Discover Oncology

Analysis

Machine learning-based detoxification enzymes-related genes prognosis model in breast cancer: immune landscape and clinical significance

Jingdi Zhang[1,2] · Wendi Zhan[1,2] · Haihong Hu[1,2] · Hongxia Zhu[1,2] · Bo Hao[2] · Siyu Wang[3] · Zhuo Li[2] · Zhiming Zhang[4] · Taolan Zhang[2,5]

Received: 25 December 2024 / Accepted: 10 May 2025
Published online: 23 June 2025

Abstract

Background Breast cancer is one of the most common malignant tumors, threatening women's health and life globally. Despite significant treatment advances, its prognosis still faces great challenges. With the rapid development of molecular biology and genomics, the role of detoxification enzymes in breast cancer occurrence, development, and prognosis has gained increasing attention. This paper aims to establish a prognostic model based on detoxification enzymes-related genes to predict breast cancer patient survival.

Methods Unsupervised clustering was used to analyze breast cancer samples based on detoxification enzymes-related genes expression. Lasso cox regression analysis and univariate and multivariate Cox analysis were used to process the data, and machine learning algorithm was used to construct breast cancer prognosis model. The effect of detoxification enzymes-related genes on breast cancer was analyzed by single cell analysis.

Results The samples were classified into two subtypes, and a breast cancer prognosis model based on detoxification enzymes-related genes was constructed and validated using TCGA and GEO cohorts. Significant differences in pathways, immune infiltration, immunotherapy response, and drug sensitivity were observed between high- and low-risk groups. Single-cell analysis revealed that SQLE, a detoxification enzymes-related gene, was highly expressed in breast cancer epithelial cells (cancer cells), where SQLE + epithelial cells primarily influenced exhausted CD8 + T cells via the MIF signaling pathway.

Conclusion In summary, the detoxification enzymes-related genes-based prognostic model developed in this study provides an effective tool for predicting breast cancer prognosis and offers new insights for diagnosis and treatment.

Keywords Detoxification enzymes · Breast cancer · Prognosis model · Single-cell analysis

Jingdi Zhang and Wendi Zhan contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12672-025-02656-3.

✉ Zhuo Li, zhuoli71@126.com; ✉ Zhiming Zhang, 404974038@qq.com; ✉ Taolan Zhang, TaolZhan@usc.edu.cn | [1]School of Pharmacy, Hengyang Medical College, University of South China, 28 Western Changsheng Road, Hengyang 421001, Hunan, China. [2]Department of Pharmacy, The First Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang 421001, Hunan, China. [3]Department of Medical Oncology, The First Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang 421001, Hunan, China. [4]Health School of Nuclear Industry, 336 South Dongfeng Road, Hengyang 421001, Hunan, China. [5]Research Center for Clinical Trial, The First Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang 421001, Hunan, China.

1 Introduction

Breast cancer is one of the most common cancers among women worldwide and one of the leading causes of cancer-related death. According to the International Agency for Research on Cancer (IARC), approximately 2 million new cases of breast cancer are diagnosed each year, and the incidence continues to rise globally [1]. The incidence of breast cancer is significantly different in different regions, which is closely related to genetic factors, lifestyle, and national screening policies. Breast cancer risk is linked to a number of factors, including age, genetic mutations (such as mutations in the BRCA1 and BRCA2 genes), hormone levels (such as long-term exposure to estrogen and progesterone), lifestyle (such as alcohol consumption, obesity, and physical inactivity), and more [2]. In terms of treatment, the management of breast cancer has developed from traditional surgery, chemotherapy and radiotherapy to targeted therapy and immunotherapy [3]. However, despite significant advances in the treatment and management of breast cancer, challenges remain with poor prognosis for advanced disease and increased need for individualized treatment. In recent years, with the rapid development of molecular biology and genomics, researchers have found that detoxification enzymes play an important role in the occurrence, development and prognosis of breast cancer [4, 5].

Detoxification enzymes, including P-oxidases cytochrome (CYP) enzymes, glutathione S-transferase (GSTs), UDP-glucuronosyltransferases (UGT), Sulfotransferases (SULT) and N- acetyl-transferase (NAT) enzymes, are the key enzymes responsible for the metabolism and removal of harmful substances in the body. These enzymes help maintain the chemical balance in the body by converting and removing environmental toxins, drugs and metabolites [6, 7]. Studies have shown that these enzymes affect the biological processes of cells such as antioxidant capacity, drug response and apoptosis through metabolic pathways [8, 9], thus having a profound impact on the occurrence and progression of tumors. Genetic polymorphisms in detoxification enzymes may influence an individual's susceptibility to cancer, especially breast cancer. Variations in certain detoxification enzymes may lead to reduced metabolic efficiency against carcinogens, thereby increasing the risk of breast tissue exposure to harmful substances. In addition, detoxification enzymes are also involved in the regulation of hormone levels in the body, and changes in hormone levels have an important impact on the occurrence and development of breast cancer. It has been found that the absence of GSTM1 is associated with the occurrence of breast cancer [10]. Therefore, the function of detoxification enzymes and their genetic variation are considered an important area of breast cancer research. However, there is no systematic study on the specific effects of detoxification enzymes-related genes on breast cancer prognosis. Therefore, the construction of breast cancer prognosis model based on detoxification enzymes-related genes has become a potential research direction. Through in-depth analysis of the expression pattern of detoxification enzymes-related genes in breast cancer patients, the potential role of detoxification enzymes-related genes in breast cancer prognosis can be revealed, so as to provide a new tool and basis for breast cancer prognosis prediction.

In this paper, the potential of detoxification enzymes-related genes in predicting the prognosis of breast cancer was investigated for the first time, and a prognostic model was constructed based on these genes. In addition, we performed a single-cell level analysis of prognostic genes to further shed light on the role of detoxification enzymes-related genes in breast cancer. To enhance its clinical value, we developed a web-based tool that combines risk scores with other clinical variables to predict overall survival (OS) and drug response in BRCA patients. By combining gene expression data with clinical information, the prognostic model not only facilitates early identification of high-risk patients, but also supports individualized treatment decisions, thereby improving the precision of clinical management.

2 Methods

The workflow of this research was shown in Fig. 1.

2.1 Data collection and processing

Download gene expression data and clinical data from breast cancer patients in the GDC TCGA breast cancer cohort from the UCSC Xena browser (https://xenabrowser.net/) [11]. After sorting out the data, the information of normal samples and duplicate patient samples were deleted, and the complete sample information of 1069 breast cancer patients was obtained. Subsequently, 1069 BRCA patients were randomly divided into TCGA train cohort (749 cases) and TCGA test cohort (320 cases) at a ratio of 7:3. To verify the accuracy of the prognostic model, the GSE20685

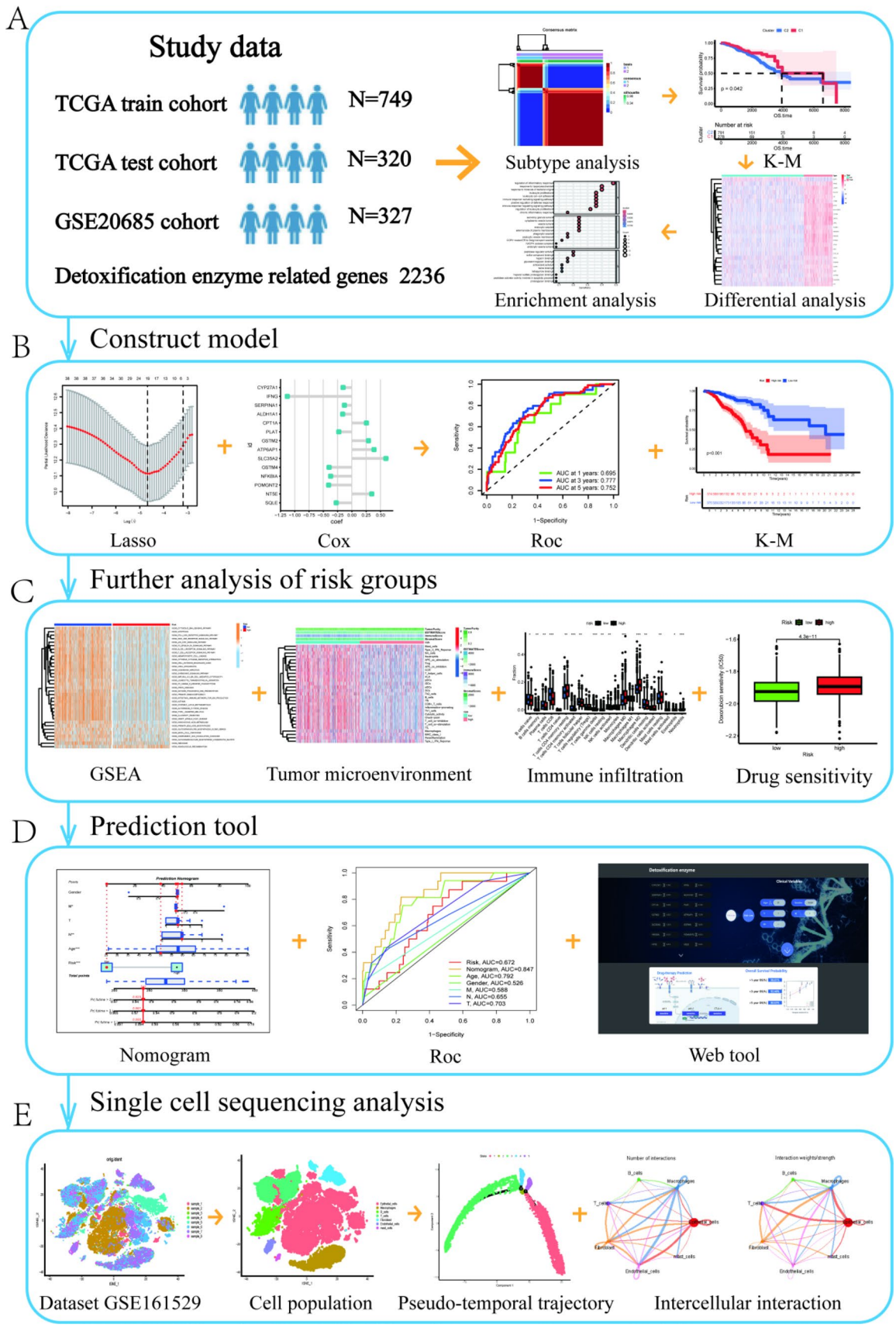

Fig. 1 Workflow diagram of the study. **A** Identification of different subtypes of breast cancer. **B** Construction of breast cancer prognostic model. **C** Immune landscape analysis of different risk groups. **D** Construction of nomogram and development of webtool. **E** Single cell sequencing analysis

dataset which contained 327 breast cancer patients was downloaded from the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) as an external validation cohort [12]. Detailed description of all cohorts was provided in Table 1.In addition, we also download the detoxification enzymes-related genes (2236)from GeneCards database (https://www.genecards.org/) for subsequent analysis. Furthermore, we searched the GEO database for breast cancer datasets (BRCA, GSE161529) and selected four triple-negative breast cancer samples (GSM4909281, GSM4909282, GSM4909283, GSM4909284) and four HER2-positive breast cancer samples (GSM4909289, GSM4909290, GSM4909291, GSM4909292) from a total of 69 samples as single-cell sequencing datasets.

2.2 Subtype analysis

We first used univariate Cox analysis to screen detoxification enzymes-related prognostic genes, and used non-negative matrix decomposition (NMF) algorithm to perform unsupervised cluster analysis, and divided patients into two different subtypes according to the expression of detoxification enzymes-related genes [13]. Then, to verify the classification results, we used principal component analysis (PCA) to analyze the distribution differences of the two subtypes. Kaplan–Meier (KM) survival analysis was used to compare the overall survival (OS) of different subtypes in the TCGA dataset, and $P \leq 0.05$ was considered significant. In addition, we also used R package "maftools" to depict waterfall maps of two different subtypes of mutant landscapes [14, 15] to explore the effect of detoxification enzymes on BRCA mutant spectrum.

Table 1 The clinical characteristics of breast cancer in TCGA cohort and GSE20685

Variables incomplete	TCGA train cohort (N = 757) N = 127	TCGA test cohort (N = 324) N = 52	GSE cohort (N = 327) 0
Age			
≥ 65 years	167	80	22
< 65 years	463	192	305
Sex			
Female	625	266	327
Male	5	6	0
M			
M0	620	265	244
M1	10	7	83
N			
N0	301	141	122
N1	220	83	102
N2	75	28	63
N3	34	20	40
T			
T1	159	78	101
T2	385	151	188
T3	68	30	30
T4	18	13	8
Stage classification			
Stage I	106	54	NA
Stage II	382	146	NA
Stage III	132	65	NA
Stage IV	10	7	NA

2.3 Differential expression analysis

The R-package "limma" was used to screen the differentially expressed genes (DEGs) of subtype 1 and subtype 2, and the differential gene expression heat map was drawn [16]. In order to explore the biological functions and pathways of related genes, we used R packages such as "ggplot2", "clusterProfiler", "DOSE" and "enrichment" [17]. The differential genes were analyzed by gene ontology (GO), disease ontology (DO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment.

2.4 Construction and verification of prognostic model

We used univariate Cox analysis to construct a breast cancer prognosis model based on differentially expressed genes (DEGs). We used glmnet package for least absolute shrinkage and selection operator (LASSO) Cox proportional risk regression to further screen characteristic genes associated with breast cancer prognosis [18]. Subsequently, in the TCGA cohort (N = 755), multivariate Cox regression analysis was used to identify the optimal genes and construct a robust risk model to predict the prognosis of BRCA patients. Regression coefficient derived from the multivariate Cox model were used to assign weights to each gene to obtain their respective contributions to the overall risk score. Finally, the calculation formula of prognostic risk score was obtained: risk score = $\sum_{n}^{i=1}$ = (expression * coefficient). "Expression" is the expression level of the corresponding gene, and "coef" is the gene regression coefficient obtained after building the model. In addition, in order to evaluate the prediction accuracy of the model, we used the TCGA test cohort as the internal validation cohort and the GSE20685 cohort as the external validation cohort to verify the model. We calculated the risk score for each breast cancer patient using the risk score formula, and divided patients into high-risk and low-risk groups based on the median risk score value. First, we plotted risk curves and survival scatter plots for all cohorts of patients. Secondly, the K-M survival curve was drawn using the R software packages "Survival" and "Survminer" to compare the survival of the two risk groups ($p < 0.05$ was considered statistically significant [19]. Then, we used the R-package "survival ROC" to construct a receiver operating characteristic (ROC) curve over time [20], and evaluated the performance of the risk model according to the ROC curve and the area under curve (AUC) predicted over 1, 3 and 5 years, respectively.

2.5 Analysis of DNA promoter methylation

In breast cancer, the promoter methylation status of some genes may affect their expression, which in turn affects the biological characteristics and clinical manifestations of breast cancer, including prognosis. So we use the university of Alabama at Birmingham cancer data analysis portal website (http://ualcan.path.uab.edu/) (UALCAN) to analyze DNA promoter methylation [21]. When TCGA breast cancer patients were selected as research objects in the database and gene names were input, the visual results of promoter methylation degree of these genes could be obtained, as well as the P-value of statistical analysis between normal tissue and primary tumor groups ($p < 0.05$ was considered statistically significant).

2.6 Gene set variation analysis and gene set enrichment analysis

To explore potential biological pathways, we used the "limma", "GSEABase" and "GSVA" packages to perform gene-set variation analysis (GSVA) for high-risk and low-risk patients in all TCGA datasets [22]. Gene set enrichment analysis (GSEA) of the tagged gene set was performed using the R packages "limma", "DOSE", "clusterprofiler" and "enrichplot" and visualized using the R package "Enrichplot" [23, 24].

2.7 Immune landscape analysis of tumor microenvironment and immune infiltration

In view of the critical role of the Tumor microenvironment (TME) in tumor immunotherapy [25, 26], we used the ESTIMATE package to calculate the ESTIMATE Score, Immune Score, Stromal Score and Tumor purity of BRCA samples [27]. The ESTIMATE Score and Tumor purity can estimate the overall purity of the tumor. The immune score and stromal score represent the infiltration level of immune cells and stromal cells, respectively. Calculating these scores can help us understand the infiltration of different cell types in the tumor microenvironment and their impact on tumor biology. In

addition, we evaluated the transcriptome data using the CIBERSORT algorithm, analyzing the levels of infiltration of 22 different immune cells in each sample [28, 29]. We also analyzed the correlation between immune cell infiltration levels and risk scores by using the R packages "limma", "reshape2", "ggpubr" and "ggExtra" [30].

2.8 Immunotherapy response and drug sensitivity analysis

The level of immune checkpoint gene expression is closely related to the therapeutic response of immune checkpoint inhibitors. To assess the potential of a prognostic model for predicting immunotherapy response in BRCA patients, we analyzed differences in the expression of immune checkpoint genes between high-risk and low-risk groups. These genes play a crucial role in regulating the immune response. To evaluate the response to immune checkpoint blockade (ICB) in different risk groups, we obtained an immunophenotypic score (IPS) from the TCIA database (https://tcia.at/) [30]. According to the expression of CTLA-4 and PD-1, the IPS in different risk groups was analyzed, and visualized with violin plots. In addition, we studied the responsiveness of different risk populations to commonly used chemotherapy drugs. We used R software package "prorophetic" [31] to calculate the maximum half inhibitory concentration (IC50) of 6 chemotherapy drugs commonly used in clinical treatment of BRCA, including doxorubicin, embelin, gemcitabine, cisplatin, paclitaxel and vinorelbine. The IC50 value is often used as an indicator of drug sensitivity. The IC50 differences between high-risk and low-risk groups were visualized using a box chart.

2.9 Construction of prediction nomogram

To improve the accuracy of the prognostic model, we used the "rms" package to construct a nomogram that included clinical variables (age, gender, TNM stage) and risk score [32]. The nomogram can clearly show the degree of influence of each prognostic factor on the outcome. Then, we used univariate and multivariate cox regression analyses to explore whether risk scores and clinical variables could be used as independent prognostic factors, and plotted forest maps to visualize the results. We not only constructed a nomogram to predict the 1-year, 3-year, and 5-year OS probabilities of the BRCA patients, but also drew a calibration curve to evaluate the nomogram's predictive power. In addition, we used the "timeROC" package for receiver operating characteristics (ROC) analysis to verify that nomogram had the highest accuracy compared to other prognostic factors.

2.10 Identification of cellular types and Pseudo-temporal trajectory analysis at the single-cell level

Single-cell analysis of eight breast cancer samples was performed using Seurat Package. Firstly, we used the Read10X function to read data from each sample and then combine it into a unified data set. To eliminate batch effects, the Harmony algorithm was used to adjust the data according to the sample source (orig.ident). Subsequently, we used TSNE technique for dimensionality reduction to identify different cell subpopulations and manually annotated cell types to accurately determine cell identities [33, 34]. Used the FindAllMarkers function to identify the Marker genes of all annotated cell subpopulations and verified the reliability of cell subpopulation annotations with the CellMarker database (http://117.50.127.228/CellMarker/). In addition, we used t-SNE dimensionality reduction to explore the expression distribution of the 14 genes used to construct the prognostic model in different cell populations, and a violin plot was created to visualize gene expression. Furthermore, in order to study the dynamics of cell changes over time during development, differentiation pathways or biological processes, we performed a pseudo-time trajectory analysis using the Monocle2 package. The "DDRTree" was used for dimensionality reduction and the cells were sequenced to construct developmental trajectories. We also used the plot_genes_jitter function to plot the expression jitter of specific genes at different time points along the pseudo-time locus, and elucidate the temporal change of gene expression. To investigate the function and distribution of prognostic genes in organisms, we collected immunohistochemical staining images from the Human Protein Atlas (HPA) (https://www.proteinatlas.org/) to observe the spatial distribution and expression levels of genes in different tissues and cells.

2.11 Analysis of interactions between cells

The interaction between ligands and cell surface receptors plays a key role in intercellular communication. To investigate the potential communication between different cell populations, we used CellChat to analyze the ligand-receptor interactions between cells [35]. During the analysis, computeCommunProb and filter functions were used to calculate

the probability of intercellular communication, ensure the reliability of the data, exclude communication relationships involving less than three cells, and construct a comprehensive cellular communication network using CellChat. Finally, we showed the number of interactions between different cell populations and their intensity through a circle diagram, and used heat maps to visualize specific signaling pathways to enhance understanding of cellular communication mechanisms.

2.12 Statistical analysis

The statistical analysis in this study was conducted using R software (version 4.3.2). To assess the differences between two groups, the Wilcoxon test was employed. Additionally, to control for false positive discoveries, the FDR (False Discovery Rate) method was utilized to adjust the p-values. For correlation analysis, the Pearson test method was applied unless otherwise specified. A significance level of $p < 0.05$ was set to indicate statistical significance.

3 Result

3.1 Identification of two breast cancer subtypes based on detoxification enzymes related-genes

We performed unsupervised cluster analysis using the TCGA-BRCA dataset to classify BRCA patients into two distinct subtypes based on Detoxification enzymes related-genes expression (Fig. 2A). Using PCA to visualize the data and it showed significant differences between the two distinct subtypes (Fig. 2B). We also performed a K-M survival analysis and found that BRCA patients in Cluster 1 had a higher survival rate than those in Cluster 2 (Fig. 2C). In addition, in order to explore the influence of detoxification enzymes on the mutational spectrum of breast cancer, we conducted mutation analysis for two subtypes. The mutation waterfall map showed that PIK3 CA, TP53, TTN and MUC16 were the most common mutated genes (mutation rate > 10%) in the two clusters (Fig. 2D,E), and the mutation rates of these genes were not significantly different in the two different subtypes. We then performed subtype differential analysis and found 25 differentially expressed genes in the two clusters (Fig. 2F). To further explore the biological functions of these differentially expressed genes (DEGs), we used KEGG, GO, and DO enrichment analyses (Fig. 2G-I). In the biological process, DEGs is mainly involved in the regulation of inflammatory response, leukocyte proliferation and immune response-activation signaling pathway. In terms of cell composition, DEGs was mainly concentrated in the outer part of the plasma membrane. In terms of molecular function, DEGs mainly affects peptidase regulator activity, heparin binding and sulfide binding (Fig. 2G). In terms of action pathway, these genes were mainly concentrated in the cell adhesion molecules and NF-κB signaling pathway (Fig. 2H). In addition, DO enrichment analysis showed that these genes were mainly enriched in primary immunodeficiency diseases (Fig. 2I).

3.2 Construction and verification of prediction model

In previous studies, we have successfully identified two breast cancer subtypes associated with detoxification enzymes-related genes, and the next goal is to model breast cancer prognosis associated with detoxification enzymes-related genes. We obtained BRCA patient data from the TCGA database and randomly divided 1069 BRCA patient data into a training cohort and an internal validation cohort at a ratio of 7:3. We first used univariate Cox regression to analyze the train cohort (N = 749) and identified 39 detoxification enzymes-related genes associated with breast cancer prognosis. Through Lasso regression and multivariate Cox regression analysis, we further screened 14 genes for the construction of prognostic models and used these genes to establish risk score formula (Fig. 3A–C). The calculation formula is: Risk Score = CYP27 A1* (− 0.153773043733361) + IFNG* (− 1.13029577293788) + SERPINA1* (− 0.130967963213151) + ALDH1 A1* (− 0.160501405610823) + CPT1 A* (0.25451256806553) + PLAT* (− 0.229065565958707) + GSTM2* (0.290013288958423) + ATP6 AP1* (0.387236315911989) + SLC35 A2* (0.595324950705634) + GSTM4* (− 0.398966850228594) + NFKBIA* (− 0.37157076126092) + POMGNT2* (− 0.406833583296618) + NT5E* (0.347167101015933) + SQLE* (− 0.274411899158172). Each patient's risk score was calculated according to the formula, and they were divided into low-risk and high-risk groups according to the median risk score value. As can be seen from the risk score distribution map, the survival scatter plot and the K-M curve both survival probability and survival time of the low-risk group were higher than those of the high-risk group in the four cohorts (Fig. 3D-F). Notably, in the training cohort (N = 749), overall survival was significantly higher in the low-risk group (N = 374) than in the high-risk group (N = 375). To evaluate the

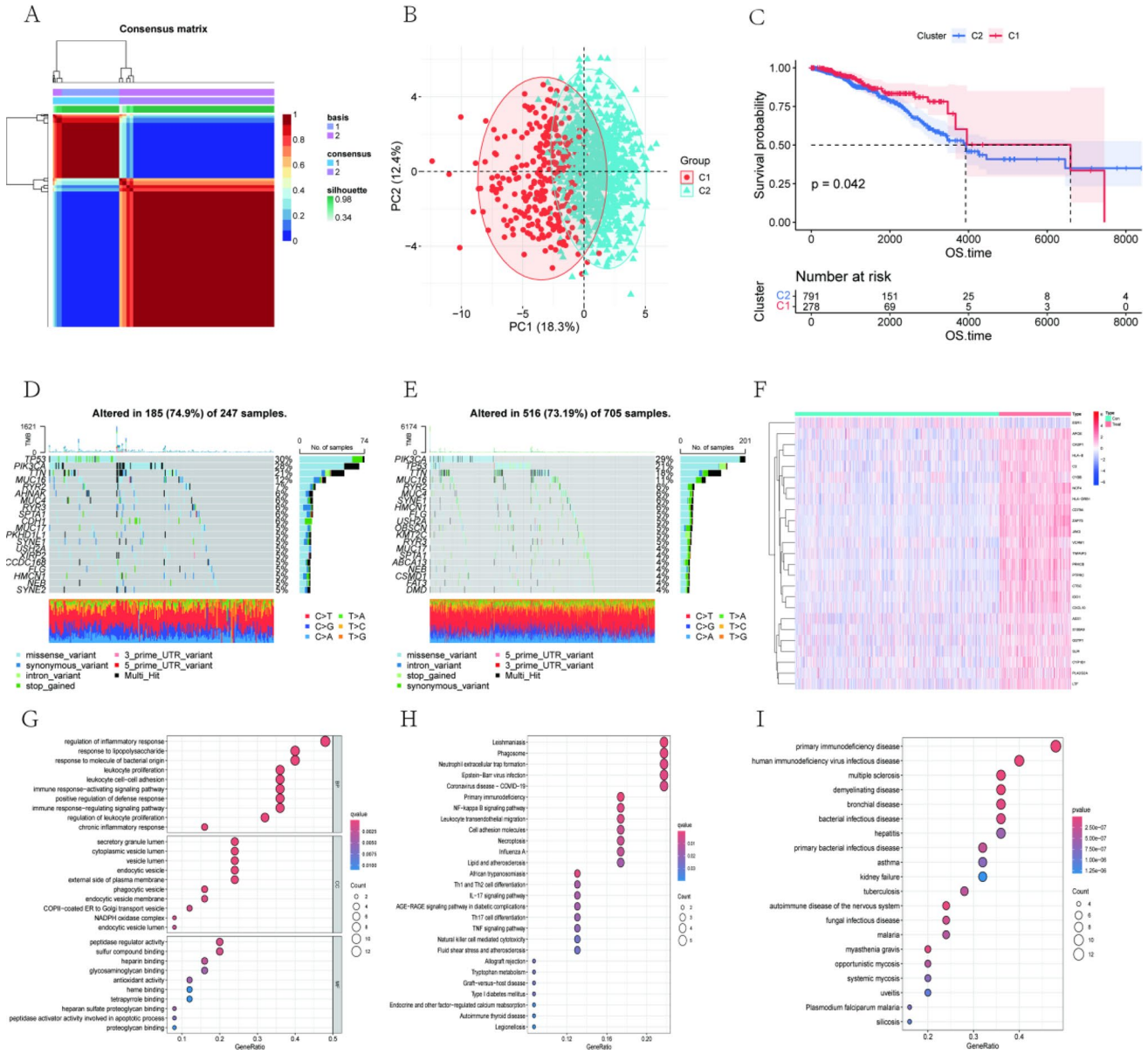

Fig. 2 Subtypes of BRCA and their characteristics. **A** Based on the expression of detoxification enzymes-related genes, the BRCA subtypes was constructed by NMF algorithm. **B** Data were visualized using PCA. **C** K-M survival analysis of OS for the two subtypes. **D, E** The mutation waterfall plots of two subtypes. **F** Differential expression profiling between two subtypes. **G** GO enrichment analysis of DEGs. **H** KEGG enrichment analysis of DEGs. **I** DO enrichment analysis of DEGs

predictive performance of the prognostic model, we performed a time-dependent ROC curve analysis and obtained encouraging results in the TCGA train cohort with auc of 0.695 (1 year), 0.777 (3 years), and 0.752 (5 years), respectively (Fig. 3G). Good results were also observed in the TCGA test cohort, with auc of 0.632 (1 year), 0.637 (3 years), and 0.605 (5 years), respectively. In addition, these good results were confirmed in the TCGA all cohort and in the GEO-GSE20685 cohort, indicating that the detoxification enzymes-related genes-based breast cancer prognosis model has a good predictive performance.

3.3 DNA promoter methylation of core genes

Promoter DNA methylation affects transcriptional inhibition and tumorigenesis [36], so we explored the methylation values of core genes in this model in both normal and BRCA tissues. The results showed that promoter methylation

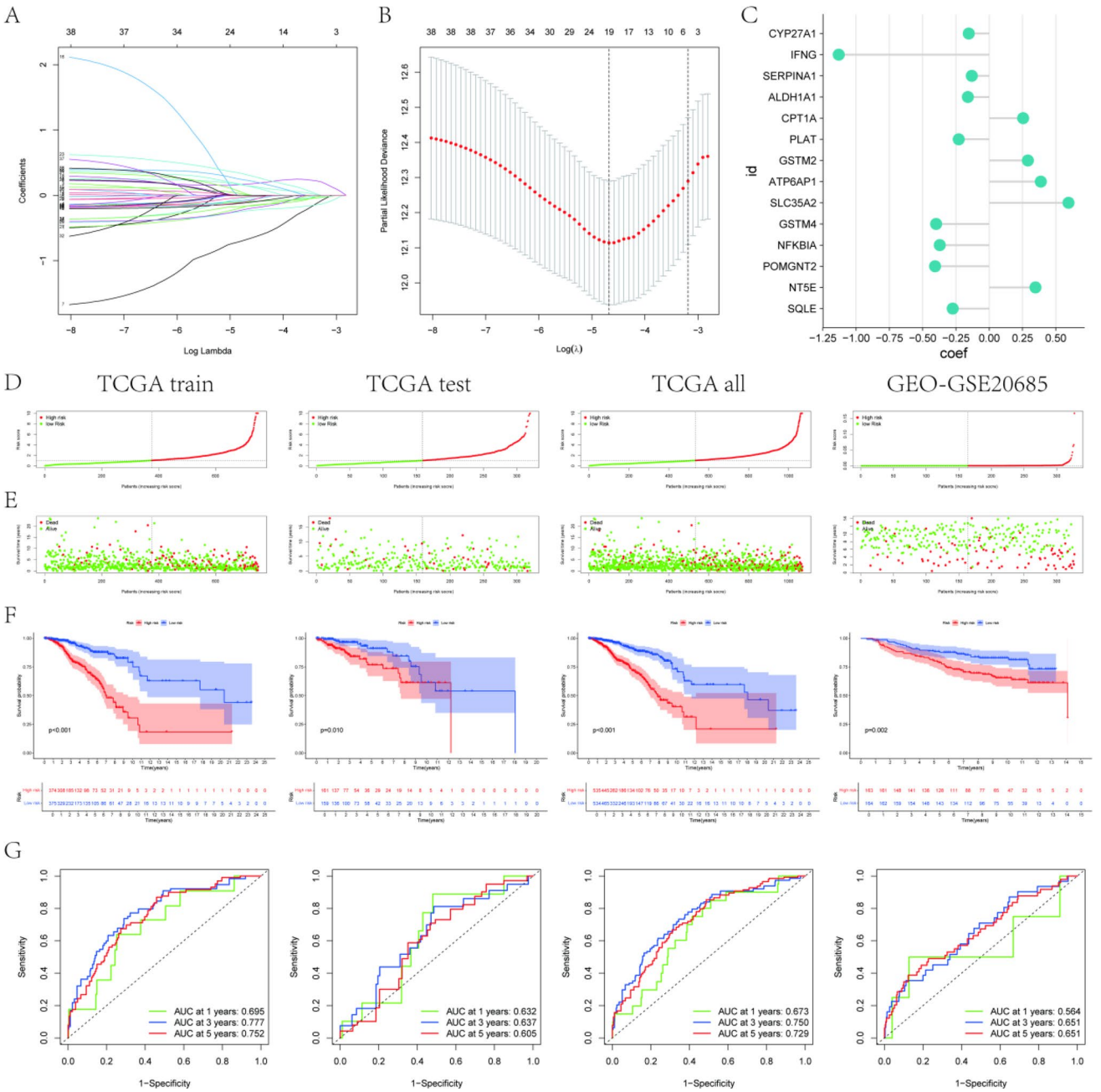

Fig. 3 Construction and validation of prognostic models. **A, B** Lasso regression screened the genes of the prognostic model. **C** Prognostic genes screened by multivariate COX regression analysis. **D** The risk score distribution plot of four cohorts. **E** The survival status diagram of four cohorts. **F** K-M survival analysis. **G** Time-related ROC curve analysis

levels of ALDH1 A1, ATP6 AP1, IFNG, NFKBIA, SERPINA1 and SLC35 A2 were significantly reduced in tumor tissues. The promoter methylation levels of CPT1 A, SQLE, GSTM2, GSTM4, CYP27 A1, PLAT and NT5E were significantly increased in tumor tissues (Fig. 4).

3.4 GSVA, GSEA analysis of high and low risk groups

To investigate the biological differences between the high-risk and low-risk groups identified by the model, we conducted Gene Set Variation Analysis (GSVA) and Gene Set Enrichment Analysis (GSEA) using all datasets. GSVA results showed significant differences in biological processes between high-risk and low-risk groups. Signaling pathways were abundant in

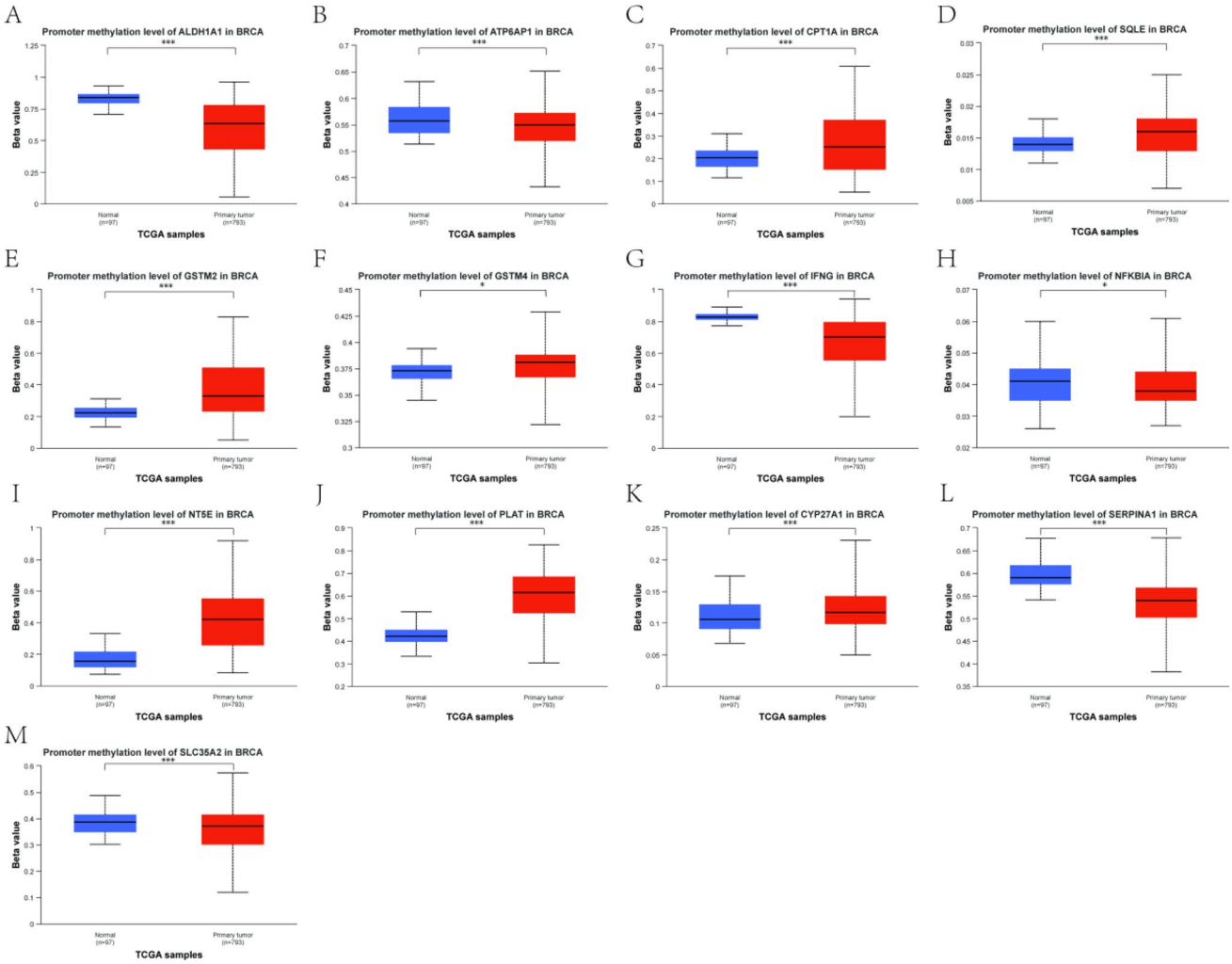

Fig. 4 Promoter methylation of core genes. Boxplots visualized the methylation levels of ALDH1 A1 (**A**), ATP6 AP1 (**B**), CPT1 A (**C**), SQLE (**D**), GSTM2 (**E**), GSTM4 (**F**), IFNG (**G**), NFKBIA (**H**), NT5E (**I**), PLAT (**J**), CYP27 A1 (**K**), SERPINA1 (**L**) and SLC35 A2 (**M**) in normal and BRCA tissues

the low-risk group, including cytosolic DNA sensing pathway, apoptosis, B cell receptor signaling pathway, T cell receptor signaling pathway and cytokine-cytokine receptor interaction (Fig. 5A). GSEA results showed that cell cycle, homologous recombination and oocyte meiosis pathway were enriched in the high-risk group (Fig. 5B), while chemokine signaling pathway, cytokine-cytokine receptor interaction, and T cell receptor signaling pathways were significantly enriched in the low-risk group (Fig. 5C). These results suggest that the link with immunity may be stronger in the low-risk group.

3.5 Immune landscape of high and low risk groups

To further investigate the relationship between high and low risk groups and immunity, we looked at features such as tumor microenvironment (TME) and immune infiltration associated with the immune landscape. The results showed that compared with the high-risk group, the ESTIMATEScore, ImmuneScore and StromalScore of BRCA patients in the low-risk group were significantly higher than the high-risk group, while the TumorPurity result was opposite (Fig. 6A,B), indicating that the content of stromal cells and immune cells in the TME of the low-risk group was higher than that of tumor cells. In addition, we investigated the relationship between different risk groups and immune cell infiltration. We found that the infiltration level of naive B cells, plasma cells, CD8 T cells, activated NK cells, resting dendritic cells and activated dendritic cells was high in the low-risk group, while the infiltration level of M0 macrophages and M2 macrophages was completely opposite. We also found that the infiltration levels of naive B cells, memory B cells, resting dendritic cells, plasma cells, activated NK cells, CD4 activated memory T cells, CD8 T cells, and γδ T cells were negatively correlated with

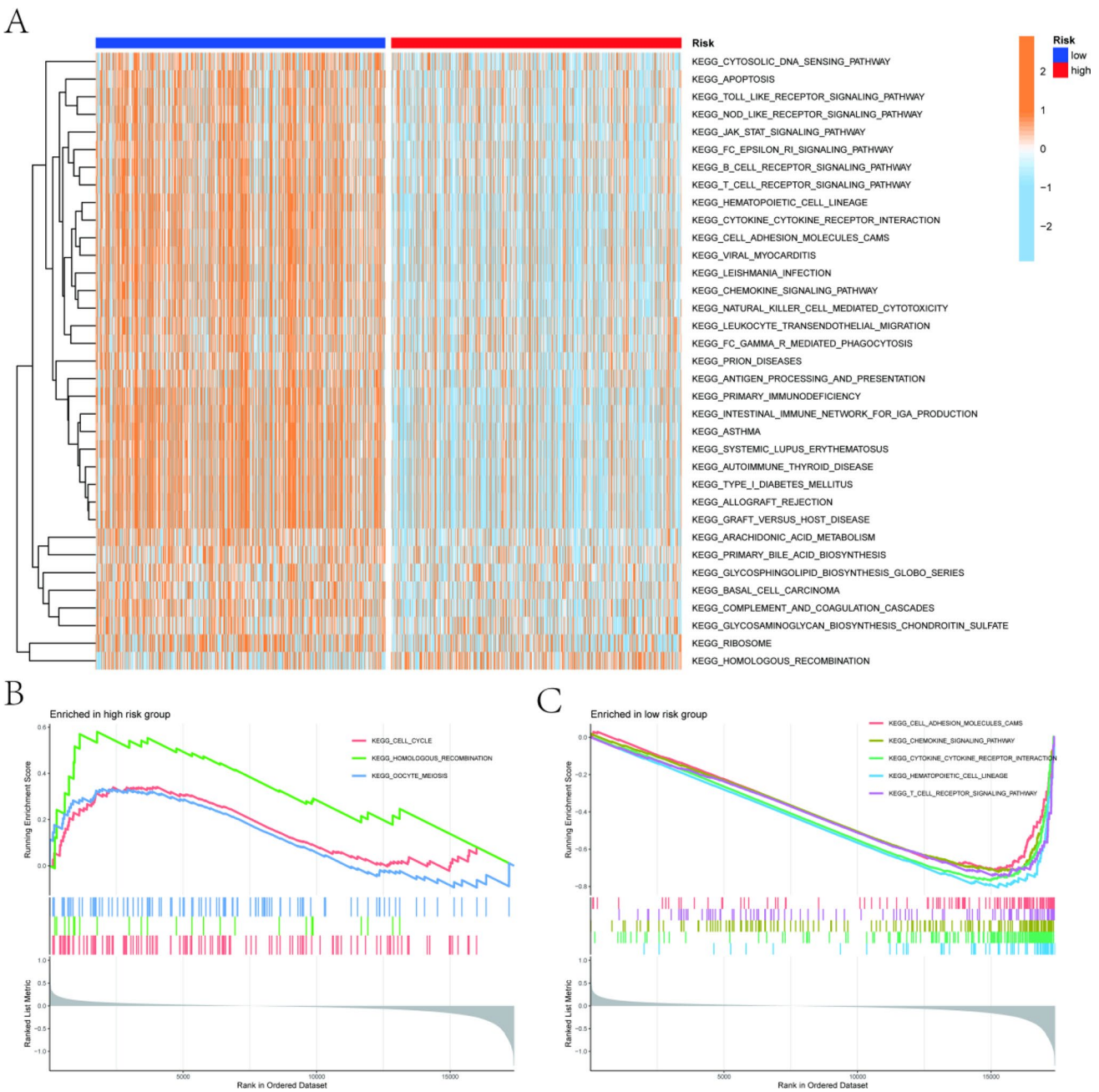

Fig. 5 GSEA and GSVA analysis of two groups. **A** GSVA heatmap of two groups. **B** GSEA enrichment analysis in high risk group. **C** GSEA enrichment analysis in low risk group

the risk score (Fig. 6C). These results suggest that the risk score of the prognostic model is closely related to immune cells, and the lower the risk score, the higher the infiltration of stromal cells and immune cells in TME.

3.6 Relationship between prognostic model and immunotherapy

To further evaluate the potential of a prognostic model for predicting immunotherapy response in BRCA patients, we analyzed the expression of immune checkpoint genes between high-risk and low-risk groups. We found that immune checkpoint gene expression was significantly higher in the low-risk group (Fig. 7A), suggesting that patients in the low-risk group may be more sensitive to immune checkpoint inhibitors. We then used IPS as indicators to evaluate the effectiveness of immune checkpoint inhibitors, and further evaluated the potential clinical efficacy of immune

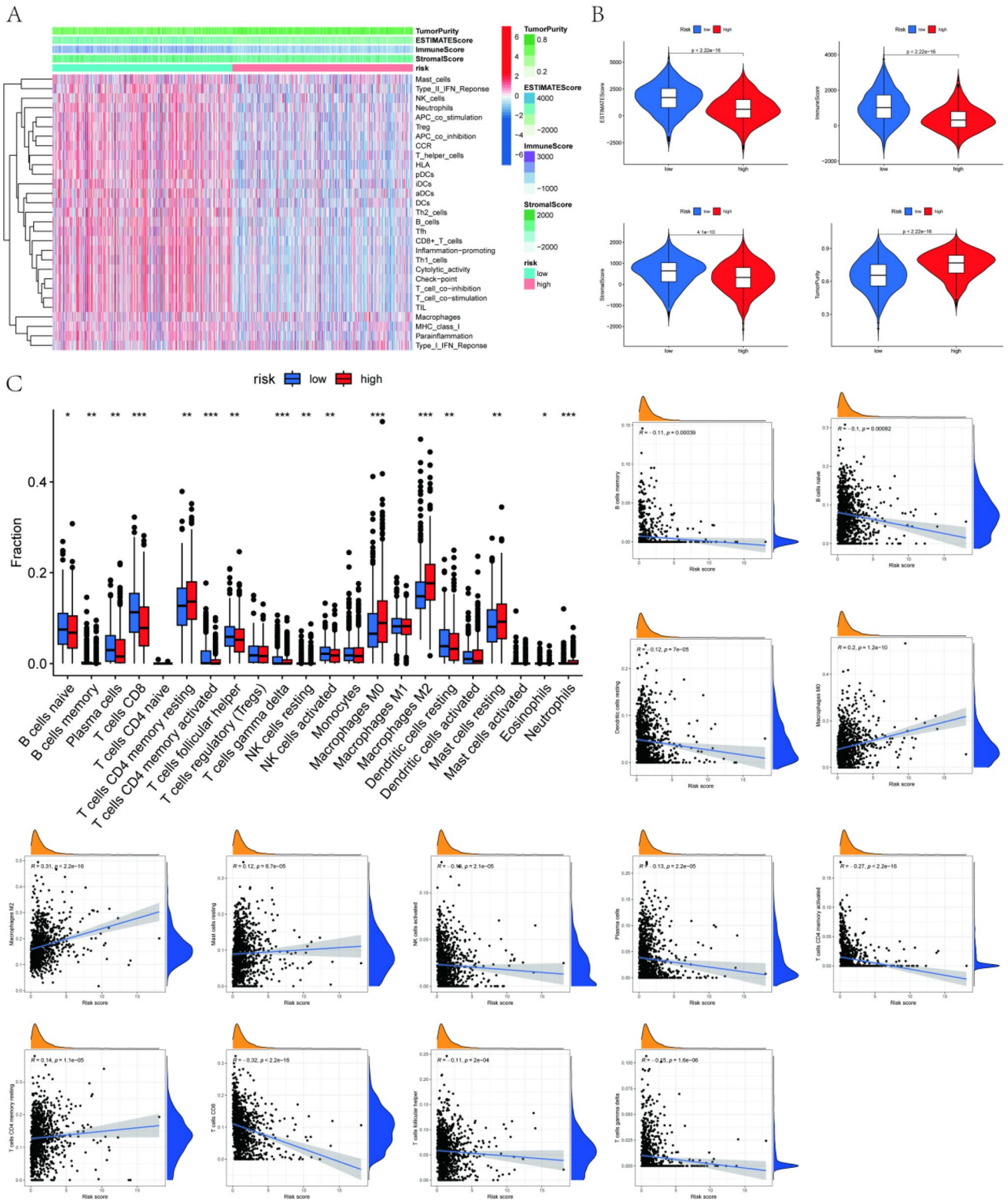

Fig. 6 Immune landscape analysis of TME and immune infiltration. **A** Heat map showing the overall immune landscape in the two risk group. **B** Differential analysis of TME between two risk groups. **C** Differential analysis of immune infiltration cells between two risk groups and correlation analysis between risk score and immune infiltration cells

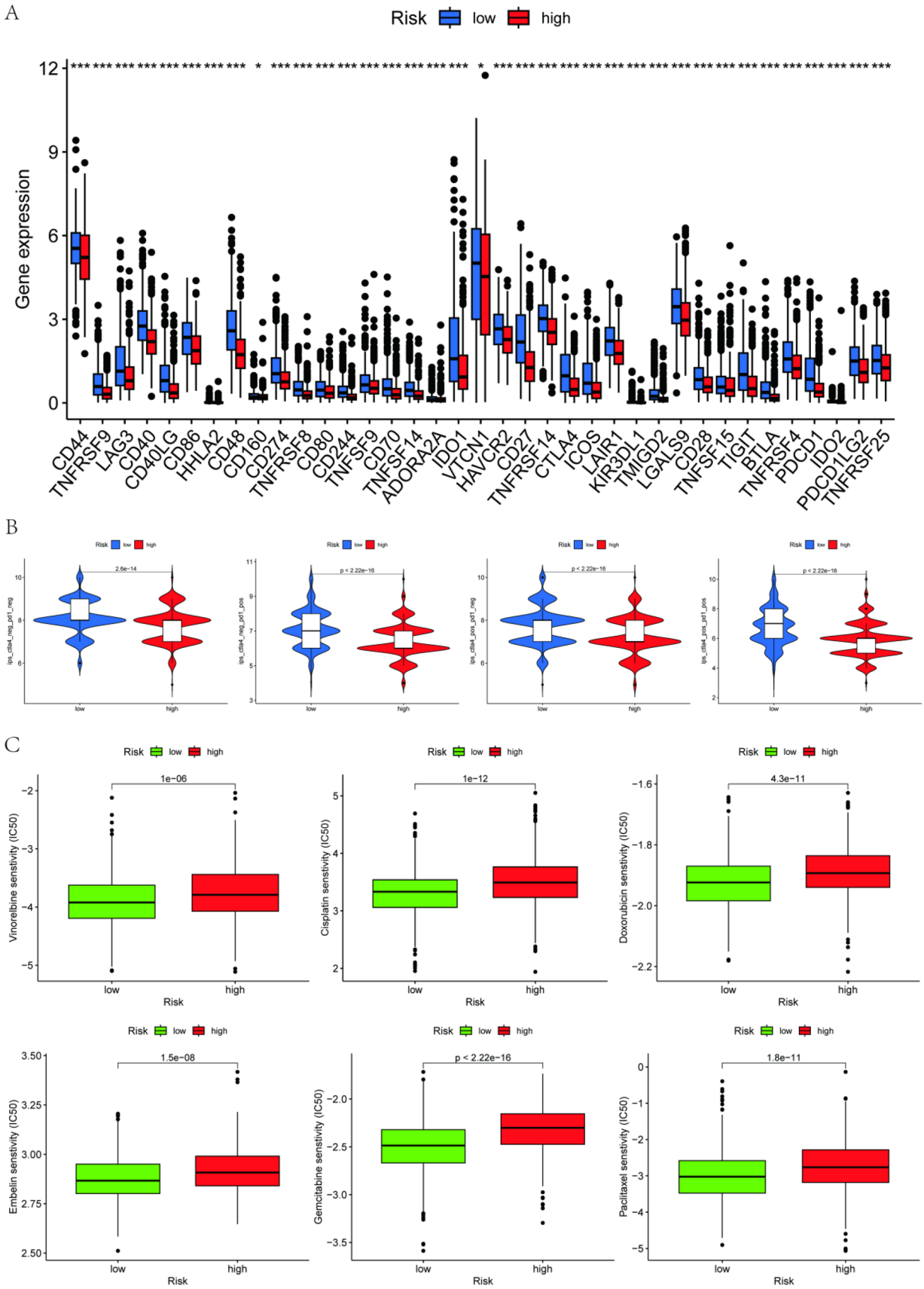

Fig. 7 Drug-therapy Prediction. **A** Expression of immune checkpoint genes in high and low risk groups. **B** IPS in high and low risk groups. **C** Drug sensitivity tests of vinorelbine, cisplatin, doxorubicin, paclitaxel, gemcitabine and embelin

checkpoint inhibitors in different risk populations. The violin plot results showed that the IPS score of the low-risk group was higher than that of the high-risk group (Fig. 7B). These results showed that patients in the low-risk group responded better to immunotherapy than those in the high-risk group. In addition, we also analyzed the relationship between the prognostic model and the commonly used chemotherapy drugs for breast cancer treatment, using the IC50 value of chemotherapy drugs as the evaluation index. The study results showed that the IC50 values of vinorelbine, cisplatin, doxorubicin, paclitaxel, gemcitabine and embelin in the low-risk group were lower than the high-risk group (Fig. 7C). This suggests that the low-risk group is more sensitive to these commonly used chemotherapy agents, and using these agents may be better for patients in the low-risk group, with a lower likelihood of resistance. Our findings suggest that the low-risk group may respond better to both immunotherapy and chemotherapy, which could have important clinical implications.

3.7 Establishment and verification of nomogram

To investigate the independent predictive power of riskscore as prognostic factors, univariate and multifactor Cox regression analyses were performed. The results showed that riskscore, age, gender, and TNM stage of breast cancer patients were significantly correlated with prognosis, and riskscore and age could be used as independent prognostic factors of breast cancer patients (Fig. 8A,B). We then combined clinicopathological features (including gender, TNM stage, and age) and riskscore to establish nomogram to predict survival in BRCA patients. The riskscore and clinical variables were combined to calculate the total score for each patient, and the higher the score, the lower the survival rate (Fig. 8C). In addition, we evaluated the accuracy of the nomogram by using the calibration curve and the area under the ROC curve. The calibration curves for 1-year, 3-year, and 5-year OS prediction showed that the predictive nomogram performed well (Fig. 8D). The 1-year, 3-year, and 5-year nomogram auc of the TCGA cohort were 0.847, 0.776 and 0.773, respectively which were significantly higher than the other parameters (Fig. 8E). This result suggests that the nomogram had better predictive accuracy than other clinical features and the original risk score.

3.8 Web tool

We have successfully integrated the prognostic model into a web-based application featuring customized algorithms designed to predict overall survival (OS) and drug therapy responses in BRCA patients based on personalized characteristics (Fig. 9) (http://wys.helyly.top/zjd/cox.html). Initially, the expression levels of detoxification enzymes-related genes are input into the application to generate a risk prediction for BRCA patients. This prediction is then combined with clinical variables, including the patient's age, gender, and TNM stage, to forecast the patient's drug therapy response and OS at 1, 3, and 5 years, which are displayed on the web interface. The tool is designed to be intuitive and easy to use, offering valuable support for individualized prognosis prediction in BRCA patients.

3.9 Cellular constitution of breast cancer and developmental trajectories of epithelial cells (cancer cells)

In this study, we analyzed eight single-cell samples from breast tumors in the GSE161529 dataset and grouped them into 23 distinct clusters (Fig. 10A,B). According to the different marker genes expressed by each cluster (Fig. 10C,D), we manually annotated these clusters into seven different cell populations, including B cells, endothelial cells, epithelial cells (cancer cells), fibroblasts, macrophages, monocytes and T cells (Fig. 10E). Subsequently, we analyzed the expression of 14 prognostic genes in breast cancer epithelial cells (cancer cells) at the single-cell level (supplementary Fig. 1) and found that SQLE was specifically highly expressed in the epithelial cells (cancer cells) of breast cancer (Fig. 10F,G). This observation highlights the important role of SQLE in tumor cells, suggesting that it may be involved in specific biological processes within the epithelial cells or the tumor microenvironment. In addition, we mapped the developmental trajectories of breast cancer epithelial cells through single-cell sequencing analysis and examined the expression patterns of SQLE in these trajectories (Fig. 10H,I). The result showed that the expression of SQLE in breast cancer varied with the stage of epithelial cell development (Fig. 10J). Immunohistochemical staining images showed that SQLE protein levels were higher in breast cancer tumor tissue than in normal tissue (Fig. 10K,L). This suggests that significant upregulation of SQLE in breast cancer may influence tumorigenesis and cancer progression.

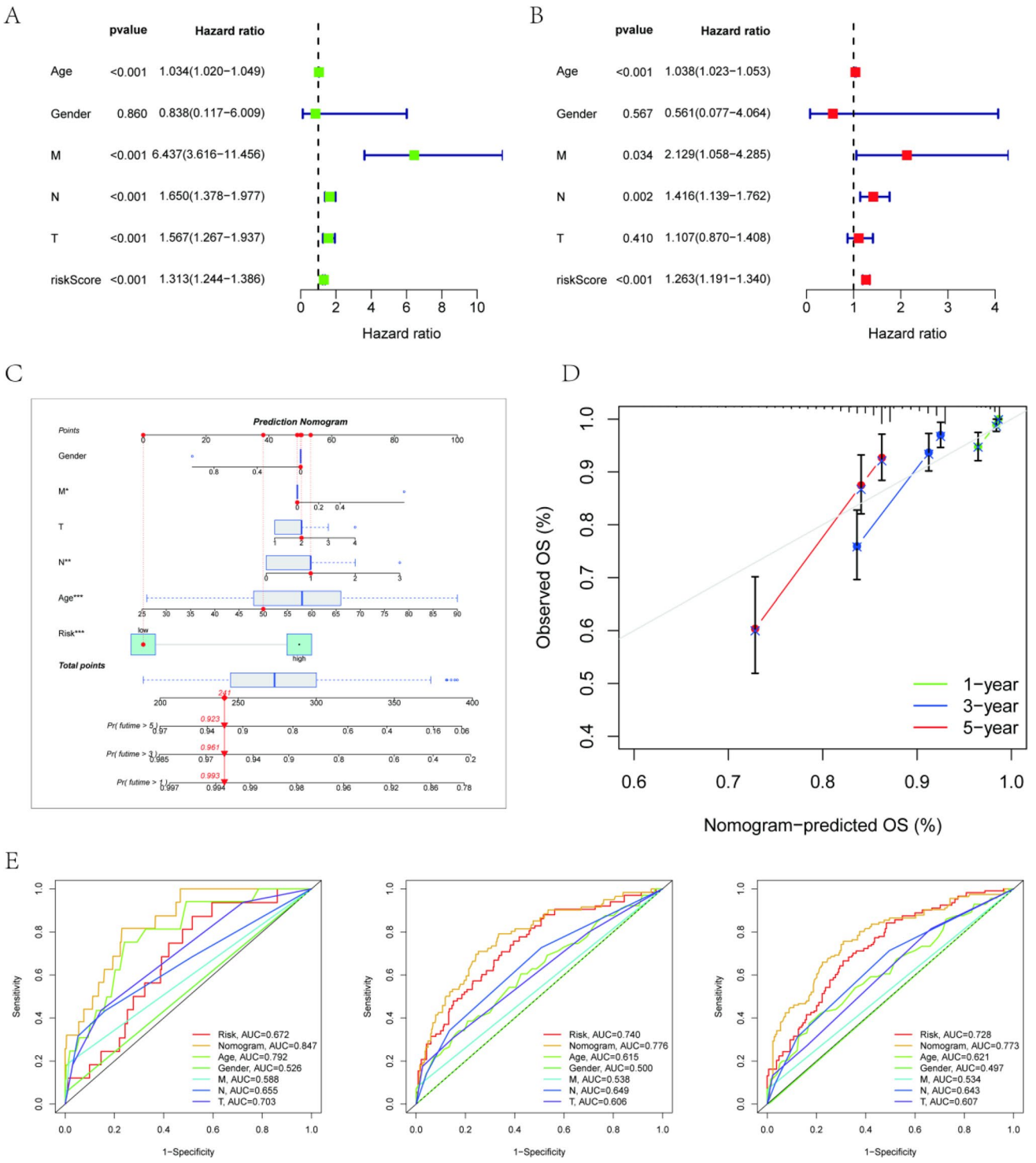

Fig. 8 Nomogram for predicting OS. **A** Univariate Cox regression analysis was used to analyze the predictive ability of age, TNM stage and riskscore. **B** Multivariate Cox regression analysis was used to analyze the predictive ability of age, TNM stage and riskscore. **C** The predictive nomogram combining clinicopathological features and risk score. When the total point was 241, the 1-year, 3-year and 5-year survival rates were 0.993, 0.961 and 0.923, respectively. **D** Calibration curve of the predictive nomogram. **E** ROC curves for 1, 3 and 5 years of the nomogram

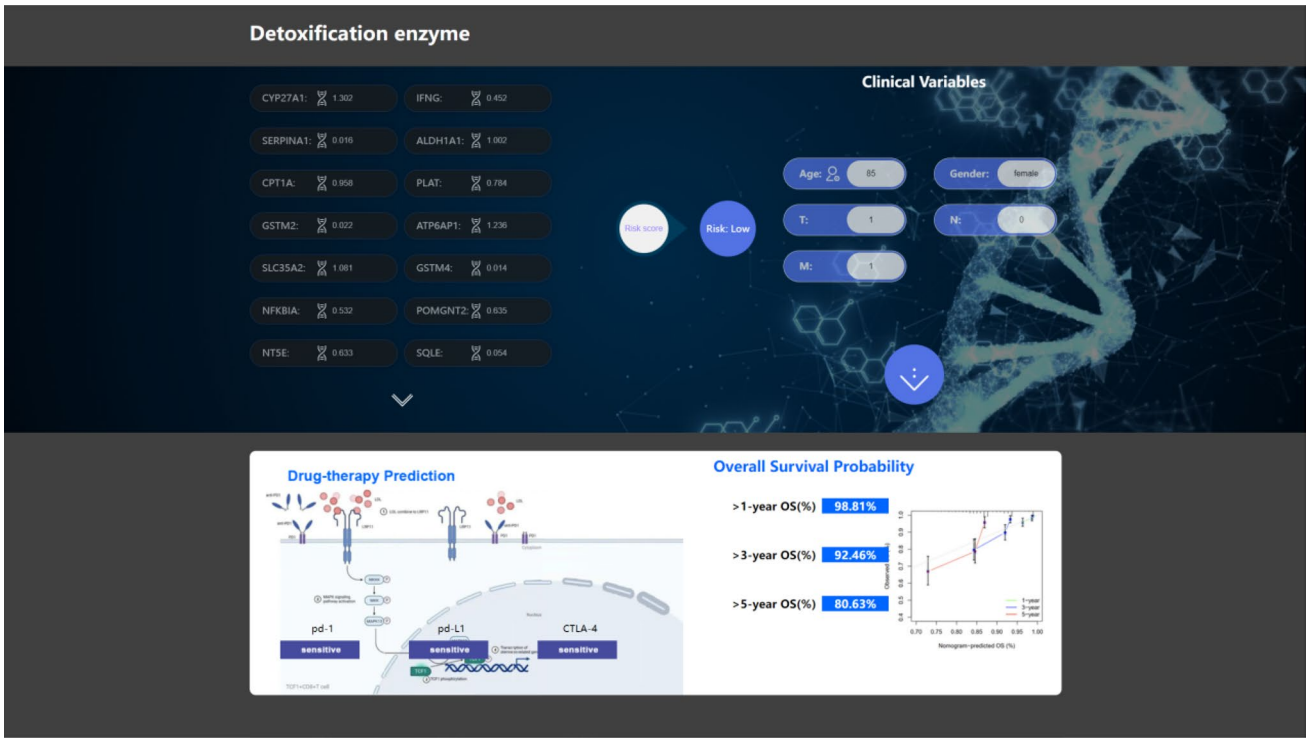

Fig.9 Webpage interface of the prediction tool. Allow users to input patient-specific data, including gene expression related to detoxification enzymes, as well as clinical variables such as age, gender, and TNM staging, through the left panel. In turn, the system generates predicted outcomes on the right panel, providing estimates for the patient's response to chemotherapy and immunotherapy, along with the 1-year, 3-year, and 5-year overall survival (OS) probabilities

3.9.1 Interaction between SQLE + epithelial cells (cancer cells) and exhausted CD8 + T cells

In this study, we analyzed the interactions between different cells in breast cancer and found that both the number and intensity of the interaction between epithelial cells (cancer cells) and T cells were stronger (Fig. 11A,B). Subsequently, We divided epithelial cells (cancer cells) into SQLE + and SQLE − epithelial cells and analyzed their interactions with different types of T cells (Fig. 11C,D). We found that compared with SQLE-epithelial cells, SQLE + epithelial cells had a stronger effect on different types of T cells, especially exhausted CD8 + T cells (Fig. 11E,F). Subsequently, we investigated the pathways by which the two interact and found that in breast cancer, SQLE + epithelial cells affect exhausted CD8 + T cells mainly through the MIF signaling pathway (Fig. 11G,H). The result suggests that SQLE may affect T cells primarily by affecting exhausted CD8 + T cells, further influencing cancer progression.

4 Discussion

This study investigated the potential role of detoxification enzymes in breast cancer by constructing a prognostic model based on the genes associated with detoxification enzymes, and verified the effectiveness of this model in predicting the prognosis of breast cancer patients. Breast cancer is one of the most common malignancies in women worldwide, and despite significant advances in treatment methods, its prognosis still faces great challenges. Therefore, establishing a model that can accurately predict the prognosis of breast cancer patients is of great significance for improving the accuracy of clinical management and implementing personalized treatment.

The role of detoxification enzymes in the development and progression of breast cancer has gradually attracted the attention of researchers. The detoxification enzymes family includes P-oxidases cytochrome (CYP) enzymes, glutathione S-transferase (GST), UDP-glucuronide transferase (UGT), sulfotransferase (SULT) and n-acetyltransferase (NAT), which are responsible for the metabolism and clearance of harmful substances in the body. It can maintain the chemical balance of the body by transforming and removing environmental toxins, drugs and metabolites

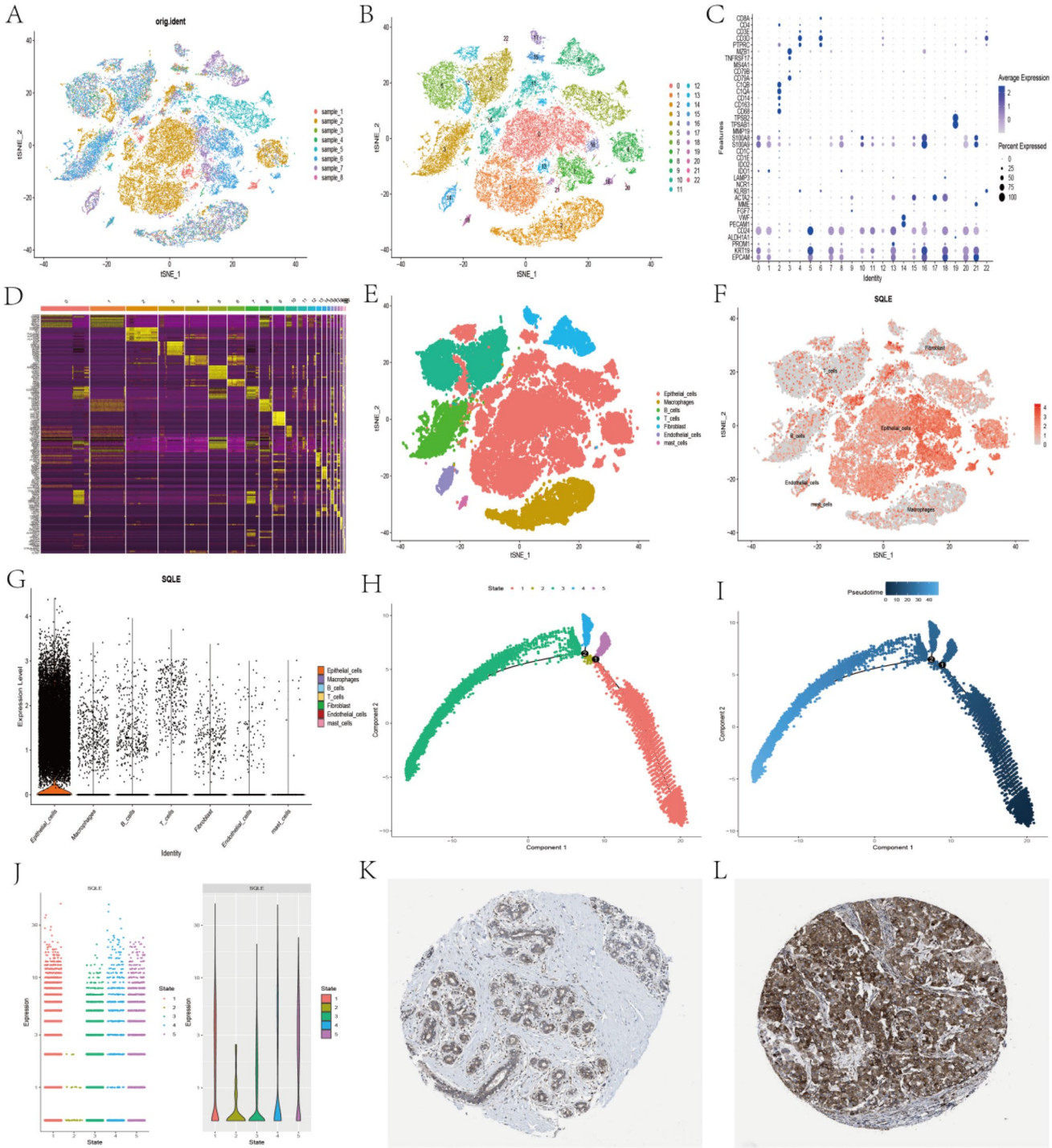

Fig.10 Single-cell atlas of breast cancer and immunohistochemical images. **A** t-SNE plot of eight sample sets. **B** t-SNE plot of 23 clusters. **C** The bubble map showed the expression of marker genes in 23 different cell clusters. **D** The heat map shows the 10 most highly expressed genes in each cell cluster. **E** t-SNE plot of annotating 7 distinct cell types in single-cell RNA sequencing. **F** t-SNE plot of SQLE expression in different cell populations. **G** Violin plot shows the expression of SQLE in different cell populations. **H** Different developmental stages of breast cancer epithelial cells. **I** Pseudo-temporal trajectory analysis of breast cancer epithelial cells. **J** Dynamic expression of SQLE along pseudo-time. **K** Expression of SQLE in normal tissues. **L** Expression of SQLE in breast cancer, Staining: High; Intensity: strong; Quantity: > 75%

[6, 7]. In breast cancer, detoxification enzymes are closely linked to processes such as antioxidant defense, drug response, and cell apoptosis [8, 9]. Furthermore, genetic polymorphisms in certain detoxification enzymes may influence an individual's ability to metabolize carcinogens, thereby increasing the risk of developing breast cancer

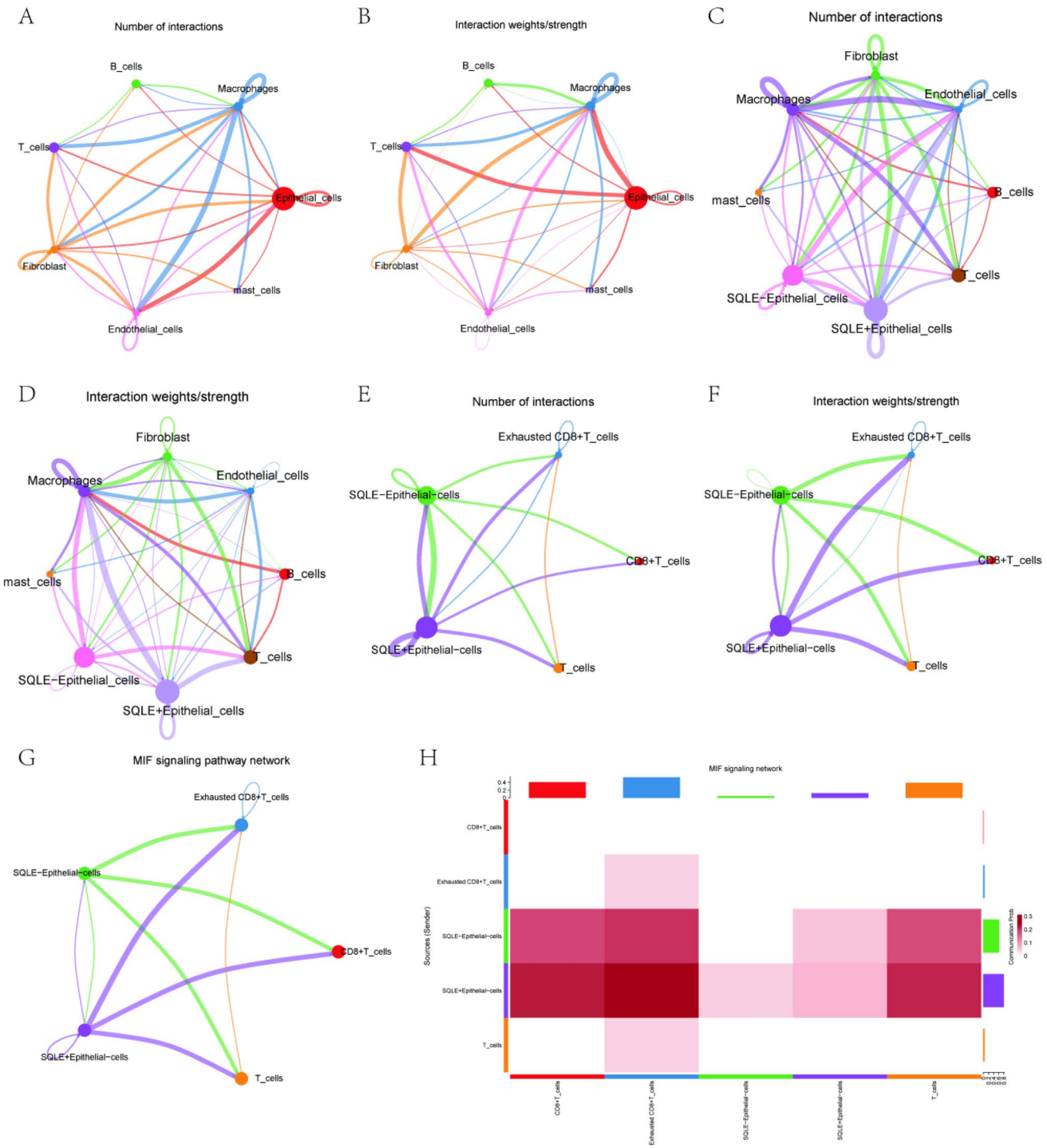

Fig. 11 Analysis of intercellular communication. **A** The number of interactions in intercellular communication networks between 7 different cell type. **B** The interaction weights in intercellular communication networks between 7 different cell types. **C** The number of interactions in intercellular communication networks between SQLE +, SQLE-epithelial cells and different cell subtypes. **D** The interaction weights in intercellular communication networks between SQLE +, SQLE-epithelial cells and different cell subtypes. **E** The number of interactions in intercellular communication networks between SQLE +, SQLE-epithelial cells and different T-cell subtypes. **F** The interaction weights in intercellular communication networks between SQLE +, SQLE − epithelial cells and different T-cell subtypes. **G** MIF signaling pathway network of SQLE +, SQLE − epithelial cells and different T-cell subtypes. **H** Heatmap depicting ligand-receptor interaction patterns in the MIF signaling pathway between SQLE +, SQLE − epithelial cells and different T-cell subtypes

[10]. For example, the deletion of the GSTM1 gene has been found to be strongly associated with the occurrence of breast cancer [10], suggesting that variations in detoxification enzymes related-genes may represent potential risk factors for the disease.

In constructing the prognostic model, we employed Lasso Cox regression analysis, univariate and multivariate Cox regression analysis, in conjunction with machine learning algorithms. Validation through the TCGA and GEO databases revealed that the model effectively distinguishes between high-risk and low-risk patient groups, with significant biological differences observed between these groups, including characteristic pathways, immune infiltration, immune therapy response, and drug sensitivity. These results suggest that the detoxification enzymes-related genes-based prognostic model not only predicts patient survival but also provides guidance for personalized treatment decisions, thereby optimizing treatment strategies and improving the quality of life and cure rates for breast cancer patients.

Additionally, this study revealed the potential role of the detoxification enzymes gene SQLE in immune evasion in breast cancer through single-cell analysis. Squalene epoxidase (SQLE) is an enzyme involved in cholesterol biosynthesis, and studies have shown that it can influence tumor cell division and proliferation by regulating the stability of CCNB1 (cyclin B1), thereby promoting breast cancer progression [37]. Our study found that SQLE is highly expressed in breast cancer epithelial cells and can affect CD8 + T cell exhaustion via the MIF signaling pathway, potentially providing a mechanistic basis for immune evasion in breast cancer. Immune evasion is a critical strategy by which tumor cells escape host immune surveillance, and CD8 + T cells are one of the key effector cells in anti-tumor immune responses. Therefore, the role of detoxification enzymes-related genes in modulating immune responses warrants further investigation.

The roles of other detoxification enzymes-related genes, such as CYP27 A1, SERPINA1, ALDH1 A1, CPT1 A, PLAT, ATP6 AP1, and SLC35 A2, in breast cancer also warrant attention. CYP27 A1, a member of the cytochrome P450 enzyme family, is primarily involved in cholesterol metabolism. Studies have shown that high expression of CYP27 A1 is associated with poor prognosis in certain breast cancer subtypes, and its metabolic product, 27-hydroxycholesterol, can modulate the tumor microenvironment by influencing the infiltration and function of immune cells, thereby affecting tumor progression and patient prognosis [38]. Similarly, SERPINA1, a gene that plays a critical role in regulating inflammation and protecting tissues, is closely associated with estrogen in breast cancer cells and acts as a direct target gene of estrogen receptors. High expression of SERPINA1 may correlate with the survival rate of breast cancer patients [39]. Aldehyde dehydrogenase 1 A1 (ALDH1 A1), a marker of tumor-initiating cells, promotes tumor growth and metastasis. Its activity in breast cancer enhances immune suppression and promotes immune evasion, thus contributing to tumor progression [40]. Therefore, ALDH1 A1 is considered a promising target for breast cancer therapy, especially in strategies targeting the tumor microenvironment. CPT1 A, a key enzyme in fatty acid metabolism, may regulate tumor cell growth and the tumor microenvironment by influencing fatty acid metabolism [41]. PLAT, a tissue-type plasminogen activator, is involved in fibrinolysis and promotes thrombus dissolution. In the tumor microenvironment, PLAT regulates cell migration and invasion. Low expression of PLAT has been linked to poor prognosis and increased immune infiltration [42]. ATP6 AP1, a gene encoding a protein involved in cellular acid–base balance and signal transduction, plays roles in immune response and cell proliferation. Upregulation of ATP6 AP1 in breast cancer tissues is associated with poorer prognosis [43]. SLC35 A2, a transporter protein, may influence breast cancer cell proliferation and migration, with some studies suggesting that its expression level correlates with patient prognosis [44]. In summary, these genes and proteins play crucial roles in the onset and progression of breast cancer and may offer new targets and strategies for early diagnosis, prognosis assessment, and treatment of breast cancer.

In the future, modulating these detoxification enzymes-related genes may provide new strategies for immunotherapy of breast cancer, especially when treating those subtypes of breast cancer with poor immune response. However, although this study has made significant progress in the construction of a breast cancer prognosis model for detoxification enzymes-related genes, there are still some limitations. First, although we validated the validity of the model in TCGA and GEO datasets, these data are mainly from publicly available databases, which may have some sample bias. Therefore, the reliability of the model needs to be further verified by multi-center and large-scale clinical data in the future. Second, although single-cell analyses provide a preliminary relationship between SQLE and immune escape, the specific mechanisms are still not fully understood. Future studies need to further explore the interaction between detoxification enzymes-related genes and immune escape through more experimental means. In addition, this study mainly focuses on the analysis of gene expression levels, and more omics data, such as proteomics and metabolomics, can be combined in the future to further explore the multidimensional role of detoxification enzymes in breast cancer.

5 Conclusion

In conclusion, this study provides a new perspective for the construction of breast cancer prognosis model using detoxification enzymes-related genes. This model not only enhances our ability to predict the prognosis of breast cancer, but also provides a valuable reference for the development of individualized treatment and prevention strategies.

Acknowledgements We would like to express our heartfelt gratitude to the TCGA database and GEO database for providing valuable resources for data analysis. We also thank Figdraw platform provides part of the drawing material.

Author contributions TLZ, JDZand WDZ conceived and designed the study.JDZ, WDZ, HXZ, BH, TY, SYW and HHH conducted data analysis. JDZ, WDZ and HXZ drafted the manuscript. TLZ, ZL and ZMZ strictly revised the manuscript. The final manuscript was reviewed and approved by all authors.

Funding The present study was supported by the National Natural Science Foundation of China (Grant number 82473965 to TLZ), the Natural Science Foundation of Hunan Province (Grant number 2023 JJ50156 to TLZ), the Clinical Research 4310 Program of the First Affiliated Hospital of the University of South China (Grant number 20224310 NHYCG04), Science and technology innovation Program of Hengyang City (Grant number 202250045223) and the project of Hunan health high-level youth backbone talents to Z.T.L.

Data availability All raw data used in this study are available in the UCSC xenabrowser (https://xenabrowser.net/) and GEO database (https://www.ncbi.nlm.nih.gov/geo/).

Declarations

Competing interests The authors declare no competing interests.

References

1. Bray F, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
2. Lukasiewicz S, et al. Breast cancer-epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-an updated review. Cancers (Basel). 2021;13(17):4287.
3. Ye F, et al. Advancements in clinical aspects of targeted therapy and immunotherapy in breast cancer. Mol Cancer. 2023;22(1):105.
4. Aiyappa-Maudsley R, et al. CYP2S1 and CYP2W1 expression is associated with patient survival in breast cancer. J Pathol Clin Res. 2022;8(6):550–66.
5. Sneha S, et al. Intratumoural cytochrome P450 expression in breast cancer: impact on standard of care treatment and new efforts to develop tumour-selective therapies. Biomedicines. 2021;9(3):290.
6. Rowland A, Miners JO, Mackenzie PI. The UDP-glucuronosyltransferases: their role in drug metabolism and detoxification. Int J Biochem Cell Biol. 2013;45(6):1121–32.
7. Zhao M, et al. Cytochrome P450 enzymes and drug metabolism in humans. Int J Mol Sci. 2021;22(23):12808.
8. Silva M, Carvalho MDG. Detoxification enzymes: cellular metabolism and susceptibility to various diseases. Rev Assoc Med Bras. 2018;64(4):307–10.
9. Lo HW, Ali-Osman F. Genetic polymorphism and function of glutathione S-transferases in tumor drug resistance. Curr Opin Pharmacol. 2007;7(4):367–74.
10. Possuelo LG, et al. Polymorphisms of GSTM1 and GSTT1 genes in breast cancer susceptibility: a case-control study. Rev Bras Ginecol Obstet. 2013;35(12):569–74.
11. Tang Y, et al. Identification and validation of a prognostic model based on three MVI-related genes in hepatocellular carcinoma. Int J Biol Sci. 2022;18(1):261–75.
12. Kao KJ, et al. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer. 2011;11:143.
13. Gong J, et al. HCC subtypes based on the activity changes of immunologic and hallmark gene sets in tumor and nontumor tissues. Brief Bioinform. 2021; 22 (5).
14. Mayakonda A, et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28(11):1747–56.
15. Grabinski VF, Brawley OW. Disparities in breast cancer. Obstet Gynecol Clin North Am. 2022;49(1):149–65.

16. Ritchie ME, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7): e47.
17. Yu G, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.
18. Zhang M, et al. An immune-related signature predicts survival in patients with lung adenocarcinoma. Front Oncol. 2019;9:1314.
19. Liu TT, et al. Identification of CDK2-related immune forecast model and ceRNA in lung adenocarcinoma, a pan-cancer analysis. Front Cell Dev Biol. 2021;9: 682002.
20. Yan C, et al. System analysis based on the cuproptosis-related genes identifies LIPT1 as a novel therapy target for liver hepatocellular carcinoma. J Transl Med. 2022;20(1):452.
21. Chandrashekar DS, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia. 2017;19(8):649–58.
22. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
23. Liberzon A, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–25.
24. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
25. Anderson NM, Simon MC. The tumor microenvironment. Curr Biol. 2020;30(16):R921–5.
26. Frankel T, Lanfranca MP, Zou W. The role of tumor microenvironment in cancer immunotherapy. Adv Exp Med Biol. 2017;1036:51–64.
27. Yoshihara K, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
28. Chen B, et al. Profiling tumor infiltrating immune cells with CIBERSORT. Methods Mol Biol. 2018;1711:243–59.
29. Newman AM, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.
30. Charoentong P, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep. 2017;18(1):248–62.
31. Geeleher P, Cox N, Huang RS. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. PLoS ONE. 2014;9(9): e107468.
32. Li L, et al. A nomogram model to predict prognosis of patients with genitourinary sarcoma. Front Oncol. 2021;11: 656325.
33. Zhou B, Jin W. Visualization of single cell RNA-seq data using t-SNE in R. Methods Mol Biol. 2020;2117:159–67.
34. Aran D, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–72.
35. Jin S, et al. Inference and analysis of cell-cell communication using Cell Chat. Nat Commun. 2021;12(1):1088.
36. Smith J, et al. Promoter DNA hypermethylation and paradoxical gene activation. Trends Cancer. 2020;6(5):392–406.
37. Tuluhong D, et al. Squalene epoxidase promotes breast cancer progression by regulating CCNB1 protein stability. Exp Cell Res. 2023;433(1): 113805.
38. Kimbung S, et al. Impact of 27-hydroxylase (CYP27A1) and 27-hydroxycholesterol in breast cancer. Endocr Relat Cancer. 2017;24(7):339–49.
39. Chan HJ, et al. SERPINA1 is a direct estrogen receptor target gene and a predictor of survival in breast cancer patients. Oncotarget. 2015;6(28):25815–27.
40. Liu C, et al. ALDH1A1 activity in tumor-initiating cells remodels myeloid-derived suppressor cells to promote breast cancer progression. Cancer Res. 2021;81(23):5919–34.
41. Das M, Giannoudis A, Sharma V. The role of CPT1A as a biomarker of breast cancer progression: a bioinformatic approach. Sci Rep. 2022;12(1):16441.
42. Wang X, et al. Low expression of PLAT in breast cancer infers poor prognosis and high immune infiltrating level. Int J Gen Med. 2021;14:10213–24.
43. Wang J, Liu Y, Zhang S. Prognostic and immunological value of ATP6AP1 in breast cancer: implications for SARS-CoV-2. Aging (Albany NY). 2021;13(13):16904–21.
44. Sun X, et al. Comprehensive analysis of SLC35A2 in pan-cancer and validation of its role in breast cancer. J Inflamm Res. 2023;16:3381–98.