASS_II_ANTIGEN_PRESENTATION (C), BIOCARTA_MHC_PATHWAY (D), REACTOME_ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PEPTIDE_LOADING_OF_CLASS_I_MHC (E), WP_FOCAL_ADHESION_PI3KAKTMTOR-SIGNALING_PATHWAY (F), WP_PI3KAKT_SIGNALING_PATHWAY (G), and REACTOME_NOTCH4_INTRACELLULAR_DOMAIN_REGULATES_TRANSCRIPTION (H). The GSEA enrichment analysis using significance filtering criteria of p. Adj <0.05 and q<0.05.

DOI: 10.54254/2753-8818/35/20240909

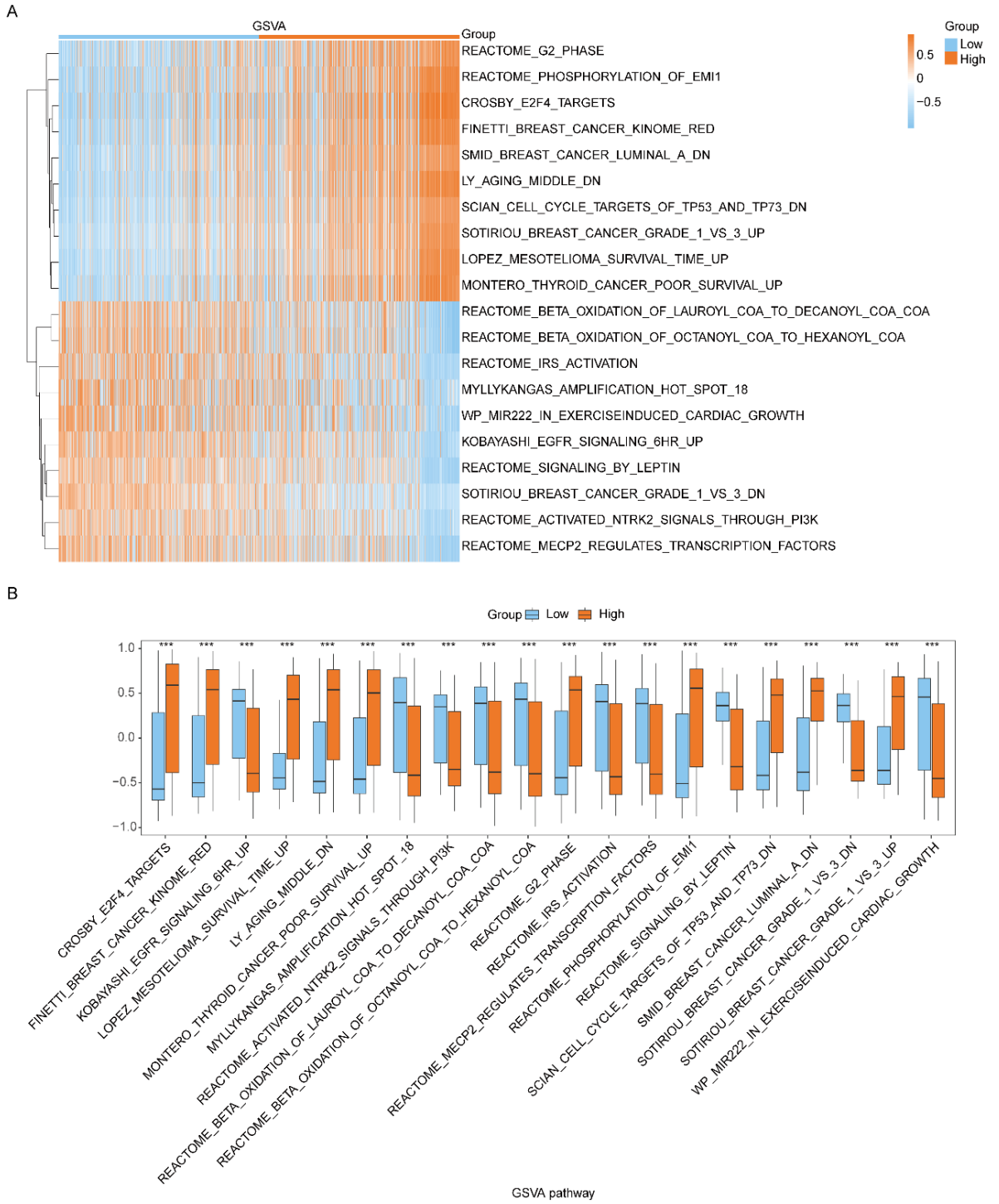

Figure 10. Displays the results of the GSVA enrichment analysis.

The intricate numeric heat map (A) displays the analysis of enrichment, while the map for group comparison (B) presents a visual depiction of the comparison. MHCRDEGs is the abbreviation for MHC Related Differentially Expressed Genes, and GSVA stands for Gene Set Variation Analysis. The term "ns" denotes a P-value greater than or equal to 0.05, signifying the absence of statistical significance. The symbol "*" signifies a P-value below 0.05, indicating statistical significance. The symbol "**"

DOI: 10.54254/2753-8818/35/20240909

suggests a P-value below 0.01, indicating considerable statistical significance. Lastly, the symbol "***" represents a P-value below 0.001, denoting highly significant findings.

3.8. ssGSEA analysis of MHCRDEGs diagnostic model between high-risk and low-risk groups in TCGA-BRCA

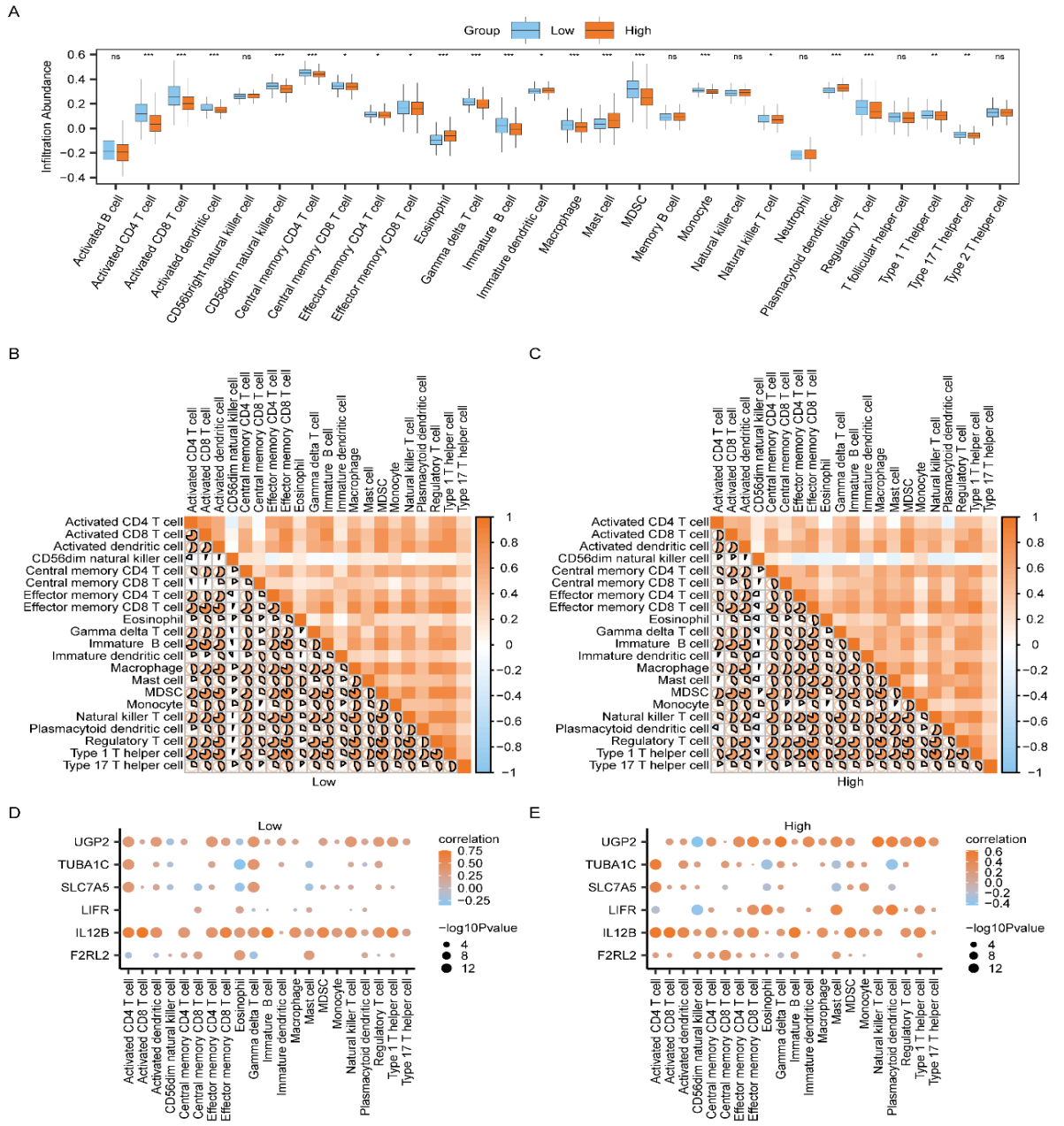

Figure 11. Analysis of Immune Infiltration in the TCGA-BRCA Dataset.

A. provides a comparison diagram that showcases the results of ssGSEA immune infiltration analysis. The presented data aims to investigate the correlation between immune cells in the low-risk (B) and high-risk (C) groups of the MHCRDEGs diagnostic model within the TCGA-BRCA dataset. Additionally, the analysis also demonstrates the correlation between immune cells and Common MHCRDEGs in the low-risk (D) and high-risk groups of the MHCRDEGs diagnostic model in the TCGA-BRCA dataset(E).

DOI: 10.54254/2753-8818/35/20240909

3.9. *Analysis of CIBERSORT in the TCGA-BRCA Dataset*

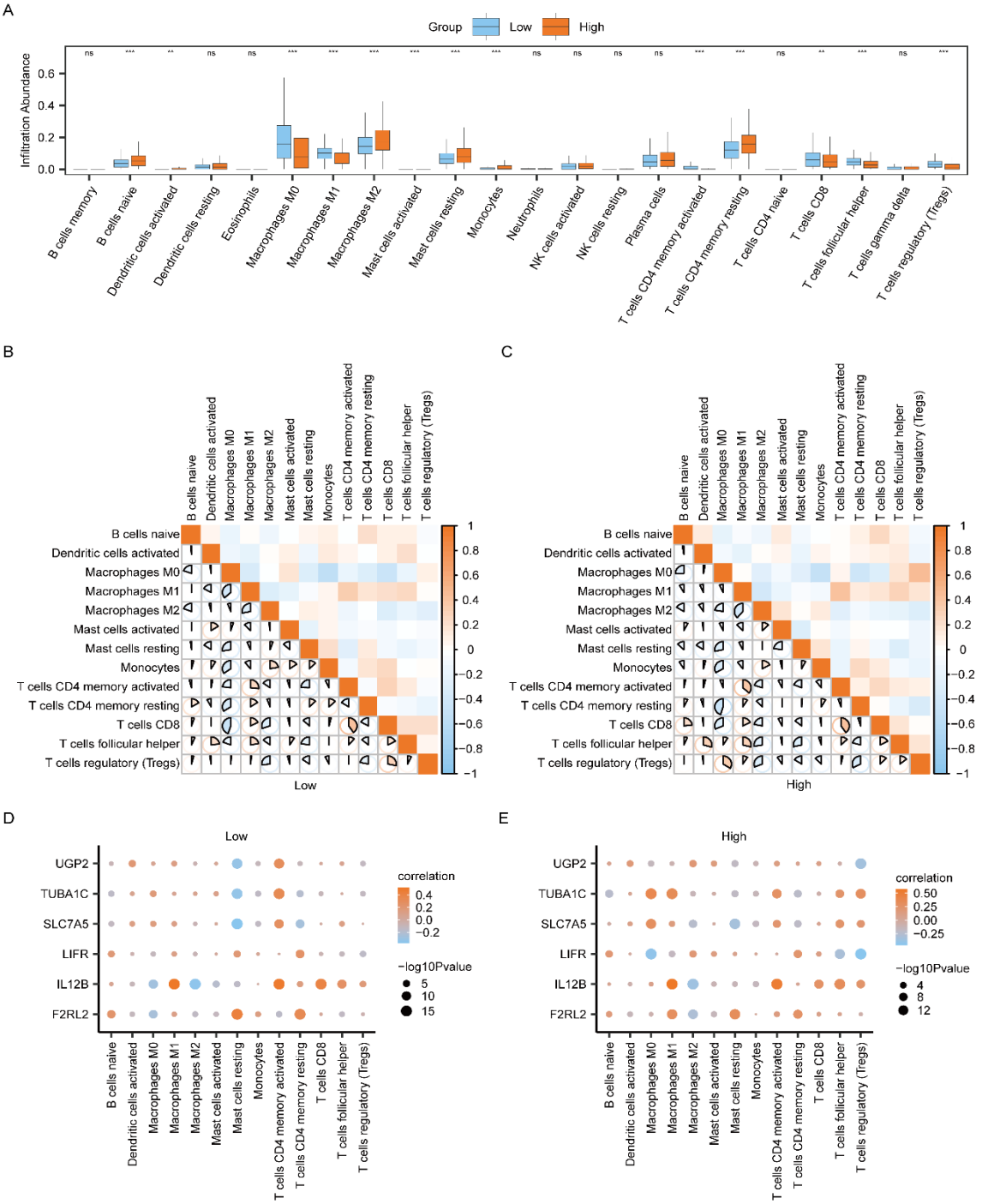

Figure 12. Analysis of CIBERSORT immune features in TCGA-BRCA dataset

(A). Comparison of immune infiltration analysis results for 22 immune cells in the TCGA-BRCA dataset was conducted to assess the risk groups (High/Low) of the MHCRDEGs diagnostic model. The correlation between immune cells in the low-risk (B) and high-risk (C) groups of the MHCRDEGs diagnostic model in the TCGA-BRCA dataset was examined. A dot plot was generated to visualize the correlation between immune cells and Common MHCRDEGs in the low-risk (D) and high-risk (E) groups of the MHCRDEGs diagnostic model in the TCGA-BRCA dataset.

DOI: 10.54254/2753-8818/35/20240909

3.10. *Immunoscore analysis of the TCGA-BRCA*

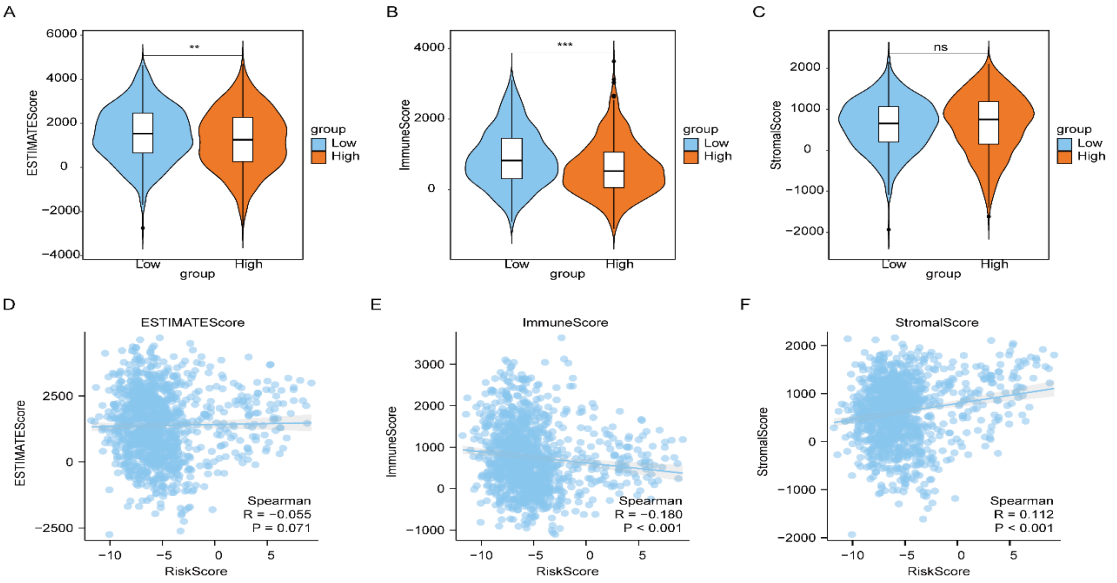

Figure 13. Shows the analysis of Immunoscore for the TCGA-BRCA dataset.

The assessment of ESTIMATE Score (A), Immune Score (B), and Stromal Score (C) in the diagnostic model MHCRDEGs is compared. By creating a scatter plot, the relationship between Riskscores in the prognostic model MHCRDEGs and ESTIMATE Score (D), Immune Score (E), as well as Stromal Score (F) in the TCGA-BRCA dataset is illustrated.

DOI: 10.54254/2753-8818/35/20240909

3.11. A comparative analysis was performed on ICPs, HLA family genes, TIDE, TMB, and MSI to distinguish in the diagnostic model MHCRDEGs.

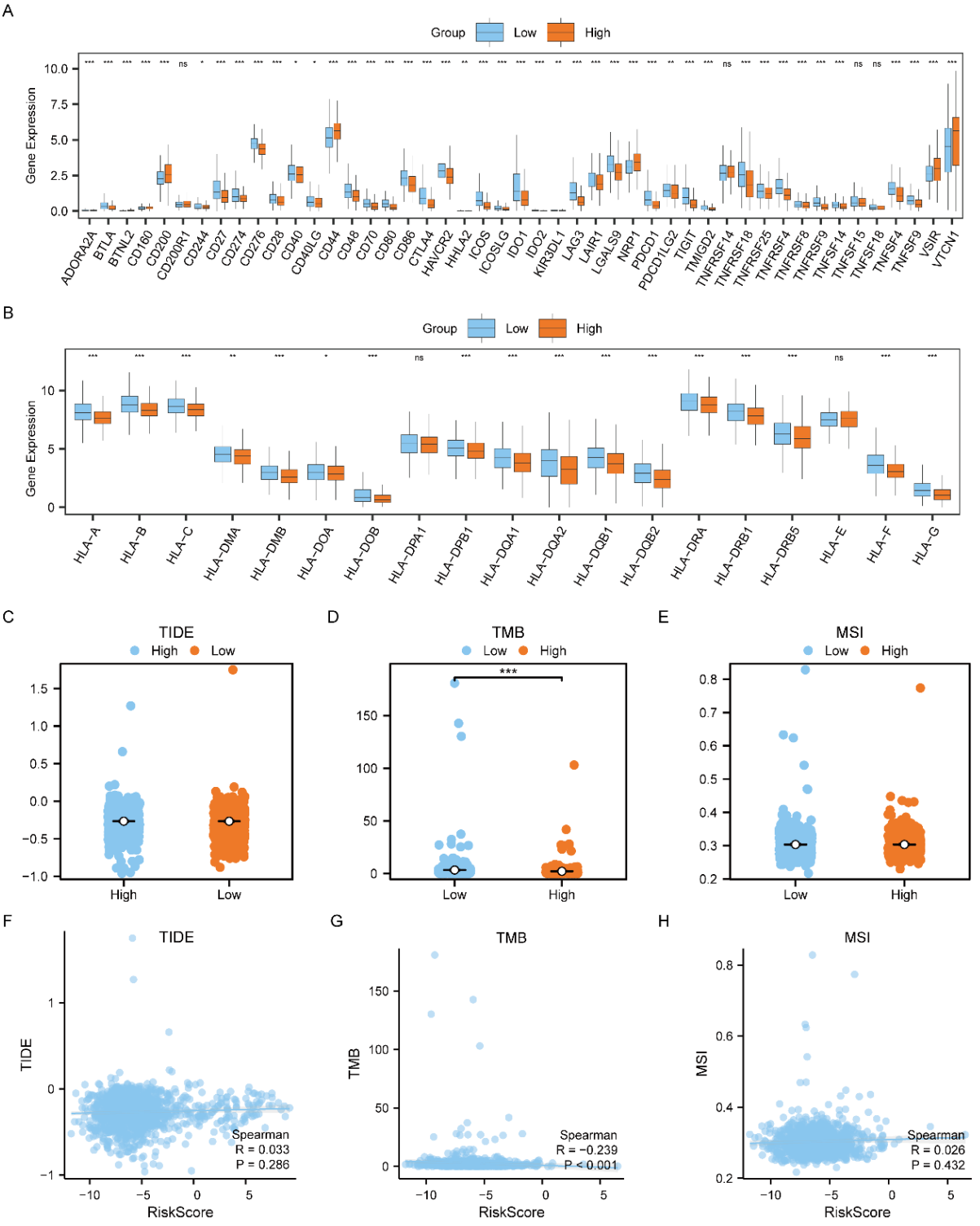

Figure 14. Demonstrates the differential analysis of Immune Checkpoints (ICPs).

Panel A illustrates a comparison between the high and low risk groups of the MHCRDEGs diagnostic model in relation to Immune Checkpoint Proteins (ICPs). Analogously, panel B presents the differential

DOI: 10.54254/2753-8818/35/20240909

analysis outcomes for the HLA family genes in the aforementioned high and low risk groups. Meanwhile, panels C, D, and E showcase the data for TIDE, TMB, and MSI, respectively. Additionally, we have provided correlation scatter plots (panels F, G, and H) demonstrating the association between TIDE, TMB, MSI, and the Riskscore of the MHCRDEGs diagnostic model.

3.12. Drug sensitivity analysis of Common MHCRDEGs

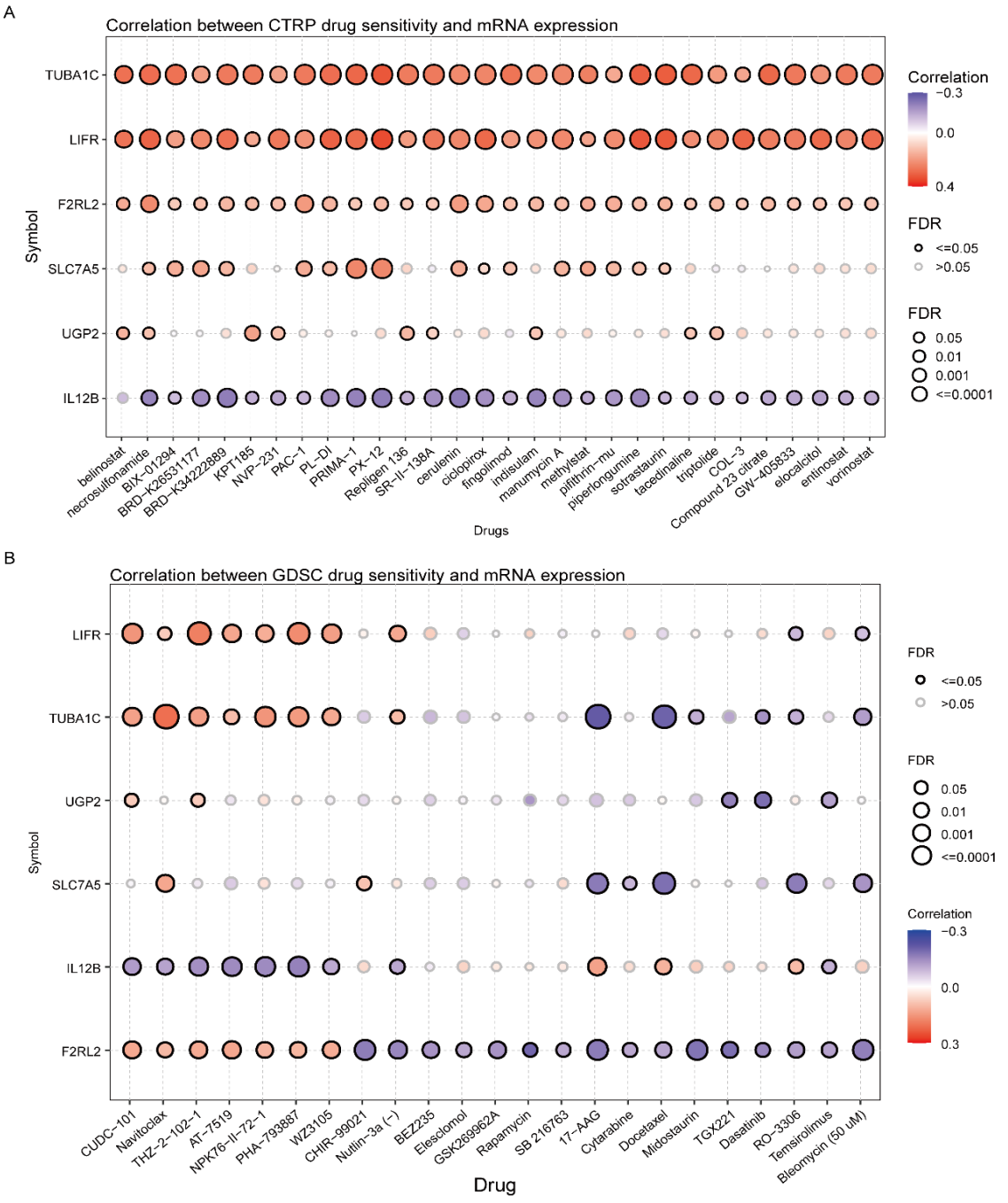

Figure 15. Drug sensitivity analysis of Common MHCRDEGs

A-B. The results of drug sensitivity analysis of Common MHCRDEGs based on CTRP database (A) and GDSC database (B) are presented.

DOI: 10.54254/2753-8818/35/20240909

3.13. *Differential expression analysis of MHCRDEGs*

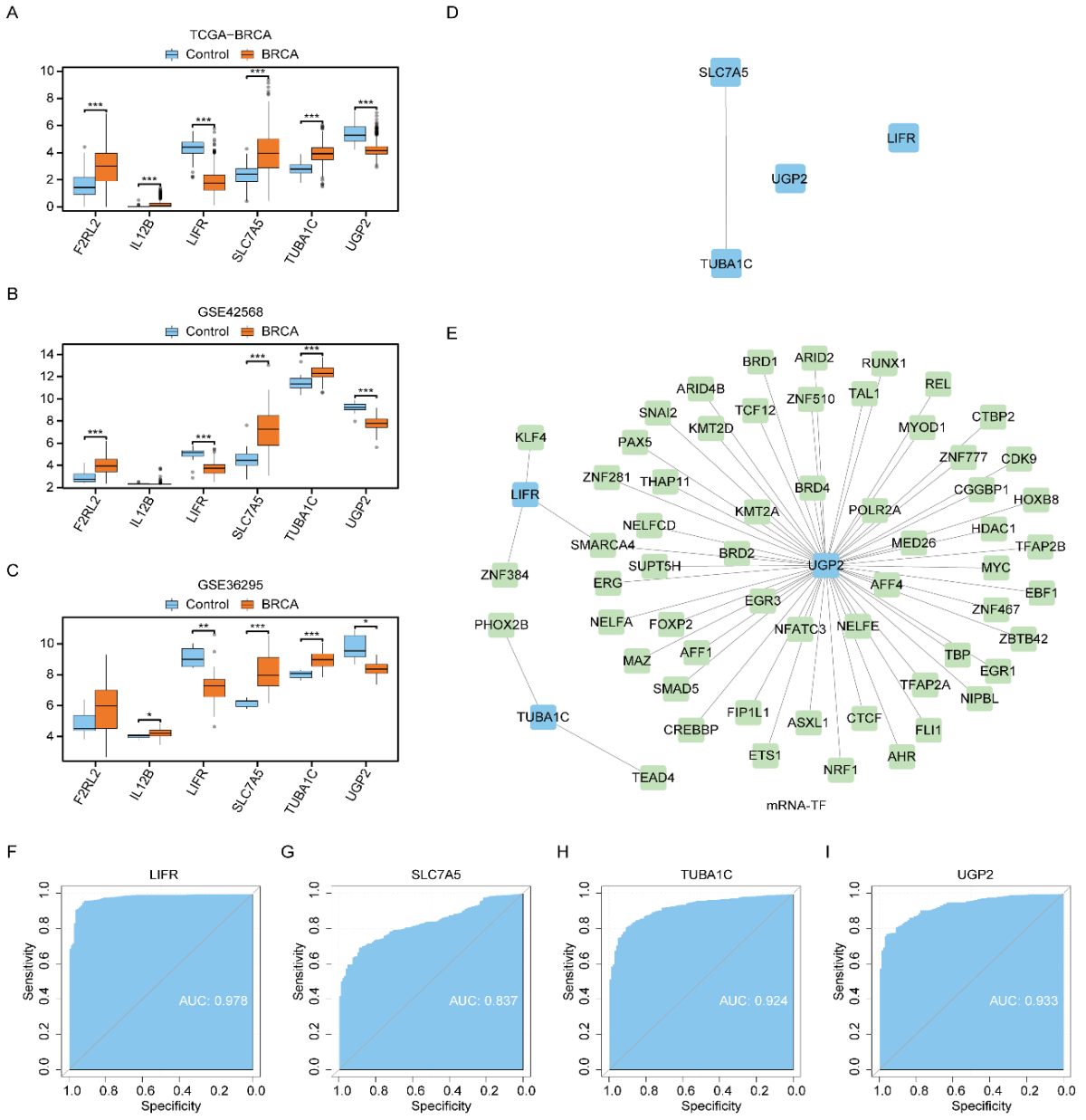

Figure 16. Differential analysis of gene expression and network of mRNA-transcription factor (TF) interactions for Common MHCRDEGs.

This figure presents the findings of group comparisons for Common MHCRDEGs in three different datasets: TCGA-BRCA (A), GSE42568 (B), and GSE36295 (C). The PPI network (D) and mRNA-TF interaction network (E) for Common MHCRDEGs are depicted. Additionally, we display the diagnostic ROC curves for LIFR (F), SLC7A5 (G), TUBA1C (H), and UGP2 (I). In order to evaluate the diagnostic effectiveness, an analysis of the area under the receiver operating characteristic (ROC) curve is conducted. The diagnostic accuracy is directly proportional to the proximity of the AUC to 1. AUC values within the range of 0.5 to 0.7 reflect a low level of accuracy, whereas AUC values ranging from 0.7 to 0.9 suggest a moderate level of accuracy. AUC values surpassing 0.9 are indicative of a high level of accuracy. Within the mRNA-TF interaction network, mRNA entities are depicted as blue squares, whereas TFs are represented by green squares.

DOI: 10.54254/2753-8818/35/20240909

3.14. Construct the prognostic model of MHCRDEGs

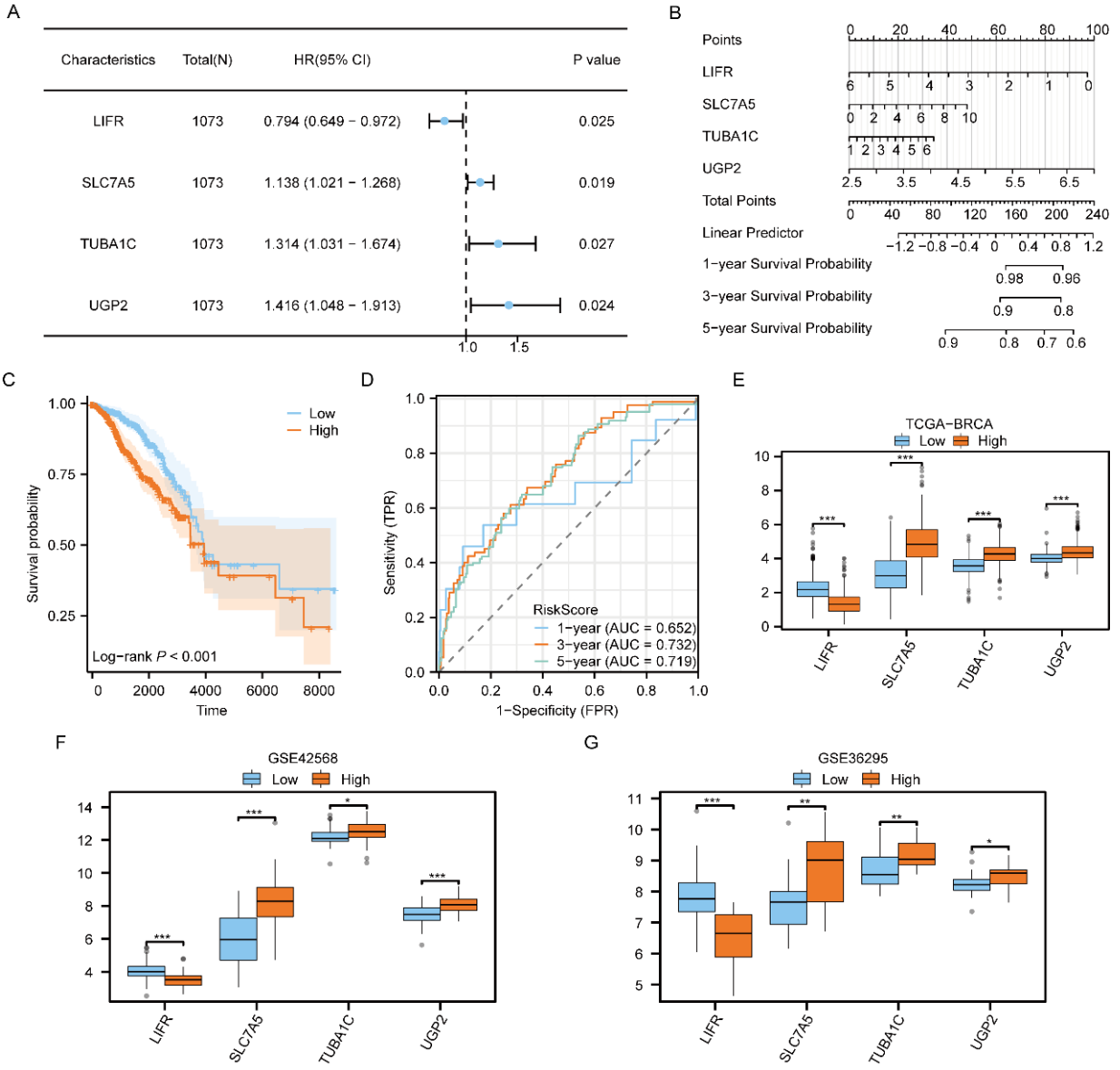

Figure.17 Construction of the Prognostic Model for MHCRDEGs

The multivariate Cox regression model for the TCGA-OV dataset was presented in the forest plot (A). Additionally, a nomogram was developed based on this multivariate Cox regression model (B) to improve prognostic predictions. To evaluate the performance of the MHCRDEGs prognostic model, KM curve analysis (C) and time-dependent ROC curve analysis (D) were conducted. The comparison of high and low-risk groups was carried out to analyze the four common MHCRDEGs in the TCGA-BRCA dataset (E), GSE42568 dataset (F), and GSE36295 dataset (G), and the results were visualized using group comparison diagrams. The significance levels were assigned as follows: p-values $\geqslant$0.05 indicated no statistical significance, p-values <0.05 indicated statistical significance, p-values <0.01 indicated high statistical significance, and p-values <0.001 indicated highly significant statistical findings.

DOI: 10.54254/2753-8818/35/20240909

3.15. *Analysis of Differences in ICPs, TIDE, TMB, MSI, and TCIA*

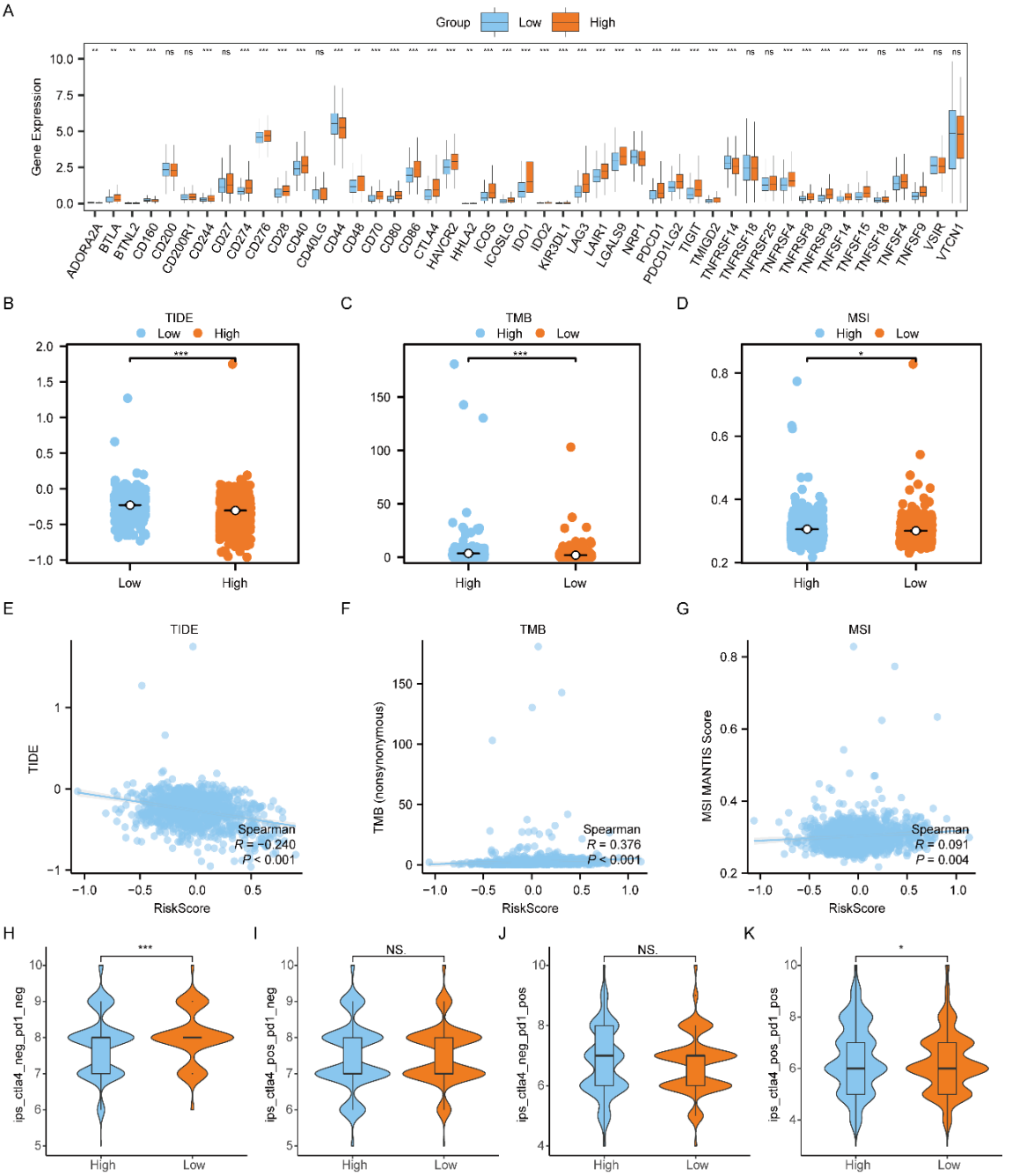

Figure 18. Shows the analysis of differences in ICPs, TIDE, TMB, MSI, and TCIA using the TCGA-BRCA dataset.

(A)The MHCRDEGs prognostic model in the TCGA-BRCA dataset was used to compare the ICPs .(B-D) Group comparison graphs were generated for TIDE (B), TMB (C), and MSI (D) in BRCA patients. (E-G) Scatter plots were used to show the correlation between TIDE (E), TMB (F), MSI (G), (H-K) Boxplot analysis was conducted for ips_ctla4_neg_pd1_neg (H), ips_ctla4_neg_pd1_pos (I), ips_ctla4_pos_pd1_neg (J) , IPS_CTLA4_NEG_PD1_pos (I), and IPS_CTLA4_POS_PD1_neg (J). An additional boxplot was shown for ips_ctla4_pos_pd1_pos (K).ICPs, which stands for immune checkpoints, were the focal point of this research.

DOI: 10.54254/2753-8818/35/20240909

3.16. *Construct a clinically relevant prognostic model*

$$\begin{aligned} \text{Risk}scores \;=\; & -0.003373269 + Pathologic_{T_{stageT2}} * 0.070412788 + PPathologic_{T_{stageT3}} \\ & * 0.150981972 + Pathologic_{T_{stageT4}} * 0.763038816 + Pathologic_{N_{stageN2}} \\ & * 0.877881836 + Pathologic_{N_{stageN3}} * 1.206011272 + Pathologic_{N_{stageN1}} \\ & * 0.501978246 + Pathologic_{T_{stageT2}} * 0.0.666333968 \\ & + MHCRDEGs\ Prognosis\ Model * 1.085307349 \end{aligned}$$

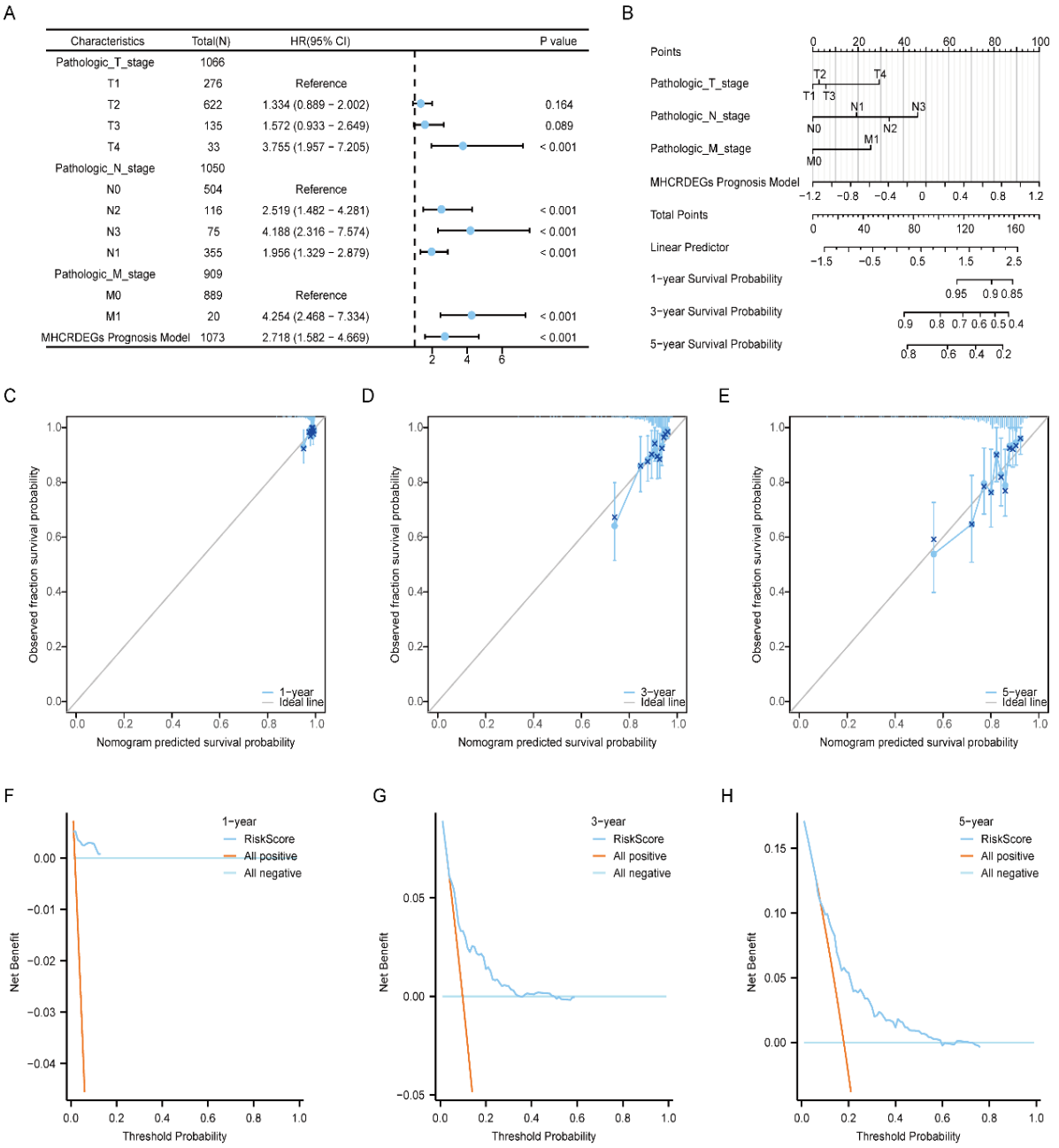

Figure 19. Development of a Prognostic Model for TCGA-BRCA Dataset

A. Presentation of the outcomes from the univariate Cox regression examination can be seen in a forest plot. B. A visual representation in the form of a nomogram is provided to depict the clinically relevant prognostic model for the TCGA-BRCA dataset.C-E. Calibration curves are utilized to analyze the

DOI: 10.54254/2753-8818/35/20240909

accuracy of the clinically relevant prognostic model. These curves specifically focus on the 1-year (C), 3-year (D), and 5-year (E) durations.F-H. Decision curve analysis (DCA) plots are employed to evaluate the performance of the clinically relevant prognostic models within different time frames. The time frames examined include 1 year (F), 3 years (G), and 5 years (H).

4. Discussion

Breast cancer may hinder the expression of MHC molecules, facilitating immune system evasion by tumor cells. Additionally, polymorphisms in MHC-related genes can impact the efficacy of antigen presentation, subsequently influencing the immune detection and elimination of tumor. Consequently, the atypical expression of MHC-related genes in breast cancer may serve as a crucial mechanism for tumor immune evasion [8-10].

In our study, a total of 99 MHCRDEGs were obtained through differential analysis, and mutation analysis was performed on them. It was found that the mutation types of the 99 MHCRDEGs in BRCA patients were mainly single nucleotide polymorphisms (SNPs). SNPs are DNA sequences . SNP may affect gene expression or function, thereby affecting disease susceptibility [31]. The gene with the most single nucleotide polymorphisms is BRCA2. The BRCA2 gene is an important tumor suppressor gene. Therefore, the BRCA2 gene is considered One of the genetic susceptibility factors for breast cancer [32, 33]. We then analyzed the CNV of 99 MHCRDEGs in breast cancer (BRCA) patients, show that the most amplified genes were PIGR and CHI3L1, while there was the most deleted SLC7A5, which indicated that MHC-related genes Gene variation is an important pathogenesis of breast cancer, and it also affects patient survival and treatment effects. Based on this, an enrichment analysis was conducted on 99 MHCRDEGs to find the functions and pathways of MHC-related genes in breast cancer. The findings displayed their notable participation in the interaction between cytokines and cytokine receptors, the signaling pathway of JAK-STAT, and the pathway of prolactin signaling, along with differentiation of Th1 and Th2 cells, and the pathway of relaxin signaling among other pathways. Remarkable discoveries propose a substantial increase in the expression of prolactin (PRL) and its receptors[14]As a result, these pathways involved in transducing signals actively contribute to the regulation of a wide range of biological phenomena, including cell proliferation, differentiation, programmed cell death, and immune response. Additionally, PRL (prolactin) possesses the ability to enhance the growth, viability, movement, and invasiveness of cells associated with breast cancer [15]. Cytokine-receptor interactions can enhance the proliferation and spread of breast cancer cells. These cytokines interact with their corresponding receptors. As a result, cancer cells experience increased proliferation and invasion, ultimately expediting tumor growth and metastasis [16-19]. Conclusively, MHCRDEGs primarily contribute to biological processes involving immune regulation, as well as growth and proliferation. Additionally, the findings demonstrate that MHC-related genes facilitate breast cancer progression by evading the immune system, promoting growth and proliferation, and impacting the prognosis.

This research used a machine learning method to identify 6 characteristic genes (LIFR, UGP2, F2RL2, SLC7A5, TUBA1C, IL12B). In order to examine the influence of gene expression levels on BRCA within these risk categories, a gene set enrichment analysis (GSEA) was conducted on the entire set of genes in the TCGA-BRCA dataset using the MHCRDEGs diagnostic model. The outcomes of the analysis indicated that pathways related to growth and proliferation, such as PI3K-AKT, displayed significant enrichment in the high-risk group. Conversely, immune-related pathways, such as MHC, were predominantly enriched in the low-risk group. Additionally, immune scores were assigned to both the high-risk and low-risk groups based on the diagnostic model. The examination revealed noteworthy discrepancies in the ESTIMATE Score and Immune Score between these two risk groups. The findings indicated that the high-risk group exhibited lower scores in both ESTIMATE Score and Immune Score. It is commonly understood that higher immune scores correlate with better prognoses, suggesting that the high-risk group of the diagnostic model may have a poorer prognosis.

In order to further establish a clinically relevant prognostic model, we conducted a multi-data set control difference analysis on 6 characteristic genes. We discovered that the levels of expression for four MHCRDEGs (LIFR, SLC7A5, TUBA1C, UGP2) remained consistent. This difference was statistically

DOI: 10.54254/2753-8818/35/20240909

significant. To examine the above-mentioned differentiating genes further, we developed a prognostic model for MHCRDEGs. The expression patterns of the four commonly occurring MHCRDEGs in both the high-risk and low-risk groups of the prognostic model demonstrated significant variances across the three data sets, reaffirming the reliability of the model. Significantly, the Riskscore prognostic model for TIDE, TMB, MSI, and MHCRDEGs exhibited significant statistical disparities between the low-risk and high-risk groups, deviating from the diagnostic model. Additionally, the TIDE score was utilized to evaluate the sensitivity of patients to immunotherapy. Through our analysis, it was found that the low-risk group presented slightly elevated scores in comparison to the high-risk group (with statistical significance). This suggests the potential for immunotherapy to further augment survival outcomes for patients within the low-risk group.

The four Common MHCRDEGs are: LIFR, SLC7A5, TUBA1C, and UGP2 [34].Breast cancer extensively disrupts the expression and operational capacity of LIFR. Research has substantiated the diminished levels of LIFR expression in breast cancer tissue , establishing a strong association with breast cancer's malignancy and prognosis [35, 36]. Downregulation of LIFR may be caused by various mechanisms such as gene deletion, promoter region methylation , or microRNA regulation [35, 37, 38]. This reduction in LIFR can result in uncontrolled cell proliferation and the inhibition of apoptosis, ultimately promoting the progression of breast cancer. These findings align with our study. Additionally, the SLC7A5 gene encodes a sodium-glucose co-transporter (SGLT) responsible for glucose absorption [24, 25]. Our study reveals a significantly higher expression of the SLC7A5 gene in breast cancer patients compared to the normal population. Consequently, it is evident that the SLC7A5 gene contributes to tumor cell proliferation and invasion [25]. Numerous studies have demonstrated that breast cancer cells exhibiting elevated expression of the SLC7A5 gene possess heightened capabilities of proliferation and invasion [26, 27]. Aberrations in the TUBA1C gene can lead to abnormal accumulation of tubulin, thereby facilitating cell division and proliferation, hindering cell apoptosis, and ultimately resulting in the onset of breast cancer [28]. Moreover, mutations in the TUBA1C gene are more frequently observed among breast cancer patients and are notably associated with an unfavorable prognosis [29]. Subsequent investigations have revealed that TUBA1C gene mutations can induce heightened proliferation, invasion, and metastasis capabilities in breast cancer cells, while simultaneously diminishing their sensitivity to chemotherapy drugs [30, 31]. Thus, it can be deduced that TUBA1C gene mutation represents a pivotal factor in the development of breast cancer. On the other hand, the UGP2 gene encodes UDP-glucuronate dehydrogenase, a pivotal enzyme operating within the glycolysis pathway. Extensive research has demonstrated that UGP2 gene expression is elevated in various cancer types, including breast cancer [32, 33]. Mechanistic investigations have established that UGP2 can trigger the proliferation, inhibition of apoptosis, and metastasis of cancer cells via the activation of the PI3K/AKT and MAPK signaling pathways [34-37].

5. Discussion

In summary, our research findings indicate substantial disparities in MHC-related genes between normal tissues and breast cancer tumor. By utilizing advanced machine learning techniques, we identified six distinctive genes and developed a diagnostic risk model capable of predicting breast cancer occurrence. To support early detection efforts, we compared these six genes with various datasets, eventually determining four genes of significance and constructing a clinical prognostic risk model. The model's credibility and precision have been confirmed across multiple datasets. Utilizing this model, we can establish a theoretical foundation for predicting patient survival prognosis.

References

[1] Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, Shi W, Jiang J, Yao PP, Zhu HP: Risk Factors and Preventions of Breast Cancer. *Int J Biol Sci* 2017, 13(11):1387-1397.

[2] Ho PJ, Khng AJ, Tan BK, Tan EY, Tan SM, Tan VKM, Lim GH, Aronson KJ, Chan TL, Choi JY *et al*: Relevance of the MHC region for breast cancer susceptibility in Asians. *Breast Cancer* 2022, 29(5):869-879.

[3] Grimholt U: MHC and Evolution in Teleosts. *Biology (Basel)* 2016, 5(1).
[4] Breast Cancer Association C, Dorling L, Carvalho S, Allen J, Gonzalez-Neira A, Luccarini C, Wahlstrom C, Pooley KA, Parsons MT, Fortuno C *et al*: Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women. *N Engl J Med* 2021, 384(5):428-439.
[5] He Q, Liu Z, Liu Z, Lai Y, Zhou X, Weng J: TCR-like antibodies in cancer immunotherapy. *J Hematol Oncol* 2019, 12(1):99.
[6] Szekely B, Bossuyt V, Li X, Wali VB, Patwardhan GA, Frederick C, Silber A, Park T, Harigopal M, Pelekanou V *et al*: Immunological differences between primary and metastatic breast cancer. *Ann Oncol* 2018, 29(11):2232-2239.
[7] Jobim MR, Jobim M, Salim PH, Portela P, Jobim LF, Leistner-Segal S, Bittelbrunn AC, Menke CH, Biazus JV, Roesler R *et al*: Analysis of KIR gene frequencies and HLA class I genotypes in breast cancer and control group. *Hum Immunol* 2013, 74(9):1130-1133.
[8] Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000, 28(1):27-30.
[9] Tsang JY, Ho CS, Ni YB, Shao Y, Poon IK, Chan SK, Cheung SY, Shea KH, Marabi M, Tse GM: Co-expression of HLA-I loci improved prognostication in HER2+ breast cancers. *Cancer Immunol Immunother* 2020, 69(5):799-811.
[10] Dejardin E, Deregowski V, Greimers R, Cai Z, Chouaib S, Merville MP, Bours V: Regulation of major histocompatibility complex class I expression by NF-kappaB-related proteins in breast cancer cells. *Oncogene* 1998, 16(25):3299-3307.
[11] Moller P, Hammerling GJ: The role of surface HLA-A,B,C molecules in tumour immunity. *Cancer Surv* 1992, 13:101-127.
[12] Walter NA, McWeeney SK, Peters ST, Belknap JK, Hitzemann R, Buck KJ: Single-nucleotide polymorphism masking. *Alcohol Res Health* 2008, 31(3):270-271.
[13] Filippini SE, Vega A: Breast cancer genes: beyond BRCA1 and BRCA2. *Front Biosci (Landmark Ed)* 2013, 18(4):1358-1372.
[14] Gracia-Aznarez FJ, Fernandez V, Pita G, Peterlongo P, Dominguez O, de la Hoya M, Duran M, Osorio A, Moreno L, Gonzalez-Neira A *et al*: Whole exome sequencing suggests much of non-BRCA1/BRCA2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. *PLoS One* 2013, 8(2):e55681.
[15] Gill S, Peston D, Vonderhaar BK, Shousha S: Expression of prolactin receptors in normal, benign, and malignant breast tissue: an immunohistological study. *J Clin Pathol* 2001, 54(12):956-960.
[16] Aksamitiene E, Achanta S, Kolch W, Kholodenko BN, Hoek JB, Kiyatkin A: Prolactin-stimulated activation of ERK1/2 mitogen-activated protein kinases is controlled by PI3-kinase/Rac/PAK signaling pathway in breast cancer cells. *Cell Signal* 2011, 23(11):1794-1805.
[17] Lopez-Mejia JA, Mantilla-Ollarves JC, Rocha-Zavaleta L: Modulation of JAK-STAT Signaling by LNK: A Forgotten Oncogenic Pathway in Hormone Receptor-Positive Breast Cancer. *Int J Mol Sci* 2023, 24(19).
[18] Standing D, Dandawate P, Anant S: Prolactin receptor signaling: A novel target for cancer treatment - Exploring anti-PRLR signaling strategies. *Front Endocrinol (Lausanne)* 2022, 13:1112987.
[19] Wu L, Sun S, Qu F, Sun M, Liu X, Sun Q, Cheng L, Zheng Y, Su G: CXCL9 influences the tumor immune microenvironment by stimulating JAK/STAT pathway in triple-negative breast cancer. *Cancer Immunol Immunother* 2023, 72(6):1479-1492.
[20] Song X, Wei C, Li X: The Signaling Pathways Associated With Breast Cancer Bone Metastasis. *Front Oncol* 2022, 12:855609.
[21] Viswanadhapalli S, Dileep KV, Zhang KYJ, Nair HB, Vadlamudi RK: Targeting LIF/LIFR signaling in cancer. *Genes Dis* 2022, 9(4):973-980.
[22] Chen D, Sun Y, Wei Y, Zhang P, Rezaeian AH, Teruya-Feldstein J, Gupta S, Liang H, Lin HK, Hung MC *et al*: LIFR is a breast cancer metastasis suppressor upstream of the Hippo-YAP pathway and a prognostic marker. *Nat Med* 2012, 18(10):1511-1517.

DOI: 10.54254/2753-8818/35/20240909

[23] Clements ME, Holtslander L, Edwards C, Todd V, Dooyema SDR, Bullock K, Bergdorf K, Zahnow CA, Connolly RM, Johnson RW: HDAC inhibitors induce LIFR expression and promote a dormancy phenotype in breast cancer. *Oncogene* 2021, 40(34):5314-5326.

[24] Woosley AN, Dalton AC, Hussey GS, Howley BV, Mohanty BK, Grelet S, Dincman T, Bloos S, Olsen SK, Howe PH: TGFbeta promotes breast cancer stem cell self-renewal through an ILEI/LIFR signaling axis. *Oncogene* 2019, 38(20):3794-3811.

[25] Xu F, Li H, Hu C: LIFR-AS1 modulates Sufu to inhibit cell proliferation and migration by miR-197-3p in breast cancer. *Biosci Rep* 2019, 39(7).

[26] Tornroos R, Tina E, Gothlin Eremo A: SLC7A5 is linked to increased expression of genes related to proliferation and hypoxia in estrogen-receptor-positive breast cancer. *Oncol Rep* 2022, 47(1).

[27] Hisada T, Kondo N, Wanifuchi-Endo Y, Osaga S, Fujita T, Asano T, Uemoto Y, Nishikawa S, Katagiri Y, Terada M *et al*: Co-expression effect of LLGL2 and SLC7A5 to predict prognosis in ERalpha-positive breast cancer. *Sci Rep* 2022, 12(1):16515.

[28] El Ansari R, Craze ML, Miligy I, Diez-Rodriguez M, Nolan CC, Ellis IO, Rakha EA, Green AR: The amino acid transporter SLC7A5 confers a poor prognosis in the highly proliferative breast cancer subtypes and is a key therapeutic target in luminal B tumours. *Breast Cancer Res* 2018, 20(1):21.

[29] Zou Y, Wang G, Fan M: Comprehensive Multiomic Analysis Identified TUBA1C as a Potential Prognostic Biological Marker of Immune-Related Therapy in Pan-Cancer. *Comput Math Methods Med* 2022, 2022:9493115.

[30] Ramos J, Yoo C, Felty Q, Gong Z, Liuzzi JP, Poppiti R, Thakur IS, Goel R, Vaid AK, Komotar RJ *et al*: Sensitivity to differential NRF1 gene signatures contributes to breast cancer disparities. *J Cancer Res Clin Oncol* 2020, 146(11):2777-2815.

[31] Wu Z, Sun S, Fan R, Wang Z: Tubulin alpha 1c promotes aerobic glycolysis and cell growth through upregulation of yes association protein expression in breast cancer. *Anticancer Drugs* 2022, 33(2):132-141.

[32] Nami B, Wang Z: Genetics and Expression Profile of the Tubulin Gene Superfamily in Breast Cancer Subtypes and Its Relation to Taxane Resistance. *Cancers (Basel)* 2018, 10(8).

[33] Kim S, Wolfe A, Kim SE: Targeting cancer's sweet spot: UGP2 as a therapeutic vulnerability. *Mol Cell Oncol* 2021, 8(6):1990676.

[34] Hu Q, Shen S, Li J, Liu L, Liu X, Zhang Y, Zhou Y, Zhu W, Yu Y, Cui G: Low UGP2 Expression Is Associated with Tumour Progression and Predicts Poor Prognosis in Hepatocellular Carcinoma. *Dis Markers* 2020, 2020:3231273.

[35] Zeng C, Xing W, Liu Y: Identification of UGP2 as a progression marker that promotes cell growth and motility in human glioma. *J Cell Biochem* 2019, 120(8):12489-12499.

[36] Wolfe AL, Zhou Q, Toska E, Galeas J, Ku AA, Koche RP, Bandyopadhyay S, Scaltriti M, Lebrilla CB, McCormick F *et al*: UDP-glucose pyrophosphorylase 2, a regulator of glycogen synthesis and glycosylation, is critical for pancreatic cancer growth. *Proc Natl Acad Sci U S A* 2021, 118(31).

[37] Li Y, Zhuang H, Zhang X, Li Y, Liu Y, Yi X, Qin G, Wei W, Chen R: Multiomics Integration Reveals the Landscape of Prometastasis Metabolism in Hepatocellular Carcinoma. *Mol Cell Proteomics* 2018, 17(4):607-618.

[38] Pescador N, Villar D, Cifuentes D, Garcia-Rocha M, Ortiz-Barahona A, Vazquez S, Ordonez A, Cuevas Y, Saez-Morales D, Garcia-Bermejo ML *et al*: Hypoxia promotes glycogen accumulation through hypoxia inducible factor (HIF)-mediated induction of glycogen synthase 1. *PLoS One* 2010, 5(3):e9644.